Innovative Customer Behavior Forecasting Framework for Subscription-based Organizations

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Leonidas St. Katelaris

Department of Digital Systems

University of Piraeus

A thesis submitted for the degree of

Doctor of Philosophy

Piraeus 2020

In memory of my beloved grandmother
Theodosia

Dedicated to
my wife, my daughter,
my parents, brothers, and sister,
for their love, endless support and encouragement

Table of Contents

Acknowledgments . . . x
Declaration . . . xi
Abstract . . . xiii

Chapter 1 1

1 Introduction 1
1.1 Introduction . . . 2
1.1.1 Customer Churn . . . 3
1.2 Research Aim and Objectives . . . 4
1.2.1 Research Aim . . . 4
1.2.2 Research Objectives . . . 4
1.3 Introduction to Research Methodology . . . 5
1.4 Thesis Outline . . . 5

Chapter 2 9

2 Systematic Literature Review 9
2.1 Introduction . . . 10
2.2 Customer Churn Causes . . . 10
2.3 The Importance of Customer Retention . . . 11
2.4 Systematic Literature Review on Customer Churn Prediction . . . 13
2.4.1 Stages of the Literature Review Process . . . 14
2.4.2 Organizing Literature Reviews . . . 16
2.4.3 Customer Churn Prediction Literature Review: Methodology and Plan 17
2.5 Customer Churn Prediction Literature Review Selected Cases . . . 25
2.5.1 Amin et al., [1, 4, 39] Research on Customer Churn Prediction . . . 28
2.5.2 Coussement et al., [3] Data Preparation Stage . . . 32
2.5.3 Verbeke et al., [40] Research Work on Data Mining Approach . . . 35
2.5.4 Kaya et al., [42] Prediction Model . . . 37
2.5.5 Vafeiadis et al., [43] Comparison Study on Machine Learning Methods 39
2.6 Findings of Literature Review on Customer Churn . . . 40
2.6.1 Similarities and Differences . . . 43

2.7 Conclusions 45

Chapter 3 **47**

3 Conceptual Framework Development **47**

3.1 Introduction 48

3.2 Research Challenges and Proposals 50

3.2.1 Research Issues Derived from Literature Review on Customer Churn in Subscription-based Domain 57

3.2.2 Proposed Solutions For Customer Churn Prediction 62

3.2.2.1 The Importance of Data Type 67

3.2.2.2 Considering Data Preparation Techniques in the Proposed Conceptual Framework 67

3.2.2.3 Churn Prediction Phase Elements 68

3.2.2.4 Outcome Interpretation 68

3.3 Proposed Customer Behavior Forecasting Framework for Subscription-based Domain 69

3.3.1 Analyzing Elements of The Proposed Customer Behavior Forecasting Framework 72

3.3.1.1 Proposed Business Case Analysis Element (E1) 72

3.3.1.2 Proposed Data Collection Element Analysis(E2) 73

3.3.1.3 Proposed Feature Selection Element Analysis (E3) 74

3.3.1.4 Proposed Prediction Model Element Analysis (E4) 75

3.3.1.5 Proposed Insights Element Analysis (E5) 76

3.4 Conclusions 77

Chapter 4 **79**

4 Research Methodology **79**

4.1 Introduction 80

4.2 Selecting Appropriate Research Approach for this thesis 81

4.2.1 Philosophical Perspectives 82

4.2.2 Qualitative vs. Quantitative Research Methods 84

4.2.3 Justifying Research Method Selected 86

4.3 Research Strategy 87

4.3.1 Case Study Research Strategy Selection 88

4.3.1.1 Single and Multiple Case Studies 89

4.4 Empirical Research Methodology 91

4.4.1 Research Design . . . 91
4.4.2 Data Collection . . . 93
4.4.2.1 Data Collection Tools . . . 93
4.4.3 Data Analysis . . . 96
4.4.3.1 Data Triangulation . . . 97
4.5 Case Study Protocol . . . 97
4.5.1 Case Study Overview . . . 98
4.5.2 Fieldwork Research Procedures . . . 99
4.5.3 Addressed Issues for Investigation and Output Format . . . 100
4.6 Conclusions . . . 100

Chapter 5 **102**

5 Empirical Data and Research Findings **102**
5.1 Introduction . . . 103
5.2 Case Study One: BlueTelco . . . 104
5.2.1 BlueTelco Background . . . 104
5.2.2 Customer Churn Prediction Actions in BlueTelco . . . 105
5.2.3 Description and Analysis of the BlueTelco Dataset . . . 119
5.2.4 Summary of BlueTelco Case Study . . . 122
5.3 Case Study Two: OrangeTelco . . . 125
5.3.1 OrangeTelco Background . . . 125
5.3.2 Customer Churn Prediction Actions in OrangeTelco . . . 125
5.3.3 Description and Analysis of the OrangeTelco Dataset . . . 127
5.3.4 Summary of OrangeTelco Case Study . . . 137
5.4 Conclusions . . . 140

Chapter 6 **142**

6 Revision of the Customer Behavior Forecasting Framework **142**
6.1 Introduction . . . 143
6.2 Lessons Learned from Case Studies . . . 144
6.3 Revised Customer Behavior Forecasting Framework . . . 145
6.4 Conclusions . . . 151

Chapter 7 **153**

7 Conclusions and Future Work 153
7.1 Research Overview . . . 154
7.2 Main Findings . . . 156
7.3 Meeting the Objectives of this Dissertation . . . 158
7.4 Novel Contribution . . . 159
7.5 Research Limitations . . . 162
7.6 Future Research Work . . . 163

Bibliography 165

Annexes 172

A Acronyms 172

B Previous Studies and Approaches on Churn Prediction 175
B.1 Churn Prediction Approaches . . . 176
B.1.1 Random Forest: . . . 176
B.1.2 Regression Analysis: . . . 177
B.1.3 Neural Networks: . . . 178
B.1.4 Support Vector Machines: . . . 178

C Queries Submission 179
C.1 IEEEXplore Keywords Submission . . . 180
C.2 SpringerLink Keywords Submission . . . 180
C.3 Google Scholar Keywords Submission . . . 181
C.4 Result Tables . . . 182

D Interview Agenda 250

E Εκτενής Περίληψη στα Ελληνικά 265

F Terminology and Notations used in Thesis 275
F.1 Terminologies Explanation used Throughout Thesis . . . 275

G Prediction Models - Explanation of Parameters 276
G.1 Description of Parameters used in Prediction Models Explanation . . . 276

List of Figures

1.1 Customer Retention vs Customer Acquisition . . . 3
1.2 Thesis Outline . . . 8

2.1 The probability of selling to existing and new customer . . . 13
2.2 Evidence-based Approach Steps in IS . . . 14
2.3 Systematic Literature Review Plan (Adapted from Brereton et al., [24] . . . 16
2.4 IEEEXplore Queries Union of Returning Results . . . 22
2.5 SpringerLink Queries Union of Returning Results . . . 22
2.6 Google Scholar Queries Union of Returning Results . . . 23
2.7 Results Union Between 3 Different Electronic Sources . . . 24
2.8 Customer churn prediction model major stages setup [Source Amin et al., [1]] 29
2.9 Oversampling Techniques in Amin et al., [4] Work . . . 31
2.10 Coussement et al., [3] Dataset Structure used for Conceptual Framework on Customer Churn . . . 33
2.11 Data Samples extracted from Main dataset of Financial Institution for Kaya et al.,[42] . . . 39
2.12 Observations Stemming From the Literature Review . . . 41

3.1 Mapping Chapter 2 Outcomes to Chapter 3 . . . 49
3.2 Abstract Level: Churn Prediction Procedure . . . 51
3.3 Input Data - Churn Prediction Procedure . . . 52
3.4 Prediction Churn - Churn Prediction Procedure . . . 52
3.5 Input Data - Sub-phases . . . 53
3.6 Use Data Type Constituents . . . 53
3.7 Use Data Preparation Techniques Structure . . . 54
3.8 Oversampling Techniques, Transformation Step and Representation Step Constituents . . . 54
3.9 Predict Churn - Churn Prediction Procedure . . . 55
3.10 Select Classification Algorithm Constituents . . . 56

3.11 Analyze Outcome - Churn Prediction Procedure 56
3.12 Evaluation Metrics and Interpret Prediction Results Constituents 57
3.13 Churn Prediction Procedure - Basic Structure 58
3.14 Classification Algorithms Usage in Research Works Studied 60
3.15 Data Types Elements 63
3.16 Merging Transformation and Representation Phases 64
3.17 Categorizing Classification Algorithms 65
3.18 Modified Version of Churn Prediction Procedure 66
3.19 Mapping Proposed Elements of the Conceptual Framework to Churn Prediction Procedure 69
3.20 Proposed Customer Behavior Forecasting Framework - Basic Elements 70
3.21 Proposed Business Case Analysis Element 73
3.22 Proposed Data Collection Element 74
3.23 Proposed Feature Selection Element Analysis 74
3.24 Proposed Prediction Model Element 75
3.25 Proposed Insight Element Analysis 76
3.26 Proposed Conceptual Customer Behavior Forecasting Framework 78

4.1 Research Methodology Components 80
4.2 Empirical Research Methodology Stages 92

5.1 Challenges faced by BlueTelco that affect Churn 105
5.2 Churners Classification at BlueTelco 107
5.3 Scenario 1: Customer Closes Subscription and Opens New Under the Same Account 108
5.4 Scenario 2: Customer Closes Subscription and a New Account is Created by Previous Subscription Owner 109
5.5 Methods and Systems used by BlueTelco to Predict Customer Churn 110
5.6 Ownership Status by Churner 113
5.7 Services by Churner 115
5.8 Proposed Customer Behavior Forecasting Framework (same as Figure 3.26) . 117
5.9 Values for Termination Reason Feature in the BlueTelco Dataset 121
5.10 OrangeTelco Churn Rate 128
5.11 Listing Churners by Demographic Features - Gender 129
5.12 Listing Churners by Demographic Features - Partnership and Dependents . . 130
5.13 Listing Churners by Service-pelated Features - Loyalty Period and Churn . . . 131

5.14 Listing Churners and non-churners by Financial / Profit-related Features - Contract Type . . . 132
5.15 Listing Churners and non-churners by Financial / Profit-related Features - Paperless Billing . . . 132
5.16 Prediction Models Performance on OrangeTelco Dataset . . . 136

6.1 Revised Business Case Analysis Element . . . 145
6.2 Revised Data Collection Element . . . 147
6.3 Revised Feature Selection Element . . . 147
6.4 Revised Prediction Model Element . . . 148
6.5 Revised Insights Element . . . 149
6.6 Revised Customer Behavior Forecasting Framework . . . 150

7.1 Novel Contribution . . . 160

B.1 Examples of Logistic and Linear Regression Graphs . . . 178

C.1 IEEEXplore Queries Union of Returning Results . . . 180
C.2 SpringerLink Queries Union of Returning Results . . . 181
C.3 Google Scholar Queries Union of Returning Results . . . 182

D.1 Modified Version of Churn Prediction Procedure . . . 257
D.2 Customer Behavior Forecasting Framework . . . 262

E.1 Προτεινόμενη Διαδικασία Πρόβλεψης Απώλειας Πελατών . . . 268
E.2 Προτεινόμενο Πλαίσιο Πρόβλεψης Απώλειας Πελατών . . . 269
E.3 Τελικό Πλαίσιο Προβλεψης Συμπεριφοράς Πελατών . . . 272

List of Tables

2.1 Taxonomy of the Literature Review Material [Source Cooper [25]] 17
2.2 Literature Review Research Findings 18
2.3 Research Questions for Systematic Literature Review 20
2.4 Literature Review Selected Cases 27
2.5 Evaluation Metrics used for Classifiers Performance 30
2.6 Homogeneous and Heterogeneous Ensemble Techniques 32
2.7 AUC and TDL Evaluation Metrics 34
2.8 DPT Framework from Coussement [3] Research Work 35
2.9 Classification Techniques Evaluated in Benchmark Study 36
2.10 Spatio-temporal Expenditure Patterns and Financial Choice Patterns 38
2.11 Research Works Studied: Aspects They Focused For Customer Churn Prediction 42
2.12 Aspects Studied in Churn Prediction Process 43
2.13 Similarities Identified in Studied Research Works 44
2.14 Differences Identified in Studied Research Works 45

4.1 Comparison of Qualitative and Quantitative Research Methods [Adapted from: [78, 79]] 85
4.2 Data Collection Tools Used Throughout This Research Work 94
4.3 Interview Agenda Overview 96
4.4 Questioning Levels adapted from Yin outline 98
4.5 Issues to be Addressed During Case Study investigation 100

5.1 BlueTelco Features in Dataset 119
5.2 BlueTelco's Churners Dataset – Unique Values per Feature 120
5.3 Clustering Churning Reasons from BlueTelco Churners Dataset 121
5.4 Findings from BlueTelco that Address Elements of the Proposed CBFF 123
5.5 OrangeTelco Dataset Features 126
5.6 Data Type Classification for OrangeTelco Features 127

5.7 Prediction Algorithms Evaluation Metric Scores on OrangeTelco 136
5.8 Findings from OrangeTelco that address elements of proposed CBFF 138

6.1 Main Findings from Case Studies Regarding CBFF Elements 146

7.1 Revisions Made on CBFF 156

A.1 Acronyms and Abbreviations 174

C.1 IEEEXplore Results 196
C.2 SpringerLink Results 219
C.3 Google Scholar Results 249

D.1 Systems for Customer Churn in Organization 256
D.2 Ongoing Projects on Customer Churn 256
D.3 Proposed procedures to be included in every churn prediction process 261

Acknowledgments

I express my sincere thanks to Professor Marinos Themistocleous, for his interest, valuable advice, and the continuous support he provided to me during my Ph.D studies.

Also, I would like to thank all of my Ph.D. Advisory Committee Professor George Vassilacopoulos and Associate Professor Dimosthenis Kyriazis, for their valuable scientific support throughout the research process, which contributed to its successful completion.

Finally, I would like to thank my wife, my parents, my brothers and sister, who believed in me and my choices, encouraging me to get to the end of this research journey.

Leonidas Katelaris

Declaration

Refereed Journal Papers / Book Chapters

Koumaditis, K., **Katelaris, L.**, and Themistocleous, M., 2015. 'A cloud based patient-centered e-Health record'. International Journal on Advances in Life Sciences, ISSN: 1942-2660, 7(1–2), 30–39

Katelaris, L. and Themistocleous, M., 2017. 'Predicting customer churn: Customer behavior forecasting for subscription-based organizations'. In Information Systems, Vol 299, ISBN: 978-3-030-11395-7, published by Springer International Publishing, Cham, Switzerland.

Katelaris, L., Themistocleous M., T. and Roberto, G., 2018. 'Mobile Number Portability Using a Reliable Cloud Database Appliance to Match Predictable Performance'. In Information Systems, Vol 341, ISBN: 978-3-030-44321-4, published by Springer International Publishing, Cham, Switzerland.

Katelaris, L., Themistocleous, M. and Roberto, G., 2019. 'CloudDBAppliance Database as a Service to Match Predictable Performance'. In Information Systems, Vol 381, ISBN: 978-3-030-11394-0, published by Springer International Publishing, Cham, Switzerland.

Refereed International Conferences

Katelaris, L., Themistocleous, M. and Koumaditis, K., 2014. 'Optimized Resource Provisioning based on SLAs in Cloud Computing'. In 11th European Mediterranean Middle Eastern Conference on Information Systems.

Pittas, G., Pittas, C., **Katelaris, L.**, 2014. 'Social media marketing and the consumer decision journey'. In Proceedings of the European, Mediterranean and Middle Eastern Conference on Information Systems, EMCIS 2014. European and Mediterranean Conference on Information Systems.

Katelaris, L., and Themistocleous, M. 2015. 'Smart E-Government: Predictive analytics in transforming cities into smart ones'. In Proceedings of the International Conferences on ICT, Society and Human Beings 2015, ICT 2015, Web Based Communities and Social Media 2015, WBC 2015 and Connected Smart Cities 2015, CSC 2015 - Part of the Multi Conference on Computer Science and Information Systems 2015 (pp. 231–234). IADIS.

Katelaris, L., and Themistocleous, M., 2016.' ICT-enabled governance innovation in public organizations'. In Proceedings of the 13th European, Mediterranean and Middle Eastern Conference on Information Systems.

Katelaris, L., Themistocleous, M., Brasca, F., 2019. 'Immediate mobility number portability: Cloud database appliance platform to provide central portability'. In CLOSER 2019 - Proceedings of the 9th International Conference on Cloud Computing and Services Science. SciTePress, pp. 621–624. doi:10.5220/0007877306210624

Journal Paper - Under Review

Katelaris, L., and Themistocleous, M., 'Proposing a customer churn prediction framework: Insights from a telco organization', In Journal of Enterprise Information Management.

Abstract

This thesis focuses on customer behavior forecasting and more specifically the customer churn faced by subscription-based organizations. Customer churn is a critical and challenging issue for subscription-based organizations in a rapidly growing and competitive market. Customer churn is inextricably linked with organizations prosperity and growth, as it is responsible for revenue and sales losses and potential negative impact to organizations due to the loss of customers. Accordingly, many organizations from the subscription-based domain conduct research resulting in the development and assessment of various methods, techniques, and solutions, that could improve churn prediction. However, there is a lack of customer churn prediction framework and a commonly accepted churn prediction procedure, that will support decision making process. Taking into account the above, this thesis focuses on: i) the proposition of a churn prediction procedure and ii) the synthesis and development of a customer churn prediction framework that can be used as a decision making tool, and may assists organizations in churn reduction.

Grounded, on the systematic literature review conducted in Chapter 2, a research gap was identified. In an attempt, to address this gap a churn prediction procedure and a customer churn prediction framework were conceptualized and proposed in Chapter 3. The proposed conceptualization, was tested in the practical arena using an interpretivism qualitative multiple case study strategy. Empirical data collected from two telecommunication subscription-based organizations, verified the proposed conceptualization and suggested minor additions. The latter, resulted in the revised customer churn prediction framework presented in Chapter 6. The research work carried in this thesis, contributes to the body of knowledge, by providing novel contribution and extending the literature. It also assist organizations to customer churn prediction accuracy and supports decision making process.

Chapter 1

Introduction

The best preparation for tomorrow is
doing your best today.

H. Jackson Brown, Jr. (1940-)

Summary

During the last years, various technologies have been used to forecast customer behavior in subscription-based organizations. Despite their, forecasting efforts, organizations failed to understand the real need of their subscribers, leading customers to churn from their service. The goal of this chapter is to introduce the area under study (customer churn), explain the research aim and objectives of this thesis and report the structure of this thesis. In doing so, it helps the reader to better understand the scope of this study and the outline of this thesis.

1.1 Introduction

Customer churn is the situation of shifting from one service provider to another [1], which is a critical and challenging issue affecting subscription-based organizations in today's rapidly growing and competitive market. According to Bingquan Huang et al., [2] churn is not a good situation for companies as it is associated with huge loss of provided services and thus it is a significant problem for the organizations. According to Coussement et al., [3], many companies suffer from churn due to market competition and its tremendous change throughout the last decade.

Customer churn affects organizations' growth with an impact on their overall performance, decrease in sales and revenue loses and a potential cause for negative impact on the organization's image in the market. The enormous digitization and the fast movement of the market to big data, shift to a subscription-based market and create the need for churn prevention best practices. Such practices will highly benefit organizations and assist them in decision making. This thesis studies this area and focuses on two κey issues: i) to investigate and analyze customer churn prediction methods, used in subscription-based organizations and ii) to emerge the effectiveness of customer behavior forecasting using state of the art techniques, that explicitly seek to reduce customer churn.

The first key issue arises from the need for improved customer behavior forecasting methods in subscription-based organizations. This is shown from related studies for its positive effects on organization's growth [4] and enhancement of customer churn prediction approaches. Retaining existing customers is more than five times less expensive than acquiring new, as it is illustrated in [4] Figure 1.1. The importance of customer churn forecasting in subscription-based oriented companies, has increased in recent years and many actions have been taken. During the last years, many subscription-based organizations focused on customer forecasting techniques in a way to predict customer churn. Moreover, many subscription-based organizations are seen to move to a more customer-oriented approach for retaining their customers and reduce churn [3]. Based on that, a common technique used by subscription-based organizations to forecast customer behavior is the analysis of historical data in Customer Relationship Management (CRM) systems [1].

The second key issue comes from the subscription-based organizations need to provide more accurate products/services according to their customers' needs. In that way, the usage of data analytic techniques will lead to retaining customers and reduced customer de-

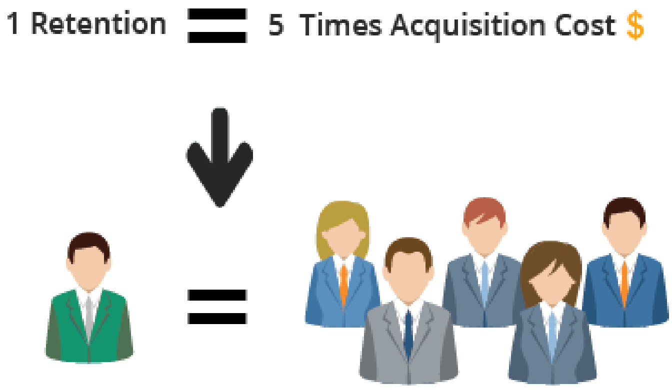


Figure 1.1: Customer Retention vs Customer Acquisition

fection. Customer behavior forecasting techniques have a positive impact on the overall performance and growth for subscription-based organizations. The normative literature reports various customer churn prediction approaches with positives and negatives. The author will analyze previous relevant approaches to better study the area in an attempt to propose more efficient ways for customer churn prediction. Possible new practices for customer behavior forecasting, arising thought this research work may help companies to understand their customers' needs better.

As mentioned above, the need for better understanding of customers' needs leads the efforts on forecasting techniques. The early signs of potential churn identified through the company's knowledge on its customers derived from their interactions across the company's channels, such as the customer service logs, store visits, purchase history, etc.

Throughout the thesis the author uses a notation to refer and describe specific roles and persons. In an attempt to eliminate misunderstandings and different meanings by the readers, the explanation of the notations is presented in Appendix 6.

1.1.1 Customer Churn

Many researchers including Amin et al., [4], and Huang et al.,[2] present customer churn as a significant problem affecting subscription-based organizations, which occurs when a customer stops collaborating with a company for another competitor. A customer shifting between providers is described as churn and is a well-known [5], challenging and critical problem faced by subscription-based companies during the last years [1, 4]. Churn rate can be described as a Key Performance Indicator (KPI) as formulated in equation 1.1, which it reports to the number of customers' lost during a specified period divided by the average total number of customers over that specific period [6].

$$ChurnRate = \frac{NumberofCustomersLost}{Timeframe} \tag{1.1}$$

As mentioned in Section 1.1, churn is inextricably linked with the company's prosperity and growth. The loss of customers leads subscription-based organizations to many negative consequences. Including the lower numbers in sales and negative impact on the organization's image [1, 7]. A customer-centric approach is used by many companies in the subscription-based domain to predict their customers' behavior, based on data stored in their CRM systems. The management and analysis of those data could give beneficial insights for possible churners. The literature reports three types of customer churns [1]:

- **Active churner (voluntary):** This category of churners, refers to the customers who decide to leave a provider for the next competitor in the market. There are multiple reasons for this, including low service experience, high cost, quality of service, etc.
- **Passive churner (non-voluntary):** This category describes churns that occur, due to the provider's action to discontinue the contract itself.
- **Rotational churner (incidental):** The last category of churners refers to both parties (company or customer), who may suddenly discontinue collaboration without prior notification. Reasons for this vary and may include financial problems, change of geographical location of the customer where there is no service from current provider.

1.2 Research Aim and Objectives

1.2.1 Research Aim

The aim of this thesis is to:

> *"Investigate customer churn in subscription-based organizations. In doing so, a conceptual framework will be developed, to enhance customer behavior forecasting and to be used as a decision-making tool for subscription-based organizations."*

1.2.2 Research Objectives

To reflect upon the research aim, the following research objectives are set:

Objective 1: To conduct a literature review (LR) in the area of customer behavior forecasting with a focus on customer churn prediction and critically evaluate literature that is relevant to customer churn in the subscription-based domain.

Objective 2: To develop a conceptual framework to support customer churn forecasting.

Objective 3: To examine established research methodologies, select an appropriate one for this research, and test the conceptual framework in the practical arena.

Objective 4: To extrapolate findings and provide a novel contribution around customer churn forecasting for the subscription-based organizations.

1.3 Introduction to Research Methodology

The study of the different aspects of customer churn in subscription-based domain, suggests the usage of an appropriate research methodology, which is able to investigate: i) how stakeholders of the domain understand the customer churn, ii) how the different techniques and methods used to predict churn rate achieved a decrease in churn rate and iii) which factors enhance churn prediction accuracy. Additionally, taking into consideration the attempt from the author to introduce a conceptual framework that aims to enhance churn prediction related actions and address open issues. To this end, the qualitative research methodology was adopted as it seem more suitable to explore the churn phenomenon.

Furthermore, the interpretivism epistemology was selected to investigate customer churn, which tries to understand phenomena based on participants' meanings instead of accepting one objective reality. Hence, a multiple case research design adopted for the empirical data collection and the evaluation of the proposed conceptual framework in a subscription-based domain. In doing so, multiple data collection tools deployed including among others: i) interviews, ii) documentation, iii) observation of participants, iv) physical artifacts and v) archival records. The various data sources used in this thesis alongside with the investigation of the phenomenon form different perspectives as part of the data triangulation approach reduced systemic biased and chance of associations.

1.4 Thesis Outline

This Ph.D. thesis's structure follows the methodology proposed by Phillips and Pugh [8] and consists of four elements as follows:

- **Background Theory:** Background theory focuses on the literature review for the area under study.
- **Focal Theory:** Refers to the proposition of the conceptual framework for this thesis.
- **Data Theory:** The third element is about the material and the sources of the research. It justifies the data collection method and the research design framework and it also covers the empirical data and its analysis.
- **Novel Contribution:** The last element underlies the significance of the current research and the novel contribution of this thesis in the research community.

This thesis is composed of seven chapters, each one of those providing findings and understanding of the critical points of this research. The thesis outline is illustrated in Figure 1.2, and synopsis of chapters' outline explained in the next paragraphs.

Chapter 1: Introduction - Background Theory

The first chapter, introduces the research problem under investigation. It presents customer churn issues faced by subscription-based domain organizations. It also presents the need for better understanding of customer behavior by subscription-based organizations to face their customers' needs, as well the aim objectives and the structure outline of this thesis.

Chapter 2: Systematic Literature Review - Background Theory

Chapter 2 provides the literature review, addressing the Objective 1 in Section 1.2.2. More specifically, the issues related to customer churn in subscription-based domain companies and how this impedes on their growth and prosperity are analyzed and discussed. Moreover, an analytical and critical review of the existing customer churn prediction research identified and highlighted.

Chapter 3: Conceptual Framework - Focal Theory

In Chapter 3, issues derived from literature review are taken into consideration and led to the development of the conceptual framework (Objective 2), of this research work. The proposed framework, namely "Customer Behavior Forecasting Framework", is described in detail, including elements that consist of the framework, through the sections of Chapter 3.

Chapter 4: Research Methodology - Data Theory

Chapter 4 describes the research methodology selected for this thesis, addressing Objective 3 of this research work. Chapter 3 focused on the development of the conceptualizations for customer churn, which was grounded on the critical literature review and research findings. This chapter describes the steps followed to justify the research methodology adopted for this research work. Among others, research strategy, research approach, and data collection methods used are presented and analyzed in detail throughout this chapter.

Chapter 5: Empirical Data and Research Findings - Data Theory

Chapter 5 focuses on Objective 4 and presents and analyses the empirical data gathered through the two case studies. In doing so, Chapter 5 tests proposed churn prediction procedure and the elements of the proposed Customer Behavior Forecasting Framework accordingly. Objective 4 is partially addressed through Chapter 5 and continues to next chapters.

Chapter 6: Revisions on Customer Behavior Forecasting Framework - Novel Contribution

The analysis of the empirical data presented in Chapter 5, denotes modifications in the proposed Customer Behavior Forecasting Framework, which are discussed and analyzed throughout this chapter. A revised version of proposed conceptualizations are introduced. Compared to the proposed framework presented in Chapter 3 (Figure 3.16, the revised framework of Chapter 6 (Figure 6.6, includes new tasks and guidelines in the five basic elements of the CBFF, which derived from the empirical data investigation made in Chapter 5.

Chapter 7: Conclusions and Future Work - Novel Contribution

The conclusions of this research are reported in this last chapter in accordance with Objective 4. Novel contribution of this research work as well as findings derived are reported and analyzed. More specifically the novelties derived from the development of the CBFF include among others the usage of the CBFF as a: i) decision making tool for churn prediction, ii) customer profiling tool that enhances customer retention, through the identification of customer needs, iii) guideline for the procedure that subscription-based organizations need to follow for churn prediction, and iv) countermeasure mechanism for churn, through the interpretation of the churn prediction outcome. Additionally, possible limitations of this research work are highlighted as well as, recommendations for future work.

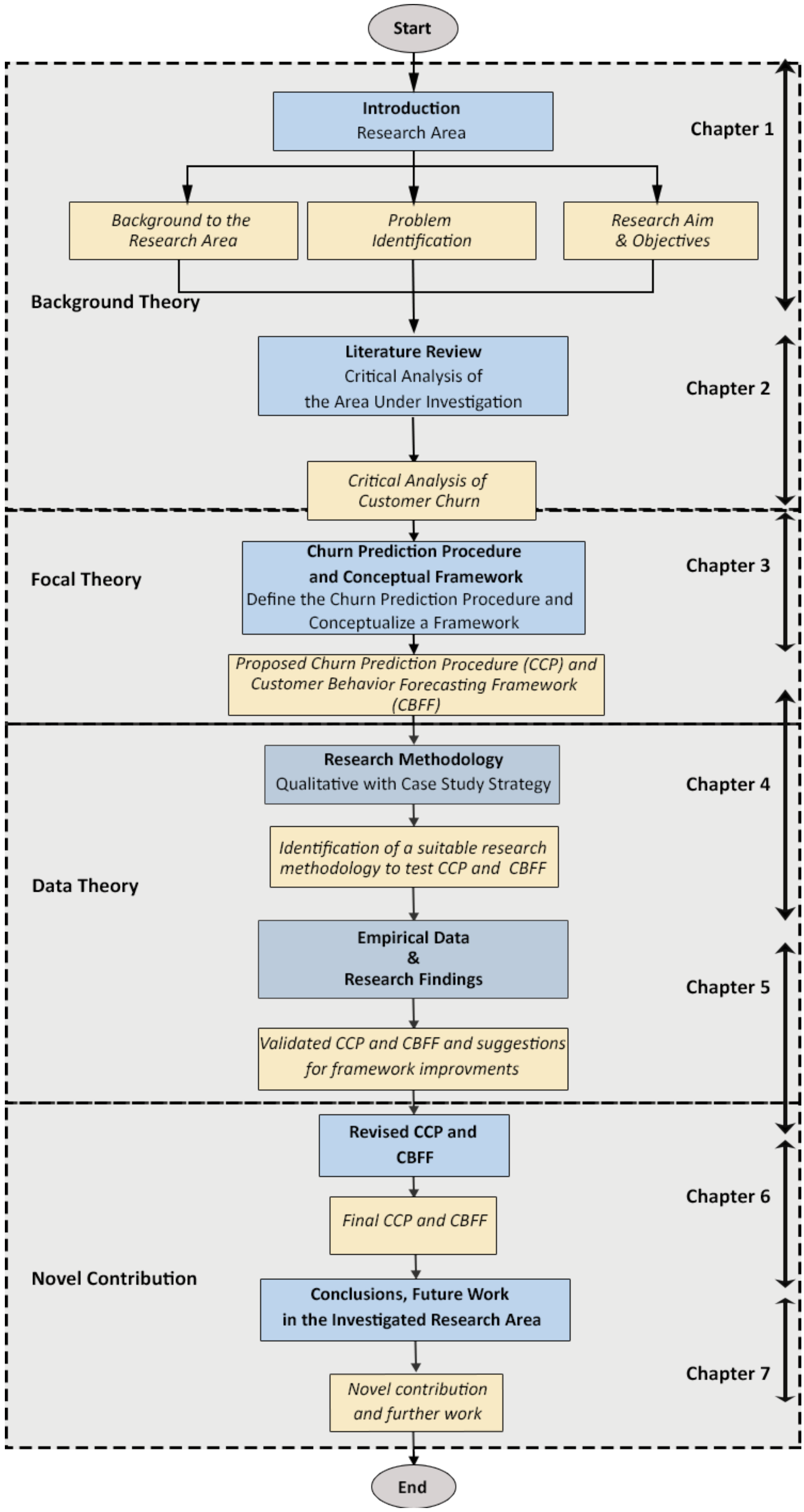


Figure 1.2: Thesis Outline

Chapter 2

Systematic Literature Review

Information is not knowledge

Albert Einstein (1879 -1955)

Summary

The previous chapter introduces the research problem, the aim, the objectives, and the outline of this Ph.D. thesis. In this chapter, a critical review and analysis of the literature are made on customer churn and forecasting methods used to predict customer behavior. The prediction of customer behavior could give meaningful insights for subscription-based organizations, helping them to retain their customers and prevent churn. The critical analysis conducted in this chapter, reveals a research gap, on the absence of a comprehensive proposal to face customer churn prediction for subscription-based organizations. Further, open issues identified in the literature review were also revealed in the area of customer churn, and covered in detail through Chapter 2.

2.1 Introduction

As stated in Chapter 1, customer churn is a crucial issue for subscription-based organizations, as the fail of forecasting is assorted with customer churn from the organization. During the past years, many attempts have been made to establish a method that predicts customers' behavior in a more accurate way. However, this issue remains under investigation for further study, which denotes plenty of space for research in this area. Throughout the next paragraphs, a systematic literature review is made covering issues around customer churn and churn prediction approaches used to prevent attrition from companies in the subscription-based domain, which is associated with negative consequences for them.

2.2 Customer Churn Causes

Churn is inextricably linked with the company's prosperity and growth. The loss of customers leads subscription-based organizations to many negative consequences, such as lower numbers in sales and negative impact on the organization's image. Reasons why customers leave a company varies, but those reasons which can be considered as most significant for churn are few. In an attempt to name those causes of customer churn, literature reports some, but the three reasons presented in detail below stand out as causes of churn [7, 9, 10].

- **Poor customer service - experience:** Studies [11, 12] highlight the importance of customer experience in the company and how that affects customer satisfaction rate. A poor customer service - experience creates unsatisfied customers who will probably abandon the company, with considerable damage to the company. Such damage could be the negative ratings from unsatisfied customers to others, including prospective customers. In an attempt to explain what customer service - experience means, we split it into two terms: i) customer service and ii) customer experience. Customer service describes a single transaction between the customer and the company, where company is obligated to offer a product / service to get the money from the customer.

  Customer experience includes all those aspects of customer interaction, such as the feelings, thoughts, and experiences from the customer. Such aspects of successful customer experience could be after-sales support, services warranty, 24/7 provision of information, etc. A qualitative customer service ideally is followed by rich customer experience. Customer service experience is connected with the company's knowledge about their customers' needs and expectations. The development of a customer-focused strategy, as this has gained the focus from companies [12], possibly could give meaningful insights in the way of delivering high-quality products/services to customers and interact against churn.

- **Unsuccessful customer on-boarding:** On-boarding is very important for organizational health and growth, as is where a company can make a positive impact on their customers. The customer life-cycle includes multiple touch-points for the company, but two could be considered as important milestones. The first is when a customer comes to the business and sings-up for the product /service, and the second milestone is when a customer achieves his first success with that specific product/service. For any company, a robust on-boarding is improving customer experience and fostering relationships. Various studies such as those from Alex Rawson, Ewan Duncan and Conor Jones [13] argue, that companies which give emphasis to customer experience enhancing their customer journey can gain benefits such as customer satisfaction, increased revenues and sales, higher retention and better collaboration inside company.

- **Weak customer relationship building:** Customer Relationship Management (CRM) is taken very seriously by companies throughout the years [11], as it could provide benefits for the company's processes, according to their customer's needs. In a highly competitive market as nowadays, building a strong relationship with customers is an essential tool against churn [13]. Throughout this process, companies could nurture a healthy relationship with their customers by receiving feedback from transactions between them and their customers. Such actions that lead to secure customer relationship building include i) business review conducting, ii) survey customers and iii) monitoring customer health, taking appropriate actions if customer health drops. According to that, possible churners could include those customers with rare communication with the company.

2.3 The Importance of Customer Retention

Subscription-based organizations which are driven by promotional actions like telecommunication service providers [1, 4, 14], banks [15, 16], online gaming industry [17] etc., are seen more vulnerable to churn, as the findings of this research report during the upcoming sections. Despite the different strategies and tactics used by them to attract new customers, they are all often impelled by insights from historical data they have collected in their CRM systems. Increased digitization of transactions, leads in quite big volumes of customer information stored in databases, where both business and academic practitioners leverage the benefits of data analytic techniques to mine useful correlations in the data for customer behavior.

While the above presented causes in Section 2.2 are leading churn, the normative literature reports three types of churners, namely are leading churn, the normative literature reports three types of churners, namely *active churners, passive churners and rotational churners* [1]. Most of the subscription-based organizations, which analyze data to define their customer behavior, seem to give priority on*active churners* and secondly on *passive churners* category. The *active churners* category seems to be more important for subscription-based organizations followed by *passive churners*, it includes those subscribers who churn for reasons that may be avoided if a trustworthy forecasting process is used in the organization. The third churner's category seems more difficult to predict, as, in that way a possible churn will be for unknown reasons or complicated to predict.

Customer retention has a quantifiable impact on revenue while acquiring customers is critical to business growth [3], makes even more sense on new companies in the subscription-based domain. Even more, an increase in retention has more positive effects on a company. Customer retention reduces explicitly churn increases revenue, and customer lifetime value (CLV) [18]. Many studies, including Kim [19] and Aksoy et al., [20] agree that is an essential attribute for subscription-based organizations, as customers seem to buy more from a company they trust. Likely, it has a positive impact on revenue for the company between 25% and 95%. Also, studies [21] show that selling service / product to existing customers is five and four times more likely to repurchase and to refer their friends respectively while selling to new customers is between 5-20% as seen in Figure 2.1. According to that, existing customers are 50% more likely to try new products and spend even 31% more compared to newcomers. Also, 89% of the companies highlight the importance of customer experience as a critical factor leading to retention rate increase [19–21].

An investigation of the market indicates that, there is a significant shift to what is called *the subscription economy* from the traditional pay-per product model. Subscription economy describes a new economy model where traditional pay-per product/service organizations are moving towards a subscription-based model. This is further highlighted by, Gartner's predictions for 2020 [22], which reports that more than 80% of software vendors will shift

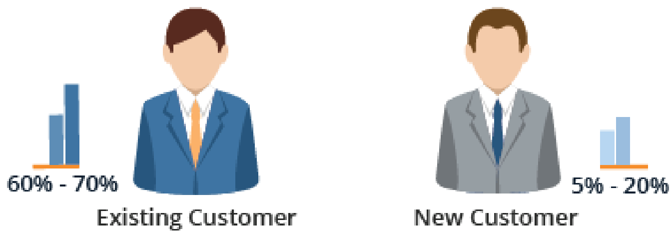


Figure 2.1: The probability of selling to existing and new customer

from traditional model to the subscription-based model. Also, International Data Corporation (IDC) predicts [23] for the same period that 50% of the world's leading companies will see the majority of their business on their ability to create digitally enhanced products, services and experiences. Taking into account the enormous shifting to the subscription economy, where the revenue of this market is estimated to billions of dollars annually, it seems that churn prevention is the key for growth and prosperity for the companies of the subscription-based economy domain.

Real-life examples to subscription economy shift, like Netflix, Spotify, Amazon, etc.,reveal that customer attrition is even more critical in their way to stay at the top of the competition. For those companies, churn translated into an annual loss of millions in revenues and even more expenses in customer acquisition campaigns. The reality is that getting new customers is an obvious win [4], but due to the associated high cost of acquiring new customers [21],many subscription-based organizations are working on how to retain existing customers. Prioritizing the above crucial issues, companies have to *retain existing customers* and to *attract new clients*.

2.4 Systematic Literature Review on Customer Churn Prediction

In an attempt to perform a sound literature review, the author adopts established systematic literature review approaches [24, 25], which are presented in detail in the upcoming paragraphs of this section. A systematic literature review is conducted for a variety of reasons, including the need for theoretical background for ongoing or future researches, deepening on a research topic of interest and providing research contribution on the research area [26]. The process for conducting a literature review, especially for novice researchers, could become a quite diligent work in an attempt to cope with continuously growing research studies in the area of interest [27].

2.4.1 Stages of the Literature Review Process

Various approaches and guidelines for conducting a rigorous and comprehensive literature review in Information Systems (IS) are met [24, 25, 27, 28]. Each approach describes the procedure to be followed by the researcher to conduct a meaningful literature review. In this regard, approaches including, those proposed by Brereton et al., [24] provide a plan for a systematic review process based on three phases as those presented in detail in the upcoming paragraphs.

Among others, Brereton's et al., [24] promote the use of evidence-based approach, which originated in medicine [29], but was also spread to other sciences including software engineering as raised by Kitchenham's et al., [30]. Kitchenham et al., [30] highlight the importance of evidence in software engineering and conclude, that the adoption and usage of techniques supported by evidence, improve the quality of software systems and help software engineers to adopt new techniques, based on scientific evidence. The steps from evidence-based practice in medicine have been redeveloped to meet evidence-based software engineering [30, 31]. Figure 2.2 presents the evidence-based approach steps in IS.

Figure 2.2: Evidence-based Approach Steps in IS

The five steps as seen in Figure 2.2 are: i) problem identification through the specification of answerable question, ii) best evidence finding which answers the question from the previous step, iii) critically evaluate evidence for its quality, iv) applying the evidence adopting it in practice and finally v) evaluating the performance of executing steps (i)-(iv), trying to improve possible gaps.The first three steps represent the systematic review of the literature, which meets one of the reasons for conducting a literature review, as those presented in Section 2.4. According to that, works included in the review presented as selected studies. Selected studies in this thesis are studies chosen to be in the main focus of this research work, which derived, based on the systematic literature review approaches adopted by this thesis. The author presents the selected studies for this specific research review in Section 2.5 of this thesis, based on Brereton's approach.

Based on the five steps of the evidence-based approach, Brereton et al., [24], stand out ten discrete activities, which grouped in three main phases as those presented in Figure 2.3.

The three phases for the systematic review process as those presented in Brereton's et al., [24] are:

Phase 1 - Planning Review:

- Purpose - Identify the research questions and goals for the review.
- Research protocol - Establish the methodology to be followed during the review.

Phase 2 - Conduct Review:

- Identify relevant research material - Searching the literature to identify relevant material for the review.
- Select use cases - State reasons for including or excluding studies from the review.
- Ensure quality of studies - Establish quality criteria for the selected studies to be included in the review.
- Extract data - Usage of data extraction forms for an accurate record of researcher outcomes.

Phase 3 - Reporting Review:

- Analyze data - Explain how the analyzed data answer the research question.
- Export review outcomes - Document the review outcomes in detail.

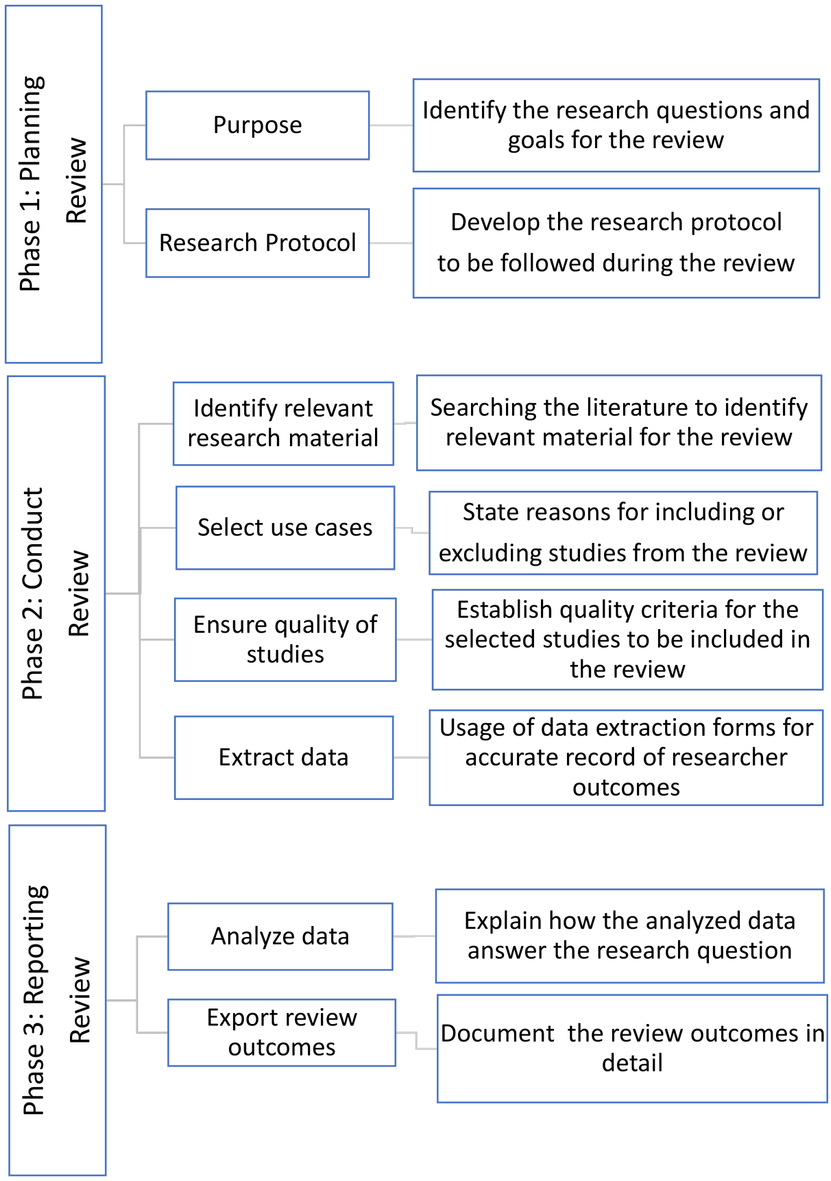


Figure 2.3: Systematic Literature Review Plan (Adapted from Brereton et al., [24]

2.4.2 Organizing Literature Reviews

In order to, conduct a rigorous and comprehensive literature review, a plan is needed for the literature review like the one provided by Brereton et al., [24], but also a research methodology approach to categorize literature reviews. Based on that, the author adopts Cooper's guidelines [25]. According to Cooper [25], a literature review has six basic characteristics namely: i) focus, ii) goals, iii) perspective, iv) coverage, v) organization and vi) audience. Each of the six characteristics is sub-divided into specific categories as presented in Table 2.1. The left column in the Table 2.1 shows the six characteristics based on Cooper's literature review [25], while the second the categories where each of the characteristic is divided.

Characteristics	Categories			
Focus	Research outcome	Research methods	Theories	Practices or applications
Goals	Integration	Criticism	Identification of central issues	
Perspective	Neutral representation		Espousal of position	
Coverage	Exhaustive	Exhaustive with selective citation	Representative	Central or pivotal
Organization	Historical	Conceptual	Methodological	
Audience	Specialized Scholars	General scholars	Practitioners or policy makers	General public

Table 2.1: Taxonomy of the Literature Review Material [Source Cooper [25]]

The first characteristic, namely Focus, concentrates on the issues that are of central interest to the reviewer. It could be sub-divided to i) research outcome, ii) research methods, iii) theories and iv) practices. Each one of these sub-categories presents the possible areas of focus for the review. The second characteristic Goals, concerns about the outcome of the research, which could be: i) integration, ii) criticism, iii) identification of central issues. The third characteristic namely Perspective addresses the impact of the reviewer's view on the literature, and it can be categorized as i) neutral representation and ii) espousal of position. Following, the fourth characteristic namely Coverage presents the way reviewers' select the suitable material and with quality material, which is categorized as i) Exhaustive, ii) Exhaustive with selective citation, iii) representative and iv) central or pivotal. Organization characteristic deals with the paper structure and arrangements like; i) historical, ii) conceptual and iii) methodological. The last characteristic namely Audience treats with the reviewer's indented audience, which can be categorized as i) specialized scholars, ii) general scholars, iii) practitioners or policymakers and iv) general public.

2.4.3 Customer Churn Prediction Literature Review: Methodology and Plan

Based on Cooper's methodology [25], the researcher can decide on the research material which will be included in the research. According to that researcher could focus on specific material as presented inTable 2.2.

As the purpose of this research is to investigate customer churn in subscription-based companies, the researcher focuses on published research, including frameworks, models, and previous attempts around customer churn prediction. According to that, the Focus of the literature review is to: i) identify proposed solutions in customer churn prediction addressing the issues raised, ii) identify methods or theories used to conduct previous customer churn approaches, and iii) identify gap and limitations of previously proposed churn prediction approaches. As a result, Table 2.2 summarizes the research findings for this literature review. The first column presents the database where the researcher looked for research material, while the second specifies the range for the published work. The third column presents the returning results based on the submitted *keywords queries*. The process followed to submit the *keyword queries* is presented in the upcoming sections of this chapter. Following is the fourth column of the table, which presents research material selected from the search results, based on their relevance to the research area of customer churn in the subscription-based domain. The last column (5^{th}) shows the final research material included in this research work.

Database	Range	Results	Considered	Selected
Google Scholar	2012-2018	298	14	4
SpringerLink	2012-2018	210	6	2
IEEExplore	2012-2018	110	8	2
Research Gate	2012-2018	40	8	2

Table 2.2: Literature Review Research Findings

Following, the second characteristic from Cooper's [25] literature review Goals, which concerns about the outcome of the research, could be the: i) identification of the main issues to be addressed in customer churn, ii) integrating them in the development of a common solution (e.g., model, framework) for predicting churn, iii) evaluate the proposed forecasting technique, as well as, iv) criticize findings and supply with challenges and points of interest around customer churn prediction.

The third characteristic, namely Perspective, distinguishes the point of view for the reviewers during the literature review. Perspective is neutral for this thesis, as the author tries to avoid bias in this research work, affecting free thinking. In doing so, the author attempts to represent in a fair way all opinions and evidence from research works cited in this work.

The fourth characteristic, Coverage, could be characterized as i) exhaustive with selective citation and ii) representative. The *exhaustive with selective citation* derives from the reviewer's choice to limit the works in a time-frame period of the last five years, for research material on approaches on customer churn. The explanation for choosing works from the last five years, it is based mainly on the tremendous development of technology (e.g., machine learning, big data, etc.) and secondly on the fact that, approaches used for churn predictions before that period, does not seem to be so essential or helpful in the development of new innovative churn prediction approaches [32]. The *representative*, which also characterizes this review, it is based on the selection for works that seem to be the representative of many other works around customer churn. The choice of limiting the period in the last five years for research works around customer churns, and the representatives' research work selection categorizes this research review, even though a comprehensive and rigorous coverage of the literature conducted.

The fifth characteristic Organization, which refers to the structure and arrangements of the research study, could be characterized as methodologically organized. The characterization as *methodologically organized* is based on the fact that the research works studied during this thesis present similar methods grouped, giving the reader a better view of aspects around customer churn prediction approaches.

The last characteristic Audience deals with indented audience. This literature review conducted with an emphasis on specialized scholars around customer churn prediction in subscription-based organizations, but also the writing style could be appreciated and welcomed from a broader range of readers.

Consequently with Cooper's [25] methodology to categorize literature reviews, Brereton et al., [24] provide a systematic literature review plan, which was adopted during this rigorous and comprehensive literature review. The following paragraphs give the literature research stages for this thesis:

Phase 1 - Planning Review:
Purpose - The purpose of phase is to define the Goals and Research Questions for the literature review.

- Goals: The goals which concern about the outcome of this research, extracted based on Cooper's [25] taxonomy characteristics for the review and are: i) identifying the main issues to be addressed in customer churn, ii) integrating them in the development of a standard solution (e.g., model, framework) for predicting churn, iii) evaluating the proposed forecasting technique and iv) criticizing findings and supply with challenges and points of interest around customer churn prediction.

- Research Questions (For the Systematic Literature Review): One of the most critical elements of a systematic review is the specification of the research question(s) [24]. Research questions are used in the development of search strings to extract the correct data from each *selected case*, limiting the results to aggregate. In addition, research questions are part of the research protocol, which is not subject to changes after submission. This literature review started with the following question:
  - **Research Question 1:** How does customer churn affect subscription-based organizations?
  - **Research Question 2:** What are the practices followed by subscription-based organizations to predict customer churn?

Research protocol - Another essential procedure of the literature review is the definition of the research protocol. The research protocol is described in a documented definition of the review process, based on Cooper's [25] approach to categorize literature reviews.

Research Questions	
Q1	How does customer churn affect subscription-based organizations?
Q2	What are the practices followed by subscription-based organizations to predict customer churn?

Table 2.3: Research Questions for Systematic Literature Review

Phase 2 - Conduct Review:

Identify relevant research material - The material studied include both academic and non-academic studies from various sources like: i) journal articles, ii) conference papers, iii) books, iv) white papers, v) websites, vi) workshops, vii) blogs, etc. While the international academic practice advice that, peer-reviewed articles like conference papers and journals are significant, white papers, websites blogs, etc., are also important and attract the author's attention. The importance to include also works from other sources rather than peer-reviewed articles derives from the fact that innovative ideas or approaches could be met in white papers first, rather than peer-reviewed articles. Based on that, peer-reviewed articles need more time to be published, as a result of the review process, which limits the existence of "fresh ideas". This inclusion helps the author to construct the concept of the research topic.

On the contrary, the combination of white papers, corporate websites, blogs, etc., alongside with the trustworthy of peer-reviewed articles, and more specifically, all sources except peer-reviewed articles highlight keywords, which used to extract all relevant material from academic sources. Furthermore, the author first attempts to extract all relevant material around the submitted keywords i) behavior forecasting, and ii) customer behavior forecasting. The results from the previous queries were referring to a specific keyword which, leads to the identification of a more relevant keyword around the domain of interest. The knowledge gained by the previous results guided the author in the construction of new keyword queries submitted in the researched databases like: i) "customer churn prediction", and ii) "churn prediction model/framework". Appendix 3 lists all results from submitted queries in the electronic data sources.

However, the submission of the above search queries in the electronic databases return thousands of results in one database (e.g., Google Scholar), while fewer results return from others(e.g., IEEExplore). The problem identified in the search engines of the databases, which are organized under different models. Because of the different models of the databases, the author was not able to submit the same set of search queries to each database. Therefore, the author constructed different queries based on the purpose and goals of this research described in *Purpose* section, to extract relevant material from the research databases. In doing so, the submitted queries for IEEEXplore database were:

- **Query 1:** *(((((customer churn prediction) AND "Abstract":customer churn prediction) OR "Author Keywords":customer churn prediction) OR "IEEE Terms":customer churn prediction) OR "Publication Title":customer churn prediction)* - Returned 80 results sorted by relevance.

- **Query 2:** *((((customer churn) AND prediction model) OR "Abstract":prediction model) AND "Abstract":customer churn)* - Returned 106 results sorted by relevance.

- **Query 3:** *(((((customer churn) AND prediction framework) OR "Abstract":prediction framework) AND "Abstract":customer churn))* - Returned 12 results sorted by relevance.

All the above queries submitted, with year filter range 2012 - present. Another point to be mentioned at this point is that the above queries resulted in 198 documents from IEEEXplore database with some duplicates. So, the author, to cope with this combined the queries and got the union of the result sets, which were 110 documents instead of 198 documents as presented in Figure 2.4.

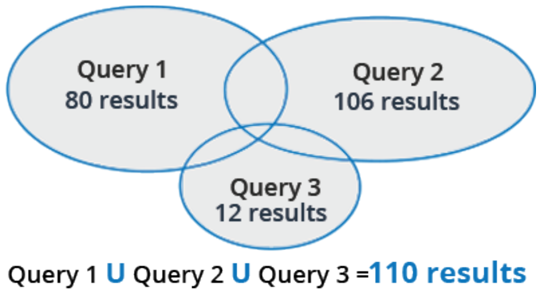


Figure 2.4: IEEEXplore Queries Union of Returning Results

Moving to the next electronic database (Springer Link), the author submitted the queries using advanced search option:

- **Query1:** *with the exact phrase "customer churn prediction"* - Returned 93 results sorted by relevance.
- **Query 2:** *customer AND churn AND "prediction model"* - Returned 173 results sorted by relevance.

All the above queries submitted, with year filter range 2012 - present. The author followed the same process as in IEEEXplore, to get the union of both result sets, which were 210 articles as presented in Figure 2.5.

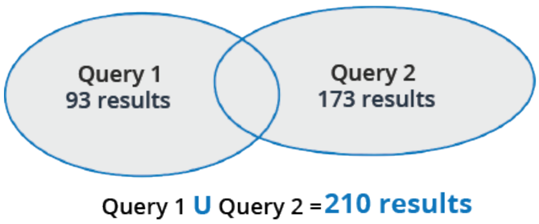


Figure 2.5: SpringerLink Queries Union of Returning Results

Following, the Google Scholar where author attempts to get any relevant article, which did not return from the other databases, as Google Scholar combines articles from many different electronic databases including among others SpringerLink and IEEEXplore. In addition to that, the author uses a software tool namely *Harzing's Publish or Perish* [33], which is listed as the free alternative software to Scopus and the Web of Science by thousands of libraries worldwide. Harzing's Publish or Perish toolkit gives the option to create more complex queries to be submitted in online databases, assigned with more metrics instead. As, the previous databases (e.g., Springer Link, IEEEXplore) have their advanced search tool, there is not an obvious benefit deriving using Harzing's Publish or Perish. In contrast, Google Scholar model on the advanced search mode does not provide such filtering options. In this way, the author strict the submitted queries to be more specific and minimize "irrelevant" results. In doing so, the author submitted the following queries:

- **Query 1:** *with the exact phrase (allintitle: "customer churn prediction")* - Returned 154 results sorted by number of citations.
- **Query 2:** *with the exact phrase ("customer churn prediction model")* - Returned 200 results sorted by number of citations.
- **Query 3:** *with the exact phrase ("customer churn prediction framework")* - Returned 3 results sorted by number of citations.

All the above queries were submitted, with year filter range 2012 - present. The first two queries have a property "allintititle", which means that the submitted keyword included in the title of the article. The author followed the same process as in previous electronic databases, to get the union of both result sets, which were 298 articles as presented in Figure 2.6.

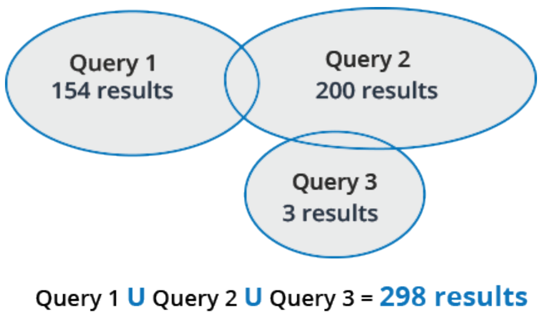


Figure 2.6: Google Scholar Queries Union of Returning Results

According to what already mention above, that Google Scholar combines articles from many different electronic databases, a new intersection between Google Scholar, IEEExplore, and SpringerLink is needed to remove duplicate articles in results. The author followed the same process as in previous electronic sources, to get the union of results, which were 548 articles, as presented in Figure 2.7.

Last but not least, another source that takes attention by the author is the Research Gate Network, which is a social networking site for researchers and scientists, where users can share research works, ask or answer questions on topics of interest. Research Gate considered as the most significant academic, social network in terms of active users, according to

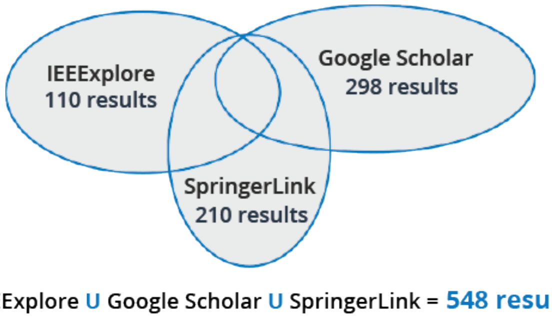


Figure 2.7: Results Union Between 3 Different Electronic Sources

studies by Nature [34] and Times Higher Education [35]. As a consequence, of the different nature (Research Gate is not an electronic data source) of Research Gate, the author is able to search directly for research works on a research topic from the authors, even unpublished works, or ask experienced researchers and scientists in the domain of customer churn. Through Research Gate, the author collected three articles, asking the authors directly.

The selected studies, as those presented in Table 2.2 are chosen based on a quality appraisal, which discussed in the following stage.

Select use cases - Based on the Purpose, of this research explained in detail in the previous phase (Phase 1 - Purpose), complying papers selected for further analysis. According to that, papers not included in this research are:

- Papers with limited scope - Regarding the focus of this literature review on specific area of customer churn such as customer churn in subscription-based organizations, papers as *"The gamma CUSUM chart method for online customer churn prediction"* [36], *"Customer Event History for Churn Prediction: How long is enough"* [37] were not included in this research, but they were used as supportive papers for the introduction to the research domain of churn prediction, due to their limited scope on research area or more recent works that overcome their findings.

- White papers and material from non-academic sources - During the first stages of this literature review, the author examined papers and material from a variety of sources. These sources, include white papers from organizations providing marketing strategies (e.g., marketing agencies) on customer attrition, agencies providing info-graphics, statistics and business reviews (e.g., advisory companies) around customer churn issues during the research period covered in this review (2012-today).

Despite the useful insights derived from those sources, the author used them more like supportive works to deepen in the research problem, rather than as a cornerstone for the development of the conceptual framework on customer churn.

- Articles in non-English language - Research material in different languages retrieved from electronic sources and examined, such as [38], but due to the lack of accurate translation excluded from literature review.

Ensure quality of studies - As of the international academia practice, the use of peer-reviewed work is the most appropriate way of judging academic publications. Based on the previous sentence, the author can make his contribution to advancing new concepts and insights with confidence based on published works of other researchers in the researched domain.

Extract data - The articles selected for further analysis were, used to extract related data on customer churn prediction.

Phase 3 - Reporting Review:
Analyze data - During this stage of literature research, the author attempt to extract knowledge stemming from research works studied using befitting qualitative and quantitative methodologies. Qualitative methods selected to examine each aspect of customer churn, including a theory that support approaches proposals (model/ framework). Quantitative methods used to identify the number of research works referring to each aspect of customer churn and deepen on common issues considerations made by the selected research works.

Export review outcomes - Sections 2.4.3 - 2.7 constitute the report of the literature review.

2.5 Customer Churn Prediction Literature Review Selected Cases

The selected cases for this research were chosen based on Brereton's et al., [24] methodology and listed in Table 2.4. The first column of Table 2.4 reports the authors conducted the research paper, the second column the title of the research paper and the third explains, in brief the focus of the research work. Based on *Phase 2- Identify relevant research material* step, where the extraction rules for the relevant research works presented.

Authors	Paper Title	Focus
Amin et al., [1]	Customer Churn prediction in telecommunication industry using the rough set approach	RST tool for mining hidden rules in the data that enhance decision-making techniques used in churn prediction approaches. Well-known rule generation techniques used to evaluate the performance of proposed predictive classifier that predicts churn behavior from collected knowledge
Amin et al., [4]	Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study	Address the imbalance problem of a dataset comparing six oversampling techniques, to conclude which of them is performing better when dealing with a two-class problem (i.e., applied in the proposed study)
Amin et al., [39]	Just-in-time customer churn prediction in the telecommunication sector	Just in time approach to improving the accuracy of traditional customer churn prediction models using the cross-company dataset in the telecommunication sector
Coussement et al., [3]	A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry	Benchmark eight state of the art data mining techniques to enhance the prediction performance emphasizing in the data preparation techniques used
Verbeke et al. [40]	New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach	Introduce a data mining approach based on profit to determine the best model to target customers with higher retention profit. Also, conducts extensive benchmarks experiment on various classification techniques
Verbeke et al., [41]	Social network analysis for customer churn prediction	Introduce multiple mining approaches for extracting information from social networks related to customer churn prediction and evaluate the significance of social network attributes in churn prediction models
Kaya et al., [42]	Behavioral attributes and financial churn prediction	Prediction model, which is based on dynamic behavioral attributes and evaluate the significant improvement in performance, when compared with traditional considered churn factors as demographics

Vafeiadis et al., [43]	A comparison of machine learning techniques for customer churn prediction	Compare popular ML methods used in customer churn in two phases. In first phase they evaluate using cross-validation while in second phase they compare a boosting version of each algorithm, based on Monte Carlo simulations to choose most efficient parameter combinations

Table 2.4: Literature Review Selected Cases

2.5.1 Amin et al., [1, 4, 39] Research on Customer Churn Prediction

Amin et al., [1] highlight the lack of efficient rule-based customer churn prediction approaches in the telecommunication domain. They conducted a literature review on customer churn classifying the customer churn categories (active churners, passive churners, rotational churners) as those explained in Section 1.1.1 of this thesis, according to the negative consequences churn associated with, for the subscription-based organization. They list previous approaches in customer churn prediction, presenting the techniques, a tool used for prediction, the records based on which the tool run, and the outcome of each of the approaches. According to their findings, they propose a benchmark empirical approach, which in their belief, behaves better in mining hidden rules in the data and after that are more efficient in predicting customer churns.

