ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Μέθοδοι Επιβλεπόμενης Μηχανικής Μάθησης για την Εκτίμηση της Πιθανότητας Αθέτησης

Χρήστος Κουρούκλης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2024

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Μέθοδοι Επιβλεπόμενης Μηχανικής Μάθησης για την Εκτίμηση της Πιθανότητας Αθέτησης

Χρήστος Κουρούκλης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη Συνέλευση του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή της σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:
- Καθηγητής Σωτήριος Μπερσίμης (Επιβλέπων)
- Αναπληρωτής Καθηγητής Δημήτριος Αντζουλάκος
- Επίκουρος Καθηγητής Σωτήριος Τασουλής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Supervised Machine Learning Methods for Estimating Probability of Default

By

Christos Kourouklis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in *Applied Statistics*

Piraeus, Greece
September 2024

To those whom I love
And to all who cherish the scientific approach

ACKNOWLEDGEMENTS

With the completion of this thesis, I would like to express my gratitude to all the people who stood by my side during my postgraduate studies and supported me through all the difficulties I faced, as the emotional strength they gave me helped me in ways I could not adequately express. Additionally, I would like to especially thank my fellow students, with whom I spent countless hours studying together, who supported me greatly -whom now are good friends to me. Finally, I would like to extend my sincere thanks to my supervising professor, Mr. Sotirios Bersimis, for the continuous interest he expressed and the guidance he provided throughout this thesis. Naturally, I must also express my thanks to the other professors of the postgraduate program whom I had the good fortune to meet, as the scientific knowledge I received from them has helped and continues to help me significantly in my professional and personal development.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία επικεντρώνεται στην εκτίμηση της Πιθανότητας Αθέτησης Υποχρεώσεων χρησιμοποιώντας μεθόδους εποπευτόμενης στατιστικής και μηχανικής μάθησης για δίτιμα προβλημάτα, που είναι απαραίτητες για την αποτελεσματική διαχείριση κινδύνου και τη συμμόρφωση με τους κανονισμούς στα χρηματοπιστωτικά ιδρύματα και για τον τραπεζικό τομέα. Τονίζεται ο κρίσιμος ρόλος της εκτίμησης της πιθανότητας αθέτησης, συζητώντας τη σημασία των χρηματοπιστωτικών ιδρυμάτων στις σύγχρονες οικονομίες, τις προκλήσεις που αντιμετωπίζουν, και τη σημασία της ακριβούς εκτίμησης της πιθανότητας αθέτησης στη μείωση του πιστωτικού κινδύνου και την εξασφάλιση της κανονιστικής συμμόρφωσης, παρέχοντας επίσης μια συνοπτική βιβλιογραφική ανασκόπηση σχετικά με τον πιστωτικό κίνδυνο και την εκτίμηση της πιθανότητας αθέτησης, από παραδοσιακές στατιστικές μεθόδους μέχρι τις πιο πρόσφατες εξελίξεις στη μηχανική μάθηση. Περιγράφεται επίσης το θεωρητικό υπόβαθρο της Λογιστικής Παλινδρόμησης (Logistic Regression), των Τυχαίων Δασών (Random Forest), της Βαθμιαίας Ενίσχυσης (Gradient Boosting), και των Νευρωνικών Δικτύων (Neural Networks), συμπεριλαμβανομένης της διαχείρισης δεδομένων, των μετρικών αξιολόγησης, και των διαδικασιών εκπαίδευσης των μοντέλων. Στο τελευταίο μέρος της διπλωματικής εργασίας, αυτές οι μέθοδοι εφαρμόζονται πρακτικά χρησιμοποιώντας πραγματικά δεδομένα πιστωτικών κινδύνου, όπου γίνονται συγκρίσεις των αντίστοιχων αποδόσεων των μοντέλων χρησιμοποιώντας τις μετρικές Kolmogorov-Smirnov, Gini, και Area Under the Curve (οι μετρικές που χρησιμοποιούνται για την διαχείρηση κινδύνου στον τραπεζικό κλάδο). Η διπλωματική εργασία στοχεύει στη παρουσίαση των πρακτικών αξιολόγησης πιστωτικού κινδύνου μέσω της εφαρμογής, σύγκρισης, αξιολόγησης, και βελτιστοποίησης αυτών των στατιστικών μοντέλων μηχανικής μάθησης.

ABSTRACT

This thesis focuses on the estimation of Probability of Default using binary supervised learning methods, which is essential for effective risk management and regulatory compliance in financial institutions and the banking sector. The crucial role of Probability of Default estimation is emphasized, discussing the importance of financial institutions, the challenges they face, and the significance of accurate Probability of Default estimation in mitigating credit risk and ensuring regulatory compliance, by also providng a concise literature review on credit risk and Probability of Default estimation, from traditional statistical methods to more recent advancements in machine learning. The theoretical background of Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks, including data management, evaluation metrics, and model training processes are also being described. In the last part of the current thesis, these methods are practically implemented using real-world credit data, with performance comparisons made using Kolmogorov-Smironv, Gini, and Area Under the Curve metrics (i.e. banking industry standards). The thesis aims to enhance credit risk assessment practices through the application, comparison, evaluation, and optimization of these statistical machine learning models.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction to Probability of Default Estimation

The field of applied statistics plays a pivotal role in numerous domains, with financial institutions and the banking sector being no exception. In these industries, the estimation of the Probability of Default (PD) using binary supervised (statistical) learning methods is crucial for effective risk management and regulatory compliance. This thesis focuses on exploring and implementing various statistical Machine Mearning (ML) methods to estimate PD, thereby contributing to the enhancement of credit risk assessment practices.

In Chapter 2, we delve into the crucial role of PD estimation in financial institutions and the banking sector. This chapter is structured to provide a comprehensive understanding of the significance of PD estimation. The Role of Financial Institutions in Modern Economies is being discussed, highlighting their integral functions and impact -overall. Special emphasis is given to the challenges that financial institutions face in managing stability and solvency. Additionally, special focus is given on credit risk and the Significance of PD, underscoring the importance of accurate PD estimation in mitigating credit risk. Finally, the fundamental role of PD in regulatory compliance is being depicted, detailing how PD estimation is essential for adhering to regulatory frameworks and maintaining financial health.

Chapter 3 presents a concise literature review on credit risk and PD, spanning from traditional statistical methods to the latest advancements in statistical supervised ML. This review provides a historical perspective and sets the stage for understanding the evolution and current state of PD estimation methodologies.

In Chapter 4, we delve into the theoretical background of the statistical ML methods studied in this thesis. This chapter covers the theoretical foundations of Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Neural Networks (NN). Additionally, we discuss critical aspects of data management, train-test split, feature evaluation, and the segmentation of potential predictors using the Information Value (IV) criterion, in the context of the statistical ML in credit risk. Special attention is given to industry-standard performance metrics such as Area Under the ROC Curve (AUC), Gini Index, and Kolmogorov-Smirnov Separation statistic (KS), which are widely used in the banking sector to evaluate model performance. The chapter also outlines the process of efficiently training ML models, including feature selection, hyperparameter tuning, and cross-validation, as well as the importance of unbiased monitoring of model performance on unseen data.

Chapter 5 involves the practical implementation of LR, RF, GB, and NN resulting in the creation of application scorecards for each model, for credit risk assessemnt purposes. The performance of these models is compared using the KS, Gini Index, and Area Under the ROC Curve (AUC) metric. Furthermore, the IV criterion is used to compare characteristics between training and test sets. Population Stability Index (PSI) test is also conducted to compare the distributions of scores on training against the test data, ensuring the robustness and stability of the developed models.

More specifically, as the financial sector has evolved, so too has the need for sophisticated methods to accurately estimate credit risk. The introduction of statistical learning models in PD estimation has been a significant development. These models not only provide a more precise understanding of risk but also enhance the decision-making process for financial institutions. The financial crisis of 2007–2008 highlighted the importance of robust risk management frameworks, pushing financial institutions to develop statistical and ML models with meticulus attention to detail, aiming for a more accurate estimation of PD, and subsequently for better estimation of potential losses.

Supervised learning methods, such as LR, RF, GB, NN, offer distinct advantages, being however accompanied by some disadvantages as well. By leveraging these such models, banks and financial institutions can accurately predict the probability of borrowers defaulting on loans, ensuring more stable credit portfolios and reducing exposure to high-risk clients. The models rely on past borrower data to identify patterns and relationships between various financial indicators and the risk of default. This data-driven approach has led to more informed credit risk management strategies, allowing institutions to balance their portfolios by minimizing risk and maximizing potential returns.

Moreover, regulatory frameworks, such as Basel III and the International Financial Reporting Standard 9 (IFRS9), have underscored the need for precise PD estimation. Basel III, for example, places strict requirements on banks to maintain adequate capital reserves against unexpected losses, with PD estimation playing a critical role in determining how much capital a bank should hold. Similarly, IFRS9 emphasizes the importance of forward-looking models in estimating Expected Credit Losses (ECL), requiring banks to account for potential losses well in advance of actual default events. This regulatory pressure has accelerated the adoption of advanced ML techniques in the banking sector, where PD estimation has become central to both regulatory compliance and strategic financial planning.

The adoption of ML models in PD estimation is not without its challenges. For instance, ensuring data quality and addressing data imbalances are critical factors that affect model performance. Imbalanced datasets, where default cases are much less frequent than non-default cases, can skew the results of ML models. To mitigate this, financial institutions employ various techniques, such as the fuzzy augmentation technique -inferring the rejected applicants as if their loan application had been approved, and incorporating them into the analyses and to the model development

proceedure. Additionally, model interpretability is another challenge, especially with more complex models like NN. While these models can provide highly accurate predictions, their "black box" nature makes it difficult for stakeholders to understand the decision-making process behind the predictions. This has led to a growing interest in explainable machine learnign techniques that allow financial institutions to interpret the predictions of ML models more effectively.

The need for transparency and interpretability in PD estimation is not just a matter of regulatory compliance but also a business imperative. Financial institutions must be able to justify their credit decisions to both regulators and customers. As a result, while more complex models such as GB and NN often offer superior predictive power, simpler models like Logistic Regression continue to be widely used due to their ease of interpretation -with their performance also reaching at high and accaptable levels. Moreover, this trade-off between model accuracy and interpretability remains a key consideration in the deployment of statistical learning methods for PD estimation.

Looking ahead, the role of statistical learning in PD estimation is likely to grow even more important as financial institutions face increasing pressure from regulators and shareholders to optimize their credit risk management practices. Advances in big data analytics, combined with the growing availability of alternative data sources (e.g. social media data, transaction data), offer new opportunities for enhancing PD estimation models. However, with these opportunities come new challenges, particularly in terms of data privacy and ethical considerations. As ML models become more pervasive in financial decision-making, ensuring that these models are fair, transparent, and compliant with data protection regulations will be critical to their continued success.

To this end, financial institutions are investing heavily in the development of robust governance frameworks to oversee the deployment of ML models. These frameworks are designed to ensure that the models used for PD estimation are not only accurate but also aligned with regulatory requirements and ethical standards. Model validation, for instance, has become a crucial step in the credit risk management process, with institutions employing independent teams to validate the performance of their models before they are deployed in real-world scenarios. This validation process typically involves stress testing the models under various economic scenarios to ensure that they can accurately predict PD even in adverse conditions.

The integration of advanced statistican ML models into the financial sector represents a paradigm shift in how credit risk is managed, with financial institutions embrace a data-driven approach that leverages the power of statistical ML to make more informed, strategic decisions. By combining the strengths of various statistical and ML models, institutions can build more resilient credit portfolios that are better equipped to withstand the uncertainties of the financial markets.

In conclusion, the adoption of statistical learning methods for PD estimation marks a significant step forward in the evolution of credit risk management. As

financial institutions continue to refine their models and integrate new data sources, the accuracy and reliability of PD estimates will only improve. This, in turn, will contribute to greater financial stability and ensure that banks are better prepared to manage the risks associated with lending in an increasingly complex global economy. Finally, this thesis aims to provide a comprehensive analysis of PD estimation methods, offering insights into their theoretical underpinnings, practical implementations, and performance evaluations. By leveraging advanced statistical ML techniques, this work contributes to the ongoing efforts to improve credit risk assessment and management in the financial sector.

CHAPTER 2

The Crucial Role of PD Estimation In Financial Institutions & Banking Sector

The estimation of the Probability of Default (PD) has established as a pivotal tool across the financial sector, underpinning critical decision-making processes with profound implications for financial stability and risk management. Originating in the world of finance, PD estimation has transcended its roots to become an indispensable metric in sectors as diverse as banking, investment, and regulatory compliance. The need for institutions -as well as for the regulatory authorities- to better apprehend the complicated field of credit risk became imperative, as many financial institutions grow big enough to cause global economic crises in case of bankruptcy -being "Too big to fail", since a single financial institution's failure could ignite a chained transmission of the crisis across different countries and continents. As a historical antecedent, PD estimation initially took shape in response to the challenges faced by lenders seeking to assess the creditworthiness of borrowers in order to maximize their profits. However, it has since grown into a multi-disciplinary instrument with far-reaching applications, ensuring the integrity and resilience of financial systems and industries worldwide. The purpose of this chapter is to present and highlight the decisive significance of PD estimation, mainly within the banking sector, as also to expound on its applications and underscore its paramount importance for regulatory authorities in the pursuit of attaining global economic stability.

2.1 The Role of Financial Institutions In Modern Economies

Banks and financial institutions play a profound role in the efficient operation of the economic system in every modern society. In their role as intermediaries, banks manage financial capital, engage in money and capital markets, and gain profits through a series of activities involving deposits, lending and investments. Consequently, banks play a vital role in shaping the factors that influence the overall money supply within an economy, as clearly stated by Saunders Cornett (2017) in their book "Financial Institutions Management: A Risk Management Approach".

Furthermore, the banking sector can be categorized into three main segments: firstly, there are the Central Banks, responsible for controlling monetary policy within the region or country in which they operate. They also oversee and regulate other banks, with ultimate goal of maintaining financial stability and soundness. Secondly, there are the Retail and Commercial Banks, which accept deposits from individuals, households and businesses, and offer various financial services such as wealth management, credit cards, and other retail banking services -by using the received deposits. Additionally,

they are subject to regulations to ensure the safety of deposits and fair lending practices. In return, depositors receive interest payments over time based on a pre-agreed interest rate (a stable interest rate) or on an interest rate that fluctuates. Finally, there are other financial institutions such as Investment Banks, Mortgage Banks, Cooperative Banks, Agricultural Banks, and Shipping Banks. These institutions focus on activities such as raising capital, asset management and trading various financial instruments like bonds, stocks, derivatives, and currencies. They generate profits through fees and commissions, interest income, trading gains, and other sources depending on their activities -each within their respective areas of expertise and influence.

The financial system can be delineated into two primary sub-segments: the Money Market and the Capital Market (Sapountzoglou G., Pentotis C. (2017)). In the Money Market segment, activities encompass the buying and selling, as well as asset management of various financial instruments such as foreign exchange transactions, treasury bills, banking bonds, repos, alongside short-term borrowing and lending, certificates of deposit, and commercial paper. These activities involve a diverse array of participants and are characterized by high levels of liquidity, leading to low default rates. Conversely, the capital market segment involves activities such as the trading of stocks, fixed and fluctuating-rate bonds, convertible bonds, derivatives, commodities, and real estate investment trusts (REITs). Unlike the money market, the capital market exhibits lower levels of liquidity and higher default rates, largely due to the relatively higher variance in the prices of negotiable assets. Furthermore, financial markets can be divided into two sub-segments, namely the primary market and the secondary market. The primary market involves the origination and initial offering of new assets by companies, governments, or other entities to raise capital, whereas the secondary market encompasses the trading of pre-existing assets among investors, acquired initially in the primary market.

Banks and financial institutions participate in various ways in all the aforementioned activities. More specifically, banks carry out the following functions:

- Serve as intermediaries connecting savers with borrowers.
- Accept deposits from the public and provide loans. This ongoing process not only improves the efficiency of the financial system but also safeguards depositors.
- Provide financing for individuals, businesses, and governments.
- Oversee their assets, which are supported by both short-term and long-term liabilities (including deposits, bank loans, and equity).
- Effectively manage and oversee payment settlements.
- Conduct checks to prevent fraudulent activities.
- Gather, process, and manage information.

- Operate within an environment characterized by uncertainty and various risks.

As can be seen, the intermediary role of banks and financial institutions is extremely important for the efficacious operation of the modern economic system. Nevertheless, the borrower typically possesses a better understanding of their own ability to fulfill their loan obligations compared to their lender and also are familiar with the financial aspects and characteristics of the institution providing the loan. This information imbalance between the borrower and lender is known as the “asymmetric information” problem. The challenge is that banks, which act as intermediaries, work hard and spend a lot of money to get the right information about borrowers, as well as to process this information and obtain significant insights about their level of credibility. The imperfections in financial markets are mainly attributed to this situation of unequal access to information between those seeking loans and those providing them. Imperfections in financial markets arise not only from information costs but also from transaction costs. These inherent imperfections in financial markets play a significant role in justifying the existence of the banking system and its role as an intermediary, since banks and financial institutions obtain the expertise and known-how to mitigate these potential issues, when getting involved into such activities.

Due to the problem of “asymmetric information”, banks and financial institutions are confronted with two more practical issues; “the adverse selection problem” and the “moral hazard problem”. The first problem refers to the potential profit loss that financial institutions and banks may face by lending money to riskier clients who are more likely to default or by rejecting safer clients, both as a result of inaccurate risk assessments. The latter, refers to a situation in which the borrower after obtaining a loan, becomes involved in undisclosed high-risk activities that negatively impact their financial stability, this subsequent moral hazard decreases the probability of loan repayment -thus increasing the probability of a defaulting event being occured.

Therefore, the need for an accurate estimation and quantification of this associated risk has risen for financial institutions, in order to make informed-based decisions and to minimize potential losses. (Atzoulatos A. (2011)).

2.2 Risk and problems encountered by Financial Institution

Before moving on to the presentation of PD and its indispensable role in quantifying credit risk, it is essential to provide a brief overview of the various types of risks that financial institutions and banks encounter. These risks are deeply interconnected, and an inaccurate assessment of credit risk could increase or introduce other potential risks that may ultimately lead to the failure of the financial institution (Saunders C. (2017)).

Nearly all the activities in which fixnancial institutions and banks engage involve various types of risks. These risks are presented epigrammatically below:

- Interest rate risk

- Credit risk
- Liquidity risk
- Foreign exchange risk
- Country sovereign risk
- Market risk
- Off-balance sheet risk
- Technological risk
- Operational risk
- Insolvency risk

All the mentioned risks are inherent to the activities of financial institutions and are interconnected with each other in various ways. A concise overview of all these risks will be presented, as the interconnection and subsequent interaction between them is perhaps the most intricate and complex challenge faced by banks and financial institutions.

- Interest rate risk: This risk arises from fluctuations in interest rates. Financial institutions may face losses if the interest rates change unfavorably, affecting the value of their assets and liabilities.
- Credit risk: This is the risk that arises for potential inability or denial of borrowers or counterparties to fulfill their financial obligations, resulting in losses for the lender or investor. Almost all kind of financial institutions encounter this risk.
- Liquidity risk: It involves the risk that an institution may not have enough cash or liquid assets to meet its short-term financial obligations. It may lead to financial distress or bankruptcy.
- Foreign exchange risk: Changes in currency values can affect the value of assets and liabilities denominated in different currencies. Hence, this risk occurs when financial institutions have exposure to fluctuations in exchange rates, by maintaining or investing in assets of foreign currencies.
- Country sovereign risk: This risk relates to investments in foreign government or corporal securities. It involves the possibility of a country defaulting on its debt or implementing policies that negatively impact investments. Thus, this is a different type of credit risk for financial institutions.
- Market risk: Arises from the trading of assets and liabilities due to fluctuations in interest rates, exchange rates, and the values of financial assets. It occurs when financial institutions trade their assets and liabilities instead of holding them for long-term investments and for risk mitigation.
- Off-balance Sheet risk: This refers to potential losses from contingent liabilities and commitments that are not reflected on a company's balance sheet. These obligations can become actual losses. Includes all the activities that create potential assets and liabilities, which result in a potential future position for the

balance sheet of the financial institution.

- Technological risk: It encompasses the risk of disruptions or vulnerabilities in a financial institution's technology infrastructure, including cyberattacks, data breaches, or system failures.
- Operational risk: This risk arises from internal processes, systems, people, and of the external environment of the financial institution. It includes risks related to errors, fraud, legal compliance, and business interruptions. Reputation and Strategic risk can also be included in this segment, as a wider definition of what operational risk is.
- Insolvency risk: It is the risk that a financial institution becomes insolvent, meaning it cannot meet its financial obligations by mitigating risks produced by a sudden and/or sharp decrease on the values of the assets that a financial institution holds. Hence, the financial institution may face bankruptcy or regulatory intervention.

While these risks are often discussed in isolation, they are intertwined in complicated and various ways. For instance, an abrupt raise in interest rates can cause turbulences through both business and consumer sectors, disrupting their ability to fulfill their financial contractual commitments. This interdependence sets in motion chained reactions that impact credit risk, interest rate risk, and off-balance-sheet risk, creating a dynamic landscape of interconnected challenges, all due to the interactions between the aforementioned individual risks.

Furthermore, a financial institution's capital is intricately linked to the capital of its counterparties, forming a symbiotic relationship crucial for effective liquidity management. Consequently, liquidity risk is tightly intertwined with credit risk and interest rate risk. When a client fails to meet its obligations, the consequences extend beyond the financial institution, affecting earnings, profits, and ultimately, the institution's overall capital position. As a result, every risk and its multi-level interactions with every other aforementioned risk, become pivotal in shaping the risk of insolvency, which may lead to bankruptcy.

Similarly, there is a strong correlation between fluctuations in foreign exchange rates and interest rates. Exchange rates may undergo changes when the central bank adjusts a key interest rate through monetary policy. Additionally, various other discrete or random risks can impact the profitability and exposure of a financial institution. These unique risks may include external events, such as sudden shifts in regulatory legislation, which can positively affect certain risk factors while negatively impacting others, with their combined effect potentially leading to an overall negative outcome for the financial institution's capital position.

Other more discrete or random risks include sudden and unexpected changes in conditions prevailing in financial markets due to a war, an uprising, or a sudden market collapse, such as the stock market crash of 2008 and the consequent global financial

crisis, the COVID-19 pandemic, the Russia-Ukraine war. These changes can have a significant impact on the risk exposure of a financial institution. Other risks of discrete or random events include fraudulent activities, extreme weather events and mismanagement. All of these could ultimately cause significant damage or even the insolvency of a financial institution. However, each of these risks is difficult to predict and be modeled. Finally, more macroeconomic or systemic risks, such as inflation's volatility and unemployment, can directly and indirectly affect the risk exposure of a financial institution to interest rate risk, credit risk, and liquidity risk. For example, the Greek unemployment rate had more than tripled between 2008 and 2012 -from 7.6 % in 2008 to 24.2% in 2012, the number of the unemployed increased from 364.3 thousand in 2008 to 1 million and 172.2 thousand in 2012. In the summer of 2013, the unemployment rate exceeded 28%. With such high unemployment, the exposure of all the Greek financial institutions to credit risk increased dramatically -as too many borrowers could not repay their loans after losing their jobs and subsequently their sources of income. (Saunders C. (2017)).

2.3 Credit risk and Probability of Default Significance

Credit risk arises from the potential inability of borrowers or counterparties to fulfill their financial contractual obligations and is the primary financial risk faced by credit institutions, as it is directly related to their core activity, namely, the provision of loan capital. These obligations may refer either to the repayment of a loan granted by the credit institution or to the regular payments arising from the issuance of a bond in which the credit institution has invested. Consequently, all financial institutions closely monitor the performance of portfolios associated with credit risk, in order to maximize their profits and be ready to mitigate the loss -in case a credit event arises. The "Too big to fail" financial institutions (i.e. "SSI – Systemically Important Institutions") are obligated by regulatory authorities to operate within and not deviate from some industry stands, as well as to closely monitoring credit risk and report the overall status of their portfolio, as a potential bankruptcy could trigger economic destabilization on the continent where the financial institution operates. Even worse, the economic crisis could spread to other countries, resulting to a generalized economic recession among countries and continents (Atzoulatos A. (2011)). The credit risk regulations which are applied by the authorities, will be analyzed in the next sub-section "2.4 The Fundamental Role of PD in Regulatory Authorities". For the rest of this sub-section, a coherent elaboration on credit risk and PD will be made.

The drive for credit expansion and fierce competition among financial institutions can occasionally lead to the financing of individuals and legal entities that fall short of fundamental solvency requirements—that is, their ability to meet long-term financial obligations. Essentially, these entities are unable to guarantee the repayment of their debts. This situation increases the risk of losses for credit institutions, as it raises the probability of default within their loan portfolios and reduces the chances of recovering the initially loaned capital—or even a portion of it (Sapountzoglou G.,

Pentotis C. (2017)).

Financial institutions aim to avoid potentially "bad" borrowers -those that eventually will not repay the borrowed amount to its full, and on the contrary, they want to invest in "good" applicants -those that will eventually fully repay the borrowed amount. This segmentation between the populations of "good" and "bad" applicants is crucial in the context of credit risk -as well as for all the financial institutions that encounter credit risk (Saunders C. (2017)). More precisely, there are many definitions for the categorization of a borrower into "good" or "bad", and this categorization depends directly to the borrower's number of days past due of each corresponding credit obligation. However, specific guidelines were given by the EBA (European Banking Authority) for the assessment & classification of a borrower (natural or legal entity) as "good" or "bad", and hence constructing the most up-to-date default definition across all firms that are subjected to the Capital Requirements Regulation (CRR) and are required to hold capital against credit activities. Regulation authorities such as EBA and NCAs (i.e. National Competent Authorities) have adopted the new default definition (PWC – EBA Credit Risk: Default Definition (2017)).

The definition of default varies for retail and non-retail financial institutions (EBA 2020/978, (2020) http://data.europa.eu/eli/guideline/2020/978/oj). It is based on the overdue period of more than 90 consecutive days, with additional materiality rules:

- For retail financial institutions, a borrower is considered as defaulted if the overdue amount is equal to or exceeds 100€ and/or if the total overdue amount exceeds the 1% of the borrower's total assets.

- For non-retail financial institutions, a borrower is considered as defaulted if the overdue amount is equal to or exceeds 500€ and/or if the total overdue amount exceeds the 1% of the borrower's total assets.

In both cases, breaching these specified thresholds (i.e. more than 90 days past due and/or breaching a materiality threshold) identifies the borrower as defaulted. Therefore, the PD is a measure that quantifies how likely it is for a borrower to become classified as defaulted.

It should be noted that the failure to meet an obligation and the subsequent classification of the borrower as defaulted does not always lead to negative consequences for the credit institution. There are situations in which, despite the initial default event, the borrower will eventually fully repay all its contractual obligations.

In a bank's credit policy, the approval of loans primarily depends on the borrower's creditworthiness. Additionally, the quality of any potential collateral or guarantees that come with the loan are considered. To this direction points the fact there is always some remaining risk, regardless of the size of the collateral. However, it would be a mistake to overlook the size and quality of collateral that accompanies a bank's claim because these often play a crucial role in recovering the provided capital

if the borrower faces bankruptcy. (Sapountzoglou G., Pentotis C. (2017)).

Common types of guarantees that come with the requirements of banks include:

- Collateral
- Guarantees from third parties
- Credit derivatives

Collateral itself is divided into two categories: physical and financial. Physical collateral includes assets like real estate properties and land. Financial collateral encompasses assets such as stocks and other investment securities. The role of collateral extends beyond merely reducing potential credit losses in case of borrower default; it also serves as a deterrent for borrowers to default on their obligations. Guarantees involve shifting the responsibility for repaying the debt to third parties (co-debtors) when the original borrowers are unable to fulfill their obligations. Providing guarantees from third parties substantially reduces the PD, as a full default only happens when both the borrower and the guarantor cannot meet their obligations. Lastly, credit institutions use credit derivatives to ensure that in specific credit events, such as a borrower's bankruptcy, they receive compensation from the counterparty to cover some or all of their credit losses. In practice, assessing collateral and guarantees usually involves estimating a Recovery Rate (RR) -this rate indicates the portion of the credit exposure (the lender's claim) that can be recovered if the borrower defaults. Estimating the recoverable capital in bankruptcy is not always certain, as actual recoveries may significantly differ from initial estimates due to factors like undervaluation of physical collateral (like real estate) or difficulties in enforcing claims against guarantors. Hence, collaterals convert proportion of the credit risk into legal and valuation risk, while third-party guarantees reduce the probability of bankruptcy.

Credit rating agencies have drawn conclusions about the potential recovery of capital from defaulted securities by analyzing historical data related to the type of loan, issuer’s creditworthiness, and the historical values of defaulted debt securities. It has been found that -in practical terms- senior debt securities (i.e. securities that have a higher repayment priority) typically exhibit high capital recovery rates in case of default, such as bonds issued by governments. Conversely, junior debt securities (i.e. collaterals with lower repayment priority) offer limited capital recovery potential to their holders. Additionally, secured bonds backed by physical assets such as land and real estate provide a substantial degree of capital recovery assurance in the event of issuer bankruptcy. On the other hand, unsecured debt (i.e. debt bound to no collateral) recovery rate relies directly to the issuer's financial health. Lastly, subordinated debt holders are at the bottom of the hierarchy in case of issuer bankruptcy, will be repaid only if the claims of higher-ranking claims have been satisfied. Recovery rates in the context of debt issuances are defined as the ratio of the bond's post-bankruptcy price to its face value (par value). During economic downturns, recovery rates for defaulted debt securities tend to decrease significantly, whereas the opposite trend occurs during periods of economic prosperity. Thus, there has been a robust positive correlation between recovery rates and the overall economic conditions.

When a banking institution evaluates its exposure to credit risk, it focuses on assessing potential losses that might occur within each segment of its portfolio. These potential credit losses are being calculated on a regular basis. The estimation of Expected Losses (EL) involves the estimation of PD , Loss Given Default (LGD), RR and the Exposure at Default (EAD) -the total amount in which the financial institution is exposed at the time of default.

Hence, EL represents the estimated average losses that the credit portfolio is expected to face. The EL of a portfolio is computed by multiplying the PD by the LGD and further by the EAD. This results to the following equation:

$$EL = PD \times LGD \times EAD$$

where LGD = 1 – RR, RR is the recovery rate of the corresponding exposure.

It should be noted that EL, which a financial institution includes in its risk management strategy, should not be considered a risk for the financial institution itself. This amount is regularly estimated and, therefore, is regarded as a known figure. The financial institution is required to have the capability to efficiently handle any potential adverse events. In addition to EL, financial institutions and banks should also assess the level of Unexpected Losses (UL), which essentially represents the true risk faced by financial institutions.

UL are defined as the maximum potential losses that a financial institution's portfolio may encounter under a certain degree of uncertainty. To quantify UL, two additional components have been introduced: Value at Risk and Expected Shortfall. (Boutsikas M. (2023)).

Value at Risk (VaR) serves the purpose of determining the most adverse potential loss an investment can incur and the corresponding amount of capital required to safeguard against it. It's important to note that this worst-case scenario typically involves complete loss of the invested amount, an event with an exceedingly low probability that would result into a huge reserve, often unrealistic. Consequently, VaR tries to answer the following question: What represents the most unfavorable conceivable outcome for the investment, under a certain level of uncertainty. The maximum potential loss for a certain level of uncertainty -set by the researcher, denoted as VaR, corresponds to the loss “L” linked to an investment characterized by stochastic gains, let $\Pi = -L$. This metric signifies the highest conceivable loss that may transpire within the least favorable interval of possible investment outcomes. In practice, VaR is the most commonly used risk measure, but it does come with notable limitations. For instance, it does not offer insights into the magnitude of potential losses should they surpass the Value at Risk (VaR) for a specified confidence level. This is why another risk metric, Expected Shortfall, finds practical application. Expected Shortfall, in essence, signifies the average loss when it falls within the range of more extreme scenarios, given that the losses have exceeded the VaR.

To conclude, following the above analyses it can easily be seen that PDs' estimation plays a pivotal role in assessing and managing credit risk within a financial institution. Both EL and UL are fundamental components in this context. EL represents the anticipated average losses that the credit portfolio may encounter and its calculation heavily relies on the sound estimation of PD. It provides insight into the regular, expected losses that the institution should be well-prepared to handle efficiently. On the other hand, UL signifies the maximum potential losses under certain levels of uncertainty, particularly in extreme scenarios, which is also heavily influenced by the estimation of PD, since the accurate estimation of PD results to a more accurate estimation of EL -and the subsequent accurate estimation of UL. Hence, a comprehensive framework is formed that aids financial institutions in understanding, measuring, and mitigating the multilevel and complex credit risks they face.

2.4 The Fundamental Role of PD in Regulatory Compliance

Two main pillars that comprise the regulations of the financial institutions are "Basel III" (i.e. updates of Basel I & II) and "IFRS9". In simple terms, IFRS9 builds a framework in which all entities that hold financial instruments, keep track of their possessions (i.e. account-wise) and portfolio's structure (i.e. model-wise) correctly and in transparency, whereas Basel rules strive for making banks not taking too many risks with their money, by maintaining a safety cushion ("Required Reserves") in case of economic turndown. Therefore, to emphasize the importance of accurate PD estimation, it is deemed suitable to offer a concise overview of the essential regulatory standards that financial institutions must strictly follow concerning credit risk, as mandated by regulatory authorities in the context of credit risk management.

IFRS 9 (International Financial Reporting Standard 9): IFRS9 replaced IAS39 (International Accounting Standard 39) and was first issued by the International Accounting Standards Board (IASB) in July 2014. It was introduced to address some of the deficiencies in IAS39 and to improve the accounting for financial instruments, particularly in response to the global financial crisis of 2007-2009, as well as to enhance transparency and comparability in financial statements. The standard became effective for annual periods beginning on or after January 1, 2018. (PWC: IFRS 9 impairment - Significant Increase in Credit Risk, 2017). IFRS9 regulations and basic concepts regard to:

- Classification and measurement of financial instruments
- Hedge accounting
- Impairment

For the "classification and measurement of financial instruments" and "hedge accounting" segments, only a brief summary will follow, since they refer to accounting modifications which are out of the scope of this thesis. However, the "impairment" segment regards to modifications in ECL modeling, and as described in the previous sub-section, PD plays an extremely crucial role for efficacious implementations and

estimations to be made.

The IFRS9 regulations are briefly described below:

- Financial instruments are classified and measured as follows:
    - Amortized cost: For stable, low-risk investments like bonds and loans held until maturity.
    - Fair Value through Other Comprehensive Income (FVOCI): For assets with the intent to be held for a while but may be sold in the future, like certain debt securities.
    - Fair Value through Profit or Loss (FVTPL): For assets primarily held for trading, such as stocks and derivatives.

- Hedge accounting under IFRS 9:
    - IFRS 9 improved hedge accounting to help companies manage financial risks more accurately and transparently. It allows them to offset the impact of market changes, like currency or interest rate fluctuations, on their financial statements, aligning their accounting with risk management strategies. This makes financial reporting clearer and more aligned with a company's actual financial health.

- Impairment:

    - Financial institutions are required to account for expected credit losses on financial instruments, considering both the PD and the overdue amount at the time of the default event -that is the magnitude of loss. Before the implementation of IFRS9, the accounting standards were based on an “Incurred Loss” model. The “Incurred Loss” model was used by banks and financial institutions in the past-recognizing impairment losses on financial instruments when there was evidence that a loss had already taken place (i.e, an estimation-given-default event approach), which often resulted in delayed recognition of credit losses. With IFRS9, the approach to recognizing credit losses shifted to an "Expected Credit Loss" -instead of the “Incurred Loss” model. Hence, under IFRS9, financial institutions need to predict and account for losses at all times, taking into account past events, current conditions, and future predictions. This is a more forward-looking approach and results into a much quicker recognition of losses, compared to that of the “Incurred Model” which predicted too little and too late.

Below are presented the main components of the IFRS9 approach -some of which have already been introduced in the current thesis:

- PD: Entities need to assess the probability that a borrower will default on their obligations. This is the PD which has already been described. The PDs that are calculated are the 12-month PDs or the Lifetime-PDs. The 12-month PD is the probability assigned to each potential or current borrower to default in the next-

12month period. On the other hand, the Lifetime-PD is the probability assigned for the whole lifetime of the asset, that is the period till maturity. For example, for a 20-year mortgage loan, the Lifetime-PD would be the probability of the borrower defaulting within the next 20 years, if calculated at the asset's origination date. In contrast, the 12-month PD would estimate the probability of default within the next year, without considering the remaining 19 years.

- LGD: The percentage of the exposed amount that the financial institution estimates that will not be recovered.
- EAD: In addition to the PD, the ECL model also considers the magnitude of loss that might be incurred if a default occurs. This involves estimating the Exposure at Default -that is the total overdue amount at the time of the default event.
- ECL: Under the ECL model, banks and financial institutions are required to consider the future credit losses that are expected to occur over a 12-month period (12-month ECL) or over the life of a financial instrument (Lifetime ECL), rather than waiting until a loss has actually happened.

There are 3 stages of impairment for each financial instrument -and each subsequent exposure is classified in one of these 3 stages -based on the impairment recognition. Specifically:

- Stage 1: By the time an exposure is recognized (i.e., origination date), the ECLs that may occur from potential defaults within the next 12 months are calculated. Hence, for each existing exposure, the financial institution is obliged to estimate the corresponding credit losses. For each subsequent reporting date (i.e. for each subsequent month), accounts that maintain their credit stability (i.e. no SICR - Significant Increase in their corresponding Credit Risk is occurred since their initial recognition), also have the 12-month ECLs applied to them. Finally, when it comes to the estimation of the interest revenue originating from an exposure classified as Stage 1, the full amount of the exposure is used (i.e. Gross amount) -that is without deduction of the ECL.

- Stage 2: These accounts are underperforming or having a specific event occurred, resulting to significant increase in credit risk. Hence, their corresponding ECL is estimated over the expected lifetime of the asset (Lifetime ECLs). The interest revenue is calculated as in Stage 1 (i.e. over the full amount of the exposure – gross amount), however the ECL are calculated by using the Lifetime-PD (and not the 12-month PD).

- Stage 3: In this category belong all the non-performing exposures, having objective evidence of impairment, and thus the financial instrument is considered as credit-impaired. The lifetime ECLs are calculated as in Stage 2, however interest revenue is determined using the exposures' amortized cost (i.e., the gross carrying amount minus the loss allowance – or else net carrying amount).

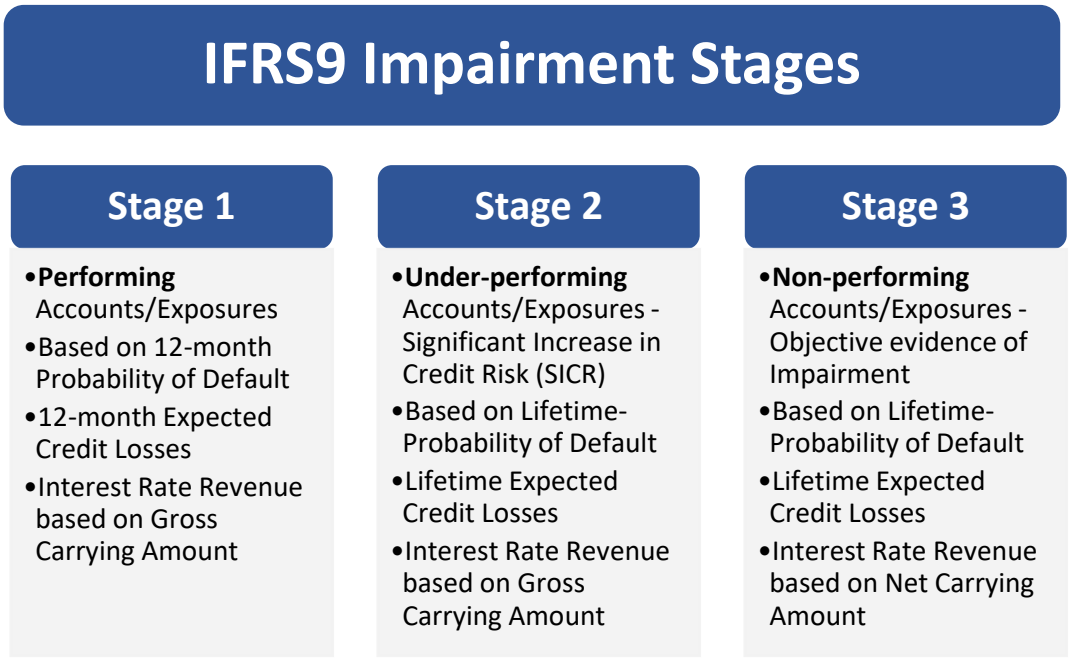


Table 2.4.1: IFRS9 Impairment Stages

It should be noted that the term "Significant Increase in Credit Risk" is not strictly defined under IFRS9, making it a judgmental area (PWC: IFRS 9 impairment - Significant Increase in Credit Risk, 2017). To assess it, entities need to consider specific factors based on the financial asset and how they manage credit risk. More specifically, the factor that give shape to SICR estimation are classified as quantitative & qualitative indicators that have to be assessed. A concise summary is presented below:

- Quantitative indicators:
    - Comparison of 12-month and Lifetime-PD: This involves comparing the 12-month PD to the Lifetime-PD. Generally, a Lifetime-PD is used, but a 12-month PD may suffice if changes in it reasonably approximate changes in the Lifetime-PD -since LPD leads to a more accurate estimation, though being more difficult to be modelled.
    - Changes in Lifetime-PD instead of monitoring the volatility of ECL: IFRS 9 requires assessing the significant increase in credit risk (SICR) based on changes in the Lifetime-PD over the assets' remaining years to maturity and not based on changes in the ECLs. Hence, if a PD model is used, the PD measure is used and not the LGD.
    - Relative assessment: It's a relative comparison between the PD at the reporting date and the PD at the initial recognition date. For example, if the PD of a loan is estimated at 1% at the initial recognition date, and the next month (i.e. reporting date) is estimated at 2%, it exhibits a percentage increase of 100%. Absolute assessments, in combination with the use of a PD threshold, will not suffice -unless it in-line with the relative assessment approach.
    - Residual life of instrument: At the initial recognition date (i.e. initial

date) of an exposure, the Lifetime-PD is estimated. It is generally and conceptually expected that for the performing exposures the Lifetime-PD will decrease month-by-month (i.e. for the subsequent reporting dates), as the remaining months till maturity will be reduced. At the origination date, additional Lifetime-PDs are estimated and assigned to all the corresponding upcoming months (i.e. upcoming reporting dates) till maturity. The Lifetime-PD of each reporting date is compared with the corresponding initial expectation for the Lifetime-PD. In other words, instead of comparing the currently estimated Lifetime-PD of each month with the Lifetime-PD of the starting date of the exposure, the comparison is made with the Lifetime-PD that was expected at the beginning for the corresponding month. Hence, the change in the Lifetime-PD is assessed by comparing the remaining Lifetime-PD at the reporting date with the Lifetime-PD that was expected at initial recognition for that point in time. Since PD often decreases as time passes, the initial Lifetime-PD might not capture the SICR.

- Variability with initial risk: What is considered a significant change depends on the initial risk (i.e. initially estimated Lifetime-PD). For example, assuming that a loan or bond was initially assessed with a 0.5% PD, and this PD later changes to 1%, resulting to double the initial risk as well as to a percentage increase of 100%. In contrast, assuming that the instrument started with a 4% PD and then increased to 4.5%, the absolute increase remained the same (i.e. 0.5%), whereas the percentage increase stood at 12.5%, which is significantly lower than that of the previous example, nevertheless the risk level is greatly higher on the latter. Therefore, when all other factors remain constant, the threshold for what is considered a significant change in default risk should be smaller for the high-grade instrument compared to the lower-grade ones.
- Stage 2 monitoring: Instruments in Stage 2 should be monitored for a significant increase in credit risk. Not all exposures in Stage 2 will transit into Stage 3, as some may revert to Stage 1. However, an underperforming exposure indicates that the probability of a default event increases, since the borrower has already shown a potential inability to meet the contractual obligations, by generating a material overdue amount.
- Reasonable thresholds: Judgment is used in determining what threshold is significant, thus creating a balance between recognizing expected losses too early or too late. More specifically, financial institutions do not want to set the risk (i.e. PD) threshold too low because it might trigger the recognition of expected losses too frequently, leading to unnecessary adjustments and increased volatility in financial reporting. On the contrary, setting the risk (i.e. PD) threshold too high will increase the probability of recognizing expected losses too late, meaning waiting

too long to account for potential future defaulted exposures, thus resulting to financial statements that do not accurately reflect the true credit risk.

- Qualitative indicators: Qualitative factors, separate from quantitative assessment, should be considered if relevant. These include factors mentioned in IFRS 9.
- Backstop indicator: If contractual payments are more than 30 days past due, there's a presumption of a significant increase in credit risk. This presumption can be rebutted with reasonable evidence to prove the contrary.

The EL model encourages financial institutions to be proactive in recognizing potential credit losses and to set aside provisions for these losses. This helps in providing a more accurate and forward-looking representation of an entity's financial health and risk exposure.

Basel I, Basel II, and Basel III:

These are international banking supervisory frameworks developed by the Basel Committee on Banking Supervision (BCBS). They were created in response to financial crises to ensure the stability and safety of the banking system. (Capgemini, The ABCs of Basel I, II & III. (2014)). Below follows a brief overview of each:

- Basel I (released rule in 1988):
    - Basel I introduced the concept of minimum capital requirements for banks. It required banks to maintain a minimum capital adequacy ratio of 8% of their risk-weighted assets. Assets were categorized into broad risk buckets, and capital requirements were set accordingly -based on the risk category in which the asset was classified.
    - It was a relatively simple framework, primarily focused on credit risk.
- Basel II (released rule in 2007):
    - Basel II was more comprehensive and risk-sensitive compared to Basel I. It introduced three pillars:
        1. Minimum capital requirements: Similar to Basel I but with more complex risk calculations.
        2. Supervisory Review Process (SRP): Encouraged banks and regulators to assess their risk management practices and set additional capital requirements based on internal assessments.
        3. Market discipline: Disclosures and transparency requirements to make banks more accountable to stakeholders.

Basel II introduced more advanced risk modeling techniques and allowed for a more detailed assessment of credit, operational, and market risks.

- Basel III (released rule in 2013):

    - Basel III was developed in response to the 2007-2009 global financial crisis. It significantly strengthened the regulatory framework by addressing various shortcomings. Key components include:

        1. Common Equity Tier 1 (CET1): Banks must maintain a higher proportion of common equity as a percentage of risk-weighted assets.

        2. Liquidity Coverage Ratio (LCR): Banks must maintain a minimum amount of highly liquid assets to meet short-term liquidity -if needed.

        3. Leverage ratio: Introduced to prevent excessive leverage.

        4. Counterparty credit risk: Improved measurement and management of counterparty credit risk.

        5. Risk management and supervision: The framework emphasizes improved risk management practices, enhanced supervision, and stress testing to identify and mitigate potential risks in the banking system.

        6. Systemically Important Banks (SIBs): Additional capital requirements and stricter regulations for globally systemically important banks (G-SIBs).

Basel III aimed to enhance the resilience and stability of the global banking system by addressing capital, liquidity, and risk management concerns.

In summary, IFRS 9 focuses on accounting, reporting and in modelling risk for financial instruments, while Basel I, Basel II, and Basel III are international banking regulatory frameworks that evolved over time to strengthen the stability and safety of the banking system by imposing more stringent capital, liquidity, and risk management requirements on banks. Both IFRS 9 and Basel III are important components of the global financial regulatory framework, contributing to the transparency, stability, and resilience of financial institutions and markets. It is apparent that financial institutions and banks, when engaged in activities associated with inherent credit risk, must assess it and quantify it as accurate as possible. This involves estimating the PD, and among others, the LGD & the EAD -which have already been described in the previous sections. By obliging financial institutions to operate by making data-oriented decisions, both financial-stability will be maintained within the economies they operate -a goal that is in accordance with their profits' maximization, and additionally, potential economic recessions will be mitigated, promoting sound investments, with the overall economic growth being fostered. In doing so, they contribute to the enhancement of

societies' standards of living. To this extend, as described in this chapter, assessing and estimating accurately the of PD is pivotal as to attain all the aforementioned goals.

CHAPTER 3

A Concise Literature Review on Credit Risk and PD

In this chapter, a concise literature Review on Credit Risk and PD will be presented, by pinpointing the importance of some specific topics and scientific research that have been done throughout the years that have contributed and given shape to the latest statistical and more complex ML methods which are used for the estimation of PD nowadays.

3.1 From Markowitz Portfolio Theory to More Sophisticated Statistical Models

For too many years, financial institutions heavily relied on subjective analysis or so-called "expert" systems, where bankers assessed credit risk based on borrower characteristics like reputation, leverage, earnings stability, and collateral, known as the "4 Cs" of credit, which led to scaled rating. As shown by Sommerville R. and Taffer R. (1995) in their work “Banker judgement versus formal forecasting models: The case of country risk assessment”, these subjective ratings tended to be overly pessimistic, prompting a shift towards more objective methods. However, the transition from relying solely on "expert judgment" to incorporating statistical models for measuring credit risk did not happen overnight and took several years and scientific studies. This shift was gradual and part of a broader movement towards quantitative risk management techniques that gained momentum in the late 20th century. The development and adoption of these statistical models were also influenced by several other factors, including advances in computing technology, the increasing availability of data, and evolving regulatory requirements.

The main problem in the approval/rejection process of loan applications -in terms of modeling- that is to approve the potential “good” borrowers (i.e. those that will eventually fulfill their contractual agreement) and subsequently reject applications of potential “bad” ones (i.e. those that will eventually not fulfill their contractual agreement) by maintaining the risk at low levels, has its roots back in the 1950s. Specifically, Markowitz H. (1952) through his groundbreaking work and through his book (Portfolio Selection: Efficient Diversification of Investments, 1959), established the foundations of modern portfolio theory. Though not studying or measuring credit-risk directly, his pioneering book and article were the first to systematically present the concepts of the portfolio optimization theory in general, laying the groundwork for future advancements in the field. Markowitz introduced the idea of approaching portfolio construction based on the mean and variance of a collection of assets. He established a key theorem; the mean-variance portfolio theory -which involves minimizing variance for a given level of expected returns, thus creating an efficient

frontier, offering investors a range of portfolios to select from based on their individual risk and return preferences. A crucial insight from Markowitz's theory is the need to consider the correlation of assets rather than evaluating them in isolation, hence, by acknowledging how each security's movements are correlated with that of others, investors can construct portfolios that achieve the desired returns with lower risk compared to portfolios that did not take into account these correlations. Additionally, Markowitz's theory emphasized the importance of diversification in reducing portfolio risk. This principle is crucial in credit risk management, where diversifying the types of credit exposures (across different sectors, geographies, borrower types) can reduce the risk of significant losses from defaults, while also aiming for the best possible returns (risk-return trade-off). Markowitz's modern portfolio theory not only revolutionized investment management but also laid the conceptual and methodological foundations for significant advancements in credit risk management.

Fair & Isaac Company - FICO (1956), introduced the FICO Score, that was the first risk score developed to evaluate credit risk, as stated by Amos T. (2019). Designed as a generic score, it assesses the risk level of potential or existing customers using information from their credit files held by credit bureaus or credit reporting agencies like Experian, Equifax, or TransUnion in the U.S., Tiresias Bank Information Systems SA in Greece. Although many vendors have since developed similar generic, one-size-might-fit-all risk scores, there is an increasing trend towards developing customized in-house risk scores due to a variety of influencing factors. For the following years, the literature on measuring credit risk mainly focused on the creation and evaluation of univariate models -that compared borrowers' accounting ratios with industry norms, and multivariate models -that combined and weighted these variables to produce credit risk scores or default probability measures. If these metrics exceeded a critical risk level threshold -that corresponds to a certain probability threshold, loan applications were either rejected or subject to closer scrutiny. Among the techniques and models studied during that time, three principal techniques gained a lot of attention and used in multi-variable credit-scoring systems: logit and probit models (part of the generalized linear models family), and discriminant analysis. Between the aforementioned statistical models, discriminant analysis and the logit model established as leading methodologies, not only due to the accuracy of their predictions, but also for their interpretable results.

The Z-score model, developed by Altman E. (1968) and based on discriminant analysis, employs a formula derived from a study of 66 companies, half of which had declared bankruptcy. This formula assesses a company's financial stability or ability to remain solvent, using a metric referred to as the Z-score:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

where

X_1 = Working Capital / Total Assets

X_2 = Retained Earnings / Total Assets

X_3 = Earnings Before Interest and Taxes (EBIT) / Total Assets

X_4 = Market Value of Equity / (Book Value of) Total Liabilities

X_5 = Sales / Total Assets

Utilizing this particular formula, Altman constructed a decision-making guideline. According to this guideline, a Z-score greater than 2.99 indicates that a client is financially stable, whereas a Z-score of 1.80 or less suggests financial instability. For Z-scores that fall between 1.8 and 2.99, the model is unable to definitively classify the client as either solvent or insolvent.

In 1977, Altman et al. (1977) developed an advanced version of the original Z-Score model, introducing significant improvements. This study aimed to create, analyze, and evaluate a novel bankruptcy prediction model that takes into account the trends -of that time- in business failures, by refining the use of discriminant statistical methods. The authors outlined multiple reasons for developing a new model, with empirical evidence supporting this initiative. They named the model ZETA, and this model successfully identified companies at risk of bankruptcy up to five years before their failure, focusing on a dataset of manufacturing and retail corporations. The Z-Score model has been further revised (Altman E., 2000), so that it could be applied to firms of the private sector as well.

Another notable research that was conducted is that of Martin D. (1977), which focused on estimating the probability of bank failure. In his work, Martin D. pinpoints the significance of various financial ratios and their combinations to predict bank failures, by utilizing historical data of Federal Reserve member banks from 1970 to 1976, where 58 banks identified as failures. As predictors were use specific financial ratios, which were categorized into asset risk, liquidity, capital adequacy, and earnings. The study compared the logit model's predictions with those from linear discriminant models, finding that while the discriminant models could classify banks into failed or non-failed categories similarly to the logit model, the latter provided more useful probability estimates for different levels of risk. Finaly, it was presented that for simple classification purposes, discriminant analysis could suffice, but for more accurate probability estimates -which is a focal point in risk assessment analyses, the logit model is preferred.

A lot of research was also done by focusing on specific problems that the credit scoring models of that time had. Specifically, as seen through the work of Eisenbeis R. A. (1978), a more holistic and statistically robust approach to credit scoring model development was critical, stating that models that not only should minimize default rates -but also optimize profitability, taking into account the broader implications and dynamics of granting credit loans. It underscored the complexity of designing statistically sound credit scoring models, by making special referral to various statistical violations that credit models of that time suffered, such as violating statistical assumptions upon certain models were based on, misinterpreting individual variables and thus misleading the decision-making process, sampling bias that might be introduced to the analysis by using non-representative samples, etc. Finaly, it pinpointed the importance of addressing these issues as crucial for creating models that are both reliable and compliant with regulatory standards of that time.

Another research that resonated on the same wavelength by further highlighting issues faced by existing models of that time, was that of Myers G. and Siera S. (1980). This study was focused on the alarming increase in student loan default rates in Mexico, with national rates rising from 4.3% in 1972 to 18.5% by 1975. The research aimed to develop and test a discriminant analysis model that could predict the likelihood of loan default based on characteristics available at the time of loan application and subsequent academic performance (i.e. GPA of each student). While the derived discriminant analysis models offered some insights, they fall short of providing reliable predictions and to accurately forecast default probabilities. A call for additional research to uncover more effective predictive variables and to refine analytical models was made.

In 1984, the Credit Valuation KMV model was introduced by Vasicek A. (1984). This innovative approach leveraged market signals and quantitative models to evaluate the PD and the ELs. Among others, it emphasized on incorporating key elements such as market-based valuation instead of subject-based valuation, EL calculation and loan pricing into the methodology for loan pricing and the analysis of systematic risk in portfolio management. This model assumed the efficiency of markets in reflecting all available information in the prices of securities and views a borrower's creditworthiness through the prism of their asset value. It enabled the calculation of loan default probability and expected loss based on a myriad of factors including the initial market value, debt amount, and the volatility of asset returns. The paper also underscores the importance of portfolio diversification in reducing the variance of expected loss.

Significant research was also made for finding specific predictors for estimating PD. Such study was that of Mona J. and Dixie L. M. (1989). This research focused on analyzing the characteristics of borrowers and loan contracts to determine factors associated with the resolution of overdue accounts. Through logit regression analysis, the study pinpoints the significance of certain traditional variables assessed at the time of the loan application, such as borrower's occupation, property age and location, in predicting PD. Additionally, it examined variables available only at the time of delinquency (i.e. performance information), like the reason for becoming delinquent and the borrower's payment history, as important indicators of default risk. The study is based on data from a major mortgage lender but suggests that its findings could be relevant for broader credit risk management practices -given the commonalities in data analysis across different lending scenarios, including the evaluation of credit history, financial status and the demographic information of the potential borrower.

Notable research was also made by Lawrence et al. (1992), analyzing the default risk in mobile home credit. This paper, building on a foundation laid by numerous studies over the previous decades, sought to enhance the understanding of factors influencing borrowers' defaulting on their obligations by analyzing both traditional variables that considered at the loan's inception -and those observable at the time of the default. Acknowledging the limitations cited in previous research regarding the scarcity of detailed, publicly available databases, this study aimed to broaden the analytical framework by incorporating a more extensive dataset. The research uniquely

contributed by focusing on the underexplored -at that time- area of mobile home lending, utilizing one of the most comprehensive loan datasets from the 1980s. The findings pinpointed the importance of traditional variables like loan-to-value ratios, debt size, and payment coverage in assessing the PD. However, they highlighted the paramount significance of borrowers' payment history in predicting default risk for older loans, including the duration of the loan (i.e. number of tenors), recent delinquency status (i.e. current/last month, previous two months, previous six months, etc. perfromance history), as well as patterns of 30-day and 60-day delinquencies. The study also stated that recent delinquency patterns (i.e. historical delinquencies within the last year of the observation date) emerge as the most reliable and strong indicators of the PD estimation.

Among the years, several studies focused on the sampling bias introduced in the credit risk models, which constituted a significant issue with the way credit scoring models are typically developed, applied and monitored. More specifically, PD models are built using historical data from individuals who have previously been accepted for loans or credit. This approach inherently introduces a bias because the data only reflects the behavior and characteristics of people who have already been approved -and thus have passed through the approval/rejection decision making process. The core problem here is that the models are intended to evaluate all the forethcomening applicants -both that will be approved and rejected as well. The rejected population is expected to be qualitatively and significantly different from the approved population, since the rejected population is of higher credit risk. Hence, if these models are based solely on the data from accepted applicants, they may not accurately predict the risk or behavior of the broader applicant pool, which includes both potentially accepted and rejected applicants. This sampling bias raised many concerns about the appropriateness and effectiveness in evaluating new applications, which could come from anyone in the general population, not just those similar to previously accepted applicants. Greene W. (1998) pinpointed this issue, stating that such sample selection biases could lead to inferior classification results.

The aforementioned issue, however, is suggested to be confronted by not only utilizing the science of statistics, but also incorporating the business credit risk logic. As Sabato G. (2008) wrote in his relevant work “Solving Sample Selection Bias in Credit Scoring: The Reject Inference, *Bank of Scotland*”, in page 1, *“... Some statisticians argue that reject inference can solve the nonrandom sample selection problem (e.g. Copas and Li (1997), Joanes (1994), Donald (1995) and Green (1998)). In particular, reject inference techniques attempt to get additional data for rejected applicants or try to infer the missing performance (good/bad) information. The most common methods explored in the literature are: enlargement, reweighting and extrapolation (see Ash and Meester (2002), Banasik et al. (2003), Crook and Banasik (2004) and Parnitzke (2005)). However, some authors (e.g. Hand and Henley (1993)) demonstrate that the reject inference methods typically employed in the industry are often not sound and rest on very tenuous assumptions. They point out that reliable reject inference is impossible and that the only robust approach to reject inference is to accept*

a sample of rejected applications and observe their behaviour", and further continued "*… in contrast with most of the available literature, we consider the business perspective more relevant than the statistical one in the financial industry context. As such, we conclude that increasing the prediction accuracy of scoring models should not be regarded as the main goal of reject inference techniques. The possibility of including rejects in the development sample should be considered, instead, as an opportunity to replicate the experience and the decision taken by underwriters, credit analysts or branch managers when assessing applicants' creditworthiness*".

Hence, throughout the years, many statistical methodologies have been implemented in the wider context of credit risk and more specifically in analyses involving the PD estimation, to mitigate this inherent sampling bias. Nevertheless, since not all of these could be covered, the most important will be mentioned. Some of the most prevalent are represented in the paper by Ehrhardt A. et al. (2021), where re-weighting, re-classification, as well as fuzzy parcelling/augmentation are examined. The fuzzy parcelling/augmentation technique is a statistical method that offers versatility and constitutes a great solution to the aforementioned problem. Specifically, the fuzzy parcelling/augmentation method is the procedure in which each rejected applicant is hypothetically considered one time as if he/she has been approved and has been classified as "good" applicant, and one time as if he/she has been rejected and has been classified as "bad". However, each inferred observation receives a specific weight -which in the fuzzy parcelling/augmentation method corresponds to a probability. This probability is a model estimate (in previous years it was mainly based on a LR model, nowadays it could be a ML model as well) based on the approved/known population, for which the financial institution has obtained enough historical data (historical payments, credit bureau information, demographic characteristics, etc). Hence, each doubled observation of the rejected population is inferred as approved and "good" applicant - with the corresponding probability of being "good", and inferred as approved and "bad" - with the corresponding probability of being bad, constituting complementary probabilities that add up to 1. Finally, by obtaining the revised obsevations with each approved (known) applicant being represented one time, and each rejected applicant being represented 2 times with their corresponding weights, a weighted modeling analysis takes place with each rejected applicant being represented as one observation. The fuzzy parcelling/augmentation, as well as the rest of the aforementioned reject inference methods, are also presented by Raymond A. (2007) through his book "The Credit Scoring Toolkit", which constitutes a great reference for credit risk assessment and for the estimation of the PD.

In his work, Amos T. O. (2019) presents consicely the model development proceedure, as well as various statistical metrics (i.e.industry standards) upon the performance of the model is being monitored. More specifically, during the initial steps of the development proceedure of a credit risk model, the event definition should be established (i.e. the criterion upon which the "good" - "bad" classification will be made) and the historical data that will be used (i.e. demographic, transactional, performance and credit bureau data, etc.). Furthermore, the sample that will be used as development

sample should be defined -that is the Known Good-Bad population (KGB)- which is nececarry to have enough time after the issuance of each loan -in order for the loans to have enough time to "mature" and make the Good-Bad classification possible. As a next step, variable analyses take place, by proceeding with qualitative and quantitave checks, variable creations (such as the debt-to-income ratio characteristic), outlier handling, variable discretization -that is to segment continuous and categorical variables into classes based into patterns which associate the corresponding predictor with response variable (i.e. good-bad classification) by using statistical metrics such as the Weight of Evidence (Douglas L Weed, 2006) and the IV criterion (Bruce L., David B., 2013), that is known as KGB variable analysis. Subsequently, the development of the KGB Scorecard takes place, where only information from the approved population is utilized. Right after, the reject inference procedure follows, where duplicate values are created for the rejected applicants with each receiving a probability/weight of being "good" and "bad" respectively -probabilities that are estimated by using the KGB scorecard. By utilizing the KIGB sample that has been created (i.e. Known Inferred Good Bad), the same procedure as previously descibed follows. Hence, by incorporating the rejected applicants to the analysis -as if they have been approved, the KIGB variable analysis and KIGB Scorecard development follows. As final step, the final Scorecard is being evaluated by utilizing out-of-sample data (i.e. data that were not included to the development sample), in order for an unbiased estimation of the generalization ability of the model to be made. Some of the most common metrics that are widely used in the context of credit scoring and upon which the validation of the model's generalization ability and stability over-time are being made, are the KS statistic, the Gini's Index, AUC, PSI and the IV criterion -that has already been mentioned.