Mainly, the approach used by Amin et al., [1] focuses on a mathematical theory namely Rough Set Theory (RST) proposed by Pawlak [44] to be used in customer behavior forecasting. In addition, they based their theory on the fact that the churn dataset (class of interest) contains a fewer number of samples compared to the non-churn dataset (majority class). This particularity of the churn dataset (also called class imbalance problem) makes it difficult for the most ML techniques to retrieve the class of interest. In that way, the elimination of outliers could lead to better accuracy for churn prediction. RST theory state that information could be associated with every object of the universe of discussion. Also, in RST, there is a distinct conception of upper and lower estimations and a boundary region. The boundary region is the region, which separates the lower with upper estimations, for those objects that cannot be classified as in upper or lower bounds following their study on RST as the base classifier.

The Amin et al., [1] RST-based classification model consist of three main phases: i) Data Pre-processing, ii) Training Process and iii) Classification Process as those presented in Figure 2.8 and explained in detail below:

- Data preprocessing: During this stage, actions like the data preparation and feature selection are taking place. The result of data preparation and feature selection, used in the development of a decision table. Following is the decision table which splits the dataset into: i) discrete training set (70%) and ii) discrete testing data (30%) for the upcoming training and classification process.
- Training process: Algorithms for calculating rule sets namely: i) Evolutionary Algorithm, ii) Genetic Algorithm, iii) Covering Algorithm and iv) Learning from Examples

Modules version 2 explained in detail in the upcoming paragraphs are taking place. The result from the rule-set algorithms passed to rough set classification approach [1]. The result of this is a classification validation, which is passed into the validator of the next phase.

- Classification process: During this phase, a discrete testing data passed to the validator, which evaluates the prediction performance for the proposed approach in predicting customer churn, using the classification validation imported from the previous phase.

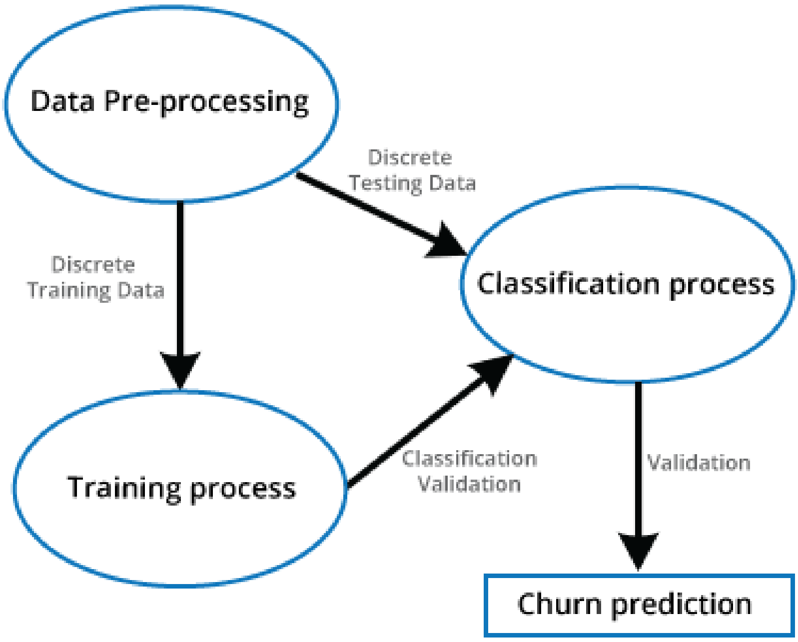


Figure 2.8: Customer churn prediction model major stages setup [Source Amin et al., [1]]

Amin et al., [1], used a benchmark methodology to evaluate their proposed classifier, by investigating four rules-generation algorithms namely:

- Evolutionary Algorithm (EA) [45]: This technique takes subsets of features over the initial population of features and returns minimal decision rules and the decreased set, based on the Boolean reasoning approach.
- Genetic Algorithm (GA) [46]: Adaptive search method based on the Darwin principle of natural selection. In regards to the rules, generation process of, work [1], it is used to decrease computational cost over a complex decision table.
- Covering Algorithm (CA) [45] Extracts the minimal set of rules which reflects on the initial set of objects.

- Learning from Examples Module version 2 (LEM2) Algorithm [1, 47]: This induction algorithm is based on the divide and conquer approach and returns a lower and upper approximation of RST.

Amin et al., [1] show that analysis on variables level has an essential impact on the churn model development and its success on customer retention strategy. Their work resulted in the fact that RST classification based on Genetic Algorithm (GA) performs better than the other three rules generation algorithms in terms of *precision, recall, accuracy*, and *F-measure*, etc., like those explained in Table 2.5.

Metric	Description
Precision(2.1)	The fraction of predicted positive instances among retrieved instances
Recall(2.2)	The fraction of predicted positive instances that have been retrieved over the total amount of relevant instances
Accuracy(2.3)	The fraction of the total number of correct predictions
F-measure(2.4)	Measure test accuracy considering both precision and recall, and it is used as a single metric for evaluating classifier performance

Table 2.5: Evaluation Metrics used for Classifiers Performance

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2.1}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2.2}$$

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + FalsePositives + TrueNegatives + FalseNegatives} \tag{2.3}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.4}$$

Furthermore, Amin et al., [4] conducted another work comparing six's oversampling techniques performance as those demonstrated in Figure 2.9, based on RST theory to address the Class Imbalance Problem (CIP). The six oversampling techniques tested were: i) Mega-Trend Diffusion Function (MTDF), Synthetic Minority Oversampling Technique (SMOTE), iii) Adaptive Synthetic Sampling Approach (ADASYN), iv) Majority Weighted Minority Oversampling Technique, v) Immune Centroids Oversampling Technique (ICOTE) and vi) Couples Top-N Reverse k-Nearest Neighbor (TRkNN). They examined their performance on four publicly available datasets from the telecommunication domain and concluded that MTDF performed better than the other compared oversampling techniques. As the core of this research is not the oversampling techniques, the researcher will not deepen in more detail in oversampling techniques.

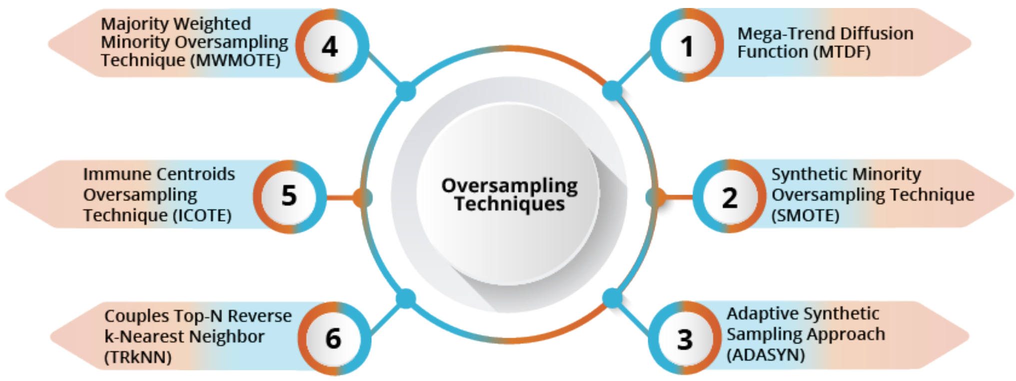


Figure 2.9: Oversampling Techniques in Amin et al., [4] Work

A more recent research work by Amin et al., [39] that summarizes previous studies on customer churn focuses on a different point of view on customer churn. In their recent work [39], they state the scenario of a company which is not having enough historical data due to either the few years in the market or possible loss in data. Their proposed approach namely Just In Time (JIT) [48], use the cross-company dataset for customer churn prediction in the telecommunication sector. The successful deployment of JIT methodology in various domains such as manufacturing industry [48], bug prediction in mobile applications [49], defect prediction [50], etc., make it suitable for newly established organizations in the competitive domain of telecommunications.

Amin et al., [39], study the relevance of a JIT approach in the telecommunication sector in case of a newly established telecommunication company where there are not enough customer behavior data to train models for churn prediction. For this reason, the JIT approach configured based on the data of an experienced telecommunication company to train dataset for the new company in the telecommunication sector. They choose SVM

(Appendix B - B.1.4 Support Vector Machines) as base classifier technique for their proposed JIT approach: i) due its ability to extrapolate well in high dimensional data, ii) the robustness in outliers, iii) rendering of global optimal solution and iv) modeling of nonlinear functional relationships. The data used for their empirical study derived from publicly available datasets, which have been used widely in the literature. Additionally, authors during this research work [39] conduct an empirical investigation and performance evaluation of homogeneous and heterogeneous ensemble models as described in Table 2.6, as ensemble methods seem to achieve better improvements in classification models when compared to single classifier [51].

Furthermore, they evaluate their prediction model performance with standard evaluation measures as those used in their previous research works above, i.e., precision, recall, and F-measure. Their outcomes indicate that using SVM as a base classifier in the heterogeneous ensemble perform better in general compared to SVM applied in the homogeneous ensemble. Although, findings from this research conclude that the use of JIT approach with cross-company data in customer churn prediction could be used in organization of the same sector (the authors evaluate their JIT model only in telecommunication sector) with significant accuracy when compared to traditional churn prediction models, solving at the same time the inefficiency in data for the newly established organizations.

Technique	**Description**
Homogeneous ensemble	A technique that uses a base classifier to achieve diversity through multiple loops
Heterogeneous ensemble	A technique that achieves diversity, in terms of matching classifiers, through the use of multiple arbitrary selection of algorithms

Table 2.6: Homogeneous and Heterogeneous Ensemble Techniques

2.5.2 Coussement et al., [3] Data Preparation Stage

The work presented by Coussement et al., [3], deal with the data preparation stage in customer churn prediction models, examining alternatives to enhance the prediction performance of prediction models. Coussement et al., [3], concluded to a conceptual data preparation treatment based framework, which benchmarks an optimized logit model compared to eight state of the art data mining techniques. To evaluate their findings, they use the residential database from a large European mobile telecommunication provider. The characteristics of the database, as presented in Figure 2.10 include 30,104 customers with churn ratio of 4.52%. As they have access to the internal data structures they have included 956 churn drivers as those are classified to: i) *156 categorical drivers* and ii) *800 continuous drivers*.

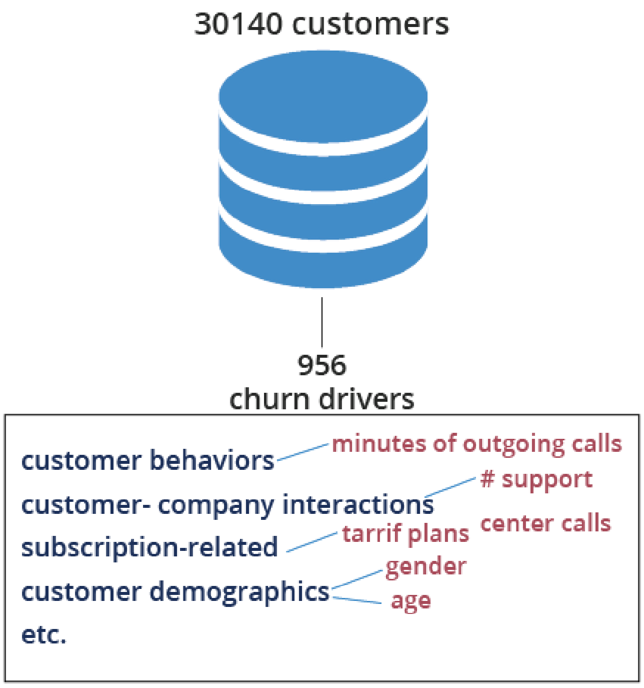


Figure 2.10: Coussement et al., [3] Dataset Structure used for Conceptual Framework on Customer Churn

Drivers included in the data set, such as customer behaviors, customer-company interaction variables, subscription-related variables, customer demographics, etc., are some of the categorical drivers. Consequently to the database from a large European mobile telecommunication provider, authors use evaluation metrics to calculate the performance of classification methods, such as the: i) *Area Under the Curve (AUC)* and ii) *Top Decile Lift (TDL)* (see Table 2.7). AUC is often met in studies around customer churn [40], likewise TDL [17].

Coussement et al., [3], emphasize in the data-preparation technique selected by analysts and state the importance in churn prediction performance it has. They focus on one stage of the established Cross-Industry Standard Process for Data Mining (CRISP-DM), which consists of six distinct stages, as those are: i) business understanding, ii) data understanding, data preprocessing, iv) modeling, v) evaluation and vi) deployment. Coussement et al., [3], give particular focus on one of the most time-consuming stages, the *data preprocessing* during their study. The data preprocessing stage, is the stage where the dataset variables are transformed to better fit in the predictive model. Data preprocessing stage comprised

Metric	**Description**
AUC	Provides an aggregate measurement of performance across all possible classification thresholds. AUC represents the probability that randomly chosen churner (random positive) is more probable than the probability of a randomly selected non-churner (random negative)
TDL	Metric that expresses the possibility of customers most possible to churn. The accuracy of the prediction Algeria is higher as TDL is higher

Table 2.7: AUC and TDL Evaluation Metrics

of two main processes: i) *Data reduction* and ii) *Data preparation*, as those presented in brief below:

- Data reduction: Data reduction techniques are used to select the most relevant variables, with the more significant meaning for separating churners and non-churners. After data reduction techniques are applied, the dimensionality of the dataset reduced.

- Data preparation: During this process, often variables are transformed into a form, that is supported by the classification algorithm to be used in the prediction model. For example, in the case of logistic regression, categorical variables cannot directly be used. For this reason, a transformation to a new form that fits logistic regression is made to the variables.

The DPT framework consists of data preparation steps: i) Transformation and ii) Representation. Transformation step subdivides into variables types of: i) Categorical and ii) Continuous. Both types of variables include their DPT methods. Categorical variables include: i) No Regrouping and ii) Decision Tree-based Remapping. Continuous variables include: i) Decision Tree-based Discretization, ii) Equal Frequency Discretization and iii) Equal Width Discretization. Second step,Representation, includes DPT methods such as: i) Dummy Coding, ii) Incidence Replacement and iii) Weight of Evidence Conversion methods.

All elements of the DPT framework from Coussement et al., [3] are presented in Table 2.8. During their study Coussement et al., [3], also referred to previous studies on customer churn prediction, summarizing them and derive to several conclusions, such as the one that shows that churn prediction literature tends to compare different churn prediction algorithms within the same paper. Besides, most of these studies acknowledge logistic regression as a benchmark algorithm [40]. Furthermore, Coussement et al., [3], give particular focus on both the academic and managerial perspective of data preparation treat-

Data Preparation Step	Variable Type	Description
Transformation	Categorical	No Regrouping
		Decision Tree-based Remapping
	Continuous	Decision Tree-based Discretization
		Equal Frequency Discretization
		Weight of Evidence Conversion
Representation		Dummy Coding
		Incidence Replacement
		Weight-of-Evidence Conversion

Table 2.8: DPT Framework from Coussement [3] Research Work

ment methods. From an academic perspective, the usage of non-proper data preparation treatment method during the development of novel classification algorithms on customer churn might fail to high- light their advantages over standard or straightforward methods, such as logistic regression. From the managerial perspective, the usage of proper data preparation treatment method, give the option to improve the accuracy of deployed churn prediction models without invading in their core procedures, alongside with the increase of customer retention management effectiveness, limiting costs of novel prediction solution implementations.

Coussement et al., [3] concluded to that Data Preparation Treatments (DTP) have an important effect on overall churn prediction performance, as those expressed as 14.5% in AUC and 34% in TDL metrics. Correspondingly, they stand out on the fact that few studies on churn prediction domain consider DTPs and that it has to be in the point of interest of data analysts, based on the results also stemming from this work that shows the essential positive effect on customer retention.

2.5.3 Verbeke et al., [40] Research Work on Data Mining Approach

Verbeke et al., [40] introduce a data mining approach based on profit to determine the best model to target customers with a higher retention profit. Also, during their research work, Verbeke et al., [40], conduct an extensive benchmarks experiment on various classification techniques evaluating them on eleven real-life datasets from telecommunication operators. In doing so, they introduce a *novel profit centric performance measure (maximum profit criterion)*, which calculates the profit of top-ranked customers, when the lasts included in a retention campaign.

The maximum profit criterion developed based on Neslin's et al., [52] formula, which calculates the profits generated by a retention campaign and defined in the equation 2.5 below, where **α** is the included fraction of customers and **Π** the profit generated by a single customer retention campaign. Additionally, moving to the classification techniques that are included in the benchmark experiments [40], those consisted of eight classification technique categories and total 19 classification techniques as presented in Table 2.9. The evaluation of performance measured using statistically-based performance measures, such as top decile lift and AUC, but also with the maximum profit criterion introduced by the researchers during this research work.

$$MP = \max_{a}(\Pi) \tag{2.5}$$

Classification Techniques Categories	**Classification Techniques**
Decision Trees	C4.5 Decision Tree, Classification and Regression Tree, Alternating Decision Tree
Ensemble Methods	Logistic Model Tree, Bagging, Boosting
Nearest Neighbors	k-Nearest Neighbor k=10, k-Nearest Neighbor k=100
Neural Networks	Radial Basis Function Network, Multilayer Perception
Rule Induction Techniques	RIPPER, PART
Statistical Classifiers	Logistic Regression, Naive Bayes, Bayesian Networks
SVM based Techniques	SVM with Linear Kernel, SVM with Radial Basis Function Kernel, LSSVM with Linear Kernel, LSSVM with Radial Basis Function Kernel

Table 2.9: Classification Techniques Evaluated in Benchmark Study

Findings stem out their work [40] include: i) strong indications on the usage of maximum profit criterion in generating profits from a retention campaign, ii) the oversampling technique in the dataset does not always lead to positive effect in performance as it depends strongly on dataset and classification technique, iii) based on the previous finding classification technique has a significant impact on predictive accuracy of the resulting model and iv) the quality in data is less expensive (in economic point of view) than take into account all possible attributes around a customer to predict churn.

Verbeke et al., [41] search for ways to mine churn associated information through social media networks and examine their effect on customer churn prediction models [41]. Their goal was to identify the use and value of social network information for customer churn prediction in the telecommunication domain. Therefore, they introduced multiple data approaches for data mining social network information for churn prediction, examining their value in customer churn prediction models and thus to improve the efficiency of retention campaigns conducted by organizations to prevent churn of their customers. The outcome of Verbeke et al., [41] research work saw a general framework for customer churn prediction for the telecommunication sector.

Verbeke et al., [40] test the performance of newly proposed techniques, based on two real-life case studies on large scale telecommunication sector datasets. The datasets used by the authors to evaluate their proposed techniques consisted of both networked (call detail logs) and non-networked(customer-related) attributes for millions of customers. Findings indicate a significant impact of social network on churn behavior on telecommunication company customers. Moreover, authors inspired by the findings of the higher-order, non-Markovian networks effects on the health of individuals [53], observe the impact on churn behavior not only on friends of the telecommunication company customers but also on friends of friends. Another finding is those non-relational classifiers used with network and non-networked information lead to the detection of different groups or types of churners, intimating the beginning of research studies around social network analysis for churn prediction [41].

2.5.4 Kaya et al., [42] Prediction Model

Kaya et al., [42], attempt to fill the gap between the dynamic behavioral pattern of clients, their financial decisions, and churning activities. Kaya et al., [42], during this research work, propose a prediction model, which is based on dynamic behavioral attributes and evaluate the significant improvement in performance, when compared with traditionally considered churn factors as demographics. The main contributions of their research work could be summarized as:

- Spatio-temporal and choice features are performing quite better than demographic features in financial churn prediction.
- Churn prediction performance of younger people is higher when using Spatio-temporal and choice features than demographic features.
- Introduce the entropy of choice features by analyzing the relative performance of each behavioral feature.

- Introduce novel financial data and churn definitions.

In contrast, to the most research works in the area of churn prediction, Kaya et al., [42], focus on behavioral patterns (Spatio-temporal, financial choice) as presented in Table 2.10 than on demographics (e.g., gender, marital status, job type, income, educational status, age etc.). They use anonymized data samples (Sample A, Sample B) as presented in Figure 2.11 from a significant financial institution in an Organization for Economic Co-operation and Development (OECD) country, which includes demographic information, credit card transactions, money transfers and electronic fund transfers of 160 thousand customers and roughly 77 millions of transactions (Sample A - 45M, Sample B - 22M). The samples are part of a larger dataset which consisted of 450 thousand customers located in a major metropolitan city.

<table>
<tr><th colspan="5">Spatio-temporal and Financial Choice Patterns</th></tr>
<tr><td>Spatio-temporal Expenditure</td><td colspan="2">Loyalty</td><td>Diversity</td><td>Regularity</td></tr>
<tr><td>Financial Choice</td><td>Money Transfer Entropy (transee)</td><td>EFT Entropy (efte)</td><td>Entropy of credit Card Transactions with Respect to Merchants (ecctmer) and Merchant Category Code (MCC)</td><td>Entropy of Offline Credit Card Transactions with Respect to Merchants (efcctmer) and MCC (efcctmcc)</td></tr>
</table>

Table 2.10: Spatio-temporal Expenditure Patterns and Financial Choice Patterns

For their study, Kaya et al., [42], adopt Random Forest (Appendix B - B.1.1 Random Forest) as the classification training method with 500 trees during training and maximum two features per tree. The model's evaluation is based on 8-fold cross-validations, where almost the same proportions of churners and non-churners involve in each iteration. As the high level of imbalance of labels (churners, non-churners), authors use the area under ROC (ROC is the application of AUC metric) curve metric (AUROC).

Also, researchers [42] noted that the AUROC could lead to a selection of a less optimal prediction model due to delusive assumptions made by the evaluation metric on wrong classified costs. However, all data sets engaged in their research study showed significantly better performance for Spatio-temporal behavioral features than the demographic features in terms of area under the ROC curve metric. More specifically, in a test dataset, that is not biased towards credit card users, and longer observation window, the AUROC for predicting churners scored 77.9% instead of 51.3% of the demographic model. The previous outcomes come to support the author's assumption made during the early stages of literature review, that churners' behavior could be predicted i a large extend by analyzing dynamic behavioral patterns followed by customers in both temporal and spatial domains. Additionally, authors express their hope for future research works to compare Spatio-temporal features adopted in this paper with traditional features used for customer churn prediction, that may lead to even better prediction models.

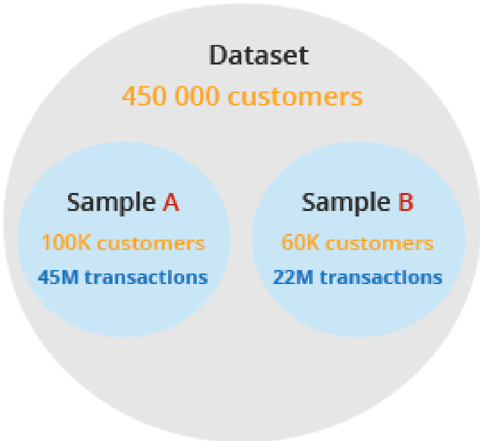


Figure 2.11: Data Samples extracted from Main dataset of Financial Institution for Kaya et al.,[42]

2.5.5 Vafeiadis et al., [43] Comparison Study on Machine Learning Methods

Vafeiadis et al., [43] compare popular machine learning methods applied in customer churn such as: i) Artificial Neural Network, ii) Support Vector Machines, iii) Naive Bayes, iv) Decision Trees Learning and v) Regression Analysis - Logistic Regression Analysis using Monte Carlo simulations [54] for every method on a wide range of variables, testing boosted versions of each churn prediction model.

They [43] show that boosted versions of machine learning methods applied in customer churn perform better than their regular versions. To evaluate their measures, they [42]use the measures of *precision*, *recall*, *accuracy* and *F-measure* as those explained in Table 2.5, over a popular public domain dataset. The dataset used to evaluate findings include variables related to service usage such as the number of months active user, the total charge of evening calls, area code, total minutes of night calls, international plan, the total charge of night calls, etc.

Vafeiadis et al., [43] main goal for using boosting algorithm, is the classification performance improvement deriving from the mixture of multiple decisions from the different classification models, as previously achieved when boosting used with success in retail [55], where improvements in classification performance met through multiple combinations of decision from many classification models. During their work, Vafeiadis et al., [43], measure the effectiveness of boosting on the improvement of *F-measure*. Weak classifiers are combined to build a stronger classifier in the train set. The boosting algorithm assigns weights to each training pattern resulting in patterns with a different contribution to the final training error and high accuracy in the test set [56]. The form of boost algorithm the authors choose for their study is a popular one called AdaBoost *AdaBoost* [57], where weak classifiers are decision trees. More specifically, they [43] use the *AdaBoost.M1*, where Decision Trees and Back-Propagation Network algorithm [58] used as weak classifiers. Hundreds of Monte Carlo simulations, based on different parameters for each classifier generated for cross-validation of the results.

The outcome of their work was an overall improvement to three of the classifiers due to boosting. The three classifiers that achieved improvements due boosting are i) *Back - Propagation Network* (popular model of artificial neural networks), ii) *Decision Trees* and iii) *Support Vector Machines* (Gaussian Radial Basis (SVM-RBF) and Polynomial (SVM-POLY) kernels used). The remaining two classifiers, *Naive Bayes* and *Logistic Regression*, due to the lack of free parameters to be tuned, they were not able to be boosted. The Boosting algorithm achieved improvements in *Accuracy* measurement between 1% and 4% and for *F-measure* between 4.5% and 15%.

2.6 Findings of Literature Review on Customer Churn

The review of the normative literature reported in Sections 2.1 - 2.5 lead to the following observations as those illustrated in Figure 2.12.

- **Observation A:** *Customer churn in subscription-based organizations is important and vital for the organizations' health and growth.* This observation is also supported by the vast number of articles extracted (i.e., 548 - see Figure 2.7) from the electronic databases during the *Phase 2 - Conduct Review* described in the previous section of the literature review.

- **Observation B:** Although, in the literature,there was an attempt from various researchers to improve customer churn prediction methods, It seems that, *there is an absence of a work (e.g., framework) that integrates the improvements proposed by all these researchers*. Clearly, this is an important research gap. Thus, there is a need for a customer churn prediction framework, that integrates all significant research works, synthesizes and orchestrates them in an efficient way that will assist organizations to predict customer churn. This observation is not only supported by the selected cases [1, 3, 4, 39–43], that are in the main focus of this research but also by the vast majority of the research works studied during this research work.

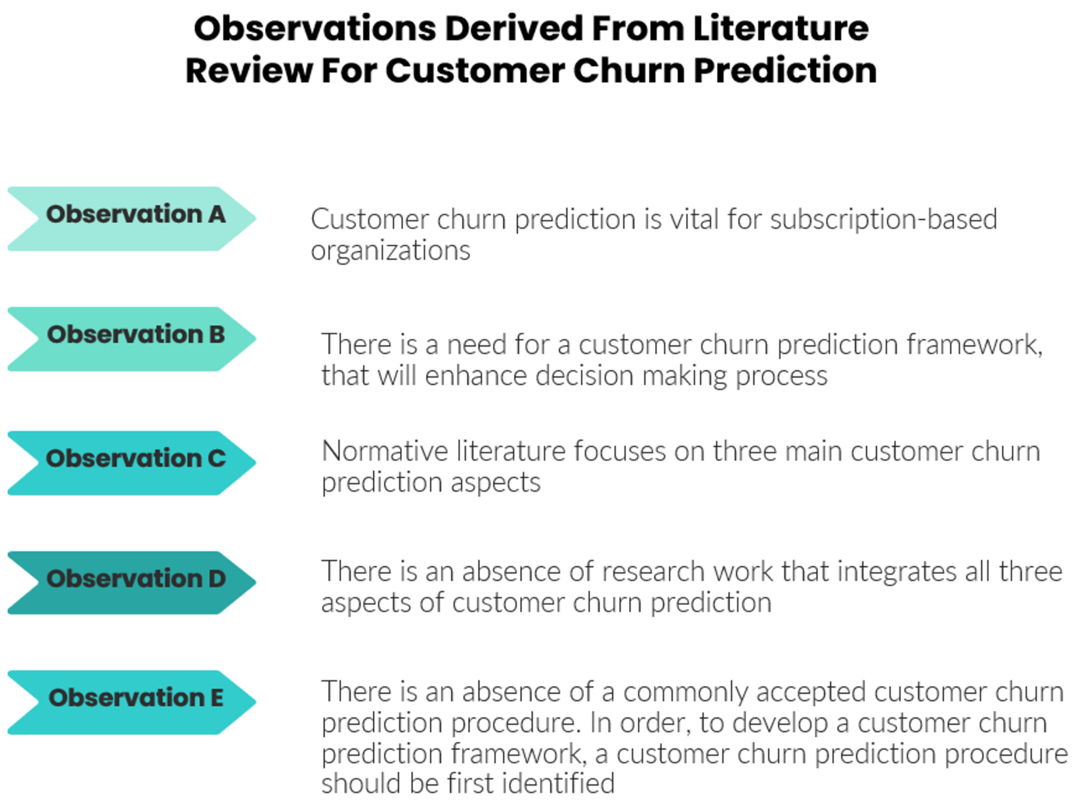


Figure 2.12: Observations Stemming From the Literature Review

- **Observation C:** A critical analysis of normative literature reveals that various aspects affect customer churn prediction, with the author classifying them into three main categories:
  - Aspect I - Data Preparation Techniques: These techniques include: i) attributes selection, ii) attributes balance and iii) sampling techniques.
  - Aspect II - Classification Algorithm Selection: Research works studied, benchmark well-known classification techniques such as, Logistic Regression, Neural Networks, Random Forest, Naive Bayes SVM, etc., to find the best that fits for a current dataset or benchmark the performance. Also, benchmark modified versions of the well-known classification techniques on the same dataset for performance.

  - Aspect III - Data Type for Churn Prediction Model: This refers to works that denoted the usage of specific data types in the improvement of churn prediction accuracy [1, 41, 42].

- **Observation D:** Based on Observation C, the improvements made by various researchers on customer churn prediction methods can be classified as illustrated in Table 2.11. Clearly, there is not an integrated work (e.g., framework) that covers all three aspects of customer churn prediction.

Research Works Studied	Aspect I: Data Preparation Techniques	Aspect II: Classification Algorithm Selection	Aspect III: Data Type for Churn Prediction Model
Amin et al., [1]		✓	
Amin et al., [4]	✓		
Amin et al., [39]			✓
Coussement et al., [3]	✓	✓	
Verbeke et al., [40]		✓	✓
Verbeke et al., [41]			✓
Kaya et al., [42]			✓
Vafeiadis et al., [43]		✓	

Table 2.11: Research Works Studied: Aspects They Focused For Customer Churn Prediction

- **Observation E:** To develop a customer churn prediction framework, a customer churn prediction procedure should be identified. The literature review conducted in this chapter indicate that there is not a commonly accepted as customer churn procedure. Instead of that, only parts of a customer churn procedure are identified in the studied research works. Clearly, this is another important research findings for further investigation.

Based on the findings derived from the literature review regarding the different aspects studied by various researchers Table 2.12 was developed to report all the techniques, methods, tools etc., identified in the literature.

Table 2.12, provides further justification and supports Observations A, B, C, D and E. A detailed review of the literature reveals the absence of an integrated customer churn prediction procedure. Also, Table 2.12, illustrates that published research works focus not only on a different aspect but also on different tools, techniques methods etc.

Research Works	Aspect I: Data Preparation Techniques	Aspect II: Classification Algorithm Selection	Aspect III: Data Type for Churn Prediction
Amin et al., [1]		RST Tool	
Amin et al., [4]	MTDF, SMOTE, ADASYN, Majority, Weighted Minority Oversampling Technique, ICOTE, TRkNN		
Amin et al., [39]			Cross-Company Data
Coussement et al., [3]	No Regrouping, Decision Tree-based Remapping, Decision Tree-based Discretization, Equal Frequency Discretization, Equal Width Discretization, Dummy Coding, Incidence Replacement, Weight-of-Evidence Conversion	Bagged CART, Bayesian Network, J4.8 Decision Tree, Multilayer Perception Neural Network, Naive Bayes, Random Forest, Radial Basis Kernel SVM, Stochastic Gradient Boosting	
Verbeke et al., [40]		Decision Trees, Ensemble Methods, Nearest Neighbours, Neural Networks, Rule Induction Techniques, Statistical Classifiers, SVM with Linear Kernel, SVM with Radial Basis Function Kernel, LSSVM with Linear Kernel, LSSVM with Radial Basis Function Kernel	Profit-related Variables
Verbeke et al., [41]			Call-detail Logs, Customer Related Attributes
Kaya et al., [42]			Spatio-Temporal Behavioral Data
Vafeiadis et al., [43]		Multi-layer Perception Artificial Neural Network, Gaussian Radial Basis Kernel SVM (SMV-RBF), Polynomial Kernel SVM (SMVPOLY), Naive Bayes, C5.0 Decision Tree, Logistic Regression	

Table 2.12: Aspects Studied in Churn Prediction Process

2.6.1 Similarities and Differences

Section 2.6, critically assesses the normative literature and presents the main outcomes as those reflected in the observations presented in Figure 2.12. In this section, a comparison of the research works studied in this chapter is presented to identify similarities and differences with Table 2.13 presenting the findings.

Regarding similarities, the majority of the research works [1, 3, 40–42], mention the importance of analysis on variables level and the benefits derived from this analysis in retention actions taken by the subscription-based organizations. Likewise, the critical role of data type, impacting the accuracy of the churn prediction, is another similarity identified in literature [3, 39, 41, 42], which addresses open issues regarding the usage of specific data type in churn prediction to enhance accuracy of the prediction.

In addition, many of the studied research works [1, 3, 4, 41, 43], are referring to evaluation metrics to calculate the performance of their proposals (e.g, models, methods etc.). Besides, the comparison of well-known classification algorithms was in the focus of some

	Research Works							
Similarities	[4]	[3]	[1]	[39]	[40]	[41]	[42]	[43]
Highlight the dataset imbalance problem; the significance of the data preparation stage on churn prediction	✓	✓			✓			
Data type enhance churn prediction accuracy (cross company dataset, mining social media networks, dynamic behavioral attributes)				✓		✓	✓	
Analysis of the variable level has an essential impact on the churn model development and customer retention strategy		✓	✓		✓		✓	
Usage of evaluation metrics to measure the performance of the proposed solution	✓	✓	✓	✓	✓			✓
Benchmark well-known classification techniques					✓			✓

Table 2.13: Similarities Identified in Studied Research Works

of the studied research works [40, 43]. Last but not least, is the similarity revealed from research works [3, 4], that highlights the significant impact of dataset imbalance problem, on the selection of appropriate features; variables for churn prediction.

Furthermore, the systematic literature review, identified differences as those are presented in Table 2.14. The first finding derives from Amin et al., [1] which despite the improvements they achieved in customer churn prediction using oversampling techniques in accordance to their proposed RST tool to balance the dataset variables, denote the lack of efficient rule-based customer churn approach. On the other hand, Verbeke et al., [40], assume that oversampling techniques on the dataset does not always lead to positive effect in the performance of churn prediction, as it depends on both the dataset and the classification algorithm used. Another research work that comes form Coussement et al., [3], differentiates their position on what is an open issues for customer churn prediction, highlighting the lack of research on data preparation techniques used in churn prediction. More specifically, Coussement et al., [3], denote that despite the increased interested using Machine Learning (ML) methods to achieve better accuracy in churn prediction, this interest is quite lower regarding data preparation, which seems to have significant impact on churn prediction.

Differences	
Research Work	**Difference**
Amin et al., [1]	1. Highlight the lack of efficient rule-based customer churn approaches
Verbeke et al., [40]	1. Denote that, oversampling techniques in the dataset does not always lead to positive effects in churn prediction
Coussement et al., [3]	1. Denote the lack of research in data preparation techniques used in churn prediction
Kaya et al., [42]	1. Compare dynamic behavior attributes with traditional demographic attributes attempting to increase churn prediction accuracy
Vafeiadis et., al [43]	1. Modification of well-known algorithms and usage of boosting versions to predict customer churn

Table 2.14: Differences Identified in Studied Research Works

Kaya et al., [42], research work adds their different perspective on what they assume it will enhance churn prediction. Kaya et al., [42] research work highlight the gap on churn prediction approaches on the usage of "traditional" demographic (e.g., age, sex, address etc.,) variables. In that manner, they proposed the comparison of dynamic behavioral attributes with traditional demographic attributes to enhance churn prediction accuracy. A different perspective on that it could enhance churn prediction outcome is presented by Vafeiadis et al., [43]. The research work of [43], compare boosted version of popular ML methods to achieve better results in churn prediction. More specifically, during their research in order to determine the combinations of the most efficient features, they executed a series of Monte Carlo simulations for each classification algorithm under study, with a wide variety of parameters accordingly. Findings from Vafeiadis et al., [43], showed that in many cases, boosted version of classification algorithms achieved improvements in churn prediction accuracy.

2.7 Conclusions

Chapter 2, presents a critical review and analysis of the literature on customer churn prediction. Initially, the systematic literature review approach is discussed and a plan for it is explained. Inclusions ad exclusions criteria for the systematic literature review are also reported and justified. Figure 2.3 presents the systematic literature review plan, adopted based on Brereton's et al., [24] and followed during this thesis. Besides the taxonomy of the literature review which is based on Cooper et al., [25], is presented in Table 2.1. Research questions, based on which the systematic literature review is conducted and used in the development of search strings to extract the correct data from data sources, are listed in Table 2.3. Moreover, Table 2.4 lists the selected cases that are in the main focus of this research.

Furthermore, observations and open issues derived from the systematic literature review on customer churn prediction are covered in detail in Section 2.6. In Section 2.6 the observations derived from the systematic literature review are highlighted and depicted in Figure 2.12. Five important observations revealed and indicate that: i) customer churn prediction is significant for subscription-based organizations, ii) there is a need for a customer churn prediction framework that will enhance decision making process, iii) literature pays attention to three main customer churn prediction aspects, iv) there is an absence of a research work that integrates all three aspects of customer churn prediction and v) there is a lack of a commonly accepted customer churn prediction procedure. All these observations demonstrate a research gap open research issues for further investigation.

The next chapter of this thesis focuses on the outcomes of the systematic literature review on customer churn and open issues derived from the literature review and led to the development of the Objective 2 of this research work.

Chapter 3

Conceptual Framework Development

Form follows Function

Louis Sullivan (1856-1924)

Summary

The previous chapter critically reviewed the normative literature on predicting customer churn. The work presented in Chapter 2 identified a research gap which, consists of two main issues. The first one reveals that there is a need for a customer churn prediction framework, that will enhance decision making process, where the second one highlights the lack of a commonly accepted customer churn prediction procedure. In order to address this gap, it is essential to firstly propose a procedure for customer churn prediction and then based on that to develop and introduce a customer churn prediction framework. This chapter attempts to address this research gap by proposing a customer churn procedure that includes five phases. The proposed procedure is then used for the development of the Customer Behavior Forecasting Framework. The latter, aims to improve decision making process and assist subscription-based organizations. Chapter 3 describes in depth the proposed conceptual framework, including its constituent elements. The framework attempts to contribute to the broad area of customer behavior forecasting, particularly in predicting customer churn in the subscription-based domain.

3.1 Introduction

The previous chapter presented a systematic literature review of the literature in the area of customer churn and critically evaluated the normative literature in customer churn in the subscription-based domain. Section 2.2, introduced the causes of customer churn and highlights, among others: i) poor customer service experiences, ii) unsuccessful customer on-boarding, and iii) weak customer relationship building. Furthermore, Section 2.3, analyzed challenges to retaining customers. Similarly, the effects of churn prevention were, in particular, the focus of the same section. Therefore, a detailed description of the benefits gained from a successful retention campaign presented in Section 2.3 , which include: i) a quantifiable impact on revenues, ii) vital insights for business growth, and iii) an increased Customer Lifetime Value (CLV).

In addition, the selection and justification of an established systematic literature review presented and analyzed in Section 2.4. The author adopted the Brereton's et al.,[24] methodology, which was used to select the cases reviewed during this research work and reflected in detail in Section 2.5. Chapter 2 concludes with the findings regarding the open issues and observations as those covered in Section 2.6. More specifically, during the literature review a research gap revealed, which consists of two main issues. The first denotes the research need for a customer churn prediction framework that will enhance decision making process regarding churn prediction actions taken by subscription-based organizations and the second one highlights the lack of a commonly accepted customer churn prediction procedure. Five important observations derived from the literature review are summarized in Figure 2.12.

Taking into consideration the outcome of the systematic literature review, Chapter 3 seeks to address them by four main investigation actions as those illustrated in Figure 3.1. More specifically, the author in an attempt to further investigate the open issues derived from literature review during Chapter 3, maps the main outcomes derived from Chapter 2 investigation actions to be taken in Chapter 3 to address the identified issues.

The main investigation actions as those demonstrated in Figure 3.1:

- Attempt to conceptualize of a framework that will enhance decision making process based on a proposed customer churn procedure. In doing so, the author seeks to address the research gap derived from the literature review and consists of two of the observations made: i) Observation B, and ii) Observation E.

- Attempt to investigate how the different aspects identified in Section 2.6, impact the churn prediction accuracy. The investigation of this seeks to address Observation A, and C.
- Attempt to investigate the integration of aspects derived Section 2.6 and to create a churn prediction procedure. This investigation action attempts to address Observation D.
- Attempt to identify the elements of a customer churn prediction framework and how the different elements can be orchestrated based on customer churn procedure for subscription-based organizations.

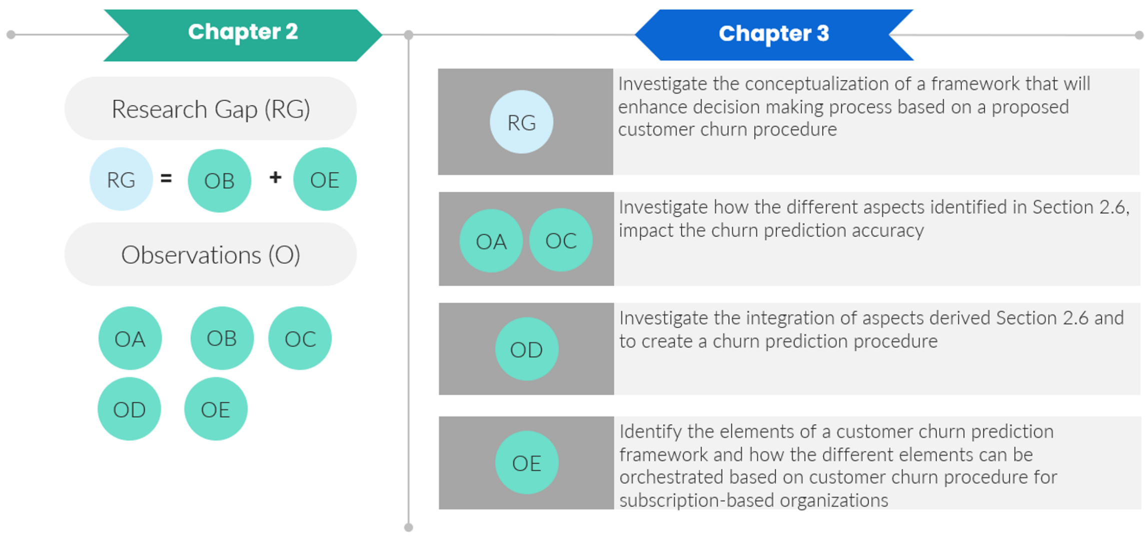


Figure 3.1: Mapping Chapter 2 Outcomes to Chapter 3

The structure of this chapter is as follows:

- **Research Challenges and Proposals:** In this section, the author analyzes and discusses the literature review findings based on observations, similarities and differences, discussed in Section 2.6, as part of the first investigation action discussed above and illustrated in Figure 3.1, to address the research gap derived from the systematic literature review. Clearly, the Chapter 2 findings indicate that there is need for a customer churn prediction framework, that will enhance decision making process and it will follow an established churn prediction procedure. Therefore, this section focuses on the synthesis of a customer churn prediction procedure based on literature review findings. In doing so, the author attempts to organize previously identified aspects of customer churn prediction and their constituents in the customer churn prediction procedure. This action will help the author to identify what aspects from the studied research works could be included in such a procedure. This section is divided into two subsections:

- Research Issues Derived from Literature Review on Customer Churn in Subscription-based Domain: This focuses on addressing one of the issues revealed from the literature review. In this section, the author attempts to address the absence of a commonly accepted customer churn prediction procedure by analyzing the aspects and their constituents, which will help him to define the phases of such a procedure. In doing so, he presents the proposed churn prediction procedure in this section and illustrates it in Figure 3.2. Furthermore, the author provides the ground for the upcoming sections by which to propose justifications on this first version of the churn prediction procedure; this should conclude in the development of the proposed conceptual framework by this thesis.
- Proposed Solutions for Customer Churn Prediction: In this section, the author proposes modifications that enhance the customer churn prediction procedure mentioned in previous paragraph. Thus, in this section, the author presents his additions and proposals, through the development of the new elements, that will enhance the entire churn prediction procedure, based on which the proposed conceptual framework will be based.

• **Proposed Customer Behavior Forecasting Framework on Subscription-based Domain:** The proposed conceptual framework referred as Customer Behavior Forecasting Framework (CBFF), addressing the second issue of the research gap revealed from the literature review, which indicates the need for a customer churn prediction framework. The elements included in the proposed conceptual framework are presented in detail throughout this section, addressing Observations A, C and E, derived from the critical analysis of the literature.

3.2 Research Challenges and Proposals

Findings and open issues derived from the literature review are discussed and analyzed in this section. The research gap revealed from the literature review and consists of Observations B and E as those summarized in Figure 2.12 is in the main focus of this section among with Observation D. Therefore, the author attempts to fill this gap, through the identification of a customer churn prediction procedure, based on which the customer churn prediction framework could be based.

The starting point for the development of the customer churn prediction procedure is Observation C (see Section 2.6), which identifies three aspects for customer churn prediction. The author investigated how the main aspects derived from literature could be integrated in a churn prediction procedure. The outcome of this investigation revealed that previous published works, despite their different perspective and focus on customer churn prediction aspects they follow in a way a similar structure as they start from the data and all related actions on them (e.g., identification of available dataset, defined types of data, cleaning dataset, etc.,). Then, they proceed to churn prediction, where actions like the selection of the classification algorithm, and the prediction model is made. The final step of this basic structure is the churn prediction outcome where related actions could take place (e.g., evaluation metrics to measure prediction model performance, goal achievement etc.,).

Based on the above findings the author developed the abstract level of the churn prediction procedure as this is demonstrated in Figure 3.2. The abstract level churn prediction procedure consists of three main phases as those are the: i) Input Data, ii) Predict Churn, iii) Analyze Outcome.

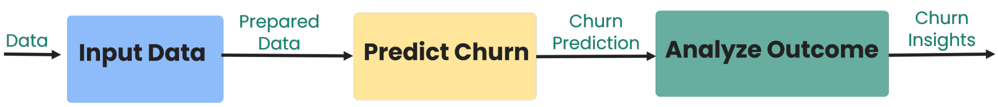


Figure 3.2: Abstract Level: Churn Prediction Procedure

Input Data includes all those actions that are related to the gathering, cleaning and defining of the data for churn prediction. To this end, the author suggests that, Aspect I- Data Preparation Techniques and Aspect III - Data Type for Churn Prediction, which include methods techniques, tools, methods, etc., could take place during this first phase as this is demonstrated in Figure 3.3.

Furthermore, the Churn Prediction phase describes all those actions related to the selection of a classification algorithm for the prediction model and the churn prediction procedure, it is suggested that the methods, algorithms, techniques and tools described in Aspect III - Classification Algorithm Selection may take place during this phase of the churn prediction procedure, as this illustrated in Figure 3.4.

The last phase of the proposed customer churn procedure, is the Prediction Outcome, which denotes the end of the procedure. In this phase, the churn prediction outcome and

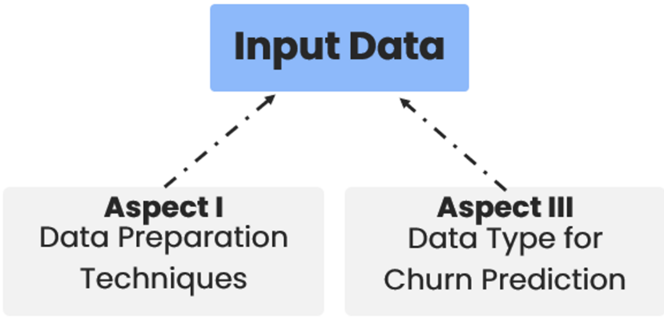


Figure 3.3: Input Data - Churn Prediction Procedure

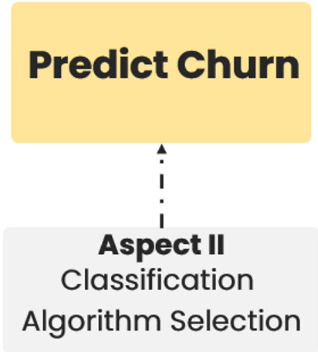


Figure 3.4: Prediction Churn - Churn Prediction Procedure

other related actions (e.g.,evaluation of prediction model used, churn rate etc.,) are taking place. This initial proposed churn prediction procedure helps the author to identify other related actions to be included in the first three phases or the modification of the current proposed phases. Therefore, the author takes each of the proposed phases of the initial churn prediction procedure and analyze it in more detail regarding its constituents. The outcome of these actions could give a first clear view of the cooperating parts of churn prediction procedure, based on which the author will develop the conceptual framework for customer churn prediction.

Starting from the Input Data phase, which includes among others the tools, methods, techniques etc., of Aspect I and Aspect III, the author attempted to categorize them under actions of this first phase. In doing so, he proposed that the two main sub-phases of the Input Data phase to be the i) Data type identification, and ii) Data Preparation Techniques, which both seem to have important impact on customer churn prediction, based on the findings derived from the systematic literature review. Figure 3.5 illustrates the two sub-phases of the Input Data phase.

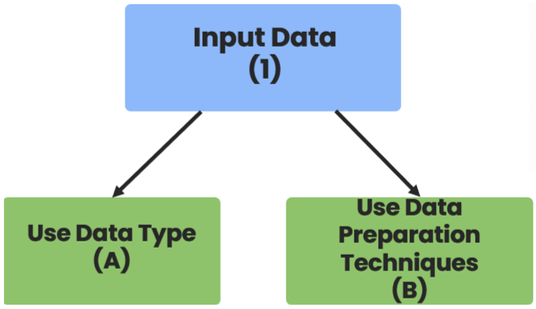


Figure 3.5: Input Data - Sub-phases

The Use Data Type sub-phase refers to the usage of specific type of data to be used during the churn prediction process and, based on the findings from the literature, this data could be: i) Cross-company, ii) Demographics, iii) Spatio-temporal Behavioral, iv) Service usage-related, or v) Profit related as those demonstrated in Figure 3.6. The Data Preparation Techniques, which indicates the second sub-phase of Input Data phase, it hosts all the methods and techniques related to dataset balancing, transforming, and representation of the data to fit better in the corresponding classification algorithm for the next phase.

Figure 3.6: Use Data Type Constituents

The Literature Review, highlighted the significance of data preparation stage on churn prediction and referred to the imbalance problem of the datasets [3, 4, 40] that may conclude in wrong churn prediction results. Therefore, Amin et al., [4] propose the usage of oversampling techniques to face the imbalance problem. Likewise, Coussement et al., [3], in their proposed framework which focuses on the data preparation aspect and covered in Section 2.5.2, as this referred by them [3], a very time consuming stage of the churn prediction, they introduced the Data Preparation steps: i) Transformation and ii) Representation. Taking them into consideration the author attempted to synthesize them under the Data Preparation Techniques sub-phase. Figure 3.7 demonstrates the synthesis of these aspects derived from the literature review under Data Preparation Techniques sub-phase.

Figure 3.8 presents the constitutes of the Oversampling Techniques, Transformation Step and the Representation Step. The oversampling techniques used in the studied research

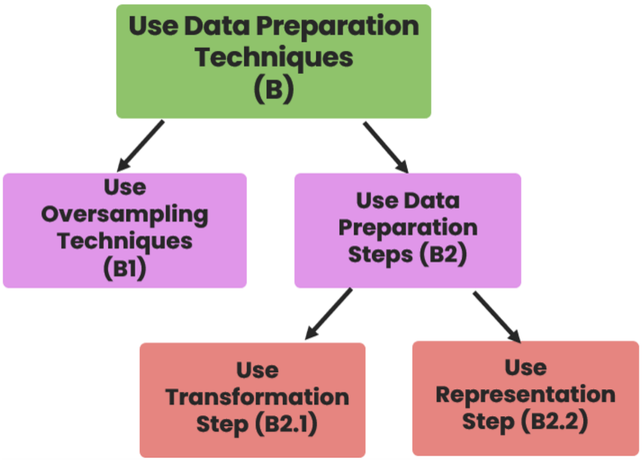


Figure 3.7: Use Data Preparation Techniques Structure

works include: i) Immune Centroids Oversampling Technique (ICOTE), ii) Majority Weighted Minority Oversampling Technique (MWMOTE), iii) Adaptive Synthetic Sampling Approach (ADASYN), iv) Mega-Trend Diffusion Function (MTDF), v) Synthetic Minority Oversampling Technique (SMOTE), and vi) Couples Top-N Reverse k-Nearest Neighbor (TRkNN). Furthermore, methods and techniques of the Transformation Step revealed from the research studied include: i) No regrouping, ii) Decision-based Remapping, iii) Decision Tree-based Discretization, iv) Equal Frequency Discretization, and v) Weight of Evidence Conversion. Moving on, to the Representation step, methods and techniques identified in the literature are: i) Dummy Coding, ii) Incidence Replacement, and iii) Weight of Evidence Conversion, which is also included in Transformation Step.

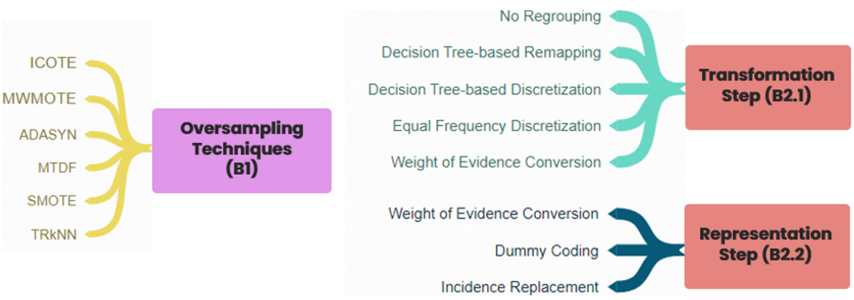


Figure 3.8: Oversampling Techniques, Transformation Step and Representation Step Constituents

The Predict Churn, denotes the second phase of the proposed customer churn procedure. During this phase methods and techniques described in Aspect II - Classification Algorithm Selection as part of Observation C reflected in Section 2.6, of the literature review, as the churn prediction related actions are included in this phase. Based on literature review findings, Predict Churn phase could be sub-divided into: i) Classification Algorithm Selection and ii) Prediction sub-phases. Figure 3.9, demonstrates the proposed structure of the Predict Churn phase.

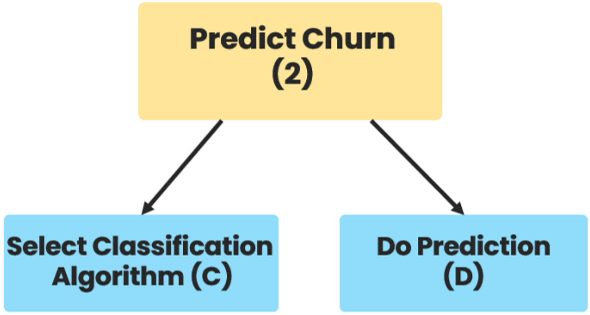


Figure 3.9: Predict Churn - Churn Prediction Procedure

The Classification Algorithm Selection sub-phase includes actions that conclude to the selection of the appropriate classification algorithm for the churn prediction model. In the Literature Review, the research studied focuses on various classification algorithms as those are: i) Bagged Classification and Regression Tree (Bagged CART), ii) k-Nearest Neighbor(k=10, k=100), iii) Alternating Decision Tree, iv) Classification and Regression Tree, v) C4.5 Decision Tree, vi) J4.8 Decision Tree, vii) Bayesian Network, viii) Multi-layer Perception Artificial Neural Network, ix) Naive Bayes, x) Random Forest, xi) Radial Basis Kernel Support Vector Machine (SVM-RBF), xii) Stochastic Gradient Boosting, xiii) Logistic Model Tree, xiv) Bagging, xv) Rough Set Theory-based (RST-based) Tool, xvi) Boosting, xvii) Polynomial Kernel Support Vector Machine (SVM- POLY), xviii) C5.0 Decision Tree and (xix) Logistic Regression, xx) Radial Basis Function Network, xxi) Repeated Incremental Pruning to Produce Error Reduction (RIPPER), xxii) Projective Adaptive Resonance Theory (PART), xxiii) Least-Squares Support Vector Machine (LSSVM) with Linear Kernel and xxiv) Least-Squares Support Vector Machine (LSSVM) with Radial Basis Function Kernel.Figure 3.10, summarizes the algorithms that they were on the focus of the studied research works.

The third phase of the customer churn procedure is the Prediction Outcome. This last phase of the churn prediction procedure includes all those actions that describe the churn prediction results and evaluation techniques. Based on the outcomes of the Literature Review, [1, 3, 4, 39, 40, 42] evaluation metrics to measure the performance of the proposals are used. In addition, the churn prediction outcome indicates the achievement of goals set by the organizations as those referred in studied research works. In doing so, the author

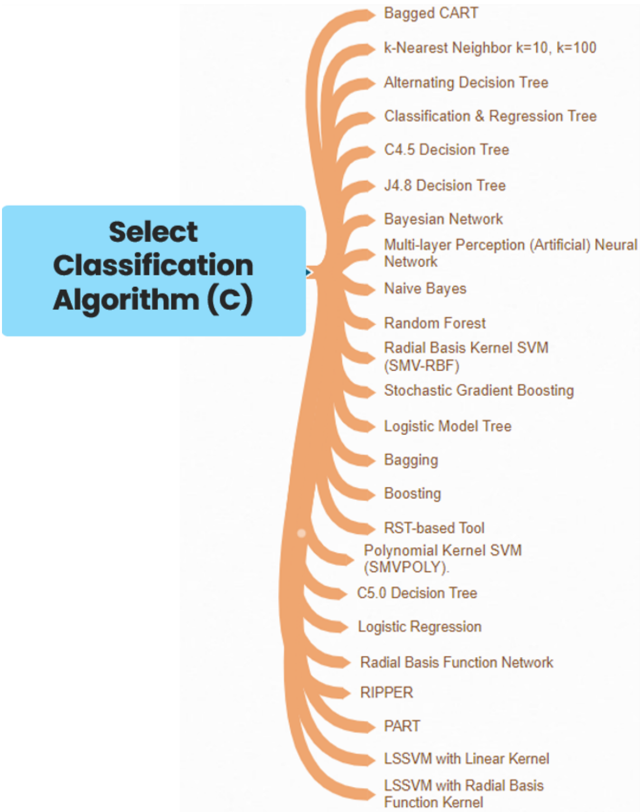


Figure 3.10: Select Classification Algorithm Constituents

proposed that the Prediction Outcome phase could be subdivided into: i) Evaluation Metrics and iii) Goals, as it is demonstrated in Figure 3.11.

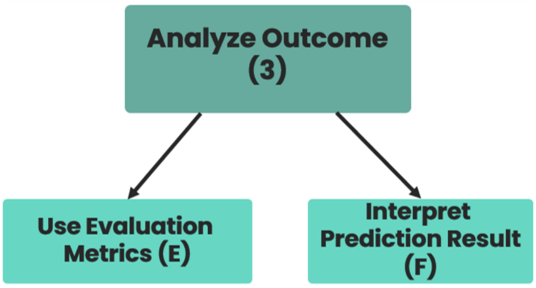


Figure 3.11: Analyze Outcome - Churn Prediction Procedure

The evaluation metrics derived from the literature review include: i) Recall, ii) Precision, iii) Accuracy, iv) F-measure, v) Area Under the Curve (AUC), vi) Top Decile Lift (TDL) and vii) Maximum Profit Criterion. As of the Goals sub-phase, the references from studied works to desired goals and targets [39, 41, 43] are summarized in: i) Churn prevention, ii) Improved Retention Campaigns or iii) Product / Service Improvements. The constituents of the two sub-phases are illustrated in Figure 3.12

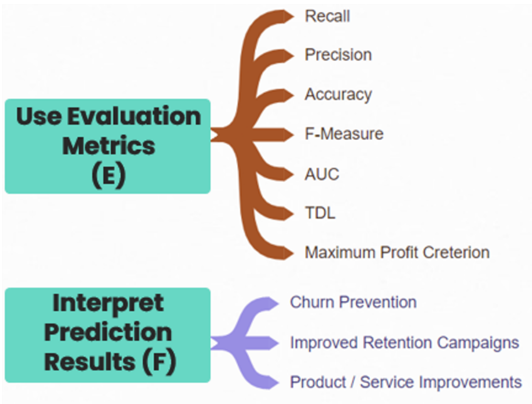


Figure 3.12: Evaluation Metrics and Interpret Prediction Results Constituents

Figure 3.13, demonstrates the synthesis of the main phases and sub-phases of the churn prediction procedure as these are presented and analyzed throughout this section. The visualization of the churn prediction procedure, attempts to support the development of a churn prediction framework.

3.2.1 Research Issues Derived from Literature Review on Customer Churn in Subscription-based Domain

Section 2.5 presented and analyzed research findings derived from the literature review investigation and are in the main focus of the author. Likewise, Section 2.6 reports observations identified in studied research, which are illustrated in Figure 2.12. The five important observations referred to: i) the importance of customer churn for the existence and prosperity of the subscription-based organizations, ii) the need for a customer churn prediction framework that will enhance decision making process, iii) the focus of the normative literature on three main customer churn prediction aspects iv) the absence of a research work that integrates all three aspects of customer churn prediction and v) the absence of a commonly accepted customer churn prediction procedure, on which the development of a customer prediction framework could be based.

Taking those observations into consideration, the author in Section 3.2 attempted to conceptualize the churn prediction procedure. The outcome of this attempt was the churn prediction procedure, which presented in Figure 3.13 and consisted of three main phases

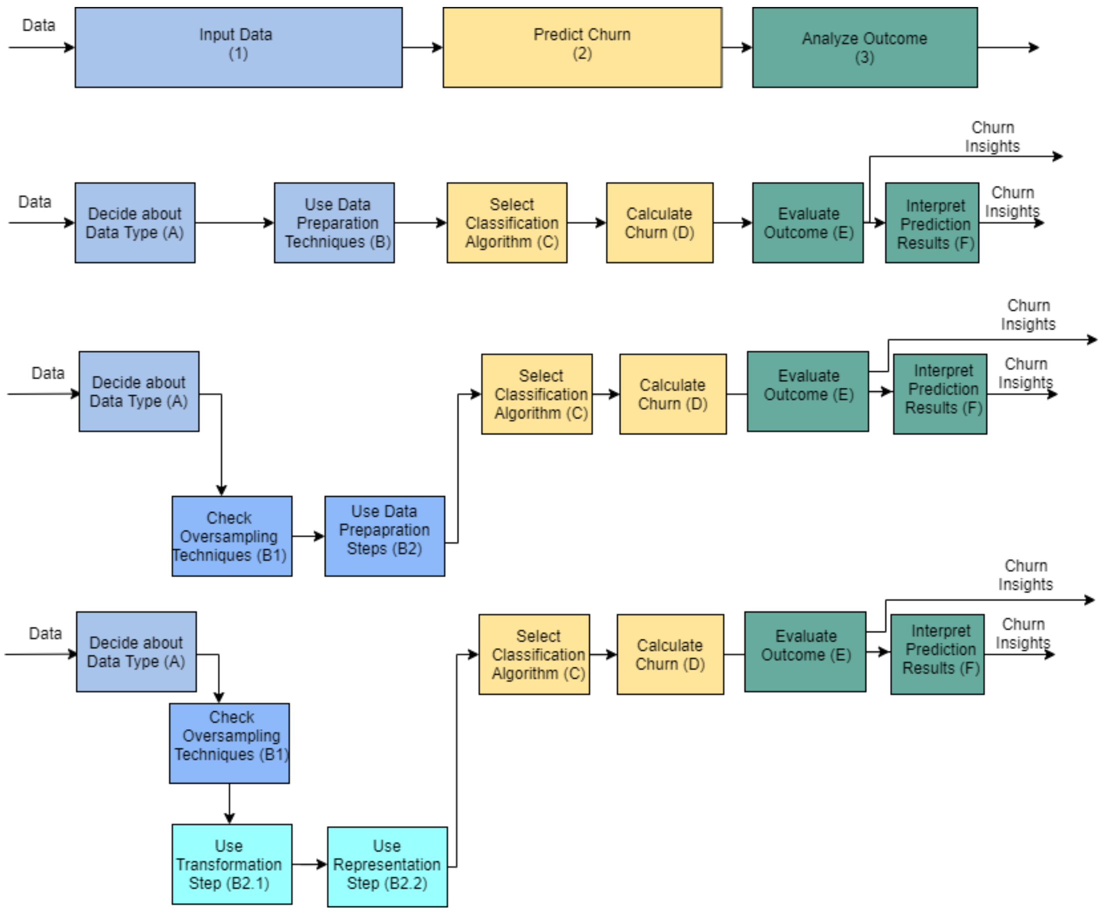


Figure 3.13: Churn Prediction Procedure - Basic Structure

as those are the: i)Input Data, ii) Predict Churn, and iii) Analyze Outcome. The three phases were based on the findings from the literature review as the majority of the studied research works seems to follow this basic structure. In addition, the author guided from the observations, regarding the three main aspects the normative literature focuses on, attempted to integrate them in the three main phases of the churn prediction procedure.