The KS statistic quantifies the separation between the score (or the probability) distributions of the "good" and "bad" borrowers (Zheng Y., Yue W. et al., 2004). The Receiver Operating Characteristic (ROC) Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various probability threshold settings. The TPR indicates the proportion of actual positives correctly identified by the model, while the FPR indicates the proportion of actual negatives incorrectly classified as positive. The area under this ROC curve -AUC (Andrew P. Bradley, 1997), quantifies the overall ability of the model to correctly classify the positives and negatives across all possible classification thresholds. An AUC of 0.5 indicates no discrimination ability -equivalent to random guessing, whereas an AUC of 1 indicates perfect discrimination. The Gini Index (Gastwirth J. L., 1972) is another popular discrimination metric used in credit risk, and is a linear transformation of the AUC of the ROC curve, and measures in essence the same thing -just in different scale. The PSI (Bilal Y. and Joshua N., 2021) is a statistical measure used to determine how much a variable's distribution has shifted over time between different samples, typically between a baseline (e.g. model development sample) and a more recent or validation sample.

3.2 Supervised Statistical Machine Learning Methods in Credit Risk: Literature Review and Recent Advancements

In the 1990s, some pioneering studies at the time focused on supervised ML methods for assessing credit risk. However, these efforts were hindered by limited computational power, scarce data availability, and challenges in interpreting the algorithms -a problem that persists even today. This prevented ML from flourishing in the area of credit risk and PD estimation, where the model intrepretation is a prerequisite, and not just and extra advantage. Nowadays, however, the scenario is different, since technological advancements and the abundance of data have ignited the growth of the ML field. Consequently, there has been a surge in research on statistical ML algorithms, particularly in the assessment of credit risk and the estimation of the PD.

In 1992, triggered by the high rate of bank failures in the US, Tam K.Y. and Kiang M. (1992) highlighted the urgent need for developing predictive models to anticipate such downturns. In their work, a comparison between discriminant analysis, LR, KNN, ID3 (i.e. an early classification tree method) and NN modeling was made. After presenting the pros and cons of each method, it was found that the discriminant analysis model was outperformed by the other, and that among the rest of the models, no single model uniformly was the best across all scenarios tested, emphasizing the need for model comparison and selection based on specific predictive requirements and constraints. Nevertheless, the paper suggested that while NN, particularly those with complex architectures, generally provided superior predictive performance for bank failures, the choice of model should be guided by -among other- the economic context, and the available data's time frame relative to the prediction target. Additionally, NN were suggested as potentially beneficial for other financial applications -such as PD estimation and loan evaluation, beyond just bankruptcy prediction.

However, Banu Y. and Jonathan C. (2000), by conducting a comparative analysis of predictive performances across LDA, NN, GAs (i.e. Genetic Algorithms), and Decision Trees -using credit scoring data, revealed LDA's superiority in terms of both accuracy and computational efficiency. The study suggested the need for further research, especially to explore the impact of different sample sizes and the cost implications of incorrect classifications, to better understand the strengths and limitations of each technique in practical applications. The study also notes that even though NN and GAs are robust enough to model non-linear relationships, they might got trapped in local optima and thus not generalize well in hold out samples.

In the following years, with the help of the groundbreaking technological innovations, the availability of data as well as the computational power both increased dramatically. So did the studies regarding the statistical ML models in the credit risk assessment. Loris N. and Alessandra L. (2009), cocnluded that ensemble methods, particularly the Random Subspace (RS) method, can effectively improve the accuracy and reliability of bankruptcy prediction and credit scoring systems. As noted, the concept of ensemble classifiers is to leverage multiple classifiers to achieve better

predictive performance than individual classifiers. In this study, four ensemble methods were assessed: Bagging (random subsets of training data are used to train individual classifiers), RS (each classifier is trained using a random subset of features), Class Switching (random switching of class labels in training data) and Rotation Forest (application of PCA to subsets of features -followed by recombination). The classifiers that were tested were NN, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). Different international credit datasets were utilized (in order to ensure the robustness of the predictions), and the evaluations were made upon using the metrics of Accuracy, AUROC and Type I/II errors. It also was shown that the Levenberg-Marquardt NN, provided the best results across the tested datasets, also suggesting that feature correlation (which RS exploits by selecting random feature subsets) is significant in bankruptcy prediction and credit scoring tasks. Finaly, this study contributed to the broader understanding of applying statistical ML techniques in the context of credit risk assessment, demonstrating the potential to surpass traditional statistical methods.

Another notable paper that strongly focused on implementations of statistical ML algorithms in the field of credit risk, is that of Stefan L., Bart B. et al., (2015). As pinpointed, previous studies often use limited datasets, neglecting newer methods and comprehensive statistical metrics. The study benchmarked 41 classification methods -across eight significant credit risk datasets, some of which were heavily imbalanced, hence evaluated the impact that imbalance datasets could have to the performance of each classifier. The need for interpretable credit scoring models is also stressed, as these models must not only be accurate -but also interpretable to ensure they meet regulatory standards, so they could effectively be managed and monitored. Additionally, the evaluation of the different statistical ML models was made upon using multiple performance metrics, which reflected the complexity of evaluating classifiers in a credit scoring context, as different metrics provide different insights.

In terms of model comparison, in the same study, Artificial NN ranked better than LR -indicating advancements since earlier studies, however the empirical evidence suggested that no single advanced method consistently outperformed simpler statistical techniques. Furthermore, as noted, the complexity or novelty of a classifier does not necessarily correlate with better predictive performance. The paper also explored how small increases in the accuracy of classifiers such as LR, artificial NN, RF , and HCES-Bag can lead to notable reductions in misclassification costs, directly impacting the bottom line of the financial data of the credit institution. When the cost of misclassifying a bad customer is significantly higher than misclassifying a good one, classifiers like RF and NN sometimes perform better than HCES-Bag. This outcome suggested that these models might be less conservative, potentially producing fewer false negatives, which become more financially punitive under these respective cost settings. Finally, a balanced and pragmatic approach to adopting new technologies is suggested -considering both statistical performance and the potential business added value of more accurate predictions in high-stakes financial environments.

In credit risk and PD assessment, the calibration of the estimated probability of the statistical ML models should also be examined. As presented by Pedro G. F. and Hugo D. L., (2017), while much attention was given to the discrimination and risk ranking abilities of these models, calibration -the accuracy of mapping a credit score to the actual (observed) probabilities of default- is less explored, especially with realistic, noisy, and imbalanced datasets, and the increasing use of ML in the field of credit risk requires careful calibration -as the more comple ML models often do not naturally produce calibrated probabilities. Instead of focusing solely on the industry standards (i.e. KS, Gini, AUROC), the Brier Score metric is utilized to assess calibration. This metric measures the mean squared difference between predicted probabilities and the actual outcomes, allowing an evaluation of how close the predictions are to reality without relying on predefined risk categories. LR, RF and GB models were compared, by using Isotonic Regression and Sigmoid function (i.e. Platt Scaling) for the probability calibration, and both the RF and GB showed significant improvements when Isotonic calibration was used.

A thorough literature review in credit risk assesment with specific focus in the predictors used among the years for the PD estimation, was made by Büşra A. Ç. and Erman C., (2021). The integration of big data -in general and not only in credit risk, have indeed flourished the field of ML, however it also raises concerns about data privacy, security, and the need for regulations to ensure fair and transparent credit risk assessments. The move towards integrating non-financial data, such as socioeconomic and behavioral factors, into credit scoring models reflects a broader trend towards more holistic approaches. The field of credit risk assessment is evolving from a model based primarily on historical financial data to one that integrates a wider spectrum of data sources, including those derived from digital footprints -that is behavioral data from mobile usage and social networks, user-generated content, and other non-structured data available through online platforms. The use of financial predictors has decreased, while psychometric and alternative variables have gained prominence, especially in the latest years. Though this shift is facilitated by advancements in big data analytics and the increasing availability of non-traditional data source, it also necessitates careful consideration of ethical implications and regulatory requirements to ensure that these innovations benefit both financial institutions and borrowers without compromising privacy or fairness.

Büşra A. Ç. and Erman C., (2021) also make special references to some predictors being consistent over-time, whereas some other seem to have lost discriminatory power or may have stopped being used that much from credit risk model developers. Factors such as home ownership, number of dependents, and wealth (i.e. socioeconomic characteristics) are consistently significant across studies over the years, highlighting their robustness in predicting the PD. Other socioeconomic factors like social class and employment status also show strong predictive power, but are not as universally significant across all studies. Furthermore, personality characteristics such as self-control, emotional stability, and intelligence are significant predictors, and factors like risk-taking (such as betting activities), and spending patterns are also

critical, indicating that how individuals perceive and handle money correlate with their financial stability. Finally, factors such as life-altering events and educational contexts (e.g. field of study, GPA) are correlated with the borrower's financial behavior, indicating that personal circumstances can profoundly impact financial decision-making.

3.3 Summary of Literature Review

The historical evolution of credit risk assessment in financial institutions has undergone a significant transformation from subjective judgment-based methods to advanced statistical models, a shift documented thoroughly in scholarly research across several decades. Initially, credit decisions were largely influenced by personal evaluations and basic financial metrics. However, the inherent biases and inefficiencies in such systems led to the development of more objective and data-driven models, driven by the need to manage portfolios of credit exposures systematically. The introduction of the FICO Score marked a critical turning point in credit risk assessment, paving the way for the development of standardized risk scoring systems that assess borrower risk based on comprehensive credit data. More sophisticated statistical models such as discriminant analysis, LR and probit models were utilized, for more precise and consistent PD estimates. Studies have also highlighted the need for models to not only predict credit risk but also optimize profitability and comply with regulatory standards. Additionally, the development of PD models -which relies on historical data from individuals previously accepted for loans or credit, introduces an inherent sampling bias that has been extensively studied throughout the years. This sampling bias, being one of the most diachronic issues confronted in credit-scoring modelling, should be confronted with the incorporation of business credit logic, in order for the credit scoring models to produce sound and robust estimates.

The performance of these models should be monitored on a regular basis. Metrics such as the KS statistic, Gini Index, AUROC, PSI, and the IV criterion help measure the model's ability to differentiate between good and bad loans, as well as to monitor its stability over time. Through these methodologies and metrics, credit scoring models aim to mitigate inherent biases and improve their predictive accuracy, ultimately enhancing financial decision-making processes.

Over the past few decades, significant technological and methodological advancements have marked the evolution of ML applications in credit risk assessment and PD estimation. The traditional statistical models that were immensly studied among earlier years, were favored by low computational need and by the inherent interpretability they offer. These traditional Statistical models made the PD assessment an imperative need for financial institutions, as investments of higher risk could be avoided, and greatly assisted financial institutions to effectively segment and rank risk. Nevertheless, groundbreaking innovations and the abudancy of various types of data, gave way to various studies in the field of ML. Various complex predictive models have

been studied and implemented in the wider context of credit risk, and based on specific requirements and constraints, the need for an appropriate model selection is now necessity. Furthermore, studies began to challenge and refine the accuracy of different credit scoring models, emphasizing the importance of model comparisons in practical and business applications. Additionally, a shift towards integrating non-financial data underscores a broader trend in credit risk assessment. This trend moves beyond using traditional financial predictors, incorporating socio-economic and behavioral data, thereby enhancing the models' predictive power and reflecting a holistic view of borrower risk. This, nevertheless, raises many concerns about data privacy issues and poses new challenges regarding ethical considerations and regulatory compliance.

CHAPTER 4

Theoretical Background of the Statistical Machine Learning Methods That Will Be Studied

4.1 Statistical and Machine Learning Models

Data science and statistics encompass a robust set of theoretically solid techniques that enable us to either learn from data or tackle significant problems in everyday life. We often deal with classification problems, where observations are identified by various features (*characteristics*) and a corresponding label (*response variable*). In such cases, we need to develop a rule that can predict the label from a predetermined set of possible values when provided with the appropriate features. These types of problems are diverse, appearing in various forms and across different sectors. However, before making a label prediction for each observation based on its features, a probability estimate is calculated for the observation belonging to each potential class label. This process is particularly important in fields like credit risk, where estimating the PD is crucial. For instance, when evaluating potential loan applicants, the focus is more on the probability estimates provided by the model for each applicant being "good" or "bad", rather than merely predicting just a label.

4.1.1 Logistic Regression

LR stands out as a particularly effective model for estimating probabilities and solving classification issues (supervised statistical learning). It belongs to the broader family of generalized linear models (GLM). These models were essential because traditional Linear Regression was either inadequate for handling a wide array of conceptuall soundness issues or failed to meet the necessary underlying assumptions for proper implementation. When tackling a classification problem where the expected label adheres more closely to a discrete distribution like that of Bernoulli, (Multinomial) LR can be used, which is adept at modelling such problems. This chapter was written upong using notes of the following professors: Iliopoulos G. (2022), Politis K. (2022), Koutras M., Boutsikas M. (2011).

LR is employed to predict the probability of an event occurring. The response distribution used is the Bernoulli distribution, which falls within the exponential family of distributions. Therefore, in the case of X following the Bernoulli distribution we find that

$$X \sim B(1, p) \Rightarrow E(X) = p$$

where p is the probability of an even occuring (or not occuring).

As a link function *Logit* is mainly used:

$$Logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

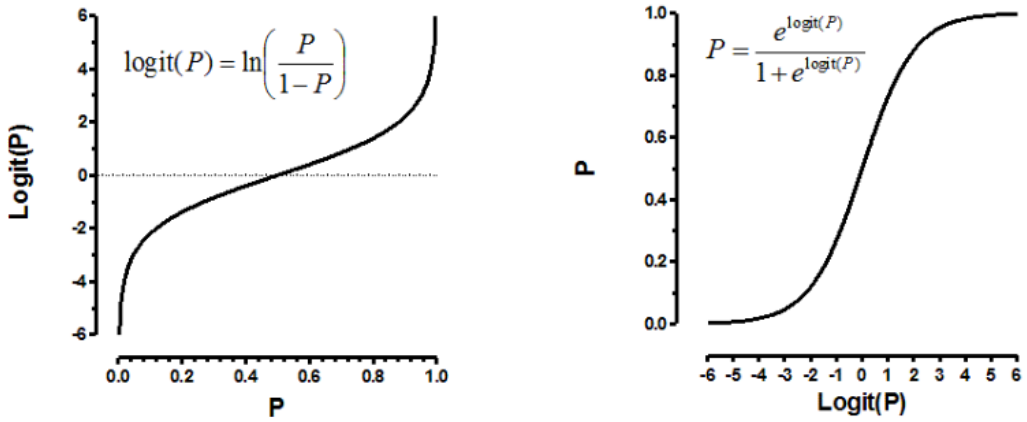


Figure 4.1.1: Logit link function [1]

Other known link functions are the probit and the Complementary log-log, that will not be further discussed throughout the current thesis.

The logit model is the

$$Logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1i} + \cdots b_k x_{ni} \qquad (4.1.1.1)$$

The need for a link function stems from the fact that in traditional regression models, based on the types and values of the explanatory variables, the response variable could potentially vary within the entire range of real numbers. In our scenario, however, where the response variable is a probability, it becomes evident that often the fundamental condition -that the values are constrained within the interval [0,1]-would be frequently violated. The link function helps us resolve this issue by appropriately adjusting the response variable's range. The parameters are estimated using the maximum likelihood method.

We aim to estimate the vector

$$\beta^T = (\beta_0, \dots, \beta)$$

The system that emerges from setting the logarithm of the derivative to zero needs to be solved

$$\frac{d\beta}{d\beta_i} = 0$$

[1] Figure from https://math.stackexchange.com/questions/3816925/how-to-adjust-logit-functions-input-domain

for each i.

For a random variable that follows the Binomial distribution with parameters n,p

$$f_Y(y;p) = \exp[ylog\left(\frac{p}{1-p}\right) + nlog(1-p) + log\binom{n}{y}]$$

Since n=1, we result to a Bernoulli distribution, thus

$$f_Y(y;p) = \exp[ylog\left(\frac{p}{1-p}\right) + log(1-p)]c(y)$$

where the function c(y) does not depend on the parameter p.

Subsequently, the likelihood of the sample is the

$$L(p;y) = \exp[\sum_{i=1}^{n} y_i\, log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \log\,(1-p_i)]c_n(y)$$

with $\mathrm{p}^{\mathrm{T}} = (p_1, \dots, p_n)$, $\mathrm{y}^{\mathrm{T}} = (y_1, \dots, y_n)$

Ignoring the c_n term -which is not connected to the parameters, the logarithm of the likelihood function is expressed as

$$l(p;y) = \sum_{i=1}^{n} y_i\, log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} \mathrm{y_i} \log\,(1-p_i) \qquad (4.1.1.2)$$

Substituting the original linear relationship (4.1.1.1) into the equation (4.1.1.2), we find that

$$l(\beta;y) = \sum_{i=1}^{n}\sum_{j=0}^{k} y_i\beta_j x_{ij} - \sum_{i=1}^{n} \log\,(1 + exp\sum_{j=0}^{k}\beta_j x_{ij})$$

By solving equation (4.1.1.2) for p_i, we find that

$$p_i = \frac{\exp\,(\sum_{j=0}^{k}\beta_j x_{ij})}{1+\exp\,(\sum_{j=0}^{k}\beta_j x_{ij})} \qquad (4.1.1.3)$$

Subsequently, by substituting equation (4.1.1.3) into equation (4.1.1.2), we obtain

$$\frac{\partial y}{\partial \beta_r} = \sum_{i=1}^{n} y_i x_{ir} - \sum_{i=0}^{n} p_i x_{ir} = \sum_{i=1}^{n} (y_i - p_i) x_{ir}$$

which, when set to zero for each r, results in a system of k+1 equations with respect to the parameters β_r contained within p_i from equation (4.1.1.3). This system can only be solved using numerical analysis methods through a statistical package. To find the

values, the iterative method of Newton-Raphson is utilized.

From the equation (4.1.1.3), once we have obtained the estimates of β_j from the statistical package, we can find the corresponding estimates of p_i through the

$$\hat{p}_i = \frac{\exp\left(\sum_{j=0}^{k} \hat{\beta}_j x_{ij}\right)}{1+ \exp\left(\sum_{j=0}^{k} \hat{\beta}_j x_{ij}\right)} \quad (4.1.1.4)$$

In the context of statistical inference, we can conduct tests on the model and its parameters. The most basic tests concern the parameters β_i. Often, we need to test whether the true value of a parameter is equal to zero, as this implies that the response variable is not statistically significant connected with the corresponding explanatory variable, or given that also other explanatory variables are present in the model, the specific one is not deemed statistically significant.

Thus, for the test of the statistical hypothesis

$$H_0: \beta_i = 0$$

$$H_1: \text{otherwise}$$

we leverage the asymptotic property of the normality of the maximum likelihood estimators and use the Wald test statistic

$$w = \frac{\hat{\beta}}{s(\hat{\beta}_i)}$$

which under the null hypothesis approximately follows the standard normal distribution N(0,1).

Therefore, we reject the null hypothesis when

$$\left|\frac{\hat{\beta}}{s(\hat{\beta}_i)}\right| > z_{\alpha/2}$$

where α is the level of significance set by the analyst and $z_{\alpha/2}$ is the upper percentile point of the standard normal distribution.

A significant concept in LR modeling is the Deviance

$$D = -2Loglikelihood(model)$$

With this quantity, and through the generalized likelihood ratio test, we can proceed to tests for the contribution of new incoming variables in nested models. Nested models are considered two models for which the set of explanatory variables of one model is a subset of the explanatory variables of the other model.

Therefore, having a model m_1 and a nested model m_2, we can calculate

$$D2 - D1 = -2[Loglikelihood(m2) - Loglikelihood(m1)]$$

where the difference in deviations follows approximately a χ^2 distribution with p degrees of freedom, that is the difference in the degrees of freedom of the regression of the 2 models.

Specifically, the following hypothesis is being tested

H_0: m1 – m2 the difference between m1 & m2 is not statistically significant

H_1: the models differ significantly

As we have seen, LR can estimate the probabilities p_i through the formula (4.1.1.4). Therefore, the way we can use the LR model as a classification tool is as follows. We choose a threshold for p_i which is between 0 and 1, and if the vector of explanatory variables for the new observation, adjusted to the estimated parameters of the true parameters of β_i, gives us an estimate above the threshold we have set, then we classify this observation as a success, hence as 1. Otherwise, we consider it as a failure and label it with the value 0. The most common threshold is 0.5, but it can significantly vary across difference industries, studies, as well as it greatly depends on the metric that need to be optimized.

4.1.2 Decision Trees

Before moving on to the RF model, an introduction to the wider family of the Decision Trees will be made. Decision Trees are adept at addressing both regression and classification (in which we will focus on) issues. The principle behind decision trees is to establish a pathway from the root to the leaves, dictated by the interconnections of the data at hand. These trees comprise several nodes, each containing conditions relevant to our observations' predictors. The process moves forward depending on whether these conditions are satisfied. The bibliography for this sub-section is Bersimis S. (2021), Iliopoulos G. (2023), Jerome H. F., Robert T., et al. (2009).

A decision tree consists of nodes, which fall into one of the following categories:

- Root Node (or *Parent Node*)
- Internal Nodes
- Leaf Nodes

The Root Node (also known as *Parent Node*) is the initial node, unique in that it feeds other nodes but is not fed by any other node. Following the Root Node, there are the Internal Nodes, which both receive input from and provide output to other nodes, branching out two or more other Internal Nodes. Lastly, the Leaf Nodes are those that only receive input from other nodes -producing no other node. The connections between the nodes are referred to as *Branches*.

Decision trees start from the root and filter through the given features (or *Characteristics* or *Predictors*) of the observation we wish to classify within the internal nodes, eventually leading to a leaf. This leaf determines the final category into which the observation will be classified. The process works through a condition related to the node; depending on whether the outcome of this condition is True or False, it decides the next step of the process. If the condition is met, we continue to the left internal node or leaf node; otherwise, we proceed to the right, as shown in Figure 4.1.2.1.

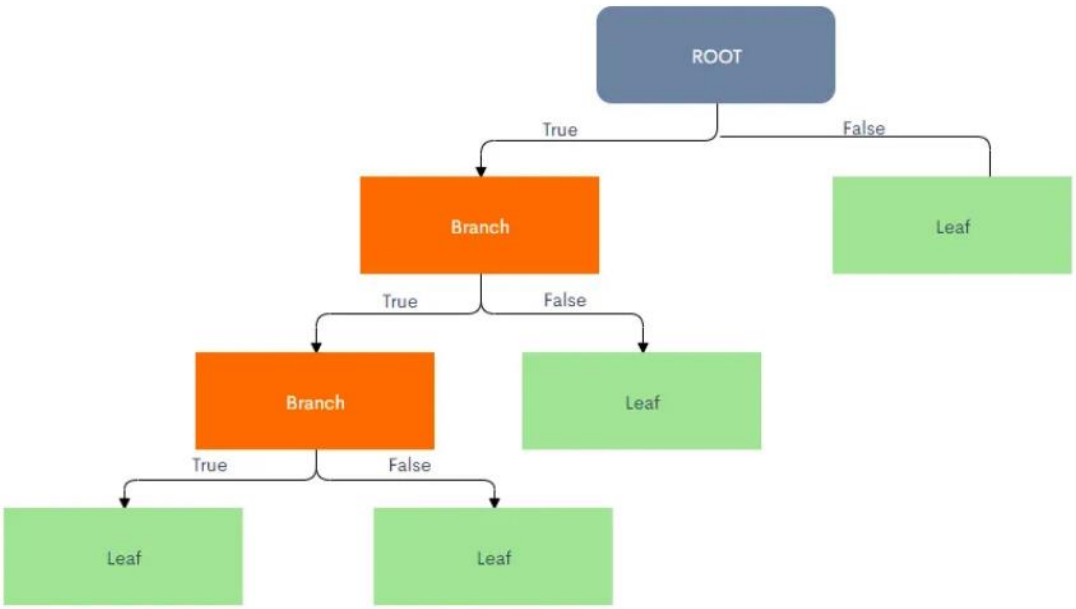


Figure 4.1.2.1: Representation of a Decesion Tree [2]

In constructing the model, consider a dataset

$$D = \{x_i, y_i\}$$

where x_i is the vector of observation features -which may be categorical or quantitative, and y_i is the label of the observation. Our objective is to develop a set of criteria applied to x_i. Fulfilling or violating these criteria directs to specific leaf nodes that provide predictions for y_i. Importantly, upper-level nodes play a crucial role because once a pathway is selected, it cannot be changed, making the placement of conditions within the nodes essential for the model's effectiveness.

One common approach for structuring the model and arranging the nodes is through the use of the Gini Impurity, which calculates the probability of k label categories appearing within a set, effectively measuring the uniformity.

Gini Impurity is calculated by using the below formula

[2] Figure from https://medium.com/geekculture/all-about-decision-trees-part-i-cfa148c75631

$$Gini_{Impurity}(D) = 1 - \sum_{j=1}^{k} p_j^2$$

Where p_j is the probability of class j appearing from a set of k classes in a dataset D consisting of n observations. When the dataset D is split into two subsets D_1 and D_2 of sizes n1 and n2 respectively, then

$$Gini_{Impurity}(D) = \left(\frac{n}{n1}\right) \times Gini_{Impurity}(D1) + \left(\frac{n}{n_2}\right) \times Gini_{Impurity}(D2)$$

meaning that the Gini Impurity is weighted according to the number of observations in each subset. The Gini Impurity is an indicator of purity or, alternatively, the degree of impurity of a node. The lower the Gini Impurity is, the "purer" the node is, indicating a good separation between the different categories.

When dealing with large datasets that contain too many observations and variables, reaching the leaf nodes in a decision tree is not straightforward. Instead, there are numerous internal nodes that repeatedly filter the observations through various conditions to reach a decision at a leaf node. Often, these decisions appear to have been determined at higher levels of the tree. Therefore, to avoid excessive complexity in the Decision Tree, the model developer might choose to "prune" some of the branches, converting the last level of internal nodes into leaf nodes. The depth and the point of pruning are determined by the analyst (or by some iterative procedures that will be presented till the end of this chapter), provided that it does not significantly reduce the performance of the model.

Another useful widely used metric which is used for the development (or training) of Decision Tree models is that of Entropy. In information theory, the entropy of a random variable represents the average amount of "information" or "uncertainty" associated with the variable's potential outcomes. Within the framework of Decision Trees, entropy is utilized to quantify the level of disorder or impurity present within a node. [3]

Entropy is calculated by using the formula

$$E = -\sum_{j=1}^{k} p_j \log_2(p_j)$$

where the p_j is the probability of randomly selecting an example in class j.

Hence, to identify the root node, a Decision Tree calculates the entropy for each variable and its (all) possible divisions. This involves determining a potential division for each variable, computing the Average Entropy across all resulting nodes, and then

[3] Source https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c

assessing the reduction in entropy compared to the parent node. This reduction in entropy is known as Information Gain, and it measures the amount of information a feature contributes towards predicting the target variable.

Subsequently, Information Gain is calculated as follows

$$Information\ Gain = Entropy_{Parent} - Entropy_{Children}$$

where “Parent” is denoted each node that undergoes a split into smaller groups -based on a certain attribute or feature, as presented previously, and as “Children” are denoted the nodes that result from splitting the parent node, hence each child node represents a subset of the data from the parent node.

4.1.3 Random Forest

RF belongs to the wide family of Decision Trees. By incorporating many statistical methods, a much more complex model can be made and produce accurate predictions -even in cases where the underlying relationships of the data are non-linear and too complex.

Ensemble learning is a statistical (machine) learning technique that combines multiple models to produce a single predictive model that typically performs better than any single contributing model. The predictions from multiple models, often achieve more accurate and robust results than using any single model alone. Ensemble methods use multiple learning algorithms or multiple instances of the same algorithm with different parameters or training data subsets. This diversity helps in making more stable and accurate predictions by reducing the bias of an individual model.

Bootstrap is a statistical technique primarily used to estimate the distribution, variance, or confidence intervals of a statistic by resampling with replacement from the original sample when direct calculation or traditional methods are not feasible. For example, there is no statistical formula to estimate the variance (and as such the standard deviation as well) of statistical functions such as the Pearson correlation coefficient. Bootstrap can make it possible, allowing for robust statistical inferences to be made through uncertainty quantification (Bootstrap Confidence Intervals).

Let T a dataset

$$T = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

where x represents the vector of p predictors and y is the label of the corresponding observation. A bootstrap sample, denoted as T*, is a random sample selected with replacement from T. Practically, this means that each pair (x_i,y_i) from the dataset, listed as

$$T^* = \{(x_1^*, y_1^*), \dots, (x_n^*, y_n^*)\}$$

is chosen with a probability of 1/N, independently from each other. This procedure (random sampling with replacement) creates T* samples which are composed of independent and interchangeable observations, with some pairs from the original dataset appearing multiple times in some of the obtained bootstrap samples, while others might not appear at all. The probability that a bootstrap sample exactly replicates the original dataset is extremely low, and the probability of a specific observation (x_i, y_i) not being included in a bootstrap sample is

$$\left(1 - \frac{1}{N}\right)^N$$

As N grows (number of observations of the original sample), this probability approaches approximately e^{-1} (or about 0.368). This implies that for a sufficiently large N, each observation from the original dataset is expected to be absent in about to 36.8% of the B bootstrap samples.

To illustrate, for different values of N, the probabilities are as follows:

- With N=25, the probability is approximately 0.360.
- With N=50, the probability is approximately 0.364.
- With N=100, the probability is approximately 0.366.
- With N=200, the probability is approximately 0.367.
- With N=500, the probability is approximately 0.368.

Consider a regression scenario where a model is fitted to predict the value of a continuous random variable $\hat{f}$(x), where $x = \{x_1, x_2, \dots, x_n\}$. Bagging, an ensemble learning technique, leverages the average of several predictions made by the same model but on different bootstrap samples to refine its estimate. By obtaining multiple bootstrap samples $T^* = \{T_1^*, T_2^*, \dots, T_B^*\}$, using each sample at a time, a calculation for the prediction of $\hat{f}$ is being made, thus calculating the corresponding predictions of $\hat{f}^{*1}$(x), $\hat{f}^{*2}$(x), ..., $\hat{f}^{*B}$(x). The bagging prediction, $\hat{f}_{bag}$(x), is then determined by taking the average of these predictions:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}^{*b}(x)$$

If f(x) is linear, practically there's no substantial difference in proceeding with the estimation directly from the original dataset and from using the bagging method. However, if f(x) involves complex or non-linear relations, then $\hat{f}_{bag}$(x) not only differs -but typically offers an improvement since it reduces the variance of the prediction.

Subsequently, given that decision trees can capture complex and non-linear relationships, bagging can be especially useful. Each bootstrap sample generates a tree based on different rules (*thresholds*) for dividing the data into different sub-populations. The bagging technique then averages the predictions from these multiple trees to produce a final estimate. For classification trees, the bagging prediction $\hat{G}_{bag}$(x), is the category most frequently selected by the B trees generated from the bootstrap samples.

Alternatively, the vector that contains the average probabilities calculated from the B bootstrap samples for the observation belonging in each class 1,2, …, K is first calculated, and subsequently the observation is classified in the class with the highest estimated probability. This approach not only offers improved estimates of the probabilities for each of the K categories -but also tends to lower the error in classification.

RF constitutes an ensemble learning technique used for both classification and regression tasks. This method employs bagging with "weak" decision trees to create a powerful model that exhibits lower variance. In this approach, trees are built from bootstrap samples, and each split in the tree is determined by selecting from a random subset of the available predictors -thus not using all the potential predictors for the construction of each tree, with each independent tree being "weaker" when compared to trees where all the potential predictors were used.

Since the B trees are constructed using the same process, each specific statistical function produced by them has an identical mean, which aligns with the mean of their corresponding average. The average of B independent and identically distributed random variables -each one with a variance of σ^2, yields a variance of σ^2/B.

In the case where these random variables are identically distributed but exhibit a positive correlation ρ, the variance calculation is as follows

$$\mathrm{Var}(\bar{Z}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

and for sufficiently large B, the contribution of the second term in the equation becomes negligible and the variance of the average being determined by the first term $\rho\sigma^2$.

The concept of RF aims to mitigate the variance of estimations generated through the bagging method by reducing the correlation among tree decisions. This reduction is achieved without significantly diminishing the variance of each tree individually. Hence, the strategic random selection of variables for splitting helps maintain a relatively low correlation between predictions derived from different trees. The optimal number of predictors to use for constructing each tree depends heavily on the nature of the problem and the dataset. However, a smaller subset of predictors typically reduces the correlation among the trees within the forest, thereby decreasing the variance in the final output.

Once B trees are constructed, they offer predictions of

$$f(x;\theta_1), f(x;\theta_2), \dots, f(x;\theta_B)$$

where x is a given point with the values of the p-predictors used and θ_b is the parameters of each trained tree.

The overall prediction from the RF is calculated as

$$\hat{f}_{\text{rf}}(\text{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}\ (x;\theta_b)$$

A crucial aspect of RF is the use of Out-Of-Bag (OOB) samples -which are available due to the bootstrap procedure during the training of each forest. Specifically, for each bootstrap sample used to construct a tree, the observations which are not included in this sample form the corresponding OOB sample comprise B such samples. For each observation (x_i, y_i), the corresponding prediction $\hat{f}_{\text{OOB}}(x_i)$ of the RF is determined by averaging the predictions from only those trees built from bootstrap samples that do not contain that observation. The contribution of this observation to the total error estimation is

$$Error_{OOB\,i} = \left(y_i - \widehat{f_{OOB}}(x_i)\right)^2$$

Subsequently, when confronting a classification problem, the contribution of each i^{th} observation to the OOB error is the majority vote of the trees that have been constructed from samples not containing that observation. OOB error is very helpful in evaluating the model because it effectively performs cross-validation, by approximating leave-one-out CV alongside the training of the model - despite overestimating the real error.

4.1.4 Gradient Boosting

GB are a class of non-parametric supervised ML algorithms that are widely used across different industries, both in regression and classification tasks. Unlike traditional methods which rely either on single model outputs (regression models such as LR), or in ensemble learning models such are RF and NN ensembles, GB construct models by sequentially adding weaker models (*base-learners*), thereby creating a robust predictive model through a stage-wise optimization process. In contrast to the ensemble methods, the main idea of GB is to add new models to the ensemble sequentially and not train them in parallel, thus each subsequent model (new "weak" base-learner) is driven by the results of the whole ensemble up until that point. This is where the term "Boosting" refers to. This methodology offers flexibility and high performance (Alexey N. and Alois K., 2013).

As supervised ML mandates, data must be pre-labeled.

The dataset is denoted as

$$(x, y)_i^N$$

where $x = (x_1, \ldots, x_n)$ represents the input variables and y represents the corresponding target labels of the response variable.

The objective is to approximate the unknown functional dependency $\text{x} \xrightarrow{f} \text{y}$ using an estimate $\hat{f}(\text{x})$ that minimizes a predefined loss function $L = (y, f)$ as follows

$$\hat{f}(x) = \arg\min_{f(x)} L\big(y, f(x)\big) \tag{4.1.4.1}$$

Since the parameters of the model need to be estimated, the functional dependency is written as

$$f(x, \theta)$$

where θ is a vector containing the parameters of the function and thus, $f(x, \theta)$ describes a parametric family of functions. This converts the problem of function optimization as a problem of parameter estimation

$$\hat{f}(\mathrm{x}) = f(\mathrm{x}, \hat{\theta}) \tag{4.1.4.2}$$

$$\hat{\theta} = \arg\min_{\theta} E_x[E_y(L(y, f(x, \theta)]|x] \tag{4.1.4.3}$$

and for these estimation, iterative numerical procedures are utilized.

For M iteration steps, the parameter estimates are written as

$$\hat{\theta} = \sum_{i=1}^{M} \hat{\theta_i} \tag{4.1.4.4}$$

The most frequently used parameter estimation procedure is the Steepest Gradient Descent. Given N data points the empirical loss function J(θ) over the observed data is aimed to be decreased

$$J(\theta) = \sum_{i=1}^{N} L(y_i, f(x_i, \hat{\theta}))$$

The classical steepest descent optimization method involves iterative improvements in the direction of the gradient of the loss function

$$\nabla J(\theta)$$

Since parameter estimates θ are updated incrementally, different notations are used for clarity. The subscript $\hat{\theta}_t$ represents the estimate at the t^{th} incremental step, while the superscript $\hat{\theta}^t$ denotes the collapsed estimate up to step t, which is the sum of all incremental updates from step 1 to step t. The steps for the steepest descent optimization are as follows:

- Initialization of the parameter estimates as θ_0 and for each iteration t, repeat:
- Calculate the cumulative parameter estimate θ^t from all previous iterations

$$\widehat{\theta^t} = \sum_{i=0}^{t-1} \hat{\theta_i}$$

Evaluate the gradient of the loss function $\nabla J(\theta)$

- using the current cumulative parameter estimates:

$$\nabla J(\theta) = \{\nabla J(\theta_i)\} = \left[\frac{\partial J(\theta)}{\partial J(\theta_i)}\right]_{\theta = \hat{\theta}^t}$$

- Compute the new incremental parameter estimate θ_t:

$$\theta_t \leftarrow -\nabla J(\theta)$$

- Add the new estimate θ_t to the ensemble.

The key attribute of boosting methods lies in the domain of optimization: Boosting is being operated in the function space. This means the function estimate f is expressed in an additive form as

$$\hat{f}(x) = \widehat{f^M}(x) = \sum_{i=0}^{M} \hat{f}_i(x) \qquad (4.1.4.5)$$

where M is the number of iterations, $\hat{f}_0$ is the initialization guess (in order for the algorithm to start searching for the optimal gradient), and $\{\hat{f}_i\}_i^M$ are the incremental functions, referred to as "boosts".

To practically implement this functional approach, a similar strategy of parameterizing the function family is used. Parameterized "base-learner" functions are denoted as

$$h(x, \theta)$$

to distinguish them from the overall ensemble function estimates of $\hat{f}(x)$. Various types of base-learners, such as decision trees or splines, can be employed. For more information about different types of base learners refer to Alexey N. and Alois K., (2013).

The "greedy stagewise" method of function incrementing with base-learners can now be formulated. At each iteration, the optimal step-size ρ needs to be determined. For the function estimate at the t^{th} iteration, the optimization rule is defined

$$f_t \leftarrow f_{t-1} + \rho_t h(x, \theta_t)$$

The parameters (ρ_t, θ_t) are such that minimize the relationship

$$(\rho_t, \theta_t) = argmin_{\rho,\theta} \sum_{i=1}^{N} L(y_i, \hat{f}_{t-1}) + \rho h(x_i, \theta)$$

In GB, both the loss function and the base-learner can be defined as needed by the model developer/analyst. When dealing with a specific loss function and/or a custom base-learner, finding the parameter estimates can be challenging. To address this, it was proposed to select a new function of $h(x,\theta_t)$ that closely aligns with the negative gradient of

$$g_t(x) = -E_y\left[\frac{\partial L((y,f(x))}{\partial f(x)} \mid x\right]_{f(x)=\hat{f}^{t-1}(x)} \quad (4.1.4.6)$$

Rather than searching for a comprehensive solution for the boost increment in the function space, one can choose the new function increment to be the most correlated with $-g_t(x)$. This transforms a potentially difficult optimization problem into a more manageable least-squares minimization problem

$$(\rho_t, \theta_t) = argmin_{\rho,\theta} \sum_{i=1}^{N} [-\mathrm{g_t(x_i)} + \rho \mathrm{h(x_i, \theta)}]^2 \quad (4.1.4.7)$$

This process continues iteratively, sequentially enhancing the model by fitting new base-learners to the residuals (errors) of the combined ensemble of previous learners, thereby improving the overall prediction accuracy. As stated previously, GB offers great flexibility in choosing the loss function and the base-learner model, thus being adept at handling various type of non-linear and complex relationships of the data at hand.

Loss functions are categorized based on the type of response variable y. Various boosting algorithms have been developed for different response families, including regression, classification. The most commonly used loss functions are organized as follows:

- For continuous response variable some of the most common ones are Gaussian, Laplace, Huber and Quantile loss functions.
- For categorical response variable some of the most common ones are Binomial and AdaBoost loss functions.

For more information about the loss functions used in GB algorithms, refer to Alexey N. and Alois K., (2013).

Below is presented the summary of (Friedman's) Gradient Boost Algorithm:

Inputs:

- Input data $(x, y)_i^N$
- Number of iterations M
- Choice of the loss function $L(y, f)$
- Choice of the base-learner model $h(x, \theta)$

Algorithm:

- Initialize f with a constant → $\hat{f}_0$.
- For t=1 to M do:

  Compute the negative gradient $g_t(x)$ (4.1.4.6)

  Fit a new base-learner function $h(x, \theta_t)$.

  Find the optimal gradient descent step-size ρ_t: (4.1.4.7)

$$(\rho_t, \theta_t) = argmin_{\rho,\theta} \sum_{i=1}^{N} [-\mathrm{g_t(x_i)} + \rho \mathrm{h(x_i, \theta)}]^2$$

Update the function estimate: $f_t \leftarrow f_{t-1} + \rho_t h(x, \theta_t)$

- End for.

In conclusion, GB can be designed with various base-learner models, and numerous types have been introduced in the literature. Base-learner models are typically categorized into three main groups: linear models, smooth models, and decision trees. An important feature of GB is the flexibility to combine different classes of base-learner models within a single GB -since a model can simultaneously include smooth additive components and decision trees. Additionally, explanatory variables can be handled with different boosted base-learner models fitted to each subspace.

4.1.5 Neural Networks

Some ML algorithms, inspired from the functioning of biological learning systems and particularly from the human brain, are comprised of complex networks of interconnected neurons. Known as artificial NN, these algorithms mimic this structure with a dense network of interconnected "simple" units. Each unit processes a set of numerical inputs and produces numerical outputs. This processing occurs across three main levels in a typical NN: the input layer, the hidden layer, and the output layer. Some NN are designed with additional layers, especially an increased number of hidden layers, and contain a vast number of neurons. The development and analysis of such highly complex NN fall under a specialized subfield of ML known as "Deep Learning". The bibliography used for this sub-chapter is Tasoulis, S., (2021), Pan Z., Jiashi F. et al., (2020).

To apprehend the properties and advantages of NN, it is essential to present the most significant architectures that have been widely used in recent years and examine their distinct characteristics. Apart from the three fundamental layers – input layer (contains the *Input Nodes*), hidden layer(s) (contain the *Hidden Layer Nodes*), and output layer (contains the *Output Nodes*) - mentioned previously, typical artificial NN include various other parameters that define their functionality and behavior. Two key parameters are the *Input* and *Activation* functions, which regulate the operations of the individual units that make up the network's layers. These functions are generally consistent across each layer of the NN.

The term "Input" to a node in a NN refers to the weighted sum of the outputs from the nodes connected to it. A typical input function usually takes the form

$$Inpt_i = \sum w_{ij} x_j + \mu_i$$

where $Inpt_i$ represents the weighted sum of the input elements x_j to unit i, w_{ij} denotes the weights connecting neuron j to neuron i, and μ_i is the constant for neuron i. The threshold acts as a constant -thus is a reference point in the absence of other input values. Each unit in the NN then takes these input values described by $Inpt_i$ and applies an activation function to them. For example, the output of the j-th unit (or *activation value*), is inserted into a function $f(.)$ as

$$f\left(\sum w_{ij}x_i\right) = a_i$$

where $f(.)$ is the activation function, and x_i is the output from the i-th unit connected to unit j, and α_i is the corresponding output value.

Some of the most well-known and widely used activation functions include:

1. Sigmoid Function:
   - Formula: f(z) = $\frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$
   - Characteristics: Maps input values to a range between 0 and 1, making it particularly useful for binary classification problems. However, it may suffer from "vanishing" gradients during backpropagation -thus slowing down the learning rate.

2. Hyperbolic Tangent (tanh) Function:
   - Formula: f(z) = $\frac{(e^z - e^{-z})}{e^z + e^{-z}}$
   - Characteristics: Maps input values to a range between -1 and 1, and also is zero-centered, which can make it easier for the network to model inputs that have strongly negative or strongly positive values. Like the sigmoid, it can also suffer from "vanishing" gradients.

3. ReLU (Rectified Linear Unit):
   - Formula: f(z) = $\begin{cases} 0 & z < 0 \\ z & otherwise \end{cases}$
   - Characteristics: Introduces non-linearity by outputting the input directly if it is positive; otherwise, it outputs zero. It is simple and computationally efficient, and also helps mitigate the vanishing gradient problem. However, it can suffer from issues as become "inactive" and only output zero values.

4. ELU (Exponential Linear Unit):
   - Formula: f(z) = $\begin{cases} z, & z > 0 \\ ae^z - 1, & z < 0 \end{cases}$
     where $\alpha > 0$ is parameter (that can be tuned).
   - Characteristics: It allows for smooth and non-zero gradients when the input is negative, which can speed up learning.

5. Softmax Function:

- Formula: $f(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$
- Characteristics: Converts logits (raw network outputs) into probabilities. It is commonly used in the output layer of classification networks where the classes are mutually exclusive.

These activation functions -having their advantages and drawbacks- are chosen depending on the specific use case. Choosing the appropriate activation function is crucial for the performance of a NN. Additionally, the goal of activation functions is to prevent extreme values in the output results, which could hinder the network's learning process.

Thus, in summary, the key components that make up the architecture and structure of a typical NN are:

- Number of hidden layers: A NN can have one or more hidden layers, enabling it to identify complex relationships within the data. Networks with just one or two hidden layers can perform well if they contain a significant number of neurons -and/or if the underlying correlations and interactions between the data are not too complex. For datasets with extremely complex underlying relationships, increasing the number of hidden layers could help the model to produce more accurate preditictions.

- Number of hidden nodes: There is no definitive method for selecting the optimal number of hidden nodes in an artificial NN. This selection requires experimental trials and performance evaluation for each test -being itself a hyperparameter to be tuned. More about the hyperparameter tuning in machine learing are presented in sub-chapter 4.5.4.

- Number of output nodes: The number of output nodes is crucial and depends on whether the problem is a regression one or a classification (binary or multiclass). In classification tasks, the number of output nodes typically corresponds to the number of classes we want to predict.

- Activation and input functions: Activation functions that convert linear relationships into non-linear -and help determine the outcome of the processing steps on the input data through the input functions.

The above presented points can be apprehended more intuitively by the following figures

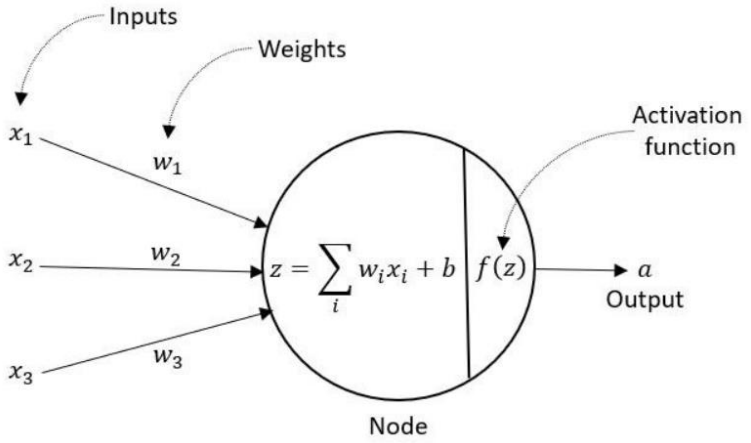


Figure 4.1.5.1: NN Node and the role of the Activation Function [4]

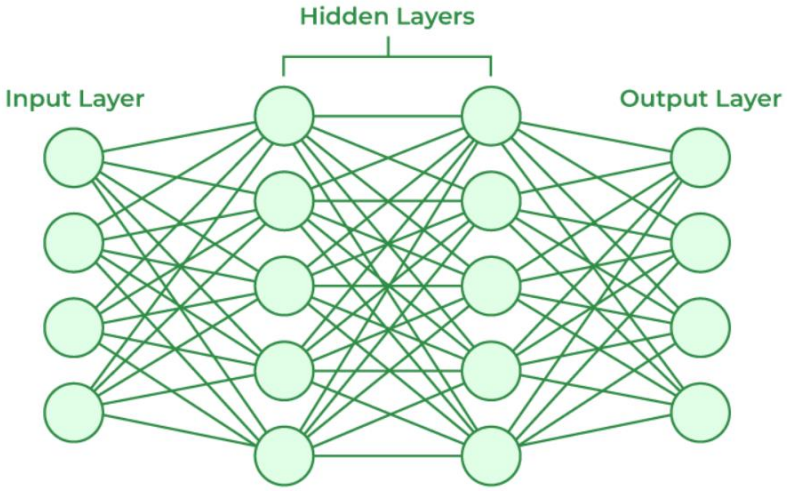


Figure 4.1.5.2: NN Architecture [5]

Thus, it becomes apparent that there is no single architecture for artificial NN; each architecture includes the components mentioned above but differs in how it processes information and produces results. Depending on the nature of the problem, an appropriate NN architecture must be chosen.

The main and most-well known types of NN (at the time this thesis is being written), are the Feedforward Neura Networks (FNN), the Recurrent NN (RNN), the Convolutional NN (CNN), and the Generative Adversial Networks (GAN). Specifically, due to their corresponding characteristics and methodology, each type is typically suited for specific tasks:

1. FNN: The simplest type of NN, where data flows in one direction from input to output. Used for pattern recognition, regression, and classification. FNN are basic and versatile, ideal for straightforward tasks without temporal dependencies.

[4] Figure from https://levelup.gitconnected.com/a-review-of-the-math-used-in-training-a-neural-network-9b9d5838f272
[5] Figure from https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

2. RNN: Designed for sequential data, these networks use loops to process data, making them suitable for time series predictions and language translation.
3. CNN: Specialize in image and pattern recognition by using convolutional and pooling layers. Commonly used in computer vision.
4. GAN: Consist of two networks, a generator and a discriminator, that work together to generate realistic data. Thus, GAN are powerful for generating new, realistic data by learning the underlying distribution of the training data. Used in image generation and data augmentation.

For the purpose of this thesis, only FNN will be presented, as the objective is to predict the PD in the context of credit risk. For example, RNN are designed for sequential data, and the data used in PD estimation are not interconnected in a sequential manner. Therefore, employing complex models like RNN would be unnecessary and too complex for this task.

A FNN has the architecture previously described. For this type of network, additional details worth mentioning include:

- Forward propagation:
  - Input layer: The input data enter the network.
  - Hidden layers: Each neuron processes its input using a weighted sum and an activation function (as already described), passing the result to the next layer.
  - Output layer: The final prediction is produced after data has passed through all layers.

- Loss function:
  - Quantifies how well the network's predictions match the actual data.
  - Common choices include Mean Squared Error (MSE) for regression tasks and cross-entropy loss for classification tasks.

- Training and backpropagation
  - Feedforward NN are trained using a process called backpropagation. Backpropagation involves computing the gradients of the loss function with respect to the network's weights and biases (errors). These gradients are then used to update the weights and biases to minimize the loss.

- Hyperparameters: The learning rate, the number of hidden layers, and the number of neurons in each layer, must be initialized before training begins. Effective initialization and continuous testing of these hyperparameters play a crucial role in the network's performance. Specifically:
  - Learning rate: Controls the step size during gradient descent updates.
  - Number of hidden layers: More layers can capture more complex patterns but may lead to overfitting (mentioned previously).
  - Neurons per layer: Affects the model's capacity to learn from the data

(mentioned previously).

The careful selection and tuning of these aspects are critical for the successful training and performance of a feedforward NN.

Finally, the weights of the interconncted nodes used in the NN model need to estimated. In linear models, simpler methods are being used such as Maximum Likelihood Estimators (MLE) and Least Squares. However, NN are more complex than linear models, and different approches have been utilized for estimating the weights of the model. Some of the most well-known and widely-used methods are the Adam and Stochastic Gradient Descent (Pan Z., Jiashi F. et al., 2020).

4.2 Data Management & Pre-Processing

Different models require different approaches when it comes to data management. Many things must be considered before training/developing a predictive statistical model. Nevertheless, some things are in common -regardless of the model that will be used. (Brown I., 2014). In specific:

- Relevance: The data collected is relevant to the problem trying to be solved.
- Quality checks: For accuracy and consistency in the data. For example, extremely high or low values for some characteristic may indicate inaccuracy and data integrity issues.
- Missing values: Identification and handling of missing values through imputation, flagging or removal. In credit risk modelling, the most common approach is to flag the missing values (i.e. to encode with special codes such as -9999), since the missing value contains itself useful information, and thus if imputed or dropped the predictive power of the corresponding variable will be negatively affected and/or be biased.
- Outlier detection: Detection of outliers through univariate and/or multivariate methods. However, some extremely high or low values might also indicate that the data are not accurate, and thus this specific data points will not constitute real observed values.
- Duplicates: removal of multiple completely identical records, to prevent bias from preventing the statistical models from overly getting trained and validated on the same patterns again-and-again.
- Data transformation:
    - Normalization/standardization, thus scaling features to a standard range or distribution. This is more common in linear models, such as LR, where the difference in scale between the used predictors may prevent

the model from identifying true and valid relationships between the data. It is also a common approach when using NN. In contrary, tree-based models are not affected by the difference in scales of the predictors used.

- Encoding: conversion of categorical variables into numerical values and moreover, in ordinal order (ordinal encoding) or into binary/dummy variables.

- Exploratory Data Analysis (EDA):

  - Descriptive statistics: Summary of the main characteristics of the dataset.

  - Data visualization: Plots and graphs to understand patterns and the distribution of the data, as well as to search for potential relationships between the used predictors as well as with the target variable.

- Over-time consistency checks: Verification of the integrity and consistency of the data over time. For example, when handling a dataset consisting of loan applications, the model developer should always check whereas the data are available for all the months of each year. If specific dates/months are missing, then sampling biased might be introduced into the analysis and the model may not be able to predict well on unseen data.

4.3 Train-test split

The train-test split procedure is a fundamental method in ML used to evaluate the performance of predictive models. It involves dividing the available dataset into two separate subsets -train and test sets, or into three -training, validation and test sets. Each subset serves a specific purpose in the model development and evaluation process, ensuring that the model generalizes well to new data (Jerome H. F., Robert T. et al., 2009). The train-test split procedure can be described in the following setps:

1. Shuffling the dataset: Random shuffling of the dataset, to ensure that the data distribution is uniform and that any inherent order does not bias the split.

2. Dividing the data: Split the shuffled dataset into three parts: the training set, validation set, and test set. A common split ratio is 70/15/15 or 60/20/20, where 70% (or 60%) of the data is allocated to the training set, 15% (or 20%) to the validation set, and the remaining 15% (or 20%) to the test set.

3. Ensuring representative distribution: For imbalanced datasets, it is crucial to maintain the same class distribution in all subsets. This can be achieved using stratified sampling, which ensures that each subset has a similar proportion of each class. In credit risk, handling imbalanced datasets is very usual, as the proportion of defaulted (i.e. “bad”) applicants is too small.

4. Separating features and the target variable: Divide each subset into features (input variables) and target variable (output variable) to prepare them for the model training, validation, and evaluation phases.

5. Training the model: Use the training set to train the ML model. The model learns patterns and relationships within the training data by adjusting its parameters to minimize errors.

6. Validating the model: The validation set is used during the training phase to fine-tune the model. By evaluating the model on the validation set, we can tune hyperparameters, prevent overfitting, and make decisions about model improvements. The validation set acts as a proxy for the test set, allowing adjustments without biasing the final evaluation.

7. Evaluating the model: After the model is trained and validated, the test set is used for the final evaluation. This involves applying the model to the test data and comparing the predicted labels with the actual labels to calculate metrics such as Accuracy, Precision, Recall, and F1 score. The test set should not be used during training to provide an unbiased assessment of the model's performance on unseen data.

By incorporating a validation set in addition to the training and test sets, the train-test split procedure ensures a robust evaluation framework. This approach helps in developing a model that not only performs well on the training data but also generalizes effectively to new, unseen data, ensuring its practical applicability. It is also extremely important to note that all the data pre-processing steps applied, as well as the feature engineering, should be done upon using only the training set, in order to obtain an unbiased estimation of the model's capability to generalize well on unseen data, hence to prevent what is called "data leakage". In specific, if Min-Max scaling is applied to the feature engineering phase, if all the dataset is being used, the Min and Max values will potentially differ from the corresponding Min-Max values of the training set alone. If done so, information from the test set (i.e. different Min-Max values) will have been used for the scaling of the features, thus the valuation metrics will be biased, by under-estimating the true error.

In credit risk and PD modelling, the training set is often called as the "Development sample", and the test set is often called as the "Validation sample" (Brown I., 2014).

4.4 Evaluation of Potential Predictors

The evaluation of potential predictors is a crucial step in the development of robust predictive models. When dealing with dataset of hundreds of potential predictors, it is essential to evaluate each individual predictor and assess its predictive power (on univariate level). This procedure should be done by having in mind not only

the predictive power of each individual characteristic, but also the conceptual soundness of the patterns captured by the potential predictor. This step can also be regarded as an initial "manual feature selection" step, as predictors with low predictive power and/or predictors with no conceptually sound interpretation will be excluded from the pool of the potential predictors, hence, pruning the number of potential features that will be assessed during the training of the model. This not only saves up valuable time, but also allows the statistical packages to identify stronger and more robust underlying relationships, thus reducing potential noise (Raymond A., 2007).

In binary problems and in PD modelling, two widely used techniques for assessing the predictive power of potential variables are the "Weight of Evidence" (WoE) and IV. These methods provide insights regarding the relationship between predictor variables and the target variable. This can be achieved by segmenting (or *discretizing*, or *binning*) the potential predictor into groups -based on its relationship with the target variable. Such methods are often called as "target-guided encoding". These methods also create models which are robust to outliers, since extremely low values of the corresponding predictor will be assigned to the lowest bin, whereas extremely high values will be assigned to the top bin.

Hence, WoE and IV are measures used to quantify the predictive power of each individual predictor in relation to a binary target variable, by transforming categorical and continuous variables into a more interpretable form, facilitating the comparison of different predictors. WoE and IV not only handle outliers by also providing a clear indication of the relationship between each bin of the predictor and the target variable -but also handle missing values effectively (as missing values can also be categorized on a specific segment) and ensure that the transformed variable has either a monotonic relationship with the target variable -or a non-linear relationship- such as "U-shape". Identifying linear and non-linear relationships is crucial in credit risk modelling. For example, the relationship of the applicant's "age" has a U-shape relationship with the default-rate (i.e. the risk-level of the applicant) (Douglas L. W., 2006), (Bruce L., David B., 2013).

For example, it is usual seen that applicants of Age 18-25 and 60+, are the most risky segments, whereas applicants between the Age of 40-50 are the best performing applicants. The other remaining age groups are more of a middle risk-level, thus having a U-shape relationship with the corresponding risk (i.e. PD). Of course, the above segmentations can change, based on the product of loan, the country in which the financial institution operates, and various other factors, whereas the relationship of the applicants age with the corresponding risk level can also have a monotonic (i.e. linear) relationship. It is now more evident that identifying the underlying relationship between the potential predictor with the target variable is crucial.

The WoE for the i-th bin of a predictor is calculated by using the formula

$$\text{WOE}_i = \ln\left(\frac{\frac{Number\ of\ Positives_i}{Total\ Positives}}{\frac{Number\ of\ Negatives_i}{Total\ Negatives}}\right) = \ln\left(\frac{\%\ of\ Positives_i\ (over\ total\ Positives)}{\%\ of\ Negatives_i\ (over\ total\ Negatives)}\right) \quad (4.4.1)$$

In the context of credit risk, WoE is calculated as follows

$$\text{WoE}_i = \ln\left(\frac{\frac{Number\ of\ Goods_i}{Total\ Goods}}{\frac{Number\ of\ Bads_i}{Total\ Bads}}\right) = \ln\left(\frac{\%\ of\ Goods_i\ (over\ total\ Goods)}{\%\ of\ Bads_i (over\ total\ Bads)}\right)$$

The percentage of goods (or positive events) of the i-th bin is the proportion of non-defaulters (e.g. good applicants) in the i-th bin of the predictor variable over the total number of good applicants (or positive events). In essence, it quantifies how many of the total good applicants (or positive events) belong to the i-th bin of the corresponding predictor. Subsequently, the percentage of bads (or negative events) of the i-th category measures the same thing -but for the population of defaulted applicants (i.e. bad applicants).

IV is a metric that summarizes the predictive power of a variable across all its bins. IV is calculated using the WoE values and the distributions of goods and bads across the bins:

$$\text{IV} = \sum\left(\frac{Number\ of\ Positives_i}{Total\ Positives} - \frac{Number\ of\ Negatives_i}{Total\ Negatives}\right) \times WoE_i \times 100$$

$$\text{IV} = \sum(\%\ of\ Positives_i - \%\ of\ Negatives_i) \times WoE_i \times 100 \quad (4.4.2)$$

In the context of credit risk, IV is calculated as follows

$$\text{IV} = \sum\left(\frac{Number\ of\ Goods_i}{Total\ Goods} - \frac{Number\ of\ Bads_i}{Total\ Bads}\right) \times WoE_i \times 100$$

$$\text{IV} = \sum(\%\ of\ Positives_i - \%\ of\ Negatives_i) \times WoE_i \times 100$$

with the notations being the same as those in WoE.

The IV value helps in ranking predictors and identifying the most significant variables for the model. Generally, the IV can be interpreted as follows:

- IV < 2: Predictors with no predictive power.
- 2 ≤ IV < 10: Predictors with weak predictive power.
- 10 ≤ IV < 30: Predictors with medium predictive power.
- IV ≥ 30: Predictors with strong predictive power.

Steps for evaluating predictors using WoE and IV:

1. Data Preparation:
    - Split of the data into bins for each predictor variable. Binning (or segmentation/discretization) can be done using techniques like equal-width binning, equal-frequency binning, or based on domain

knowledge. The optimal binning is to proceed manually, thus segmenting the features based on the corresponding outcome variable -also incorporating the domain knowledge.
    - Calculation of the proportion of goods and bads for each bin.
2. Calculation of WoE:
    - Computation of the WoE for each bin of the predictor variables using the formula of (4.4.1).
3. Calculate IV:
    - Using the WoE values and the proportions of positives (goods) and negatives (bads), the IV for each predictor variable is being calculated.
4. Interpretation of the results:
    - Analyzation of the WoE/IV values, to understand the relationship between predictor variables and the target variable.
    - By utilizing the IV values, the predictors can be ranked in descending order based on the predictive power -which is now quantified. The most significant variables are being selected as potential predictors for the training of the model.

In conclusion, by systematically applying WoE and IV, we can effectively evaluate and select the most relevant predictors for our model. These techniques not only help in understanding the impact of each predictor but also improve the overall predictive performance of the model by focusing on the most informative variables.

4.5 Model Training

Model training is a critical phase in ML where a chosen algorithm learns from data to make predictions or decisions. The objective is to optimize the model's performance by adjusting its parameters to minimize errors. This involves several key steps, including data preprocessing, selecting appropriate features, choosing an optimization metric, and tuning hyperparameters. Techniques such as cross-validation ensure the model generalizes well to new data, preventing overfitting. Effective model training leads to robust, accurate, and reliable models capable of solving real-world problems. As already described, it is crucial that all the data pre-processing and feature engineering have to be implemented using only the training set. The validation/test sets must be used only for fine-tuning and for unbiased measure of the model’s performance respectively.