In this section, an attempt to identify issues on the initial Churn Prediction procedure as this derived from the literature review investigation is made. Taking the first phase of the churn prediction procedure, as seen in Figure 3.13, the Input Data phase is subdivided into two sub-phases: i) Data Type and ii) Data Preparation Techniques. The first sub-phase, Data Type, includes five data types identified in research works studied, and those are: i) cross-company, ii) demographics, iii) spatio-temporal behavioral, iv) service usage related and v) profit related. While ii), iv), and v) clearly express different types of data, and attributes of a dataset could be characterized as one of those categories, the: i) cross-company and iii) spatio-temporal behavioral data type categories do not give a clear view of what kind of attributes could be characterized as such. Notably, the term "cross-company" employed in the work of Amin et al., [39] is used to describe a data type that includes variables such as international plan, day charges, international minute, etc. In their work [39] they use two well-known publicly available datasets from the telecommunication sector, and based on their findings, cross-company datasets can be used for newly established companies in the telecommunication sector that do not have historical data yet. Likewise, the variables included in "cross-company" data type by Adnan Amin et al., [39] could be also be characterized as service usage related and cause confusion regarding what is classified as a data type; thus, it needs more investigation.

Similarly, what is called the spatio-temporal behavioral data type, referred to in Kaya et al., [42], includes attributes related to the expenditure behavior of customers. The confusion is identified in the usage of the term spatio-temporal in Kaya et al., [42] research work, which characterizes data with spatial relations (distance, direction, shape, etc.) and temporal relations (occurrence time, duration, etc.). The issue is identified in the usage of the term Spatio-temporal to characterize attributes similar to the ones found in the service usage related data type, which Kaya et al. [42] call "behavioral" and thus refer to as spatio-temporal behavioral in the churn prediction procedure presented in Figure 3.13. This issue needs to be clearly described and the author paying attention to it during the development of the proposed conceptual framework.

Furthermore, the Data Preparation Techniques step which subdivides into two sub-phases: i) Oversampling Techniques and ii) Data Preparation Steps. Oversampling Techniques include six methods and techniques as those are: i) ICOTE, ii) MWMOTE, iii) ADASYN, iv) MTDF, v) SMOTE and vi) TRkNN. The second sub-phase, Data Preparation Steps, also subdivides into the i) Transformation Step and ii) Representation Step, like those described in Coussement et al., [3]. The issue found at this point is related to the subdivision of Data Preparation Steps into two sub-elements (Transformation, Representation), which transform and represent dataset variables into a suitable form that better fits the classification algorithm used in the upcoming steps of the churn prediction procedure.

Taking into consideration the issues revealed from the study of Coussement et al., [3] research work, regarding the correct placement of the referred techniques listed under Transformation and Representation steps as shown in Figure 3.13 in a churn prediction procedure the author attempt to resolve the issue. As such, the stages of Transformation and Representation presented in Figure 3.8 should be merged under the parent element of

Data Preparation Steps. Based on that, it can be seen in Figure 3.8 that one element, Weight of Evidence Conversion, met in both steps (Transformation, Representation), which does not resolve the confusion regarding when to use it as a transformation method and when it is more suitable for representation purposes. In order to clarify such overlaps encountered in literature, this thesis investigates the processes taking place in that specific sub-phase of the churn prediction procedure (Data Preparation Techniques), to simplify and clarify methods and elements included in each sight of the Data Preparation Techniques step accordingly.

The second phase of the initial churn prediction procedure, the Predict Churn, and more specifically the sub-phase Classification Algorithm Selection, holds twenty four classification algorithms used by researchers of the studied research works, and those are: i) Bagged CART, ii) k-Nearest Neighbor(k=10, k=100), iii) Alternating Decision Tree, iv) Classification and Regression Tree, v) C4.5 Decision Tree, vi) J4.8 Decision Tree, vii) Bayesian Network, viii) Multi-layer Perception Artificial Neural Network, ix) Naive Bayes, x) Random Forest, xi) Radial Basis Kernel SVM (SVM-RBF), xii) Stochastic Gradient Boosting, xiii) Logistic Model Tree, xiv) Bagging, xv) RST-based Tool, xvi) Boosting, xvii) Polynomial Kernel SVM (SVM-POLY),xviii) C5.0 Decision Tree, xix) Logistic Regression, (xx) Radial Basis Function Network, xxi) RIPPER, xxii) PART, xxiii) LSSVM with Linear Kernel, and xxiv) LSSVM with Radial Basis Function Kernel. The classification algorithms mentioned above, used in research works studied under different circumstances, are shown in Figure 3.14 and explained below:

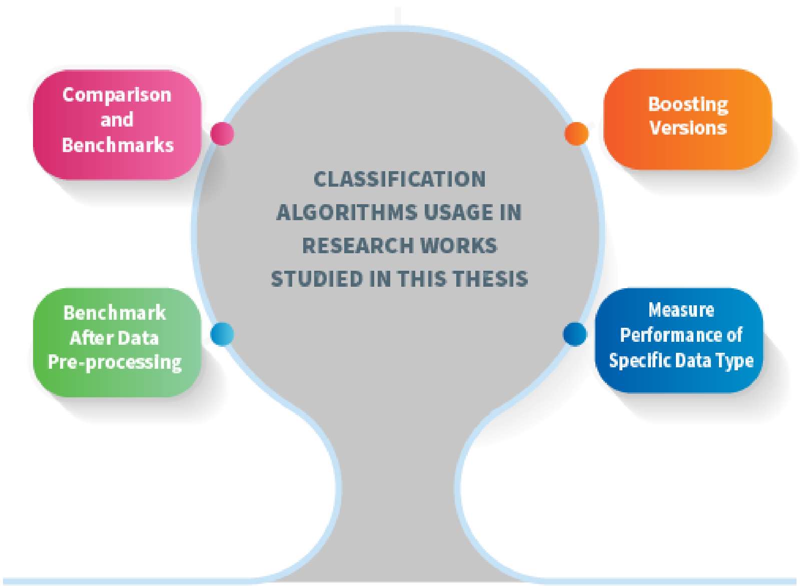


Figure 3.14: Classification Algorithms Usage in Research Works Studied

- Comparison and Benchmarks: Many of the research works studied during the literature review compare different classification algorithms in an attempt to determine which show better performance. The research of Vafeiadis et al., [43] compares well-known algorithms from those mentioned above, which are: i) Multi-layer Perception Artificial Neural Network, ii) Radial Basis Kernel SVM (SVM-RBF), iii) Polynomial Kernel SVM (SVMPOLY), iv) Naive Bayes, v) C5.0 Decision Tree, and vi) Logistic Regression. Equally, Verbeke et al., [40] conduct extensive benchmark experiments on various classification technique categories, including i) Decision Trees, ii) Ensemble Methods, iii) Nearest Neighbours, iv) Neural Networks, v) Rule Induction Techniques, vi) Statistical Classifiers, vii) Statistical Classifiers, and viii) SVM-based Techniques; these are presented in detail in Table 2.9.

- Boosting Classification Algorithms: Vafeiadis et al., [43], benchmark boosting versions of four algorithms compared during the first phase of their research. Boosted versions of algorithms compared include four that were able to be boosted: i) Back-Propagation Network (popular model of artificial neural networks), ii) Decision Trees iii) Support Vector Machines (Gaussian Radial Basis (SVM-RBF), and iv) Polynomial (SVM-POLY). The two that were not able to be boosted were i) Naive Bayes and ii) Logistic Regression due to the lack of free tuning parameters, as described and analyzed in Section 2.5.5.

- Measure Performance of Specific Data Type: Some of the research presented in Section 2.5 measure the performance of classification algorithms based on the specific data type. Kaya et al., [42] attempt to fill the gap between the dynamic behavioral pattern of clients, their financial decisions, and churning activities. In doing so, they propose a prediction model based on the usage of Spatio-temporal behavioral data. Kaya et al. [42] used Random Forest as a classification algorithm for their dataset. Similarly, Verbeke et al., [40] compare seven categories of algorithms, as presented in Table 2.9, to measure their performance on service usage related data type. Verbeke et al., [41], in another study included in this thesis, benchmark the performance of various state-of-the-art data mining classification algorithms.

- Benchmark Classification Algorithms After Data Pre-processing: Research by Coussement et al., [3] examine alternatives to enhance prediction performance. In doing so, it benchmarks the eight state-of-the-art data mining techniques presented in Table 2.8. To measure their findings performance, they compared eight classifiers, which are included in Figure 3.10: i) Bagged CART, ii) Bayesian Network, iii) J4.8 Decision

Tree, iv) Multilayer Perception Neural Network, v) Naive Bayes, vi) Random Forest, vii) Radial Basis Kernel SVM, and viii) Stochastic Gradient Boosting.

However, instead of circumstances in which classification algorithms are typically used in research, there is a gap regarding the different versions of well-known classification algorithms that are referred to in the research works studied. For example, are the different versions of SVM classification algorithms, such as SVM-POLY, LSSVM, and SVM-RBF? During the upcoming sections, the author attempts to address these issues by providing his proposals, while he discusses derived challenges. Another point that has to be examined is the confusion regarding the term used by Coussement et al., [3] called "Multi-layer Perception Neural Network" and by Vafeiadis et al., [43] called "Multi-layer Artificial Neural Networks" which refers to same classification algorithm but with slightly different terminology.

Issues like these create misunderstandings regarding customer churn prediction approaches. Based on Objective 1 of this thesis—that is, to investigate the previously proposed customer prediction approaches to identify issues and challenges and thus enhance future actions taken regarding customer churn prediction in subscription-based organizations—these issues need to be clarified.

3.2.2 Proposed Solutions For Customer Churn Prediction

This section deals with the modification of the churn prediction procedure based on issues identified and discussed in the previous sections. Moreover, this section proposes additions to customer churn prediction approaches as a preamble to the upcoming CBFF that is proposed in this chapter.

Modifications to the churn prediction procedure are essential to clarify unclear points and issues identified and discussed during section Section 3.2. The five data types classified in the Use Data Type sub-phase in Figure 3.6 give rise to concerns, like those discussed in Section 3.2.1. More specifically, the cross-company data type presented in Amin et al., [39], which includes attributes like the international plan, day charges, international minutes, etc., describes the usage of a service and thus could be characterized as service usage related. Therefore, the cross-company characteristic of a dataset that holds service usage related data than a data type. Similarly, the spatio-temporal behavioral data type as presented by Kaya et al., [42] refers to attributes related to the expenditure behavior of the customers, together with spatial and temporal information around those attributes. The expenditure-related attributes mentioned by Kaya et al., [42], describe the usage of a financial product together with spatio-temporal information with service-usage-related attributes.

Similar to the case of Admin et al., [39], Kaya et al.'s, [42] spatio-temporal behavioral are referring to characteristics of a data type, than a standalone data type. Owing to this, the Use Data Type sub-phase in the modified version of the churn prediction procedure includes three data types, as follows: i) demographics, ii) profit related, iii) service usage related. Data types previously characterized as cross-company and spatio-temporal behavioral, are now characteristics of a dataset and a service usage related data type accordingly, as shown in Figure 3.15.

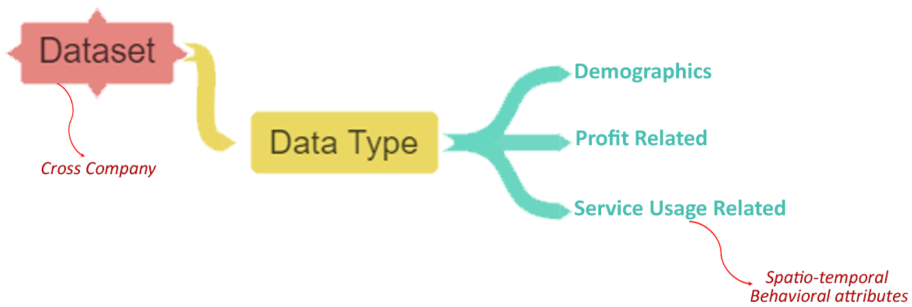


Figure 3.15: Data Types Elements

Moreover, issues were also encountered in the subdivision of Data Preparation sub-phase into two sub-elements (Transformation, Representation), covered in Section 3.2.1. The author proposes the modification of the Data Preparation Techniques sub-phase into Data Preparation Treatments, which better describes the elements included in that phase. Based on the findings and discussion of the proposed DPT framework by Coussement et al., [3] which shows improvements in the churn prediction, the author transformed it to include overlaps identified in the research works studied during the literature review process. Figure 3.16 presents the modified Data Preparation Treatments step. The modified Data Preparation Treatments step holds seven elements, as follows: i) No Regrouping, ii) Decision Tree-based Remapping, iii) Decision Tree-based Discretization, iv) Equal Frequency Discretization, v) Weight of Evidence Conversion, vi) Dummy Coding and Incidence Replacement. As seen in Figure 3.16, the Weight of Evidence Conversion element is considered as a Data Preparation Treatment, eliminating confusion encountered in the previous version of the element.

Furthermore, the Classification Algorithms Selection sub-phase shown in Figure 3.13, which holds the twenty-four algorithms identified in the Literature Review, needs a justification

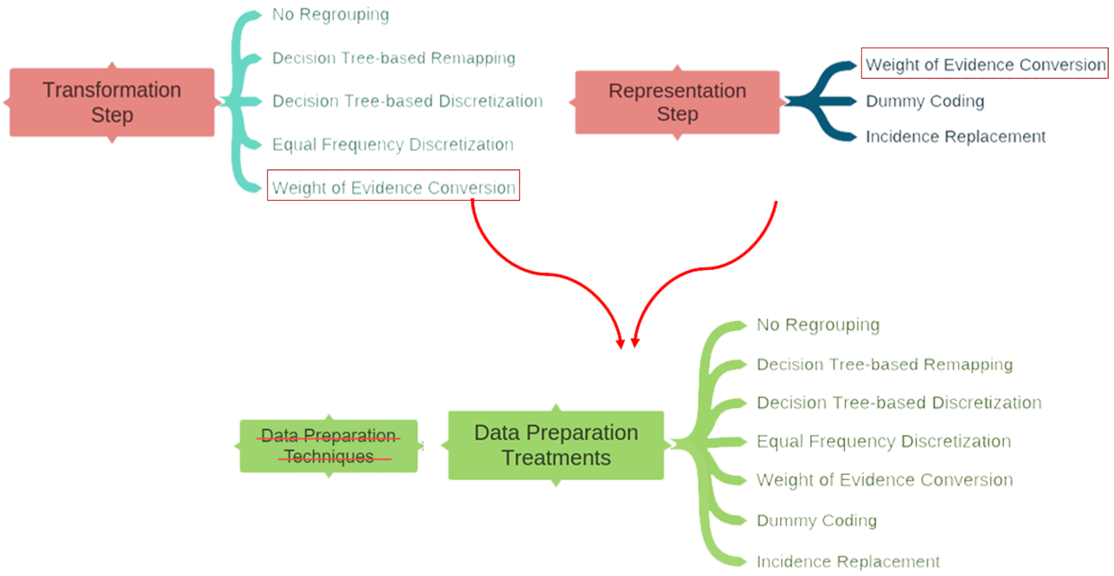


Figure 3.16: Merging Transformation and Representation Phases

for a better presentation of the classifications algorithms. Verbeke et al., [40] use an efficient way to summarize the classification techniques evaluated during their study by grouping the classification algorithms in categories. Given that, the author, groups classification algorithms in the categories mentioned above, justifying the Classification Algorithms Selection sub-phase of the churn prediction procedure accordingly as presented in Figure 3.17.

Categories of classification are based on Verbeke et al., [40], as already mentioned and a brief description of them is as follows:

- Decision Tree Approaches: Recursively grown trees, from upside down. Training records are partitioned into branches. The end of the branch that is not able to split again is the decision. The feature's importance is evident, and relations are easily observed. Different decision tree methods measure the splits and homogeneity of groups in a different way [40].
- Ensemble Methods: Ensemble methods create multiple models, which then are combined improving predictive performance compared to a single model [59]. This fact makes the ensemble methods to be preferred by top machine learning competitions such as Kaggle [60] and KDD Cup [61]. Random forests incorporates CART as base learner, a logistic model tree utilizes logit, and both bagging and boosting use decision trees [40].

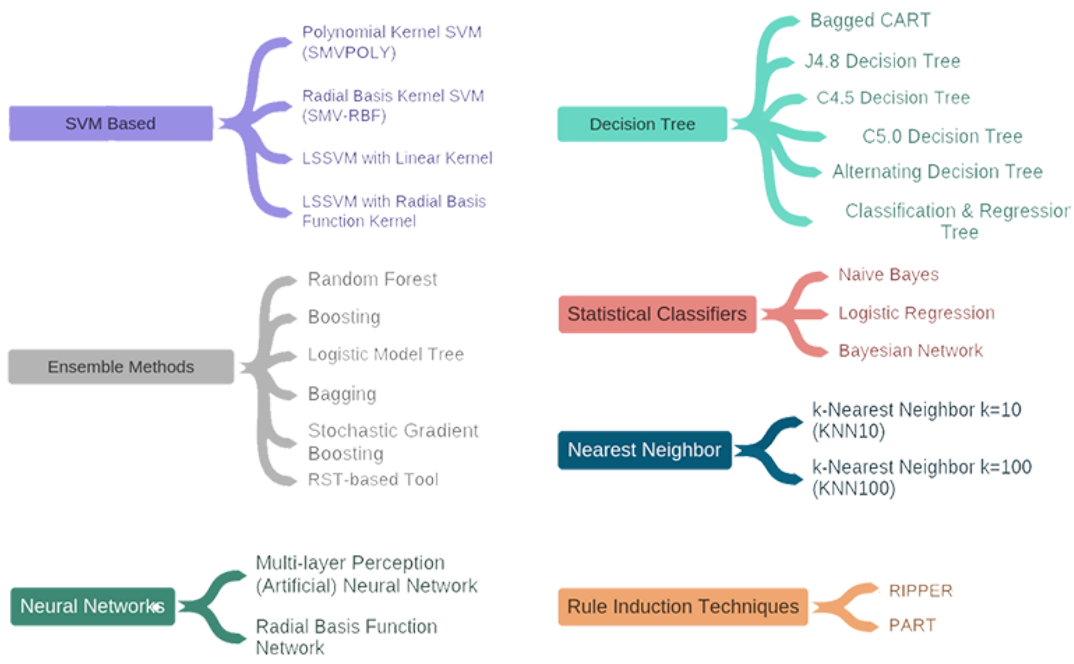


Figure 3.17: Categorizing Classification Algorithms

- Nearest Neighbor Methods: Classification methods that find the most similar data points in the training data, and make their prediction based on their classifications. Classifications of attributes are based on the k-most similar or nearest attributes [40].

- Neural Networks: In neural networks, the idea is that each variable is associated with a weight. Then a combination of weighted variables is run to develop the prediction model [43].

- Rule Induction Techniques: This technique uses a clear set of if-then rules to predict the minority class, as the majority class is assigned by default [40, 62].

- Statistical Classifiers: Are considered all those techniques, that simulate probabilistic relationships between the class variable and the attribute set [40]. Naive Bayes is considered as a statistical classifier. In Naive Bayes the estimation of the class-conditional probability is made based on the assumption that attributes are conditionally independent.

- SVM-based: Describes all methods based on the SVM classification method. In SVM, each variable is separated into two categories. Then the Support Vector Machine training algorithm develops a model where it assigns new variables to one or the other category. The output of the Support Vector Machine prediction model is points in space divided by a clear gap [63].

The justified version of the Classification Algorithms selection sub-phase aims to provide a better understanding of which of the well-known algorithms is used to churn prediction approaches and allows to recognize the specific version or modification of it used in each of the research works. Accordingly, the author, established a better perspective on how to implement the classification algorithms selection sub-phase in the proposed customer behavior forecasting framework.

As a result of the above modifications, the author altered the churn prediction procedure of Figure 3.13, to propose a refined one as seen in Figure 3.18, which summarizes the justifications presented and analyzed in this section. The procedure presented in Figure 3.18, decreases overlaps and other issues discussed previously in this section. Figure 3.18 is used as the basis for the development of the proposed conceptual framework for customer churn prediction (see Section 3.3). The proposed conceptual framework should take into account the justifications of the customer churn procedure phases. In doing so, the proposed framework aims to enhance previously taken actions in customer churn prediction for subscription-based domain and provide a clearer overview of the process taking place during a churn prediction procedure.

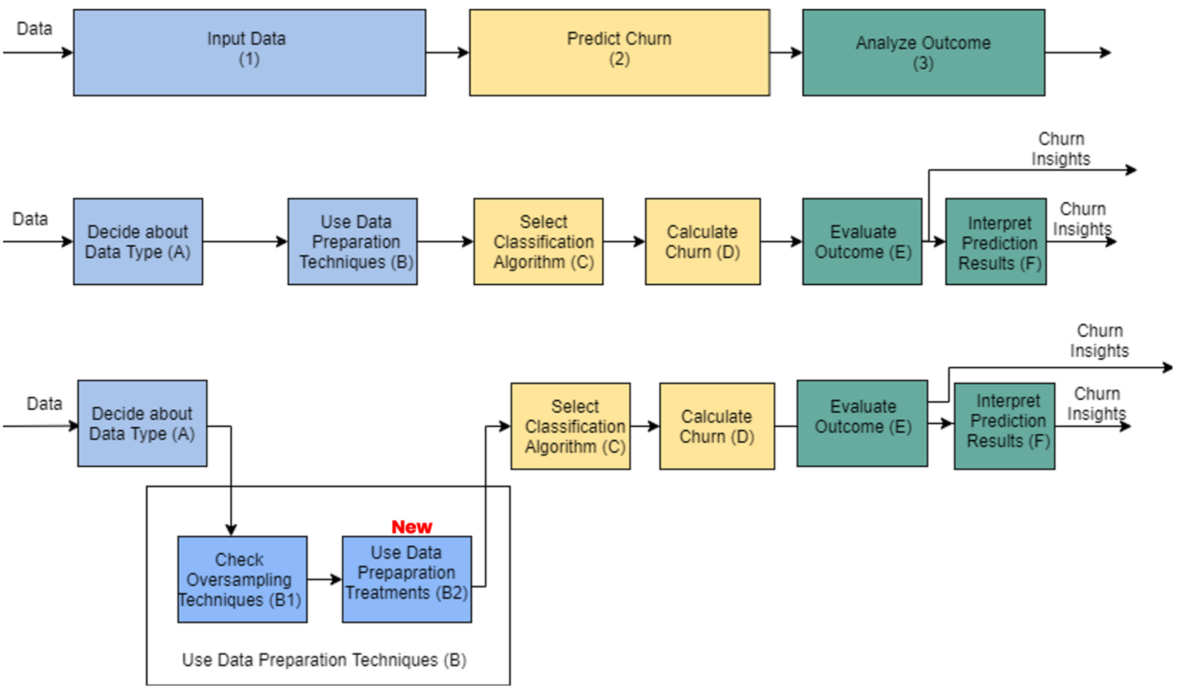


Figure 3.18: Modified Version of Churn Prediction Procedure

3.2.2.1 The Importance of Data Type

The following paragraphs present proposals to customer churn prediction approaches. The proposals presented are derived from literature review findings and knowledge gained throughout the analysis of literature review findings. The requisite elements of the churn prediction procedure have been identified in previous sections. In an attempt to enhance the knowledge and guidelines around customer churn, the author presents and discusses proposals for the upcoming conceptual framework for churn prediction. The following paragraphs present proposals on the essential elements of the churn prediction procedure identified in previous sections of this chapter.

One of the aspects of great importance in the research works studied is the data type for churn prediction. Different researchers, such as Kaya et al., [42] and Verbeke et al., [40, 41], grounded their findings on the specific type of data used during the churn prediction. More specifically, those authors improved the churn prediction accuracy based on the different types of data they use. The usage of a specific data type in the churn prediction dataset has a significant impact on the churn prediction outcome. As a result, the author suggests that Use Data Type sub-phase, is significant. In addition to this, the author assumes that it is essential at this stage to better understand the organization under study and its needs. Thus, an analysis of important organizations KPIs could enhance knowledge and will lead to the identification of the most appropriate data for the churn prediction. In doing so, the author suggests the development of an element - Element 1(E1) - which can work as a predecessor of data collection actions and which can include tasks related to the collection of those KPIs from the organization and conducting an analysis on them to extract the appropriate knowledge that it can be used during the collection of the available data records from the organizations. In that manner, the author assumes that the data type used in the churn prediction dataset is based on the desired goals and impact feature selection during the churn prediction procedure. More details, regarding the elements that take place during the Input Data phase of the churn prediction procedure are given in Section 3.3.

3.2.2.2 Considering Data Preparation Techniques in the Proposed Conceptual Framework

Following, the suggestion of the Section 3.2.2.1, current section seeks to consider the usage of Data Preparation Techniques in the proposed conceptual framework for churn prediction. Most of the real-life classification problems include a grade of class imbalance, as this is also noted in the research works studied during the literature review [3, 4, 41]. Therefore, any actions taken to improve current churn prediction techniques may take into consideration this issue, adjusting the dataset to avoid fault classification and identification of the class of importance. The constituents of the Data Preparation Techniques sub-phase as this resented in Figure 3.18, refer to: i) Oversampling Techniques and ii) Data Preparation Treatments can be used to improve the possibility of correct feature selection and identification of the class of importance in the proposed conceptual framework.

As a result, the adoption of two different elements (E2, E3), is suggested, which will also take place during Input Data phase, but after the Use Data Type sub-phase. The first of the two proposed elements: i) will take into consideration the knowledge extracted from the element suggested in Section 3.2.2.1, ii) can identify all available data resources in the company that are can be used for churn prediction and iii) can proceed with data cleaning and data preparation actions on them. Oversampling techniques as those presented in Figure 3.8 can be used to mitigate the class imbalance problem in a dataset, either by oversampling the minority class or under-sampling the majority class. Then, a second proposed element under this sub-phase, can take the input from the previous element, which can be a cleaned and balanced dataset to proceed with feature selection actions.

3.2.2.3 Churn Prediction Phase Elements

Having, the churn prediction dataset balanced and defined, with the appropriate features for the churn prediction, the churn prediction process can begin. During this phase the selection of the appropriate classification algorithm precedes the churn prediction process. Based on the normative literature that refers to prediction models [1, 4, 39–43] to describe the churn prediction stage, the author suggests the inclusion of an element (E4) that can include the actions related to the selection of the appropriate classification algorithm and the churn prediction, under the Churn Prediction phase.

3.2.2.4 Outcome Interpretation

The Analyze Outcome denotes the end of the churn prediction procedure, but also the phase where important tasks and actions take place. The churn prediction outcome can ring different alarms for the subscription-organization such as specific churn countermeasures, differentiation in retention strategy, product development etc. Taking these situations alongside with the literature review findings, the author suggest the use of an element (E5) in the last phase of the churn prediction procedure, which could include tasks and actions related to churn prediction outcome interpretation for achieving organization's desired goals, as of the evaluation of the procedure.

Grounded on the discussion carried out in Sections 3.2.2.1-3.2.2.4, the author suggests, five elements to be included in the conceptual framework. The five suggested elements can be mapped in the churn prediction procedure as seen in Figure 3.20, seeking to enhance decision making process from subscription-based organizations.

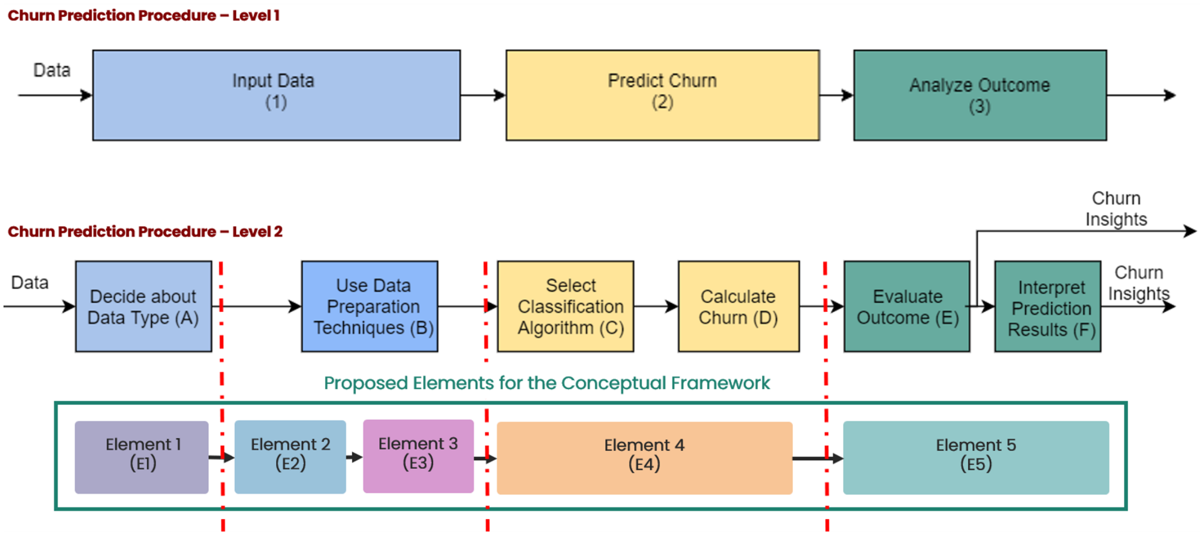


Figure 3.19: Mapping Proposed Elements of the Conceptual Framework to Churn Prediction Procedure

3.3 Proposed Customer Behavior Forecasting Framework for Subscription-based Domain

To introduce the reader smoothly to the content of this section, a subscription-based organization definition is given. Subscription-based organizations are all those companies that acquire customers once and then have a long-term paying relationship with them. Based on that, companies of the subscription-based domain have a "more predictable revenue". The focus of this thesis is on the subscription-based domain and the improvements in churn prediction; as already observed from the literature review, churn is vital for their health and prosperity.

This section presents author's proposals to improve churn prediction approaches in subscription-based organizations by introducing the proposed conceptual Customer Behavior Forecasting Framework, which takes into consideration those elements (discussed in Section 3.2.2). Taking into consideration the churn prediction procedure as presented in Figure 3.18, the author developed an abstract level of the forthcoming conceptual framework. Figure 3.19, maps the proposed customer churn prediction procedure to five elements identified as significant for customer churn prediction as reported in, Section 3.2.2 Figure 3.20 presents an abstract level of the proposed conceptual framework, namely, the Customer Behavior Forecasting Framework and incorporates the five elements presented in Figure 3.19.

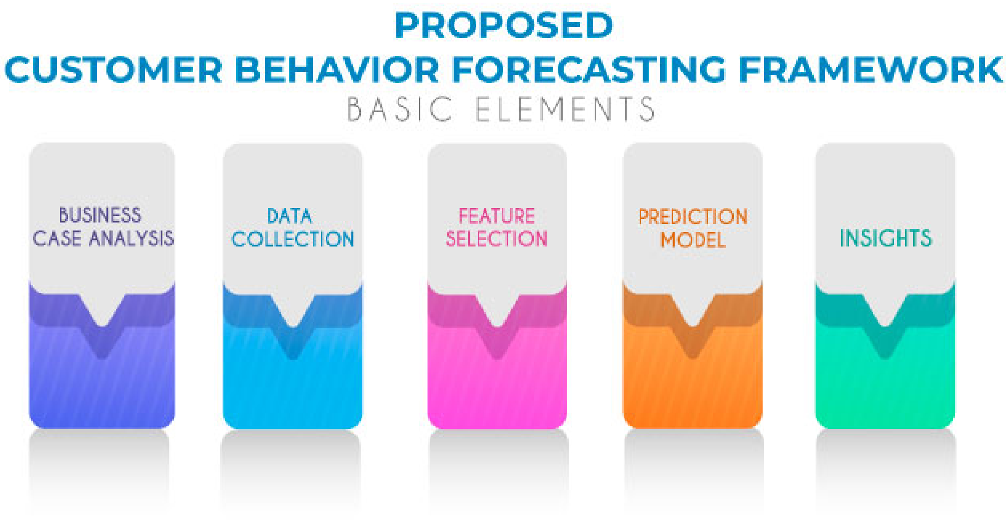


Figure 3.20: Proposed Customer Behavior Forecasting Framework - Basic Elements

Looking at the abstract level of the Customer Behavior Forecasting Framework [7] illustrated in Figure 3.20 the author proposes the inclusion of five essential elements, which are:

- **Business Case Analysis (E1):** This element should encapsulate all business case related actions that should support a successful churn prediction process. More specifically, this element is responsible to collect all needed information from the organization under study. This information is related to the business KPIs, desired outcomes, annual goals, etc. The analysis of the business case may help actors to develop an enhanced knowledge about their churners' profiles and lead them to the identification of their churning customers more accurately. Identifying churners and reducing customer churn rates requires a clear understanding of the business desired outcomes.
- **Data Collection (E2):** This element describes all actions taken to identify suitable data sources to be used by feature selection processes. Taking into account what

the discussion in Section 3.2.2, regarding the data type, this element copes also with data cleaning and data preparation actions. Moreover, methods such as those described in the same section (Section 3.2.2) and more specifically for the data preparation techniques element of the modified churn prediction procedure in Figure 3.18 are included.

- **Feature Selection (E3):** The Feature Selection Element describes the process taking place on cleaned data, during the feature selection process. Techniques and methods for this element should include, among others: i) features proposed from the Business Case Analysis Element, ii) features derived as an output of the Data Collection Element, and iii) tools that enhance the creation or selection of relevant features in the dataset.

- **Prediction Model (E4):** This element may play an important role in accurate churn prediction process. The Prediction Model element refer to all those actions related to the prediction of the possible churners. More specifically the Prediction Model Element could include: i) classification algorithms selection and ii) the churn prediction actions.

- **Insights (E5):** The last element of the proposed conceptual framework, namely Insights, could consist of actions that lead to the interpretation of the churn prediction outcome. Such actions could be the inclusion of evaluation metrics, which should support the imprint of the performance of tools, methods, and techniques used during a churn prediction process. The author gives special attention to actions related to the interpretation of results, as many of the churn prediction approaches studied during the systematic literature review concluded to the classification to churners and non-churners, without the details on how they concluded to that result. The author assumes that those details should lead to uncovered information related to the profile of the churners, which should be taken into consideration in future churn prediction attempts. Outcome interpretation denotes one of the important innovations of the proposed Customer Behavior Forecasting Framework this thesis proposes. Actions, methods and techniques included in these elements are described in the Analyze Outcome phase in Figure 3.18 and more specifically during Evaluation Metrics and Goals sub-phases.

Regarding the focus of this thesis on churn prediction for subscription-based companies, the proposed conceptual framework takes into consideration all those aspects that could improve the accuracy of churn prediction approaches.

3.3.1 Analyzing Elements of The Proposed Customer Behavior Forecasting Framework

The author conducted this research work in an attempt to clarify steps and procedures to be followed by subscription-based companies in order to identify possible churners. In doing so, he proposes the Customer Behavior Forecasting Framework (CBFF), which aims to enhance actions related to customer churn prediction improving the way organizations work towards the identification of churning customers while at the same time archiving higher accuracy in their predictions. During this section, a more detailed presentation of each element of the CBFF takes place. The five elements of the CBFF are discussed and analyzed and gives the opportunity to the reader to understand the functionality and interactions between elements of the conceptual framework this research work proposed in this section, to face challenges around churn prediction in subscription-based domain.

3.3.1.1 Proposed Business Case Analysis Element (E1)

The Business Case Analysis element corresponds to significant actions taken during the process and its effect on the success of the CBFF. Each organization in the subscription-based domain has desired outcomes, and market targets, which need to achieve during a time-frame, even this are semester, annual or long-term goals. Based on that, the CBFF should take into consideration the different parameters for each of the organizations to customize the churning process to that organization needs. In doing so, during this element essential market indicators, desired outcome targets and related information should be taken into consideration in that early stage of churn prediction, that will enhance decision making for the upcoming tasks in the CBFF.

To achieve that, processes proposed to consist of the Business Case Analysis Element, should suggest the selection of appropriate features based on insights and other related information derived either from previous churn prediction attempts or from the business analysis as this illustrated in Figure 3.21. The selection of the suitable features from the data set for churn prediction requires a deep understanding of business aims and objectives. Issues derived from the literature review conducted on churn prediction domain, show the limited focus given to prior activities, such as those included in a business case analysis, before the feeding of the dataset into a classification algorithm. This limited focus, often leads to less accurate churn prediction results. Therefore, one of the aims of the Business Case Element is to produce all those information related to the organization under study, which should support the churn prediction accuracy enhancement.

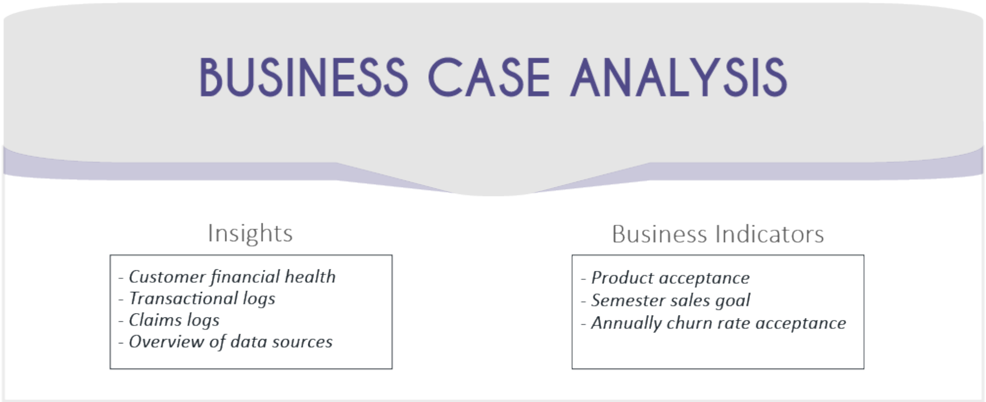


Figure 3.21: Proposed Business Case Analysis Element

3.3.1.2 Proposed Data Collection Element Analysis(E2)

The second element of CBFF, is the Data Collection Element which is illustrated in Figure 3.22 and includes all those actions that support the selection of all appropriate data sources from the organization, which should be used as the churn prediction dataset. Following, the analysis of the goals, important KPIs and other related information of the organization, which derived during the Business Case Analysis Element, the Data Collection Element is responsible to collect all available data sources from the organization taking into consideration many different aspects to develop the churn prediction dataset. Actions that are enclosed in the Data Collection Element include among others: i) the identification of the different data sources that exists, ii) the possible merging of the different data sources into one and iii) the cleaning and balancing activities on the final dataset's variables.

The data sources may vary and my include reports with business KPIs, CRM database and archive files, that will be taken into consideration during data sources identification actions. Following the identification of the data sources should be actions related to merging the different data sources into a single dataset that should be used for the prediction of the churners. Merging actions are often accompanied with the conversion to a single format for all the data sources, to enable further process on the data. Moreover, processes related to cleaning and balancing of the dataset may include among others: i) oversampling techniques, ii) data reduction, iii) fixing of missing variables and iii) variable balancing techniques. In summary the outcome of the Data Collection Element should be a balanced dataset, adjusted, and ready to be imported in the next element of the proposed conceptual framework.

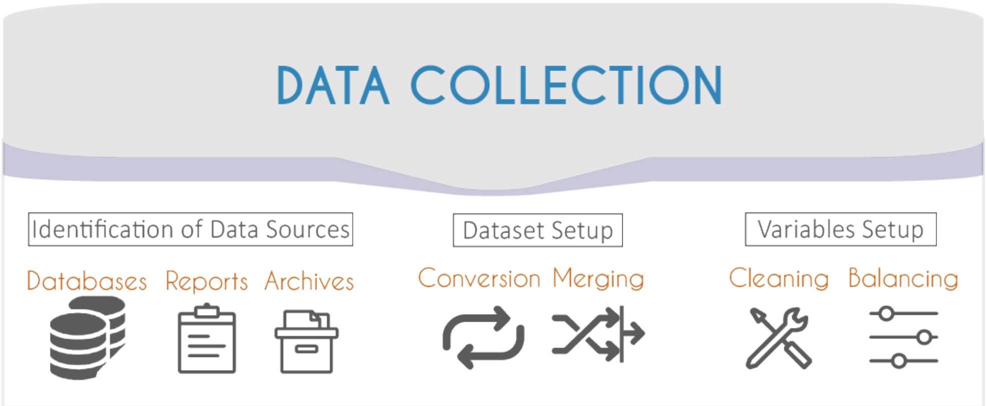


Figure 3.22: Proposed Data Collection Element

3.3.1.3 Proposed Feature Selection Element Analysis (E3)

The third element of the proposed conceptual framework is the Feature Selection Element, where processes regarding the selection of appropriate features take place. Feature selection Element which is presented in Figure 3.23 and it may play a critical role in the accuracy of churn prediction model. The previous assumption is based on the literature review findings which support that if the dataset variables are falsely selected as features for churn prediction model then the prediction process outcome may conclude in fault predictions and insights. The Feature Selection Element takes a balanced dataset as an input and should results in the features that consist of the prediction dataset.

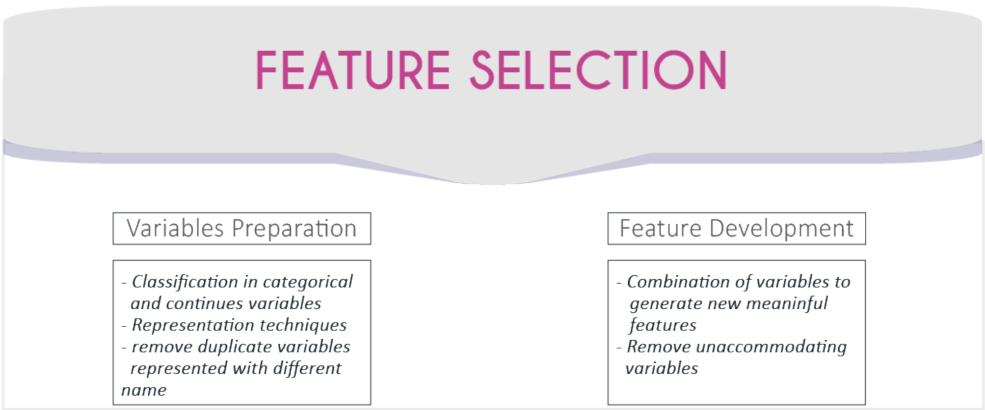


Figure 3.23: Proposed Feature Selection Element Analysis

The Feature Selection Element could include actions related to variables preparation. More specifically processes that cope with the classification of variables into categorical and continued, may support feature analysis processes. Representation techniques could be part of these actions, as the removal of duplicate variables, which exist in the dataset with different names. Moreover, in Feature Selection Element, may be included actions related to feature development. In more detail, processes associated with the combination of variables, which lead to the generation of new meaningful features are proposed in that element. Likewise, processes bound up with the removal of unaccommodating variables may be part of Feature Selection Element.

3.3.1.4 Proposed Prediction Model Element Analysis (E4)

While previous elements of the conceptual framework deal with KPIs, the identification of available data sources and the selection of the suitable features for the churn prediction dataset, this element is responsible to choose a suitable classification algorithm which will be used to predict churners. The Prediction Element takes the dataset eliminated from missing values and balanced variables to perform churn prediction.

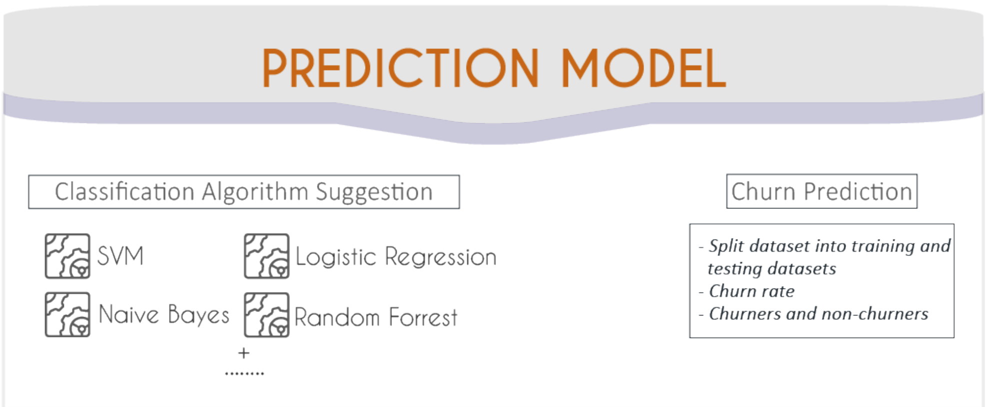


Figure 3.24: Proposed Prediction Model Element

With the aim to identify possible churners, the Prediction Model Element may include actions associated with the selection of a suitable classification algorithm as those presented in Figure 3.24 and the churn prediction. More specifically, the author incorporates processes that may lead to the selection of the appropriate classification algorithm based on the parameters of the churn prediction dataset. Following, the selection of the suitable algorithm, should take place processes related to the splitting of the dataset into training and testing parts and proceed to the churn prediction process. The outcome of Prediction Model Element, should be the churn rate for the subscription-based under investigation, associated with the classification of churners and non-churners.

3.3.1.5 Proposed Insights Element Analysis (E5)

The fifth and last element of the CBFF is the Insights Element, which is illustrated in Figure 3.25. The existence of this element in the proposed conceptual framework aims to add additional value to the whole process of churn prediction. To achieve that, Insights Element should incorporate actions that should be used to evaluate classification algorithm performance. Such evaluation metrics should be among others: i) Recall, ii) Precision, iii) Recall, iv) F-Measure and v) Area Under the Curve (AUC).

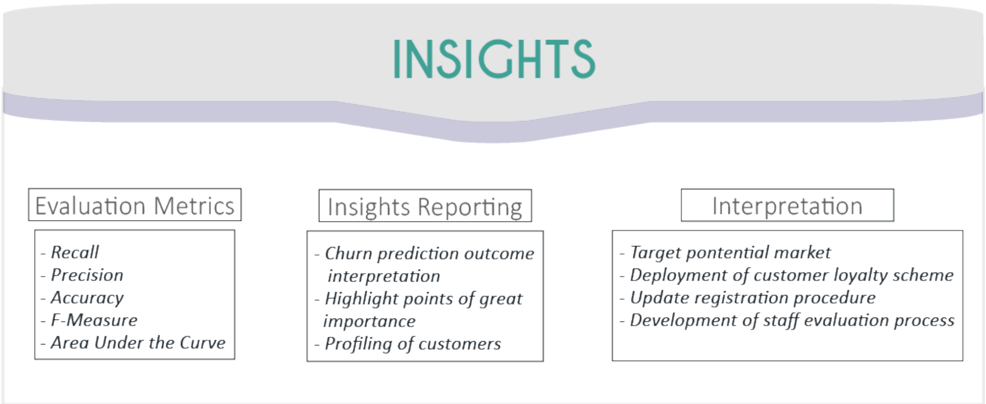


Figure 3.25: Proposed Insight Element Analysis

Furthermore, the most important feature of this element is be the insights reporting. Through the insight reporting the subscription-based organization may be able to get meaningful insights regarding churners behavior and achieved goals. Insights such as the comparisons between desired churn rate KPI noted during the Business Case Analysis and the actual churn rate derived from the last churn prediction attempt should be noted. In addition important points should be highlighted in the report, as well as insights that work towards the development of the churners profile. Insights related to churners profile should function as useful tool in the organization's hands. Such insights, should enhance multiple processes inside the organization including but not limited to: i) better understanding of churners behavior, ii) altering of processes during contracting customers, iii) redesign of retention campaigns, iv) Re-targeting its potential customers and v) incorporation of new features in future churn prediction attempts that work towards the enhancement of churn prediction accuracy.

In summary, the previous paragraphs of this section presented a detailed view of the five elements of the proposed conceptual framework. In doing so, actions associated with each element were discussed and analyzed. The five elements, are synthesized and result in the proposed Customer Behavior Forecasting Framework (CBFF) as this illustrated in Figure 3.26. CBFF, which aims to clarify the processes to be followed by subscription-based organizations for a successful churn prediction.

3.4 Conclusions

Chapter 3 presented the proposed conceptual framework for this thesis, which aims to support the decision making process related to customer churn prediction for subscription-based organizations, as illustrated in Figure 3.26. In Chapter 3, the research challenges derived from Chapter 2 were analyzed thoroughly and discussed. Additionally, detailed proposed suggestions for the open issues were presented. As a result, the churn prediction procedure, is proposed and depict in Figure 3.18. The development of the churn prediction procedure, provides an overview of the cooperating parts of the proposed churn prediction.

The proposed conceptual framework (CBFF) is presented in detail in Section 3.3 of Chapter 3. The five proposed elements namely: i) Business Case Analysis, ii) Data Collection, iii) Feature Selection, iv) Prediction Model, and v) Insights. The proposed Business Case Analysis Element, encapsulates all the business case-related actions and may support a successful churn prediction process, is reflected in Section 3.3.1.1. Moreover, the proposed Data Collection Element, which is responsible for identifying all available and relevant data sources to combine the churn prediction dataset, is covered in Section 3.3.1.2. The third proposed element is the Feature Selection Element, which includes all those actions related to the development of the features list for the churn prediction dataset; this is explained in Section 3.3.1.3. The proposed Prediction Model Element, which hosts the processes related to the identification of the classification algorithm and the churn prediction task, is presented in Section 3.3.1.4. Finally, the proposed Insights Element, which is responsible for the interpretation of the outcomes, is demonstrated in Section 3.3.1.5. The proposed conceptual framework CBFF, is illustrated in Figure 3.26. The next chapter (Chapter 4), will present and justify the research methodology adopted to test the proposed CBFF.

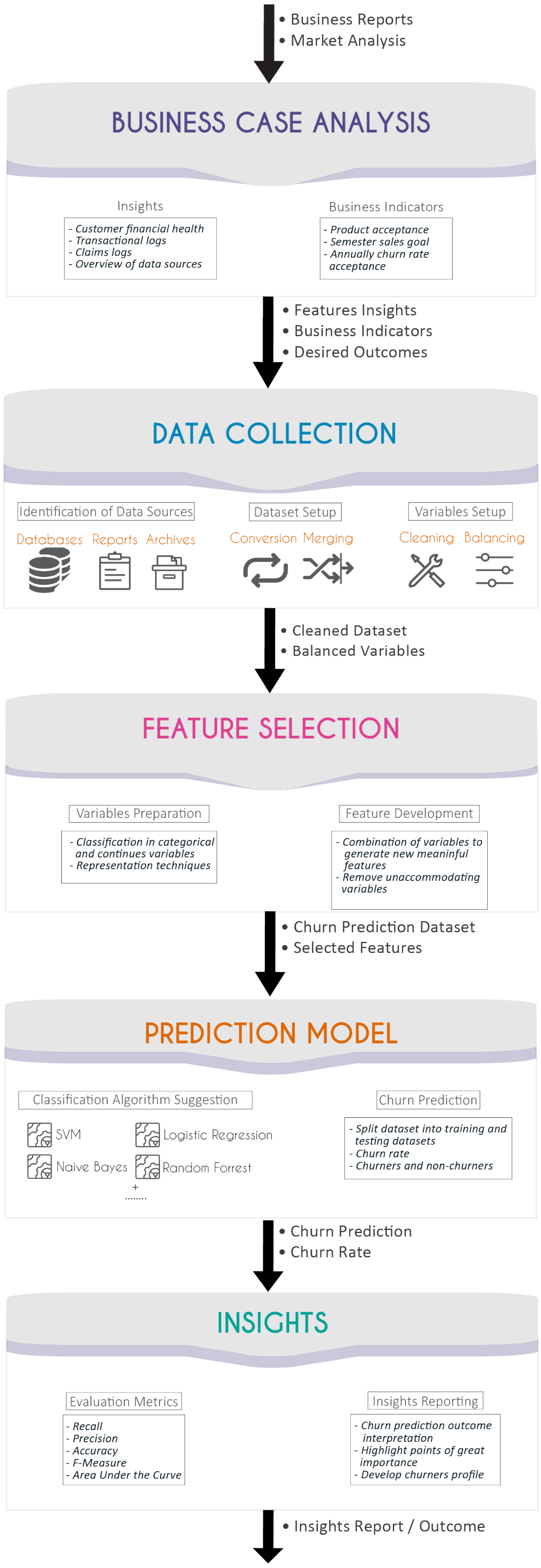


Figure 3.26: Proposed Conceptual Customer Behavior Forecasting Framework

Chapter 4

Research Methodology

Absence of evidence is not evidence of absence

Carl Sagan (1934-1996)

Summary

Based on the research gaps identified in Chapter 2, Chapter 3 introduced a conceptual framework for customer churn prediction. Chapter 4 aims to identify a suitable research methodology with which to test the proposed conceptual framework. The chapter investigates research approaches and strategies and justifies the selection of the qualitative case study strategy that is used in this thesis.

4.1 Introduction

This chapter aims to assess the research methodologies to identify an appropriate one to be used in this thesis. The structure of Chapter 4 is presented in Figure 4.1, which is inspired by Wilson's research methodology honeycomb [64]. Each cell of the honeycomb in Figure 4.1 illustrates the sections comprising Chapter 4 that are followed to define the research methodology adopted for this thesis. This section outlines the chapter objective and structure, and it is followed by a presentation of research methodologies, accompanied by paradigms, used in the course of this research work. The selected methodology is based on commonly used research methodologies in social sciences.

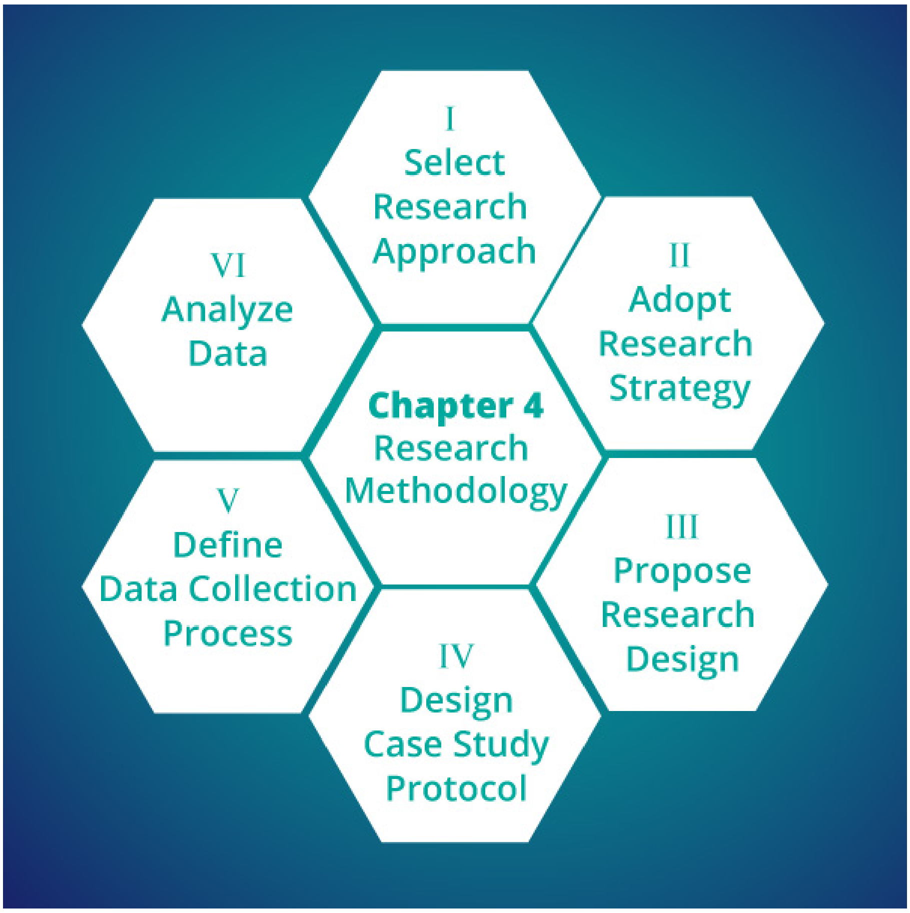


Figure 4.1: Research Methodology Components

The first cell in Figure 4.1, which also denotes the first section of this chapter, covers the selection of the research approach. In this section, the philosophical stances are explained to justify the selection of the stance adopted for this thesis. Additionally, in the same section, a presentation of research methods is made, concluding the justification of the research method adopted during this thesis. Next is the research strategy adoption section, in which suitable research strategies that are used in social sciences, are presented and explained. Also, in this section, the selection and justification of the adopted research strategy are made. Thereafter is the research design section that deals with the proposal of the research design to be followed including decisions such as i) the addressing of research questions during the empirical study, ii) the relevant data, iii) the type of data to be collected during the data collection process, and iv) how to interpret the outcomes to conclude with justified results and conclusions. The author designs the case study protocol for this Thesis based on the decisions taken during the proposal of the research design; this is presented in detail in the Case Study Protocol section of this chapter. Moreover, in the Case Study Protocol section, a case study overview is presented, in which the rules and regulations to be followed in deciding which data would be collected, are set out. The last section, Conclusions, summarizes the discussions, findings, and analysis made in the chapter.

Finally is the data collection process, which explores available data collection methods in the literature and presents the data collection methods used during this thesis. The last step is the analysis of the data, as illustrated in Figure 4.1 and presented in the Data Analysis section of Chapter 4. In the Data Analysis section, the author presents the approaches that are followed during data interpretation including, among others, the data triangulation approach. Data triangulation deals with another challenging issue of the adopted philosophical approach in this research, i.e., the validity and reliability of research findings. Chapter 4 ends with the Conclusions section, in which the outcomes derived from the sections of this chapter are summarized.

4.2 Selecting Appropriate Research Approach for this thesis

This section presents and explains the philosophical stances to justify the selection of the stance adopted during this research work. Various research studies in social sciences, such as those of Galliers and Land [65] and Walsham [66], agree that the selection of the appropriate research methodology is a critical factor for the success of the research work. This agreement is based on the existence of multiple research frameworks that could be selected while researching in social science [67]. The research on customer churn is classified under social science, making the selection of the appropriate research methodology a critical process. In particular, novice researchers could become more experienced by studying the multidisciplinary nature of social science and all of the different fields related to i) natural sciences, ii) mathematics, and iii) behavioral sciences [65]. From this perspective, the following section presents the research methods and examples used in conducting the current research's work framework.

4.2.1 Philosophical Perspectives

Different philosophical perspectives also exist in the social sciences. Selecting an appropriate one to follow during research is a critical and complex process [66]. The different philosophical approaches available in social science are those identified in Orlikowski and Baroudi [68], Goldkuh [69] and Ryan [70]. These research studies include, among others: i) Critical, ii) Interpretive, iii) Positivism, iv) Post-Positivism, and v) Pragmatism. The following paragraph briefly describes the different philosophical stances available in social science to give the reader a better understanding of the selection of the epistemological approach adopted in this thesis:

- **Critical:** The Critical stance emphasizes the social and historical origins and contexts of meaning [71]. The critical approach aims to analyze in detail how knowledge is gained. Accordingly, knowledge is not the result of objective inquiry but is gained through debate and critical disclosure [68, 71]. The critical research approach uses questions that lead to a tool of knowledge being used for the practical transformation of society [71].

- **Interpretivism:** The Interpretive philosophical approach mainly focuses on understanding human experience and actions. The interpretive stance is based on the fact that people are creating and associating their opinions based on their interaction with the world around them and what this is like for them [68, 71]. Researchers that adopt interpretivism try to understand phenomena based on participants' meanings. Interpretivism stands on the non-existence of an objective reality that researchers could discover and others could replicate [66].

- **Positivism:** The Positivist epistemology stance assumes the previous existence of structured relationships in the phenomena under investigation [68]. Positivism is more common in studies that increase the predictive understanding of the phenomena.

- **Post-Positivism:** The Post-positivist epistemology emphasizes the creation of new knowledge, based on the recognition that all observations and theories are reversible. The main characteristics of the post-positivist epistemology include, among others: i) the recognition of the research as being broad rather than specialized, ii) theory and practice being kept together.

- **Pragmatism:** The Pragmatism research approach takes into consideration phenomena that support action. Pragmatism is not aligned with any of the philosophical stances and accepts the value of both the physical and social world [64]. Pragmatism copes with the ”what” and ”how” questions of the research problem [72]. Researchers using pragmatism put the research problem and research questions in the center of the research, using methods that seem to lead to the generation of the most significant insights from their research [64].

Findings from the literature [68] show that the positivism approach was proposed for research in social science [66, 67]. Also, Orlikowski and Baroudi [68] accept the positivism perspective as being suitable for research in social science if testing a hypothesis that is on fixed a priori relationships. Despite the dominance of the positivism epistemology shown in the research studies of the past twenty years [68], recent researchers [64] argue about the loss of interesting insights in the adoption of a positivism stance.

The aim of a critical stance is critiquing the status quo on how the knowledge is gained, making this perspective unsuitable for this thesis, which deals with the understanding of customer churn through the experience of stakeholders of the subscription-based domain. Accordingly, the pragmatism approach is more like a mixture of methods and is more suitable for phenomena that support action.

Taking into consideration the above, together with the fact that the author has to understand the social world of the stakeholders included in the research study, such as the i) subscription-based organizations, ii) key personnel from those organizations, and the iii) situations under which they carry out business, makes the interpretivist approach more suitable for investigating the churn prediction in the subscription-based domain.

The reasons that lead the author to adopt the interpretivist stance in this thesis are justified below:

- Chapter 2 presents the literature review on customer churn, in which issues on how stakeholders of the domain understand the customer churn in organizations and the different techniques used achieved a decrease in churn rate, are factors that could be investigated under an interpretive perspective.

- The proposed conceptual framework presented in Chapter 3 identifies complex managerial (e.g., decision-making from different persons inside a subscription-based organization) and technical (e.g., classification algorithms, many different techniques) issues.
- The nature of the research problem under current investigation, such as the customer churn in subscription-based companies, does not present a clear hypothesis for testing, which makes the positivist approach inappropriate for this research study. Similarly, the critical approach could not be selected for this research work as this research emphasizes the customer churn domain stakeholders' experience and their actions rather than mutual learning and self-reflection [71].

4.2.2 Qualitative vs. Quantitative Research Methods

Various research methodologies exist in literature and are classified in various ways, but the most common classification is between *"Qualitative"* and *"Quantitative"* research methods [64, 73]. Multiple definitions for Qualitative research methodology can be found in the literature, such as that provided by Corbin and Strauss [74] who describe qualitative research methodology as:

> *" a form of research in which the researcher or the designated co-researcher collects and interprets data, making the researcher as much a part of the research process as the participants and the data they provide".*

Furthermore, Fossey et al., [71] inspired by the research work of Corbin and Strauss [74], refer to qualitative research methods as an umbrella term for research methodologies that describe and explain experiences, behaviors, interactions, and social contexts. Additionally, Fossey et al. [71] highlight the importance of research on the question of greatest importance.

By contrast, quantitative research methods are described as the research methods that typically gather data in a numerical form, classified into categories or ranked, or measured in units of measurement. Quantitative research methods support or refuse a hypothesis regarding a phenomenon and its relationships by testing it, quantifying opinions, behaviors, and attitudes. In doing so, answers to questions on relationships are made using measurable variables [75]. Similarly, Denzin and Lincoln [76] refer to the quantitative research methodology as the methodology that emphasizes:

> *" the measurement and analysis of causal relationships between variables, not processes"*

Furthermore, in studies using quantitative research methodology, there is an attempt to construct generalizations that consider the population of interest as a whole [77]. Usually, research based on quantitative research methodology ends with a confirmation or refusal of the hypothesis investigated. Nevertheless, there are researchers that support the combination of both research methods, rather than using them exclusively [64].

Mack et al., [78] state the significant differences between quantitative and qualitative research methods as: i) the different structure in questions they follow, ii)their analytical objectives, iii) the types of data collection tools they use, iv) the forms of data they produce, and v) the degree of freedom in the research study design, as presented in Table 4.1.

	Qualitative	**Quantitative**
General Framework	Explore phenomena from the point of view of the actor	Tests hypothesis about phenomena and ultimately supports or rejects it
	Highly structured methods including questionnaires, surveys and structured observation	Semi-structure methods such as in-depth interviews, focus groups, and participant observation
Analytical Objectives	Describe variation	Attempt to predict casual relationship
	Describe and explain relationships	Quantify variation
	Describe individual experiences	Describe characteristics of a population
	Describe group norms	
Types of Questions	Open-ended	Close-ended or Fixed
Forms of Data Produced	Related to a specific temporal or spatial domain	Emphasis on the collection of metric data
Freedom in Research Study Design	Flexible aspects of study design (e.g., wording of particular interview questions, additions etc.,	Study design based on statistical assumptions and conditions
	Interviewing procedure affect the order and which of the questions the researcher ask	Stable research study design from the start

Table 4.1: Comparison of Qualitative and Quantitative Research Methods [Adapted from: [78, 79]]

A comparison in terms of Qualitative and Quantitative Methods reveals a key difference in terms of flexibility in understanding the research problem [78]. Quantitative research methods often use surveys and questionnaires for all participants (e.g., the same questions and the same order of questions). Inflexibility, in the response categories for participants between "close-ended" or "fixed", allows the researcher the option of meaningful comparison of responses across all participants. Nevertheless, the choice of the right questions to include in the questionnaire is a critical task of the quantitative research method.

Conversely, qualitative research methods are more flexible, allowing the researcher the opportunity to better adapt to the interaction between the participants of the research.

Moreover, qualitative research method questionnaires are often characterized as "open-ended". Open-ended questions include more complex responses than the "yes" or "no" met in the close-ended or fixed questionnaires of quantitative research methods. Additionally, interviews in qualitative research methods enable a less formal relationship between the researcher and the participant of the study [78]. Additionally, qualitative research methods provide identification of intangible factors including: i) social norms, ii) socioeconomic status, iii) gender roles, iv) ethnicity, and v) religion, which may not be readily noticeable [78].

4.2.3 Justifying Research Method Selected

This thesis aims to investigate the churn prediction in subscription-based organizations. In doing so, a conceptual framework will be developed to support churn prediction related actions, such as the churn prediction accuracy increment and the enhancement of decision-making related to churn outcome. Based on the characteristics of the Qualitative and Quantitative methods, the qualitative research method is adopted in this thesis as it is more appropriate for this research work. The main reasons for selecting the qualitative research method and the interpretivism as the epistemological stance are summarized below:

- Taking into consideration the aim and objectives presented in Section 1.2 of this research work, where the author has: i) to review the literature related to churn prediction for the subscription-based domain, ii) to develop a conceptual framework, iii) to evaluate the proposed conceptual framework and iv) to extrapolate empirical findings, the author investigated the literature to choose an appropriate research method and has decided that the adoption of a qualitative research method is more suitable for this thesis. More specifically, the experiences of stakeholders from subscription-based organizations should be collected in a way that better describes the individual characteristics of each specific case. Such characteristics are collected using a qualitative research method instead of a quantitative one.

- Customer churn is a pertinent and challenging issue that affects not only the growth but also the existence, of the subscription-based organizations, as is discussed and analyzed in Section 2.2. In addition, the literature review findings which were presented in Section 2.6, show that investigation of churn prediction needs a better understanding and analysis of the related processes that take place during the development of a new churn prediction approach.

- The various methods and techniques used for previous churn prediction approaches, as discussed and analyzed in Section 2.5, show that the examination of the phenomenon under its natural setting is recommended, together with the need to learn from practice [80], making the qualitative research method more suitable for this research work.

4.3 Research Strategy

Following the justification of the epistemology stance as interpretivism and the qualitative research methodology for this research work, this section presents the available research strategies. This section aims to select an appropriate research strategy. Benbasat et al. [81] refer to goals and objectives of research work as the nature of the research topic and as the reasons that lead to the selection of a specific research strategy.

Various research strategies can be used when conducting qualitative research such as those summarized below:

- **Action Research:** This type of research strategy is used in applied research, attempting to identify an effective way to demonstrate the knowledge gained in a partially controlled environment [82]. Action research combines observations with participation, as the researcher in the study has to be part of the research instead of merely observing and recording [83].

- **Ethnography:** This research strategy is derived from anthropology, where the researcher attempts to identify patterns in human activities based on socially acquired knowledge [64]. Ethnography researchers interpret the social world as being members of the participant group studied [64]. More specifically, the Ethnographic Research Strategy participants are studied in their cultural setting, aiming to understand how participants behave in their own terms.

- **Participative Inquiry:** During a participative study, participants are involved as much as possible in the research study, which takes place in the same organization or its own group [64]. Participants in the participative research study drive the direction and process during the study that is conducted from them [64].

- **Grounded Theory:** Grounded Theory strategy constructs a theory based on the actual data. More specifically, the data collection process precedes the development of the hypothesis. Therefore, the theories and hypotheses derive from the collection and analysis of the data [64]. This strategy was first conceived by Glaser and Strauss in 1967 [84] as a reaction to positivist studies [64].

4.3.1 Case Study Research Strategy Selection

According to, Collis and Hussey [82] and Benbasat et al. [81], the case study strategy attempts to study a phenomenon (the case) in its natural environment using a set of methodologies to gather data and information from one or a few stakeholders. More specifically, Benbasat et al., [81] summarize the main characteristics of case research are:

- The phenomenon under investigation is examined within its real-life context.
- Various methods are supported for data collection.
- The focus is on recent events.
- One or a few entities is/are examined.
- There is no need for prior specifications of the variables to be tested.
- No experimental controls or manipulation are involved.
- The complexity of the unit is studied intensively.
- It seems more suitable for the exploration, classification, and hypothesis development stages of the knowledge-building process. The investigator should have a receptive attitude towards exploration.
- The results derived depend heavily on the integrative powers of the investigator.
- Questions ”why” and ”how” become useful as they deal with operational links to be traced over time rather than with frequency or incidence.

Similarly, Yin [85] characterizes the case study as an empirical inquiry that:

- Examines in-depth and within its natural form a phenomenon, and it seems more suitable in the case where the boundaries between the phenomenon and context are not easily observable.
- Data is collected via multiple methods such as interviews, observation, questionnaires, and prior research material.
- The data collection and analysis is guided based on the prior development of theoretical propositions.

Ryan et al. [70] and Yin [85] identify four different types of case studies, depending on the research questions they are used to answer such as *what, why,* and *how* as presented in brief below:

- **Exploratory:** An existing theory is used to understand and analyze what is happening. Exploratory research focuses on a research problem with little, or without, prior work. An exploratory case aims to develop better insights for a particular topic, concluding with the development of a set of hypotheses to test and future research directions [64].
- **Descriptive:** In this kind of research, the study attempts to describe already known phenomena or past phenomena through observation [64]. Descriptive case studies are used to describe current practice [70].
- **Illustrative:** This kind or case study is used where the research aims to illustrate new and possibly innovative practices adopted by particular companies [70]
- **Experimental:** Experimental case studies are used where the research examines difficulties in introducing new procedures and techniques in an organization and evaluating the value from this [70].

Taking into consideration the aim of this thesis, which is to study customer churn prediction in subscription-based organizations and the evaluation of the proposed framework, exploratory case studies are adopted in this research work. The justification for the adoption of exploratory case studies is based on reasons such as i) the suitability of an exploratory case study to understand and analyze the churn prediction challenge, and its valuable support in the development of concepts for future investigation, ii) the aim of this research study, which is to answer *what* questions on customer churn prediction (e.g., what are the effects of churn in subscription-based organizations), making the exploratory cases more suitable, and iii) the various approaches identified during the literature review, which did not provide a clear solution but confused things, regarding customer churn prevention, raised the need for an in-depth understanding of the research topic to provide suitable solutions.

4.3.1.1 Single and Multiple Case Studies

During the research strategy selection, a critical step is considered to be the decision for the case study design to be adopted. Therefore, following, the selection and justification of the research strategy as an exploratory case study, the author decided between single or multiple case designs. This section presents and discusses the available case study research designs, and justifies the selection of multiple case study research as being suitable for adoption in this research work. Therefore, a case study research strategy should be: i) single or ii) multiple [86]. A brief, description of each case study design is given in the following paragraphs.

A single case study design allows more in-depth investigation of a phenomenon, providing rich primary data from the case under investigation [86]. For that reason, a single case study: i) helps in the deeper and more detailed description of a rare phenomenon [86] that contributes to knowledge [85], ii) supports theory building by rectifying and developing concepts [87], and iii) for testing theory, deciding on the rejection or confirmation of the theory [88]. Despite the single case study design providing more detail, there is always the possibility of there being insufficient data [85]. In that case, the single case study is not able to justify the findings for this research work.

By contrast, using multiple cases may be more suitable for verifying findings derived from a research study as they do not include the idiosyncrasies of the research setting [80]. In addition, the adoption of the multiple cases design enables the analysis of empirical data across cases [86]. Moreover, there is not a pre-defined number of cases to study, as this depends on the knowledge gained from the study of a case and also from the contribution in terms of information that may be derived from the study of further cases [89].

Therefore, the current research work adopts the investigation of multiple cases as this seems more suitable. Multiple cases should offer insights to describe, develop, and test a theory for customer churn prediction that the single case research design could not achieve. However, the usage of the exploratory type and multiple cases strategy should support the study of customer churn in the subscription-based domain in more depth, helping the author to understand and analyze the phenomenon of customer churn. To that end, the evaluation of the proposed framework should be tested across the two case organizations selected for study during this thesis, which are presented in the next chapter.

Despite the challenges faced during the data collection process, the author faced a new obstacle that needs to be mentioned at this point, as it also supports the adoption of the multiple case design for this research. The new GDPR EU regulation [90] for data protection raises new challenges for researchers. Due to GDPR [90] regulation, researchers face complex obstacles in the collection of empirical data, such as those explained briefly in Section 5.1 upcoming Chapter 5. These obstacles may lead to extreme delays during the data collection process, problematic access to rich data, or even the rejection of the selected case for reasons that are not related to the suitability of the case in contributing to the evaluation of the proposed conceptual framework, but are due to GDPR [90] regulation disagreement. As a result, the adoption of a single case design for the investigation of customer churn prediction includes additional risks related to the availability of sensitive data for analysis from the case organization. To that end, the adoption of the multiple cases design minimizes the risk of the nonavailability of rich data due to the GDPR [90] regulation.

In conclusion, the multiple cases study research design is adopted for this thesis, as it seems more suitable to support the aim and objectives of this research work on customer churn prediction in subscription-based domain organizations. For the purpose of this research work, two case organizations were studied, encoded with the names BlueTelco and OrangeTelco; these are presented in detail in Chapter 5. The next section presents and discusses the empirical research methodology stages.

4.4 Empirical Research Methodology

Based on Themistocleous [91, 92], empirical research methodology includes three stages: i) Research Design, ii) Data Collection, and iii) Data Analysis as presented and analyzed in this section. The following sections present and analyze each of the steps referred to. Additionally in Figure 4.2.

4.4.1 Research Design

The research design indicates the first stage of the empirical research methodology as this presented in Figure 4.2. During the research design the author:

- Conducts a literature review to build an understanding of the research topic and identify the relevant literature to study. Chapter 1 introduces the research theme, while Chapter 2 reviews and classifies the literature to be studied in relation to customer churn in subscription-based organizations.
- Identifies issues derived from the literature review that need to be addressed. Chapter 3 presents the issues of customer churn that have been derived from the literature review in Section 3.2.1.

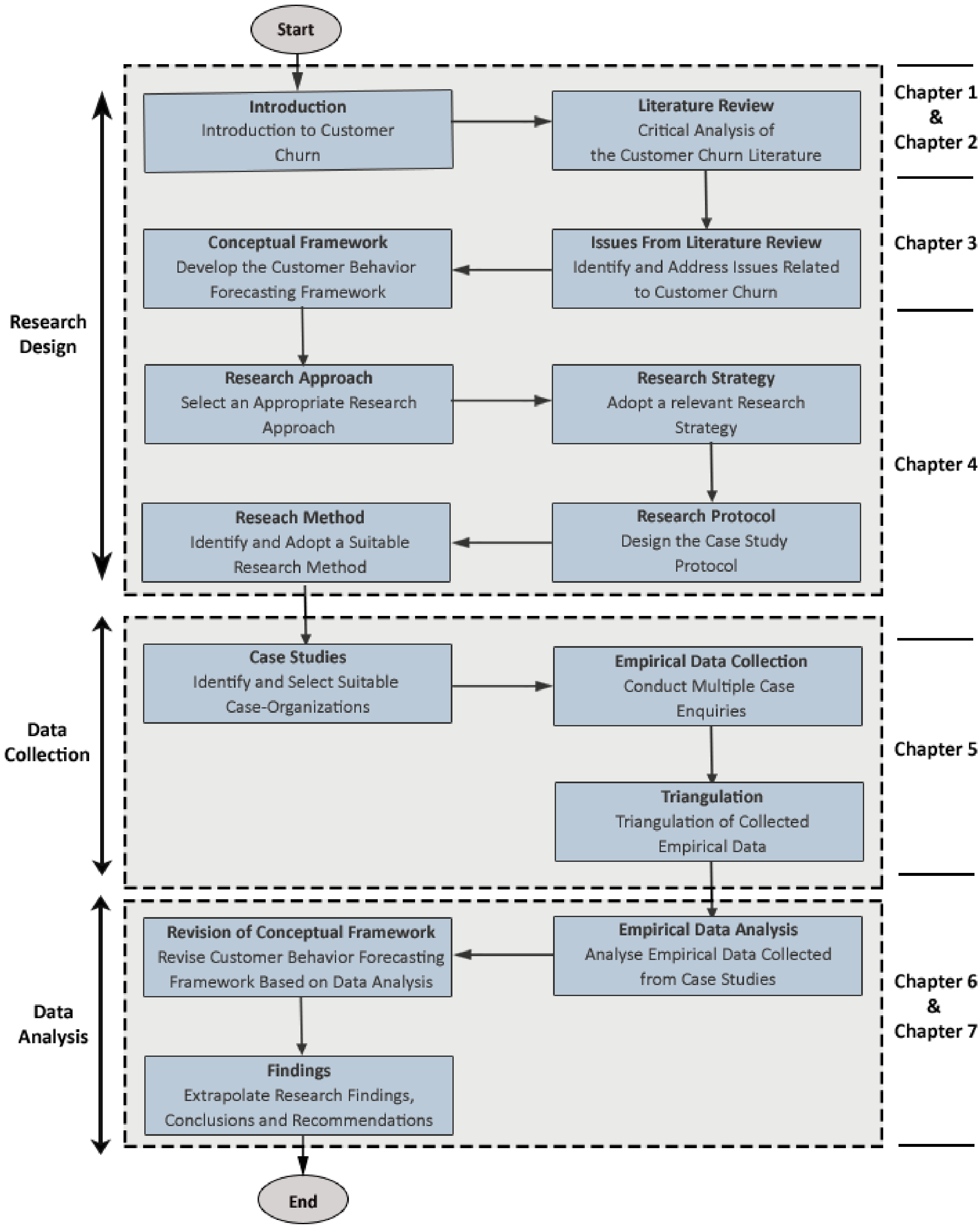


Figure 4.2: Empirical Research Methodology Stages

- Proposes a solution to address the issues derived from the literature. The proposed conceptual framework is reflected on in Section 3.3, namely CBFF with five elements that need to be tested and evaluated.

- Decides on the way that they are going to collect the empirical data and how to analyze the findings. Chapter 4 presents and discusses the steps followed by the author to i) select an appropriate research approach, which is presented in Section 4.2, ii) to adopt a research strategy as reflected on in Section 4.3, and iii) to design the research protocol, which is presented in Section 4.5

4.4.2 Data Collection

The second stage of the empirical research methodology is the data collection, which is illustrated in Figure 4.2. Throughout the data collection stage, actions take place such as:

- Identification and selection of case studies. The author based this on the research protocol that was designed during the research design stage, selecting the case organizations referred to as BlueTelco and OrangeTelco. The two case organizations are presented and analyzed in the next chapter of this thesis and, more specifically, in Case Study One: BlueTelco and Case Study Two: OrangeTelco.

- Empirical data collection is carried out using suitable data collection tools as described in thedata collection tools section below. Qualitative research includes multiple data collection tools for the collection of the primary data; many of these are used in this research work. By the term primary data, the author refers to the data which have been collected for this research work [64] and are combined with the secondary data [64] (e.g., data already published, other research works about customer churn prediction and customer behavior forecasting) to conclude the research findings of this thesis.

- Use of data triangulation approaches, such as those reflected on in the Data Triangulation section, is to reduce the risk of systemic biases and the chance of associations with the case under investigation [64].

4.4.2.1 Data Collection Tools

The use of multiple data collection tools by the author enhances support of research findings and also reduces the risk of bias. Among others, Yin [85] refers to multiple data collection tools as those that include: i) documentation, ii) archival records, iii) interviews, iv) observation and v) physical artifacts. Furthermore, Table 4.2 summarizes the standard data collection tools referred to in the literature [85], together with the use of each tool in the current research work.

Data Collection Tools	Description	Employed in this Research Work
Documentation	Documentation data collection technique is stable, which means that it could repeatedly be review. Additionally, it is broad in coverage and span in an extended period, including many events and settings. The negatives in reporting, identified in the incomplete documentation or bias of the author in selectivity.	Organization report, reference material from web sources, white papers
Archival Records	This data collection technique is similar to documentation technique, hence includes positives and negatives identified in documentation data collection method. Archival records may refer to organization charts, financial or personnel documents.	Charts, datasets, presentations, layouts
Interviews	One of the most common data collection tools used in research studies. Interviews, enable access to information from the research topic using key stakeholders of the domain under investigation. The communication with people working on the subscription-based domain, give to the author a perspective to customer churn based on people's inside the organization experience although the bias generated during the interviews should be given special attention from the author.	Structure interviews, semi-structure interviews, unstructured interviews
Direct Observations	The aim of this data collection tool is the capture of users actions. The researcher, observes subscription-based workers or stakeholders and notices actions, events, and any other relevant information from their real environment, such as their offices.	Formal and informal meetings with subscription-based stakeholders.
Participant Observation	Similar to direct observation data collecting method, but the researcher, in that case, is identified as the researcher filling a real role in the environment under investigation accordingly.	Simple observation
Physical Artefacts	These data collection methods, attempt to gain insights into cultural features and technical operations.	Infrastructure components, software

Table 4.2: Data Collection Tools Used Throughout This Research Work

Interviews

In consideration of the significance given to qualitative reviews throughout the literature [64, 73], the author gives more detail of interviews as a data collection tool. Interviews allow the researcher to have a close relationship with people close to subscription-based organizations. Similarly, Walsham [66] addresses the essential benefits of interviews as being: i) the access they allow to the participants' understanding of actions and events on the research problem and ii) the perspective that the participants and other stakeholders of subscription-based organizations have about customer churn.

Among others, Robson [93] identifies three, distinct interview types as i) unstructured interviews, ii) semi-structured interviews, and iii) fully structured interviews. A brief description of each type is presented in the following paragraphs.

- **Unstructured Interviews:** This type of interview consists of questions that are not previously arranged. The questions are formulated by the researcher, based on their interest or the discussion topic.

- **Semi-structured Interview:** Semi-structured interviews include prearranged questions, but the order in which they are asked is not predefined or limited. During the interview process, the researcher decides the order, or the questions, from the question list. This kind of interview supports the exploration of studied issues focusing on the specifications of each case.

- **Fully-structured Interview:** This type of interview includes prearranged questions that follow the same order as that planned. Fully-structured interviews are similar to the survey questionnaires used in other data collection methods.

Additionally, the form under which the interview takes place varies and includes, among others: i) face-to-face interviews, ii) focus group interviews, and iii) telephone interviews. [64, 78]. Accordingly, the duration of the interview process is also specific and could last from five minutes, if it is a telephone interview, to multiple sessions with stakeholders of the case organization [64]. Regarding the interview type adopted throughout this thesis, the author based the interviews on the unstructured type.

The interview agenda for this research work is summarized in Appendix 3. Regarding the semi-structured type of interview followed by the author, each section consists of multiple questions, formed in a way to guide the interviewer to the semi-structured interview format, as well as allowing the interviewee to express their opinions without biased responses. The interview agenda focused on collecting data from specific segments as follows: i) general organization data, ii) participant data, iii) customer churn prevention, iv) proposed CBFF, and v) comments. Each of the segments referred to is analyzed in the following section, and questions comprising each segment are presented in Table 4.3.

- **General Organization Data:** The purpose of this section is to collect data related to the organization under investigation including, among others, data such as i) the nature of the organization, ii) the critical targets of the organization, ii) the number of employees, and iii) the number of departments.

- **Participant Data:** The purpose of this section is to collect data regarding the interviewee such as i) age, ii) sex, iii) position in the organization, and iv) contact details.

- **Customer Churn:** The purpose of this section is to collect data related to churn and how the organization understands this term. More specifically this section aims to collect data such as i) software and methods used for churn prevention, ii) ongoing projects on churn predictions, and iii) any other actions used for measuring or facing churn in the organization. Questions in this section are customized for each case study to meet its unique parameters.

- **Customer Behavior Forecasting Framework:** During this set of questions the researcher aims to collect data related to i) elements consisting of the proposed conceptual framework, namely CBFF, ii) suitability of the proposed framework within the organization, and iii) any other information regarding similar frameworks followed by the organization.

- **Comments:** This section aims to collect general comments related to customer churn prediction.

Section	Section Title	Questions
A	General Organization Data	A.1 - A.6
B	Participant Data	B.1 - B.4
C	Customer Churn	C.1 - C.7
D	Customer Behavior Forecasting Framework (CBFF)	D.1 - D.10
E	Comments	E.1

Table 4.3: Interview Agenda Overview

4.4.3 Data Analysis

The last stage of the empirical research methodology is the data analysis involved with the interpretation of the collected empirical data as illustrated in Figure 4.2. Actions taken during data analysis include, among others:

- The empirical data analysis. Following, the data collection process the author had to extract the transcription of recordings and field notes taken as soon as possible. In doing so, the author rearranged the transcript data to better serve the research study (e.g., grouping together answers related to the same element of the proposed CBFF), simultaneously maintaining data coherence. Moreover, the nature of the qualitative research methodologies, analysis, and comparison of the raw qualitative data is challenging [78]. However, to ensure data validity and study outcomes, the author had to organize the data in a rigorous and standardized way [78]. Concerns about the validity and reliability of the derived conclusions are often related to ”Triangulation”, which aims to validate the findings from the research work [94, 95]. Triangulation is explained below in theData Triangulation section.

- The refining of the conceptual framework. Interpretation of the empirical data denotes new guidelines and updates to the proposed conceptual framework, which are covered in detail in Section 6.3.
- Findings extrapolation. During the last stage of the empirical research methodology, the author extrapolated the research findings of this research work. In Chapter 7 the author presents the contributions, achievements, and recommendations for future research works.

4.4.3.1 Data Triangulation

The data triangulation approach during the data collection reduces the risk of systemic biases and the chance of associations [64]. Miles et al., [80], describe triangulation, giving an example of working habits followed by police detectives, among others. More specifically, Miles et al. [80] present the procedure followed by a police detective who collects fingerprints, hair samples, alibis, and eyewitnesses that lead to the conclusion that one suspect fits far better than the others. The findings of the previous case involve pattern matching using multiple data sources.

Denzin [94] and Janesick [96] refer to five types of triangulation namely: i) data triangulation; referring to the variety of data sources used in a study [94], ii) investigator; referring to the usage of multiple researchers or emulators [96], iii) theory; referring to the investigation of the phenomenon from different perspectives [94], iv) methodological; referring to the incorporation of multiple data methods to study a single research problem [94], and v) interdisciplinary; referring to investigation issues that are related to more than one discipline [96]. For this research, the author employed methodological, investigator, and interdisciplinary triangulation in both the case studies investigated during the empirical data collection process.

4.5 Case Study Protocol

When carrying out case study research, an essential step is the development of the case study protocol, which enhances the reliability and validity of the case. More specifically, a case protocol maintains the focus on the research topic, as well as containing questions, regulations, and settings for the case under investigation [85]. Moreover, Yin [85] suggests that the outline of the questions should direct the interview process at five levels, as presented in Table 4.4, concerning the interview agenda of this research work presented in Appendix 4. The case study protocol begins with "Case Study Overview", which deals with background information related to the research study, addressing the focus and direction of the research accordingly. Additionally, a set of procedures defined to face "real-life" events often occur in case studies research design. Finally, questions addressed to the author, to support him during the case study investigation and data collection process, are included in the case study protocol.

Level	Case Study Questions	Section Reference
1	Asked for specific interviewees	(Interview Agenda)
2	Asked of individual case study	4.5.1, 4.5.2, 4.5.3
3	Asked across multiple cases inquires	4.5.3
4	Asked of entire case study	(Interview Agenda)
5	About the recommendations and conclusions beyond the scope of the study	(Interview Agenda)

Table 4.4: Questioning Levels adapted from Yin outline

4.5.1 Case Study Overview

This part of the case protocol supports the attempts to cover the background information for this research work. More specifically, details on the issues under investigation during the case study are used to support the author in focusing on the right direction. Considering the purpose of this research study, which aims to investigate customer churn in the subscription-based domain and collect data that support this investigation, the author of this thesis presents the issues to be addressed as follows:

- Identify customer churn countermeasures used in case organizations and the cost associated with these activities.
- Identify departments, people, systems, methods, and techniques associated with churn prediction approaches.
- Identify the benefits of successful churn prediction activities adoption within investigated organizations.

- Identify the suitability of the proposed CBFF in subscription-based organizations.
- Test the elements comprising the proposed CBFF and their relationships.
- Identify the criteria used for customer churn prediction models' evaluation.

4.5.2 Fieldwork Research Procedures

Carrying out a case study research, which deals with the study of a phenomenon in its natural setting, as analyzed in Section 4.3.1, is usually accompanied by "unexpected" events, such as the inaccessibility of interviewees and data documentation. For the author to face these issues, a well-defined set of procedures to follow during the investigation of case studies was drawn up and included the following actions:

- **Define organizations and key persons for interview:** Since customer churn prediction is an emerging research field, knowledge within the organization's personnel is limited or multidisciplinary. Accordingly, the author suggests that marketing department managers, as well as IT managers, should be included among the interviewees. Additionally, managers from the financial department are needed to provide support with data related to the billing information on products/services.
- **Interview appointments and documents requests:** Interview appointments should be defined, and documents requests should be made a respectable time before appointments. There is always the possibility of interviews being rescheduled due to interviewee inaccessibility or the lack of documents, but these events should not prevent the author from collecting data.
- **Define a clear procedure for data collection:** The interview agenda defined and presented in (Appendix 4) will be used for primary data collection, through structured interviews. Organization documents, archival data, meeting minutes, reports, and website data should be used as additional data sources to enrich the data collection.
- **Define confidentiality between involved parties:** Establish confidentiality of shared information between the author and the organizations' interviewees, even with the whole organization being used as a case study. The establishment of confidentiality should be done by signing a confidentiality agreement document between the involved parties that ensures the scientific approach and explains what information might be published for scientific reasons.

4.5.3 Addressed Issues for Investigation and Output Format

During the data collection process, the author needs to focus on the target. In doing so, a set of proposed issues should be addressed for investigation during the case study, to support the author i) to develop the interview agenda, ii) to maintain concentration in the right direction while collecting data, and iii) to remind them of the aim and goal of this research work. Table 4.5 summarizes the addressed issues for investigation. The proposed issues for investigation are exclusively for the researchers, and for that reason, they are not disclosed to the interviewee. The author reviews issues proposed for investigation before each particular interview.

Following the data collection process, the author, in Chapter 5, presents and analyses the empirical data analysis, accompanied by the analysis, outline, and format of the case study outcome. Additionally, issues relating to large amounts of data collected from each case study were taken into consideration by the author. The author attempted to demonstrate empirical results based on the conventional methods used by other IS research studies, while also using a style that a broader range of readers could understand.

Issue	**Issue Description**
Customer Churn Rate	How customer churn rate is measured in an organization.
Customer Churn Prediction	Customer churn prediction adoption. Identify systems, methods, factors influence adoption decisions.
Retention vs New customers	Identify actions taken by the organization to keep current customers (retention) and actions for attracting new customers, including benefits, barriers, cost associated with churn prediction adoptions.
Customer Behavior Forecasting Framework	Proposed framework elements should be tested validated, through multiple case studies' scenarios.

Table 4.5: Issues to be Addressed During Case Study investigation

4.6 Conclusions

Chapter 4 presented and analyzed the steps followed by the author to identify a suitable research methodology for testing and evaluating the proposed conceptual framework. The chapter started with the selection of an appropriate research approach in Section 4.2 by investigating the epistemological stances used in the social sciences. Then, in Section 4.2.1, the author justified the selection of interpretivism as the most suitable approach for the research theme. In Section 4.2.2, a description was given of two of the most common classifications of research methodologies: qualitative and quantitative. Section 4.2.3 presented a justification of the qualitative research methodology as being the most suitable approach for fulfilling the defined objectives of this research.

The available research strategies used in the social sciences were described in Section 4.3, together with a justification for the adoption of the appropriate research strategy for the exploratory case study in this thesis. More specifically, the author started Section 4.3 with a presentation of the research strategies used in the social sciences; then, in Section 4.3.1, he referred to the case study research strategy adopted in this thesis and justified the reasons for his decision.

The next section of Chapter 4 covered the empirical research methodology phases followed during this thesis, which were also illustrated in Figure 4.2. In more detail, the author presented and analyzed the empirical research methodology stages in Section 4.4 as: i) Research Design (Section 4.4.1), ii) Data Collection (Section 4.4.2), and iii) Data Analysis (Section 4.4.3). In Section 4.4.2.1, the available data collection tools were described, and Table 4.2 summarized those used in this research. The interview, which is considered as one of the most used data selection tools in qualitative research studies, was covered in more detail in Section 4.4.2.1, together with a justification for the semi-structured interview being used in this research.

The last part of Chapter 4 addressed the case study protocol considered in Section 4.5. An overview of the case study topics was presented in Section 4.5.1, in which the author described the issues to be addressed by the case-study organizations. Then, Section 4.5.2 presented the fieldwork research procedure actions that could be taken to address the real-life problems identified during the case studies, such as the inaccessibility of interviewees and data documentation. Finally, in Section 4.5.3, issues were proposed for investigation (exclusively for the author) to support the data gathering procedure during interviews, and the format of the expected research output was described.

Chapter 5

Empirical Data and Research Findings

To know, is to know that you know nothing. That is the meaning of true knowledge

Socrates (469 BC - 399 BC)

Summary

Chapter 5 deals with the empirical data collection and the evaluation of the proposed conceptual framework. Empirical data gathered from the case studies are described, and analyzed in this section. The proposed conceptual framework presented in Figure 3.26 is evaluated through the use of two case studies Both case studies, come form telecommunications providers and the author refers to them using the coded names BlueTelco and OrangeTelco, respectively. The findings validate the conceptual framework as well as the customer churn prediction procedure (Figure 3.18) and they also suggest further additions to it. As a result, the next chapter (Chapter 6) will deal with the revisions of the proposed framework.

5.1 Introduction

Chapter 3 focuses on the investigation of open issues derived from the literature review and resulted in the proposition of a conceptual framework for customer churn prediction, namely CBFF. CBFF is a synthesis of five elements reported in the various articles in the literature with new elements proposed by the author. All these elements, were orchestrated in a novel way to result in an innovative framework. The proposed framework, seeks to enhance and extend practices followed by subscription-based organizations on customer churn prediction. In doing so, CBFF enriches the information, knowledge and visibility that organizations have on potential churners. Thus, it enhances and support the decision making process. Chapter 4 resulted in the identification of a suitable research methodology. In doing so, the author adopted the interpretivism qualitative research methodology and the multiple cases study research strategy as being more appropriate to test CBFF.

In Chapter 5, the author focuses on the empirical data collected from the two case organizations. Based on the case study protocol that was described in Section 4.5, the author identified and selected two suitable case organizations to be investigated regarding customer churn prediction, to test and evaluate the proposed conceptual framework (i.e., BlueTelco and OrangeTelco). The BlueTelco case study is covered in Section 5.2. Likewise, OrangeTelco, is presented in Section 5.3. Both sections of the two case organizations cover the empirical data collection process followed. The chapter ends with conclusions derived from each case study analysis.

The author, during the empirical data collection process, faced multiple obstacles. The most important, which is worth mentioning, is the GDPR EU regulation [90] for data protection. The GDPR [90] regulation inside EU has forced companies to reexamine their practices regarding data sharing. Due to that, the author started a long period of discussions with the legal departments of the multiple organizations selected for this research work and, unfortunately, all organizations declined to participate except for two. The author signed a Non-Disclosure Agreement with each organization but, unfortunately, one canceled the collaboration because of the fear of breaking the GDPR [90] rule. This caused additional delays and, finally, after a year, the author managed to collect data from another organization. It is evident that GDPR [90] regulations raise new and complex challenges for researchers in collecting rich data for their research studies, especially in cases where sensitive clients' data need to be shared. The challenges of the GDPR [90] are beyond the scope of this research, which is why the author will not expand on it but merely state it as being a concern for future research studies.

5.2 Case Study One: BlueTelco

This section focuses on the first case study investigated during this thesis. The next subsections seek to: i) present background information related to the organization under investigation, ii) describe customer churn related actions, accompanied with benefits and barriers associated with decisions taken, and iii) test the suitability of the proposed conceptual framework (CBFF) and its elements. Due to confidentiality reasons, the author uses the coded name BlueTelco to refer to the first case organization.

5.2.1 BlueTelco Background

BlueTelco is located in an Eastern European country and is a key player in the telecommunications market. BlueTelco is one of the fastest-growing telecommunications providers in the country, offering broadband, fixed telephony, television services, and, more recently, mobile telephony through the cable infrastructure that it owns and the mobile telephony network. The company has over 22% of the market share in the broadband market and 10% in the fixed telephony market, based on the 56% coverage it had at the time this research work was written.

BlueTelco invests significant amounts in building its pioneering Hybrid Fiber Coaxial (HFC) cable network, which is considered to be a ”next-generation network” to offer broadband speeds that can exceed 400Mpbs per connection. Based on that, BlueTelco provides the fastest broadband connectivity in this specific Eastern European country, with Internet speeds reaching up to 300Mbps. During its first steps, the main focus was the provision of cable TV, which has become the leading service along with its high-speed Internet packages. It currently offers very competitive television content packages, including broadcast rights for the most popular football teams in this Eastern European country. It has recently paid more attention to its mobile phone service, which is its newest service.

Its objectives include, among others: i) to be ranked in the most innovative digital Telcos, ii) to continue to roll-out a super-fast fixed-line network to cover all urban areas in the country, iii) to continue to invest in TV content and local sports, iv) to invest in next-generation TV experiences, v) to invest heavily in establishing its mobile network and increase its mobile market share, and vi) to significantly improve its overall customer experience.

5.2.2 Customer Churn Prediction Actions in BlueTelco

During the data collection process, the author collected: i) business reports, ii) archival records, iii) datasets, and iv) interviews for the BlueTelco case study. In addition to this, the author conducted interviews with key stakeholders in BlueTelco. Based on this data, the organization's challenges and how they are related to customer churn are presented and summarized in Figure 5.1.

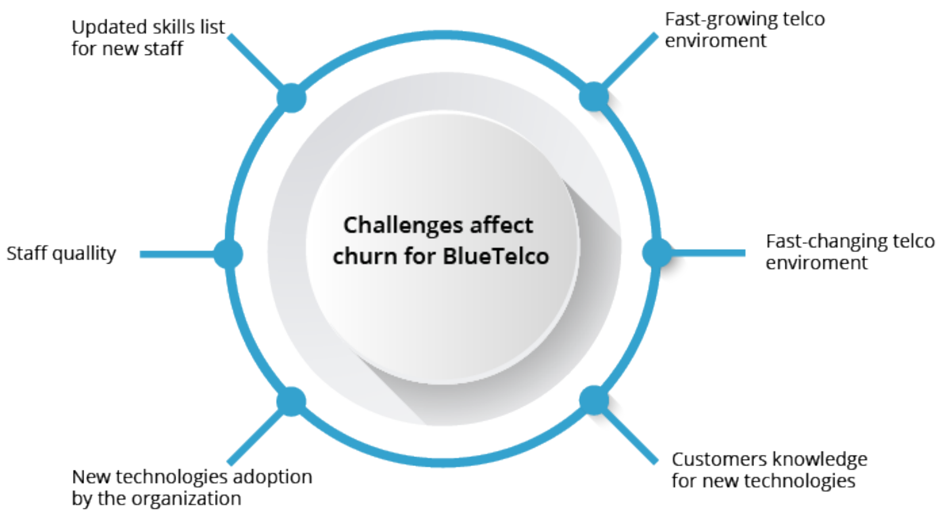


Figure 5.1: Challenges faced by BlueTelco that affect Churn

Initially, the author asked the interviewees about the current challenges. The *Intelligent Manager (head of market research, business intelligence, reporting, and product development)* at BlueTelco stated:

> *"The current challenges faced by the organization are IT and digital transformation. The IT challenge tests the readiness of the company, and the digital transformation tests the readiness of the customer to change and adapt to a new environment and approaches."*

On the same subject, the *Operation Manager* stated:

> *"The main challenges faced are: (i) the national roll-out of our fixed and mobile networks (nationwide), (ii) the continuous improvement and updating of our operations and business sub-systems (as telecommunication's domain technology is changing fast), (iii) keeping up with Human Resource (HR) requirements (new skills and re-learning skills), and (iv) maintaining a relevant and value-for-money product and service proposition (highly competitive market landscape)."*

Analyzing the replies from key personnel in BlueTelco, it is evident that the fast-growing and fast-changing environment of the telecommunication domain is very challenging for organizations and their efforts to retain their customer base. Additionally, the difference between new technology adoption by telecommunication organizations and customers' knowledge about new technologies is an important indicator for BlueTelco in its churn prevention efforts.

The empirical data (interview data, archival records, etc.) indicates that for BlueTelco, staff quality is essential in achieving the organization's goals. Likewise, one of the interviewees (the Operation Manager) referred to the need for continuous staff development and working with HR to identify the new skills needed when hiring future staff.

Another issue that was discussed with the interviewees was the existence of a definition for account health (loyalty). All of the interviewees replied negatively but shared the information for ongoing processes to analyze customer lifetime statistics to define a health index. More specifically, the *Intelligent Manager* responded:

> *"Currently, we don't have a definition for customer loyalty, but we are in the process of researching our customer lifetime statistics, and based on this, we will attempt to define a health index at the subscription level."*

Despite the current nonexistence of a definition for customer health, the organization is working hard to classify it because such an indicator would provide meaningful insights regarding retention strategy. It is important to mention here that for every organization, the definition of such an indicator (i.e., customer health) differs depending on the domain in which the subscription-based organization is located and on preferences like location, competition, maturity of the market, etc.

Therefore, essential insights from BlueTelco's ongoing work to define its health indicator were shared as these insights show the impact on churning customers. BlueTelco classifies its churning customers into healthy and unhealthy churners, as shown in Figure 5.2.

Based on the knowledge gained from BlueTelco about its thought interactions and experiences with its customers classified as healthy churners, those who left the organization did so because of: i) relocation, ii) a better offer in the market, iii) end of the current contract, or iv) new subscription on the same account. Before moving to unhealthy churners,

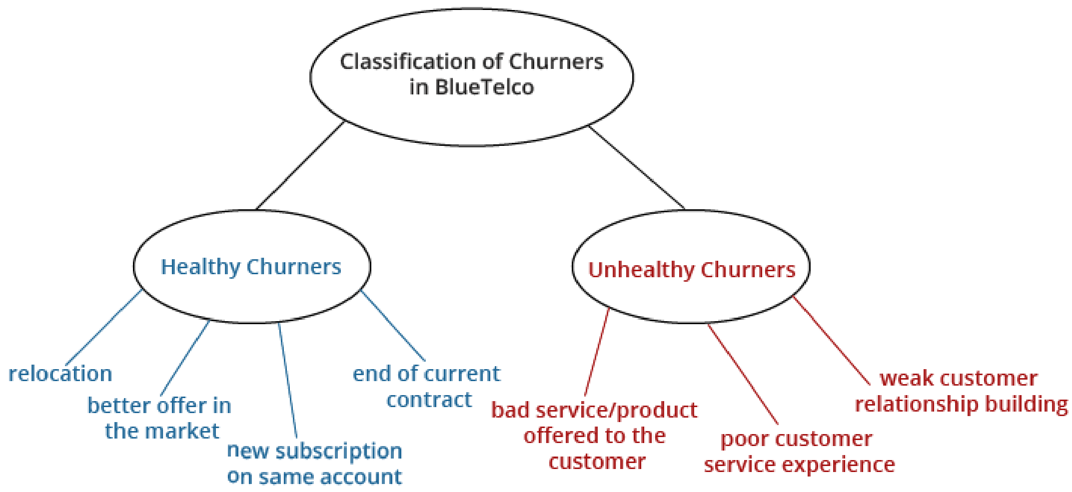


Figure 5.2: Churners Classification at BlueTelco

it is worth clarifying some of the questions raised about healthy churners. Telecommunication providers use multiple and complex systems to support their operations, which leads to additional problems in synchronizing their systems or logging all the actions. To explain these issues, the author presents an example, as described by key personnel from BlueTelco, alongside the reports gathered during the data collection process.

When a customer relocates and there are no BlueTelco services in the new location, the customer terminates the contract with the organization as there is no other option. The churner is identified as a rational or incidental churner based on findings from the literature review [1]. When customers receive a better offer in the market, BlueTelco classifies them as healthy churners because the organization cannot offer the same or better than their competitor. Customers are also classified as healthy churners when their BlueTelco contract comes to an end because the customer will no longer use the service or product. Finally, healthy churners can also include those who terminate a subscription but sign up for a new one under the same account.

To clarify the last reason BlueTelco classifies a churner as healthy, the author created a scenario, which is illustrated in Figure 5.3. In the scenario, a father initially creates a new subscription in his account for his children during their studies (Subscription ID3). He then decides to close this subscription after his children complete their studies but creates a new subscription (Subscription ID3A) for his daughter's new rental house. In this case, BlueTelco's systems count the closure of Subscription 3 as a churn, which is wrong as the customer does not actually churn from the organization.

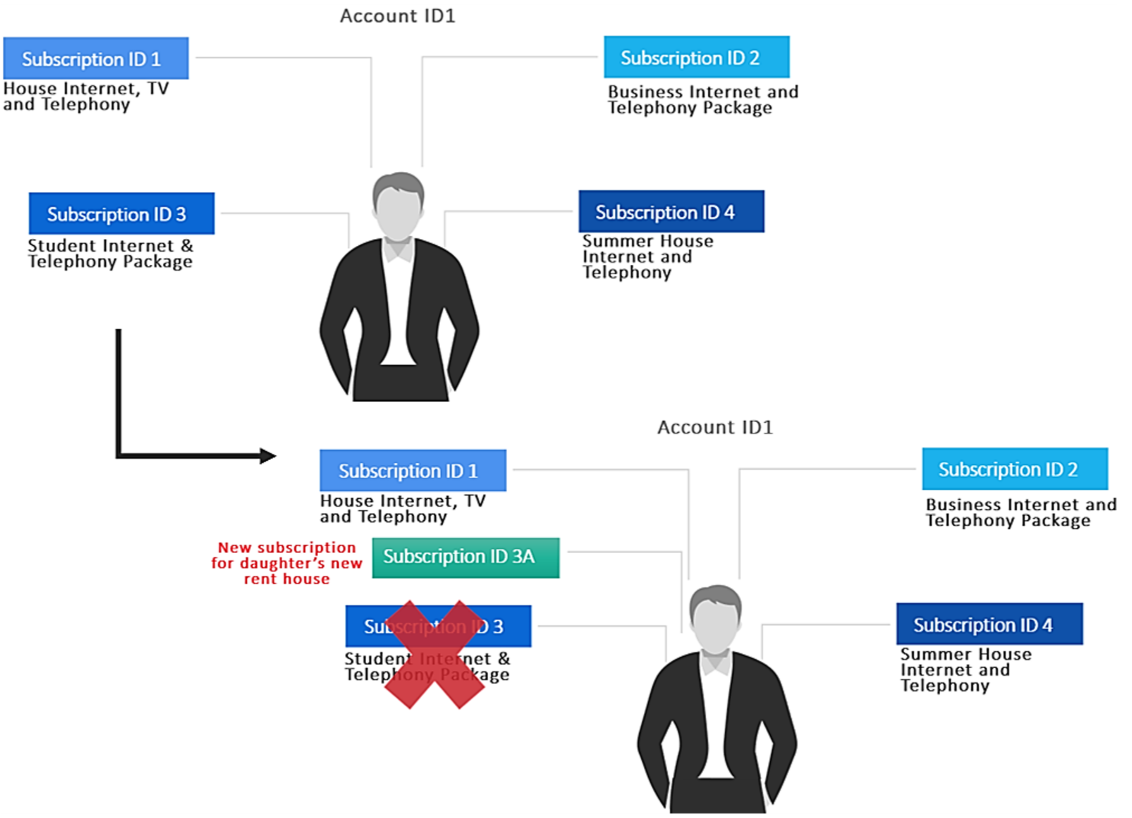


Figure 5.3: Scenario 1: Customer Closes Subscription and Opens New Under the Same Account

Additionally, a second scenario, which is illustrated in Figure 5.4, was created to refer to integration issues in BlueTelco. In this scenario, a father terminates his daughter's subscription due to graduation. The daughter then creates a new account and a new subscription once she moves to her new location. In this case, BlueTelco's IT systems count one churn and one new customer acquisition, instead of one new customer acquisition only as they cannot connect the accounts and subscriptions held in their IT systems.

BlueTelco clarified that an unhealthy churner can include someone who has received: i) a bad service/product offer, ii) poor customer service experience, or iii) weak customer relationship building. When the offered product/service fails to meet customer expectations, they will probably leave and become a negative churner, which reflects negatively on BlueTelco. A customer can also leave because of poor customer service or have a poor relationship with the company, which derives from a lack of monitoring of customer health indicators.

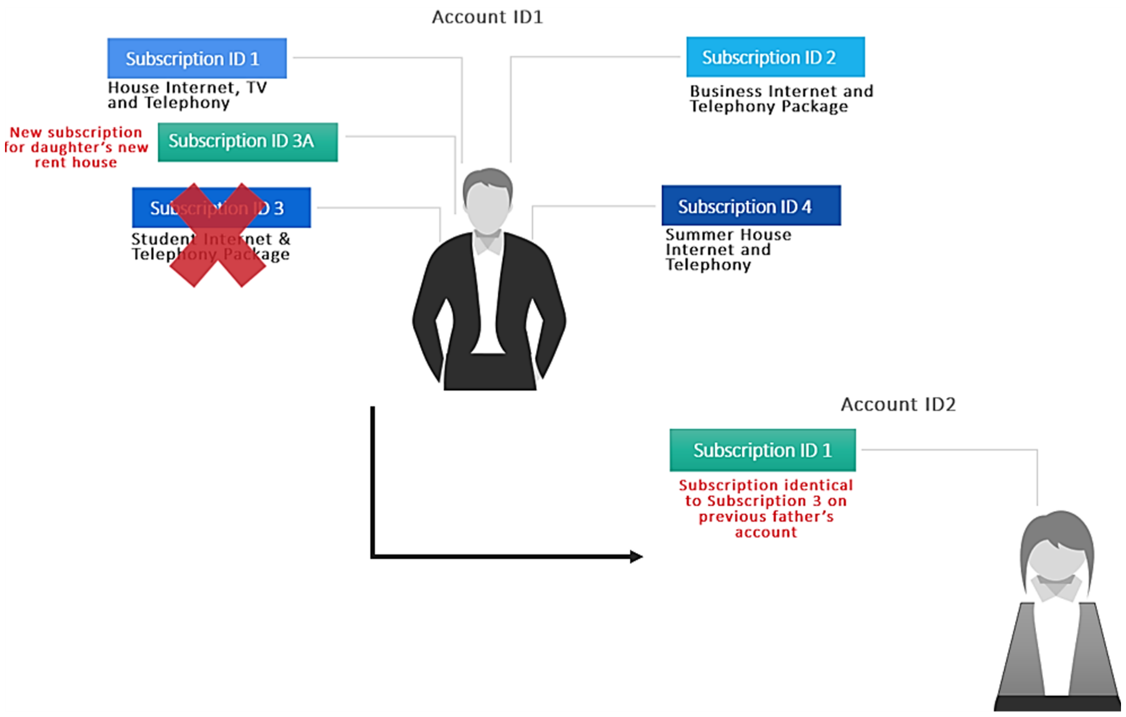


Figure 5.4: Scenario 2: Customer Closes Subscription and a New Account is Created by Previous Subscription Owner

Furthermore, based on findings derived from the literature review on customer churn and reflected in Section 2.5, organizations use multiple methods and IT systems to track their churn rate. Therefore, the author attempted to investigate what methods and systems BlueTelco used to predict their churn rate. In doing so, empirical data was collected and analyzed, including archival records, business reports, and interview replies from critical personnel responsible for managing such efforts inside the organization. The *Intelligent Manager* provided detailed answers when responding to the question about the existence of methods and systems for customer churn. More specifically, the *Intelligent Manager* noted that they predict churn at the aggregate level using a multiplicative seasonality model, which is described as:

$$Yt = Tt * St * It \tag{5.1}$$

T=Secular Trend, S=Seasonal Trend, I=Irregular Value

The market research department is responsible for running this model at BlueTelco and plans to involve the business intelligence department in the upcoming year. In terms of the parameters that constitute the multiplicative seasonality equation, seasonality trends are shown as time-series data [97] that i) show a long-term up or downstream in the data that is not always linear and might change direction as time passes, and ii) are affected by seasonal factors, such as the time of the year or the day of the week, respectively. The next parameter, namely, the irregular value, is also used in the time series and describes data that follows a temporal sequence with varying time intervals [98].

In addition to the multiplicative seasonality model, BlueTelco uses the SAS Business Intelligence (BI) [99] system to obtain information about customer activities. The organization's strategy department looks at the SAS BI's customer retention framework [99] to gain insights that can be included in their marketing strategy.

Depending on BlueTelco's findings, which are demonstrated in Figure 5.5, the predicted churn rate is based on seasonality parameters as they seem to have a significant effect on the behavior of their churning customers. Furthermore, SAS BI is fed with seasonal parameters to provide useful insights for the structure of retention campaigns. BlueTelco also compares the multiplicative seasonality method with SAS BI to predict churn rate inside the organization.

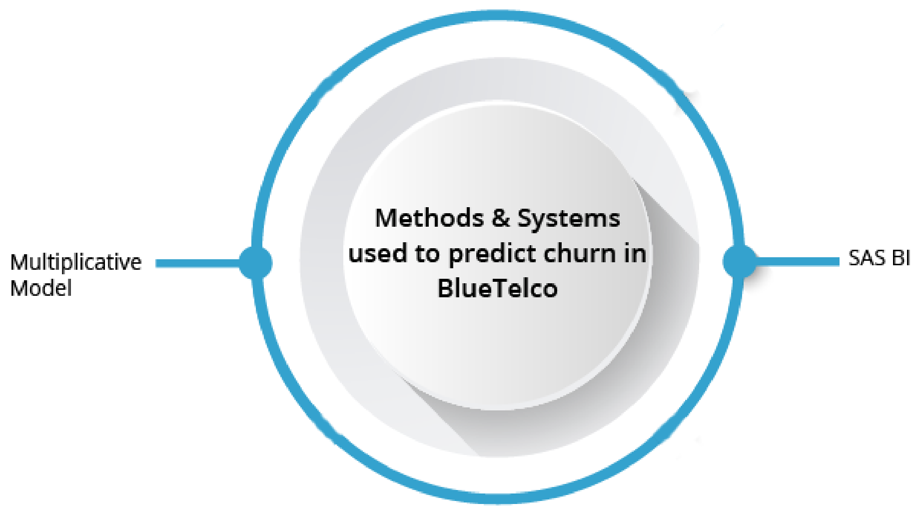


Figure 5.5: Methods and Systems used by BlueTelco to Predict Customer Churn

Following the methods and systems used to predict churn in BlueTelco, the importance of the data used for customer behavior forecasting by subscription-based organizations was investigated as these were proposed in Chapter 3 and the Business Case Element of the proposed conceptual framework. To retrieve all available material from BlueTelco, the author denoted the importance and significance of data analytics as a current research tool. BlueTelco key personnel were asked for their opinions on the following question and also on the proposed *Business Case Analysis, Data Collection, Feature Selection, and Prediction Model Elements* from the proposed conceptual framework of this thesis:

> *"We are living in a data-driven industrial revolution, and data collection and analysis is referred to as 'the new gold' [100, 101] by important players in the market, including, among others, the 'Big Four' [102]. Experienced data analysts strongly believe that no machine learning model can give meaningful insights if the data is not useful. Based on this assumption, how important are business indicators, KPIs, and the organization's desired goals? Should these be taken into account in a customer churn prevention approach and, more specifically, should they be considered in a feature selection decision? Could you provide your opinions on that?"*

Both managers replied, giving meaningful insights on the evaluation of the proposed components. The *Intelligent Manager* responded:

> *"There is some credibility given to the idea that the most valuable predictive data on the issue of churn is not being captured by organizations in the telecommunication sector, which is further complicated by the data restrictions imposed by GDPR. Most subscriptions are terminated without any predictable warning due to the very nature of the telecommunication systems, seasons, and competitive landscape, while only a tiny percentage of churn can be classified as predictable due to the quality of the customer experience delivery with the said product or service. Despite those mentioned earlier, our experience and research also indicate a very high degree of churn predictability when analyzed at the aggregate level by factoring in annual seasonality and secular trends. Churn begins to look a lot more predictable at this layer of abstraction, with supportive analysis into the observable reasons."*

Findings from BlueTelco support the proposed elements of CBFF. The proposed Business Analysis Element suggests a path to identify the most appropriate features to be included in the churn prediction dataset. Based on these findings, if the company's business indicators and desired outcomes are outlined clearly in the proposed Business Case Analysis, then BlueTelco can identify and collect the features to be included in the churn prediction dataset.

This step is essential in the churn prediction life cycle because the proposed Data Collection Element is dependent on identifying and collecting such data. In more detail, the outcome of the proposed Business Case Analysis Element may navigate the actions taken by the proposed Data Collection Element on the collection of more relevant data but also give important insights to the responsible department to define guidelines for the features to be included in the churn prediction dataset. The evaluation of CBFF's proposed elements is based on their importance and value as these were analyzed and backed by interviewees, and additional empirical data was collected from BlueTelco.

The evaluation of the proposed Data Collection and Feature Selection Elements from the proposed CBFF was performed. The role of data types in BlueTelco was investigated, and key staff members were asked for their opinion if there was a best or preferred data type for churn prediction. Replies included essential insights, including this response from the *Intelligent Manager*:

> *"My personal experience points to a weak argument made for the ability of telecoms companies to be in a position to capture best churn data. Typically, a customer profile/segment is compiled based on the collectible data available that would assign a probability of future churn (e.g., tenants have a significantly higher likelihood of churn than property owners) without this being a churn predictor by default. Such segments help companies better target more stable, long-term market segments. However, collecting data to assign a 'vulnerability/loyalty' rating at the subscription level will prove to be much more challenging due to the elusive nature of churn."*

The *Operation Manager* also highlighted two essential categories of data promoted as more valuable for churn prediction in BlueTelco: i) customer life cycle data and ii) customer interaction data. Customer life cycle data includes all features related to the steps the customers go through before they make their first purchase from the organization. Likewise, customer interaction data includes all data related to the customer service experience that takes place just after customer acquisition. Data referred to as customer interaction includes but is not limited to: i) customer service call logs, ii) technical support, and iii) bill-related support.

The above statements are in line with other data collected from BlueTelco and support the proposed elements of CBFF, the Data Collection Element, and the Feature Selection Element. The Data Collection Element's importance derives from the high value of specific data types in the churn prediction process added from BlueTelco. Accordingly, the proposed Feature Selection Element is assigned with a crucial task during the CBFF; it selects the most appropriate features that would lead the proposed Prediction Model Element in the next stage to make accurate and meaningful predictions.

Figure 5.6 demonstrates the ownership status of churners as those identified in the BlueTelco churners' dataset. The data shows that tenants are more frequent churners by 58% than 42% for the owners' category. This statement supports the response from the *Intelligent Manager* that house owners seem to be more loyal compared to tenants. Taking this into account, BlueTelco has developed its future retention strategy.

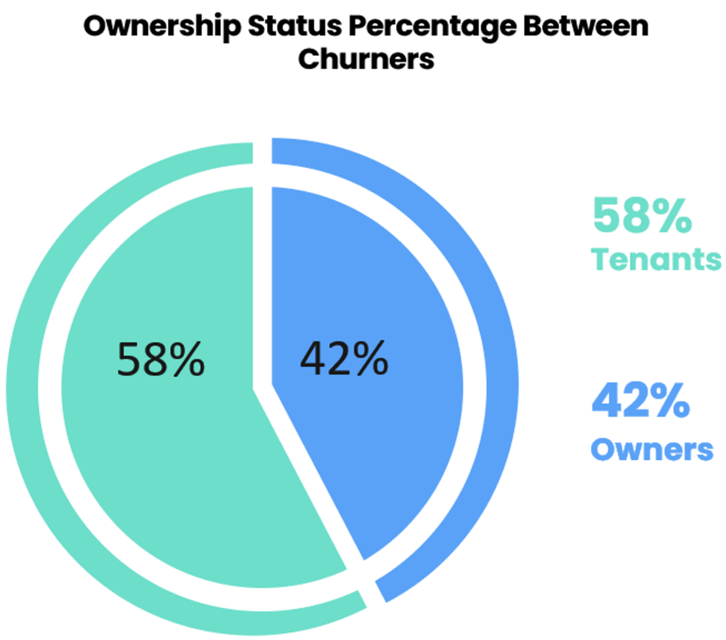


Figure 5.6: Ownership Status by Churner

The author deepened the research on the existence of specific preferred algorithms in churn prediction models used by subscription-based organizations. In terms of BlueTelco's favored prediction algorithm, the *Intelligent Manager* stated:

> *"Again, my personal experience points to a weak argument made for the ability of telecoms companies to be in a position to capture best churn data, which, in turn, can be modeled to predict churn probability (such as using a form of discriminant analysis based on a robust dataset of learning data). This year may yield some better breakthroughs in this type of methodology based on innovative thinking around the topic/model."*

This statement supports other findings from collected material and previous discussions with key personnel from BlueTelco. Once again, it was highlighted that telecommunication organizations and maybe other organizations in the subscription-based domain cannot capture all the valuable data needed for their churn prediction process. Their inefficiency in capturing all the data is because of: i) a lack of knowledge about what data to collect, ii) data privacy regulations (e.g., [90]), iii) incapable staff, and iv) incapable software and methods used inside the organization (e.g., old-fashioned methodologies that are not capable for current trends).

Based on the previous findings related to the identification of a suitable algorithm for churn prediction, the author attempted to identify what parameters led to selecting a specific algorithm to be used for the churn prediction model. When the *Intelligent Manager* was asked about it, he mentioned that:

> *"Parameters we take into consideration are: i) the number of services subscribed to within the company, ii) the status of property ownership, and iii) the usage intensity across broadband and telephony services as a starting point."*

The empirical data reveals that churning happens less often with customers who have more than two services, but owners also show a significant increase in loyalty compared to renters, as discussed in the previous finding regarding the ownership status of customers. Last but not least, the usage of broadband services compared to telephony services denotes a critical indicator for churn. More specifically, the company's most known product is its broadband network offering the highest speed in the country. Therefore, BlueTelco needs to understand how important it is for their customers to have a different provider for broadband services and a different one for telephony services.

More recent data collected from the company showed that customers pay more attention to the quality and speed of their broadband service instead of the fixed telephony service, so customers who choose BlueTelco for their broadband services instead of their fixed telephony products are considered more loyal. Figure 5.7 demonstrates a visualization of the distribution of churners based on the services they owned with the company at the time of the churn. Significant insights derived from the data are the correlation between the number of services owned by churners, where the majority of churners (81%) have two or fewer services with BlueTelco, compared to 19% of those who owned three or more services with BlueTelco.

After investigating the parameters taken into consideration by the organization in selecting the most appropriate prediction algorithm, the author attempted to evaluate the last element of the proposed CBFF, the Insights Element. Therefore, key personnel were asked

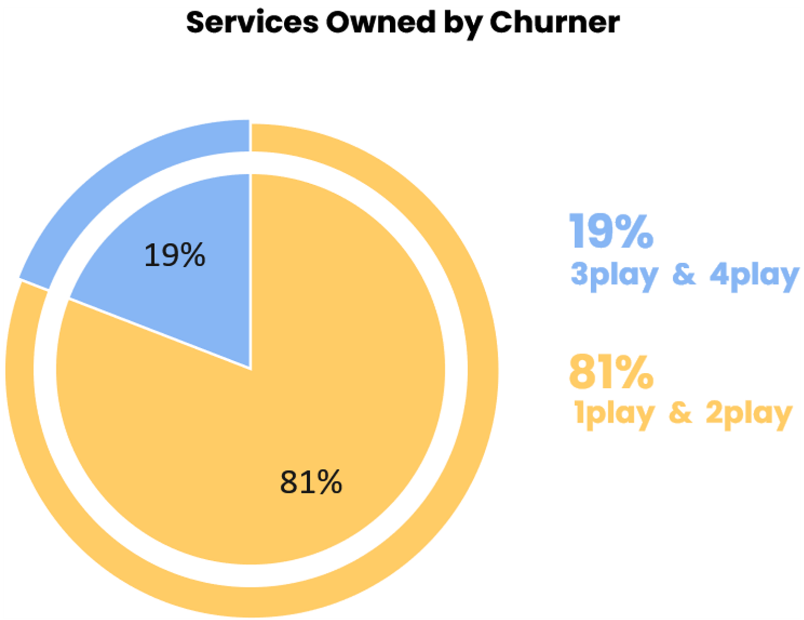


Figure 5.7: Services by Churner

about the importance of insight interpretation in the success of an accurate selection of features for future churn prediction processes. Both the Intelligent Manager and the Operation Manager provided meaningful information:

> *"Churn outcome interpretations are very critical. Interpreting and differentiating between what can be classified as healthy and unhealthy churn is very important. It is clear, for instance, that a significant share of our churn is due to systemic or procedural churn, while an unhealthy churn (i.e., a churn that results from a bad service experience) is critical to be identified as a particular type of churn from all the remaining types."*

> *"If the algorithm to predict churn was there, it would be invaluable to the development of new and innovative features to improve the stickiness of customers and reduce churn."*

Comparing the two replies, it was evident that the two managers understand interpretation differently. Previous evidence is ordinary, and it is related to their positions in BlueTelco. The reply from the *Intelligent Manager* highlights the value of interpretation in identifying healthy and unhealthy churning categories, like those used by BlueTelco to classify leavers. A meaningful interpretation for the *Intelligent Manager* is to classify the churners as healthy or unhealthy, then BlueTelco can design its marketing strategy based on these findings.

On the other hand, the *Operation Manager* had a different point of view regarding the procedures during the churn prediction process. He assumed that if BlueTelco knew the most suitable churn prediction algorithm, then it would not be necessary to develop a new innovative tool or method for prediction outcome interpretation.

Based on the findings from the literature review and empirical data collected from this case study, it is assumed that, currently, there is no comprehensive solution that achieves the best results for every situation in the subscription-based domain. Findings related to the value of interpretation support the existence of the Insights Element of the proposed CBFF.

Following the investigation of BlueTelco, the proposed conceptual framework, as shown in Figure 5.8, was presented to the interviewees. Every element and step contributing to customer churn prediction was explained to them along with the aim of the proposed conceptual framework, which was to connect and orchestrate the different activities taken from various departments and systems inside a subscription-based organization to better understand customer behavior, be proactive in decreasing the churn rate, and help retain customers. The additional benefits of using CBFF to gain insights that would enhance acquisition strategy design were also highlighted.

The data collected from the *Intelligent Manager*, *Operation Manager*, and other sources inside BlueTelco regarding the elements of the proposed CBFF helped the author review and refine parts of the proposed framework, which is presented in Chapter 6. The *Intelligent Manager* commented:

> *"First, to construct a robust churn prediction model, the business case analysis phase should carefully study churn trends at the aggregate level and over the past few years. Correct assumptions must first be made about the nature of churn before any model can be conceived. Second, once a realistic picture of churn can be established, the data capture and collection phase would need to provide an honest account of what data is available to support the churn hypotheses and what data is missing or simply cannot be collected. This should yield the reliability of any churn model to be conceived. If the aforementioned phases can be performed robustly and completely, then the likelihood of success in the later phases will be significantly higher."*

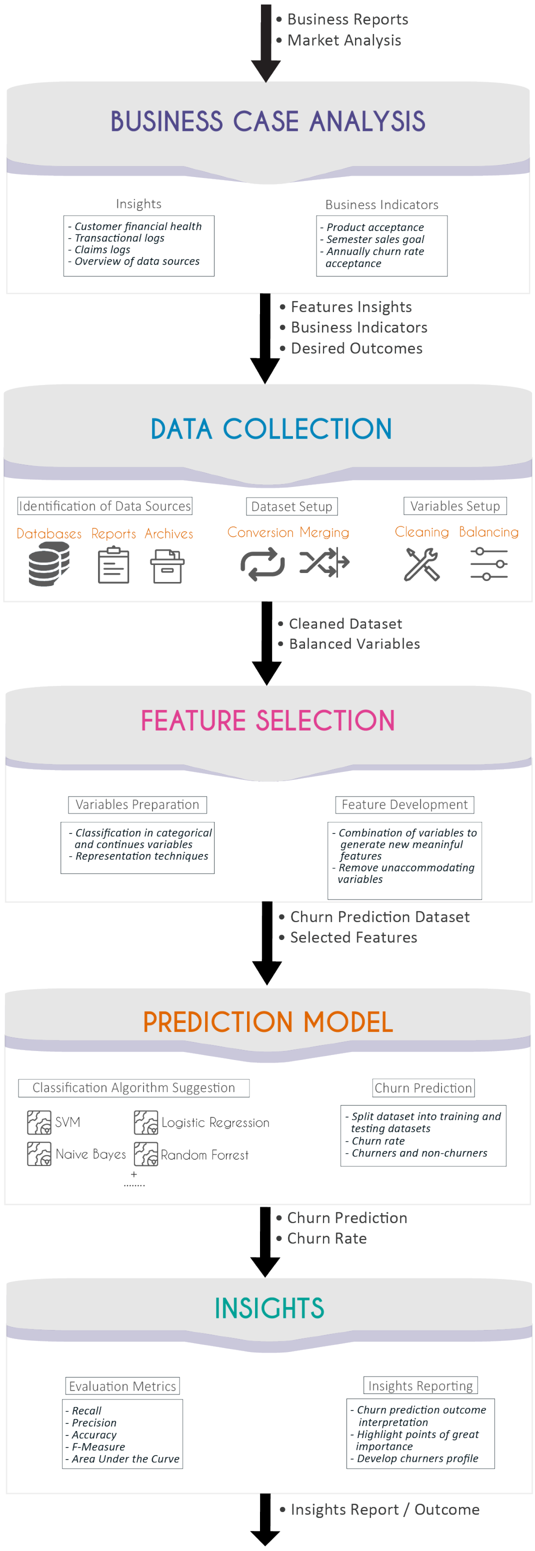


Figure 5.8: Proposed Customer Behavior Forecasting Framework (same as Figure 3.26)

The valuable role of the elements included in the proposed conceptual framework derived from the analysis of data provided by the *Intelligent Manager*. More specifically, the *Intelligent Manager* highlighted the crucial role of the first element (i.e., proposed Business Case Analysis), denoting the correlation between churn trends and different years. Interpreting empirical data collected from BlueTelco alongside interviewees' data, there is strong support in the statement that churn trends vary over time. To support that statement, the author referred to a customer who churned from a telecommunication service provider five years ago due to the best offer in SMS and call minutes.