4.5.1 Optimization Metrics

In model training, various metrics can be optimized to evaluate and enhance the performance of predictive models. KS, Gini, and AUROC are widely used in credit

scoring to measure the discriminatory power of the models, and all three metrics constitute industry standards. In the following section are presented some of the most well-known and widely used metrics (Zheng Y., Yue W., et al. 2004), (Joseph L. G., 1972), (Andrew P. B., 1997).

4.5.1.1 Kolmogorov-Smirnov (KS) Statistic

The KS statistic measures the maximum difference between the cumulative distribution functions (CDF) of the predicted positive (good) and negative (bad) events (i.e. defaults, non-defaults).

$$KS = \max |F_1(x) - F_0(x)|$$

where $F_1(x)$ and $F_0(x)$ are the CDFs of the positive and negative classes, respectively.

4.5.1.2 Gini Index

The Gini Index has already been presented. Nevertheless, for completeness purposes, the formula is presented below.

$$Gini = 2 \times \text{AUC} - 1$$

where AUC is the Area Under the Receiver Operating Characteristic Curve.

4.5.1.3 Area Under the ROC Curve - AUROC

In binary classification, predictions are often based on a continuous score for each instance, such as an estimated probability from LR. A threshold parameter T is used to make the classification: if the score X is greater than T (X > T), the instance is classified as "positive"; otherwise, it is classified as "negative". The score X follows a probability density function $f_1(x)$ for positive instances and $f_2(x)$ for negative instances. Therefore, the true positive rate is given by

$$TPR(T) = \int_T^{\infty} f_1(x)\, dx$$

and the false positive rate is given by

$$FPR(T) = \int_T^{\infty} f_0(x)\, dx$$

The ROC curve plots parametrically TPR(T) versus FPR(T) with T as the varying parameter. The AUROC is the Area Under the ROC curve and is calculated as follows

$$\text{AUC} = \int_0^1 ROC(t)dt$$

where t is the false positive rate.

AUROC provides a single metric to evaluate the model's performance across all classification thresholds.

4.5.1.4 Accuracy – Precision – Recall - F1 Score - Confusion Matrix

Precision measures the proportion of true positive predictions among all positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

where TP is True Positives and FP is False Positives.

Hence, it indicates the accuracy of positive predictions.

Recall, or Sensitivity, measures the proportion of true positive predictions among all actual positive instances

$$\text{Recall (or Sensitivity)} = \frac{TP}{TP+FN}$$

where FN is False Negatives.

Hence, it reflects the model's ability to capture all actual positive instances.

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of those two

$$\text{F1-Score} = \frac{2Precision*Recall}{Precision+Recall}$$

It constitutes a balanced metric of precision and recall, particularly useful in imbalanced datasets.

Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

where TN is True Negatives.

Hence, it provides an overall "correctness" measure, though it can be misleading in imbalanced datasets.

4.5.2 Cross-Validation

In many ML and statistical models, parameters must be selected by the model developer according to the specific case being examined. Therefore, the developer must choose a value for each parameter that will affect the overall performance of the model. A common technique for selecting such parameter values is cross-validation (Jerome H. F., Robert T. et al., 2009). A quick summary of the procedure follows:

1. Data partitioning: Split of the training set into k mutually exclusive subsets (folds).

2. Model tuning:
    - Choose a metric (e.g., accuracy) to evaluate performance.
    - For a parameter (e.g. β), select a range of possible values.
    - Train k models using k−1 folds for training -and the remaining fold for validation of the performance on unseen data.
    - Repeat this process k times -each time with a different fold as the validation set.

3. Evaluation: computation of the average performance metric (e.g. accuracy) across all k models for each parameter value.

4. Selection: Choose the parameter value that yields the highest average performance metric.

Hence, cross-validation provides a more reliable estimate of model performance by using multiple train-test splits. It also helps in identifying the parameter values that optimize the model's performance. By employing cross-validation, model developers can ensure that their models are both accurate and generalizable to unseen data. This technique is essential for fine-tuning model parameters and achieving the best possible performance.

4.5.3 Feature Selection Methods

When building a model, it is often necessary to select a subset of features that either maximize performance according to a metric or maintain performance while minimizing the number of variables used. This is especially crucial in problems with a large number of explanatory variables. Selecting the right features can enhance model performance or keep it satisfactory while limiting the number of variables, thus making it more interpretable (Jerome H. F., Robert T. et al., 2009).

Specifically, feature selection enhances the model's performance, by selecting the most relevant features, the model can achieve higher accuracy and generalizability. Additionally, it reduces overfitting by limiting the number of features and helps prevent the model from learning noise in the training data. Thus, a model with fewer features is easier to understand and interpret. Finally, reducing the number of features can lead to faster training and prediction times, thus saving up valuable computational time.

Here, we describe three classic feature selection techniques: Forward Selection, Backward Elimination and Stepwise Selection.

4.5.3.1 Forward Selection

In forward selection, the statistical package is initialized with an “empty” model and iteratively adds the variable that maximizes model performance according to a chosen metric. The process continues until adding more variables does not significantly improve the model's performance or until a pre-defined limit on the number of variables is reached.

Steps:

- Begin with no variables.
- Add the variable that provides the highest performance improvement.
- Repeat until adding more variables yields minimal performance gains or a limit is reached.

4.5.3.2 Backward Elimination

Backward elimination starts with the full model containing all variables. The statistical package iteratively removes variables that do not meet a statistical or mathematical criterion or whose removal negligibly impacts model performance. This process continues until the desired number of variables is achieved or significant information loss occurs.

Steps:

- Start with all variables.
- Remove the least significant variable.
- Repeat until performance degrades significantly or a target number of variables is reached.

4.5.3.3 Stepwise Selection

Stepwise selection is a combination of forward selection and backward elimination. It allows variables to be added or removed at each step based on specific criteria. This method iteratively evaluates both the inclusion and exclusion of variables to optimize the model.

Steps:

- Begin with an empty model.
- Add the variable that adds the most significant information (based on the optimization criterion).

- Right after the addition of a variable, check whether the removal of each variable contained on the model -on the corresponding step, will improve the performance metric.
- Continue until no further significant improvements can be made by adding or removing variables.

4.5.4 Hyperparameter Tuning

Hyperparameter tuning is essential for optimizing ML models. Hyperparameters are configuration settings used to control the learning process and model architecture, such as learning rate, number of layers, and regularization parameters. Proper tuning of these hyperparameters can significantly improve model performance, prevent overfitting, and ensure the model generalizes well to unseen data. It involves searching for the best set of hyperparameters that result in the highest model performance based on a specified evaluation metric (Petro L. and Pavlo L., 2019), (Jia W., Xiu-Yun C. et al., 2019).

4.5.4.1 Grid Search

Grid Search is an exhaustive search method that evaluates all possible combinations of a predefined set of hyperparameters. It systematically builds and assesses a model for every combination of hyperparameter values provided in a grid-like structure.

Advantages:

- Simple to implement.
- Guarantees finding the optimal combination within the provided pre-defined search space.

Disadvantages:

- Computationally expensive, especially with large datasets and many hyperparameters.
- Can be extremely inefficient if the search space is large or contains many irrelevant combinations.

4.5.4.2 Randomized Search

Randomized search, unlike grid search, samples a fixed number of hyperparameter combinations from the specified range randomly. It does not try every possible combination but rather a random subset.

Advantages:

- More efficient than grid search, as it evaluates fewer combinations.
- Can potentially find a good combination faster, especially in large search spaces.

Disadvantages:

- Does not guarantee finding the optimal combination.
- The results may vary between runs due to its random nature.

4.5.4.3 Bayesian Optimization

Bayesian Optimization is a more advanced method that builds a probabilistic model of the function mapping hyperparameters to the objective function (model performance). In the following part of this sub-section, a description of Bayesian optimization will follow. For more detailed information refer to Jia W., Xiu-Yun C. et al. (2019).

Bayesian optimization is an effective method for solving functions that are computationally expensive to find the extrema. It can be applied for solving a function which does not have a closed-form expression, is expensive to calculate or has derivatives that are hard to evaluate. The optimization goal in bayesian optimization is to find the maximum value at the sampling point for an unknown function f(x)

$$x_+ = argmax_{x \in A} f(x)$$

where A denotes the search space of x. Bayesian optimization derives from Bayes' theorem, which states that the posterior probability P(M|E) of a model M -for given the data E- is proportional to the conditional likelihood P(E|M) of observing E given model M, multiplied by the prior probability P(M). Thus, the following relationship describes the main idea of Bayesian Optimization

$$P(M|E) \propto P(E|M)P(M)$$

The principle of Bayesian optimization is to combine the prior distribution of the function f(x) with the sample information E, to obtain the posterior of the function. The posterior information is then used to find where the function f(x) is maximized according to a specified criterion. This criterion is represented by a utility function u(.) that is also known as the *Acquisition Function*. This function is used to determine the next sample point to maximize the expected utility.

In bayesian optimization, the prior distribution of the function f(.) is often assumed to be a gaussian process due to its flexibility and ease of handling. A gaussian process is a collection of random variables, any finite number of which have a joint gaussian distribution, and this particular characteristic makes it suitable for modeling the distribution over functions. The gaussian process is used to fit the data and update the posterior distribution as new data points are observed.

The acquisition function u(.) is a crucial component in bayesian optimization - it determines the next point to sample by balancing exploration (sampling from areas of high uncertainty) and exploitation (sampling from areas with high predicted values). This balance helps reduce the number of samples needed to find the maximum of the function, especially when the function has multiple local maxima. Common acquisition functions include Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB).

Due to its nature of exploring areas with high uncertainty and exploiting areas with high predictive values, bayesian optimization is extremely suitable for hyperparameter tuning in ML models due to its efficiency in finding optimal values with fewer samples and without needing the explicit expression of the function. This method is applied to various complex models such as RF, artificial NN. Experiments carried out on standard datasets, demonstrate that bayesian optimization can achieve high accuracy and significantly reduce runtime compared to manual search (Jia W., Xiu-Yun C. et al., 2019).

Advantages:

- Much more sample-efficient than grid and randomized search.
- Can find better hyperparameter combinations with fewer evaluations.

Disadvantages:

- More complex to implement efficiently -since it has many different aspects that the developer should take into account
- Requires more computational overhead to maintain and update the probabilistic model.

4.6 Performance Monitoring

Monitoring the performance of statistical and ML models is crucial for several reasons. Effective monitoring ensures that the model remains accurate, reliable, and relevant over time, especially when deployed in real-world applications where data can evolve (Amos T. O., 2019), (Bruce L. and David B., 2013), (Andrew P. B., 1997), (Bilal Y. and Joshua N., 2021), (Brown I., 2014).

Below are the key points highlighting the importance of monitoring model performance:

- The model's performance should be regularly validated using out-of-sample data (i.e. data not used during training). In credit risk, industry-standard metrics include the KS, Gini Index, and AUC. The goal of PD-model validation is to ensure the model continues to effectively discriminate between "good" and "bad" applicants. Performance validation quantifies how accurately the model's predictors produce estimates by comparing metrics from the training (development) sample with those from the test (validation) sample

(multivariate-level analysis).

- Additionally, apart from evaluating the overall performance of the model (multivariate level), it is crucial to validate the performance of each independent predictor that is used in the model. In PD modelling, this can be achieved by using the IV criterion (Bruce L. and David B., 2013). Thus, the performance of each characteristic is examined upon comparing the distributions and the defaults versus non-default events within each segment of the used predictors, in both the training (development) and test (validation) samples.
- Apart from the above mentioned, it is crucial to check whether the used characteristics are being stable over-time. If the distribution(s) of the used predictor(s) change over-time, then the model's estimations are likely to change as well. To achieve this, the PSI is used, by comparing the distributions of the characteristics that were used during the training of the model, with the distributions of the characteristics on the test set (Bilal Y. and Joshua N., 2021).

In summary, these monitoring activities help stakeholders determine if the model continues to perform as expected or if there is any deviation. If performance deterioration is detected, IV and PSI analyses provide insights into why the model is underperforming and what actions can be taken to retrain or recalibrate it to restore satisfactory performance.

CHAPTER 5

Implementation of Supervised Statistical and Machine Learning Methods for PD Estimation

5.1 Introduction to the dataset and the analyses that will follow

The data used for the following analyses are sourced from LendingClub.com, where they were publicly shared. Below is a dictionary providing descriptions for the corresponding columns of the dataset that were used. Some variables lacked descriptions, and those that were not self-explanatory were omitted. Additionally, some variables that were not relevant to the analysis (i.e. the information was unavailable at the time of the application and/or was retrospectively updated) were further dropped. Further analyses were made (conceptual soundness & correlation checks) before concluding to the final set of variables that were used for the analysis.

Below follows a description of each characteristic contained on then dataset:

Column	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
chargeoff_within_12_mths	Number of charge-offs within 12 months
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.

Column	Description
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
fico_range_high	The upper boundary range the borrower’s FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower’s FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
issue_d	The month which the loan was funded
last_fico_range_high	The upper boundary range the borrower’s last FICO pulled belongs to.
last_fico_range_low	The lower boundary range the borrower’s last FICO pulled belongs to.
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
mort_acc	Number of mortgage accounts.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_inq	Months since most recent inquiry.
mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_rev_accts	Number of revolving accounts
open_acc	The number of open credit lines in the borrower's credit file.
open_acc_6m	Number of open trades in last 6 months

Column	Description
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
out_prncp	Remaining outstanding principal for total amount funded
policy_code	publicly available policy_code=1 new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
total_acc	The total number of credit lines currently in the borrower's credit file
total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment accounts
total_cu_tl	Number of finance trades
total_pymnt	Payments received to date for total amount funded
total_rec_prncp	Principal received to date
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

Table 5.1.1: Dataset Dictionary

5.1.1 Target Variable

The target variable was created based on the “loan_status” column. Specifically, the mapping of good-bad applicants (i.e. the creation of the target variable) is based on the descriptions provided by LendingClub for the values of the “loan_status” field:

- Current: Loan is up to date on all outstanding payments.
- In Grace Period: Loan is past due but within the 15-day grace period.
- Late (16-30): Loan has not been current for 16 to 30 days.
- Late (31-120): Loan has not been current for 31 to 120 days.

- Fully paid: Loan has been fully repaid, either at the expiration of the 3-year or 5-year term or as a result of a prepayment.
- Charged Off: Loan for which there is no longer a reasonable expectation of further payments.

Thus, borrowers are defined as "Defaulted" if the "loan_status" field is equal to "Charged Off ", "Late (16-30 days)", or "Late (31-120 days)". The following mapping is applied:

Current	**Good**
In Grace Period	
Fully paid	

Charged Off	**Bad**
Late (16-30)	
Late (31-120)	

Table 5.1.1.1: Target variable mapping – G/B definition

5.1.2 Dropped Variables

Some variables eirther had too many values or were constant/quasi-constant and thus were further dropped. Specifically:

- Zip_code – It has too many values and no concentration is identified in any specific value. If kept, the ML models would probably overfit.
- policy_code – Constant, having only one value, and thus was dropped.
- chargeoff_within_12_mths - Quasi-constant, having only two values, and thus was dropped.

chargeoff_within_12_mths	Frequency	Percentage	Cumulative Percentage
0.0	237159	99.381066	99.381066
1.0	1374	0.575772	99.956838
2.0	81	0.033943	99.990781
3.0	13	0.005448	99.996229
4.0	6	0.002514	99.998743
6.0	1	0.000419	99.999162
7.0	1	0.000419	99.999581
9.0	1	0.000419	100.000000

Table 5.1.2.1: Chargeoff_within_12_months – Quasi-constant

- acc_now_delinq - Quasi-constant, having only two values, and thus was dropped.

acc_now_delinq	Frequency	Percentage	Cumulative Percentage
0.0	238610	99.989105	99.989105
1.0	26	0.010895	100.000000

Table 5.1.2.2: acc_now_delinq– Quasi-constant

The resulted dataset contained 60 columns and 238,638 obsevations, with applications spawning from 01/2018 to 2018/06.

5.2 Train-Test Split

In credit risk and PD model development, the training and test sets cover different time periods, with the test set representing data from a later period than the training set. This means the test set is an out-of-period sample rather than being drawn from the same timeframe as the training data. As a result, the test set will include data from months that follow the period covered by the training set. Due to the very large volumes of the dataset, and for time-efficiency purposes, as training set were selected applications from to 01/2018 to 03/2018, covering a period of 3 months and containing 107,864 approved applications. As a test set, applications from 04/2018 to 05/2018 were selected, covering a period of 2 months and containing 42,928 approved applications. The train-test-split was made prior to any data management/feature engineering actions, as to prevent any data leakage from the training to the test set.

5.3 Correlation Data Analysis and more Data pruning

In any analysis multicollinearity can be a significant issue. Multicollinearity occurs when two or more predictors are highly correlated, leading to redundancy and potentially misleading the results of a model. To address this, one common approach is to drop one feature from each pair of highly correlated predictors. Here, we will also consider the conceptual soundness of the analysis to ensure that we retain the most meaningful predictors.

Below, are depicted the features that are correlated with a pearson's statistic of 0.85 or above -in absolute terms:

Feature 1	Feature 2	Correlation
loan_amnt	funded_amnt	1.000000
loan_amnt	installment	0.944957
funded_amnt	installment	0.944957
fico_range_low	fico_range_high	1.000000
mths_since_last_delinq	mths_since_recent_revol_delinq	0.860612
total_pymnt	total_rec_prncp	0.975128
last_fico_range_high	last_fico_range_low	0.876065
total_bal_il	total_bal_ex_mort	0.907411
mths_since_recent_bc_dlq	mths_since_recent_revol_delinq	0.893902

Table 5.3.1: Correlated Pairs (absolute value)

From the above pairs, "fico_range_low", "mths_since_last_delinq", "total_pymnt", mths_since_recent_bc_dlq" were kept. Hence, the following variables were dropped:

- loan_amnt
- funded_amnt
- fico_range_high
- mths_since_recent_revol_delinq
- pub_rec_bankruptcies
- total_rec_prncp

- last_fico_range_low
- total_bal_il
- total_bal_ex_mort

5.4 Feature Engineering & Information Value Segmentation Analysis

Figures for the distributions and the outliers of the (unsegmented) potential predictors are presented in Appendix.

Before proceeding with the IV analyses, missing values were encoded as “-1” for continuous variables, and as “MISSING” for string values.

Furthermore, rounded versions of the continuous variables were created, in order to be processed more efficiently, by capturing potential higher concentrations in certain values. Thus, the following variables were created:

- dti_rounded
- instlmnt_round
- Annual_Inc_round
- instlmnt_to_Annual_Inc
- revol_bal_round
- revol_util_round
- total_pymnt_round
- recoveries_round
- avg_cur_bal_round

Another variable was created, named as “years_with_Credit_line”, which counts the years since the first credit line of the applicant, based on the existing variable of the dataset, that of “earliest_cr_line”.

To analyze the distribution of "Good" and "Bad" borrowers in a training dataset, a Python function that outputs these distributions into separate excel files was created. Additionally, this function calculates the WoE and IV for each value of the corresponding variable. The primary goal of this analysis is to group the variable values into meaningful categories (being conceptually sound and not misaligned), with each category consisting from population with similar default rates (a process known as binning or discretization). The objective is to assign borrowers with similar risk levels (i.e. default rates) to the same category while maximizing the Information Value (IV), which helps improve the predictive power of the model. Hence, the aim is to create a categorical version of the initial variable that better discriminates “Good” and “Bad” applicants. This approach allows for a direct categorization of predictors based on their observed relationship with the target variable. Hence, the IV analysis is conducted for each variable individually, ensuring optimal segmentation of their characteristics. Moreover, this methodology helps in creating variables that are robust to outliers, thereby enhancing the stability of predictive algorithms. By mitigating the risk of overfitting, this approach ensures more reliable and consistent predictions over time.

Finally, 26 potential predictors were optimally segmented. The categories for each predictor, along with their corresponding bad rates and the percentage of the total distribution within each category, are depicted in the figures below:

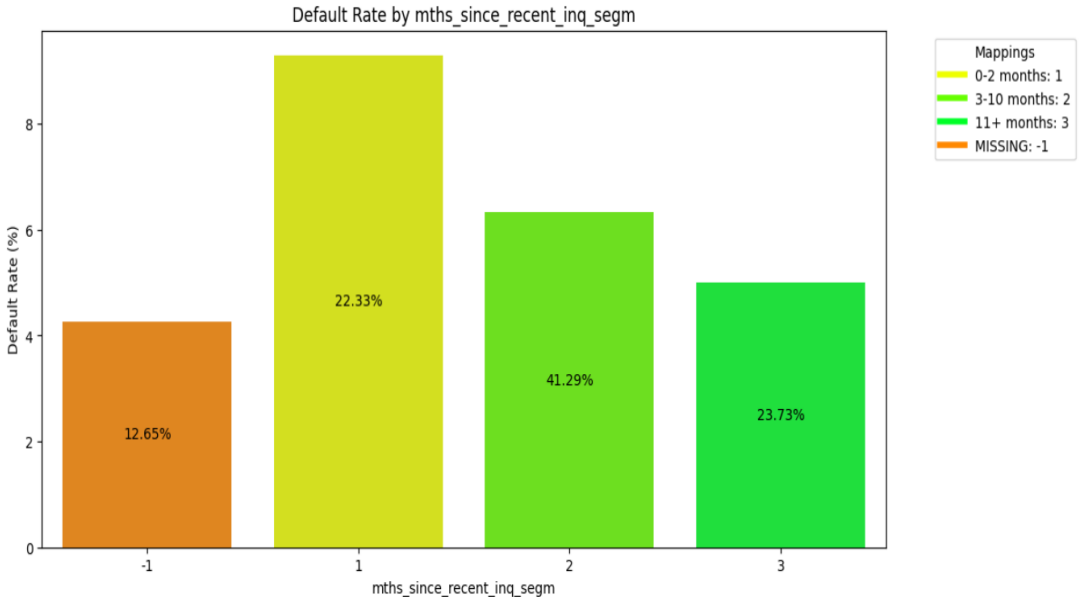


Figure 5.4.1: Months Since Last Inquiry

The predictor "Months Since Last Inquiry" is segmented into four categories based on the time elapsed since the borrower's last credit inquiry. The segments and their corresponding bad rates indicate that borrowers with more recent inquiries tend to have higher default rates. For instance, borrowers with inquiries within the last 2 months exhibit a higher bad rate compared to those with inquiries older than 3-10 months, and also to those of 11 months or more. This segmentation helps to identify higher-risk borrowers and contributes to more accurate risk assessments.

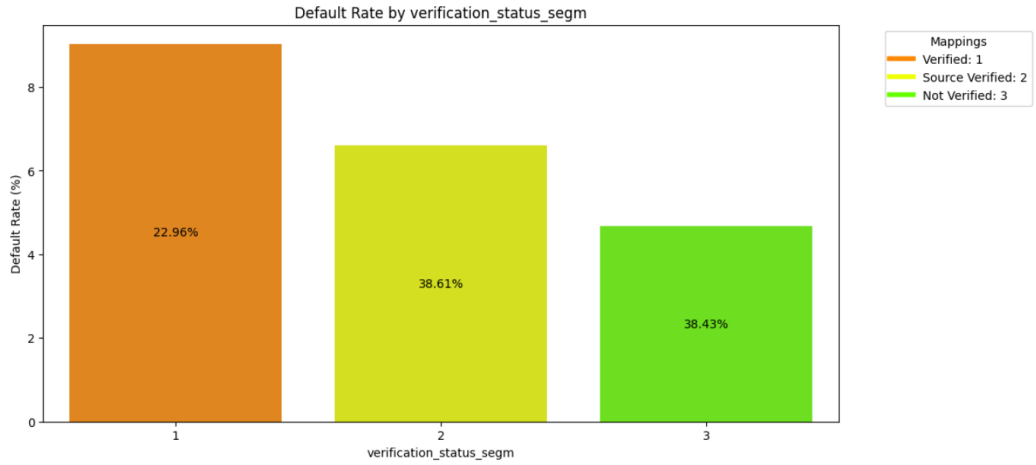


Figure 5.4.2: Verification Status

The "Verification Status" predictor is divided into Verified, Source Verified and Non-verified categories. Typically, verified borrowers show a lower bad rate compared to non-verified borrowers. However, in our case, the opposite is observed. This is

probably the outcome of a targeted verification from the financial institution. Specifically, the bank seems that does not look to verify the potential good profiles, whereas it does try to verify the riskier clients. Verification Status is a useful predictor in credit risk modeling.

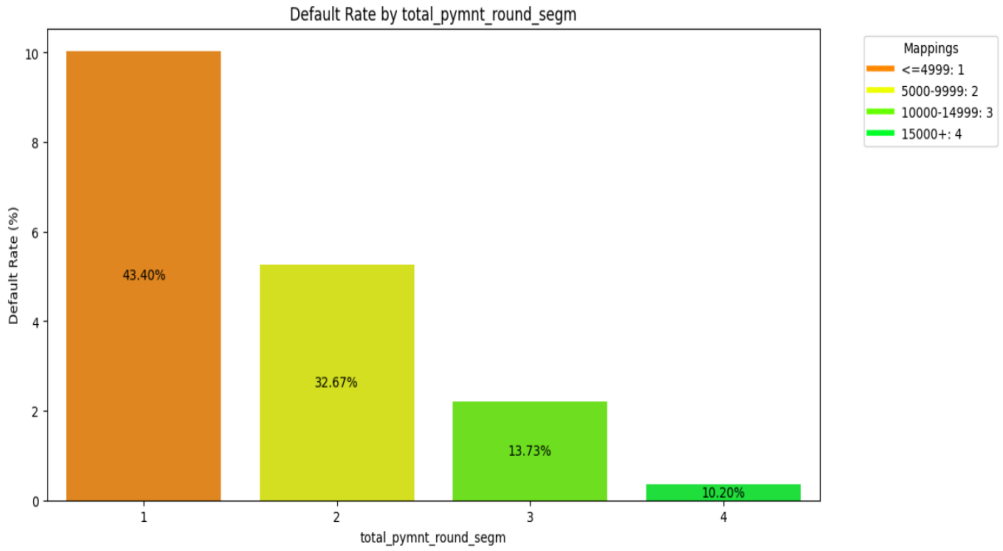


Figure 5.4.3: Total Payment Rounded

The "Total Payment Rounded" predictor segments borrowers based on their total payment amount, categorized into four distinct ranges. These segments are associated with varying default rates, reflecting the relationship between the total amount paid and the PD. By analyzing these segments, lenders can better understand how different payment levels impact default risk.

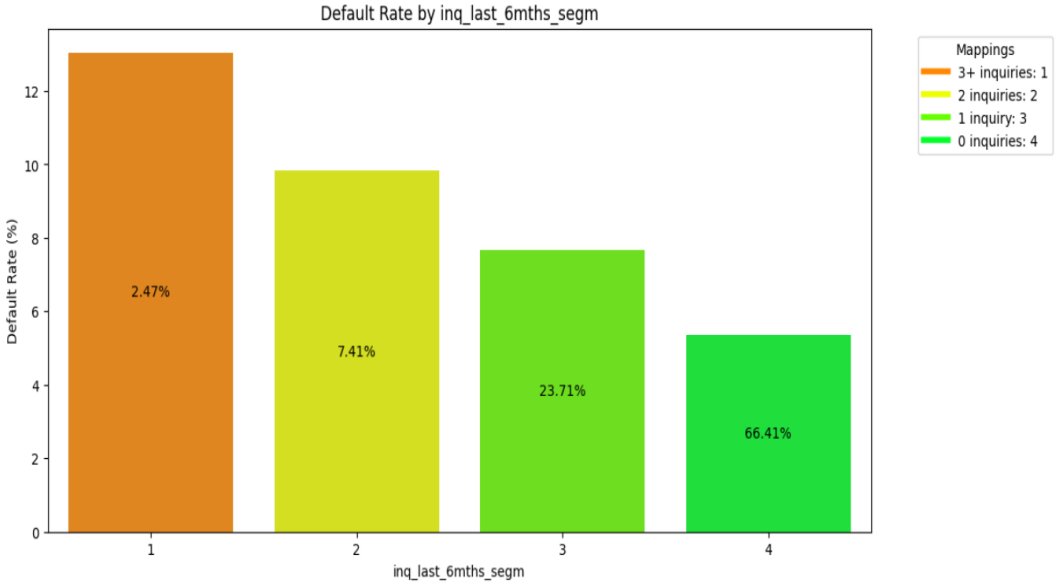


Figure 5.4.4: Inquiries Last 6 Months

The "Inquiries Last 6 Months" characteristic segments borrowers based on the number of credit inquiries made within the last six months. Higher numbers of inquiries

correlate with increased bad rates, indicating that frequent credit seeking behavior is associated with higher risk. Borrowers with no recent inquiries exhibit the lowest bad rates.

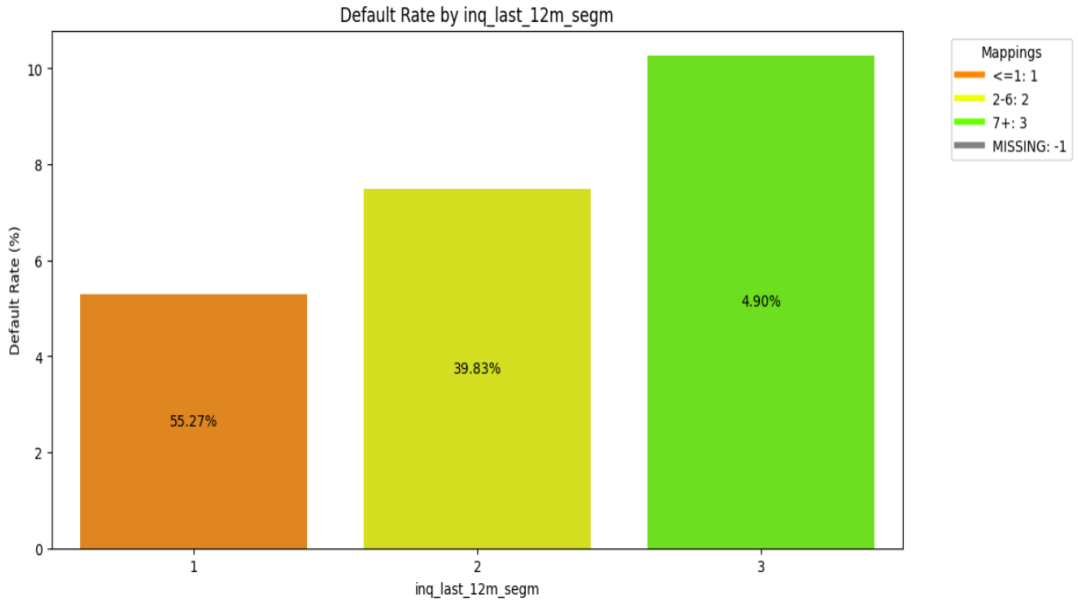


Figure 5.4.5: Inquiries Last 12 Months

Similar to the 6-month inquiry predictor, the "Inquiries Last 12 Months" predictor segments borrowers by the number of credit inquiries over the past year. Borrowers with multiple inquiries within this period tend to have higher bad rates, underscoring the risk associated with frequent credit applications.