Nowadays, the same customer churns for another provider who offers more internet data and no SMS. This happens because of the important penetration of the internet into our lives. As a result, internet data has become more useful to people, and telecommunication providers gain more from data roaming services instead of calls and SMS services. This fact is also supported by the continuous improvements in speed for data roaming in mobile networks and the large amounts invested in the new generation of mobile networks with super high transfer rates on data roaming.

Going further in analyzing the *Intelligent Manager's* response, the crucial role given to the proposed Business Analysis Element of CBFF was identified in its success ratio. More specifically, the assumptions made during the proposed Business Case Element determined what percentage of data was received in the proposed Data Collection Element for churn prediction. Assuming that the proposed Business Case Analysis Element outcome provided a comprehensive input to the proposed Data Collection Element, all upcoming elements (e.g., the proposed Feature Selection, Prediction Model, Insights) will have a significant positive probability to predict churners correctly.

Furthermore, the *Operation Manager* tested the CBFF in a real scenario developed for BlueTelco and commented:

> *"Despite having no involvement in the scientific approach to churn management, the framework proposed seems logical."*

Despite his limited involvement in the scientific approach to churn management, the *Operation Manager* identified each element's usage in a real use case scenario for BlueTelco. During the scenario, he advised checking the proposed CBFF to identify each element's value in a customer churn prediction life cycle.

In summary, this section presented and analyzed valuable empirical data collected from BlueTelco, that supported the evaluation of the proposed conceptual framework. The empirical data indicate that, each of the elements included in the proposed CBFF provided value to the churn prediction process of the organization. The following section presents and analyzes other sources of empirical data collected during the BlueTelco investigation; these resources were combined to create the BlueTelco churners dataset, which was used to evaluate additional aspects of the proposed conceptual framework.

5.2.3 Description and Analysis of the BlueTelco Dataset

Various sources were used when collecting empirical data from BlueTelco, such as i) reports, ii) archival data, iii) CRM data, and iv) data from interviews. The empirical data was then described and analyzed to evaluate elements from the proposed conceptual framework. The data were combined to create BlueTelco's churners dataset, which was used to investigate and evaluate different aspects of the proposed CBFF. Table 5.1 illustrates the 11 features included in the dataset.

BlueTelco Features			
RandomID	Segment	PackageName	PackageType
Activated	TerminatedOn	OwnershipStatus	Medium
TerminationType	TerminationReason	ARPU	

Table 5.1: BlueTelco Features in Dataset

The characteristics of this dataset include, among others, data used to predict customer churn. More specifically, the 11 features included in the BlueTelco dataset are described as:

- *Random ID:* The unique customer identifier for each row in the dataset. BlueTelco has 20,365 individual customers.
- *Segment:* The type of area the customer lives in. It has two values: Residential and Rural.
- *PackageName:* The name of the customer's product package. This category includes 37 different packages offered by BlueTelco.
- *PackageType:* Refers to the number of combined services acquired by each customer. This feature has the following values: i) 1play, ii) 2play, iii) 3play, and iv) 4play, as shown in Figure 5.7.

- *Activated:* The date the customer activated their account with the company.
- *TerminatedOn:* The date the customer closed their account with the company.
- *OwnershipStatus:* The ownership status of the customer's residence. Values for this category are Residential and Tenant.
- *Medium:* The medium through which BlueTelco provides its services to the customer. Values included in this category are Cable and DSL.
- *TerminationType:* The termination type. Values for this feature are Termination by Customer and Termination by BlueTelco.
- *TerminationReason:* The actual reason the subscription terminated. This feature's values include 17 reasons as demonstrated in Figure 5.9 and are clustered in three main categories.
- *ARPU:* The Average Revenue Per User (ARPU) in euros (€) for each subscription. ARPU [103] is defined as the total revenue divided by the number of subscribers.

Feature	**Unique Values**
RandomID	20365
Segment	1
PackageName	37
PackageType	4
Activated	2456
TerminatedOn	605
Medium	2
ARPU	26
TerminationType	2
OwnershipStatus	2
TerminationReason	17

Table 5.2: BlueTelco's Churners Dataset – Unique Values per Feature

Table 5.2 summarizes the unique values for each of the 11 features of the BlueTelco dataset. Taking the RandomID unique values, 20,365 denotes the unique churners included in the BlueTelco dataset. When examining the unique values in Table 5.2, the reasons for termination stand out because of the findings presented in Section 5.2.2. There are 17 termination reasons, as illustrated in Figure 5.9, and these are clustered in the following three

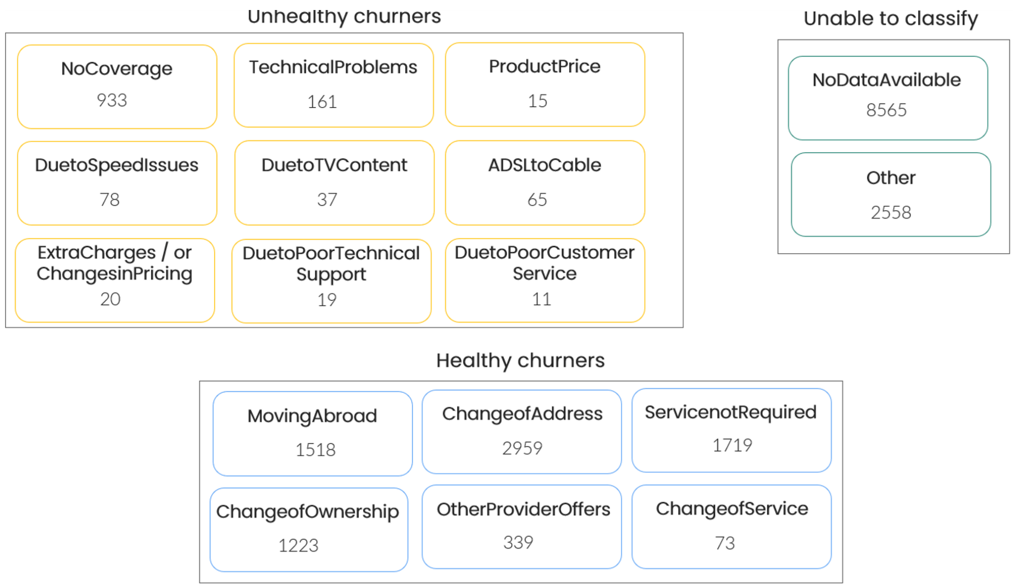


Figure 5.9: Values for Termination Reason Feature in the BlueTelco Dataset

categories: i) unhealthy churners, ii) healthy churners, and iii) unable to classify. The number below each churning cause signifies the appearances in the dataset for each churning reason.

An overview of the churning reasons are summarized in Table 5.3 and shows the percentage of each churning cluster. The three categories are i) Undefined, ii) Unhealthy churners, and iii) Healthy churners.

Undefined / Not Able to Classify - 55%	**Unhealthy Churn - 7%**	**Healthy Churn - 38%**
1. No data available	1. No Coverage	1. Change of address
2. Other	2. Technical problems	2. Service not required
	3. Due to speed issues	3. Moving Abroad
	4. Due to TV content	4. Change of ownership
	5. Extra charges / or changes in pricing	5. Other provider offers
	6. Due to poor technical support	6. Change of Service
	7. Product price	
	8. ADSL to Cable	
	9. Due to poor customer service	

Table 5.3: Clustering Churning Reasons from BlueTelco Churners Dataset

The most significant category is undefined/unable to classify, which includes churners feedback that is described by the dataset's values as other and no data available and represents 55% of the churning reasons. The healthy churner represents 38% of the dataset and is the second most significant category. There are six churning reasons in this category: i) Change of address, ii) Service not required, iii) Moving abroad, iv) Change of ownership, v) Other provider offers, and vi) Change of service. The unhealthy churner represents 7% of the total churning reasons and includes nine reasons: i) No coverage, ii) Technical problems, iii) Speed issues, iv) TV content, v) Extra charges/Change of service, vi) Poor technical support, vii) Product price, viii) ADSL to cable, and ix) Poor customer service.

5.2.4 Summary of BlueTelco Case Study

The empirical data investigation of BlueTelco reveals that the proposed CBFF could be applied to subscription-based organizations to enhance customer behavior forecasting, which could lead to more accurate predictions about churning customers. The findings of this case are summarized in Table 5.4, and demonstrated that, each of the elements constituting the proposed conceptual framework was supported and validated in the case of BlueTelco BlueTelco case.

The proposed Business Case Analysis Element is considered crucial for the success of the customer churn prediction procedure as followed during the CBFF. As reported in Table 5.3, the majority of churning reasons - 55%, remains undefined. This demonstrates the necessity of the Business Case Analysis Element of the proposed conceptual framework, as this will assist organizations in minimizing undefined data and thus support decision making. In addition to this finding, the staff's inappropriate skills and lack of expertise were identified as an important indicator in this element along with the adoption of new technologies and adapting to the fast-growing and evolving telco environment.

Regarding, Data Collection Element, the findings supported its importance and value in a customer churn prediction procedure. Consequently, the outcomes from BlueTelco regarding the proposed Data Collection Element highlighted the impact of data preparation techniques in the enhancement of churn prediction outcome. Likewise, the data type was highlighted as important indicator for an accurate churn prediction, referred to two essential categories of data (referred as customer life cycle data and customer interaction data). Additionally, comparing multiple resources, including data from the corresponding categories, led to more accurate and meaningful predictions for the organization.

The findings for the Data Collection Element are also related to the next element in CBFF (Feature Selection Element). To make a meaningful and valuable prediction, the organization needs to include all appropriate features in the churn prediction process. For this

Proposed Element	Findings in BlueTelco
Business Case Analysis	1. Organizations in the telecommunication sector are not capturing the most valuable predictive data on the issue of churn. 2. Staff quality: Skills required / acquired from staff, are very important in succeeding desired business goals. 3. New technologies adoption and customer knowledge for new technologies affecting successful churn prediction processes. 4. Fast-growing and changing telco environment.
Data Collection	1. Data preparation techniques have significant impact in the accuracy of churn prediction outcome. 2. Both Customer life cycle data and customer interaction data are seen to have high value for BlueTelco.
Feature Selection	1. Meaningful predictions derive from appropriate feature selection. 2. Ownership status matters in churn. 3. Customers with more than two services / products seem more loyal.
Prediction Model	1. There is not a comprehensive solution that achieves the best results for every case in subscription-based domain. 2. The Multiplicative Seasonality Model is used to identify the churning periods of interest. 3. SAS BI uses the outcome of Multiplicative Seasonality Model, alongside with Customer Interaction Data and Customer Life-cycle data to identify churning customers.
Insights	1. The value of interpretation is crucial in success of future churn prediction attempts.

Table 5.4: Findings from BlueTelco that Address Elements of the Proposed CBFF

phase to be successful, the outcome of the proposed Business Case Element needs to include the suitable assumptions and KPIs, and based on the output, the combination of the most appropriate data sources takes place during the proposed Data Collection Element.

As a result of the procedures in the previous elements, the proposed Feature Selection Element may identify all valuable features; for example, ownership status and the number of services acquired from each customer are imperative to telecommunication providers. More specifically, as discussed in section Section 5.2.2 and Figure 5.6, house owners seem to be more loyal than renters. Additionally, the customers who own three or more services from an organization seem to be more trustworthy than those who own one or two services by 81% and 19%, respectively.

Based on the findings derived from BlueTelco for first three elements (i.e., Business Case Analysis, Data Collection, Feature Selection), of the CBFF it appears that the Data Input phase of the proposed churn prediction process is also validated. This also reveals a link or correlation between the Data Input phase (including sub-phases A and B) and the element E1-E3 of the Figure 3.19.

The testing of the proposed Prediction Model Element shows that the multiplicative seasonality model was used to identify the periods throughout the year where churn increased. This model helps BlueTelco to predict churn which implies the validation of this element. In this case, BlueTelco fed the SAS BI system with: i) customer interaction data, ii) customer life-cycle data, and the outcome of the multiplicative seasonality model to identify possible churners. Although, this seasonality model works for this organization, its suitability is unclear for other subscription-based organizations and thus further research is required on this topic. The evaluation of the proposed Prediction Model Element suggested that this element should be revised in terms of the implementation of supporting tools for the element. A detailed description for the revised proposed Prediction Model are given in Section 6.3.

The findings related to the proposed Insights Element denote the value of interpreting the results of future churn prediction to develop an organization's retention campaign. Furthermore, IT systems need to integrate and update to correctly count churning customers and new acquisitions in the case of BlueTelco. Action taken during customer acquisition were also identified as important, including collecting additional data during customer registration, customer service, or other related activity. These actions are driven by the insights derived from a customer churn prediction procedure and more specifically during the Prediction Outcome phase which is also validated from this finding.

In summary, the proposed conceptual framework was evaluated successfully and provided the necessary feedback to consider it a valuable tool for subscription-based organizations against churn. Moreover, the empirical study of each CBFF element revealed possible modifications, which are summarized in the conclusion in Section 5.4.

5.3 Case Study Two: OrangeTelco

This section outlines the second case study included in this research and includes: i) information and a description of OrangeTelco, ii) information related to telecommunication churn dataset derived from OrangeTelco, and iii) testing elements of the proposed CBFF using multiple scenarios. Due to confidentiality reasons, the coded name OrangeTelco is used to refer to the second case organization.

5.3.1 OrangeTelco Background

Customer churn prediction is a critical issue in subscription-based companies, such as Internet service providers, pay-TV companies, insurance firms, and telecommunication service providers. OrangeTelco also comes from the telecommunication sector and has been used to evaluate more technical elements of the proposed CBFF, such as the classification algorithms task included in the proposed Prediction Model Element. More specifically, various classification algorithms were tested using the data collected from OrangeTelco. This case study refers to a company outside the EU, which opens up the possibility of the author accessing more sensitive empirical data without the restriction of GDPR. Despite the absence of GDPR, private and personal data for research purposes is subject to confidentiality terms. The author followed all procedures needed to establish this confidentiality and ensure the records were kept safe during the research.

OrangeTelco is a telecommunications provider that provided customer activity data through the Kaggle platform [104]. The results from the previous case study found that additional testing was needed on specific parts of the CBFF, such as the correlations between specific data types used in the proposed Data Collection Element, the importance of the proposed Insights Element, and the suggested tasks included in the proposed Prediction Model Element.

5.3.2 Customer Churn Prediction Actions in OrangeTelco

OrangeTelco also included churn labels specifying whether a customer canceled their subscription, which is used to identify churners in churn prediction models. Similarly, the features included in the OrangeTelco dataset were classified into the following three categories: i) service-related, ii) demographics, and iii) financial/profit related. More details on these three categories are provided in this section. The OrangeTelco dataset consists of 18 features, which are summarized in Table 5.5.

OrangeTelco Features					
CustomerID	Gender	Partner	Dependents	LoyaltyPeriod	PhoneService
MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
StreamingTV	ContractType	PaperlessBilling	MonthlyCharges	TotalCharges	Churn

Table 5.5: OrangeTelco Dataset Features

- *CustomerID:* The unique customer identifier for each row in the dataset. OrangeTelco has 7,043 individual customers.
- *Gender:* Refers to the sex of the customer. Values for this feature are Male and Female.
- *Partner:* The customer's living status and if they have a partner, with Yes and No as acceptable values.
- *Dependents:* The existence of dependents, with Yes and No as acceptable values.
- *LoyaltyPeriod:* The length of time the customer has stayed with the company (in months).
- *PhoneService:* The existence of phone service, with Yes or No as acceptable values.
- *MultipleLines:* The existence of multiple phone lines. This feature takes values Yes, No, and No Phone Service if the customer does not have phone service.
- *InternetService:* The type of internet service the customer has (or not). Values for this feature include Fiber optic, DSL, and No in the customer's absence of internet service.
- *OnlineSecurity:* The existence of online security service, with Yes and No as acceptable values.
- *OnlineBackup:* The existence of online backup service, with Yes and No as acceptable values.
- *DeviceProtection:* Activation of a device protection plan in case the customer has internet service. This feature takes all three values of Yes, No, and No Internet Service.
- *TechSupport:* The technical support plan for internet service products. Acceptable values for this feature are Yes, No, and No Internet Service.
- *StreamingTV:* The existence of a streaming service, which also requires internet service. Acceptable values for this feature are Yes, No, and No Internet Service.

- *ContractType:* The duration of the signed contract between the organization and the customer. This feature allows Month-to-month, One year, and Two years as acceptable values.
- *PaperlessBilling:* The activation of a paperless billing option activated by the customer for their billing information. Values for this feature are True and False.
- *MonthlyCharges:* Monthly charges for each customer.
- *TotalCharges:* Total charges by the organization for this customer.
- *Churn:* The churn status of the customer. Values for this feature are Yes and No.

The features included in the OrangeTelco dataset can be classified into three basic categories, as summarized in Table 5.6. The service-related category includes nine features, the demographics category has five features, and the financial/profit-related category has four features. The following section includes an analysis of the OrangeTelco data and the results from this analysis, which support the critical role of each element included in the proposed CBFF.

Data Type Classification		
Service Related	**Demographics**	**Financial - Profit Related**
PhoneService	CustomerID	ContractType
MultipleLines	Gender	MonthlyCharges
InternetService	Partner	TotalCharges
OnlineSecurity	Dependents	PaperlessBilling
OnlineBackup	Churn	-
DeviceProtection	-	-
StreamingTV	-	-
TechSupport	-	-
LoyaltyPeriod	-	-

Table 5.6: Data Type Classification for OrangeTelco Features

5.3.3 Description and Analysis of the OrangeTelco Dataset

The analysis of the empirical data in the OrangeTelco dataset reveals essential insights for OrangeTelco's churner profiles. The results show a churn rate of around 27%, while 73%

are still customers, as illustrated in Figure 5.10. More specifically, one out of three people is going to churn from the organization. The following paragraphs will attempt to develop the profile of an OrangeTelco churner.

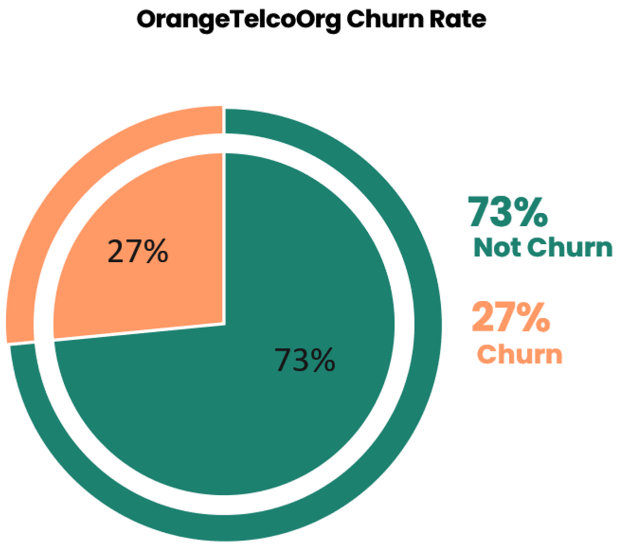


Figure 5.10: OrangeTelco Churn Rate

The gender ratio of the OrangeTelco churner is almost split in half, with 939 female churners (50.24%) and 930 male churners (49.76%), as demonstrated graphically in Figure 5.11. Other demographic features, such as partnership and dependent status provide meaningful insights for the development of a churner's profile. Figure 5.12 illustrates the significance of these churner features and the correlation between them.

In terms of partnership status, from the total of 1,869 identified churners in the dataset, 64.12% declared they did not have a partner, in contrast to 35.79% who declared they did have a partner. Taking the dependent status feature, which is demonstrated in the top and bottom charts on the right in Figure 5.12, most of the churners did not have dependents (82.56%) compared to those who had at least one dependent (17.44%). The assumption derived from these observations is that customers without partners are more vulnerable to churn. The significance of churn increases even more in those who do not have dependents.

The two bottom diagrams in Figure 5.12 demonstrate a combination of the previous features that support the previous observation for churners. The left chart shows that churners without partners or dependents together represent 93.58% of all churners. The right

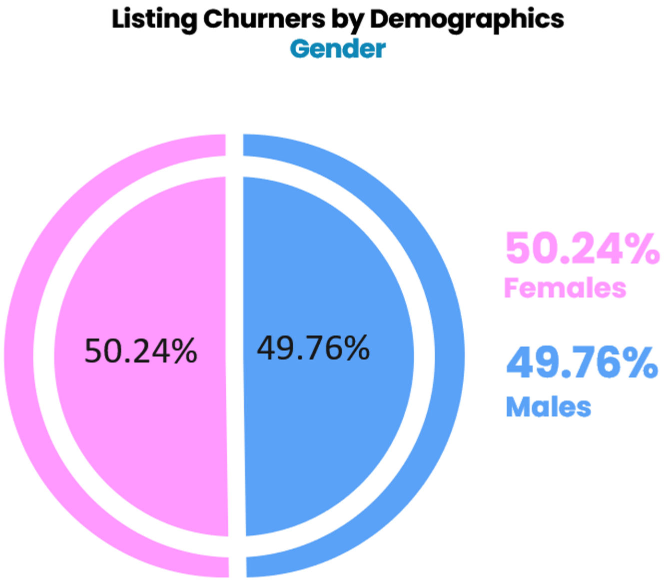


Figure 5.11: Listing Churners by Demographic Features - Gender

chart combines the same features (i.e.., partnership and dependent status) and shows that 37.22% of churners have partners and dependents, compared to 62.78% who have a partner but no dependents. In summary, the OrangeTelco churner could be male or female, possibly without a partner, and is more likely to leave by themselves.

As part of the empirical study, service-related features were also investigated and revealed additional features of the OrangeTelco churner profile. Figure 5.13 illustrates the correlation between the loyalty period in months (horizontal axis) and churners leaving the organization after a specific period (vertical axis). The top-left point indicates that 380 customers left the company during the first month of their contract, and this trend decreases as the customer stays longer with the company; the bottom right corner of the chart only has six churners after 72 months or six years after becoming OrangeTelco customers.

The significant number of customers who churn from the organization during the first month provides essential information for possible churners as this churn is correlated with the onboarding stage, which was covered in previous chapters of this thesis and specifically in Section 2.2. During the onboarding stage, there two critical milestones for the company. The first is when the customer signs up for the contract, and the second is when the customer achieves their first success with a specific product/service. Therefore, 380 churners during the first month with the company denote unsuccessful onboarding. Consequently, OrangeTelco needs to take action to decrease unsuccessful onboarding, includ-

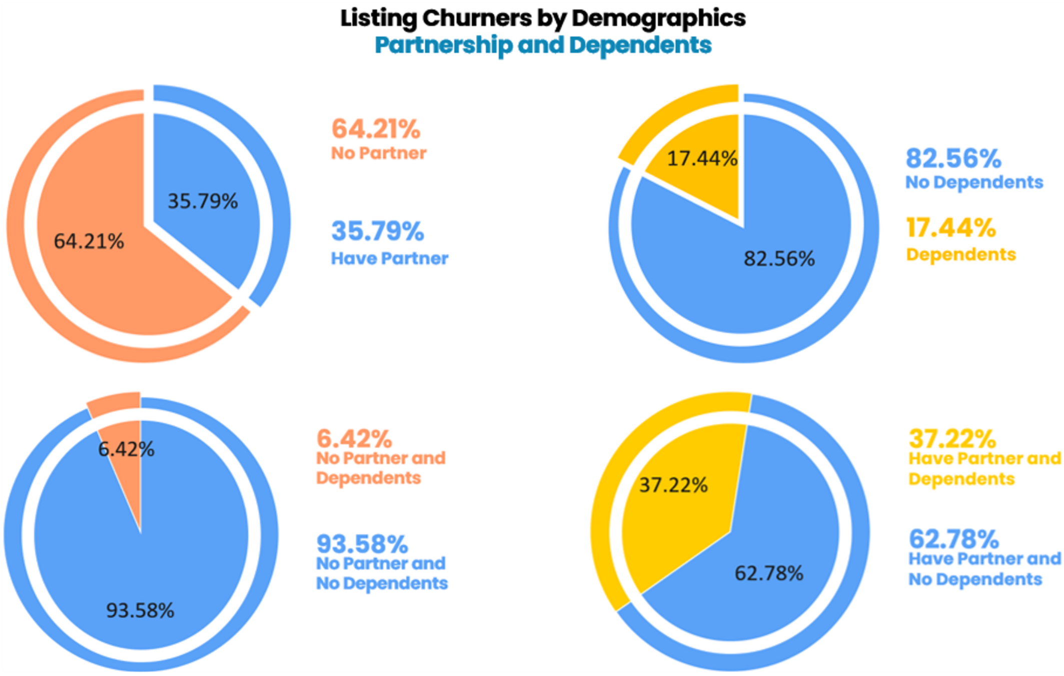


Figure 5.12: Listing Churners by Demographic Features - Partnership and Dependents

ing redesigning the product/service packages and contract duration on specific services.

The analysis of the data that is represented graphically in Figure 5.13 reveals that customers who pass the 24-month threshold of their contract are less vulnerable to churn compared to customers in the first two years of their contract. This finding should enhance actions related to marketing department offerings. Additionally, the loyalty-related data analysis illustrated in Figure 5.13 shows the important role of the proposed Insights Element in the CBFF as it suggests that future churn prediction processes should include the collection and analysis of loyalty-related data or the developing of a loyalty scheme if one does not exist at OrangeTelco.

Another feature that should be highlighted is the type of preferred contract of customers and churners, which is demonstrated in Figure 5.14. The vast majority of churners (88.55%) and just under half (42.91%) of the customers preferred the month-to-month contract. This observation is essential as it gives a meaningful insight into OrangeTelco customers, and it is possible that the majority of current customers who own a month-to-month contract churned at the end of their contract instead of renewing it.

Another finding that derives from the previous observation is that the current packages

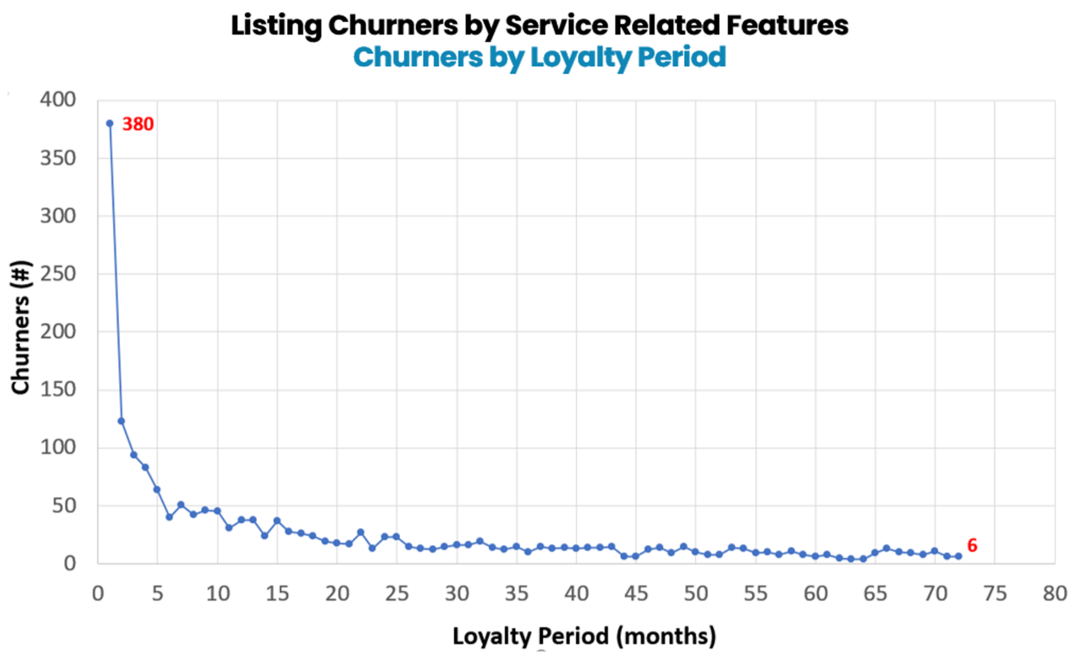


Figure 5.13: Listing Churners by Service-pelated Features - Loyalty Period and Churn

aligned with a longer contract duration are not as attractive, and these need to be identified and analyzed by OrangeTelco. This lesser commitment is often translated to churn, as shown in the churner data on the left chart in Figure 5.14. Considering this parameter, OrangeTelco could benefit from using the findings to redesign its service packages, and it should offer more benefits for extended duration contracts than it currently does. These actions could work towards decreasing the month-to-month contacts and the number of churners, accordingly.

The paperless billing feature was the next feature to be analyzed, and the preferences of churners and customers are graphically presented in Figure 5.15. The left chart demonstrates churners' preferences while the right demonstrates customer preferences. Looking at churners' preferences, there is significant interest in paperless billing (74.91%) compared to 25.09% for those who chose to receive paper bills. Likewise, just over half of customers (53.56%) prefer paperless billing. This data highlights the valuable role of the proposed conceptual framework as a decision-making tool. More specifically, following the steps of the proposed CBFF, a subscription-based organization could build a more detailed profile of the possible churner considering multiple features, KPIs, seasonal trends, and insights provided by previous churn prediction life cycles. For example, in a scenario where OrangeTelco uses this feature to recognize possible churners, it could lead to wrong decisions inside the organization. Such bad decisions can be a result of the overestimation

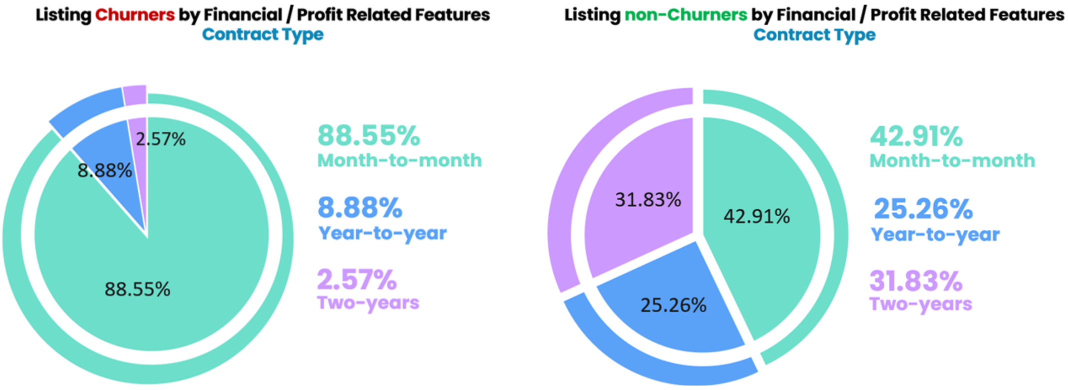


Figure 5.14: Listing Churners and non-churners by Financial / Profit-related Features - Contract Type

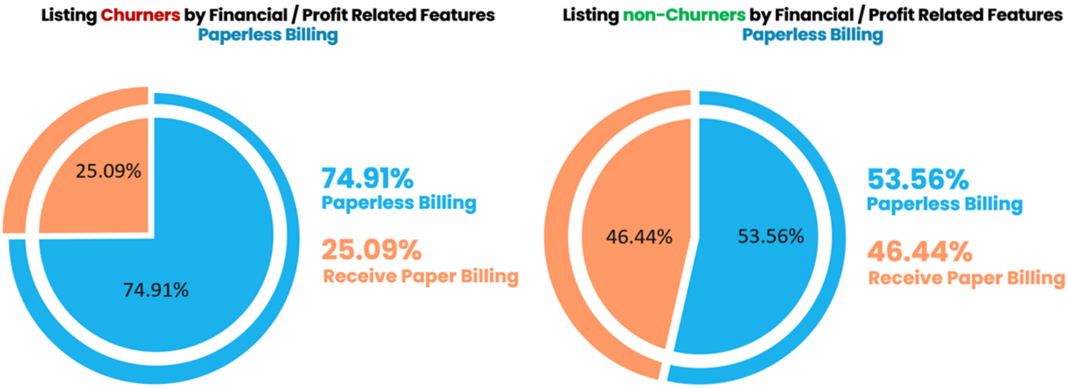


Figure 5.15: Listing Churners and non-churners by Financial / Profit-related Features - Paperless Billing

of this feature to the detriment of other more important features.

Following the actions proposed in the CBFF, OrangeTelco can avoid bad decisions because the proposed Business Case Analysis Element can denote what is essential for the organization by analyzing the data from the organization's perspective. The proposed Data Collection Element can include and transform data from the organization based on the input from the proposed Business Case Analysis Element. Then, the proposed Feature Selection Element can create the feature ideation list based on inputs from the previous elements and conclude the final features list consisting of the dataset, which will be forwarded to the proposed Prediction Model Element and, finally, extrapolate the outcomes using the proposed Insights Element.

Furthermore, after analyzing the data types, the proposed Prediction Model Element from the proposed conceptual framework was evaluated using the empirical data collected. In doing so, the OrangeTelco dataset was pre-processed before feeding the prediction models as this practice is followed for the classification of algorithms based on [105–107]. During the pre-processed phase, the OrangeTelco dataset was split into training and validation sets, which took care of the categorical variables included in the dataset (e.g., gender, internet service, etc.,) and dataset normalization.

Pre-processing the dataset was essential as it is related to the model's success, as noted in [106, 107]. Splitting the dataset into training and validation sets was used to evaluate the model before it was deployed. Next, the categorical features were converted into integers. Finally, the dataset was normalized, and values were translated to a common scale without distorting the difference between the range of values.

Some of the most known prediction models were used in the proposed Prediction Model Element to test their performance on OrangeTelco data. The role of the proposed Prediction Model Element, as described in Section 3.3.1.4, is to select the most suitable classification algorithm to perform churn prediction. In doing so, well-known classification algorithms are used to check which one performs better on the given dataset. More details and descriptions regarding the prediction models shown below can be found in Appendix 2, and the prediction models tested with their specific parameters are described in Appendix 7.

- **Logistic Regression - Baseline Model:**

$$\begin{aligned} &(C = 1.0, class\_weight = None, dual = False, fit\_intercept = True,\\ &intercept\_scaling = 1, l1\_ratio = None, max\_iter = 100,\\ &multi\_class =' ovr', n\_jobs = 1, penalty =' l2', random\_state = None,\\ &solver =' liblinear', tol = 0.0001, verbose = 0, warm\_start = False) \end{aligned} \tag{5.2}$$

- **Logistic Regression - Synthetic Minority Oversampling Technique (SMOTE):**

$$\begin{aligned} &(C = 1.0, class\_weight = None, dual = False, fit\_intercept = True,\\ &intercept\_scaling = 1, l1\_ratio = None, max\_iter = 100,\\ &multi\_class =' ovr', n\_jobs = 1, penalty =' l2', random\_state = None,\\ &solver =' liblinear', tol = 0.0001, verbose = 0, warm\_start = False) \end{aligned} \tag{5.3}$$

- **Logistic Regression - Recursive Feature Elimination (RFE):**

$$\begin{aligned}(C = 1.0, class\_weight = None, dual = False, fit\_intercept = True,\\ intercept\_scaling = 1, l1\_ratio = None, max\_iter = 100,\\ multi\_class =' ovr', n\_jobs = 1, penalty =' l2', random\_state = None,\\ solver =' liblinear', tol = 0.0001, verbose = 0, warm\_start = False)\end{aligned} \tag{5.4}$$

- **Decision Tree:**

$$\begin{aligned}(ccp_alpha = 0.0, class\_weight = None, criterion =' entropy',\\ max\_depth = 3, max\_features = None, max\_leaf\_nodes = None,\\ min\_impurity\_decrease = 0.0, min\_impurity_split = None,\\ min\_samples\_leaf = 1, min_samples\_split = 2,\\ min\_weight\_fraction\_leaf = 0.0, presort =' deprecated',\\ random\_state = None, splitter =' best')\end{aligned} \tag{5.5}$$

- **Random Forrest:**

$$\begin{aligned}(bootstrap = True, ccp\_alpha = 0.0, class\_weight = None,\\ criterion =' entropy', max\_depth = 3, max\_features =' auto',\\ max\_leaf\_nodes = None, max\_samples = None,\\ min\_impurity\_decrease = 0.0, min\_impurity_split = None,\\ min\_samples_leaf = 1, min\_samples\_split = 2,\\ min\_weight\_fraction\_leaf = 0.0, n_estimators = 100,\\ n\_jobs = None, oob\_score = False, random\_state = None,\\ verbose = 0, warm\_start = False)\end{aligned} \tag{5.6}$$

- **Naive Bayes:**

$$(priors = None, var\_smoothing = 1e - 09) \tag{5.7}$$

- **KNN:**

$$\begin{aligned}(algorithm =' auto', leaf\_size = 30, metric =' minkowski',\\ metric\_params = None, n\_jobs = 1, n\_neighbors = 5, p = 2,\\ weights =' uniform')\end{aligned} \tag{5.8}$$

- **SVM - Linear:**

$$\begin{aligned}(C = 1.0, break\_ties = False, cache\_size = 200, class\_weight = None,\\ coef0 = 0.0, decision\_function\_shape =' ovr', degree = 3, gamma = 1.0,\\ kernel =' linear', max\_iter = -1, probability = True,\\ random\_state = None, shrinking = True, tol = 0.001, verbose = False)\end{aligned} \tag{5.9}$$

- **SVM - Radial Basis Function (RBF):**

$$\begin{aligned}(C = 1.0, break\_ties = False, cache\_size = 200, class\_weight = None,\\ coef0 = 0.0, decision\_function\_shape =' ovr', degree = 3, gamma = 1.0,\\ kernel =' rbf', max\_iter = -1, probability = True, random\_state = None,\\ shrinking = True, tol = 0.001, verbose = False)\end{aligned} \tag{5.10}$$

- **XGBoost Classifier:**

$$\begin{aligned}(base\_score = 0.5, booster =' gbtree', colsample\_bylevel = 1,\\ colsample\_bynode = 1, colsample\_bytree = 1, gamma = 0,\\ learning\_rate = 0.9, max\_delta_step = 0, max\_depth = 7,\\ min\_child\_weight = 1, missing = None, n\_estimators = 100,\\ n\_jobs = 1, nthread = None, objective =' binary : logistic',\\ random\_state = 0, reg\_alpha = 0, reg\_lambda = 1,\\ scale\_pos_weight = 1, seed = None, silent = True,\\ subsample = 1, verbosity = 1)\end{aligned} \tag{5.11}$$

The prediction models outlined above tested the OrangeTelco dataset, and their performances were evaluated using the following metrics: i) Accuracy, ii) Area Under the Curve (AUC), iii) Precision, iv) Recall, and v) F-measure. Evaluation metrics were previously mentioned and described in Customer Churn Prediction Literature Review Selected Cases - Chapter 2. The data summarized in Table 5.7 shows the performance of each prediction algorithm on the OrangeTelco dataset. Regarding the accuracy metric, the logistic regression algorithm based on the baseline model achieved the highest score with a 0.80 probability of correct predictions of churners. The baseline model refers to the simplest form of logistic regression. Comparing the logistic regression algorithm using the Synthetic Minority Oversampling Technique (SMOTE), which involves sampling before feeding the data into the model, achieved a lower accuracy of 0.76. The same accuracy (0.76) was also achieved from the logistic regression algorithm where Recursive Feature Elimination (RFE) was applied before feeding the model. The KNN algorithm achieved the lowest accuracy score at 0.70.

The AUC is one of the most widely used metrics for evaluation and is used to rank a randomly chosen churner (positive example) to a randomly chosen non-churner (negative example). The highest AUC score was 0.77 and was achieved by three algorithms: i) Logistic regression (RFE), ii) Logistic regression (SMOTE), and iii) SVM-Linear. On the other hand, the

Model	Accuracy	AUC	Precision	Recall	F-Measure
Log. Regression (Baseline Model)	0.80	0.72	0.69	0.53	0.60
Random Forrest	0.79	0.69	0.67	0.48	0.56
XGBoost	0.78	0.70	0.61	0.54	0.57
Log. Regression (RFE)	0.76	0.77	0.55	0.79	0.65
Log. Regression (SMOTE)	0.76	0.77	0.55	0.80	0.65
SVM - RBF	0.76	0.64	0.61	0.36	0.46
SVM - Linear	0.75	0.77	0.53	0.80	0.64
Decision Tree	0.74	0.68	0.53	0.54	0.53
Naïve Bayes	0.73	0.76	0.51	0.82	0.63
KNN	0.70	0.71	0.48	0.73	0.58

Table 5.7: Prediction Algorithms Evaluation Metric Scores on OrangeTelco

lowest AUC score was achieved by SVM-RBF. Figure 5.16 compares the performances of the different models. While the logistic regression model in its simplest mode achieved the highest accuracy score with more correct predictions, other prediction models achieved similar scores.

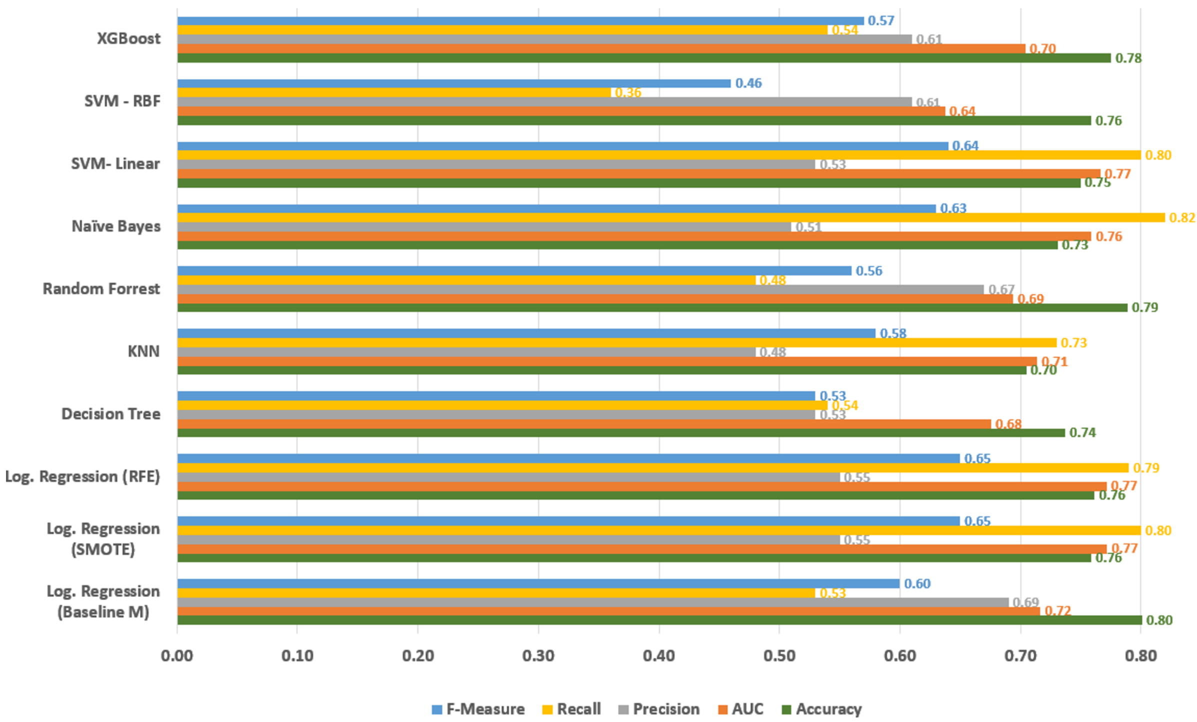


Figure 5.16: Prediction Models Performance on OrangeTelco Dataset

The comparison of the classification algorithms revealed that there is not an overall prediction model that suits all cases, but each case can identify the best fit in terms of the features, size, and categories of features (e.g., categorical, numeric) included in the dataset. The previous assumption was also evaluated during the empirical data analysis of BlueTelco, the first case study included in this research.

5.3.4 Summary of OrangeTelco Case Study

Empirical data collected from OrangeTelco were used to test, the proposed conceptual framework. The results demonstrated that the proposed CBFF can be applied in the case of OrangeTelco to enhance the actions taken to prevent customer churn. The adoption of the proposed CBFF by OrangeTelco shows that critical steps during the customer's journey through the organization are interpreted and lead to more accurate customer retention campaigns and strategies. Table 5.8, summarizes the main findings of this case. These are analyzed in the next paragraph.

The proposed Business Case Analysis Element was evaluated by defining important KPIs for the organization, which were the proportion of unsuccessful onboarding of customers during the first month of the contract. The action taken by OrangeTelco to decrease the amount of unsuccessful onboarding led to less churn from the organization, as identified from the empirical data collected. Actions that led to successful onboarding were the redesign of packages currently offered by the organization and the assignment of more benefits to extended duration contracts. The interpretation of insights from previous customer churn prediction processes provided essential input to the proposed Business Case Analysis Element, such as: i) designing new goals for the company to achieve in a specific timeline (e.g., semester, annual, etc.), ii) redesigning processes during the registration of customers (e.g., collecting data relevant to residential status), or iii) re-targeting potential customers using the knowledge gained from their churners behavior. Clearly, this finding suggests a revision to the proposed CBFF and Business Case Analysis Element should receive input from Insights Element.

The proposed Data Collection Element was evaluated with data derived from OrangeTelco empirical analysis. This element played an important role in successfully identifying the churn trend inside OrangeTelco. The study of the demographic data identified a trend between churners, showing that most churners did not have dependents or partners; this was strengthened by data related to residential status. Additionally, the combination of different data types enhanced churner profiling. In addition during the evaluation of the

Proposed Element	Findings in OrangeTelco
Business Case Analysis	1. Define important KPIs, such as the accepted proportion of accepted unsuccessful on-boarding during the first month of the contract. 2. Update or include new goals for the company to achieve in a specific timeline. 3. Re-target the company's potential customers.
Data Collection	1. Collection of data relevant to residential status, as this type of data show a trend in churners. 2. Different oversampling techniques on the dataset do not always give better results. 3. Different data type combination, enhance the profiling of the churner.
Feature Selection	1. Combination of features enhance the profiling of the churner.
Prediction Model	1. Well-known prediction algorithms tested on OrangeTelco. 2. Logistic Regression (Baseline Model) achieved the highest score based on the Accuracy metric (0.80).
Insights	1. The value of churners' interpretation during the first month of their contract should improve the on-boarding process. 2.Re-define company's market.

Table 5.8: Findings from OrangeTelco that address elements of proposed CBFF

proposed Data Collection Element in the OrangeTelco case, the different data types were collected, combined, and analyzed, as summarized in Table 5.6. The findings derived from this analysis showed that the combination of different data types enhanced actions related to churners' identification and led to more accurate churner profiling. The balancing processes included in the variables setup task of the proposed Data Collection Element showed that different oversampling techniques do not always increase the prediction model's accuracy. More specifically, the logistic regression model was tested in three different versions: i) the basic version (baseline model), ii) using SMOTE, and iii) by applying RFE to features before feeding the dataset into the model. The results concluded that the baseline model achieved the highest accuracy score of 0.80, while the other two modifications of SMOTE and RFE achieved a higher score in the F-measure metric (0.65), according to 0.60 reached in the baseline model. The F-measure metric provided a single score that balanced both precision and recall metrics.

The proposed Feature Selection Element was evaluated by applying SMOTE and RFE methods on the dataset before feeding the data into the prediction model. Before running the logistic regression model, the SMOTE oversampling technique produced the second-highest score of 0.80 in the recall metric. The recall metric shows the proportion of true positive predictions made out of all positive predictions that could have been made on the dataset. Similarly, the RFE method was applied to the features and achieved 0.79 in the recall metric. Both methods achieved the same score in the accuracy metric (0.76), which is the fourth-highest among the tested models. The outcome of the evaluation showed that the proposed Feature Selection Element is important and should give additional accuracy to predictions given that all of the required features are included in the dataset. Moreover, the empirical study revealed that tasks and processes in the proposed Feature Selection Element needed to be described more substantially. More details regarding the new guidelines and modifications of the proposed Feature Selection Element can be found in Section 6.3. Similar to the BlueTelco case, there is a correlation between elements 1-3 and Data Input phase and its sub-phases of the churn prediction procedure of Figure 3.19.

The proposed Prediction Model Element was tested using OrangeTelco data by using 10 machine-learning algorithms and modifications to predict churners. The results of the predictions are listed in Table 5.7 and Figure 5.16, where the different evaluation metrics are compared to give the reader an additional overview of their performance. The results analysis suggests that no one prediction algorithm performs better for every case, but each case can find the best fit depending on the data included in the prediction dataset. For the OrangeTelco case, logistic regression in its simplest form (e.g., baseline model) achieved the highest accuracy in the prediction of churners. In addition, the author observes a correlation between element 4 of Churn Prediction phase and its sub-phases of the churn prediction procedure of Figure 3.19.

The last element in the proposed CBFF, the Insights Element, was evaluated using OrangeTelco data and was identified as having a crucial role in the success of churner prediction. The empirical data interpretation showed that OrangeTelco needs to redefine its market by including additional features during the customers' registration, such as residential status. Insights also noted a spike in churns during the first month of the contract between the company and the customer; this was assigned to unsuccessful onboarding changes in the package offering.

The findings derived from the evaluation of each element of the proposed conceptual framework summarized in Table 5.8. Overall, the evaluation showed that the proposed conceptual framework could be applied in OrangeTelco to support and enhance customer churn prediction. Despite the successful evaluation of the proposed conceptual framework at OrangeTelco, the empirical study revealed additional points that need to be investigated further, and these are covered in the conclusion in Section 5.4.

5.4 Conclusions

This chapter has presented the empirical data collected from the two case studies selected for this thesis. It has also described the evaluation of the proposed conceptual framework and its elements. The findings derived from the two case studies have been presented and analyzed and suggest that: i) conceptualizations (i.e., churn prediction procedure, CBFF) can be used to improve decision making regarding customer churn prediction and ii) the proposed framework requires some revisions as these were reveled during the case studies investigation. In particular, the empirical study denoted the inclusion of additional guidelines and tasks in the five elements, which are summarized below.

- The proposed Business Case Analysis Element was successfully evaluated in both case studies and was shown to have a valuable role in the accurate churn prediction process. The empirical data gathered from the two case studies revealed that the inclusion of a Reports task would be helpful for domain stakeholders to evaluate staff skills, potential markets, and onboarding scores. The inclusion of loyalty scheme insights under the Insights task was also suggested.

- Regarding the proposed Data Collection Element, the data gathered from BlueTelco revealed that the organization was using different terms to refer to the same thing between its systems. This issue was tracked by the author who suggested the addition of a Terminology Checking process to address the issue.

- The proposed Feature Selection Element was evaluated and its essential role in the development of a suitable features list was highlighted. However, it was noted that tasks and processes included in this element should be presented more substantially, helping stakeholders to understand clearly what steps to follow to conclude the features list. The addition of a Feature Engineering task was proposed to host all subtasks and processes related to feature proposals, development, and selection. Furthermore, new direct input from the Business Case Analysis Element into the Feature Selection Element is suggested in the case of BlueTelco.

- A similar direct input from Insights Element to Feature Selection Element is also suggested from the findings of both cases (BlueTelco, OrangeTelco). In addition, the findings of the OrangeTelco suggested that output from the Insights Element goes to Business Case Analysis Element.

- The proposed Prediction Model evaluation uncovered an issue regarding the selection of the appropriate classification algorithm between the available options. In an attempt to address this issue, new guidelines based on research findings were proposed for the use of a meta-learning toolkit [105] to support the classification algorithm suggestion task.

- The proposed Insights Element was valuable in interpreting churn prediction outcomes and should be used to make decisions inside the organization about future churn predictions. The addition of a task that focuses on the interpretation of insights should enhance the usage of CBFF as a decision-making tool.

The outcome of the empirical investigation showed that CBFF and churn prediction procedure are suitable for the two investigated subscription-based organizations in assisting their efforts to predict churn, enhance churn prediction accuracy, and reduce churn. The next chapter considers these research findings and the concerns derived from the empirical investigation to provide a revised version of the proposed conceptual framework. Clearly, the empirical data from the two organizations suggest that: i) the CBFF can bu used for the customer churn prediction in these subscription-based organizations, and ii) the customer churn prediction procedure is also validated and can be used to enhanced customer churn prediction related actions.

Chapter 6

Revision of the Customer Behavior Forecasting Framework

Having knowledge but lacking the power to express it clearly is no better than never having any ideas at all

Pericles (495 BC - 429 BC)

Summary

The results derived from the empirical data analysis presented in Chapter 5, demonstrated that CBFF and churn prediction procedure are useful and can assist both case organizations to better predict customer churn and improve decision making process. Moreover, there are findings and suggestions from the BlueTelco and OrangeTelco that highlight minor changes and revisions to the CBFF. The proposed revisions improve CBFF and enhance churn prediction actions. These revisions of CBFF results in a more accurate framework that meets the aim of this research work, which is to develop a framework that could be used to enhance customer forecasting techniques and predict customer churn for subscription-based organizations.

6.1 Introduction

There is an increasing need to enhance tactics to more accurately predict churn for companies in the subscription-based domain. The Literature Review conducted in this thesis (Chapter 2), showed that there is a lack of efficient churn prediction procedure [1] in terms of subscription-based organizations. During the literature review, various methods and techniques were identified that aimed to solve churn prediction. However it was revealed that there is a lack of a standardized procedure that assists churn prediction. In addition, there is a lack of customer churn prediction framework that can be used as a decision making tool and that will help organizations making more informed decisions and reducing churn.

Therefore, this thesis attempted to identify issues and challenges around customer churn by providing a comprehensive solution that aimed to identify the stages to be included in a customer churn prediction procedure (Figure 3.18). Based on the proposed churn prediction procedure, a conceptual framework was proposed, namely, the Customer Behavior Forecasting Framework (CBFF) (Figure 3.26), that aims to support the churn prediction efforts undertaken by subscription-based organizations. In Chapter 4, reviewed were research methodologies justify to the selection of an appropriate research methodology for this thesis. As a result an interpretivism research stance and a qualitative multiple case study research strategy was adopted to test the conceptual framework.

In Chapter 5, the proposed conceptual framework was tested through the collection and analysis of empirical data from BlueTelco and OrangeTelco organizations. This evaluation revealed important findings regarding the value of the proposed CBFF and showed that it could be applied to the two cases of subscription-based organizations to support and enhance customer churn prediction.

However, the findings derived from the empirical data analysis revealed additional issues that need further investigation. Furthermore, empirical findings, suggests new guidelines and modifications to the elements of the CBFF. Chapter 6 begins with the lessons learned from the case studies, as reflected in Section 6.2, and it then amends the proposed conceptual framework, as covered in Section 6.3. The chapter ends with the conclusions drawn in Section 6.4.

6.2 Lessons Learned from Case Studies

During Chapter 5, the customer churn prediction in subscription-based organizations was investigated through two case studies selected for this research work. More specifically, empirical data from BlueTelco and OrangeTelco were collected and test the elements of the Customer Behavior Forecasting Framework. In doing so, lessons learned are derived, and aim to build a better understanding around customer churn prediction for the two subscription-based companies. Help researchers and organizations. The lessons learned from the two case organizations are:

- **Lesson 1:** In subscription-based companies, churn is vital for prosperity, and there is a need for standardized procedures that identify important stages to predict churn rate and retain customers.
- **Lesson 2:** There is not a commonly accepted churn prediction procedure, based on which a churn prediction framework will be based to enhance decision making process for those organizations.
- **Lesson 3:** The proposed conceptual framework provided important value to the two subscription-based organizations tested in this thesis as it assists their efforts to understand their customers' behavior and decrease churn.
- **Lesson 4:** The proposed conceptual framework captures the important elements in the customer churn prediction procedure.
- **Lesson 5:** The combination of different data types that made up each case study's dataset revealed essential details about the profiling of the churner.
- **Lesson 6:** Customers' profiles differ in characteristics, despite t they became both from telecommunication sector. This lesson denotes the valuable impact of Business Case Analysis of actions supporting the implementation of Business Analysis Element in CBFF.
- **Lesson 7:** In many cases, subscription-based organizations do not collect valuable data that is useful for churner identification.
- **Lesson 8:** Both churn prediction procedure and CBFF were verified by the two case organizations and can support the decision making process.

6.3 Revised Customer Behavior Forecasting Framework

The empirical data investigation allowed, the author to better understand the functionality and interactions between elements of the conceptual framework in real case scenarios. The outcome of those actions resulted in useful lessons being learned; these are listed in Section 6.2. Additionally, during the evaluation of the proposed CBFF, minor revisions were proposed.

This section presents and analyses the proposed revisions for the CBFF along with the main findings from the two case organizations as a result of the empirical data investigation. The main findings are summarized in Table 6.1 and are organized into two columns representing the two case studies and five rows representing the elements of the framework. The results from the case studies suggested the addition of a loyalty scheme into the insights category as empirical data revealed that loyalty-related data, such as acquired services by customers and the contract period are important in the profiling of churners.

Furthermore, a new task under the label Reports was included in the Business Case Analysis Element. This task generated reports on staff skills, the engagement of new technologies by the company, the analysis of the potential market, and on-boarding scores. Both the Report task and loyalty scheme insights were included in the revised version of the Business Case Analysis Element as illustrated in Figure 6.1 and are based on the main findings summarized in Table 6.1.

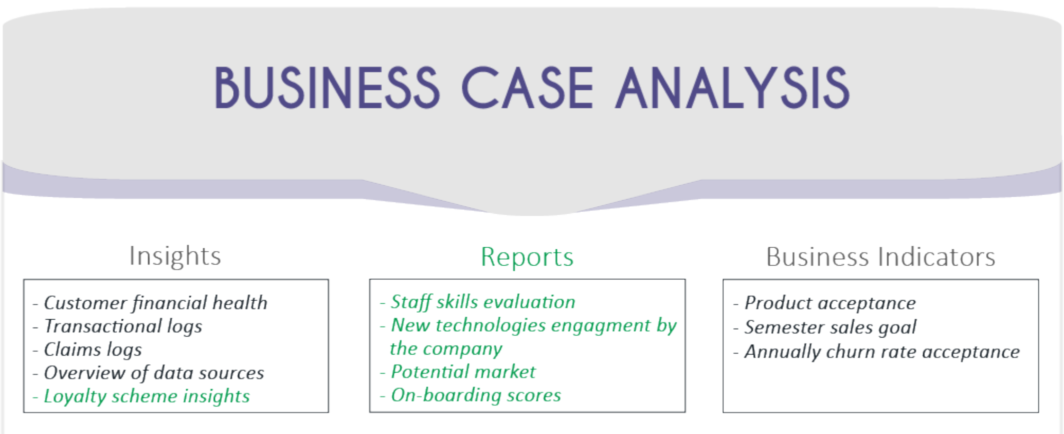


Figure 6.1: Revised Business Case Analysis Element

Element of the Conceptual Framework	Main Findings From Case Studies	
	BlueTelco	OrangeTelco
Business Case Analysis	1. Collection of loyalty insights enhance churners profile development. 2. Analysis of staff skills evaluation are important in succeeding desired business goals. 3. Tracking of new technologies engagement. Fast-growing and changing telco environment.	1. Capture of on-boarding score works towards churn rate decrease. 2. Update or include new goals for the company to achieve in a specific timeline.
Data Collection	1. Usage of the same terminology throughout the systems, to avoid false logging of churns.	1. Combination of different data types, enhance the profiling of the churner.
Feature Selection	1. Usage of feature engineering techniques lead to the selection of the appropriate features. 2. Profiling of customers enhance churn prediction accuracy. 3. Considering loyalty insights in the selection of features.	1. A flow that provides input to Feature Selection Element from Insights Element, provides a more dynamic and robust usage of the CBFF
Prediction Model	1. Usage of the Multiplicative Seasonality Model to identify the churning periods of interest. 2. Deployment of SAS BI to identify churning customers.	1. Comparison of multiple classification techniques to choose an appropriate for churn prediction dataset.
Insights	1. The value of interpretation is crucial in success of future churn prediction attempts. 2. Development of a loyalty scheme to collect customer health related data. 3. Evaluation process for staff, enables the collection of appropriate data for churn prediction.	1. A flow from Insights Element to Business Case Analysis Element enhances decision making in targeting the potential market for the organization. 2. Provide insights for the update or deployment of new processes that enhance retention.

Table 6.1: Main Findings from Case Studies Regarding CBFF Elements

The Data Collection Element includes a minor revision in the variable setup process as demonstrated in Figure 6.2. The new guideline is because the same terminology was used between all the systems in an organization, and this new task avoids falsely logged churning events. In the case of BlueTelco, there was a mismatch between the subscription and account terminologies used by the systems inside the organization, and as a result, churning events were logged incorrectly. For this reason, a terminology check was added to the variables setup task, where terminology in the system is checked and changes made where necessary.

The third row in Table 6.1 summarizes the main findings regarding the Feature Selection Element. This element addressed issues regarding actions related to the suitable selection of features during a churn prediction life cycle. The findings revealed that actions hosted by this element should be described more substantially to enable stakeholders, the reader,

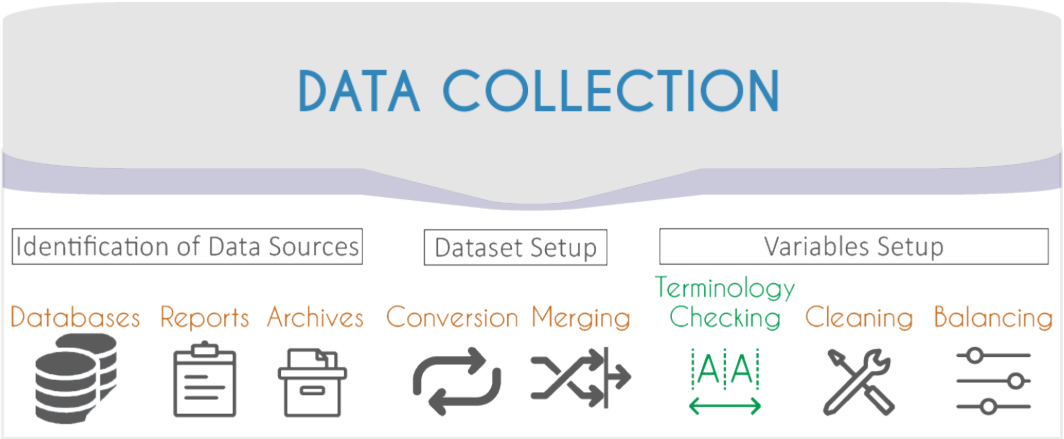


Figure 6.2: Revised Data Collection Element

and future researchers to clearly understand the actions followed to establish the appropriate features for the churn prediction dataset. Subsequently, the feature engineering task was added to the variables preparation task.

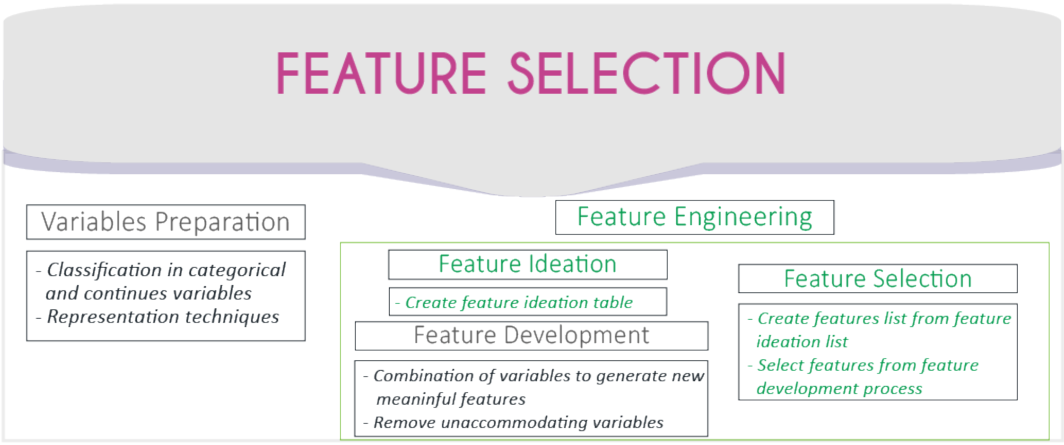


Figure 6.3: Revised Feature Selection Element

The feature engineering task hosts all the actions needed to create the features list, which includes all suitable features used during the churn prediction process. The feature development task was demoted under the feature engineering task, and the feature ideation and feature selection tasks were added. Figure 6.3 illustrates the revised version of the Feature Selection Element. The features ideation task creates the features list based on the input taken from both the Data Collection and Business Case Analysis Elements as shown in Figure 6.6. The feature selection task creates the features list based on the feature ideation and feature development tasks.

Although the Prediction Model Element was found to be both useful and valuable in the success of a churn prediction process, results showed that there was not a comprehensive prediction model that scored the highest churn accuracy for every case. Therefore, a new guideline was added in the classification algorithm suggestion task, that used meta-learning techniques during the selection of the appropriate classification algorithm. In brief, meta-learning [105] is a sub-field of machine learning where systematical observation of different machine learning approaches achieves faster learning based on prior experience. This addition was suggested in the revised version of the Prediction Model Element to enhance the selection of the most suitable classification algorithm by the organization, as shown in Figure 6.4.

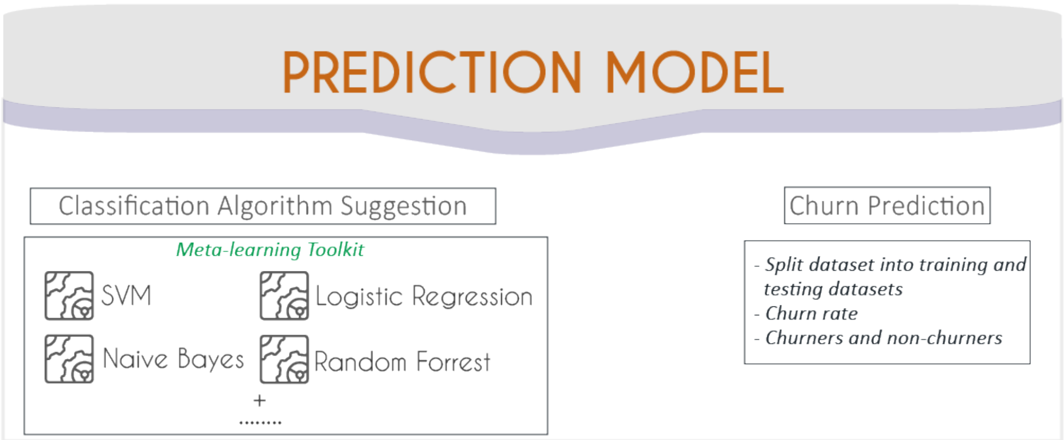


Figure 6.4: Revised Prediction Model Element

The last element in the proposed conceptual framework is the Insights Element, and its findings are summarized in Table 6.1. Its role in interpreting churn prediction was highlighted, but a revision was suggested that included an interpretation task, as presented in Figure 6.5. Examples that denote the importance of churn prediction interpretation are the outcomes of future churn prediction attempts in profiling customers. It should also help in the deployment of new processes, such as the design of a customer loyalty scheme, which should work towards the decrease of churn, as revealed from empirical data analysis. Likewise, the development of the staff evaluation process could result from insights interpretation.