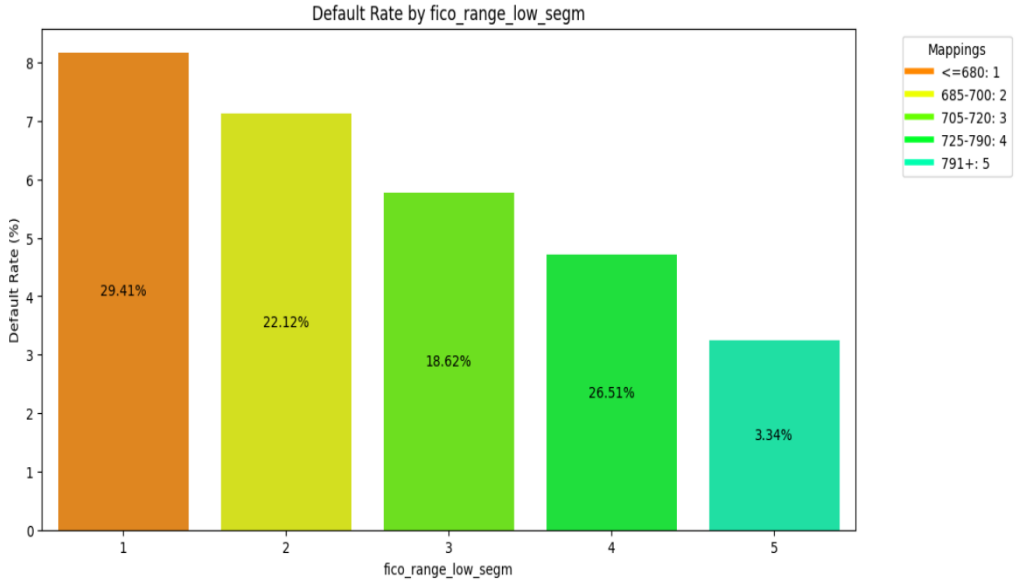


Figure 5.4.6: Fico Range Low

The "Fico Range Low" predictor segments borrowers based on their FICO scores (i.e. scored credit bureau information). Lower FICO score ranges correspond to higher bad rates, reflecting the well-established relationship between credit scores and

default risk. This segmentation is robust in identifying high-risk borrowers who might require closer monitoring or stricter lending criteria.

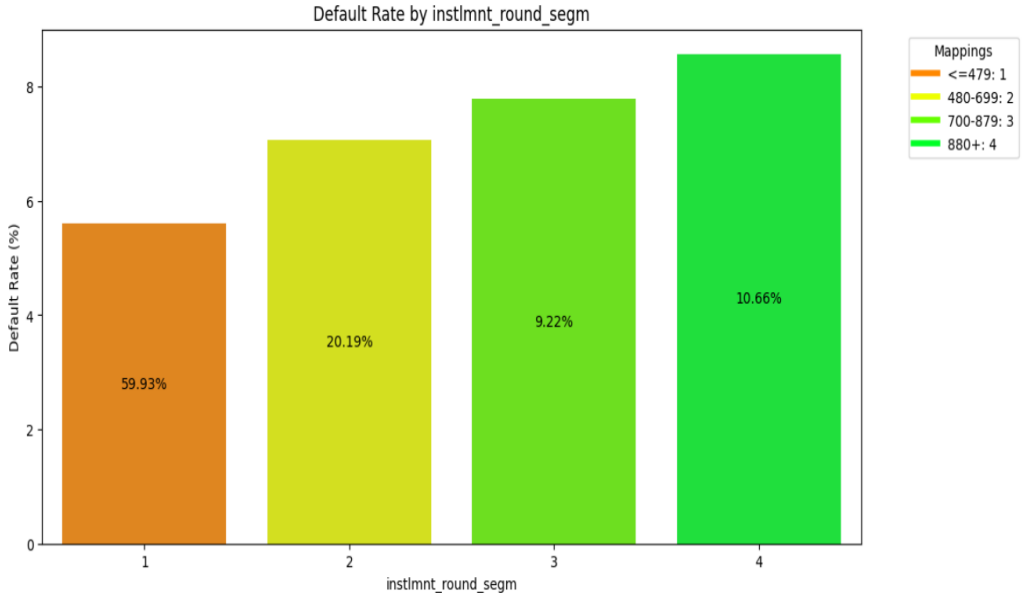


Figure 5.4.7: Instalment Rounded

The "Installment Rounded" characteristic groups borrowers by the rounded amount of their loan installments. Segments with higher installment amounts have higher bad rates, suggesting that the burden of larger loan repayments can impact the PD of the borrower.

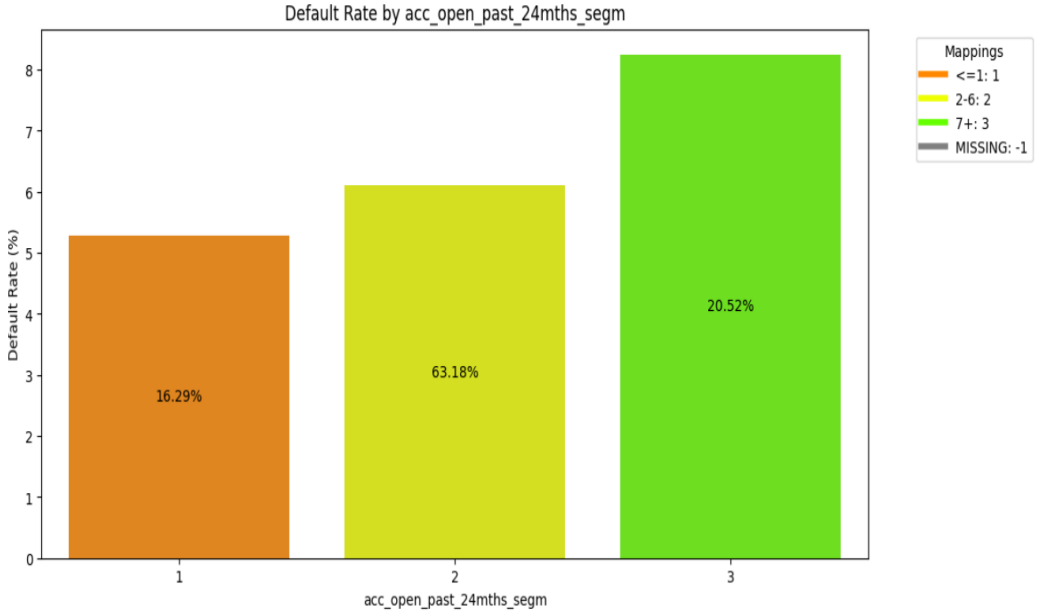


Figure 5.4.8: Accounts Open Last 24 Months

This predictor, "Accounts Open Last 24 Months" segments borrowers by the number of new accounts opened in the past two years. Higher numbers of newly opened accounts correlate with increased bad rates, indicating that aggressive credit seeking

behavior within this period is a risk factor.

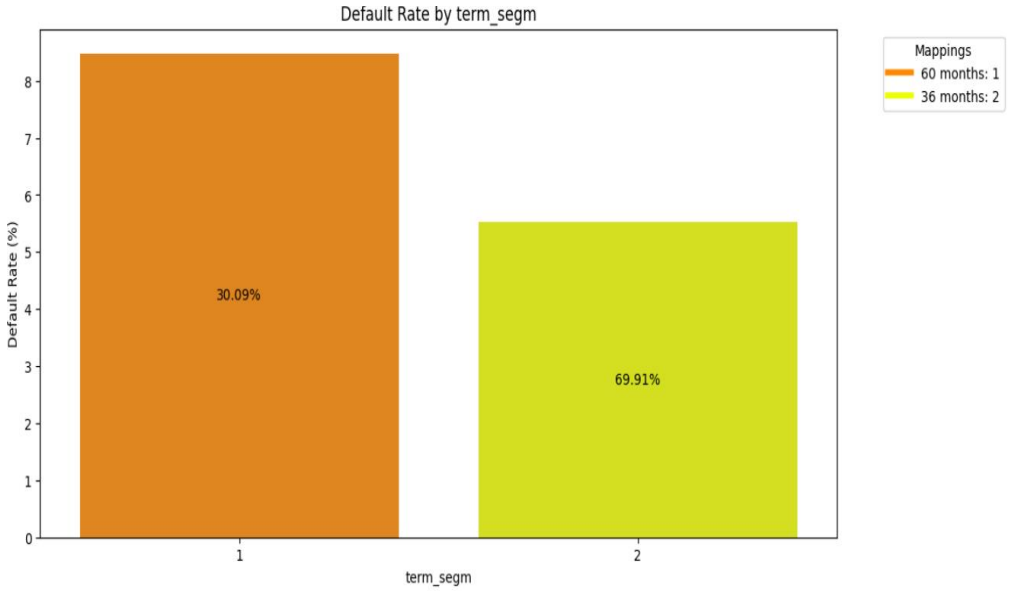


Figure 5.4.9: Loan Term

The "Loan Term" predictor categorizes borrowers based on the duration of their loans (e.g. 36 months, 60 months). Longer loan terms tend to have higher bad rates compared to shorter terms, suggesting that longer repayment periods is associated with greater uncertainty and risk.

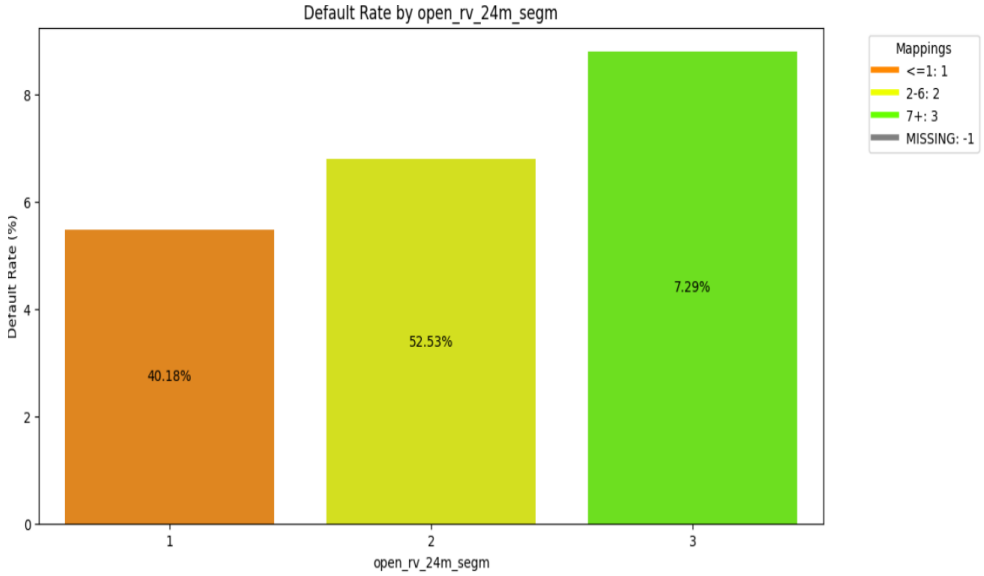


Figure 5.4.10: Open Revolving Accounts Last 24 Months

The "Open Revolving Accounts Last 24 Months" characteristic segments borrowers by the number of revolving credit accounts opened in the last two years. Higher numbers of open revolving accounts are associated with higher bad rates, reflecting the risk of “over-leveraging” through revolving credit.

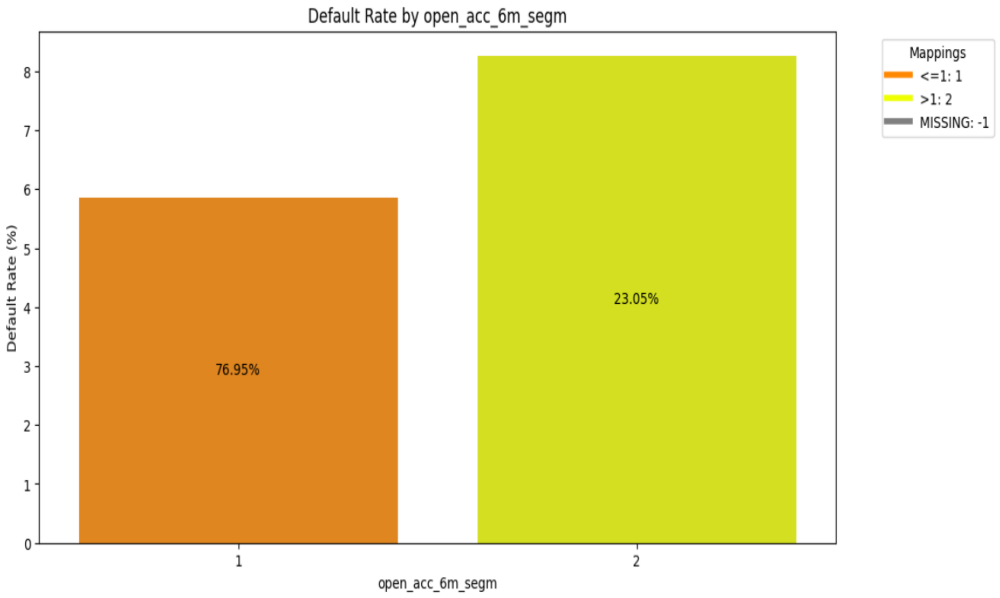


Figure 5.4.11: Open Accounts Last 6 Months

This predictor segments borrowers by the number of new accounts opened in the last six months. Similarly, a higher number of new accounts in this period correlates with higher bad rates, indicating increased risk.

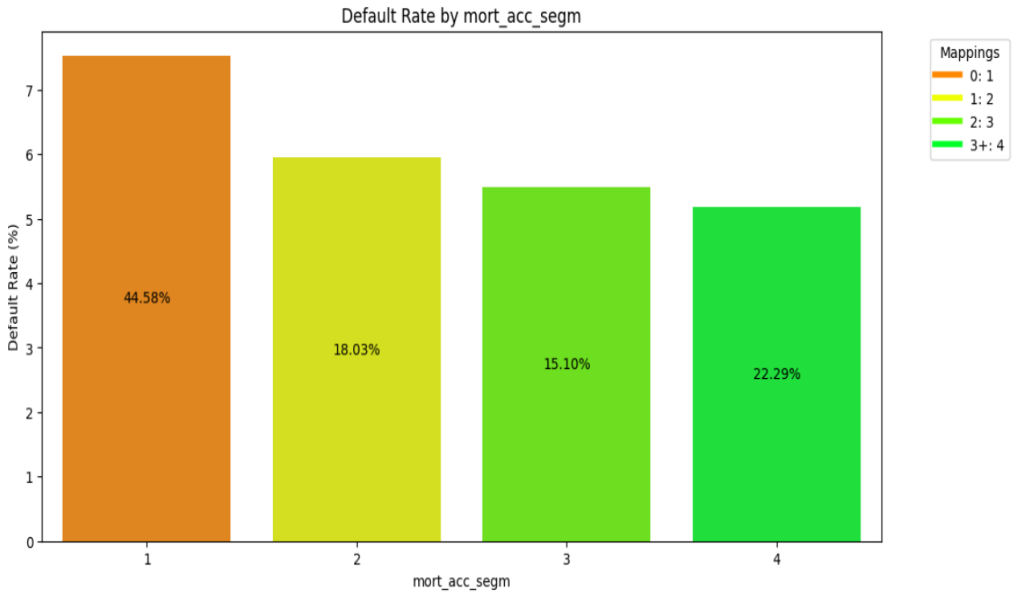


Figure 5.4.12: Mortgage Accounts

The "Mortgage Accounts" predictor categorizes borrowers based on the number of active mortgage accounts. Borrowers with multiple mortgage accounts tend to have lower bad rates, suggesting that owning multiple properties might be associated with greater financial stability. In specific, having more mortgage accounts indicates that the borrower has been approved for a number of mortgage loans in the past, indicating good repayment history.

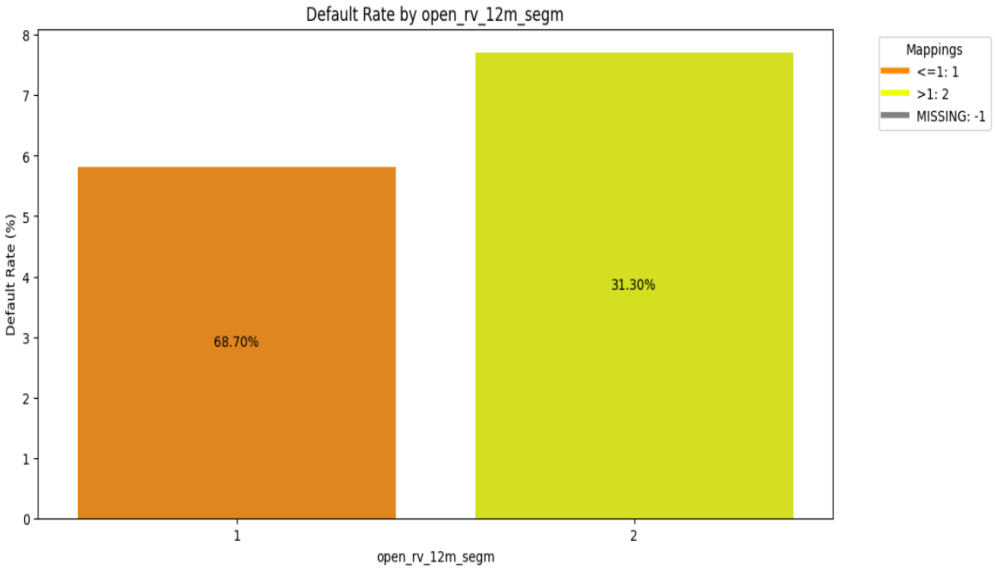


Figure 5.4.13: Open Revolving Accounts Last 12 Months

Similar to the 24-month predictor, the "Open Revolving Accounts Last 12 Months" segments borrowers by revolving accounts opened in the past year. More revolving accounts opened in this period are linked to higher bad rates, indicating increased credit risk.

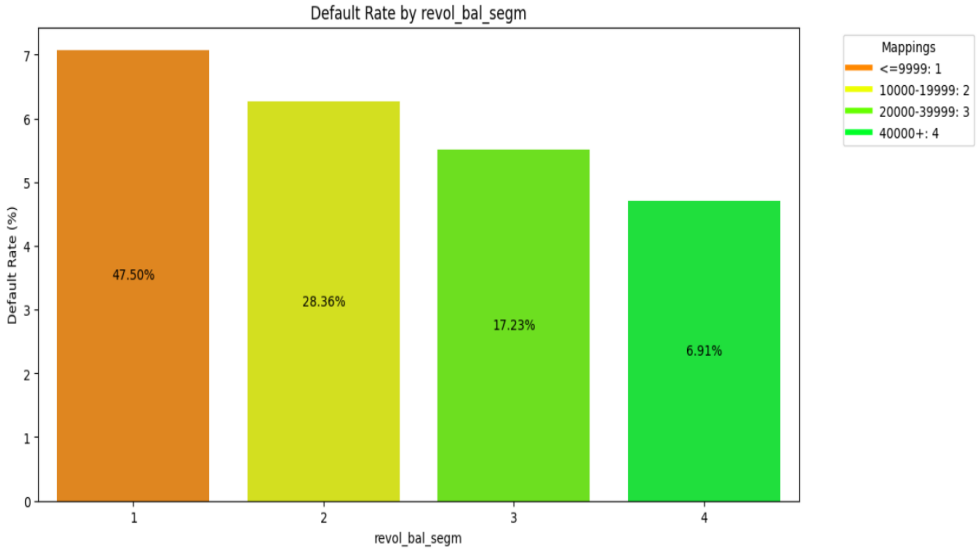


Figure 5.4.14: Balance of Revolving Accounts

The "Balance of Revolving Accounts" predictor segments borrowers based on the total balance across their revolving credit accounts. Higher balances are often associated with higher bad rates, reflecting the risk of having significant revolving debt.

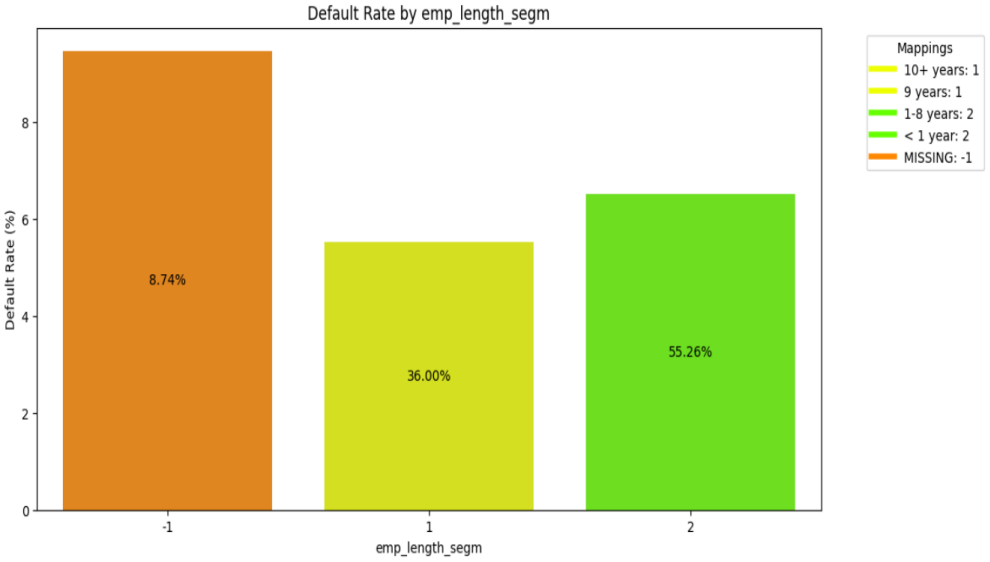


Figure 5.4.15: Years of Employment

The "Years of Employment" characteristic segments borrowers by their length of employment. Longer employment durations typically correlate with lower bad rates, indicating that stable employment is a positive indicator of creditworthiness.

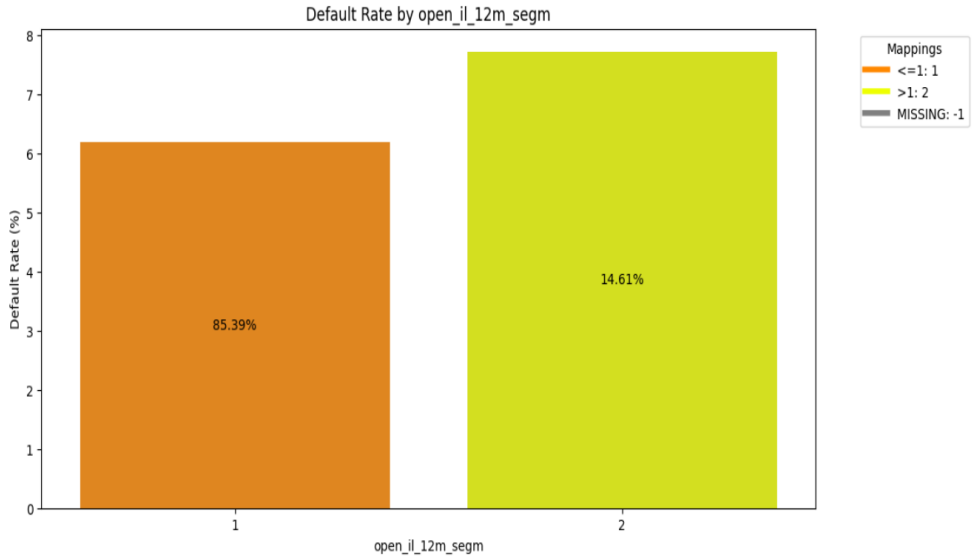


Figure 5.4.16: Installment Accounts Opened in Last 12 Months

This predictor segments borrowers by the number of installment accounts opened in the past year. More recent installment accounts are linked to higher bad rates, indicating that taking on multiple installment loans within a short period is a risk factor.

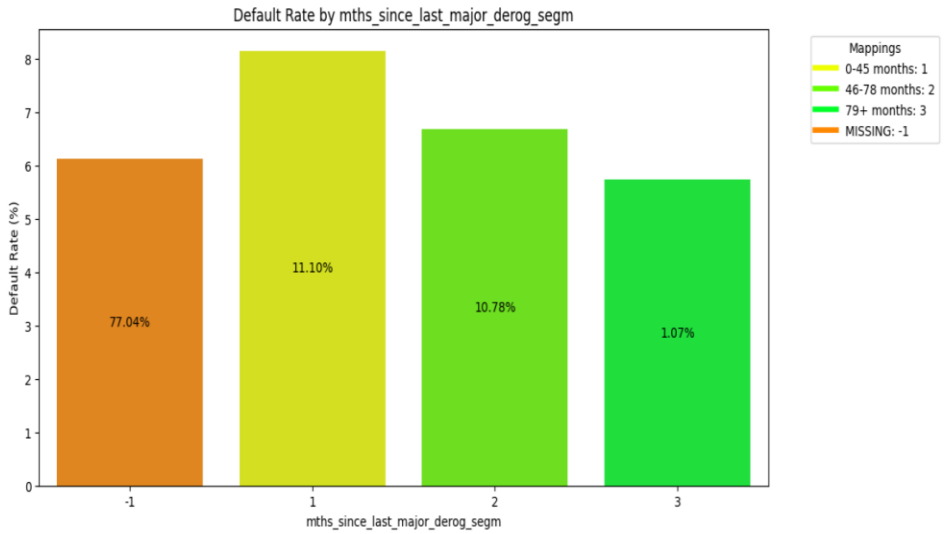


Figure 5.4.17: Months Since Last Major Derogation

The "Months Since Last Major Derogation" predictor segments borrowers based on the time elapsed since their last major derogatory mark. Longer periods since the last major derogation correspond to lower bad rates, highlighting the decreasing impact of past derogatory events over time.

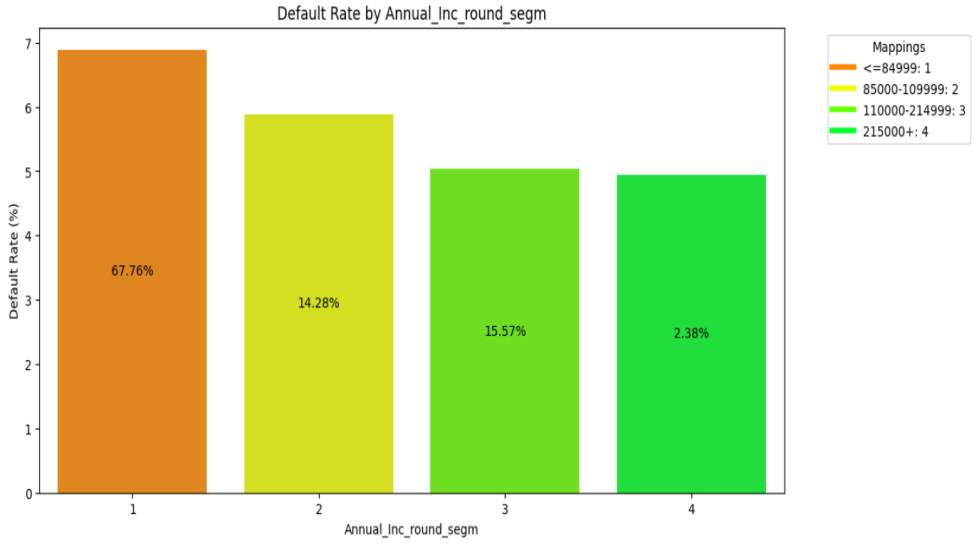


Figure 5.4.18: Annual Income Rounded

The "Annual Income Rounded" predictor categorizes borrowers based on their rounded annual income. Higher income brackets generally exhibit lower bad rates, suggesting that higher income is associated with lower default risk.

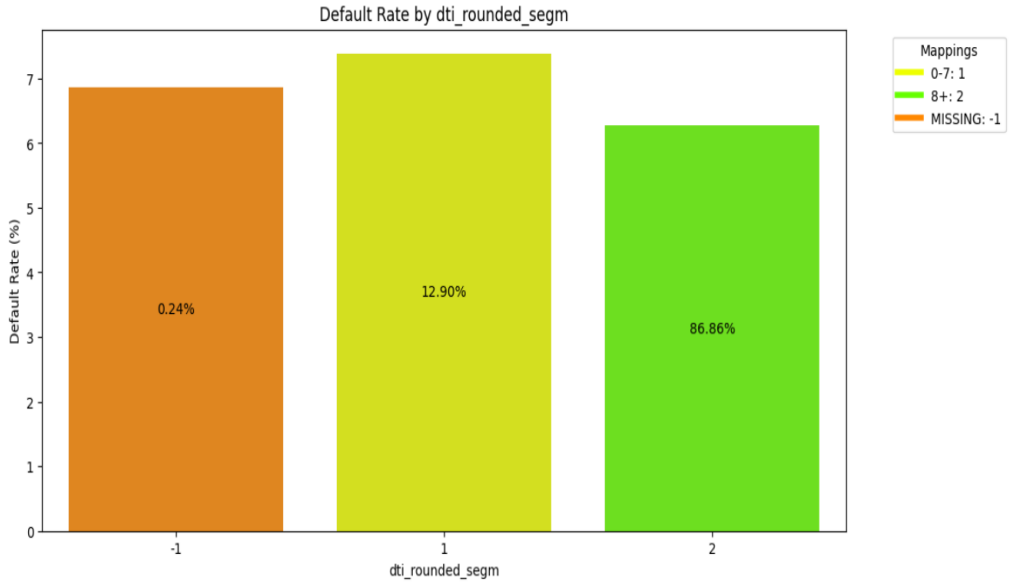


Figure 5.4.19: Debt-to-Income Rounded

The "Debt-to-Income Rounded" characteristic segments borrowers by their rounded debt-to-income ratios. Higher debt-to-income ratios are linked to higher bad rates, indicating that borrowers with higher debt relative to their income are at greater risk of default.

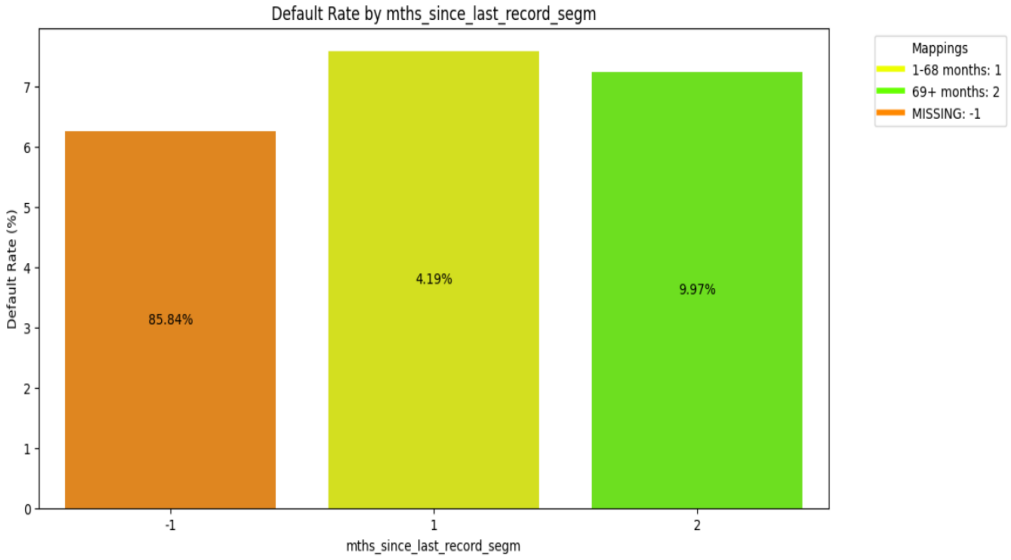


Figure 5.4.20: Months Since Last Record

The "Months Since Last Record" predictor segments borrowers by the time since their last record (e.g. bankruptcy etc). Longer periods since the last record are associated with lower bad rates, indicating that the impact of past records being decreased over time.

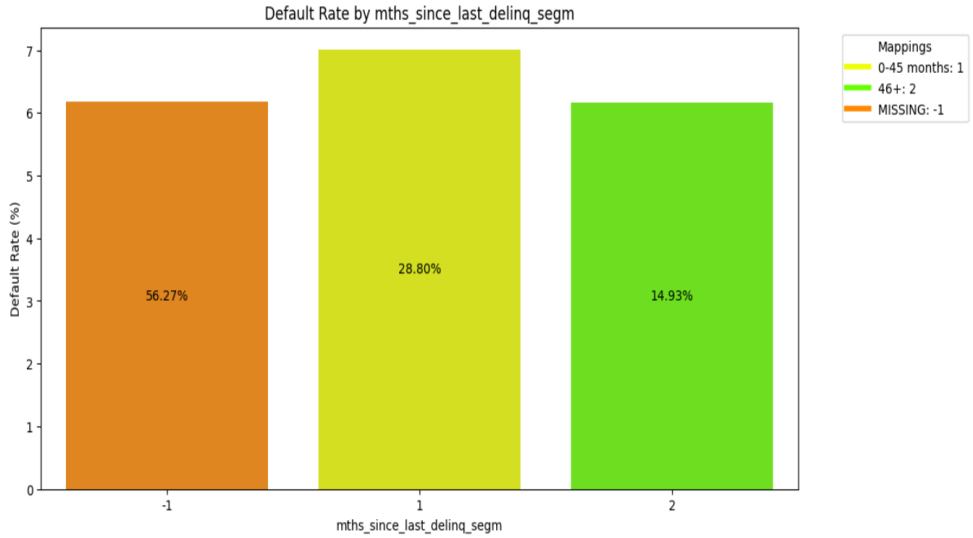


Figure 5.4.21: Months Since Last Delinquency

The "Months Since Last Delinquency" predictor categorizes borrowers based on the time since their last delinquency. Longer periods since the last delinquency correspond to lower bad rates, suggesting that the risk associated with past delinquencies decreases over time.