The revised version of the conceptual framework includes new guidelines and tasks that were derived from the analysis of the empirical data collected from the two case studies, and it is summarized in Figure 6.6 with the new flows illustrated in green. The first one dis-

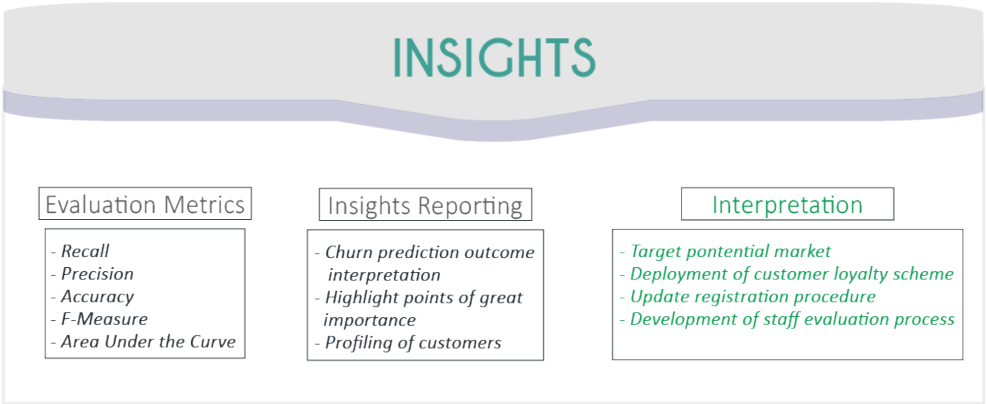


Figure 6.5: Revised Insights Element

plays the output from the Business Case Analysis Element, which includes loyalty scheme indicators that are imported to the Feature Selection Element. The second flow shows the outcome from the Insights Element, which transfers the interpretation task's insights regarding new features that should be included in the features list. The third flow provides Business Case Analysis Element with the outcome of churn prediction, as this derives from Insights Element.

The three new flows added to the revised CBFF denote the more dynamic and robust usage of the conceptual framework in its attempt to enhance churn prediction actions. More specifically, the Business Case Analysis Element has two outgoing flows, one for the Data Collection Element and the other for the Feature Selection Element. Furthermore, the Insights Element can retrofit feature-related insights into the Feature Selection Element. It also provides input to the Business Case Analysis with the outcome of the previous churn prediction procedure.

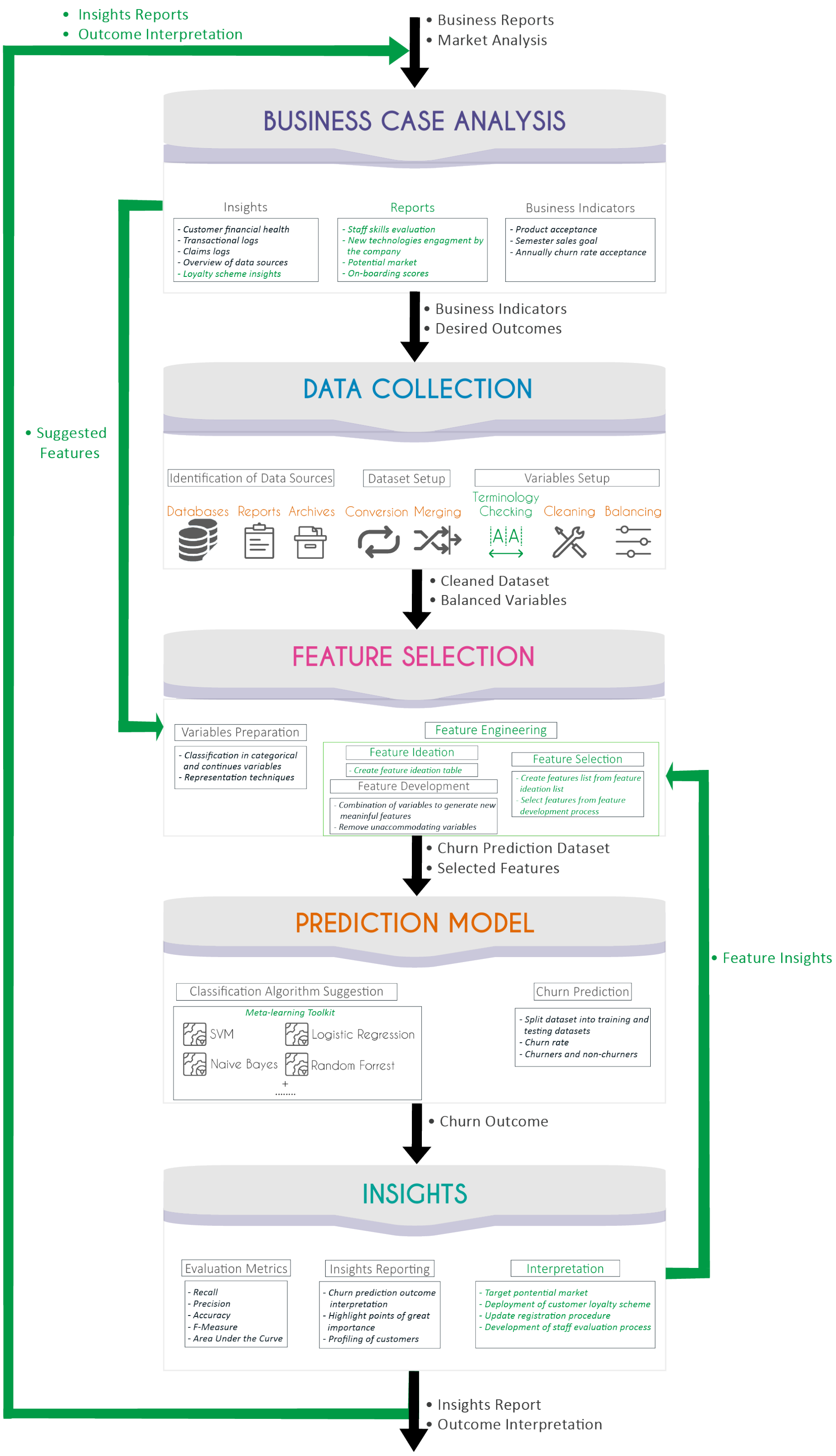


Figure 6.6: Revised Customer Behavior Forecasting Framework

6.4 Conclusions

This thesis investigates customer churn prediction for subscription-based organizations that aims to support related actions that enhance accuracy in predictions and decrease churn rate. In doing so, the author proposed the Customer Behavior Forecasting Framework (CBFF), which works toward that goal. The proposed conceptual framework was developed according to findings and observations derived from the literature review, and it was tested and evaluated through two case organizations as reported in Chapter 5. Chapter 5 verified CBFF and churn prediction procedure revealed new findings that provide enough justification for the revision of the conceptual framework.

This chapter has presented the revised version of the framework, starting with the lessons learned from the research, which are listed in Section 6.2. Compared to the proposed conceptual framework the revised one includes the following additional components:

- In the Business Analysis Element the component Reports is added for the following reasons that revealed from empirical data investigation:
  - Collection of loyalty insights enhances churners profile development.
  - Analysis of staff skills evaluation are important in succeeding desired business goals.
  - Analysis of in formations regarding on-boarding scores could work towards churn rate.
  - Analysis of insights regarding new technologies engagement by the subscription-based organization support growth and enhances tools they have against churn.
- In Data Collection Element a minor revision of the Variables Setup component is made and includes the addition of the Terminology Checking guideline to resolve the following issues identified:
  - The mismatch between the systems of the organizations when referring to an ontology with different names increases falsely logged churning events. The implementation of a terminology check in the Variables Setup component, evaluates the usage of the same terminology between all the systems in the organization.
- In Feature Selection Element the empirical data investigation lead to the addition of the Feature Engineering Component, which hosts the two new Feature Ideation

Components and Feature Selection Components, with the previous said Feature Development Component. The reasons that lead to this modifications in Feature Selection Element are:

  - Actions taken during this phase should be described in a mores substantial way to enable stakeholders, the reader and future researchers to clearly understand what actions followed to establish appropriate features for the churn prediction dataset.
  - Profiling of the customers enhances churn prediction accuracy. To achieve that a sustainable feature engineering procedure could be followed.

- In Prediction Model Element a guideline was added in the Classification Algorithm Suggestion Component for the following reason:

  - The usage of meta-learning techniques from the sub-field of machine learning domain, during the selection of the appropriate classification algorithm could achieve faster learning based on prior experience.

- In the Insights Elements the Interpretation Component is added to support the Insights Element in the following cases:

  - Interpretation of information regarding the possible development of a loyalty scheme from the organizations, could derive from this component.
  - Evaluation process for staff, which will enable the collection of appropriate data for churn prediction could derive from this component.
  - Potential market for the subscription-based organization could be defined from the usage of this component.

In addition to these there are three new flows that were added. One new outgoing flow from Business Case Analysis Element to Feature Selection Element, which provides the last element with suggested features derived from the analysis of previous insights. The second new flow is from Insights Element to Feature Selection Element, which provides the Feature Selection Element, with additional insights that enhance success of Feature Engineering component to select the appropriate features for churn prediction. The third flow is from Insights Element to Business Case Analysis Element, which provides the new attempt of churn prediction with the outcome and insights derived from previous churn prediction attempt. These new flows improve the usability of the framework and assist subscription-based organizations to take more accurate decisions regarding churn prediction related activities.

Chapter 7

Conclusions and Future Work

It does not matter how slowly you go
as long as you do not stop

Confucius (551 BC - 479 BC)

Summary

Chapter 7 summarizes the research findings along with the novel contributions and limitations of this thesis. A critical evaluation of the research work is also presented in this chapter. The research conducted presented in this research work meets the aim and objectives of this thesis (i.e., "Study customer churn in subscription-based organizations.") and results in five novel theoretical or practical contributions. These novel contributions support organizations to better understand, predict and minimize customer churn which is highly associated with their growth and prosperity. Moreover, further research work is identified in this complex and demanding area of customer churn prediction in subscription-based organizations.

7.1 Research Overview

This thesis investigated customer churn prediction in subscription-based organizations. The thesis began with an overview of the research problem as reflected in Section 1.1. The research aims and objectives were outlined in Section 1.2, the thesis structure was described in Section 1.4 and the thesis outline in Figure 1.2.

In Chapter 2, a critical analysis of the literature on customer churn prediction was presented. The causes of churn and benefits of churn prediction were outlined in Section 2.2 and Section 2.3, respectively. In Section 2.4, established systematic literature review methodologies were investigated and a plan was chosen for this thesis based on Cooper et al., [25] and Brereton et al., [24]. As a result, a systematic literature review on customer churn prediction was concluded with the literature review findings and the critical analysis reported in Section 2.6. Chapter 2, makes five important observations that form a valid and timely research gap, as this was depicted in Figure 2.12. The research gap highlights: i) the lack of a commonly accepted churn prediction procedure and ii) the absence of a customer churn prediction framework that can be used as a decision making tool. Open issues, regarding what aspects of customer churn could be included in such a framework, as well as how the different aspects could be orchestrated together as a decision making tool for the organizations, identified for further research.

The outcomes of Chapter 2 as those summarized in Figure 2.12, mapped with investigation actions in Chapter 3, which attempted to address them. Figure 3.1, demonstrated this mapping. Grounded on Figure 2.12, the author developed and proposed a churn prediction procedure, which was then used to build the conceptual framework seeks to be used as a customer churn prediction. The proposed conceptual framework as a decision-making tool to enhance customer churn prediction in subscription-based organizations. The proposed conceptual framework was referred to as Customer Behavior Forecasting Framework (CBFF) and consisted of five basic elements: i) Business Case Analysis, ii) Data Collection, iii) Feature Selection, iv) Prediction Model, and v) Insights. The breakdown, of CBFF, the analysis and justification of its elements is described in Section 3.3.The proposed CBFF, attempts to address customer churn open issues, as this illustrated in Figure 3.26.

In Chapter 4, available research methodologies were investigated and the selection of an appropriate research methodology for this thesis was justified. The interpretivism stance and the qualitative multiple case study research strategy were considered as suitable to test the proposed conceptual framework. The empirical research methodology followed during this research is reflected in Section 4.4, with Figure 4.2 demonstrating the empirical research methodology.

The proposed conceptual framework was evaluated in Chapter 5. The two case organizations, with coded names BlueTelco and OrangeTelco, were presented and analyzed in Section 5.2 and Section 5.3, respectively. The preliminary research findings derived from the collected empirical data were described and demonstrated along with the issues under investigation. The outcome of the empirical data investigation in Chapter 5, tested and evaluated the elements of the CBFF and concluded that CBFF can be used to improve decision making regarding customer churn prediction. In addition, the testing of CBFF elements lead to additions in the proposed framework with the most significant to include: i) a new component - "Reports" in Business Case Analysis Element, ii) modification of "Variables Setup" component in Data Collection Element, iii) new components - "Feature Engineering", "Feature Ideation", and "Feature Selection" added on Feature selection Element, iv) a guideline for using meta-learning techniques during "Classification Algorithm Suggestion", component in Prediction Model Element and v) a new component - "Interpretation" in Insights Element, and three new flows of information that result in a more robust and dynamic usage of the CBFF.

Chapter 6 focused on the interpretation of the empirical data from which: lessons learned were drawn from the two case studies, also there was a realization that revisions would have to be made to the proposed conceptual framework, to reflect empirical findings. The empirical findings revealed a few additional guidelines, modifications, and proposals regarding the elements of CBFF as described in Section 6.3, and the revised version of the conceptual framework was illustrated in Figure 6.6.

The proposed conceptual framework contributed to the body of knowledge as it: i) clarifies the stages followed during a churn prediction procedure for subscription-based organizations, ii) introduces guidelines and specified additions beyond the normative literature review, iii) recognizes the basic aspects of the churn prediction procedure by introducing five elements, iv) highlights the value of each proposed element in the churn prediction procedure, and vi) combines the five elements to propose the Customer Behavior Forecasting Framework (CBFF).

The additions / revisions that were made on the proposed framework are summarized in Table 7.1. In conclusion, the empirical data evaluated the proposed framework and indicated that it could be used to address churn prediction issues in these two case subscription-based organizations and as a decision-making tool for stakeholders to enhance churn prediction accuracy and reduce the churn rate in their organizations. The main findings of this research work are presented in the next section.

Proposed Element	Additions/Revisions
Business Case Analysis	**Guidelines** 1. Inclusion of loyalty scheme insights **Tasks & Processes** 1. Adding Reports task, which analyzes data from staff evaluation reports, new technology engagement by the company, potential market reaching and on-boarding scores achieved.
Data Collection	**Tasks & Processes** 1. Adding Terminology Checking process, which checks the usage of the same terminology throughout the organization.
Feature Selection	**Tasks & Processes** 1. Deploys Feature Engineering Task, which hosts Feature Ideation, Feature Development and Feature Selection Tasks. 2. Adding Feature Ideation task, which creates the feature ideation table based on the input taken from both the Data Collection Element and Business Case Analysis Element. 3. Adding a new flow from Insights Element to Feature Selection Element, which provides features insights. 4. Adding a new flow from Business Case Analysis Element to Feature Selection Element, which provides suggestions for features.
Prediction Model	**Guidelines** 1. Proposes the usage of a Meta-learning Toolkit, which supports the classification algorithm suggestion task.
Insights	**Tasks & Processes** 1. Adding Interpretation Task, which interprets churn prediction outcome to take various decisions for the organizations such as: i) targeting of potential market, ii) deployment of customer loyalty scheme, iii) update of registration procedures and iii) development of staff evaluation process. 2. Adding a new flow from Insights Element to Business Case Analysis, which provides, the churn prediction outcome interpretation for a new churn prediction attempt.

Table 7.1: Revisions Made on CBFF

7.2 Main Findings

This research, aimed to investigate customer churn in subscription-based organizations. In doing so, it proposed a conceptual framework that could be used as a decision-making tool to enhance churn prediction actions for subscription-based companies. Various findings derived from the investigation of the two case study organizations research and are listed below:

- **Finding 1:** The systematic literature review led to some important observations as demonstrated in Figure 2.12. These include: i) the importance of customer churn in the growth and prosperity of the subscription-based organizations, ii) the need for a customer churn prediction framework that will enhance decision making process, iii) the addressing of three main aspects of churn prediction from the normative literature iv) the absence of a research work, that integrates all three aspects of customer churn prediction and v) the absence of a commonly accepted customer churn prediction procedure, which can be used for the development of a customer churn prediction framework.

- **Finding 2:** Observations B and E (Figure 2.12), indicate that there is a research gap, and thus there is a need for: i) the development of a commonly accepted churn prediction procedure and ii) a customer churn prediction framework. Both of them, will assist in reducing customer churn and thus improve decision making and subscription-based organizations results.

- **Finding 3:** The investigation of the research gap identified in Chapter 2, initially led to the development and proposition of a churn prediction procedure (Figure 3.18).

- **Finding 4:** Based on the proposed churn prediction procedure the conceptual framework for customer churn prediction was build as reported in Section 3.3 and summarized in Figure 3.26.

- **Finding 5:** Two telecommunication providers, with the coded names BlueTelco and OrangeTelco, were investigated as part of the multiple case study research strategy adopted in this thesis. Both organizations validated the proposed churn prediction procedure and CBFF. In addition, the empirical findings recommend revisions to CBFF as these are summarized in Section 6.3 and depicted in Figure 6.6.

- **Finding 6:** The revised version of the conceptual framework shown in Figure 6.6, consisted of five basic elements: i) Business Case Analysis, ii) Data Collection, iii) Feature Selection, iv) Prediction Model, and v) Insights.

- **Finding 7:** The proposed framework was suitable for the two case organizations investigated in this thesis.

- **Finding 8:** The CBFF can be used as a decision-making tool by the two case subscription-based organizations to support actions that decrease churn rates, such as marketing strategy design, retention campaigns, the registration process update, and loyalty scheme development.

7.3 Meeting the Objectives of this Dissertation

This research focused on customer churn and developed a conceptual framework that could be used to enhance customer behavior forecasting and support churn prediction in subscription-based organizations. To achieve the aim of this thesis, several objectives were defined and presented in Section 1.2.2. The objectives were accomplished as follows:

- **Objective 1:** *To conduct a literature review (LR) in the area of customer behavior forecasting with a focus on customer churn prediction, and critically evaluate literature that is relevant to customer churn in the subscription-based domain.*

In the literature review, the research problem of customer churn, its impact on subscription-based organizations growth and prosperity and challenges in the area of customer churn prediction were investigated and examined, as presented and discussed in Chapters 1 and 2. A systematic literature review was adopted to offer a more accurate and complete literature review. The results of this effort are summarized in Figure 2.12, which presents the main observations, the research gap and the open issues for further investigation.

- **Objective 2:** *To develop a conceptual framework to support customer churn forecasting.*

The outcome of the literature review identified a research gap in the absence of a commonly accepted churn prediction procedure and a customer churn prediction framework. Grounded on the findings of literature review, a churn prediction procedure was developed through the utilization and synthesis of literature review observations. The proposed churn prediction procedure was then used to build the CBFF, which orchestrates in a novel way elements derived from the literature which the author's own ideas and understanding. The proposed churn prediction procedure and the CBFF aim to address open issues identified in the customer churn area and become a decision making tool in the hands of subscription-based organizations.

- **Objective 3:** *To examine established research methodologies, select an appropriate methodology for this research, and test the conceptual framework in the practical arena.*

The author investigated the literature to identify a suitable research methodology to test the proposed conceptual framework. Chapter 4 presents the steps followed to select and justify the research methodology adopted in this research. The author also justifies the selection of the interpretivism stance and the adoption of qualitative research strategy as more appropriate for the research theme. Furthermore, the selection of the exploratory case study research strategy was justified.

- **Objective 4:** *To extrapolate findings and provide a novel contribution to customer churn forecasting for subscription-based organizations.*

The empirical data collected from the two case organizations were presented and analyzed in Chapter 5. During Chapter 5 the proposed conceptualizations were tested and evaluated. The outcome of the empirical data investigation concluded that the churn prediction procedure and the CBFF could be used as a decision making tool from subscription-based organizations to enhance their attempt to predict churn. Besides, additions were added in the proposed conceptual framework as those covered thoroughly during Chapters 6 and 7. Moreover, the novel contribution of this research is reported and analyzed in detail in Chapter 7. In brief, the main novelties of the CBFF include: i)the definition of a churn prediction procedure which enhances the efforts taken by subscription-based organizations to face customer churn, ii) its usage as a decision making tool for subscription based organizations against churn, iii) the data analysis on customer data helping organizations make more accurate customer profiles, that increases customer retention, iv) the support on selecting a suitable algorithm for the classification process of customer churn through the usage of meta-learning techniques, v) the provision of dynamic customer churn prediction through the interpretation of the churn prediction outcome and vi) the proposal of churn countermeasures from insights interpretation.

The development, testing and revision of the conceptual framework (CBFF), denotes the accomplishment of the above objectives, which became possible by examining the limitations of the established norms in churn prediction for subscription-based organizations. This thesis contributes to both theory and practice regarding the churn prediction-related actions of subscription-based organizations.

7.4 Novel Contribution

The outcome of this thesis contributes in extending the body of knowledge in the domain of customer churn in subscription-based organizations. In particular, the proposed framework could be used as a decision making tool from subscription-based organizations in their attempts against churn. Moreover, CBFF could be used by future researchers who need to investigate the churn prediction domain. The novel contributions resulted from this research are summarized in Figure 7.1 and described in detail in this section.

Figure 7.1: Novel Contribution

- **Novelty 1 – Provides a Churn Prediction Procedure:** A churn prediction procedure provides the steps to be followed by the organizations to predict customer churn. The churn prediction procedure fills the gap identified in the absence of a defined procedure for churn prediction.

- **Novelty 2 – Offers a decision-making tool for churn prediction:** The CBFF can be used as a decision-making tool to reduce customer churn and increase the effectiveness of the two case organizations. Although, CBFF is innovative in its entirety, its building blocks are also innovative as they suggest new ways of studying and predicting customer churn.

- **Novelty 3 - Enhances customer profiling through data analysis:** The proposed framework introduces the Business Case Analysis Element, that allows organizations to analyze business parameters, (e.g., business needs, goals, desired outcomes, etc.). The novelty derives from the inclusion of the Business Case Analysis Element that performs the data analysis for the subscription-based organization, which has an important impact on the success of the churn prediction procedure.

- **Novelty 4 - Supports the selection of a suitable classification algorithm:** The research on customer churn revealed the absence of a comprehensive classification algorithm that serves every case. Taking that into consideration, the implementation of meta-learning techniques [105] was suggested in the proposed Prediction Model Element of the CBFF. Meta-learning techniques are ML techniques that, achieve faster learning based on prior experience and decide the most appropriate classification algorithm to be used. In summary, the CBFF fills this gap by a *priori* suggestion for a suitable classification algorithm that better fits each case based on the feed from business analysis and a combination of data sources deriving from them.

- **Novelty 5 - Provides dynamic and robust customer churn prediction:** The conceptual framework implements a novelty feature that evolves previously proposed churn prediction approaches. More specifically, the interpreted results deriving from the proposed Insights Element during a churn prediction cycle retrofit the proposed Feature Selection Element, as well as the Business Case Analysis Element. This feature supports the development of new features or the inclusion of additional meaningful features for the next churn prediction attempt. Additionally, this feature enhances the prediction process as it establishes more accurate predictions and actions for churn prevention.

- **Novelty 6 - Suggests churn prevention countermeasures:** Most of the previously proposed research on churn prediction remains at the stage of churn prediction results. The fundamental purpose of making a churn prediction is first, to identify a possible churning customer and second, to react before they confirm they are leaving the organization. The outcome of the churn prediction process should be taken into account by organizations as a method of proposing possible countermeasures against customer churn. The customer behavior framework includes a feature that takes into consideration the result of customer churn and proposes countermeasures to decrease the churn rate. The countermeasures are calculated using business insights (e.g., previous churn prediction process, if they exist) alongside business-case desired outcomes and generally desired outcomes for similar businesses in the same area.

7.5 Research Limitations

This research on customer churn prediction for subscription-based organizations was conducted by investigating two telecommunication providers; one in the EU and one outside the EU. Despite the differences in structure and market, the limitation was that they were both from the telecommunication industry so they had more similar characteristics compared to other subscription-based organizations.

Furthermore, limitations derived from the research methodology adopted for this research, which is presented and justified in Chapter 4. The qualitative interpretative multiple case study was adopted as it allowed the investigation of the phenomenon in a real-life context and to collect rich data using multiple methods. However, collecting and analyzing the empirical data collection proved very time-consuming. Furthermore, the interpretation of outcomes from a subject point of view without any degree of bias was also a limitation; however, the author used data triangulation during empirical data collection to address this. In addition to this, the results can not be generalized.

Finally, GDPR [90] meant that collecting the empirical data needed for the investigation was extremely difficult, which was a limitation. Several subscription-organizations within the EU who matched the research protocol outlined in Section 4.5 were contacted and despite the positive early signs of willingness to collaborate, with two signing a Non-disclosure Agreement (NDA), one of the companies refused to participate at the last minute due to fear of breaking GDPR [90]. However, the author reacted immediately and replaced the company with another organization from the subscription-based domain based on the research protocol. These changes cost additional time and delayed the completion of the empirical data investigation.

7.6 Future Research Work

Along with the limitations of this research, the author suggests that further research is needed on customer churn prediction. Therefore, recommendations for future researchers are:

- This empirical data investigation was based on two case organizations from the telecommunication sector. Telecommunication providers are part of the subscription-based domain but have specific characteristics that cannot be generalized to all subscription-based companies. Thus, extending this research in more subscription-based organizations such as content-platforms (e.g., Netflix, HBO, etc.,), could be very interesting.
- The online game industry has a significant amount of subscriptions, which makes online game companies one of the largest stakeholders in the subscription-based domain. Therefore, the same research could be conducted with the online game industry, but also with other domains and industries.
- The outcomes of this research (e.g., churn prediction procedure, CBFF), can be tested through the patent proposed by Miles and Huberman [80] and includes a sequential process where qualitative research is followed by quantitative and then again by qualitative. Such a process will offer a better understanding of the phenomenon under study and allow generalization.
- This research revealed the absence of a comprehensive guide for churn prediction. The conceptual framework aims to fill this gap and a churn prediction process for subscription-based organizations was conducted. Future research may transform the Customer Behavior Forecasting Framework and its elements into a large-scale questionnaire and attempt to collect new data. Such an approach should offer the researcher the opportunity to examine attributes, tasks, and relationships between the elements of the Customer Behavior Forecasting Framework based on data collected from a significant proportion of the subscription-based domain. Therefore, the elements of the Customer Behavior Forecasting Framework could be better verified, understood, and clarified. It is therefore suggested that the revised framework should be tested using quantitative research methods.
- Another recommendation is an in-depth analysis of each element in the proposed conceptual framework. Such an analysis should enrich the tasks and processes included in each of the elements and the whole structure of the Customer Behavior Forecasting Framework.

- Based on the above recommendation it may be investigated the option to convert the CBFF in a software tool.

References

[1] Adnan Amin et al. "Churn prediction in telecommunication industry using rough set approach". In: *New Trends in Computational Collective Intelligence*. Springer, 2015, pp. 83–95.

[2] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. "Customer churn prediction in telecommunications". In: *Expert Systems with Applications* 39.1 (2012), pp. 1414–1425.

[3] Kristof Coussement, Stefan Lessmann, and Geert Verstraeten. "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry". In: *Decision Support Systems* 95 (2017), pp. 27–36.

[4] Adnan Amin et al. "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study". In: *IEEE Access* 4 (2016), pp. 7940–7957.

[5] Sandra Mitrovi et al. "Scalable RFM-Enriched Representation Learning for Churn Prediction". In: *Proceedings of the fourth IEEE International Conference on Data Science and Advanced Analytics (DSAA2017)*. 2017.

[6] Techtarget.com. *churn rate*. Feb. 2017. URL: http://searchcrm.techtarget.com/definition/churn-rate.

[7] L. Katelaris and Themistocleous M. "Predicting Customer Churn: Customer Behavior Forecasting for Subscription-Based Organizations". In: vol. 299. Cham: Springer International Publishing, Aug. 2017, pp. 128–135.

[8] Estelle Phillips, Derek Pugh, et al. *How to get a PhD: A handbook for students and their supervisors*. McGraw-Hill Education (UK), 2010.

[9] V Kumar and Werner Reinartz. "Creating enduring customer value". In: *Journal of Marketing* 80.6 (2016), pp. 36–68.

[10] Alonso Raul Melgarejo Galvan and Katerine Rocio Clavo Navarro. "Big Data Architecture for Predicting Churn Risk in Mobile Phone Companies". In: *Annual International Symposium on Information Management and Big Data*. Springer. 2015, pp. 120–132.

[11] Soumit Sain and Silvio Wilde. *Customer knowledge management*. Springer International Publishing, 2014.

[12] Janet R McColl-Kennedy. "Customer Satisfaction, Assessment, Intentions and Outcome Behaviors of Dyadic Service Encounters: A Conceyfual Model". In: *Proceedings of the 1998 Academy of Marketing Science (AMS) Annual Conference*. Springer. 2015, pp. 48–54.

[13] Alex Rawson, Ewan Duncan, and Conor Jones. "The truth about customer experience". In: *Harvard Business Review* 91.9 (2013), pp. 90–98.

[14] L. Katelaris and Themistocleous M. “Mobile Number Portability Using a Reliable Cloud Database Appliance to Match Predictable Performance”. In: vol. 341. Cham: Springer International Publishing, Jan. 2019, pp. 185–190.

[15] Josee Bloemer, Ko De Ruyter, and Pascal Peeters. “Investigating drivers of bank loyalty: the complex relationship between image, service quality and satisfaction”. In: *International Journal of bank marketing* 16.7 (1998), pp. 276–286.

[16] Mohammed Abdul Haque Farquad, Vadlamani Ravi, and S Bapi Raju. “Churn prediction using comprehensible support vector machine: An analytical CRM application”. In: *Applied Soft Computing* 19 (2014), pp. 31–40.

[17] Kristof Coussement and Koen W De Bock. “Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning”. In: *Journal of Business Research* 66.9 (2013), pp. 1629–1636.

[18] Makoto Abe. “5 Deriving Customer Lifetime Value from RFM Measures: Insights into Customer Retention and Acquisition”. In: (2015).

[19] Jin Baek Kim. “An empirical study on consumer first purchase intention in online shopping: integrating initial trust and TAM”. In: *Electronic Commerce Research* 12.2 (2012), pp. 125–150.

[20] Lerzan Aksoy et al. “A cross-national investigation of the satisfaction and loyalty linkage for mobile telecommunications services across eight countries”. In: *Journal of Interactive Marketing* 27.1 (2013), pp. 74–82.

[21] invesp. *invesp*. Sept. 2017. URL: https://www.invespcro.com/blog/customer-acquisition-retention.

[22] Gartner. *Market Trends: Prepare Now to Smooth the Inevitable Transition to a Subscription-Based Business Model*. G00330396. http://www.gartner.com/, Apr. 19, 2017, p. 13. URL: http://info.zuora.com/rs/602-QGZ-447/images/market_trends_prepare_now_to_330396.pdf.

[23] IDC. *IDC FutureScape: Worldwide IT Industry 2017 Predictions*. www.idc.com, Nov. 2016, p. 20. URL: http://www.ineturl.com/bcBoI.

[24] Pearl Brereton et al. “Lessons from applying the systematic literature review process within the software engineering domain”. In: *Journal of systems and software* 80.4 (2007), pp. 571–583.

[25] Harris M Cooper. “Organizing knowledge syntheses: A taxonomy of literature reviews”. In: *Knowledge, Technology & Policy* 1.1 (1988), pp. 104–126.

[26] Chitu Okoli and Kira Schabram. “A guide to conducting a systematic literature review of information systems research”. In: (2010).

[27] Sebastian K Boell and Dubravka Cecez-Kecmanovic. “On being ‘systematic’in literature reviews in IS”. In: *Journal of Information Technology* 30.2 (2015), pp. 161–173.

[28] Yair Levy and Timothy J Ellis. “A systems approach to conduct an effective literature review in support of information systems research.” In: *Informing Science* 9 (2006).

[29] David L Sackett. *Evidence-based Medicine How to practice and teach EBM*. WB Saunders Company, 1997.

[30] Barbara A Kitchenham, Tore Dyba, and Magne Jorgensen. "Evidence-based software engineering". In: *Proceedings of the 26th international conference on software engineering*. IEEE Computer Society. 2004, pp. 273–281.

[31] Tore Dyba, Barbara A Kitchenham, and Magne Jorgensen. "Evidence-based software engineering for practitioners". In: *IEEE software* 22.1 (2005), pp. 58–65.

[32] Yaya Xie et al. "Customer churn prediction using improved balanced random forests". In: *Expert Systems with Applications* 36.3 (2009), pp. 5445–5449.

[33] Anne-Wil Harzing et al. "Publish or perish". In: (2007).

[34] Richard Van Noorden. "Online collaboration: Scientists and the social network". In: *Nature* 512.7513 (2014), pp. 126–129.

[35] David Matthews. "Do academic social networks share academics' interests". In: *Times Higher Education* (2016), pp. 7–13.

[36] Ssu-Han Chen. "The gamma CUSUM chart method for online customer churn prediction". In: *Electronic Commerce Research and Applications* 17 (2016), pp. 99–111.

[37] Michel Ballings and Dirk Van den Poel. "Customer event history for churn prediction: How long is long enough?" In: *Expert Systems with Applications* 39.18 (2012), pp. 13517–13522.

[38] Lin Xu et al. "Research on Telecom Customer Churn Prediction Based on Customer Value Classification in 3G Environment". In: (2016).

[39] Adnan Amin et al. "Just-in-time customer churn prediction in the telecommunication sector". In: *The Journal of Supercomputing* (2017), pp. 1–25.

[40] Wouter Verbeke et al. "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach". In: *European Journal of Operational Research* 218.1 (2012), pp. 211–229.

[41] Wouter Verbeke, David Martens, and Bart Baesens. "Social network analysis for customer churn prediction". In: *Applied Soft Computing* 14 (2014), pp. 431–446.

[42] Erdem Kaya et al. "Behavioral attributes and financial churn prediction". In: *EPJ Data Science* 7.1 (2018), p. 41.

[43] Thanasis Vafeiadis et al. "A comparison of machine learning techniques for customer churn prediction". In: *Simulation Modelling Practice and Theory* 55 (2015), pp. 1–9.

[44] Zdzisław Pawlak. "Rough sets". In: *International journal of computer & information sciences* 11.5 (1982), pp. 341–356.

[45] HS Nguyen and SH Nguyen. "Analysis of stulong data by rough set exploration system (RSES)". In: *Proceedings of the ECML/PKDD Workshop*. 2003, pp. 71–82.

[46] Jakub Wróblewski. "Genetic algorithms in decomposition and classification problems". In: *Rough Sets in Knowledge Discovery 2*. Springer, 1998, pp. 471–487.

[47] Ewa Orlowska. *Incomplete information: Rough set analysis*. Vol. 13. Physica, 2013.

[48] Mladen Radisic. "Just-in-time concept". In: *Retrieved May* 20 (2009), p. 2009.

[49] Gemma Catolino. "Just-in-time bug prediction in mobile applications: the domain matters!" In: *Mobile Software Engineering and Systems (MOBILESoft), 2017 IEEE/ACM 4th International Conference on*. IEEE. 2017, pp. 201–202.

[50] Takafumi Fukushima et al. "An empirical study of just-in-time defect prediction using cross-project models". In: *Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM. 2014, pp. 172–181.

[51] Adnan Idris and Asifullah Khan. "Ensemble based efficient churn prediction model for telecom". In: *Frontiers of Information Technology (FIT), 2014 12th International Conference on*. IEEE. 2014, pp. 238–244.

[52] Scott A Neslin et al. "Defection detection: Measuring and understanding the predictive accuracy of customer churn models". In: *Journal of marketing research* 43.2 (2006), pp. 204–211.

[53] Nicholas A Christakis and James H Fowler. "The spread of obesity in a large social network over 32 years". In: *New England journal of medicine* 357.4 (2007), pp. 370–379.

[54] Christopher Z Mooney. *Monte carlo simulation*. Vol. 116. Sage Publications, 1997.

[55] M Clemente, V Giner-Bosch, and S San Mati□s. "Assessing classification methods for churn prediction by composite indicators". In: *Manuscript, Dept. of Applied Statistics, OR & Quality, UniversitatPolitècnica de València, Camino de Vera s/n* 46022 (2010).

[56] Robert E Schapire. "A brief introduction to boosting". In: *Ijcai*. Vol. 99. 1999, pp. 1401–1406.

[57] Michael Collins, Robert E Schapire, and Yoram Singer. "Logistic regression, AdaBoost and Bregman distances". In: *Machine Learning* 48.1-3 (2002), pp. 253–285.

[58] S-I Horikawa, Takeshi Furuhashi, and Yoshiki Uchikawa. "On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm". In: *IEEE transactions on Neural Networks* 3.5 (1992), pp. 801–806.

[59] Vadim Smolyakov. *Ensemble Learning to Improve Machine Learning Results*. 2017. URL: https://blog.statsbot.co/ensemble-learning-d1dcd548e936 (visited on 02/2019).

[60] Kaggle. *Kaggle Competitions*. 2019. URL: https://www.kaggle.com/competitions (visited on 12/09/2019).

[61] ACM Special Interest Group. *KDD Cup*. 2019. URL: https://www.kdd.org/kdd2019/kdd-cup (visited on 02/19/2019).

[62] Wouter Verbeke et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques". In: *Expert systems with applications* 38.3 (2011), pp. 2354–2364.

[63] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[64] Jonathan Wilson. *Essentials of business research: A guide to doing your research project*. Sage, 2014.

[65] Robert D Galliers and Frank F Land. "Choosing appropriate information systems research methodologies". In: *Communications of the ACM* 30.11 (1987), pp. 901–902.

[66] Geoff Walsham. "Interpreting information systems in organizations". In: (2015).

[67] Robert D Galliers. "Research issues in information systems". In: *Journal of information technology* 8.2 (1993), pp. 92–98.

[68] Wanda J Orlikowski and Jack J Baroudi. "Studying information technology in organizations: Research approaches and assumptions". In: *Information systems research* 2.1 (1991), pp. 1–28.

[69] Göran Goldkuhl. "Pragmatism vs interpretivism in qualitative information systems research". In: *European journal of information systems* 21.2 (2012), pp. 135–146.

[70] Bob Ryan, Robert William Scapens, and Michael Theobald. "Research method and methodology in finance and accounting". In: (2002).

[71] Ellie Fossey et al. "Understanding and evaluating qualitative research". In: *Australian and New Zealand journal of psychiatry* 36.6 (2002), pp. 717–732.

[72] John W Creswell et al. "Advanced mixed methods research designs". In: *Handbook of mixed methods in social and behavioral research* 209 (2003), p. 240.

[73] Michael D Myers and David Avison. *Qualitative research in information systems: a reader*. Sage, 2002.

[74] Juliet Corbin, Anselm Strauss, and Anselm L Strauss. *Basics of qualitative research*. sage, 2014.

[75] Thomas R Black. *Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics*. Sage, 1999.

[76] Norman K Denzin, Yvonna S Lincoln, and Linda Tuhiwai Smith. *Handbook of critical and indigenous methodologies*. Sage, 2008.

[77] Kenneth F Hyde. "Recognising deductive processes in qualitative research". In: *Qualitative market research: An international journal* 3.2 (2000), pp. 82–90.

[78] Natasha Mack. "Qualitative research methods: A data collector's field guide". In: (2005).

[79] John Van Maanen. "Reclaiming qualitative methods for organizational research: A preface". In: *Administrative science quarterly* 24.4 (1979), pp. 520–526.

[80] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook*. sage, 1994.

[81] Izak Benbasat, David K Goldstein, and Melissa Mead. "The case research strategy in studies of information systems". In: *MIS quarterly* (1987), pp. 369–386.

[82] Jill Collis and Roger Hussey. *Business research: A practical guide for undergraduate and postgraduate students*. Macmillan International Higher Education, 2013.

[83] Stephen Kemmis and Robin McTaggart. *Participatory action research: Communicative action and the public sphere*. Sage Publications Ltd, 2005.

[84] Barney G Glaser and Anselm L Strauss. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.

[85] Robert K Yin. *Applications of case study research*. SAGE Publications, 2011.

[86] Angèle LM Cavaye. "Case study research: a multi-faceted research approach for IS". In: *Information systems journal* 6.3 (1996), pp. 227–242.

[87] Andrew M Pettigrew. "Contextualist research and the study of organizational change processes". In: *Research methods in information systems* 1 (1985), pp. 53–78.

[88] M Lynne Markus. "Case selection in a disconfirmatory case study". In: *The information systems research challenge: Qualitative research methods* 1 (1989), pp. 20–26.

[89] Kathleen M Eisenhardt. “Better stories and better constructs: The case for rigor and comparative logic”. In: *Academy of Management review* 16.3 (1991), pp. 620–627.

[90] European Union Official Website European Union. *GDPR EU Law*. URL: https://ec.europa.eu/info/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en#abouttheregulationanddataprotection.

[91] Marinos Themistocleous. “Justifying the decisions for EAI implementations: a validated proposition of influential factors”. In: *Journal of Enterprise Information Management* 17.2 (2004), pp. 85–104.

[92] Marinos G Thermistocleous. “Evaluating the adoption of enterprise application integration in multinational organisations”. PhD thesis. Brunel University, School of Information Systems, Computing and Mathematics, 2002.

[93] Colin Robson. *Real world research*. Vol. 3. Wiley Chichester, 2011.

[94] Norman K Denzin. *The research act: A theoretical introduction to sociological methods*. Routledge, 2017.

[95] Sandra Mathison. “Why triangulate?” In: *Educational researcher* 17.2 (1988), pp. 13–17.

[96] Valerie J Janesick. “The choreography of qualitative research design”. In: *Handbook of qualitative research.* (2000), pp. 379–399.

[97] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[98] *Regular and irregular time series*. URL: https://www.ibm.com/support/knowledgecenter/en/SSCRJU_4.0.0/com.ibm.streams.toolkits.doc/doc/tkcom.ibm.streams.timeseries/tkcom.ibm.streams.timeseries2.html.

[99] SAS official Website. *SAS Business Intelligence & Analytics*. URL: https://www.sas.com/en_us/solutions/business-intelligence.html.

[100] Delloite Official Website Deloitte. *Data is the new gold*. URL: https://www2.deloitte.com/global/en/pages/real-estate/articles/future-real-estate-data-new-gold.html.

[101] KPMG Official Website Kristen Morgan. *Future Ready: Data is the new gold*. URL: https://info.kpmg.us/news-perspectives/future-ready/future-ready-data-is-the-new-gold.html.

[102] Wikipedia Website Wikipedia. *Big Four Accounting Firms*. URL: https://en.wikipedia.org/wiki/Big_Four_accounting_firms.

[103] Investopedia Website Investopedia. *ARPU*. URL: https://www.investopedia.com/terms/a/arpu.asp.

[104] Online community of data scientists Kaggle and machine learners. *Kaggle*. URL: https://www.kaggle.com/.

[105] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, eds. *Automatic Machine Learning: Methods, Systems, Challenges*. In press, available at http://automl.org/book. Springer, 2018.

[106] Sergio Ramírez-Gallego et al. "A survey on data preprocessing for data stream mining: Current status and future directions". In: *Neurocomputing* 239 (2017), pp. 39–57.

[107] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015.

[108] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[109] Christopher Whitrow et al. "Transaction aggregation as a strategy for credit card fraud detection". In: *Data Mining and Knowledge Discovery* 18.1 (2009), pp. 30–55.

[110] Bethany Percha, Yael Garten, and Russ B Altman. "Discovery and explanation of drug-drug interactions via text mining". In: *Pacific symposium on biocomputing. Pacific symposium on biocomputing*. NIH Public Access. 2012, p. 410.

[111] Carolin Strobl et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1 (2007), p. 25.

[112] Jigar Patel et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques". In: *Expert Systems with Applications* 42.1 (2015), pp. 259–268.

[113] Hyunju Noh, Minjung Kwak, and Ingoo Han. "Improving the prediction performance of customer behavior through multiple imputation". In: *Intelligent Data Analysis* 8.6 (2004), pp. 563–577.

[114] Cathy Lawson and Douglas C Montgomery. "Logistic regression analysis of customer satisfaction data". In: *Quality and reliability engineering international* 22.8 (2006), pp. 971–984.

[115] Yu Zhao et al. "Customer churn prediction using improved one-class support vector machine". In: *Advanced data mining and applications* (2005), pp. 731–731.

[116] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Appendix A

Acronyms

Acronym	Definition
ADASYN	Adaptive Synthetic Sampling Approach
ARPU	Average Revenue Per User
AUC	Area Under the Curve
Bagged CART	Bagged Classification and Regression Tree
CA	Covering Algorithm
CBFF	Customer Behavior Forecasting Framework
CLV	Customer Lifetime Value
CIP	Class Imbalance Problem
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
DSL	Digital Subscriber Line
DTP	Data Preparation Treatments
EA	Evolutionary Algorithm
EU	European Union
GA	Genetic Algorithm
HFC	Hybrid Fiber Coaxial
JIT	Just in Time
ICOTE	Immune Centroids Oversampling Technique
ICT	Information and Communication Technology
IDC	International Data Corporation
IoT	Internet of Things
IS	Information Systems
IT	Information Technology
KNN	K-nearest Neighbors Algorithm
LEM	Learning from Examples Modules
LR	Literature Review
LSSVM	Least Square Support Vector Machines
ML	Machine Learning
MTDF	Mega-Trend Diffusion Function

MWMOTE	Majority Weighted Minority Oversampling Technique
NDA	Non-Disclosure Agreement
OECD	Organization for Economic Co-operation and Development
PART	Projective Adaptive Resonance Theory
RBF	Radial Basis Function
REF	Recursive Feature Elimination
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
ROI	Return On Investment
RST	Rough Set Theory
SMOTE	Synthetic Minority Oversampling Technique
SMS	Short Message Service
SVM	Support Vector Machine
SVM-POLY	Support Vector Machine Polynomial
SVM-RBF	Support Vector Machine Radial Basis
TDL	Top Decile Lift
TRkNN	Couples Top-N Reverse k-Nearest Neighbor

Table A.1: Acronyms and Abbreviations

Appendix B

Previous Studies and Approaches on Churn Prediction

B.1 Churn Prediction Approaches

During the last decades, various attempts made from companies in the subscription-based domain to forecast customer behavior and prevent churn, including methods and techniques which differ in their statistical approaches, amount of variables, in the prediction model etc,. Following an overview of the most known churn prediction methods and techniques based on Machine Learning (ML) is given.

B.1.1 Random Forest:

Random forest method, has been introduced by Breiman [108] and present a new and powerfully statistical classifier, which found to perform very well compared to many other classifiers. Random forest introduced a solution to decision trees' problem where small change in the data in most of the times results in very different series of splits, which in turn lowers accuracy when validating the training data. Random forest works as a big collection of decorrelated decision trees. Based on that, the word "forest", is used to describe the multiple random trees used by this method. More specifically, random forest uses a randomly selected subset of *m* predictors to grow trees on bootstrap sample of the training data. Followed, the large number of randomly generated trees, each tree votes for the most popular class (*m* which present the number of the predictors is much smaller than the total number of variables in the model). Aggregating all votes from different trees, conclude to class label prediction. This technique is easy to implement because of the small number of parameters need to be set (*m*:predictors, total number of trees to be generated). Below (see equation 2.1) there is a matrix *S* representing training samples to be submitted in the algorithm to create the classification model. The *f* representing features of different samples (e.g., f_{A1} represents the feature A of the 1st sample). The last column in the matrix C_1.. C_N represents the lots of features and training class in the training samples. The aim is to create a random forest to classify the sample set. To achieve that, random subsets are created from *S* (see equations 2.2 and 2.3) with random values. The S_1, S_2, S_M subsets are used to create decision trees, with different variations of the main classification. Followed, the different decision trees will be used to create a ranking of the classifiers.

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & \dots & C_1 \\ & \multicolumn{3}{c}{\dots\dots\dots\dots\dots} & \\ f_{AN} & f_{BN} & f_N & \dots & C_N \end{bmatrix} \tag{B.1}$$

$$S_1 = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & \dots & C_1 \\ f_{A5} & f_{B5} & f_{C5} & \dots & C_5 \\ & \multicolumn{3}{c}{\dots\dots\dots\dots\dots} & \\ f_{A25} & f_{B25} & f_{C25} & \dots & C_{25} \end{bmatrix} S_2 = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & \dots & C_4 \\ f_{A6} & f_6 & f_{C6} & \dots & C_6 \\ & \multicolumn{3}{c}{\dots\dots\dots\dots\dots} & \\ f_{A38} & f_{B38} & f_{C38} & \dots & C_{38} \end{bmatrix} \tag{B.2}$$

$$S_M = \begin{bmatrix} f_{A9} & f_{B9} & f_{C9} & \dots & C_9 \\ f_{A12} & f_{12} & f_{C12} & \dots & C_{12} \\ & \dots & \dots & \dots & \\ f_{A23} & f_{B23} & f_{C23} & \dots & C_{23} \end{bmatrix} \tag{B.3}$$

Taking the advantages of random forest it can be said that, random forest could be used for both classification and regression task, it is able to handle the missing values and maintaining accuracy for missing data the same time. Also, if there more trees in the model, random forest classifiers wont over-lift the model. Last but not least, random forest has the ability to handle large data sets with higher dimensionality. Despite the advantages random forest has also some disadvantages. Random forest could be used for both classification and regression tasks as said before, however is not as good as regression. To explain that, in case of regression it does not give predictions beyond the range of trained data. Another disadvantage of random forest, is the limited control on what the model does.

There are various applications in different domains where random forest could be used. Such applications but not limited to are in bank sector defining loyal customers or fraud detection discovering in transactions [109]. Also, in medicine where it could be used to identify interactions between drugs [110] or predicting a certain disease [111]. Another sector where random forest could be used is the stock market identifying stock behavior [112].

B.1.2 Regression Analysis:

Regression analysis is considered to be an effective technique for customer behavior forecasting [113, 114]. Regression analysis apart from others, includes linear and logistic regression methods where each one is suitable in data-sets with specific characteristics. Both of the regression analysis methods (linear, logistic) have been used for prediction purposes. Linear regression is useful in problems where the values are continues, as it can give results between 0 and 1 ($0 < x < 1$). Linear regression attempts to model the relationship between two variables (e.g., x,y) by setting a linear equation to the observed data. Between the two variables, one (x) is the explanatory variable and the other (y) is the depended variable. Linear regression could be applied in two ways. The first way is to define a possible relationship between two variables (e.g., x and y), or to find a statistically significant relationship between the two variables. Linear regression could be used to forecast the stock prices overtime, the return on investment (ROI) for a product, the bitcoin price for the next months etc. Logistic regression in the other hand, is a well-known classification method for

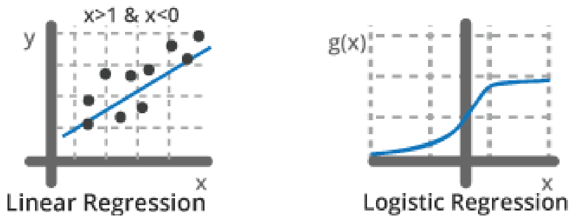


Figure B.1: Examples of Logistic and Linear Regression Graphs

predictions in data-sets with binary attributes.

B.1.3 Neural Networks:

Neural Networks are common in finding complexity of no-linear functions. Neural Networks have been taken into serious consideration by researchers and practitioners because they provide a prediction alongside with its possibility. In neural networks the idea is that each variable is associated with a weight. Then a combination of weighted variables process is running to develop the prediction model. The main disadvantage of neural networks in prediction process is that they need a large volume of data set and are time consuming while trying to weight the predictor variables. However, the achieved accuracy of prediction model in most of the times, overcomes in performance other prediction techniques like logic regression.

B.1.4 Support Vector Machines:

Support vector machine is a very popular discrimination classifier [16, 115] introduced by Cortes et., al [63], that is able in binary classification problems. Using support vector machine in a set of variables, each variable separated to each of two categories. Then the support vector machine training algorithm develops a model where it assigns new variables to one or the other category. The output of support vector machine prediction model is points in space divided by a clear gap. Support vector machine as prediction method is equivalent for data-set in which data are linearly separable, but the truth is that in real world data is often not linearly separable. In addition to performing linear classification,support vector machines use kernel function to perform non-linear classification end enhance feasibility of linear separation.

Appendix C

Queries Submission

C.1 IEEEXplore Keywords Submission

Submitted queries in IEEEXplore database:

- **Query 1:** *(((((customer churn prediction) AND "Abstract":customer churn prediction) OR "Author Keywords":customer churn prediction) OR "IEEE Terms":customer churn prediction) OR "Publication Title":customer churn prediction)* - Returned 80 results sorted by relevance.
- **Query 2:** *((((customer churn) AND prediction model) OR "Abstract":prediction model) AND "Abstract":customer churn)* - Returned 106 results sorted by relevance.
- **Query 3:** *(((((customer churn) AND prediction framework) OR "Abstract":prediction framework) AND "Abstract":customer churn))* - Returned 12 results sorted by relevance.

All the above queries were submitted, with year filter range 2012 - present. The queries resulted in 198 documents from IEEEXplore database with some duplicates. So, the author, to cope with this combined the queries and got the union of the result sets which were 110 documents instead of 198 documents as presented in Figure C.1 and Table C.1.

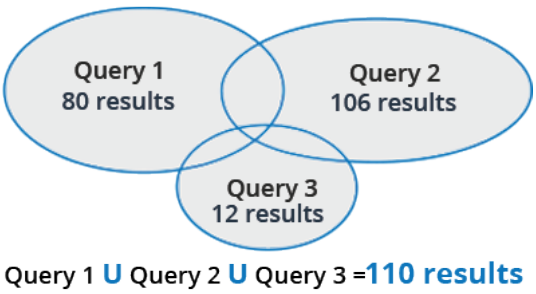


Figure C.1: IEEEXplore Queries Union of Returning Results

C.2 SpringerLink Keywords Submission

Submitted queries using advance search option:

- **Query1:** *with the exact phrase "customer churn prediction"* - Returned 93 results sorted by relevance.
- **Query 2:** *customer AND churn AND "prediction model"* - Returned 173 results sorted by relevance.

All the above queries were submitted, with year filter range 2012 - present . The author followed the same process as in IEEEXplore, to get the union of both result sets, which were 210 articles as presented in Figure C.2 and Table C.2.

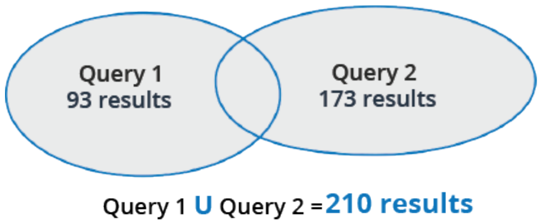


Figure C.2: SpringerLink Queries Union of Returning Results

C.3 Google Scholar Keywords Submission

Submitted queries:

- **Query 1:** *with the exact phrase (allintitle: "customer churn prediction")* - Returned 154 results sorted by number of citations.
- **Query 2:** *with the exact phrase ("customer churn prediction model")* - Returned 200 results sorted by number of citations.
- **Query 3:** *with the exact phrase ("customer churn prediction framework")* - Returned 3 results sorted by number of citations.

All the above queries were submitted, with year filter range 2012 - present . The first two queries have a property "allintititle", which means that the submitted keyword have to be included in the title of the article. The author followed the same process as in previous electronic databases, to get the union of both result sets, which were 298 articles as presented in Figure C.3 and Table C.3.

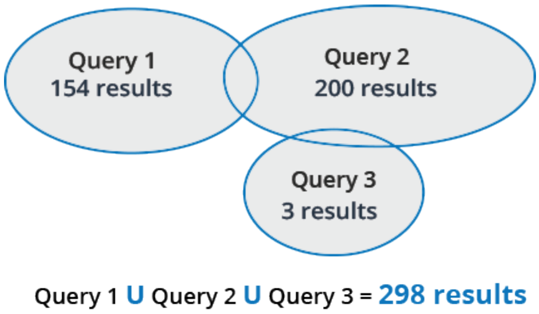


Figure C.3: Google Scholar Queries Union of Returning Results

C.4 Result Tables

IEEEXplore Results

Paper Title	Published in	Authors	Publication Year
An application of the CORER classifier on customer churn prediction	6th International Symposium on Telecommunications (IST)	İ. O. Yiğit; H. Shourabizadeh	2012
Customer Churn Prediction for Telecom Services	2012 IEEE 36th Annual Computer Software and Applications Conference	D. Do; P. Huynh; P. Vo; T. Vu	2012
Genetic Programming and Adaboosting based churn prediction for Telecom	2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)	A. Aditsania; Adiwijaya; A. L. Saonard	2012
Impact of social attributes on Predictive Analytics in telecommunication industry	2012 15th International Multitopic Conference (INMIC)	Q. Shen; H. Li; Q. Liao; W. Zhang; K. Kalilou	2012
Mobile phone customers churn prediction using elman and Jordan Recurrent Neural Network	2012 7th International Conference on Computing and Convergence Technology (ICCCT)	J. Zhang; J. Fu; C. Zhang; X. Ke; Z. Hu	2012
Predicting customer churn with extended one-class support vector machine	2012 8th International Conference on Natural Computation	A. P. Mauricio; J. M. M. Payawal; M. A. Dela Cueva; V. C. Quevedo	2012
(Case Study: Solico Food Industries Group)	The 5th Conference on Information and Knowledge Technology	W. Bi; M. Cai; M. Liu; G. Li	2013
A Dynamic Transfer Ensemble Model for Customer Churn Prediction	2013 Sixth International Conference on Business Intelligence and Financial Engineering	T. Verbraken; W. Verbeke; B. Baesens	2013
A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models	IEEE Transactions on Knowledge and Data Engineering	A. Ahmed; D. M. Linen	2013
Analytical Model of Customer Churn Based on Bayesian Network	2013 Ninth International Conference on Computational Intelligence and Security	M. Atkinson; R. Baxter; P. Brezany; O. Corcho; M. Galea; M. Parsons; D. Snelling; J. van Hemert	2013

Table C.1: IEEEXplore Results (continued)

Analytical Platform for Customer Relationship Management	The Data Bonanza:Improving Knowledge Discovery in Science, Engineering, and Business	R. Mohanty; K. J. Rani	2013
Churn prediction in subscriber management for mobile and wireless communications services	2013 IEEE Globecom Workshops (GC Wkshps)	W. M. C. Bandara; A. S. Perera; D. Alahakoon	2013
Churn prediction methodologies in the telecommunications sector: A survey	2013 International Conference on Advances in ICT for Emerging Regions (ICTer)	J. Semrl; A. Matei	2013
Data-Intensive Analysis	The Data Bonanza:Improving Knowledge Discovery in Science, Engineering, and Business	R. M. R. Dintakurthi; B. Venkatraman; P. Mahendran; S. Siddappa	2013
Forecasting bloggers' online behavior based on improved Pareto/NBD model	2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings	A. Idris; A. Khan; Y. S. Lee	2013
Predicting customers' future demand using data mining analysis: A case study of wireless communication customer	The 5th Conference on Information and Knowledge Technology	A. Akyamac; C. Phadke; D. Kushnir; H. Uzunalioglu	2013
Sociocentric and egocentric measures for identifying the key players in telecom social network	2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)	H. Li; D. Yang; L. Yang; YaoLu; X. Lin	2013
Telecom customer churn prediction based on imbalanced data resampling method	Proceedings of 2013 2nd International Conference on Measurement, Information and Control	Peng Li; Siben Li; Tingting Bi; Yang Liu	2013
Telecommunication subscribers' churn prediction model using machine learning	Eighth International Conference on Digital Information Management (ICDIM 2013)	K. Shapoval; T. Setzer	2013

Table C.1: IEEEXplore Results (continued)

Using Eye-Tracking Data of Advertisement Viewing Behavior to Predict Customer Churn	2013 IEEE 13th International Conference on Data Mining Workshops		2013
A Customer Churn Prediction Model in Telecom Industry Using Boosting	IEEE Transactions on Industrial Informatics	J. Xiao; Y. Wang; S. Wang	2014
Churn analysis: Predicting churners	Ninth International Conference on Digital Information Management (ICDIM 2014)	S. Jamil; A. Khan	2014
Customer Churn Prediction in Telecommunication Industry: With and without Counter-Example	2014 European Network Intelligence Conference	S. Yuan; S. Bai; M. Song; Z. Zhou	2014
Ensemble Based Efficient Churn Prediction Model for Telecom	2014 12th International Conference on Frontiers of Information Technology	R. Vadakattu; B. Panda; S. Narayan; H. Godhia	2014
Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest	2014 2nd International Conference on Information and Communication Technology (ICoICT)	P. Rothenbuehler; J. Runge; F. Garcin; B. Faltings	2014
Improving churn prediction in telecommunications using complementary fusion of multilayer features based on factorization and construction	The 26th Chinese Control and Decision Conference (2014 CCDC)	P. Casas; S. Wassermann	2014
Inferring potential users in mobile social networks	2014 International Conference on Data Science and Advanced Analytics (DSAA)	A. Chouiekh; E. H. I. E. Haj	2014
Predicting the influencers on wireless subscriber churn	2014 IEEE Wireless Communications and Networking Conference (WCNC)	A. Degbotse; A. K. Ang; N. Q. Vuong; J. S. K. Tan	2014

Table C.1: IEEEXplore Results (continued)

Telecom customer churn prediction method based on cluster stratified sampling logistic regression	International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014	S. A. Qureshi; A. S. Rehman; A. M. Qamar; A. Kamal; A. Rehman	2014
A case study for the churn prediction in Turksat internet service subscription	2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)	A. Mishra; U. S. Reddy	2015
Application of Computational Intelligence to Predict Churn and Non-Churn of Customers in Indian Telecommunication	2015 International Conference on Computational Intelligence and Communication Networks (CICN)	G. e. Xia; H. Wang; Y. Jiang	2015
Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty	2015 IEEE International Congress on Big Data	N. Forhad; M. S. Hussain; R. M. Rahman	2015
Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach	IEEE Transactions on Computational Intelligence and AI in Games	U. Yabas; H. C. Cankaya	2015
Combining local and social network classifiers to improve churn prediction	2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)	K. Gajowniczek; T. Ząbkowski; A. Orłowski	2015
Comparison of decision trees with Rényi and Tsallis entropy applied for imbalanced churn dataset	2015 Federated Conference on Computer Science and Information Systems (FedCSIS)	W. Zhang; L. Zhu	2015

Table C.1: IEEEXplore Results (continued)

Customer churn analysis in telecom industry	2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)	Q. Yang; X. Li; S. Kumar	2015
Customer Churn Aware Resource Allocation and Virtual Machine Placement in Cloud	2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems	K. Şen; N. G. Bayazıt	2015
Customer churn modelling in banking	2015 23nd Signal Processing and Communications Applications Conference (SIU)	U. Yabas; H. C. Cankaya; T. Ince	2015
Customer churn prediction in telecommunication	2015 23nd Signal Processing and Communications Applications Conference (SIU)	A. Amin; C. Khan; I. Ali; S. Anwar	2015
Customer Churn Prediction in Virtual Worlds	2015 IIAI 4th International Congress on Advanced Applied Informatics	I. M. M. Mitkees; S. M. Badr; A. I. B. ElSeddawy	2015
Echo State Network with SVM-readout for customer churn prediction	2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)	Xiaojun Wu; Sufang Meng	2015
Enterprise subscription churn prediction	2015 IEEE International Conference on Big Data (Big Data)	C. Sung; C. Y. Higgins; B. Zhang; Y. Choe	2015
Hidden Markov models for churn prediction	2015 SAI Intelligent Systems Conference (IntelliSys)	O. Z. Hashmi; S. Sheikh	2015

Table C.1: IEEEXplore Results (continued)

One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection	2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)	S. M. Donadelli; Y. C. Zhu; P. C. Rigby	2015
Organizational Volatility and Post-release Defects: A Replication Case Study Using Data from Google Chrome	2015 IEEE/ACM 12th Working Conference on Mining Software Repositories	A. Ahmad; A. Floris; L. Atzori	2015
Predicting behavior	IEEE Intelligent Systems	W. H. Lin; S. H. Chen	2015
Predicting home network problems using diverse data	2015 36th IEEE Sarnoff Symposium	P. Casas; A. D'Alconzo; F. Wamser; M. Seufert; B. Gardlo; A. Schwind; P. Tran-Gia; R. Schatz	2015
Profit Maximization Analysis Based on Data Mining and the Exponential Retention Model Assumption with Respect to Customer Churn Problems	2015 IEEE International Conference on Data Mining Workshop (ICDMW)	E. Stripling; S. vanden Broucke; K. Antonio; B. Baesens; M. Snoeck	2015
Profit maximizing logistic regression modeling for customer churn prediction	2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)	A. Mubeen; N. D. Abhinav; C. V. S. Swamy; K. H. Swetha; H. Rakesh	2015
Term of Contract and Portfolio Aware Churn Modeling in Telecommunication Campaigns	2015 IEEE 17th Conference on Business Informatics	F. Guo; H. L. Qin	2015
The Analysis of Customer Churns in e-Commerce Based on Decision Tree	2015 International Conference on Computer Science and Applications (CSA)	E. Diaz-Aviles; F. Pinelli; K. Lynch; Z. Nabi; Y. Gkoufas; E. Bouillet; F. Calabrese; E. Coughlan; P. Holland; J. Salzwedel	2015
Towards real-time customer experience prediction for telecommunication operators	2015 IEEE International Conference on Big Data (Big Data)	F. Napitu; M. A. Bijaksana; A. Trisetyarso; Y. Heryadi	2015

Table C.1: IEEEXplore Results (continued)

A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn	IEEE Transactions on Industrial Informatics	M. Gök; T. Özyer; J. Jida	2016
A comparative study of social network classifiers for predicting churn in the telecommunication industry	2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)	N. Lu; H. Lin; J. Lu; G. Zhang	2016
A simulator for generating realistic simulations of telecom customer behaviors	2016 24th Signal Processing and Communication Application Conference (SIU)	J. Basiri; F. Taghiyareh	2016
Analysis of churn prediction: A case study on telecommunication services in Macedonia	2016 24th Telecommunications Forum (TELFOR)	P. K. Dalvi; S. K. Khandge; A. Deomore; A. Bankar; V. A. Kanade	2016
Analysis of customer churn prediction in telecom industry using decision trees and logistic regression	2016 Symposium on Colossal Data Analysis and Networking (CDAN)	P. Sun; X. Guo; Y. Zhang; Z. Wu	2016
Application of customer churn prediction based on weighted selective ensembles	2016 3rd International Conference on Systems and Informatics (ICSAI)	R. Manongdo; G. Xu	2016
Applying client churn prediction modeling on home-based care services industry	2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)	V. Umayaparvathi; K. Iyakutti	2016
Attribute selection and Customer Churn Prediction in telecom industry	2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)	M. R. Khan; J. Manoj; A. Singh; J. Blumenstock	2016
Churn comprehension analysis for telecommunication industry using ALBA	2016 International Conference on Emerging Technologies (ICET)	H. C. Karapinar; A. Altay; G. Kayakutlu	2016

Table C.1: IEEEXplore Results (continued)

Churn detection and prediction in automotive supply industry	2016 Federated Conference on Computer Science and Information Systems (FedCSIS)	G. Esteves; J. Mendes-Moreira	2016
Churn perdiction in the telecom business	2016 Eleventh International Conference on Digital Information Management (ICDIM)	I. Franciska; B. Swaminathan	2016
Churn prediction by finding most influential nodes in social network	2016 International Conference on Computing, Analytics and Security Trends (CAST)	J. Xiao; X. Jiang; C. He; G. Teng	2016
Churn Prediction in Customer Relationship Management via GMDH-Based Multiple Classifiers Ensemble	IEEE Intelligent Systems	E. G. Castro; M. S. G. Tsuzuki	2016
Customer behavior patterns analysis in Indian mobile telecommunications industry	2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)	P. Wanchai	2016
Data warehouse based analysis on CDR to retain and acquire customers by targeted marketing	2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)	M. Atkinson; R. Baxter; P. Brezany; O. Corcho; M. Galea; M. Parsons; D. Snelling; J. van Hemert	2016
Decision support system for identifying customer churn based on buying patterns in a discrete manufacturing industry	2016 IEEE Annual India Conference (INDICON)	C. Chu; G. Xu; J. Brownlow; B. Fu	2016
Deployment of churn prediction model in financial services industry	2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)	S. H. Dolatabadi; F. Keynia	2016
E-commerce customer churn prediction based on improved SMOTE and AdaBoost	2016 13th International Conference on Service Systems and Service Management (ICSSSM)	K. B. Subramanya; A. Somani	2016

Table C.1: IEEEXplore Results (continued)

Methods for churn prediction in the pre-paid mobile telecommunications industry	2016 International Conference on Communications (COMM)	Z. Kasiran; Z. Ibrahim; M. S. Mohd Ribuan	2016
Not Too Late to Identify Potential Churners: Early Churn Prediction in Telecommunication Industry	2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT)	G. G. Sundarkumar; V. Ravi; V. Siddeshwar	2016
Pattern Recognition, Part 1 [Guest editors' introduction]	IEEE Intelligent Systems	A. Abbasi; R. Y. K. Lau; D. E. Brown	2016
Predicting Customer Lifetime Value through Data Mining Technique in a Direct Selling Company	2016 International Conference on Industrial Engineering, Management Science and Application (ICIMSA)	E. Gharavi; M. J. Tarokh	2016
Research of indicator system in customer churn prediction for telecom industry	2016 11th International Conference on Computer Science & Education (ICCSE)	Q. Yanfang; L. Chen	2016
Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction	2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), SocialCom, SustainCom, BDCloud-SocialCom-SustainCom	Li Peng; Yu Xiaoyang; Sun Boyu; Huang Jiuling	2016
A comparative study of customer churn prediction in telecom industry using ensemble based classifiers	2017 International Conference on Inventive Computing and Informatics (ICICI)	M. Óskarsdóttir; C. Bravo; W. Verbeke; C. Sarraute; B. Baesens; J. Vanthienen	2017
A review and analysis of churn prediction methods for customer retention in telecom industries	2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)	Ç. K. Kontacı; O. çetintürk; G. G. Polat; K. C. Özkısacık; A. A. Salah	2017
An approach for predicting employee churn by using data mining	2017 International Artificial Intelligence and Data Processing Symposium (IDAP)	A. Tiwari; R. Sam; S. Shaikh	2017

Table C.1: IEEEXplore Results (continued)

Analysis and prediction of churn customers for telecommunication industry	2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)	A. J. Petkovski; B. L. R. Stojkoska; K. V. Trivodaliev; S. A. Kalajdziski	2017
Churn prediction analysis using various clustering algorithms in KNIME analytics platform	2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)	R. Pagare; A. Khare	2017
Churn prediction model for effective gym customer retention	2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)	A. Idris; A. Khan	2017
Churn Prediction System for Telecom using Filter–Wrapper and Ensemble Classification	The Computer Journal	A. Backiel; Y. Verbinnen; B. Baesens; G. Claeskens	2017
Computer Simulation of Electronic Commerce Customer Churn Prediction Model Based on Web Data Mining	2017 International Conference on Smart Grid and Electrical Automation (ICSGEA)	R. Mahajan; S. Som	2017
Customer churn analysis : A case study on the telecommunication industry of Thailand	2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)	Yizhe Ge; Shan He; Jingyue Xiong; D. E. Brown	2017
Customer churn analysis for a software-as-a-service company	2017 Systems and Information Engineering Design Symposium (SIEDS)	K. Dahiya; S. Bhatia	2017
Customer churn prediction in an internet service provider	2017 IEEE International Conference on Big Data (Big Data)	M. Yıldız; S. Albayrak	2017
Customer Churn Prediction in the Online New Media Platform: A Case Study on Juzi Entertainment	2017 International Conference on Platform Technology and Service (PlatCon)	H. Y. Liao; K. Y. Chen; D. R. Liu; Y. L. Chiu	2017
Customer churn prediction model using data mining techniques	2017 13th International Computer Engineering Conference (ICENCO)	S. Cui; N. Ding	2017

Table C.1: IEEEXplore Results (continued)

Customer churn prediction using improved FCM algorithm	2017 3rd International Conference on Information Management (ICIM)	N. L. R. Machado; D. D. A. Ruiz	2017
Customer: A novel customer churn prediction method based on mobile application usage	2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)	G. Maji; S. Sen	2017
Designing of customer and employee churn prediction model based on data mining method and neural predictor	2017 2nd International Conference on Computer and Communication Systems (ICCCS)	A. Rodan; H. Faris	2017
Enhanced feature mining and classifier models to predict customer churn for an E-retailer	2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence	A. O. Akyuz; M. Uysal; B. A. Bulbul; M. O. Uysal	2017
Ensemble approach for time series analysis in demand forecasting: Ensemble learning	2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)	A. Idris; A. Khan	2017
Evaluating deep learning in chum prediction for everything-as-a-service in the cloud	2017 International Joint Conference on Neural Networks (IJCNN)	C. Rui; Q. Jia-yin; W. Mian	2017
Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm	2017 3rd International Conference on Science in Information Technology (ICSITech)	V. Effendy; Adiwijaya; Z. K. A. Baizal	2017
Improving QoE prediction in mobile video through machine learning	2017 8th International Conference on the Network of the Future (NOF)	T. H. Hsu; C. C. Chen; M. F. Chiang; K. W. Hsu; W. C. Peng	2017
Machine learning techniques applied to prepaid subscribers: Case study on the telecom industry of Morocco	2017 Intelligent Systems and Computer Vision (ISCV)	M. Ahmed; I. Siddiqi; H. Afzal; B. Khan	2017
MCS: Multiple classifier system to predict the churners in the telecom industry	2017 Intelligent Systems Conference (IntelliSys)	I. Brânduşoiu; G. Toderean; H. Beleiu	2017

Table C.1: IEEEXplore Results (continued)

OTT-ISP joint service management: A Customer Lifetime Value based approach	2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)	C. Wang; R. Li; P. Wang; Z. Chen	2017
Partition cost-sensitive CART based on customer value for Telecom customer churn prediction	2017 36th Chinese Control Conference (CCC)	C. L. Liu; B. Lovell; D. Tao; M. Tistarelli	2017
Predicting Customer Churn Using the Cumulative Quantity Control Chart	2017 International Conference on Industrial Engineering, Management Science and Application (ICIMSA)	Y. Xu	2017
Predicting QoE in cellular networks using machine learning and in-smartphone measurements	2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)	S. Motahari; T. Jung; H. Zang; K. Janakiraman; X. Y. Li; K. S. Hoo	2017
Predictive analytics in reverse supply chain management commodity life expectancy for quality engineering	2017 IEEE 19th Electronics Packaging Technology Conference (EPTC)	Z. Zhang; R. Wang; W. Zheng; S. Lan; D. Liang; H. Jin	2017
Reducing the risk of customer migration by using bigdata clustering algorithm	2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)	Q. Yihui; Z. Chiyu	2017
Research on E-commerce user churn prediction based on logistic regression	2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)	A. Hanif; N. Azhar	2017
Resolving Class Imbalance and Feature Selection in Customer Churn Dataset	2017 International Conference on Frontiers of Information Technology (FIT)	Pushpa; G. Shobha	2017

Table C.1: IEEEXplore Results (continued)

Twitter opinion mining predicts broadband internet's customer churn rate	2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)	M. Fiedler; K. D. Moor; H. Ravuri; P. Tanneedi; M. Chandiri	2017
Users on the Move: On Relationships Between QoE Ratings, Data Volumes and Intentions to Churn	2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops)	M. Ballings; D. V. d. Poel	2017
MCS: Multiple classifier system to predict the churners in the telecom industry	2017 Intelligent Systems Conference (IntelliSys)	M. Ahmed; I. Siddiqi; H. Afzal; B. Khan	2017
Resolving Class Imbalance and Feature Selection in Customer Churn Dataset	2017 International Conference on Frontiers of Information Technology (FIT)	A. Hanif; N. Azhar	2017
Predictive churn analysis with machine learning methods	2018 26th Signal Processing and Communications Applications Conference (SIU)	M. Günay; T. Ensarı	2018
An efficient hybrid clustering to predict the risk of customer churn	2018 2nd International Conference on Inventive Systems and Control (ICISC)	U. N. Dulhare; I. Ghori	2018
Just-in-time Customer Churn Prediction: With and Without Data Transformation	2018 IEEE Congress on Evolutionary Computation (CEC)	A. Amin; B. Shah; A. M. Khattak; T. Baker; H. u. Rahman Durani; S. Anwar	2018
Deep Neural Pipeline for Churn Prediction	2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet)	A. SIMION-CONSTANTINESCU; A. I. DAMIAN; N. ŢĂPUŞ; L. PICIU; A. PURDILĂ; B. DUMITRESCU	2018
Context Aware Telco Churn Prediction Powered By Temporal Feature Engineering	2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)	R. Bai; W. Rao; M. Yuan; J. Zeng; J. Yan	2018

Table C.1: IEEEXplore Results (continued)

Measuring YouTube QoE with ITU-T P.1203 Under Constrained Bandwidth Conditions	2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)	W. Robitza; D. G. Kittur; A. M. Dethof; S. Görin; B. Feiten; A. Raake	2018
HMM of Telecommunication Big Data for Consumer Churn Prediction	2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation	X. Xia; L. Zeng; R. Yu	2018
Profit Optimizing Churn Prediction for Long-term Loyal Customer in Online games	IEEE Transactions on Games	E. Lee; B. Kim; S. Kang; B. Kang; Y. Jang; H. K. Kim	2018
A Comparative Study of Employee Churn Prediction Model	2018 4th International Conference on Science and Technology (ICST)	A. Alamsyah; N. Salma	2018
A Novel Decision Tree Based on Profit Variance Maximization Criterion for Customer Churn Problem	2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)	X. Zhang; Z. Zhang; D. Liang; H. Jin	2018
Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning	2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)	S. Agrawal; A. Das; A. Gaikwad; S. Dhage	2018
Cost-Sensitive Semi-Supervised Ensemble Model for Customer Churn Prediction	2018 15th International Conference on Service Systems and Service Management (ICSSSM)	J. Xiao; L. Huang; L. Xie	2018
Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques	2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)	M. Ali; A. U. Rehman; S. Hafeez; D. M. U. Ashraf	2018

Table C.1: IEEEXplore Results

SpringerLink Results

Paper Title	Published in	Authors	Publication Year
3D coverage analysis of LTE urban heterogeneous networks with dense femtocell deployments	EURASIP Journal on Wireless Communications and Networking	Florian LetourneuxYoann CorreErwan SuteauYves Lostanlen	2012
A Graph-Based Churn Prediction Model for Mobile Telecom Networks	Advanced Data Mining and Applications	M. SaravananG. S. Vijay Raajaa	2012
ABML Knowledge Refinement Loop: A Case Study	Foundations of Intelligent Systems	Matej GuidMartin MožinaVida GroznikDejan GeorgievAleksander SadikovZvezdan PirtošekIvan Bratko	2012
An Integrated Model for Financial Data Mining	Multi-disciplinary Trends in Artificial Intelligence	Fan CaiN-A. LeKhacM-Tahar Kechadi	2012
Can File Level Characteristics Help Identify System Level Fault-Proneness?	Hardware and Software: Verification and Testing	Thomas J. OstrandElaine J. Weyuker	2012
Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm	Emerging Research in Artificial Intelligence and Computational Intelligence	Guangqing LiXiuqin Deng	2012
Customer relationship management and Web mining: the next frontier	Data Mining and Knowledge Discovery	Alexander Tuzhilin	2012
Data Mining for Churn Prediction: Multiple Regressions Approach	Computer Applications for Database, Education, and Ubiquitous Computing	Mohd Khalid AwangMohd Nordin Abdul RahmanMohammad Ridwan Ismail	2012
Data-intensive architecture for scientific knowledge discovery	Distributed and Parallel Databases	Malcolm AtkinsonChee Sun LiewMichelle GaleaPaul MartinAmrey KrauseAdrian MouatOscar CorchoDavid Snelling	2012

Table C.2: SpringerLink Results (continued)

Forecasting Consumer Interest in New Services Using Semantic-Aware Prediction Model: The Case of YouTube Clip Popularity	Agent and Multi-Agent Systems. Technologies and Applications	Luka VrdoljakVedran PodobnikGordan Jezic	2012
Improving Customer Churn Prediction by Data Augmentation Using Pictorial Stimulus-Choice Data	Management Intelligent Systems	Michel BallingsDirk Van den PoelEmmanuel Verhagen	2012
Integrating Heterogeneous Prediction Models in the Cloud	E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life	Hung-Chen ChenChih-Ping WeiYu-Cheng ChenCi-Wei Lan	2012
Peer Selection in P2P Service Overlays Using Geographical Location Criteria	Computational Science and Its Applications – ICCSA 2012	Adriano FioresePaulo SimõesFernando Boavida	2012
Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences	Advances in Data Analysis and Classification	Vera L. MiguéisDirk Van den PoelAna S. CamanhoJoão Falcão e Cunha	2012
Risk Prediction Framework and Model for Bank External Fund Attrition	Advances on Computational Intelligence	Hua LinGuangquan Zhang	2012
Special issue on data analysis and classification in marketing—preface by the guest editors	Advances in Data Analysis and Classification	Daniel BaierReinhold Decker	2012
The impact of social network-based segmentation on customer loyalty in the telecommunication industry	Journal of Database Marketing & Customer Strategy Management	Evangelos XevelonakisPatrick Som	2012
Using Social Network Classifiers for Predicting E-Commerce Adoption	E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life	Thomas VerbrakenFrank GoethalsWouter VerbekeBart Baesens	2012
A Brief History of Network Programmability and Related Fields	Network-Embedded Management and Applications	Ralf Wolter	2013

Table C.2: SpringerLink Results (continued)

A Computational Intelligence based Approach to Telecom Customer Classification for Value Added Services	Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)	Abhay BhadaniRavi ShankarD. Vijay Rao	2013
A Hybrid Machine Learning Method and Its Application in Municipal Waste Prediction	Advances in Data Mining. Applications and Theoretical Aspects	Emadoddin LivaniRaymond NguyenJörg DenzingerGünther RuheScott Banack	2013
A multi-objective optimization and decision algorithm for locator layout continuous searching in checking fixture design	The International Journal of Advanced Manufacturing Technology	Kun JiangXionghui ZhouMing LiXiao Kong	2013
A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models	Service Business	Gitae KimBongsug Kevin ChaeDavid L. Olson	2013
An Exploration of Customer-Centric Cloud Service Design	Design Science: Perspectives from Europe	Brian O'FlahertySimon WoodworthColm ThorntonYvonne O'Connor	2013
Building actions from classification rules	Knowledge and Information Systems	Ronan TréposAnsaf Salleb-AouissiMarie-Odile CordierVéronique MassonChantal Gascuel-Odoux	2013
Business Process Management (Process Life Cycle, Process Maturity)	Business Process Reengineering	Sanjay Mohapatra	2013
CDR Analysis Based Telco Churn Prediction and Customer Behavior Insights: A Case Study	Web Information Systems Engineering – WISE 2013	Natwar ModaniKuntal DeyRitesh GuptaShantanu Godbole	2013

Table C.2: SpringerLink Results (continued)

Comparison of Multi-objective Algorithms Applied to Feature Selection	Towards Advanced Data Analysis by Combining Soft Computing and Statistics	Özlem TürkşenSusana M. Vieira-José F. A. MadeiraAyşen Apaydin-João M. C. Sousa	2013
CrimeFighter Investigator: Criminal Network Sense-Making	Handbook of Computational Approaches to Counterterrorism	Rasmus Rosenqvist PetersenUffe Kock Wiil	2013
Data Mining Using Parallel Multi-objective Evolutionary Algorithms on Graphics Processing Units	Massively Parallel Evolutionary Computation on GPGPUs	Man Leung WongGeng Cui	2013
Data Warehousing	Encyclopedia of Operations Research and Management Science	Paul Gray	2013
Decision trees: a recent overview	Artificial Intelligence Review	S. B. Kotsiantis	2013
Experience, socialization and customer retention: Lessons from the dance floor	Marketing Letters	Gianna GiudicatiMassimo RiccaboniAnna Romiti	2013
From Packets to People: Quality of Experience as a New Measurement Challenge	Data Traffic Monitoring and Analysis	Raimund SchatzTobias HoßfeldLucjan JanowskiSebastian Egger	2013
Game Data Mining	Game Analytics	Anders Drachen Ph.D.Christian Thurau Ph.D.Julian Togelius Ph.D.Georgios N. Yannakakis Ph.D.Christian Bauckhage Ph.D.	2013
How Long Will She Call Me? Distribution, Social Theory and Duration Prediction	Machine Learning and Knowledge Discovery in Databases	Yuxiao DongJie TangTiancheng LouBin WuNitesh V. Chawla	2013
Improved Feature Selection Based on Particle Swarm Optimization for Liver Disease Diagnosis	Swarm, Evolutionary, and Memetic Computing	Gunasundari SelvarajJanakiraman S.	2013
Improving the Quality of Software by Quantifying the Code Change Metric and Predicting the Bugs	Computational Science and Its Applications – ICCSA 2013	V. B. SinghK. K. Chaturvedi	2013

Table C.2: SpringerLink Results (continued)

Improving Tree-Based Classification Rules Using a Particle Swarm Optimization	Advances in Production Management Systems. Competitive Manufacturing for Innovative Products and Services	Chi-Hyuck JunYun-Ju ChoHyeseon Lee	2013
Influence of confirmation biases of developers on software quality: an empirical study	Software Quality Journal	Gül ÇalıklıAyşe Başar Bener	2013
Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification	Applied Intelligence	Adnan IdrisAsifullah KhanYeon Soo Lee	2013
Neural Approaches to Economic Modeling	Economic Modeling Using Artificial Intelligence Methods	Tshilidzi Marwala	2013
Preventing Churn in Telecommunications: The Forgotten Network	Advances in Intelligent Data Analysis XII	Dejan RadosavljevikPeter van der Putten	2013
Research on Current Situation Analysis and Solving Strategies of Large Inflows and Large Outflows of Telecom Customers	The 19th International Conference on Industrial Engineering and Engineering Management	Yang GaoQun Yuan	2013
Studying the impact of social interactions on software quality	Empirical Software Engineering	Nicolas BettenburgAhmed E. Hassan	2013
The limited impact of individual developer data on software defect prediction	Empirical Software Engineering	Robert M. BellThomas J. OstrandElaine J. Weyuker	2013
Trade-Offs in Social and Behavioral Modeling in Mobile Networks	Social Computing, Behavioral-Cultural Modeling and Prediction	Yaniv AltshulerMichael FireNadav AharonyZeev VolkovichYuval EloviciAlex (Sandy) Pentland	2013

Table C.2: SpringerLink Results (continued)

A Balanced Transfer Learning Model for Customer Churn Prediction	Proceedings of the Eighth International Conference on Management Science and Engineering Management	Bing ZhuJin XiaoChangzheng He	2014
A data mining approach for segmentation-based importance-performance analysis (SOM–BPNN–IPA): a new framework for developing customer retention strategies	Service Business	Seyed Yaghoub HosseiniAlireza Ziaei Bideh	2014
A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry	Computational Collective Intelligence. Technologies and Applications	Hossam FarisBashar Al-ShboulNazeeh Ghatasheh	2014
A New Method of Identifying Individuals' Roles in Mobile Telecom Subscriber Data for Improved Group Recommendations	Multidisciplinary Social Networks Research	Saravanan MohanManisha Subramanian	2014
An Analysis of Power Trading Agent Competition 2014	Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets	Jurica BabicVedran Podobnik	2014
Customer Analyst for the Telecom Industry	Large-Scale Data Analytics	David KonopnickiMichal Shmueli-Scheuer	2014
Customer Churn Prediction in Telecommunication Industry: With and without Counter-Example	Nature-Inspired Computation and Machine Learning	Adnan AminChangez KhanImtiaz AliSajid Anwar	2014

Table C.2: SpringerLink Results (continued)

Customer Lifetime Value and Defection Possibility Prediction Model Using Machine Learning: An Application to a Cloud-Based Software Company	Intelligent Information and Database Systems	Niken PrasastiMasato OkadaKatsutoshi KanamoriHayato Ohwada	2014
From Sample to Population	Engaging Customers Using Big Data	Arvind Sathi	2014
Mining Telecommunication Networks to Enhance Customer Lifetime Predictions	Artificial Intelligence and Soft Computing	Aimée BackielBart BaesensGerda Claeskens	2014
Modeling and predicting cascading removal phenomenon over social networks	Social Network Analysis and Mining	Amir H. RazaviDyah AnggrainiRokia MissaouiJean VaillancourtMohamed Talbi	2014
No-reference image and video quality assessment: a classification and review of recent approaches	EURASIP Journal on Image and Video Processing	Muhammad ShahidAndreas RossholmBenny LövströmHans-Jürgen Zepernick	2014
One-Step Classifier Ensemble Model for Customer Churn Prediction with Imbalanced Class	Proceedings of the Eighth International Conference on Management Science and Engineering Management	Jin XiaoGeer TengChangzheng HeBing Zhu	2014
Predicting Online Community Churners Using Gaussian Sequences	Social Informatics	Matthew Rowe	2014
Predictive Policing	Encyclopedia of Criminology and Criminal Justice	Craig D. Uchida	2014
Proactive Data Mining Using Decision Trees	Proactive Data Mining with Decision Trees	Haim DahanShahar CohenLior RokachOded Maimon	2014
Proactive Data Mining: A General Approach and Algorithmic Framework	Proactive Data Mining with Decision Trees	Haim DahanShahar CohenLior RokachOded Maimon	2014

Table C.2: SpringerLink Results (continued)

RFID-based intelligent shopping environment: a comprehensive evaluation framework with neural computing approach	Neural Computing and Applications	Chia-Chen Chen	2014
Software defect prediction using Bayesian networks	Empirical Software Engineering	Ahmet OkutanOlcay Taner Yıldız	2014
Software Product Quality Models	Evolving Software Systems	Rudolf FerencPéter HegedűsTibor Gyimóthy	2014
Temporal Development of Quality of Experience	Quality of Experience	Benjamin WeissDennis GuseSebastian MöllerAlexander RaakeAdam BorowiakUlrich Reiter	2014
Training and assessing classification rules with imbalanced data	Data Mining and Knowledge Discovery	Giovanna MenardiNicola Torelli	2014
Utilizing Customers' Purchase and Contract Renewal Details to Predict Defection in the Cloud Software Industry	Knowledge Management and Acquisition for Smart Systems and Services	Niken Prasasti MartonoKatsutoshi KanamoriHayato Ohwada	2014
Visualization of trends in subscriber attributes of communities on mobile telecommunications networks	Social Network Analysis and Mining	Daniel ArchambaultNeil Hurley	2014
A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction	New Contributions in Information Systems and Technologies	Adnan AminFaisal RahimImtiaz AliChangez KhanSajid Anwar	2015
A diffusion model for churn prediction based on sociometric theory	Advances in Data Analysis and Classification	Uroš DroftinaMitja ŠtularAndrej Košir	2015
A Fuzzy-Based Approach to Survival Data Mining	Fifty Years of Fuzzy Logic and its Applications	Mark LastHezi Halpert	2015

Table C.2: SpringerLink Results (continued)

A Prudent Based Approach for Customer Churn Prediction	Beyond Databases, Architectures and Structures	Adnan AminFaisal RahimMuhammad RamzanSajid Anwar	2015
A Sensitivity Analysis of Multi-objective Cooperative Planning Optimization Using NSGA-II	Multiphysics Modelling and Simulation for Systems Design and Monitoring	Wafa Ben YahiaOmar AyadiFaouzi Masmoudi	2015
A survey of results on mobile phone datasets analysis	EPJ Data Science	Vincent D BlondelAdeline DecuyperGautier Krings	2015
Application of Function Points and Data Mining Techniques for Software Estimation - A Combined Approach	Software Measurement	Przemysław PospiesznyBeata Czarnacka-ChrobotAndrzej Kobyliński	2015
Applications of Big Data	Big Data	Hareesh Boinepelli	2015
Churn Prediction in Telecommunication Industry Using Rough Set Approach	New Trends in Computational Collective Intelligence	Adnan AminSaeed Shehzad-Changez KhanImtiaz AliSajid Anwar	2015
Consuming and Publishing Models on Azure Marketplace	Predictive Analytics with Microsoft Azure Machine Learning	Roger BargaValentine FontamaWee Hyong Tok	2015
Empirical analysis of factors affecting confirmation bias levels of software engineers	Software Quality Journal	Gul CalikliAyse Bener	2015
Enhancing Telco Service Quality with Big Data Enabled Churn Analysis: Infrastructure, Model, and Deployment	Journal of Computer Science and Technology	Hui LiDi WuGao-Xiang LiYi-Hao KeWen-Jie LiuYuan-Huan ZhengXiao-La Lin	2015
Examples of Database Marketing Models	Modeling Markets	Peter S. H. LeeflangJaap E. WieringaTammo H. A. BijmoltKoen H. Pauwels	2015
Feature-selection-based dynamic transfer ensemble model for customer churn prediction	Knowledge and Information Systems	Jin XiaoYi XiaoAnqiang Huang-Dunhu LiuShouyang Wang	2015

Table C.2: SpringerLink Results (continued)

From research to purchase: an empirical analysis of research-shopping behaviour in the insurance sector	Zeitschrift für die gesamte Versicherungswissenschaft	Stefan MauIrena Pletikosa CvijikjJoël Wagner	2015
Implementation of Two Stages k-Means Algorithm to Apply a Payment System Provider Framework in Banking Systems	Artificial Intelligence Perspectives and Applications	Omid Mahdi Ebadati E.Sara Sadat Babaie	2015
Innovation That Matters - The IBM Research Way	Innovative Technologies in Management and Science	Oded Cohn	2015
Intelligent Classification-Based Methods in Customer Profitability Modeling	Intelligent Techniques in Engineering Management	Yeliz EkinciEkrem Duman	2015
Introduction	Real World Data Mining Applications	Mahmoud Abou-NasrStefan LessmannRobert StahlbockGary M. Weiss	2015
Introduction to the special issue on “Data analytics for marketing intelligence”	Information Systems and e-Business Management	Ching-Chin ChernAnthony J. T. LeeChih-Ping Wei	2015
Mapping and Usage of Know-How Contributions	Information Systems Engineering in Complex Environments	Arnon SturmDaniel GrossJian WangSoroosh NalchigarEric Yu	2015
Predicting customer churn from valuable B2B customers in the logistics industry: a case study	Information Systems and e-Business Management	Kuanchin ChenYa-Han HuYi-Cheng Hsieh	2015
Predicting Defect Prone Modules in Web Applications	Information and Software Technologies	Mehmet Serdar BiçerBanu Diri	2015
Predicting defective modules in different test phases	Software Quality Journal	Bora CaglayanAyse Tosun Misirli-Ayse Basar BenerAndriy Miranskyy	2015

Table C.2: SpringerLink Results (continued)

Preventing Customers from Running Away! Exploring Generalized Additive Models for Customer Churn Prediction	The Sustainable Global Marketplace	Kristof CoussementDries Frederik BenoitDirk Van den Poel	2015
Segmentation: Clustering and Classification	R for Marketing Research and Analytics	Chris Chapman Elea McDonnell Feit	2015
Sentiment Analysis in Transcribed Utterances	Advances in Knowledge Discovery and Data Mining	Nir Ofek Gilad Katz Bracha Shapira Yedidya Bar-Zev	2015
Software Mining Studies: Goals, Approaches, Artifacts, and Replicability	Software Engineering	Sven Amann Stefanie Beyer Katja Kevic Harald Gall	2015
Studying the relationship between logging characteristics and the code quality of platform software	Empirical Software Engineering	Weiyi Shang Meiyappan Nagappan Ahmed E. Hassan	2015
Towards Cognitive BPM as the Next Generation BPM Platform for Analytics-Driven Business Processes	Business Process Management Workshops	Hamid R. Motahari Nezhad Rama Akkiraju	2015
Towards improving statistical modeling of software engineering data: think locally, act globally!	Empirical Software Engineering	Nicolas Bettenburg Meiyappan Nagappan Ahmed E. Hassan	2015
A Feature Extraction Method Based on Stacked Auto-Encoder for Telecom Churn Prediction	Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems	Ruiqi Li Peng Wang Zonghai Chen	2016
A Fuzzy Rule-Based Learning Algorithm for Customer Churn Prediction	Advances in Data Mining. Applications and Theoretical Aspects	Bingquan Huang Ying Huang Chongcheng ChenM. -T. Kechadi	2016

Table C.2: SpringerLink Results (continued)

A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems	Computing	Abdul Hameed Alireza Khoshkbarforoushha Rajiv RanjanPrem Prakash Jayaraman Joanna KolodziejPavan Balaji Sherali Zeadally Qutaibah Marwan Malluhi Nikos Tziritas Abhinav Vishnu Samee U. Khan Albert Zomaya	2016
An Investigation on Online Versus Batch Learning in Predicting User Behaviour	Research and Development in Intelligent Systems XXXIII	Nikolay Burlutskiy Miltos Petridis Andrew FishAlexey ChernovNour Ali	2016
Artificial immune network with feature selection for bank term deposit recommendation	Journal of Intelligent Information Systems	Xiao-Yong LuXiao-Qiang Chu Meng-Hui ChenPei-Chann ChangShih-Hsin Chen	2016
Artificial Neural Network	Intelligent Techniques for Data Science	Rajendra AkerkarPriti Srinivas Sajja	2016
Assessment and cross-product prediction of software product line quality: accounting for reuse across products, over multiple releases	Automated Software Engineering	Thomas DevineKaterina Goseva-Popstojanova Sandeep Krishnan Robyn R. Lutz	2016
Automatic Discovery and Recommendation for Telecommunication Package Using Particle Swarm Optimization	Advances in Nature and Biologically Inspired Computing	Shanshan LiuBo YangLin WangAjith Abraham	2016
Classification Methods	Data Science Using Oracle Data Miner and Oracle R Enterprise	Sibanjan Das	2016
Classification Models	Data Mining with SPSS Modeler	Tilo Wendler Sören Gröttrup	2016
Cognitive Things in an Organization	Cognitive (Internet of) Things	Arvind Sathi	2016
Customer Churn Models: A Comparison of Probability and Data Mining Approaches	Looking Forward, Looking Back: Drawing on the Past to Shape the Future of Marketing	Ali Tamaddoni JahromiStanislav StakhovychMichael Ewing	2016

Table C.2: SpringerLink Results (continued)

Data mining for software engineering and humans in the loop	Progress in Artificial Intelligence	Leandro L. MinkuEmilia MendesBurak Turhan	2016
Developing a prediction model for customer churn from electronic banking services using data mining	Financial Innovation	Abbas KeramatiHajar GhaneeiSeyed Mohammad Mirmohammadi	2016
Editor's introduction	Financial Innovation	Gang Kou	2016
Final Remarks on Big Data Analysis and Its Impact on Society and Science	Big Data Analysis: New Algorithms for a New Society	Jerzy StefanowskiNathalie Japkowicz	2016
Improvement of a Web Browser Game Through the Knowledge Extracted from Player Behavior	Knowledge, Information and Creativity Support Systems: Recent Trends, Advances and Solutions	João AlvesJosé NevesSascha LangeMartin Riedmiller	2016
Improving Behavior Prediction Accuracy by Using Machine Learning for Agent-Based Simulation	Intelligent Information and Database Systems	Shinji HayashiNiken PrasastiKatsutoshi KanamoriHayato Ohwada	2016
Index to volume 67, 2016	Journal of the Operational Research Society		2016
Introduction	Data Mining with SPSS Modeler	Tilo WendlerSören Gröttrup	2016
Introduction to Data Science	Intelligent Techniques for Data Science	Rajendra AkerkarPriti Srinivas Sajja	2016
Making Digital Freemium Business Models a Success: Predicting Customers' Lifetime Value via Initial Purchase Information	Business & Information Systems Engineering	Dipl.-Wirtsch.-Inform. Sebastian VoigtProf. Dr. Oliver Hinz	2016
Mechanical Reliability	Reliability and Safety Engineering	Ajit Kumar VermaSrividya AjitDurga Rao Karanki	2016
Mining Customer Behavior in Trial Period of a Web Application Usage—Case Study	Artificial Intelligence Perspectives in Intelligent Systems	Goran MatoševićVanja Bevanda	2016

Table C.2: SpringerLink Results (continued)

Optimization of welding process parameters by combining Kriging surrogate with particle swarm optimization algorithm	The International Journal of Advanced Manufacturing Technology	Ping JiangLongchao CaoQi ZhouZhongmei GaoYoumin RongXinyu Shao	2016
Predicting Probability of Customer Churn in Insurance	Modeling and Simulation in Engineering, Economics and Management	Catalina BolancéMontserrat GuillenAlemar E. Padilla-Barreto	2016
Predicting Replacement of Smartphones with Mobile App Usage	Web Information Systems Engineering – WISE 2016	Dun YangZhiang WuXiaopeng WangJie CaoGuandong Xu	2016
Predicting Telecommunication Customer Churn Using Data Mining Techniques	Internet and Distributed Computing Systems	Diana AlOmariMohammad Mehedi Hassan	2016
Predicting time-to-churn of prepaid mobile telephone customers using social network analysis	Journal of the Operational Research Society	Aimée BackielBart BaesensGerda Claeskens	2016
Preprocessor-based variability in open-source and industrial software systems: An empirical study	Empirical Software Engineering	Claus HunsenBo ZhangJanet SiegmundChristian KästnerOlaf LeßenichMartin BeckerSven Apel	2016
Proximal gradient method for huberized support vector machine	Pattern Analysis and Applications	Yangyang XuIoannis AkrotirianakisAmit Chakraborty	2016
Quantifying structural attributes of system decompositions in 28 feature-oriented software product lines	Empirical Software Engineering	Stefan SobernigSven ApelSergiy KolesnikovNorbert Siegmund	2016
Social Network Data and Predictive Mining (Business Intelligence 2)	Social Media Management	Amy Van Looy	2016
Studying high impact fix-inducing changes	Empirical Software Engineering	Ayse Tosun MisirliEmad ShihabYasukata Kamei	2016

Table C.2: SpringerLink Results (continued)

A churn prediction model for prepaid customers in telecom using fuzzy classifiers	Telecommunication Systems	Muhammad AzeemMuhammad UsmanA. C. M. Fong	2017
A New Customer Churn Prediction Approach Based on Soft Set Ensemble Pruning	Recent Advances on Soft Computing and Data Mining	Mohd Khalid AwangMokhairi MakhtarMohd Nordin Abd RahmanMustafa Mat Deris	2017
A new transferred feature selection algorithm for customer identification	Neural Computing and Applications	Bing ZhuYongge NiuJin XiaoBart Baesens	2017
A study examining relationships between micro patterns and security vulnerabilities	Software Quality Journal	Kazi Zakia SultanaByron J. WilliamsTanmay Bhowmik	2017
Advanced Analytics Using Machine Learning and R	Business in Real-Time Using Azure IoT and Cortana Intelligence Suite	Bob FamiliarJeff Barnes	2017
An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing	Cluster Computing	J. VijayaE. Sivasankar	2017
Applicability and effectiveness of classifications models for achieving the twin objectives of growth and outreach of microfinance institutions	Computational and Mathematical Organization Theory	Manojit ChattopadhyaySubrata Kumar Mitra	2017
Benchmarking sampling techniques for imbalance learning in churn prediction	Journal of the Operational Research Society	Bing ZhuBart BaesensAimée BackielSeppe K. L. M. vanden Broucke	2017
Big Data Analytics and Visualization: Finance	Big Data and Visual Analytics	Shyam PrabhakarLarry Maves	2017

Table C.2: SpringerLink Results (continued)

Big Data Architecture for Predicting Churn Risk in Mobile Phone Companies	Information Management and Big Data	Alonso Raul Melgarejo GalvanKaterine Rocio Clavo Navarro	2017
Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention	Neural Computing and Applications	Anil KumarGaurav KabraEswara Krishna MussadaManoj Kumar Dash-Prashant Singh Rana	2017
Cost-sensitive Association Rule Modeling for Predicting Sequential Event	Proceedings of the 23rd International Conference on Industrial Engineering and Engineering Management 2016	Hai-yan YUMeng-li YANGJie JIAN	2017
Credit Card Applications Pending – Who Are Our Best Prospect Cardholders? Improved Decisions through Business Analytics and Business Intelligence	Journal of Information Technology Teaching Cases	Jongsawas Chongwatpol	2017
Customer Segmentation by Various Clustering Approaches and Building an Effective Hybrid Learning System on Churn Prediction Dataset	Computational Intelligence in Data Mining	E. SivasankarJ. Vijaya	2017
Der Einsatz von Text Mining zur Bestimmung des Diffusionsprozesses von Produkten	Moderne Methoden der Marktforschung	Dr. Stefan EbenerMonika Ebener	2017
Do bugs foreshadow vulnerabilities? An in-depth study of the chromium project	Empirical Software Engineering	Nuthan MunaiahFelivel CamiloWesley WighamAndrew Meneely-Meiyappan Nagappan	2017
Elastic Optical Networking for 5G Transport	Journal of Network and Systems Management	Raouf BoutabaNashid ShahriarSiavash Fathi	2017

Table C.2: SpringerLink Results (continued)

Handling Imbalanced Data in Churn Prediction Using RUSBoost and Feature Selection (Case Study: PT.Telekomunikasi Indonesia Regional 7)	Recent Advances on Soft Computing and Data Mining	Erna DwiyantiAdiwijayaArie Ardiyanti	2017
High Accuracy Predictive Modelling for Customer Churn Prediction in Telecom Industry	Machine Learning and Data Mining in Pattern Recognition	R. PrashanthK. DeepakAmit Kumar Meher	2017
Identifying Prime Customers Based on MobileUsage Patterns	Advances on P2P, Parallel, Grid, Cloud and Internet Computing	Deepali AroraKin Fun Li	2017
Improved credit card churn prediction based on rough clustering and supervised learning techniques	Cluster Computing	R. RajamohamedJ. Manokaran	2017
Improving Financial Distress Prediction Using Financial Network-Based Information and GA-Based Gradient Boosting Method	Computational Economics	Jiaming LiuChong WuYongli Li	2017
Industrial Applications of Machine Learning	Machine Learning for Decision Makers	Patanjali Kashyap	2017
Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling	Cluster Computing	Adnan IdrisAksam IftikharZia ur Rehman	2017
Intelligent data analysis approaches to churn as a business problem: a survey	Knowledge and Information Systems	David L. GarcíaÀngela NebotAlfredo Vellido	2017
Just-in-time customer churn prediction in the telecommunication sector	The Journal of Supercomputing	Adnan AminFeras Al-ObeidatBabar ShahMay Al TaeChangez Khan-Hamood Ur Rehman DurraniSajid Anwar	2017

Table C.2: SpringerLink Results (continued)

K- local maximum margin feature extraction algorithm for churn prediction in telecom	Cluster Computing	Long ZhaoQian GaoXiangJun DongAimei DongXue Dong	2017
More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services	Quality and User Experience	A. SacklR. SchatzA. Raake	2017
Multi Criteria Decision Making in Financial Risk Management with a Multi-objective Genetic Algorithm	Computational Economics	Sujatha SrinivasanT. Kamalakannan	2017
Multi-episodic perceived quality for one session of consecutive usage episodes with a speech telephony service	Quality and User Experience	Dennis GuseBenjamin WeissFrank HaaseAnna WunderlichSebastian Möller	2017
Performance prediction of parallel computing models to analyze cloud-based big data applications	Cluster Computing	Chao ShenWeiqin TongKim-Kwang Raymond ChooSamina Kausar	2017
Predicting Customer Churn: Customer Behavior Forecasting for Subscription-Based Organizations	Information Systems	Leonidas KatelarisMarinos Themistocleous	2017
Predicting direct marketing response in banking: comparison of class imbalance methods	Service Business	Vera L. MiguéisAna S. CamanhoJosé Borges	2017
Predictive Analytics	Machine Learning and Cognition in Enterprises	Rohit Kumar	2017
Predictive Computing and Information Security: A Technical Review	Predictive Computing and Information Security	P. K. GuptaVipin TyagiS. K. Singh	2017

Table C.2: SpringerLink Results (continued)

RACER: accurate and efficient classification based on rule aggregation approach	Neural Computing and Applications	Javad BasiriFattaneh TaghiyarehHeshaam Faili	2017
Real-Time Analytics with Storm	Processing Big Data with Azure HDInsight	Vinit Yadav	2017
Research on Modeling Customer Churn in Video Players	Proceedings of the Fourth International Forum on Decision Sciences	Qixing QuXiaoyue LiuLi ZhangLin Wang	2017
Run-Time Models for Online Performance and Resource Management in Data Centers	Self-Aware Computing Systems	Simon SpinnerAntonio FilieriSamuel KounevMartina MaggioAnders Robertsson	2017
Temporal patterns behind the strength of persistent ties	EPJ Data Science	Henry NavarroGiovanna MiritelloArturo CanalesEsteban Moro	2017
THE “NUTS AND BOLTS” OF BEHAVIORAL INTERVENTION DEVELOPMENT: STUDY DESIGNS, METHODS AND FUNDING OPPORTUNITIES	Annals of Behavioral Medicine		2017
Time for Addressing Software Security Issues: Prediction Models and Impacting Factors	Data Science and Engineering	Lotfi Ben OthmaneGolriz ChehraziEric BoddenPetar TsalovskiAchim D. Brucker	2017
Towards a Large Scale Practical Churn Model for Prepaid Mobile Markets	Advances in Data Mining. Applications and Theoretical Aspects	Amit Kumar MeherJobin WilsonR. Prashanth	2017
Towards an Efficient Method of Modeling “Next Best Action” for Digital Buyer’s Journey in B2B	Machine Learning and Data Mining in Pattern Recognition	Anit BhandariKiran RamaNandini SethNishant NiranjanParag ChitaliaStig Berg	2017
Towards Distributed Cognitive Expert Systems	Designing the Digital Transformation	Schahin TofangchiAndre HaneltLutz M. Kolbe	2017

Table C.2: SpringerLink Results (continued)

Using association rules to assess purchase probability in online stores	Information Systems and e-Business Management	Grażyna SuchackaGrzegorz Chodak	2017
A dimensionality reduction-based efficient software fault prediction using Fisher linear discriminant analysis (FLDA)	The Journal of Supercomputing	Anum KalsoomMuazzam Maqsood-Mustansar Ali GhazanfarFarhan AadilSeungmin Rho	2018
A model for the mobile market based on customers profile to analyze the churning process	Wireless Networks	Mario Rogelio Flores-MéndezMarcos Postigo-BoixJosé Luis Melús-MorenoBurkhard Stiller	2018
ABC Based Neural Network Approach for Churn Prediction in Telecommunication Sector	Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2	Priyanka PaliwalDivya Kumar	2018
An Efficient Software Defect Analysis Using Correlation-Based Oversampling	Arabian Journal for Science and Engineering	K. Nitalaksheswara RaoCh. Satyananda Reddy	2018
Big Data Analytics: Applications, Prospects and Challenges	Mobile Big Data	Konstantinos VassakisEmmanuel PetrakisIoannis Kopanakis	2018
Business Analytics Capabilities and Use: A Value Chain Perspective	Analytics and Data Science	Torupallab GhoshalRudolph T. BedeleyLakshmi S. IyerJoyendu Bhadury	2018
Client Churn Prediction with Call Log Analysis	Database Systems for Advanced Applications	Nhi N. Y. VoShaowu LiuJames BrownlowCharles ChuBen Culbert-Guandong Xu	2018
Closing the Gap Between Experts and Novices Using Analytics-as-a-Service: An Experimental Study	Business & Information Systems Engineering	Jasmien LismontTine Van Calster-María ÓskarsdóttirSeppe vanden BrouckeBart BaesensWilfried LemahieuJan Vanthienen	2018

Table C.2: SpringerLink Results (continued)

Comparative Study of Different Data Mining Techniques in Predicting Forest Fire in Lebanon and Mediterranean	Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016	Nizar HamadehAli KarouniBassam DayaPierre Chauvet	2018
Cost-Sensitive Churn Prediction in Fund Management Services	Database Systems for Advanced Applications	James BrownlowCharles ChuBin FuGuandong XuBen CulbertQinxue Meng	2018
Customer Churn Analysis for Telecom Operators Based on SVM	Signal and Information Processing, Networking and Computers	Runsha DongFei SuShan YangXinzhou ChengWeiwei Chen	2018
Customer Churn Prediction in Superannuation: A Sequential Pattern Mining Approach	Databases Theory and Applications	Ben CulbertBin FuJames BrownlowCharles ChuQinxue MengGuandong Xu	2018
Effective Customer Relationship Management at ATB Financial: A Case Study on Industry-Academia Collaboration in Data Analytics	Highlighting the Importance of Big Data Management and Analysis for Various Applications	Ian Hargreaves Dylan Roth Muhammad Rezaul KarimMaleknaz NayebiGünther Ruhe	2018
Enhanced Prediction Model for Customer Churn in Telecommunication Using EMOTE	International Conference on Intelligent Computing and Applications	S. BabuN. R. Ananthanarayanan	2018
Future of CRM	Customer Relationship Management	V. KumarWerner Reinartz	2018
Handling Imbalanced Data: A Survey	International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications	Neelam Rout Debahuti Mishra Manas Kumar Mallick	2018
How Banks Can Better Serve Their Customers Through Artificial Techniques	Digital Marketplaces Unleashed	Armando Vieira Attul Sehgal	2018

Table C.2: SpringerLink Results (continued)

Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network	Neural Computing and Applications	E. SivasankarJ. Vijaya	2018
Improved Churn Prediction Based on Supervised and Unsupervised Hybrid Data Mining System	Information and Communication Technology for Sustainable Development	J. Vijaya E. Sivasankar	2018
In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions	Customer Needs and Solutions	Eva Ascarza Scott A. Neslin Oded Netzer Zachery Anderson Peter S. FaderSunil GuptaBruce G. S. HardieAurélie Lemmens Barak Libai David NealFoster ProvostRom Schrift	2018
Machine Learning Basics	Practical Machine Learning with Python	Dipanjan SarkarRaghav BaliTushar Sharma	2018
Particle classification optimization-based BP network for telecommunication customer churn prediction	Neural Computing and Applications	Ruiyun YuXuanmiao AnBo JinJia ShiOguti Ann MoveYonghe Liu	2018
Prediction of Customer Satisfaction Using Naive Bayes, MultiClass Classifier, K-Star and IBK	Soft Computing Applications	Sanjiban Sekhar RoyDeeksha KaulReetika RoyCornel Barna-Suhasini MehtaAnusha Misra	2018
Privacy-preserving condition-based forecasting using machine learning	Journal of Business Economics	Fabian TaigelAnselme K. TuenoRichard Pibernik	2018
Sleeping Customer Detection Using Support Vector Machine	Frontier Computing	Tsun KuPin-Liang ChenPing-Che Yang	2018
The Price of Privacy	Business & Information Systems Engineering	Annika BaumannJohannes Haupt-Fabian GebertStefan Lessmann	2018

Table C.2: SpringerLink Results (continued)

Towards improving decision making and estimating the value of decisions in value-based software engineering: the VALUE framework	Software Quality Journal	Emilia MendesPilar RodriguezVitor FreitasSimon BakerMohamed Amine Atoui	2018

Table C.2: SpringerLink Results

Google Scholar Results			
Paper Title	**Published in**	**Authors**	**Publication Year**
A classification algorithm of detect prediction for software modules based on fuzzy support vector machine	Journal of Nanjing University (Natural Sciences)	G Li-Na, Y Yang	2012
A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data	European Journal of operational research	ZY Chen, ZP Fan, M Sun	2012
A social network-based inference model for validating customer profile data	MIS quarterly	SH Park, SY Huh, W Oh, SP Han	2012
A Study on Prediction of Telecom Customer Churn Based on Dynamic Selection, Optimization and Integration of Cost Sensitivity	Chinese Journal of Management	L Guoen	2012
Airlines Marketing Analysis Based on Customer Churn Prediction	Proceedings of the 2012 3rd International Conference	G Feng	2012
An application of the CORER classifier on customer churn prediction	Telecommunications (IST), 2012 6th	J Basiri, F Taghiyareh	2012
Analyzing the Relationship between Cardiac and Cerebral Vascular Diseases and Meteorological Conditions [J]	Hebei Medicine	Z Dahui	2012
Application of business intelligence techniques in retail promotion optimization	Journal of Shandong University of Technology (Natural	J BI, H WEI	2012
Churn Prediction in Telecommunication by Boosting Regression Models	Analysis and Modeling	E Della Valle, R Lombardo, B Masiello, S Kon,	2012

Table C.3: Google Scholar Results (continued)

Customer Churn Prediction for Telecom Services	Computer Software	U Yabas, HC Cankaya, T Ince	2012
Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers	Multitopic Conference (INMIC), 2012 15th	A Idris, A Khan	2012
Customer churn prediction in commercial bank	Information, An In-ternational Interdisci-plinary Journal	G Wang, Y Peng, NE Yaw	2012
Customer churn prediction in telecommunications	Expert Systems with Applications	B Huang, MT Kechadi, B Buckley	2012
Customer churn prediction of china telecom based on cluster analysis and decision tree algorithm	Emerging research in artificial intelligence and	G Li, X Deng	2012
Customer churn prediction using back propagation neural network with evolution strategies		A Luthfiarta	2012
Customer event history for churn prediction: How long is long enough?	Expert Systems with Applications	M Ballings, D Van den Poel	2012
De Bock, Dirk Van den Poel, Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized	Expert Systems with Applications	W Koen	2012
Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction: Case Study Insurance Industry	International Conference on Information and Computer	RA Soeini, KV Rodpysh	2012

Table C.3: Google Scholar Results (continued)

Graph classification algorithm based on cost sensitivity [J]	Journal of Fuzhou University (Natural Science	G XIAO, X CHEN	2012
Impact of social attributes on Predictive Analytics in telecommunication industry	Multitopic Conference (INMIC), 2012	OZ Hashmi, S Sheikh	2012
Improved Support Vector Machine Algorithm for Fraudulent Financial Statements Detection	Management Review	K Baokui, L Zhixin, S Xiaodong, Y Zhong	2012
Improving customer churn prediction by data augmentation using pictorial stimulus-choice data	Management intelligent systems	M Ballings, D Van den Poel, E Verhagen	2012
New insights into churn prediction in the telecommunication sector: A profit driven data mining approach	European Journal of	W Verbeke, K Dejaeger, D Martens, J Hur	2012
Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models	Expert Systems with Applications	KW De Bock, D Van den Poel	2012
Research on cloud-based data mining and its prospect [J]	Computer and modernization	L DENG, H YU	2012
Research on Customer Churn Detainment in Telecom Based on Budget Restraint and Customer Detainment Value Maximization [J]	Chinese Journal of Management	L Jinyao, LIUDXIA Guoen	2012
Research on Customer Churn Model with Least Square Support Vector Machine	Applied Mechanics and Materials	TW Xia	2012

Table C.3: Google Scholar Results (continued)

Research on Customer Churn Prediction and Analysis System of ETC Based on Multilayer Feedforward Neural Network	Proceedings of the 2012 3rd International	P Jia-Hong, P Jia-Wen	2012
Research on Domain Knowledge Discovery Based on Micro-structure of Library	Library Work and Study	W Xinyun	2012
Research on E-business Customer Churning Modeling and Prediction	Computer Simulation	J REN, X ZHANG	2012
Research on Optimizing Data Mining Algorithms Based on Decision Tree [J]	Modern Computer	Z LIN, W WANG	2012
Review on customer churn issue	Computer Integrated Manufacturing	XB Yu, J Cao, ZW Gong	2012
Stock Analysis Software Based on Association Rules: Design and Implement	Journal of Tonghua Normal University	J LIU, W ZHAO, H WEN	2012
The Explanation of Support Vector Machine in Customer Churn Prediction	Contemporary Logistics	Y Li, G Xia	2012
The Research on Random Forests and the Application in Customer Churn Prediction	Management Review	Y Weiyun	2012
UYGHUR TEXT CLASSIFICATION BASED ON NAIVE BAYES AND ITS PERFORMANCE ANALYSIS	Computer Applications and Software	AATTA Hamdulla	2012
Computing, Information Systems & Development Informatics Journal	Editorial Board	CCU both Descriptive	2012
Customer Churn Prediction Model using Machine Learning Algorithms	sites.ijrit.com	I Mamvura	2012

Table C.3: Google Scholar Results (continued)

A Dynamic Transfer Ensemble Model for Customer Churn Prediction	Business Intelligence and Financial	J Xiao, Y Wang, S Wang	2013
A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques	International Journal of	H Abbasimehr, M Setak, J Soroor	2013
A neural network based approach for predicting customer churn in cellular network services	arXiv preprint arXiv:1309.3945	A Sharma, D Panigrahi, P Kumar	2013
A novel profit maximizing metric for measuring classification performance of customer churn prediction models	IEEE Transactions on	T Verbraken, W Verbeke	2013
A shilling attacks detection method of recommender systems based on hybrid strategies	Computer Engineering & Science	L Cheng-shu, W Wei-guo	2013
A Special Issue on Smart Sensors and Applications	Sensor Letters	Y Yang, B Honary	2013
Association Rule Mining Based on External Environment	Computer Technology and	F WAN, M TANG, X CHEN	2013
Business oriented data analytics: theory and case studies.		T Verbraken	2013
Case Study: Solico Food Industries Group	… Technology (IKT), 2013 5th Conference on	S Nabavi, S Jafari	2013
Cdr analysis based telco churn prediction and customer behavior insights: A case study	International Conference on Web	N Modani, K Dey, R Gupta, S Godbole	2013

Table C.3: Google Scholar Results (continued)

Churn prediction methodologies in the telecommunications sector: A survey	Advances in ICT for	WMC Bandara, AS Perera	2013
Combined algorithm of GAAC and K-means for Uyghur text clustering	Computer Engineering	T Tohti, A Ablat, M Aniwar	2013
Constraint mining in business intelligence: A case study of customer churn prediction	International Journal of	N Kerdprasop, P Kongchai, K Kerdprasop	2013
Cost-sensitive kernel principal component analysis with its application in fault diagnosis	of Central South University (Science	YB Tang, WH Gui, T Peng	2013
Customer churn data mining based on neural network prediction model	Journal of Computer Applications	LI Yang	2013
Customer churn prediction in telecommunication a decade review and classification	International Journal of	N Hashmi, NA Butt, M Iqbal	2013
Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning	Journal of Business Research	K Coussement, KW De Bock	2013
Customer churn prediction using a hybrid genetic programming approach	Research and Essays	R Obiedat, M Alkasassbeh, H Faris	2013
Customer churn prediction using a hybrid method and censored data	Management	B Bahmani, G Mohammadi	2013
Data□Intensive Analysis	The DATA Bonanza: Improving	O Corcho, J van Hemert	2013
Distributed Customer Churn Prediction Model Based on Multi-Agent in Sensor Networks	Sensor Letters	C Ju, F Guo, Q Lu	2013

Table C.3: Google Scholar Results (continued)

E-commerce customer churn prediction model combined with individual activity	Systems Engineering-Theory & Practice	CH Ju, QB Lu, FP Guo	2013
E-commerce customer churn prediction model combined with individual activity [J]	Systems Engineering-Theory & Practice	HJU Chun, BLU Qi, FP Guo	2013
Electronic Commerce Based on Self-Organizing Data Mining Customer Churn Prediction Model	2013 International Conference on Advances in	A Ren, W Zhao	2013
Evaluating the Added Value of Pictorial Data for Customer Churn Prediction		M Ballings, D Van den Poel, E Verhagen	2013
FACT-AN ADAPTIVE CUSTOMER CHURN RATE PREDICTION METHOD USING FUZZY MUTI-CRITERIA CLASSIFICATION APPROACH FOR DECISION	Asian Journal of Science and Technology	S Subramanian	2013
Handling Class Imbalance in Mobile Telecoms Customer Churn Prediction	International Journal of Computer	C Kirui, L Hong, E Kirui	2013
Hierarchical neural regression models for customer churn prediction	Journal of	G Mohammadi, R Tavakkoli-Moghaddam	2013
Hybrid data feature selection algorithm and its application in customer churn prediction	Journal of Jiangsu University of Science and	Z Junyi, M Shaohui	2013
in Churn Prediction Setting	… Retrieval Methods for	H Abbasimehr, J Mohammad	2013
Incentive Mechanism of Telecom Service Agents with Service Cost Sharing Contract	Systems Engineering	S YANG, W DUAN, Y JING, Y SUN	2013

Table C.3: Google Scholar Results (continued)

Intelligent churn prediction for telecommunication industry	International Journal of	I Khan, I Usman, T Usman, GU Rehman	2013
Knowledge Management in CRM using Data mining Techniques	… Journal of Scientific & Engineering Research	S Yadav, A Desai, V Yadev	2013
Mathematical modeling and analysis on bank customer churn prediction [J]	Journal of Changchun University of Technology	C YU, Z ZHAO	2013
Prediction of dynamic requirements of mechanical product based on Multi-variable Grey Model	Journal of Machine Design	H WANG, C ZHU	2013
Providing a customer churn prediction model using random forest and boosted tree techniques	Journal of Basic and Applied Scientific Research	S Nabavi, S Jafari	2013
Providing a Customer Churn Prediction Model Using Random Forest and SVMsTechniques (Case Study: Solico Food Industries Group)	Journal of Basic and Applied Scientific Research	S Nabavi, S Jafari	2013
Providing a Customer Churn Prediction Model using Random Forest Technique	Proc. 5th IEEE-Conference on Information and	S Nabavi, S Jafari	2013
Service profit chain model for Telecom Industry of Oman: development and empirical validation		M Ansari, S Ahmad	2013
Study of Cox Model in the Analysis of Telecom Customer Churn Causes	Journal of Hainan Normal University (Natural	C LI, X WEI, S LIU	2013
Suggestions on Steps to Solve the Problems for Hypothesis Testing	College Mathematics	Z HE, L XIA	2013
Telecom customer churn prediction based on imbalanced data resampling method	… , Information and Control	L Peng, Y Xiaoyang, S Boyu	2013

Table C.3: Google Scholar Results (continued)

The Study and Realization of Customer-Churn Model Based on Date Mining in Telcom	Applied Mechanics and	YB Jia, Q Zhang, QQ Ding, DL Liu	2013
Using Neural Networks For Customer Churn Prediction Modeling: Preliminary Findings From The Italian Electricity Industry	Using Neural Networks For Customer Churn Prediction Modeling: Preliminary Findings From The Italian Electricity Industry	N Gordini, V Veglio	2013
Cloud, Grid and Custer Computing	ieeexplore.ieee.org	M Ghorbani, MR Meybodi, AM Saghiri, FK Khalaji	2013
Customer churn prediction: Integration of clique sociometric theory in diffusion model	researchgate.net	U Droftinaa, A Koširb	2013
Evaluation and Optimization on Applicability of Big Data Technologies in Smart Grid	International Journal of Computational and	C Sun, J Zhou	2013
PREDICT CUSTOMER CHURN BY USING ROUGH SET THEORY AND NEURAL NETWORK	researchgate.net	R Qiasi, Z Roozbehani, B Minaei-Bidgoli	2013
2014 Index IEEE Transactions on Industrial Informatics Vol. 10	IEEE Transactions	B Amin, E Ancillotti, K Anderson, T Andre	2014
A Balanced Transfer Learning Model for Customer Churn Prediction	Proceedings of the Eighth International Conference	B Zhu, J Xiao, C He	2014
A comparative assessment of the performance of ensemble learning in customer churn prediction.	Int. Arab J. Inf. Technol.	H Abbasimehr, M Setak, MJ Tarokh	2014
A customer churn prediction model in telecom industry using boosting	IEEE Transactions on Industrial	N Lu, H Lin, J Lu, G Zhang	2014

Table C.3: Google Scholar Results (continued)

A genetic programming based framework for churn prediction in telecommunication industry	International Conference on	H Faris, B Al-Shboul, N Ghatasheh	2014
A New Framework for Churners' Influence Analysis Using Call Data Records	Advanced Materials Research	OO Georges, SQ Cai, Q Yuan	2014
A Study of the Application of SVM in Prediction about Decrease in Bank's Customers	Finance Forum	HE Ben-lan	2014
Application of Decision Tree Algorithm in University Employment	Computer Knowledge and Technology	LIU Zhen	2014
Customer Churn Prediction Based on Feature Clustering and Nonparallel Support Vector Machine	International Journal of	X Zhao, Y Shi, J Lee, HK Kim, H Lee	2014
Customer Churn Prediction in Mobile Operator Using Combined Model	… of the 16th International Conference on	J Mamźenko, J Gasimov	2014
Customer churn prediction in telecommunication industry: With and without counter-example	Mexican International Conference on	A Amin, C Khan, I Ali, S Anwar	2014
Customer churn prediction using recurrent neural network with reinforcement learning algorithm in mobile phone users	International Journal of	Z Kasiran, Z Ibrahim	2014
Customer Churn Prediction, Segmentation and Fraud Detection in Telecommunication Industry		ARA Ahsan Rehman	2014

Table C.3: Google Scholar Results (continued)

Customer churn prediction: an application of support vector machines in the Italian electricity industry	European Journal of Operational Research	N Gordini, V Veglio	2014
Data mining and analysis of pesticide residue in vegetables based on decision tree algorithm	Guangdong Agricultural	YM Fang, LJ Sun, JK Liao, ZG Yang	2014
Developing Churn Models Using Data Mining Techniques and Social Network Analysis		G Klepac	2014
Enhanced Boosted Tree Technique for Customer Churn Prediction Model	IOSR Journal of Engineering (IOSR-JEN)	K Kaur, S Vashisht	2014
Enhanced customer churn prediction using social network analysis	… of the 3rd Workshop on Data	MN Abd-Allah, A Salah, SR El-Beltagy	2014
Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest	Information and Communication	V Effendy, ZKA Baizal	2014
Hierarchical alpha-cut fuzzy c-means, fuzzy artmap and cox regression model for customer churn prediction	International Journal of	M Mohammadi, SH Iranmanesh	2014
Improved churn prediction in telecommunication industry using data mining techniques	Applied Soft	A Keramati, R Jafari-Marandi, M Aliannejadi	2014
Inference of Telecom Customers' Behavior Based on Rough Entropy	Industrial Engineering Journal	W Rui, W Bin	2014
Modeling of churn behavior of bank customers using logistic regression		AW Ndung'u	2014

Table C.3: Google Scholar Results (continued)