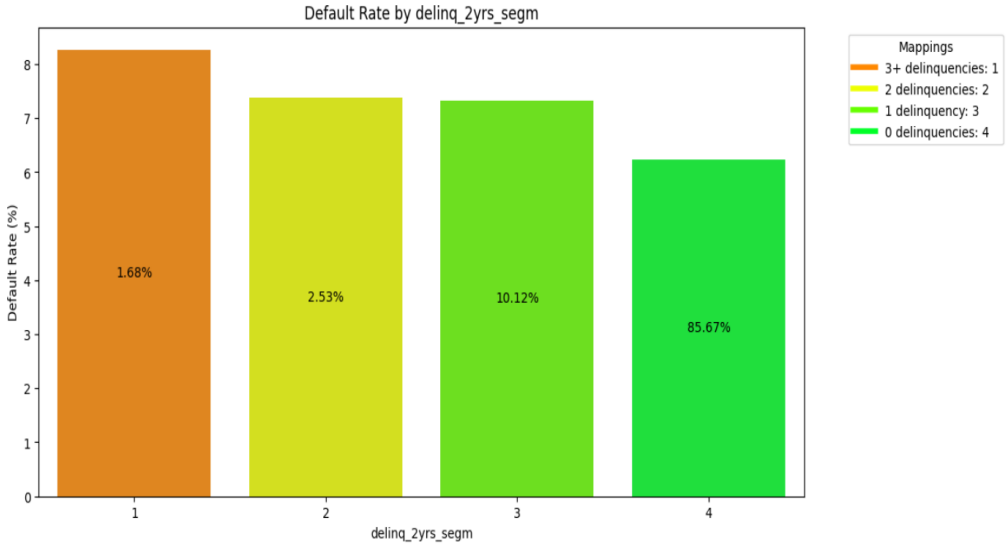


Figure 5.4.22: Delinquency Last 2 Years

This predictor segments borrowers based on the number of delinquencies in the past two years. More frequent delinquencies in this period are linked to higher bad rates, indicating that recent delinquency behavior is a significant risk factor for future delinquencies.

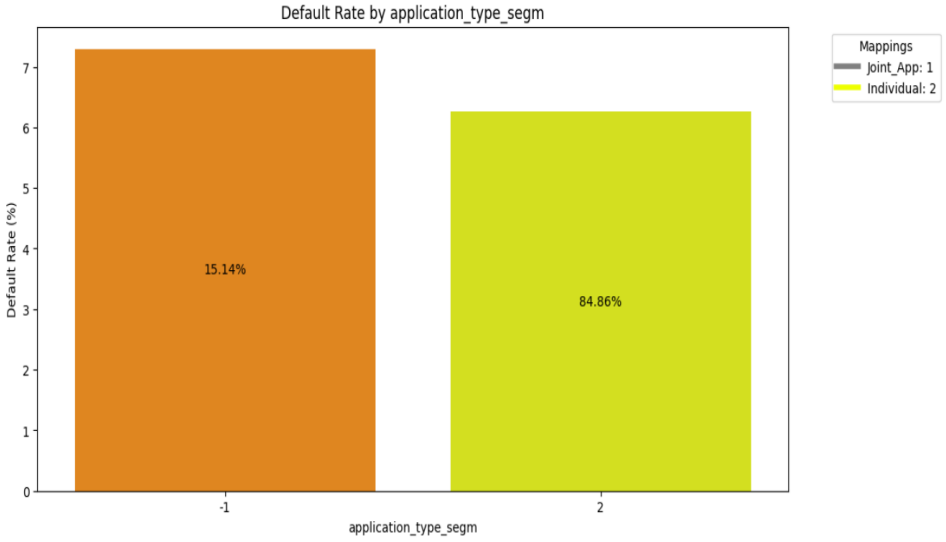


Figure 5.4.23: Application Type

The "Application Type" predictor categorizes borrowers based on the type of credit application (e.g. individual, joint). Joint applications tend to have lower bad rates compared to individual applications, suggesting that shared financial responsibility might reduce default risk.

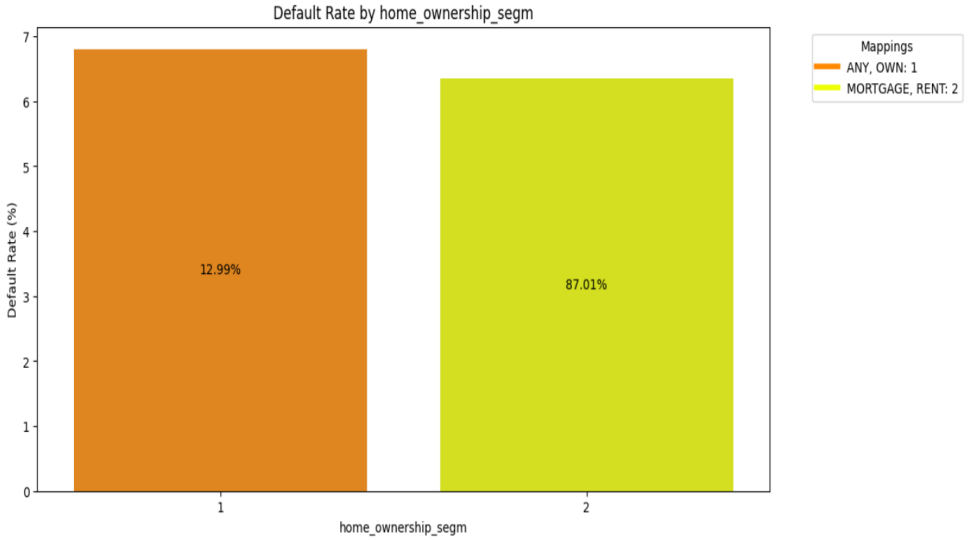


Figure 5.4.24: Home Ownership

The "Home Ownership" characteristic segments borrowers by their home ownership status (e.g. own, rent, mortgage). Homeowners generally exhibit lower bad rates compared to renters, reflecting the stability and financial security associated with owning a home.

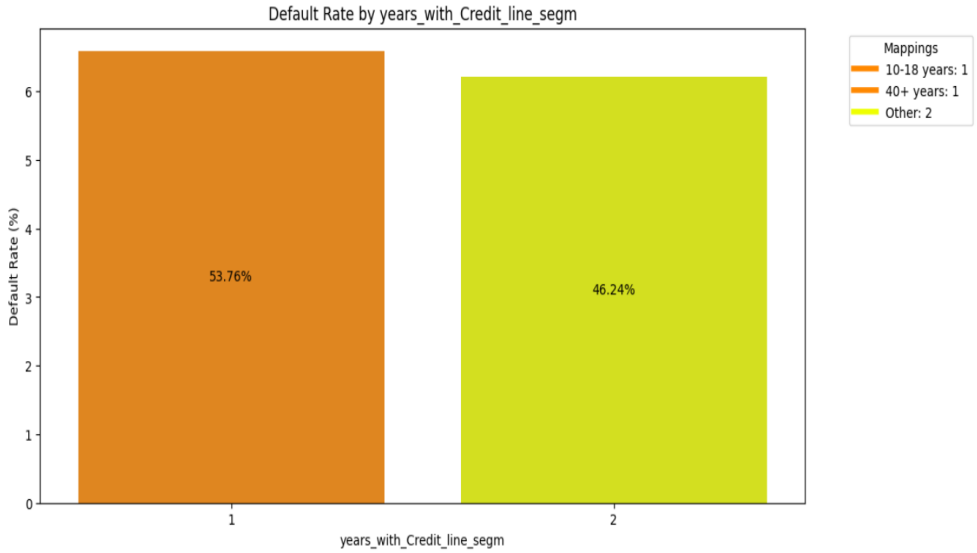


Figure 5.4.25: Years With Credit Line

The "Years With Credit Line" predictor segments borrowers based on the length of time they have had credit lines. Longer credit histories correlate with lower bad rates, indicating that more extensive credit experience is associated with lower risk.

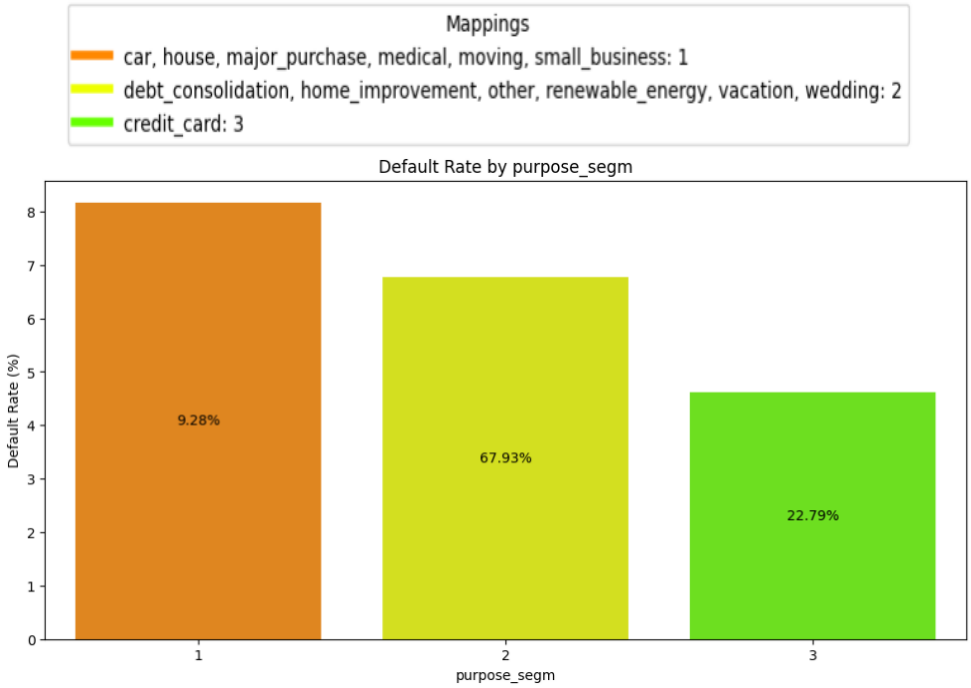


Figure 5.4.26: Purpose of Loan

The "Purpose of Loan" predictor segments borrowers based on the stated purpose of their loan (e.g. debt consolidation, home improvement, major purchase, etc). As indicated by the corresponding groups, credit card applications are of the lowest risk, whereas “Debt Consolidation”, “Home Improvement”, “Renewable Energy”, “Vacation” and “Wedding” are of mid-level risk. Finally, loans for “Car”, “House”, “Medical”, “Moving” and “Small business” purposes, are of the highest risk.

It is important to be noted that many of the above relationships can vary for different financial institutions, different types of loan products, that can also vary across different countries and continents.

Finally, the IV criterion is a univariate measure used to assess the predictive power of each predictor in relation to the target variable -as it has already been presented in the current thesis, which in this context is the PD. A higher IV indicates a stronger predictive capability. Below is an analysis of the IVs for the potential predictors, in ascending order:

#	Potential Predictor	IV
1	total_pymnt_round	55.60
2	mths_since_recent_inq_segm	7.76
3	verification_status_segm	7.54
4	inq_last_6mths_segm	7.34
5	fico_range_low_segm	6.50
6	inq_last_12m_segm	4.93
7	term_segm	4.77
8	purpose_segm	3.34
9	instlmnt_round_segm	3.04
10	mort_acc_segm	2.97
11	emp_length_segm	2.74
12	open_acc_6m_segm	2.63
13	acc_open_past_24mths_segm	2.51
14	open_rv_24m_segm	2.24
15	open_rv_12m_segm	2.04
16	revol_bal_segm	1.64
17	Annual_Inc_round_segm	1.61
18	mths_since_last_major_derog_segm	1.04
19	open_il_12m_segm	0.76
20	delinq_2yrs_segm	0.50
21	mths_since_last_delinq_segm	0.40
22	mths_since_last_record_segm	0.39
23	dti_rounded_segm	0.37
24	application_type_segm	0.37
25	years_with_Credit_line_segm	0.10
26	home_ownership_segm	0.06

Table 5.4.1: IVs on the Training Set (Segmented Variables)

5.5 Statistical and Machine Learning Model Implementations

In this section, we focus into the implementation of four statistical and ML models -aimed at predicting PD. These models are chosen for their diverse approaches to handling complex datasets and their proven efficacy in predictive analytics in general. Each model will be implemented upon using the segmented predictors, in the

versions that previously presented, aiming to extract meaningful patterns and relationships that can forecast PD with high accuracy. The models implemented are LR, RF, GB and NN. For the LR, extra handling was done, by creating binary variables (or dummy variables) from the already segmented ones, with each variable with k categories producing k-1 new binary variables. This handling was determined by the inefficiency of linear models capturing non-linearities. On the contrary, RF, GB and NN are models that can capture complex and non-linear relationships -nevertheless being much more prune to overfitting issues.

For each of the implementations that will follow, the estimated PD will be converted into Credit Score, by applying the following formula:

$$\text{Credit Score} = 400 + 28.85 \times \log\left[\frac{\text{ProbGood}}{\text{ProbBad}}\right]$$

where the second addend has been rounded (in order to obtain only integers as the Credit Score) and 400 is used as a constant to adjust the heights of the score. The coefficient of 28.85 is used in order to create a score-odds correspondence, in which for each +/-20 points of score, the Good:Bad-odds will be doubled or cut in half respectively.

5.5.1 Logistic Regression Implementation

As mentioned, for the LR, dummy variables were created. By proceeding to feature selection (between the binary predictors). The forward feature selection process, which aimed to maximize the ROC-AUC metric, identified several key dummy variables that are critical in predicting PD. These dummy variables represent various segments of categorical predictors, with each segment providing unique insights into the risk profile of borrowers. The selected dummy variables include various segments of predictors such as "total payment", "verification status", "months since recent inquiry", "inquiries in the last six months", "FICO range", "loan purpose", "installment amount", "accounts opened in the past 24 months", "loan term", "open accounts in the last six months", "mortgage accounts", "employment length", "annual income", "months since last delinquency", and "application type".

More specifically, each dummy variable corresponds to a specific level of a categorical variable. For instance, "total_pymnt_round_segm_1" indicates the first level of the total payment rounded predictor. If this dummy variable is equal to 1, it means the borrower falls into the first segment of total payment rounded, which is <= 4999, otherwise, it is 0. Similarly, "total_pymnt_round_segm_2" represents the second level of this predictor, indicating total payments in the range of 5000-9999. The same logic applies to other dummy variables, where the suffix "_segm_X" denotes the Xth segment of the categorical variable.

All the dummy variables that were selected for LR throughout the stepwise procedure, are presented in the following table:

Dummy Variable	Segment	Description
total_pymnt_round_segm_1	Segment 1	Total payment rounded ≤ 4,999
total_pymnt_round_segm_2	Segment 2	Total payment rounded between 5,000 - 9,999
total_pymnt_round_segm_3	Segment 3	Total payment rounded between 10,000 - 14,999
verification_status_segm_1	Segment 1	Verification status: Verified
verification_status_segm_2	Segment 2	Verification status: Source Verified
mths_since_recent_inq_segm_1	Segment 1	Months since recent inquiry: 0-2 months
inq_last_6mths_segm_4	Segment 4	0 inquiries in the last 6 months
inq_last_6mths_segm_3	Segment 3	1 inquiry in the last 6 months
fico_range_low_segm_1	Segment 1	FICO score ≤ 680
fico_range_low_segm_2	Segment 2	FICO score between 685-700
fico_range_low_segm_4	Segment 4	FICO score between 725-790
fico_range_low_segm_5	Segment 5	FICO score ≥ 791
purpose_segm_2	Segment 2	Loan purpose: debt consolidation, home improvement, other, renewable energy, vacation, wedding
purpose_segm_1	Segment 1	Loan purpose: car, house, major purchase, medical, moving, small business
instlmnt_round_segm_1	Segment 1	Installment amount rounded ≤ 479
instlmnt_round_segm_2	Segment 2	Installment amount rounded between 480-699
instlmnt_round_segm_3	Segment 3	Installment amount rounded between 700-879
acc_open_past_24mths_segm_3	Segment 3	Accounts opened in the past 24 months: 7+
term_segm_2	Segment 2	Loan term: 36 months
open_acc_6m_segm_1	Segment 1	Open accounts in the last 6 months ≤ 1
mort_acc_segm_1	Segment 1	Mortgage accounts: 0
emp_length_segm_2	Segment 2	Employment length: 0-8 years
emp_length_segm_1	Segment 1	Employment length: 9+ years
Annual_Inc_round_segm_3	Segment 3	Annual income rounded between 110,000-214,999
mths_since_last_delinq_segm_2	Segment 2	Months since last delinquency: 46+
application_type_segm_2	Segment 2	Application type: Individual

Table 5.5.1.1: Logistic Regression – Dummy Variables

Below are presented the KS, AUC and Gini metrics on the Training and Test sets respectively:

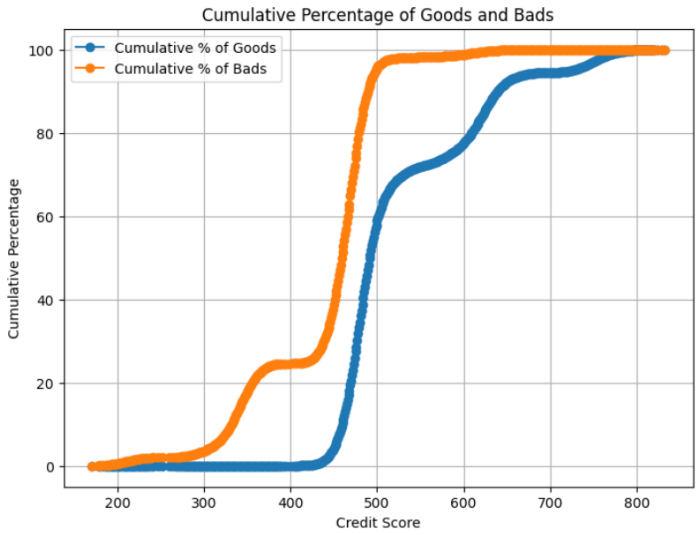


Figure 5.5.1.1: KS Logistic Regression – Training Set

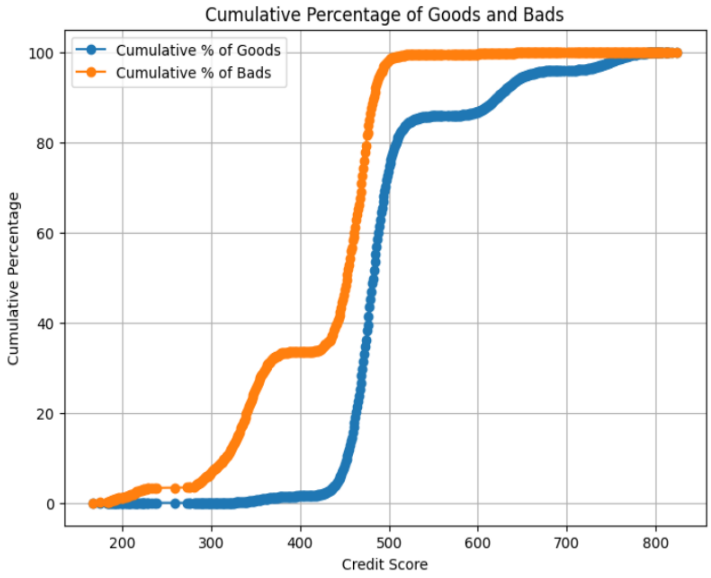


Figure 5.5.1.2: KS Logistic Regression – Test Set

The KS statistic is slightly higher in the training set when compared to the test set (46.86% vs 43.40%), suggesting that the model performs well in separating "Goods" and "Bads" in the training data. However, it still indicates good separation on unseen data. The model generalizes well to the test set with a moderate reduction in performance, which is normal and indicates stability. Furthermore, from the cumulative distributions of “goods” and “bads”, it can be observed that there is a slight difficulty in discriminating the population of the middle credit scores. That means that the model separates quite satisfactorily the goods and bads in general, but the distribution of the hard-to-classify cases (middle scores) indicate that there is room for improvement.

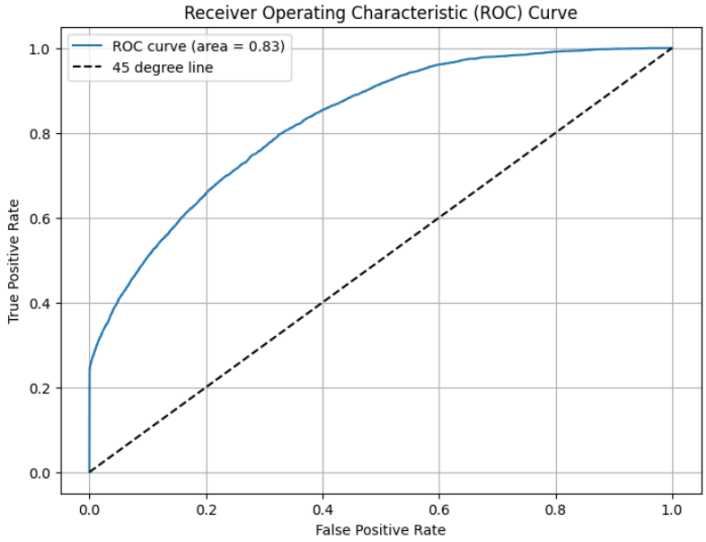


Figure 5.5.1.3: ROC Curve Logistic Regression – Training Set

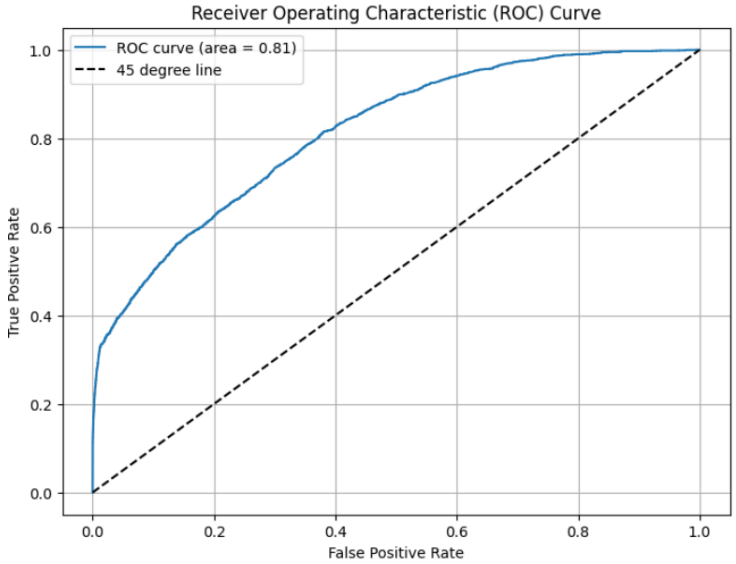


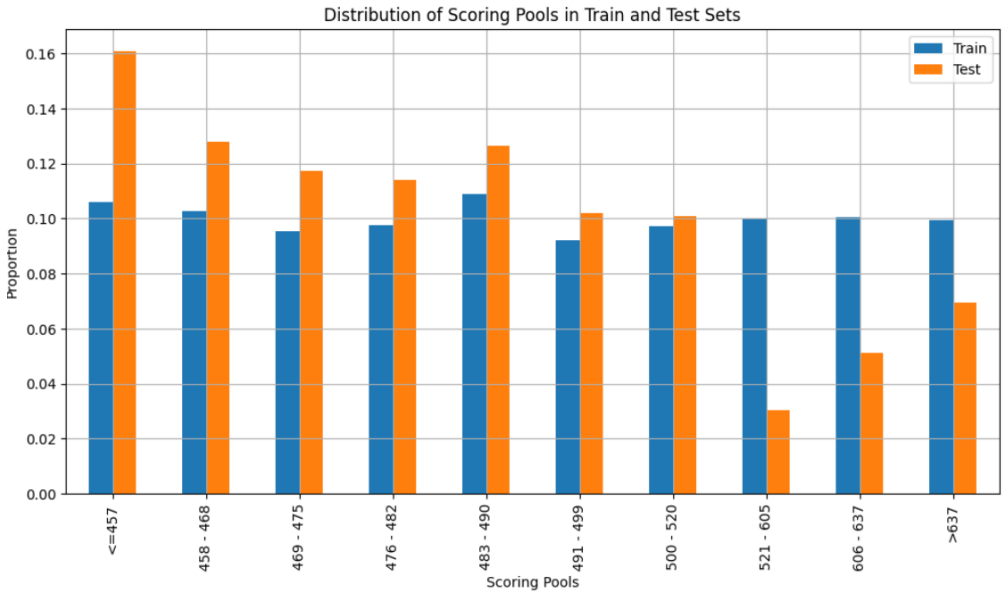
Figure 5.5.1.4: ROC Curve Logistic Regression – Test Set

The AUC on the training set stands at 0.83, whereas on the test set is equal to 0.81. This suggests that the model is doing a good job in distinguishing between "Good" and "Bad" borrowers in the test data as well, keeping the overfit in minimal levels.

```
The Kolmogorov-Smirnov statistic on the TRAINING data is: 46.86
AUC metric on the TRAINING data is: 0.83
Gini metric on the TRAINING data is: 0.66

The Kolmogorov-Smirnov statistic on the TEST data is: 43.40
AUC metric on the TEST data is: 0.81
Gini metric on the TEST data is: 0.63
```

Figure 5.5.1.5: Output Metrics - Logistic Regression – Training Vs Test Set




```
The PSI statistic between TRAINING and TEST sets is: 0.166
Moderate shift in the population (PSI = 0.166)
```

Figure 5.5.1.6: Credit Score PSI - Logistic Regression – Training Vs Test Set

The PSI statistic of 0.166 indicates a moderate shift between the training and test set distributions. Although there is some difference between the scoring distributions, it is not extreme, implying that the model performs consistently across both datasets. However, it might still need further monitoring or refinement to improve stability. Finally, this shift could be attributed to the fact that some of the characteristics that entered the model, over-penalize the applicants of the test set -thus increasing the concentration to the lower credit scores. Alternatively, the population of the test set might be indeed riskier, and thus is correctly ranked by the model to the lower scoring pools.

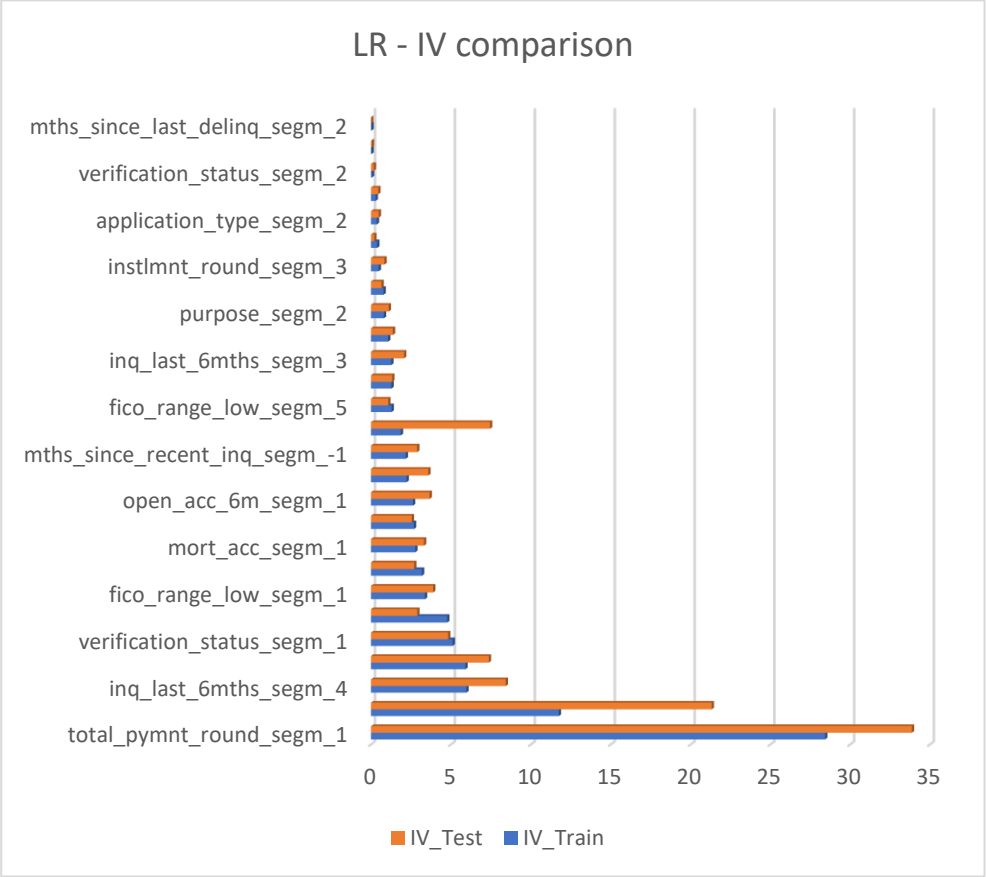


Figure 5.5.1.7: IV Comparison - Logistic Regression – Training Vs Test Set

"Total Payment Rounded Segment 1" has the highest IV in both training and test sets, making it (by far) the most significant variable in this model. There is a noticeable increase in IV for some variables in the test set (e.g. fico_range_low_segm_5, inq_last_6mths_segm_4, fico_range_low_segm_5), indicating increased predictive power in the test set, without noticing any notable drop in the IV of all the other variables that were used in the model. Overall, the important variables retain their influence across both sets, indicating that the model captures the most meaningful predictors.

5.5.2 Random Forest Implementation

RF can handle categorical variables exceptionally well and is robust to correlated features. Hence, the variables will be inserted to the model in their segmented formats, and not through dummy variables as in LR.