Modelling for the optimal product to offer a financial services customer		JS Mukomberanwa	2014
One-step classifier ensemble model for customer churn prediction with Imbalanced Class	… of the Eighth International Conference on	J Xiao, G Teng, C He, B Zhu	2014
Profit Maximization Using Prediction Model	International Journal Of	D Jayvant, P Vrushali, A Varsha, PAC Sheetal	2014
Profit optimizing customer churn prediction with Bayesian network classifiers	Intelligent Data Analysis	T Verbraken, W Verbeke	2014
Research on Customers Churn Prediction Model Based on Logistic	Advanced Materials Research	M Jiang, N Chu, XM Bi	2014
Research on E-Commerce Customer Churning Modeling and Prediction	Open Cybernetics & Systemics Journal	X Zhao	2014
Social network analysis for customer churn prediction	Applied Soft Computing	W Verbeke, D Martens, B Baesens	2014
Survey on Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models	International Journal	NV Patil, AM Dixit	2014
Telecom customer churn prediction method based on cluster stratified sampling logistic regression		P Li, S Li, T Bi, Y Liu	2014
The effect of social affinity and predictive horizon on churn prediction using diffusion modeling	Social Network Analysis and Mining	D Baras, A Ronen, E Yom-Tov	2014

Table C.3: Google Scholar Results (continued)

The Factors to Influence the Customer Running-Off of Commercial Banks——A Study Based on Survival Analysis Method	Finance Forum	W Wei-qing, YAO Rao, LIU Cheng	2014
The importance of social embeddedness: Churn models at mobile providers	Decision Sciences	G Benedek, Á Lublóy, G Vastag	2014
Using a decision tree for customer churn prediction at a hotel-casino property.	… " Hospitality and Tourism Futures", Dubai 6-9	EJ Suh, M Alhaery	2014
H-Infinity Stabilization for Singular Networked Cascade Control Systems With State Delay and Disturbance…………	researchgate.net	J Wang, D Liu, WH Ip, W Zhang, R Deters	2014
A case study for the churn prediction in Turksat Internet service subscription	Proceedings of the 2015 IEEE/ACM International	M Gök, T Özyer, J Jida	2015
A comparison of machine learning techniques for customer churn prediction	… Modelling Practice and	T Vafeiadis, KI Diamantaras, G Sarigiannidis	2015
A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction	New Contributions in Information	A Amin, F Rahim, I Ali, C Khan, S Anwar	2015
A multi-layer perceptron approach for customer churn prediction	International Journal of	MR Ismail, MK Awang, MNA Rahman	2015

Table C.3: Google Scholar Results (continued)

A Novel Approach for Providing the Customer Churn Prediction Model using Enhanced Boosted Trees Technique in Cloud Computing	International Journal of Computer	K Kaur, S Vashisht	2015
A Study on Application of Association Rule Mining Base on the Feature of Images' RGB Colors	Journal of Chongqing Three-Gorges	B ZHU, S HUO, H Wu	2015
Apparatus and method for predicting the behavior or state of a negative occurrence class	US Patent App. 14/305,246	F Ramberg, J Bala	2015
Application of clustering for customer segmentation in private banking	… Conference on Digital	X Yang, J Chen, P Hao	2015
Application of Data Analysis for in the Convergence of Three Networks	Journal of Liaoning University (Natural Sciences	ZHU Yue	2015
Customer churn prediction for an insurance company		C Huigevoort, R Dijkman	2015
Customer churn prediction in telecommunication	Signal Processing and Communications	M Yıldız, S Albayrak	2015
Customer churn prediction in telecommunication industries using data mining techniques-a review	Int. J. Adv. Res. Comput. Sci. Softw. Eng	K Dahiya, K Talwar	2015
Customer churn prediction in virtual worlds	… Applied Informatics (IIAI	HY Liao, KY Chen, DR Liu	2015
Customer churn prediction intelecommunications	Simulation Modelling	T Vafeiadis, KI Diamantaras, G Sarigiannidis	2015
Customer Churn Prediction: A Cognitive Approach	World Academy of	D Senanayake, L Muthugama, L Mendis	2015

Table C.3: Google Scholar Results (continued)

Echo state network with SVM-readout for customer churn prediction	… (AEECT), 2015 IEEE Jordan Conference on	A Rodan, H Faris	2015
FEATURE SELECTIONS AND CLASSIFICATION MODEL FOR CUSTOMER CHURN.	… of Theoretical &	S Nafis, M Makhtar, MK Awang	2015
Feature-selection-based dynamic transfer ensemble model for customer churn prediction	Knowledge and information	J Xiao, Y Xiao, A Huang, D Liu, S Wang	2015
Improving Churn Prediction with Voice of the Customer		Y Wang, K Satake, T Onishi, H Masuichi	2015
Initializing genetic programming using fuzzy clustering and its application in churn prediction in the telecom industry	Malaysian Journal of	B Al-Shboul, H Faris, N Ghatasheh	2015
Mining shoppers data streams to predict customers loyalty	Intelligent Systems and	V Nikulin, TH Huang, JD Lu	2015
Negative correlation learning for customer churn prediction: a comparison study	The Scientific World	A Rodan, A Fayyoumi, H Faris, J Alsakran	2015
On the method for data streams aggregation to predict shoppers loyalty	Neural Networks (IJCNN), 2015 International Joint	V Nikulin	2015
Predicting customer churn from valuable B2B customers in the logistics industry: a case study	Information Systems and e-Business	K Chen, YH Hu, YC Hsieh	2015
Preventing customers from running away! Exploring generalized additive models for customer churn prediction	The Sustainable Global	K Coussement, DF Benoit, D Van den Poel	2015

Table C.3: Google Scholar Results (continued)

Profit maximization analysis based on data mining and the exponential retention model assumption with respect to customer churn problems	… (ICDMW), 2015 IEEE	Z Zhang, R Wang, W Zheng, S Lan	2015
Profit maximizing logistic regression modeling for customer churn prediction	Data Science and	E Stripling, S vanden Broucke, K Antonio	2015
Research model of churn prediction based on customer segmentation and misclassification cost in the context of big data	Journal of Computer and Communications	Y Liu, Y Zhuang	2015
Risk Identification of Service Quality in Elderly Service Supply Chain Based on SLC-SVM	Chinese Journal of Systems Science	Z ZHANG, J ZHAO, Y SHI	2015
Telco churn prediction with big data	Proceedings of the	Y Huang, F Zhu, M Yuan, K Deng, Y Li, B Ni	2015
A Prudent Based Approach for Customer Churn Prediction	researchgate.net	S Anwar, A Amin, F Rahim	2015
Analysis on the Influence Factors of Purchase Choice on Online Book-Shop Platform	or.nsfc.gov.cn	H ZHANG, X LI	2015
Customer Churn Prediction Using Decision Tree Method in Electronic Banking Industry (Case Study: Ansar Bank)	□□□□□ □□□□ □□□□□ ,□□□□□□ 2015	□□□ □□□□ ,□□□□□□□□ □□□□ □□□□□	2015
3G □□□□□□□□□□□□□□□□□□□□□□ Research on Telecom Customer Churn Prediction Based on Customer Value Classification in 3G	Hans Journal of Data Mining	□□, □□□, □□□, □□	2016

Table C.3: Google Scholar Results (continued)

A Bagging Approach to Customer Churn Prediction	2nd International Conference on	S Tavassoli, H Koosha, ER Nik	2016
A big data clustering algorithm for mitigating the risk of customer churn	IEEE Transactions on Industrial	W Bi, M Cai, M Liu, G Li	2016
A data-driven framework for investigating customer retention		CS Mgbemena	2016
A Fuzzy Rule-Based Learning Algorithm for Customer Churn Prediction	Industrial Conference on Data	B Huang, Y Huang, C Chen, MT Kechadi	2016
A Metric for Measuring Customer Turnover Prediction Models		D Kambhampati, MR Kommidi, P Metla	2016
A New Customer Churn Prediction Approach Based on Soft Set Ensemble Pruning	… Conference on Soft	MK Awang, M Makhtar, MNA Rahman	2016
A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics	… Journal of Engineering and Technology (IRJET	V Umayaparvathi, K Iyakutti	2016
Analysis of customer churn prediction in telecom industry using decision trees and logistic regression	… Data Analysis and	PK Dalvi, SK Khandge, A Deomore	2016
Application of customer churn prediction based on weighted selective ensembles	Systems and Informatics (ICSAI)	G Xia, H Wang, Y Jiang	2016
Application of Improved Decision Tree Method based on Rough Set in Building Smart Medical Analysis CRM System	International Journal of Smart	H Xu, L Wang, W Gan	2016
Artificial neural network	Intelligent Techniques for Data Science	R Akerkar, PS Sajja	2016

Table C.3: Google Scholar Results (continued)

Attribute selection and Customer Churn Prediction in telecom industry	Data Mining and Advanced	V Umayaparvathi, K Iyakutti	2016
Bank Customer Churn Prediction Based on Imbalanced Learning	Artificial Intelligence	G WANG, Y LIN	2016
Churn prediction in customer relationship management via gmdh-based multiple classifiers ensemble	IEEE Intelligent Systems	J Xiao, X Jiang, C He, G Teng	2016
Churn prediction system for telecom using filter–wrapper and ensemble classification	The Computer Journal	A Idris, A Khan	2016
Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study	IEEE	A Amin, S Anwar, A Adnan, M Nawaz, N Howard	2016
Cost Sensitive Feature Selection in Decision-Theoretic Rough Set Models for Customer Churn Prediction: The Case of Telecommunication Sector Customers	World Academy of Science	EK Aydogan, M Ozmen, Y Delice	2016
Customer churn prediction using Big Data analytics		NNPP Tanneedi	2016
Deployment of churn prediction model in financial services industry	Behavioral, Economic and	C Chu, G Xu, J Brownlow, B Fu	2016
Developing a prediction model for customer churn from electronic banking services using data mining	Financial Innovation	A Keramati, H Ghaneei, SM Mirmohammadi	2016

Table C.3: Google Scholar Results (continued)

Development of a decision support system using data analytics for customer churn prediction for an online retailer	Decision Support Systems	B Yüceoğlu, Şİ Birbil, I Öztürk	2016
Discovering usage patterns of telecommunication subscribers based on polytomous logistic regression	… of the 2nd International Conference on	RJM Cabauatan, BD Gerardo	2016
E-commerce customer churn prediction based on improved SMOTE and AdaBoost	Service Systems and Service Management	X Wu, S Meng	2016
Effective Customer Churn Prediction on Large Scale Data using Metaheuristic Approach	Indian Journal of Science and Technology	K Sivasankar	2016
Identifying Prime Customers Based on MobileUsage Patterns	International Conference on P2P, Parallel, Grid, Cloud	D Arora, KF Li	2016
Improved Customer Churn and Retention Decision Management Using Operations Research Approach	Emerging Markets	SO Adebiyi, EO Oyatoye	2016
Intelligent Approach for Attracting Churning Customers in Banking Industry Based on Collaborative Filtering	Journal of Industrial and Systems	E Shafiei Gol, A Ahmadi, A Mohebi	2016
Leveraging purchase history and customer feedback for CRM: a case study on eBay's" Buy It Now"		J Chen	2016
Machine Learning Techniques for Customer Churn Prediction in Banking Environments		V Avon	2016

Table C.3: Google Scholar Results (continued)

Predicting customer churn in banking industry using neural networks	Interdisciplinary Description of Complex Systems	A Bilal Zorić	2016
Prediction of the shoppers loyalty with aggregated data streams	Journal of Artificial Intelligence and Soft Computing	V Nikulin	2016
Relevant Drivers for CustomersChurn and Retention Decision in the Nigerian Mobile Telecommunication Industry	Journal of	OS Sulaimon, OE Emmanuel	2016
Research of indicator system in customer churn prediction for telecom industry	… Science & Education (ICCSE), 2016 11th	Q Yihui, Z Chiyu	2016
Research on Telecom Customer Churn Prediction Based on Customer Value Classification in 3G Environment		L Xu, Z Zhu, H Li, M Li	2016
Research on the Prediction of the E-commerce Profit Based on the Improved Parallel PSO-LSSVM Algorithm in Cloud Computing Environment	International Journal of Grid and Distributed	Z Fu	2016
RFM-Based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies	International Academic Journal of	M Mohammadian, I Makhani	2016
Semi-Supervised Churn Clustering for Fault and Constraints Prediction in Telecom Industry	Indian Journal of Science and Technology	N Kamalraj, A Malathi	2016
Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction	Big Data and Cloud Computing	H Li, D Yang, L Yang, X Lin	2016

Table C.3: Google Scholar Results (continued)

The gamma CUSUM chart method for online customer churn prediction	Electronic Commerce Research and Applications	SH Chen	2016
USING DATA MINING IN TELECOMMUNICATION INDUSTRY: CUSTOMER'S CHURN PREDICTION MODEL	Journal of Theoretical and Applied	R Masoud, TM Ahmed	2016
A Review on Customer Churn Prediction in Telecommunication Using Data Mining Techniques	pdfs.semanticscholar.org	S Babu, NR Ananthanarayanan	2016
A Survey on Customer Churn Prediction using Machine Learning Techniques	pdfs.semanticscholar.org	A Saran Kumar, D Chandrakala	2016
Churn Prediction in Telecommunication Industry Using Decision Tree	Streamed Info-Ocean	MN Saini	2016
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY HYBRID APPROACH OF BOOSTED TREE FOR CHURN	Citeseer	N Kaur, N Singh	2016
Stochastic Customer Loyalty and Satisfaction Prediction using the SEM and SVM	pdfs.semanticscholar.org	M Kaur, MK Sidhu	2016
Survey on Prediction of Customer Churn Analysis in a Telecommunications Industry	ijream.org	S Monani, S Gupta, S Jain, V Mishra, A Singh	2016
5. The Research On Prediction Of Credit Card Customer Churn Based On The Combined Integration Model	Revista de la Facultad de Ingeniería	Q Zhang	2017

Table C.3: Google Scholar Results (continued)

A churn prediction model for prepaid customers in telecom using fuzzy classifiers	Telecommunication Systems	M Azeem, M Usman, ACM Fong	2017
A Combination of Multiperiod Training Data and Ensemble Methods in Churn Classification: the Case of Housing Loan Churn		T Le	2017
A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry	Decision Support Systems	K Coussement, S Lessmann, G Verstraeten	2017
A Machine Learning Approach for Churn Prediction in Telecommunication		K Mishra, RG Rani	2017
A new transferred feature selection algorithm for customer identification	Neural Computing and Applications	B Zhu, Y Niu, J Xiao, B Baesens	2017
A review and analysis of churn prediction methods for customer retention in telecom industries	Advanced Computing and	A Ahmed, DM Linen	2017
A Survey of Evolution in Predictive Models and Impacting Factors in Customer Churn	Advances in Data Science	M Ahmed, H Afzal, A Majeed, B Khan	2017
An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing	Cluster Computing	J Vijaya, E Sivasankar	2017

Table C.3: Google Scholar Results (continued)

Analysis and Conclusion on the Heat Model of Customer Service Behavior of Mail System	Computer & Telecommunication	Z Chao	2017
Analysis and Research on SPSS Clustering of Mobile Communication Data under Mobile Signal Interference	Boletín Técnico, ISSN: 0376-723X	H Du	2017
Automated Feature Selection and Churn Prediction using Deep Learning Models		V Umayaparvathi, K Iyakutti	2017
Churn prediction on huge telecom data using hybrid firefly based classification	Egyptian Informatics Journal	AAQ Ahmed, D Maheswari	2017
Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior	Journal of Systems and	S Khodabandehlou	2017
Computer Simulation of Electronic Commerce Customer Churn Prediction Model Based on Web Data Mining	Smart Grid and Electrical Automation	W Zhang, L Zhu	2017
Customer Churn Analysis for Telecom Operators Based on SVM	International Conference On	R Dong, F Su, S Yang, X Cheng, W Chen	2017
Customer Churn Prediction Based on BG/NBD Model		H Li, Z Guan, Y Cui	2017
Customer churn prediction in an internet service provider	Big Data (Big Data), 2017 IEEE	D Do, P Huynh, P Vo, T Vu	2017
Customer Churn Prediction in the Online New Media Platform: A Case Study on Juzi Entertainment	Platform Technology and	S Yuan, S Bai, M Song, Z Zhou	2017

Table C.3: Google Scholar Results (continued)

Customer churn prediction in the telecommunication sector using a rough set approach	Neurocomputing	A Amin, S Anwar, A Adnan, M Nawaz, K Alawfi	2017
Customer churn prediction model using data mining techniques	… (ICENCO), 2017 13th	IMM Mitkees, SM Badr	2017
Customer churn prediction using improved FCM algorithm	… (ICIM), 2017 3rd International Conference on	S Cui, N Ding	2017
Customer Churn Prediction: A Survey	… Journal of Advanced Research in Computer	MPP Mathai	2017
Customer management model for telecommunication services	Science Technology	B Udaya, T Indhumathi, RA Varshini	2017
Customer Segmentation by Various Clustering Approaches and Building an Effective Hybrid Learning System on Churn Prediction Dataset	Computational Intelligence in Data Mining	E Sivasankar, J Vijaya	2017
Customer: A novel customer churn prediction method based on mobile application usage	Wireless Communications and	NLR Machado, DDA Ruiz	2017
Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e	Industrial Marketing Management	N Gordini, V Veglio	2017
Data Mining Techniques for Customer Relationship Management	Journal of Physics: Conference Series	F Guo, H Qin	2017

Table C.3: Google Scholar Results (continued)

Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors	arXiv preprint arXiv:1703.03869	P Spanoudes, T Nguyen	2017
Designing of customer and employee churn prediction model based on data mining method and neural predictor	Computer and Communication	SH Dolatabadi, F Keynia	2017
DyadChurn: Customer Churn Prediction using Strong Social Ties	Proceedings of the 21st	MN Abd-Allah, SR El-Beltagy, A Salah	2017
Effects of Macroeconomic Factors on Apartment Price Fluctuations	□□□□□□□ ICCC □□□	SW Lee	2017
Evaluating Prediction of Customer Churn Behavior Based On Artificial Bee Colony Algorithm	International Journal Of Engineering And Computer	RRSR Sachdeva	2017
Experimental Parameter Tuning of Artificial Neural Network in Customer Churn Prediction	Trendy Ekonomiky a Managementu	M Fridrich	2017
Fuzzy Poisson Naive Bayes (FPNB) Model for Customer behavior Analysis and Hybrid Cuckoo Harmony Search (HCHS) based Feature Selection for Churn		S Induja, VP Eswaramurthy	2017
High accuracy predictive modelling for customer churn prediction in telecom industry	… Machine Learning and Data Mining in	R Prashanth, K Deepak, AK Meher	2017
Hyperparameter Optimization of Artificial Neural Network in Customer Churn Prediction using Genetic Algorithm	Trendy Ekonomiky a Managementu	M Fridrich	2017

Table C.3: Google Scholar Results (continued)

IMPROVED CUSTOMER CHURN BEHAVIOUR BY USING SVM		D Mahajan, R Gangwar	2017
Improving Accuracy and Performance of Customer Churn Prediction Using Feature Reduction Algorithms	Journal of Telecommunication, Electronic and	MK Awang, M Makhtar, MNA Rahman	2017
Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling	Cluster Computing	A Idris, A Iftikhar, Z ur Rehman	2017
Just-in-time customer churn prediction in the telecommunication sector	The Journal of	A Amin, F Al-Obeidat, B Shah, M Al Tae, C Khan	2017
Klasifikasi Pelanggan pada Customer Churn Prediction Menggunakan Decision Tree	PRISMA, Prosiding	A Nurzahputra, AR Safitri	2017
K-local mamum margin feature extraction algorithm for churn prediction in telecom	Cluster Computing	L Zhao, Q Gao, XJ Dong, A Dong, X Dong	2017
Managing churn to maximize profits		A Lemmens, S Gupta	2017
Methods and systems for customer churn prediction	US Patent App. 14	A Sharma, N Annamaneni, D Gupta	2017
Mitigating the Risk of Customer Churn Using K-Means Clustering	Data Mining and Knowledge Engineering	P Sharmila, J Ilakkiya	2017
Optimizing Coverage of Churn Prediction in Telecommunication Industry	INTERNATIONAL	A Anjum, A Zeb, IU Afridi, PM Shah	2017
Partition cost-sensitive CART based on customer value for Telecom customer churn prediction	Control Conference (CCC)	C Wang, R Li, P Wang, Z Chen	2017

Table C.3: Google Scholar Results (continued)

Predicting Customer Churn in Mobile Football Application	… of the 2017 International Conference on	P Sripawatakul, D Sutivong	2017
Predicting user churn on streaming services using recurrent neural networks		H Martins	2017
Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier.	International Journal of	TY Fei, LH Shuan, LJ Yan, G Xiaoning	2017
Profit Driven Business Analytics: A Practitioner's Guide to Transforming Big Data into Added Value		W Verbeke, B Baesens, C Bravo	2017
Profit maximizing logistic model for customer churn prediction using genetic algorithms	Swarm and Evolutionary	E Stripling, S vanden Broucke, K Antonio	2017
Random forest em dados desbalanceados: uma aplicação na modelagem de churn em seguro saúde		GC Lento	2017
Reducing the risk of customer migration by using bigdata clustering algorithm	Recent Trends in	A Mubeen, ND Abhinav, CVS Swamy	2017
Research on optimal model construction of electronic supersovereign currency	Industrial Economics System and	J Liao, X Chen, T Chen	2017
Research on the Application of DataWarehouse and Data Mining in Logistics System	Computer & Telecommunication	C Jia	2017
Research on the Protection of Consumers' Privacy in Electronic Commerce	Computer & Telecommunication	Z Junjie	2017

Table C.3: Google Scholar Results (continued)

Review of Customer Churn Analysis Studies in Telecommunications Industry	Karaelmas Fen ve Mühendislik Dergisi	F Kayaalp	2017
Review on Prediction of Churn Customer Behavior	International Journal of Engineering	RR Sharma, R Sachdeva	2017
Statistics-based CRM approach via time series segmenting RFM on large scale data	Knowledge-Based Systems	M Song, X Zhao, E Haihong, Z Ou	2017
Survey of Data Mining Techniques Used for Real Time Churn Prediction	International	JC Turiho, W Cheruiyot, A Kibe	2017
Telecom Churn Prediction Model Using Data Mining Techniques	Bahria University Journal of	S Albadawi, MM Fraz, F Kharbat	2017
The Research on the Prediction of Enterprise Customer Churn on Voice Services		Y Lin	2017
Towards an efficient method of modeling "Next Best Action" for Digital Buyer's journey in B2B	… Conference on Machine	A Bhandari, K Rama, N Seth, N Niranjan	2017
Two-layer Clustering Model for Mobile Customer Analysis	IT Professional	CI Chang, JC Ho	2017
A Predictive Analytics Model for Maximising Profit in e-commerce Companies	researchgate.net	S Das, P Singh, G Puri	2017
Author name Paper 1 Paper 2 Paper 3	ieeexplore.ieee.org	A Accent, CU K-Means	2017
Benchmarking analytical techniques for churn modelling in a B2B context	lib.ugent.be	J Van Haver	2017

Table C.3: Google Scholar Results (continued)

Building Customer Churn Prediction Models in Fitness Industry with Machine Learning Methods	umu.diva-portal.org	M Shan	2017
CUSTOMER CHURN PREDICTION–A PILOT PROJECT PERSPECTIVE	researchgate.net	M Mirkovic, A Anderla, D Stefanovic, D Culibrk	2017
3-Z-Network Boost Converter	Designing Impedance Networks Converters	G Zhang, B Zhang, Z Li	2018
A model for the mobile market based on customers profile to analyze the churning process	Wireless	MR Flores-Méndez, M Postigo-Boix, JL Melús-Moreno	2018
A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees	European Journal of	A De Caigny, K Coussement, KW De Bock	2018
A stability constrained adaptive alpha for gravitational search algorithm	Knowledge-Based Systems	G Sun, P Ma, J Ren, A Zhang, X Jia	2018
Analyzing and Investigating the Use of Electronic Payment Tools in Iran using Data Mining Techniques	Journal of AI and Data Mining	F Moslehi, AR Haeri, AR Moini	2018
Automated call classification	US Patent 9,961,202	F Bellosi, LM Newnham, T Ram	2018
Automatic detection of anomalies in electronic communications	US Patent App. 15	ML Carlough, CM Curtin, RK Parkin	2018
Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context	International Journal of	AD Rachid, A Abdellah, B Belaid	2018
Cost-Sensitive Churn Prediction in Fund Management Services	… Conference on Database	J Brownlow, C Chu, B Fu, G Xu, B Culbert	2018

Table C.3: Google Scholar Results (continued)

Customer Churn Prediction in Superannuation: A Sequential Pattern Mining Approach	Australasian Database	B Culbert, B Fu, J Brownlow, C Chu, Q Meng	2018
Customer churn prediction in telecommunication industry using data certainty	Journal of Business	A Amin, F Al-Obeidat, B Shah, A Adnan, J Loo	2018
Designing Impedance Networks Converters		G Zhang, B Zhang, Z Li	2018
Particle classification optimization-based BP network for telecommunication customer churn prediction	Neural Computing and	R Yu, X An, B Jin, J Shi, OA Move, Y Liu	2018
Retention analysis based on a logistic regression model: A case study	… , Sensing and Control	M Ghahramani, MC Zhou, CT Hon	2018
Tele Comm. Customer Data Analysis using Multi-Layer Clustering Model		Y Gopi, V Sumalatha	2018
Time-sensitive Customer Churn Prediction based on PU Learning	arXiv preprint arXiv:1802.09788	L Wang, C Chen, J Zhou, X Li	2018
Analysis Of Customer Churn Prediction In Telecom Sector Using Random Forest	ijire.org	D Prajapati, RK Dubey	2018
Falling the Risk of Consumer Relocation by Using Big data Clustering Algorithm	papers.ssrn.com	P Nirupama, EM Reddy	2018
Key-Attributes-Based Ensemble Classifier for Customer Churn Prediction	Journal of Electronic Science	Y Qian, LQ Li, JR Ran, PJ Shao	2018

Table C.3: Google Scholar Results

Appendix D

Interview Agenda

The interview questionnaire developed as part of the interview process followed during this research work consists of five segments.

Interview Questions Segments

- **Segment A: General Organization Data**
- **Segment B: Participant Data**
- **Segment C: Customer Churn**
- **Segment D: Customer Behavior Forecasting Framework (CBFF)**
- **Segment E: Comments**

Before moving to interview sections i would like to clarify the notation for the terms below, to have the same perspective and eliminate misunderstandings during the interview:

Feature selection: The interviewer refer to the process that from datasets, archival data or any other organization's records derive the KPIs, that the churn process should be based on to identify churning customers.

Data Type: The interviewer refer to categories of data as those presented in Data Type element in Figure D.1.

Customer Churn: Occurs, When a customer stops collaboration with a company for another competitor in the market.

Segment A: General Organization Data

A.1 Organization Name and Address

Organization Name: . . . .

Organization Address: . . . .

A.2 How many employees working at the organization approximately ?

- ☐ < 5
- ☐ 5 - 10
- ☐ 11 - 50
- ☐ 51 - 500
- ☐ 501 - 1000
- ☐ 1001 - 5000
- ☐ 5001 or more

A.3 What is the key business of your organization ?

. . . .
. . . .
. . . .
. . . .
. . . .
. . . .
. . . .

A.4 Is a multinational organization ?

☐ Yes

☐ No

A.4.1 If above question is ”Yes” in how many countries does your organization operates ?

. . . .

A.5 What services does your organizations offer ?

. . . .
. . . .
. . . .

A.6 How many customers do you have approximately ?

. . . .
. . . .

Segment B: Participant Data

B.1 Interviewee Contact Details

Full Name: ....

Telephone Number: ....

E-mail Address: ....

B.2 Interviewee Age

☐ 18 - 25

☐ 26 - 35

☐ 36 - 45

☐ 46 - 55

☐ 56 - 65

☐ 66 - 75

☐ 75 or more

B.3 Interviewee Gender

☐ Male

☐ Female

B.4 Interviewee Role in Organization

....

....

....

....

Segment C: Customer Churn

C.1 What are the challenges faced by the organization? Could you name them and give a brief description ?

. . . .
. . . .
. . . .
. . . .

C.2 Do you have a definition for account health, (loyalty) for your customers ?

. . . .
. . . .
. . . .
. . . .

C.3 Could you identify the benefits (e.g., economic, business target related etc.,) from a successful retention campaign than a new customer targeting campaign ?

. . . .
. . . .
. . . .
. . . .

C.4 Does your organization run Marketing Campaigns ? Could you name the total amount spent by your on Marketing Campaigns ?

. . . .
. . . .

C.4.1 What percentage of this corresponds to each of the following categories ?

Attract new customers: . . . .
Existing customers: . . . .
Other marketing activities: . . . .

C.5 Do you use any system or method to predict customer churn ?

☐ Yes

☐ No

C.5.1 If you answered "Yes" in the previous question, could you: 1) name the systems / methods used for customer churn prevention, 2) the responsible department running each system/ method in the organization and 3)persons involved running the system /method ?

System	Type	Responsible Department	Persons Involved

Table D.1: Systems for Customer Churn in Organization

C.6 Are there any ongoing projects for customer churn ?

☐ Yes

☐ No

C.6.1 If there are any, could you name them ?

Project Name	Project Scope	Budget

Table D.2: Ongoing Projects on Customer Churn

C.7 In an attempt to visualize the steps of churn prediction process the author combine the elements identified in literature. Could you comment on the churn process steps presented in high abstraction level below ?

. . . .
. . . .
. . . .

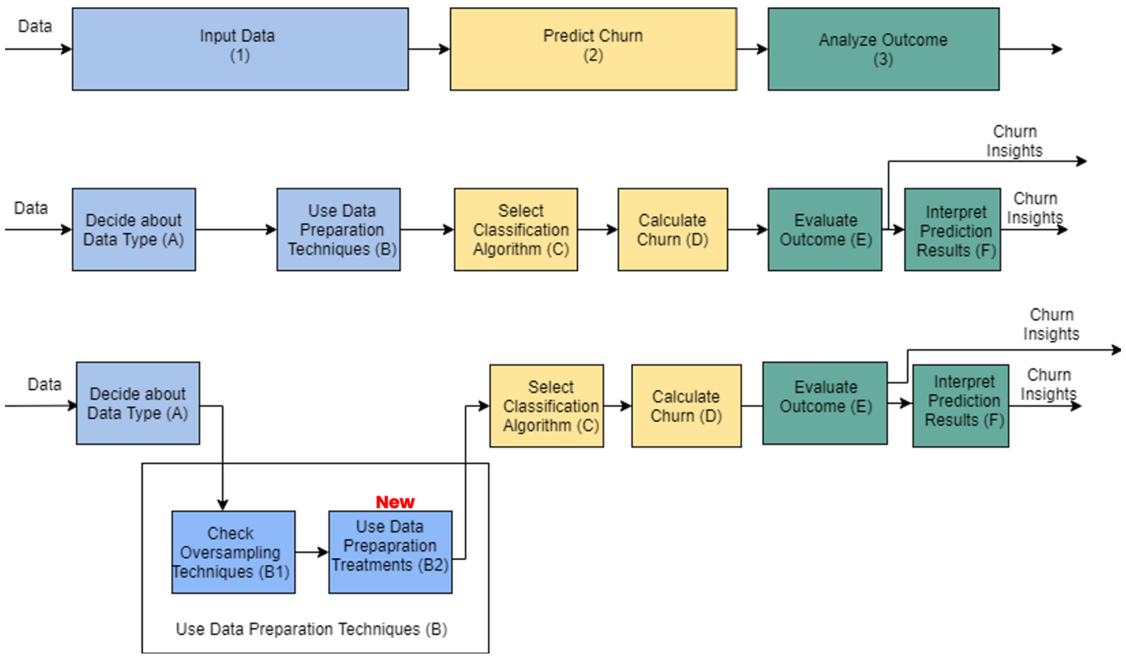


Figure D.1: Modified Version of Churn Prediction Procedure

Segment D: Customer Behavior Forecasting Framework (CBFF)

D.1 Data Analytics is one of the trends of nowadays. Experienced Data Analysts strongly belief that none machine learning model could give meaningful insights if the data (the heart of the model) are not useful. Based on the previous assumption, how important are business indicators, KPIs desired goals etc., for the organizations should be taken into account in a customer churn prevention approach and more specifically should be taken into consideration in a feature selection decision? Could you provide your opinions on that ?

....

....

....

....

....

....

D.2 What is your opinion on the type of data used for churn prediction - Is there a "best" data type to be used for churn prediction ? ?

. . . .
. . . .
. . . .
. . . .
. . . .
. . . .

D.3 Multiple learning algorithms used to predict customer churn. Do you believe that there is a "best algorithm" for churn prediction? Is your answer based on your organizations specific boundaries ?

. . . .
. . . .
. . . .
. . . .
. . . .

D.4 What parameters do you take into consideration to decide for classification algorithm ?

. . . .
. . . .
. . . .

. . . .
. . . .

D.5 How important is churn prediction outcome interpretation in the development of new features or exclusion of features in future churn prediction attempts ?

. . . .
. . . .
. . . .
. . . .
. . . .

D.6 What parameters do you take into consideration to decide for classification algorithm ?

. . . .
. . . .
. . . .
. . . .
. . . .

D.7 Churn prediction models focus on churn rate prediction as single metric or a group of metrics related to churn rate and evaluation metrics. Do you believe that churn prediction model outcome interpretation, compared to organizations desired goals, KPIs, give the opportunity to organization to gain real value from churn prediction ?

. . . .
. . . .
. . . .
. . . .
. . . .
. . . .

D.8 How important is churn prediction outcome interpretation in the development of new features or exclusion of features in future churn prediction attempts ?

. . . .
. . . .
. . . .
. . . .
. . . .
. . . .

D.9 In an attempt to compose the customer behavior forecasting framework the author identified and proposed a list of elements to consist this procedure. Could you please fill the following list if you identify the suitability of proposed element in a churn prediction process scenario you create ? Please fill any comments you have as well.

Procedures	Description	Identified	Comment
Business Indicators, KPIs in feature selection	Analysis of business indicators, KPIs and desired goals of the company to be part of the process that decide for features to be included in churn prediction dataset.	Yes / No	
Data preparation techniques	Data pre-processing is crucial for high churn prediction accuracy (e.g., oversampling, variables balancing etc.,). The existence of a procedure that process data before proceed them into churn prediction model should exist.	Yes / No	
Type of data to use for customer churn	Different organizations need different kind of data for churn prediction dataset. The existence of a procedure that defines the more suitable data type to be used for churn prediction.	Yes / No	
Data source combination	If multiple data sources are available in an organization, a procedure that checks available data resources and defines if one or more should be combined to generate churn prediction dataset	Yes / No	
Learning Algorithm	A procedure that selects the best fit classification algorithm for specific churn dataset based on features included, size, type of variables, previously used in organization - similar etc.	Yes / No	
Evaluation Metrics	A procedure that measure the performance of learning algorithms used for churn prediction.	Yes / No	
Outcomes Reporting	For the organization to gain real value a procedure that interprets the outcome of churn prediction attempt, instead of basic churn rate metric, should be part of a churn prediction process.	Yes / No	
Insights Retrofit	Interpretation of churn prediction results should be used for updating parameters(e.g., new features in dataset, remove old features, other classification algorithm etc.,) in future churn prediction attempts	Yes / No	

Table D.3: Proposed procedures to be included in every churn prediction process

D.10 The figure Figure D.2 presents the Customer Behavior Forecasting Framework this research work propose (I will explain you each element in details). Could you provide any comments on the propositions based on your experience ? Could you create a customer behavior forecasting scenario and identify the usage of the proposed CBFF in your organization ?

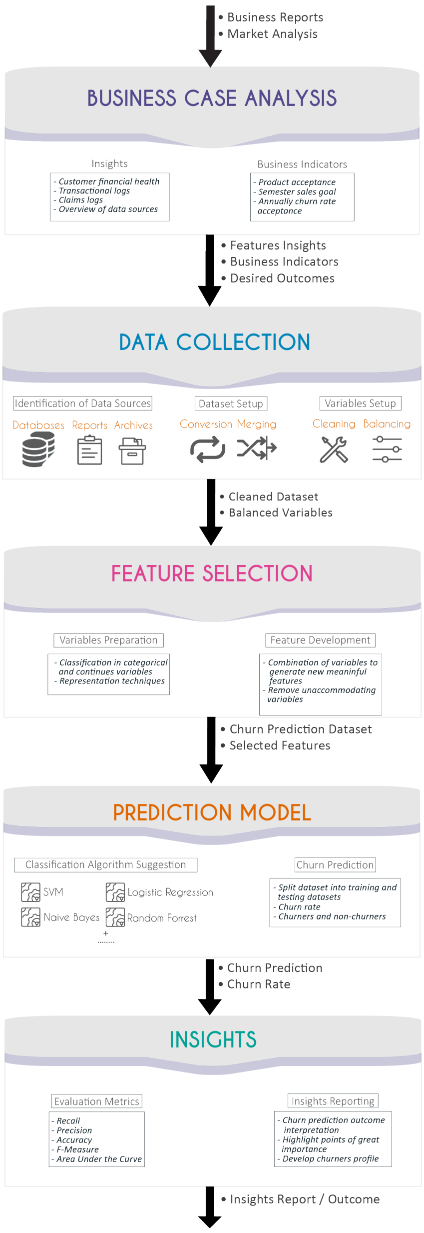


Figure D.2: Customer Behavior Forecasting Framework

Segment E: Comments

E.1 Do you have any further comments about this research ?

....
....
....
....
....
....
....
....

Appendix E

Εκτενής Περίληψη στα Ελληνικά

Ορισμός Προβλήματος

Συνδρομητικοί οργανισμοί όπως εταιρίες λογισμικού, τηλεπικοινωνιακοί οργανισμοί, συνδρομητικές πλατφόρμες τηλεόρασης, οργανισμοί παροχής ηλεκτρικής ενέργειας και άλλοι, βρίσκονται σε μία διαρκή αναζήτηση τρόπων διατήρησης και αύξησης του πελατολογίου και του τζίρου τους καθώς αυτά αποτελούν θέματα ζωτικής σημασίας για τη βιωσιμότητα και την ανάπτυξη τους. Η παρούσα διπλωματική διατριβή εστιάζει στη μελέτη της απώλειας πελατών (customer churn) συνδρομητικών οργανισμών και στη δημιουργία ενός πλαισίου(framework) που μπορεί να χρησιμοποιηθεί ως εργαλείο λήψης αποφάσεων για τη μείωση της απώλειας και την αύξηση της αποτελεσματικότητας.

Η απώλεια πελατών έχει αναδειχθεί σε μέιζον ζήτημα για τους συνδρομητικούς οργανισμούς, καθώς έχει αρνητικό αντίκτυπο στην αγορά ενώ παράλληλα επηρεάζει την ανάπτυξη και την κερδοφορία τους. Η μη κατανόηση της απώλειας αυτής μπορεί να μετατραπεί σε συστηματική και να απειλήσει την ύπαρξη του ιδίου του οργανισμού. Για το λόγο αυτό, αρκετοί πάροχοι συνδρομητικών υπηρεσιών έχουν επενδύσει σημαντικούς πόρους (π.χ., οικονομικούς, ανθρώπινο δυναμικό) στην αναζήτηση λύσεων για την αντιμετώπιση του εν λόγω προβλήματος. Το φαινόμενο αυτό έχει προσελκύσει το ενδιαφέρον πληθώρας ερευνητών, από τον ακαδημαικό και επιχειρηματικό κόσμο, που αποσκοπούν στη καλύτερη κατανόηση και βελτίωση της πρόβλεψης της πιθανής απώλειας πελατών με σκοπό την εγκαίρη αντιμετώπιση και πρόληψη της.

Ερευνητικός Σκοπός και Στόχοι

Η παρούσια διδακτορική διατριβή σκοπεύει στη:

> *"Μελέτη του φαινομένου της απώλειας πελατών σε συνδρομητικούς οργανισμούς. Η μελέτη αυτή θα οδηγήσει στη δημιουργία ενός πλαισίου κατανόησης και πρόβλεψης της απώλειας των πελατών που θα μπορεί να χρησιμοποιηθεί ως εργαλείο λήψης αποφάσεων."*

Για την επίτευξη του ανωτέρου σκοπού, καθορίστηκαν οι ακόλουθοι ερευνητικοί στόχοι:

Ερευνητικός Στόχος 1: Εκτενής ανασκόπηση και κριτική αξιολόγηση, της σχετικής με την απώλεια πελατών σε συνδρομητικούς οργανισμούς, βιβλιογραφίας.

Ερευνητικός Στόχος 2: Δημιουργία εννοιολογικού πλαισίου για την υποστήριξη της πρόβλεψης της απώλειας των πελατών συνδρομητικών οργανισμών.

Ερευνητικός Στόχος 3: Ανασκόπηση επιστημονικών μεθόδων έρευνας και επιλογή της καταλληλότερης για την αξιολόγηση του προτεινόμενου πλαισίου.

Ερευνητικός Στόχος 4: Ανάλυση των ευρημάτων της παρούσας έρευνας και πρόταση καινοτόμων προτάσεων που θα συνεισφέρουν στην καλύτερη κατανόηση του φαινομένου της απώλειας πελατών σε συνδρομητικους οργανισμούς και θα υποστηρίζει το μηχανισμό λήψης αποφάσεων.

Αποτελέσματα Βιβλιογραφικής Ανασκόπησης

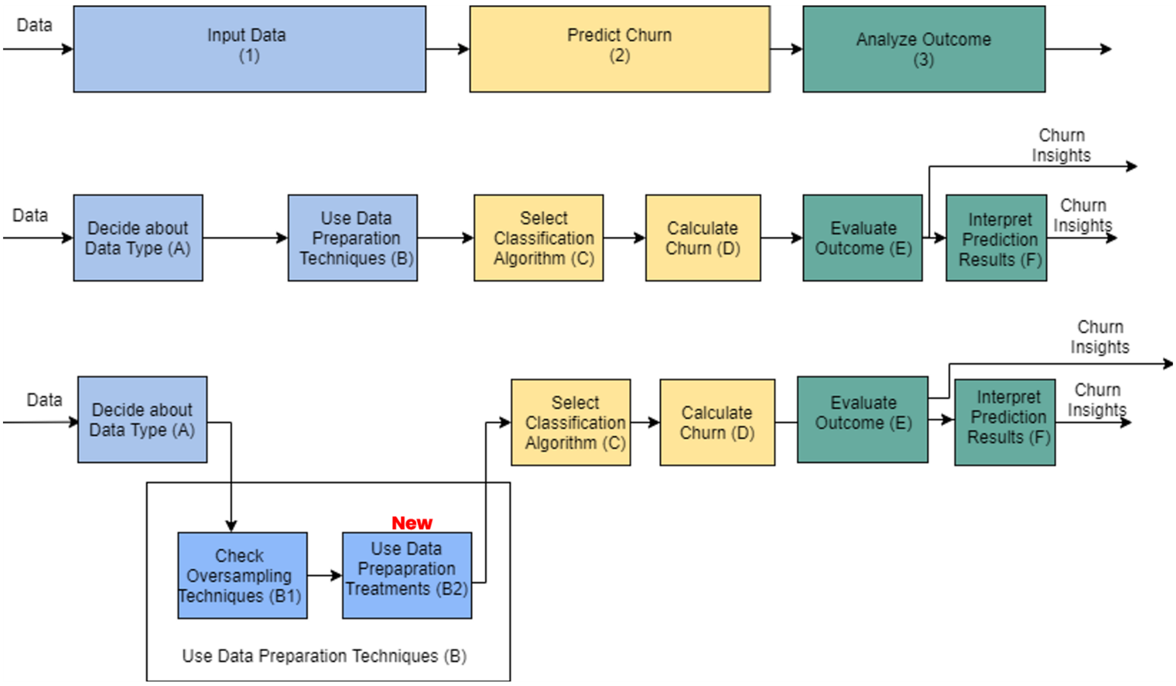
Στα πλαίσια της παρούσας διατριβής πραγματοποιήθηκε μία μεθοδολογική συστηματική βιβλιογραφική ανασκόπηση υιοθετώντας την προσέγγιση των Brereton et al., [24] και τις κατευθηντήριες γραμμές που πρότεινε ο Cooper [25]. Η ολοκλήρωση της βιβλιογραφικής ανασκόπησης οδήγησε στις ακόλουθες παρατηρήσεις (Observations):

- **Παρατήρηση 1:** Η απώλεια πελατών είναι ζωτικής σημασίας για τις επιχειρήσεις και ιδιαίτερα τους συνδρομητικούς οργανισμούς.
- **Παρατήρηση 2:** Απαιτείται η δημιουργία ενός πλαισίου πρόβλεψης της απώλειας των πελατών καθώς οι ερευνητικές προσπάθειες που έχουν γίνει μέχρι σήμερα, δεν λαμβάνουν υπόψη όλες τις παραμέτρους του προβλήματος.
- **Παρατήρηση 3:** Η σχετική με το θέμα αυτό βιβλιογραφία, εστιάζει κυρίως σε τρείς βασικούς τομείς του εν λόγω προβλήματος: (α) τεχνικές προετοιμασίας δεδομένων, (β) επιλογή αλγορίθμου κατηγοριοποίησης και (γ) τύπος δεδομένων πρόβλεψης απώλειας πελατών.
- **Παρατήρηση 4:** Δεν υπάρχει μία ερευνητική εργασία που να ολοκληρώνει και τους τρεις βασικούς τομείς.
- **Παρατήρηση 5:** Απουσιάζει μία κοινά αποδεκτή διαδικασία πρόβλεψης της απώλειας των πελατών.

Οι ανώτερω πέντε παρατηρήσεις συνιστούν την υπάρξη ερευνητικού κενού και ανοικτών θεμάτων προς περαιτέρω διερεύνηση.

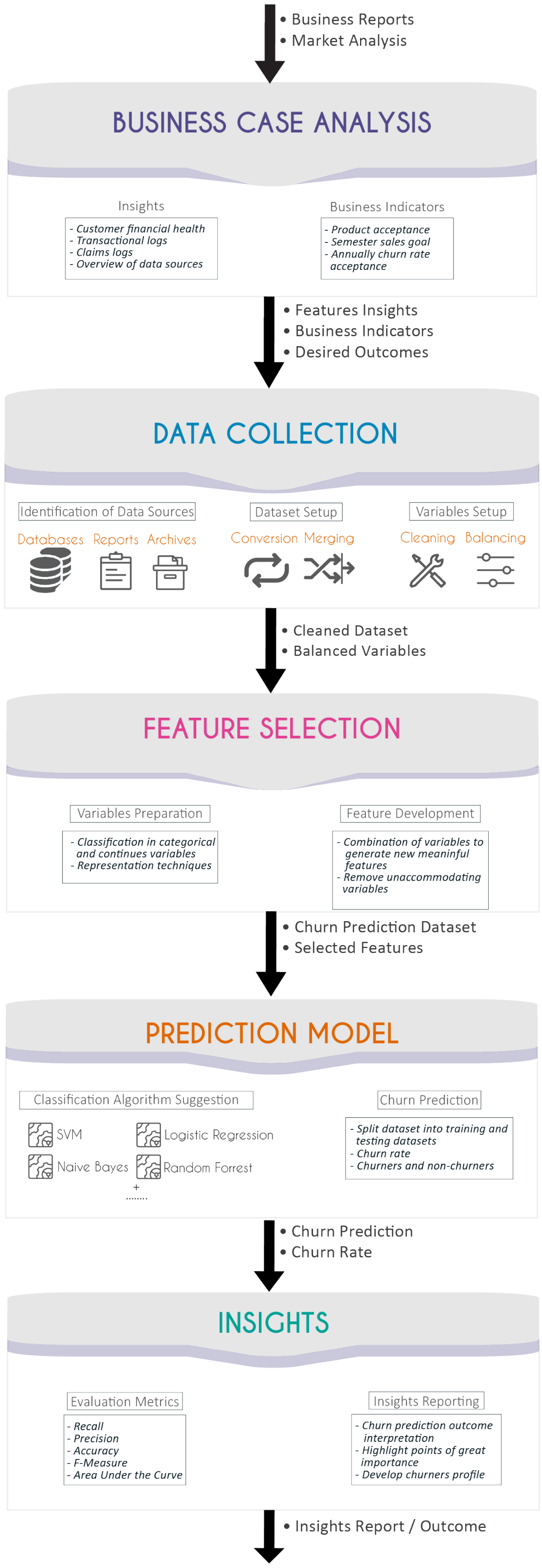
Εννοιολογικές Προτάσεις

Λαμβάνοντας υπόψη τα αποτελέσματα της βιβλιογραφικής ανασκόπησης, η εργασία εστίασε αρχικά στη δημιουργία μίας κοινώς αποδεκτής διαδικασίας για την πρόβλεψη της απώλειας των πελατών (Σχήμα Ε.1). Η προτεινόμενη διαδικασία, περιλαμβάνει τρία κύρια στάδια: i) επεξεργασία και εισαγωγή δεδομένων, ii) υπολογισμός πρόβλεψης και iii) διαμόρφωση και έκδοση αποτελεσμάτων. Τα στάδια αυτά περιλαμβάνουν τους τρεις βασικούς άξονες που αναφέρθηκαν στην Παρατήρηση 3 της προηγούμενης ενότητας, και τους ολοκληρώνει μαζί με άλλα συστατικά που προέκυψαν από τη βιβλιογραφία ή που κατά τη γνώμη του συγγραφέα κρίνονται ως απαραίτητα.



Σχήμα Ε.1: Προτεινόμενη Διαδικασία Πρόβλεψης Απώλειας Πελατών

Ακολούθως, η προτεινόμενη διαδικασία χρησιμοποιείται για τη δημιουργία του προτεινόμενου εννοιολογικού πλασίου το οποίο συνθέτει και ενορχηστρώνει βιβλιογραφικά ευρήματα ή στοιχεία που ο συγγραφέας θεωρεί ως σημαντικά. Το προτεινόμενο πλαίσιο αποτελείται από πέντε βασικά στοιχεία και πιο συγκεκριμένα: i) Ανάλυση Επιχειρησιακών Στόχων, ii) Συλλογή Δεδομένων, iii) Επιλογή Χαρακτηριστικών, iv) Επιλογή Μοντέλου Πρόβλεψης και v) Ανάλυσης Αποτελεσμάτων. Η τελική διαμόρφωση του προτεινόμενου μοντέλου συνοψίζεται στο Σχήμα Ε.2.



Σχήμα Ε.2: Προτεινόμενο Πλαίσιο Πρόβλεψης Απώλειας Πελατών

Μεθοδολογία Έρευνας

Προκειμένου να ελεγχθούν επιστημονικά οι εννοιολογικές προτάσεις της εργασίας αυτής, ο συγγραφέας μελέτησε μεθοδολογίες έρευνας που εφαρμόζονται στο χώρο των πληροφοριακών συστημάτων έτσι ώστε να επιλέξει την καταλληλότερη. Ως εκ τούτου, επιλέχθηκε η ερμηνευτική, ποιοτική προσέγγιση, με τη χρήση πολλαπλών μελετών περίπτωσης ως μία κατάλληλη στρατηγική για την αξιολόγηση των προτεινόμενων εννοιολογικών προτάσεων. Η επιλογή της συγκεκριμένης μεθοδολογίας έρευνας στηρίχθηκε στην καλύτερη υποστήριξη του σκοπού και των στόχων της συγκεκριμένης διδακτορικής διατριβής, καθώς και στην φύση των προτεινόμενων εννοιολογικών προτάσεων.

Μελέτες Περίπτωσης και Εμπειρικά Δεδομένα

Στα πλαίσια της παρούσας εργασίας, δύο μελέτες περίπτωσης τηλεπικοινωνιακών συνδρομητικών οργανισμών διερευνήθηκαν διεξοδικά. Για εμπιστευτικούς λόγους, ο συγγραφέας αναφέρεται στους οργανισμούς αυτούς χρησιμοποιώντας τις ονομασίες BlueTelco και OrangeTelco.

Η επιλογή των συγκεκριμένων περιπτώσεων, έγινε με βάση την καλύτερη εξυπηρέτηση του βασικού σκοπού και των επιμέρους στόχων αυτής της διδακτορικής διατριβής, καθώς και στη βάση του πρωτοκόλλου έρευνας το οποίο περιγράφεται αναλυτικά στην ενότητα Ενότητα 4.5. Η πρώτη μελέτη περίπτωσης που αναφέρεται ως BlueTelco, επενδύει σε νέες τεχνολογίες και έρευνα, που είχε μέχρι στιγμής ως αποτέλεσμα την γρήγορη εξέλιξη της ως οργανισμού με αρκετά σημαντικό μερίδιο της αγοράς σε σχέση με το διάστημα κατά το οποίο δραστηριοποιείται. Σε αντίθεση με την BlueTelco, που βρίσκεται στην Νοτιο-Ανατολική Ευρώπη, η OrangeTelco, βρίσκεται εκτός Ευρώπης. Λόγω του Ευρωπαικού κανονισμού GDPR, ο συγγραφέας αντιμετώπισε πάρα πολλά προβλήματα στη συλλογή δεδομένων από επιχειρήσεις που δραστηριοποιούνται εντός Ε.Ε. Έτσι στράφηκε σε οργανισμό που βρίσκεται εκτός Ε.Ε. προκειμένου να συλλέξει τα απαραίτητα δεδομένα.

Τόσο η BlueTelco, όσο και η OrangeTelco έκριναν ότι η προτεινόμενη διαδικασία και το προτεινόμενο πλαίσιο για την πρόβλεψη απώλειας πελατών, τους παρέχουν σημαντική βοήθεια και μπορούν να τους υποστηρίξουν στη λήψη αποφάσεων σχετικά με αυτό το θέμα. Ως αποτέλεσμα, οι εν λόγω οργανισμοί επαλήθευσαν και υποστηρίζουν τις ανωτέρω εννοιολογικές προτάσεις. Παράλληλα πρότειναν και μερικές μικρές τροποποιήσεις στο προτεινόμενο πλαίσιο πρόβλεψης της απώλειας πελατών. Τα εμπειρικά δεδομένα αξιολογήθηκαν και οι προτεινόμενες τροποποιήσεις οδήγησαν στη βελτιωμένη έκδοση του εννοιολογικού πλαισίου όπως αυτό παρουσιάζεται στο Σχήμα E.3.

Ευρήματα και Καινοτόμα Αποτελέσματα Έρευνας

Τα κυριότερα συμπεράσματα της συγκεκριμένης ερευνητικής εργασίας συνοψίζονται ως εξής:

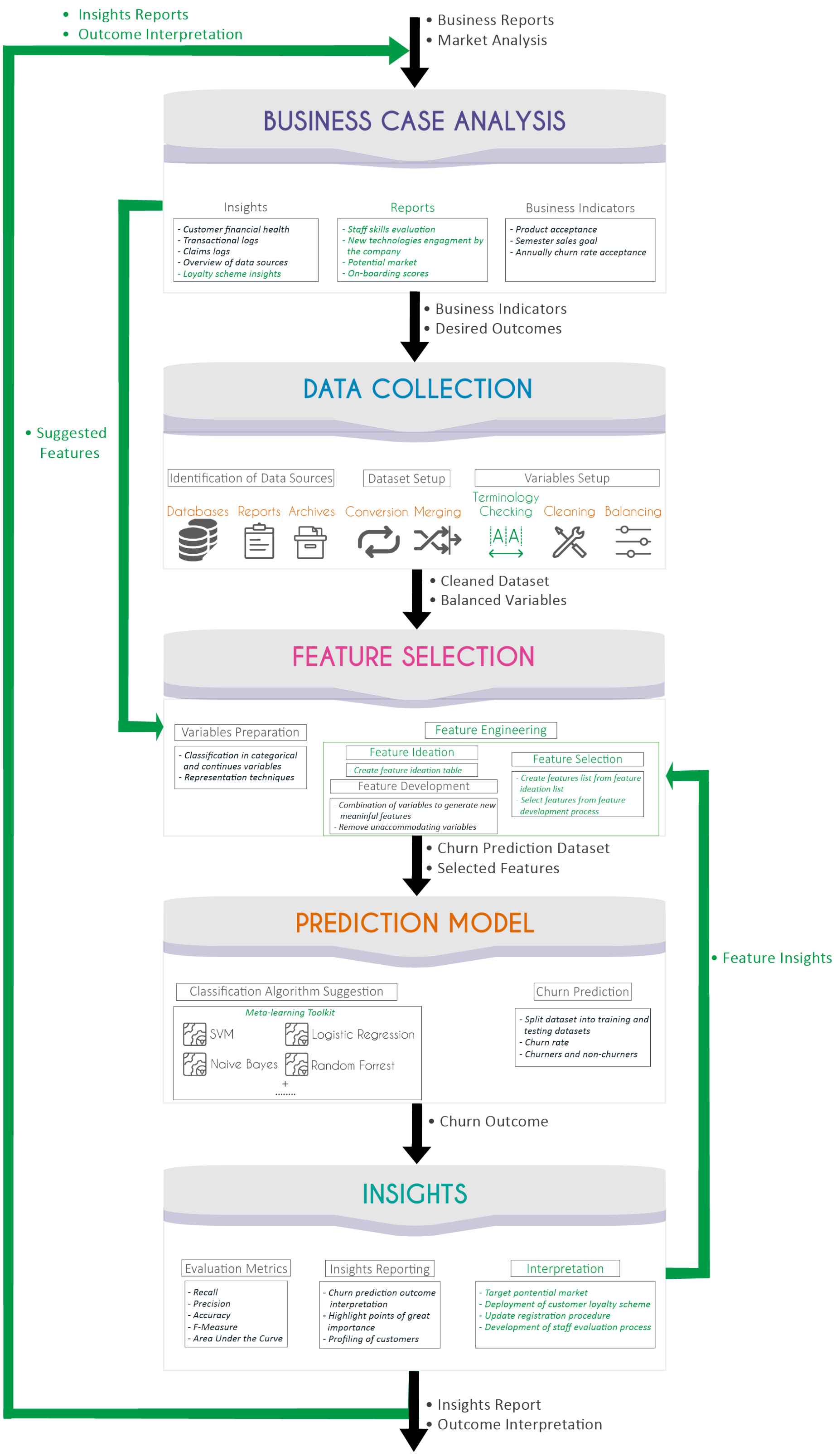
Εύρημα 1: Μέσα από τη συστηματική ανασκόπηση της βιβλιογραφίας που πραγματοποιήθηκε στην εργασία αυτή, προέκυψαν πέντε σημαντικές παρατηρήσεις-διαπιστώσεις. Οι παρατηρήσεις αυτές συνιστούν ερευνητικό κενό και επικεντρώνονται στην έλλειψη ενός πλαισίου καθώς και μιας κοινώς αποδεκτής διαδιακασίας για την πρόβλεψη της εκκροής πελατών από συνδομητικούς οργανισμούς.

Εύρημα 2: Η μελέτη αυτού του ερευνητικού κενού οδήγησε στη δημιουργία έννοιολογικών προτάσεων που αντιμετωπίζουν το εν λόγω πρόβλημα γεγονός που αποδεικνύεται από την παρουσίαση και ανάλυση των εμπειρικών δεδομένων των τηλεπικοινωνιακών συνδρομητικών οργανισμών BlueTelco και OrangeTelco. Επιπροσθέτως, τα εμπειρικά δεδομένα υπέδειξαν τροποποιήσεις στο αρχικό εννοιολογικό πλαίσιο πρόβλεψης απώλειας πελατών όπως αυτές αναλύονται στην Ενότητα 6.3 και συνοψίζονται στο Σχήμα E.3.

Εύρημα 3: Η παρούσα εργασία οδήγησε στη δημιουργία καινοτόμας γνώσης η οποία εμπλουτίζει τη βιβλιογραφία και υποστηρίζει τη βελτίωση των πρακτικων των συνδρομητικών οργανισμών. Συγκεκριμένα η εργασία παρέχει τις ακόλουθες καινοτομίες:

Καινοτομία 1: Καθορισμός διαδικασίας πρόβλεψης απώλειας πελατών η οποία μπορεί να χρησιμοποιηθεί απο τους οργανισμούς στις ενέργειες τους σχετικά με την πρόβλεψη απώλειας πελατών.

Καινοτομία 2: Παροχή πλαισίου πρόβλεψης εκκροών πελατών που μπορεί να χρησιμοποιηθεί ως εργαλείο λήψης αποφάσεων για τον περιορισμό των απωλείων των πελατών και αύξησης της αποτελεσματικότητας των οργανισμών. Παρά το γεγονός ότι το προτεινόμενο πλαίσιο είναι καινοτόμο στο σύνολο του, δομικά στοιχεία αυτού είναι επίσης καινοτόμα καθώς προτείνουν νέους τρόπους μελέτης και πρόβλεψης της απώλειας πελατών.



Σχήμα Ε.3: Τελικό Πλαίσιο Πρόβλεψης Συμπεριφοράς Πελατών

Καινοτομία 3: Ενίσχυση και εμπλουτισμός της διαδιακασίας κατανόησης της συμπεριφοράς των πελατών και δημιουργία σχετικής γνώσης για αυτούς μέσα από την ανάλυση δεδομένων και την κατά περίπτωση εφαρμογής προηγμένων τεχνικών.

Καινοτομία 4: Υποστήριξη επιλογής κατάλληλων αλγορίθμων ταξινόμησης μέσω της χρήσης τεχνικών μηχανικής μάθησης (π.χ., meta-learning).

Καινοτομία 5: Παροχή ενός δυναμικού τρόπου υπολογισμού της πρόβλεψης της εκκροής των πελατών μέσω της χρήσης μηχανισμών επανατροφοδότησης και εμπλουτισμού της γνώσης.

Καινοτομία 6: Παροχή μηχανισμού πρότασης προληπτικών μέτρων για την αντιμετώπιση πελατών που παρουσιάζουν μεγάλη πιθανότητα τερματισμού της συνεργασίας με τον οργανισμό.

Ερευνητικοί Περιορισμοί και Προτάσεις για Μελλοντική Έρευνα

Παρά την καινοτόμα συνεισφορά της παρούσας έρευνας διαπιστώνονται οι ακόλουθοι περιορισμοί:

- Τα εμπειρικά δεδομένα προέρχονται μόνο από δύο συνδρομητικούς τηλεπικοινωνιακούς οργανισμούς
- Για τον έλεγχο των εννοιολογικών προτάσεων χρησιμοποιήθηκε ποιοτητική έρευνα γεγονός που δεν επιτρέπει την γενίκευση των αποτελεσμάτων
- Η εφαρμογή του κανονισμού GDPR της Ε.Ε. καθιστά εξαιρετικά δύσκολη και σε αρκετές περιπτώσεις αδύνατη τη συλλογή δεδομένων από Ευρωπαικούς οργανισμούς γεγονός που δημιουργεί αρκετά προβλήματα στην ερευνητική κοινότητα.

Παράλληλα με τους ανωτέρω περιορισμούς καταγράφεται μία σειρά από προτάσεις για μελλοντική έρευνα που μεταξύ άλλων περιλαμβάνουν:

- Την επέκταση της παρούσας έρευνας με τη μελέτη συνδρομητικών οργανισμών από άλλους κλάδους (π.χ. τηλεοπτικές πλατφόρμες όπως το Netflix, εταιρίες παροχής συνδρομητικών παιχνιδιών (gaming industry) κλπ).
- Υιοθέτηση της πρότασης των Miles and Huberman[78] για την εναλλαγή της ποσοτικής με την ποιοτική έρευνα (π.χ. ποιοτική, ποσοτική, ποιοτική). Κάτι τέτοιο θα απαίτησει βέβαια αρκετό χρόνο για να ολοκληρωθεί αλλά θα παρέχει μεγαλύτερες δυνατότητες γενίκευσης των αποτελεσμάτων.

- Η περαιτέρω μελέτη των στοιχείων του προτεινόμενου πλαισίου μέσω της διεξαγωγής μίας μεγάλης κλίμακας ποσοτικής έρευνας.
- Η μετατροπή του προτεινόμενου πλαισίου πρόβλεψης της απώλειας πελατών σε λογισμικό.

Appendix F

Terminology and Notations used in Thesis

F.1 Terminologies Explanation used Throughout Thesis

- **"author":** The term is used to describe the owner of this research work.
- **"researchers":** The term is used basically to describe other authors, which their research works referred during the chapters of this thesis.
- **"traditional data":** This term is used to describe data met in the majority of the churn prediction models, such as demographics, support call log, service usage related etc.,[1, 3, 4].
- **"συγγραφέας":** Ο όρος αυτός αναφέρεται στον συγγραφέα της εν λόγω διδακτορικής διατριβής.

Appendix G

Prediction Models - Explanation of Parameters

G.1 Description of Parameters used in Prediction Models Explanation

- **C: float, default=1.0 -** Regularization parameter.
- **kernel: 'linear', 'poly', 'rbf', 'sigmoid', 'pre-computed' -, default='rbf' -** Specifies the type of the kernel used in the model
- **degree: int, default=3 -** Degree of the polynomial kernel function. Ignored by all other kernels.
- **penalty: ('l1', 'l2', 'elasticnet', 'none'), default='l2' -** Specifies the norm used in the penalization.
- **dual: bool, default=False -** Signals dual or primal formulation. Dual formulation is implemented for l2 penalty with liblinear solver.
- **fit_intercept: bool, default=True -** Specifies if a constant should be added to the decision function.
- **intercept_scaling: float, default=1 -** Useful only when the solver 'liblinear'.
- **l1_ratio: float, default=None -** Elastic-Net mixing parameter, with 0 <= l1_ratio <= 1. Setting l1_ratio=0 is equivalent to using penalty='l2'
- **max_iter: int, default=100 -** The maximum iterations number taken for the solvers.
- **multi_class: ('auto', 'ovr', 'multinomial'), default='auto' -** With the option 'ovr', a binary problem is fit to each label.

Source: [116]