The forward feature selection process, applied within the RF algorithm, identified several segmented variables for PD estimation. This method iteratively added predictors to the model based on their contribution to maximizing the ROC-AUC metric, ensuring that each selected variable significantly enhanced the model's discriminatory power. The variables identified as significant are the following:

Segmented Variable	Segments	Description
total_pymnt_round_segm	1: ≤ 4,999	

Segmented Variable	Segments	Description
	2: 5,000-9,999	Total payment rounded into different ranges
	3: 10,000-14,999	
	4: ≥ 15,000	
verification_status_segm	1: Verified	Borrower's verification status
	2: Source Verified	
	3: Not Verified	
mths_since_recent_inq_segm	1: 0-2 months	Months since the borrower's most recent inquiry
	2: 3-10 months	
	3: 11+ months	
	-1: MISSING	
inq_last_6mths_segm	1: 3+ inquiries	Number of inquiries in the last 6 months
	2: 2 inquiries	
	3: 1 inquiry	
	4: 0 inquiries	
fico_range_low_segm	1: ≤ 680	Borrower's FICO credit score range
	2: 685-700	
	3: 705-720	
	4: 725-790	
	5: ≥ 791	
instlmnt_round_segm	1: ≤ 479	Installment amount rounded into ranges
	2: 480-699	
	3: 700-879	
	4: ≥ 880	
term_segm	1: 60 months	Loan term length
	2: 36 months	
open_acc_6m_segm	1: ≤ 1	Number of open accounts in the last 6 months
	2: > 1	
	-1: MISSING	
mort_acc_segm	1:00	Number of mortgage accounts
	2:01	
	3:02	
	4: 3+	
revol_bal_segm	1: ≤ 9,999	Revolving balance amount ranges
	2: 10,000-19,999	
	3: 20,000-39,999	
	4: ≥ 40,000	
emp_length_segm	1: 10+ years	Employment length segmented into ranges
	2: 1-8 years	
	-1: MISSING	
dti_rounded_segm	1: 0-7	Debt-to-income ratio rounded into segments
	2: ≥ 8	
	-1: MISSING	
delinq_2yrs_segm	1: 3+ delinquencies	Number of delinquencies in the past 2 years
	2: 2 delinquencies	
	3: 1 delinquency	
	4: 0 delinquencies	

Segmented Variable	Segments	Description
application_type_segm	1: Joint Application	Type of loan application
	2: Individual Application	
home_ownership_segm	1: ANY, OWN	Home ownership status
	2: MORTGAGE, RENT	

Table 5.5.2.1: Random Forest Variables

Below are presented the KS, AUC and Gini metrics on the Training and Test sets:

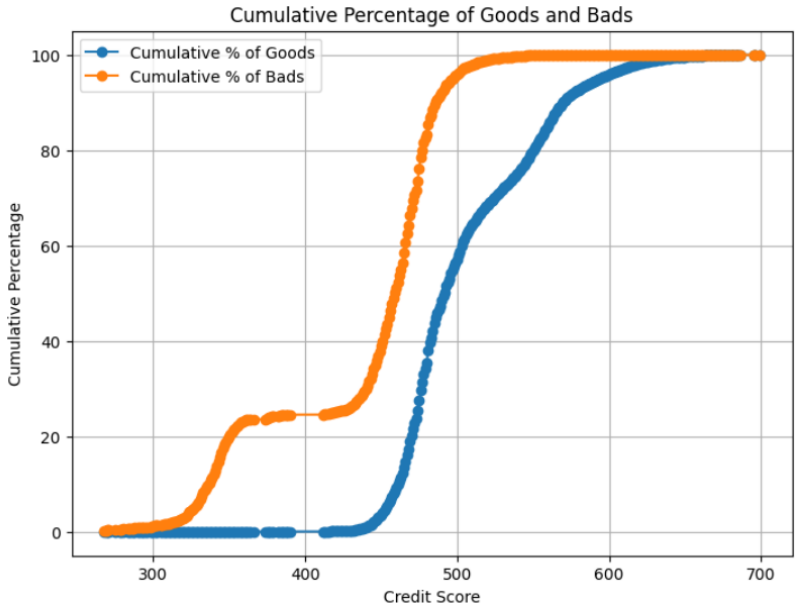


Figure 5.5.2.1: KS -Random Forest – Training Set

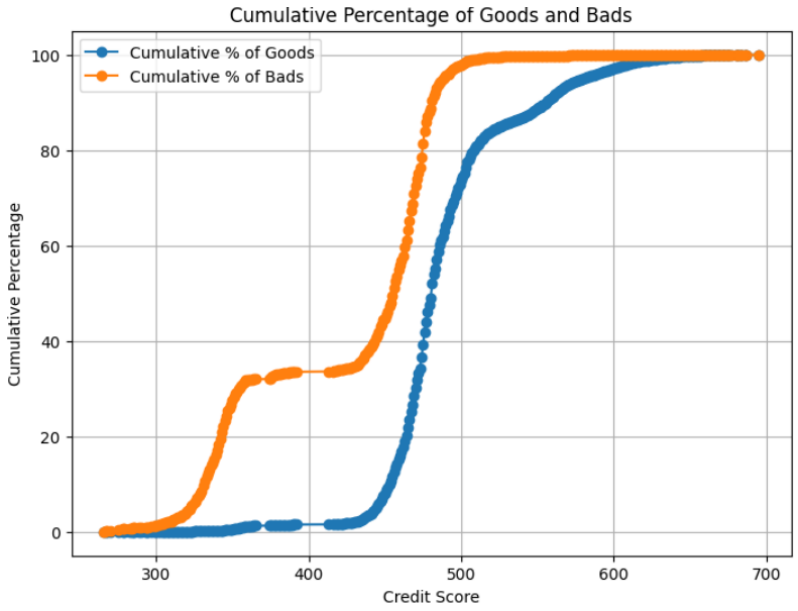


Figure 5.5.2.2: KS – Random Forest– Test Set

The KS statistic is quite higher in the training set when compared to the test set

(48.74% vs 42.39%), suggesting that the model has overfitted to the training set. However, it still indicates good separation on unseen data though, with the KS remaining in high levels in the test set as well. Again, same as in the LR implementation, from the cumulative distributions of "goods" and "bads", it can be observed that there is a slight difficulty in discriminating the population of the middle credit scores. That means that the model separates quite satisfactorily the goods and bads in general, but the distribution of the hard-to-classify cases (middle scores) indicate that there is room for improvement.

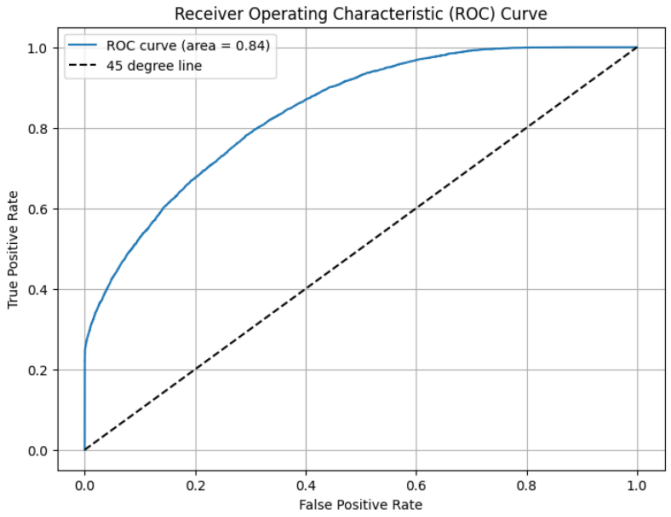


Figure 5.5.2.3: ROC Curve – Training Set

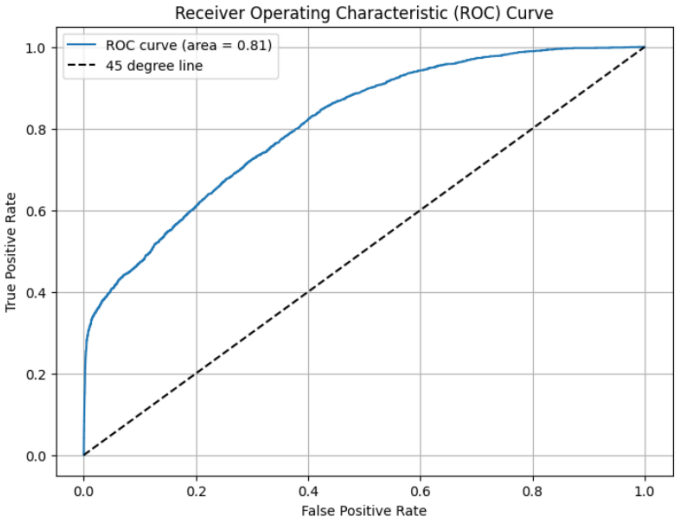


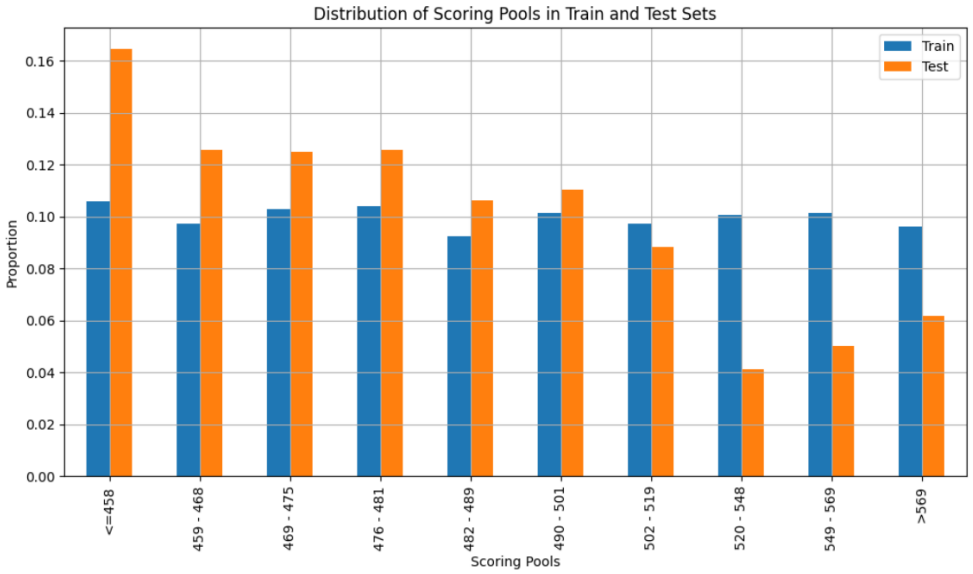
Figure 5.5.2.4: ROC Curve – Test Set

The AUC metric (0.84 training vs 0.81 test set) also indicates a slight overfit, with the generalization remaining in high levels nevertheless.

```
The Kolmogorov-Smirnov statistic on the TRAIN data is: 48.74
AUC metric on the TRAIN data is: 0.84
Gini metric on the TRAIN data is: 0.68

The Kolmogorov-Smirnov statistic on the TEST data is: 42.39
AUC metric on the TEST data is: 0.81
Gini metric on the TEST data is: 0.62
```

Figure 5.5.2.5: Output Metrics - Random Forest – Training Vs Test Set




```
The PSI statistic between TRAINING and TEST sets is: 0.150
Moderate shift in the population (PSI = 0.150)
```

Figure 5.5.2.6: Credit Score PSI – Random Forest – Training Vs Test Set

Again, the PSI statistic of 0.150 indicates a moderate shift between the training and test set distributions. The above figure depicts the shift in the score that was also observed in the RF model. As mentioned, this shift could be attributed to the reasons that were also discussed in the LR implementation (either some characteristics over-penalize the population of the test set, or the population of the test set is indeed of higher risk).

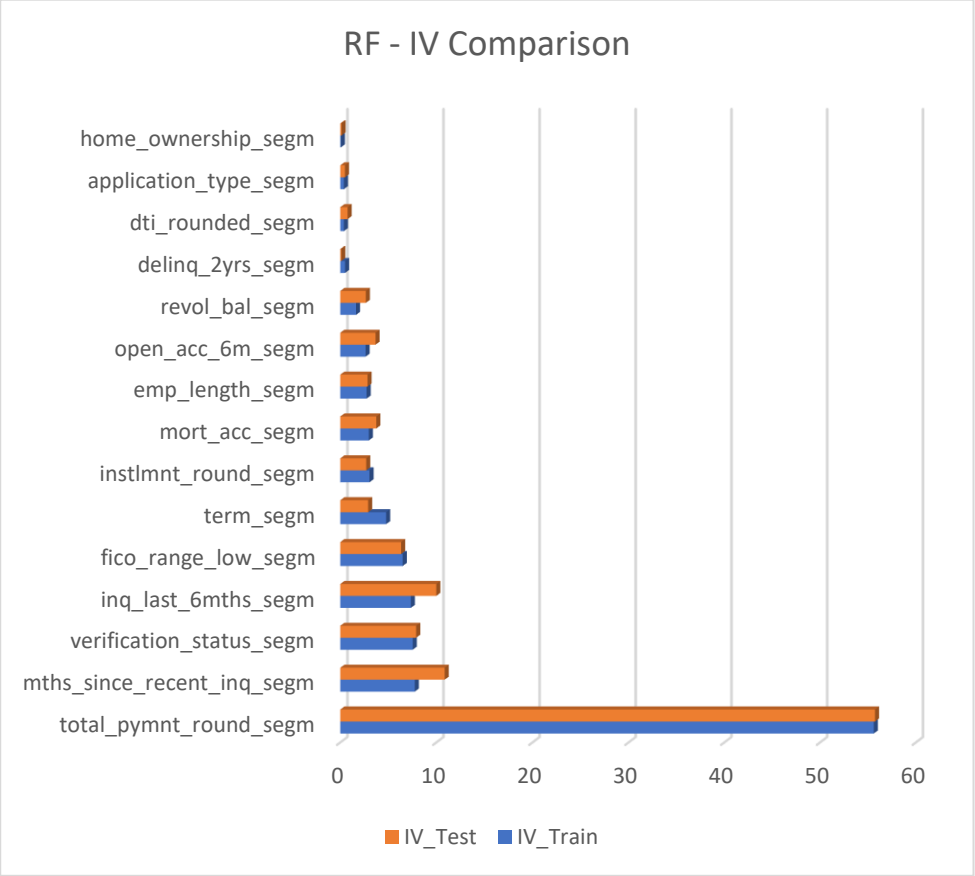


Figure 5.5.2.7: IV Comparison – Random Forest – Training Vs Test Set

No significant shifts are observed in the predictive power of the characteristics used by the model, indicating that these characteristics continue to capture the patterns that were observed on the training set on unseen data. However, it should be mentioned that a slight increase in the predictive power is observed in the characteristics of "inq_last_6_mths_segm" and "mths_since_recent_inq_segm".

5.5.3 Gradient Boosting Implementation

GB can also handle categorical variables exceptionally well -as RF does, and is also robust to correlated features. Hence, the variables will be inserted to the model in their segmented formats, and not through dummy variables as in LR.

The forward feature selection process was also used in this implementation, in order to assess the contribution of predictors into maximizing the ROC-AUC metric, ensuring that each selected variable significantly enhanced the model's discriminatory power. The variables identified as significant are the following:

Segmented Variable	Segments	Description
total_pymnt_round_segm	1: ≤ 4,999	

Segmented Variable	Segments	Description
	2: 5,000-9,999	Total payment rounded into different ranges
	3: 10,000-14,999	
	4: ≥ 15,000	
verification_status_segm	1: Verified	Borrower's verification status
	2: Source Verified	
	3: Not Verified	
mths_since_recent_inq_segm	1: 0-2 months	Months since the borrower's most recent inquiry
	2: 3-10 months	
	3: 11+ months	
	-1: MISSING	
inq_last_6mths_segm	1: 3+ inquiries	Number of inquiries in the last 6 months
	2: 2 inquiries	
	3: 1 inquiry	
	4: 0 inquiries	
fico_range_low_segm	1: ≤ 680	Borrower's FICO credit score range
	2: 685-700	
	3: 705-720	
	4: 725-790	
	5: ≥ 791	
purpose_segm	1: car, house, major purchase, medical, moving, small business	Loan purpose categories segmented
	2: debt consolidation, home improvement, other, renewable energy, vacation, wedding	
	3: credit card	
instlmnt_round_segm	1: ≤ 479	Installment amount rounded into ranges
	2: 480-699	
	3: 700-879	
	4: ≥ 880	
term_segm	1: 60 months	Loan term length
	2: 36 months	
open_acc_6m_segm	1: ≤ 1	Number of open accounts in the last 6 months
	2: > 1	
	-1: MISSING	
mort_acc_segm	1:00	Number of mortgage accounts
	2:01	
	3:02	
	4: 3+	
open_rv_12m_segm	1: ≤ 1	Number of open revolving accounts in the last 12 months
	2: > 1	
	-1: MISSING	
revol_bal_segm	1: ≤ 9,999	Revolving balance amount ranges
	2: 10,000-19,999	
	3: 20,000-39,999	

Segmented Variable	Segments	Description
	4: ≥ 40,000	
emp_length_segm	1: 10+ years	Employment length segmented into ranges
	2: 1-8 years	
	-1: MISSING	
application_type_segm	1: Joint Application	Type of loan application
	2: Individual Application	
home_ownership_segm	1: ANY, OWN	Home ownership status
	2: MORTGAGE, RENT	

Table 5.5.3.1: Gradient Boosting Variables

Below are presented the KS, AUC and Gini metrics on the training and test sets respectively:

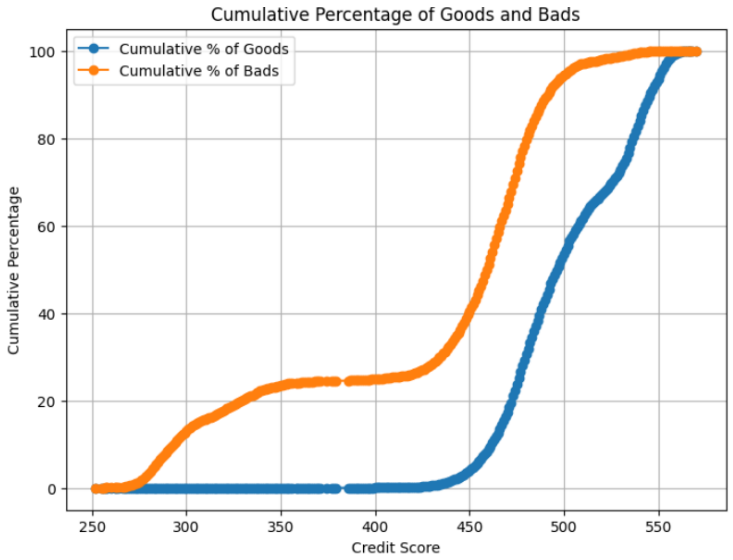


Figure 5.5.3.1: KS – Gradient Boosting – Training Set

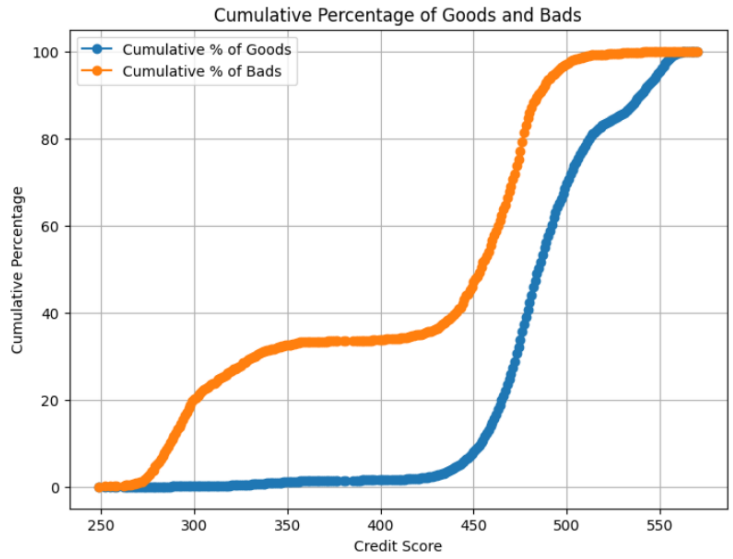


Figure 5.5.3.2: KS – Gradient Boosting – Test Set

The KS statistic is quite higher in the training set when compared to the test set (49.23% vs 44.15%), suggesting that the model has overfitted to the training set. However, it still indicates good separation on unseen data though, with the KS remaining in high levels in the test set as well. Furthermore, from the above distributions, it seems that GB is discriminating satisfactorily the good-bad populations in the middle scores (i.e. hard-to-classify cases), with the difference between the two cumulative distributions being distinguishable.

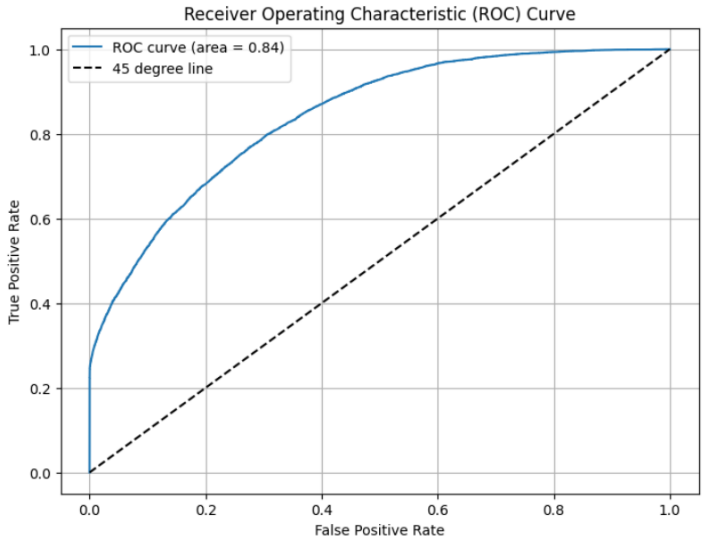


Figure 5.5.3.3: ROC Curve – Gradient Boosting – Training Set

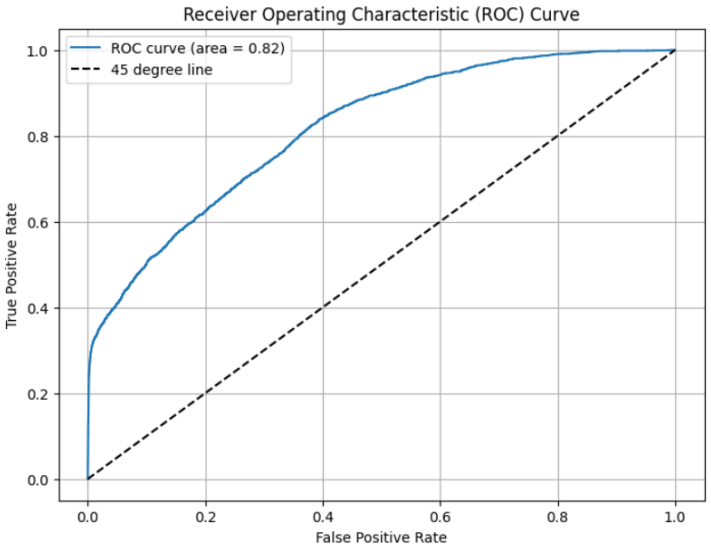


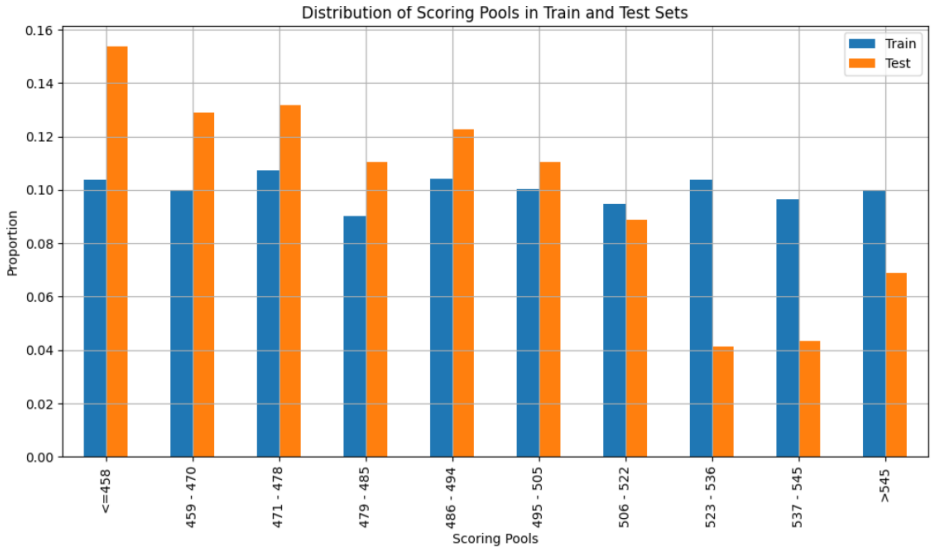
Figure 5.5.3.4: ROC Curve – Gradient Boosting – Test Set

The AUC metric (0.84 training vs 0.82 test set) indicates a minimal overfit, with the generalization on unseen data remaining in high levels.

```
The Kolmogorov-Smirnov statistic on the TRAIN data is: 49.23
AUC metric on the TRAIN data is: 0.84
Gini metric on the TRAIN data is: 0.68

The Kolmogorov-Smirnov statistic on the TEST data is: 44.15
AUC metric on the TEST data is: 0.82
Gini metric on the TEST data is: 0.63
```

Figure 5.5.3.5: Output Metrics – Gradient Boosting – Training Vs Test Set




```
The PSI statistic between TRAINING and TEST sets is: 0.152
Moderate shift in the population (PSI = 0.152)
```

Figure 5.5.3.6: Credit Score PSI – Gradient Boosting – Training Vs Test Set

Again, same as in the previous model implementations, the PSI statistic of 0.152 indicates a moderate shift between the training and test set distributions. The above figure depicts the shift in the score that was also observed in the GB model. As mentioned, this shift could be attributed to the reasons that were also described in the previous two implementations (i.e.LR and RF).

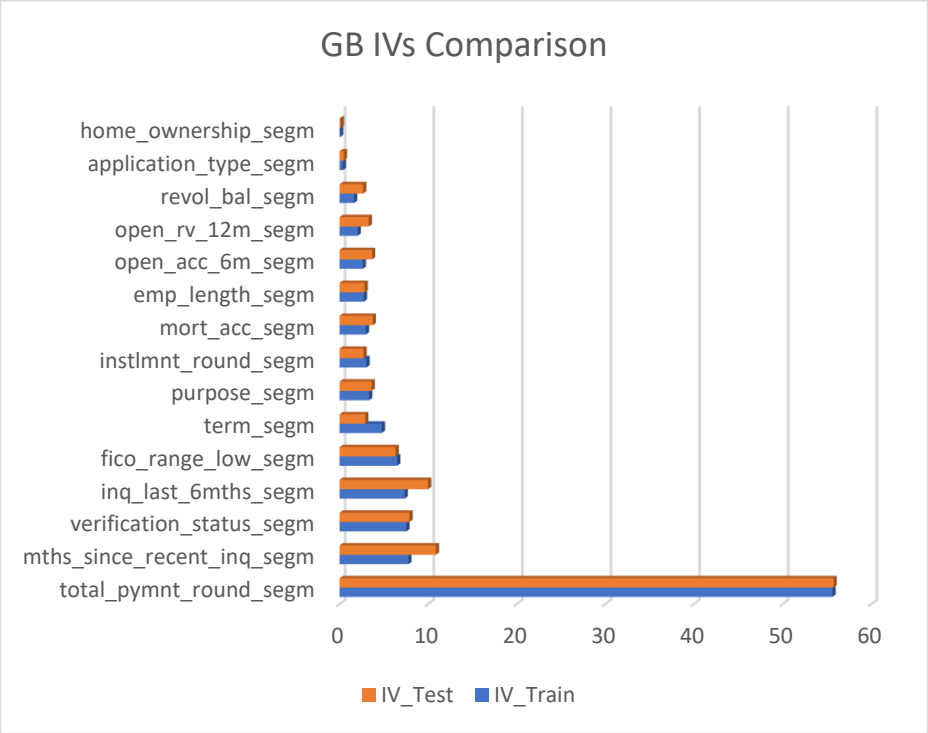


Figure 5.5.3.7: IV Comparison – Gradient Boosting – Training Vs Test Set

No significant shifts are observed in the predictive power of the characteristics used by the model, indicating that these characteristics continue to capture the patterns that were observed on the training set on unseen data. However, it should be mentioned that a slight increase in the predictive power is observed in the characteristics of "inq_last_6_mths_segm" and "mths_since_recent_inq_segm".

5.5.4 Neural Networks Implementation

For the implementation of the NN model, the training set has been further segmented into training and validation sets, with the validation set being 20% of the initial training set. Validation set will be used during the training of the model, in order to reduce the potential overfit of the algorithm to the training data. All the 26 potential predictors will be used, as NN are known for their ability to automatically detect and leverage important features from the data. Of course, this can lead to overfitting, especially if the model becomes too complex relative to the amount of training data, but the segmentation of the characteristics also aim to reduce the complexity of the predictors used in the model.

The architecture of the implemented NN is designed to balance complexity and regularization, aiming to provide robust predictions while minimizing overfitting. Here's a detailed overview of the architecture:

Input Layer:

- Shape: The input shape is set to (26,) to match the 26 predictors.

- This layer feeds the input data into the network, ensuring that each feature is considered.

Hidden layers:

- First hidden layer:
  - Nodes: 16
  - Activation function: ReLU
  - Dropout: 20% dropout rate to prevent overfitting by randomly setting 20% of the nodes to zero during training.

- Second hidden layer:
  - Nodes: 32
  - Activation function: ReLU
  - Dropout: 20% dropout rate.

- Third hidden layer:
  - Nodes: 64
  - Activation function: ReLU
  - Dropout: 20% dropout rate.

Output layer:

- Nodes: 1
- Activation function: Sigmoid, producing a single output between 0 and 1 -suitable for binary classification problems.

Optimizer:

- Adam optimizer: The Adam optimizer is chosen for its efficiency and adaptive learning rate capabilities, which help in faster convergence and better handling of noisy gradients.
- Learning Rate: Specified within the Adam optimizer settings, tuned conservatively to avoid overfitting.

Loss function:

- Binary crossentropy: Used for binary classification tasks, this loss function measures the performance of the model by comparing the predicted probabilities to the actual binary labels.

Metrics:

- AUC: Monitored during training and evaluation to optimize the model's performance.

The model is trained with a focus on to prevent overfitting. Dropout layers are used to randomly deactivate neurons during training, encouraging the model to generalize better to unseen data. Additionally, conservative hyperparameter tuning aims to make the model not too simple nor too complex thus aiming to minimize overfitting.

The training process involves running the model for 50 epochs with a batch size of 16, updating weights after processing small groups of data, which improves computational efficiency and stability. Performance is monitored on a validation set to prevent the model from overfitting to the training data. These methods collectively help create a robust model that generalizes well to new data while avoiding overfitting.

The characteristics that entered the model are presented below:

Segmented Variable	Segments	Description
home_ownership_segm	1: ANY, OWN	Home ownership status
	2: MORTGAGE, RENT	
years_with_Credit_line_segm	1: 10-18 years, 40+ years	Number of years with a credit line
	2: Other	
application_type_segm	1: Joint Application	Type of loan application
	2: Individual Application	
dti_rounded_segm	1: 0-7	Debt-to-income ratio rounded into segments
	2: ≥ 8	
	-1: MISSING	
mths_since_last_record_segm	1: 1-68 months	Months since the borrower's last public record
	2: 69+ months	
	-1: MISSING	
mths_since_last_delinq_segm	1: 0-45 months	Months since the borrower's last delinquency
	2: 46+ months	
	-1: MISSING	
delinq_2yrs_segm	1: 3+ delinquencies	Number of delinquencies in the past 2 years
	2: 2 delinquencies	
	3: 1 delinquency	
	4: 0 delinquencies	
open_il_12m_segm	1: ≤ 1	Number of installment loans opened in the last 12 months
	2: > 1	
	-1: MISSING	
mths_since_last_major_derog_segm	1: 0-45 months	Months since the last major derogatory mark on credit
	2: 46-78 months	
	3: 79+ months	

Segmented Variable	Segments	Description
	-1: MISSING	
Annual_Inc_round_segm	1: ≤ 84,999	Annual income rounded into ranges
	2: 85,000-109,999	
	3: 110,000-214,999	
	4: ≥ 215,000	
revol_bal_segm	1: ≤ 9,999	Revolving balance amount segmented into ranges
	2: 10,000-19,999	
	3: 20,000-39,999	
	4: ≥ 40,000	
open_rv_12m_segm	1: ≤ 1	Number of revolving accounts opened in the last 12 months
	2: > 1	
	-1: MISSING	
open_rv_24m_segm	1: ≤ 1	Number of revolving accounts opened in the last 24 months
	2: 2-6	
	3: 7+	
	-1: MISSING	
acc_open_past_24mths_segm	1: ≤ 1	Number of accounts opened in the past 24 months
	2: 2-6	
	3: 7+	
	-1: MISSING	
open_acc_6m_segm	1: ≤ 1	Number of open accounts in the last 6 months
	2: > 1	
	-1: MISSING	
emp_length_segm	1: 10+ years	Employment length segmented into ranges
	2: 1-8 years	
	-1: MISSING	
mort_acc_segm	1:00	Number of mortgage accounts
	2:01	
	3:02	
	4: 3+	
instlmnt_round_segm	1: ≤ 479	Installment amount rounded into ranges
	2: 480-699	
	3: 700-879	
	4: ≥ 880	
purpose_segm	1: car, house, major purchase, medical,	Loan purpose segmented into categories

Segmented Variable	Segments	Description
	moving, small business	
	2: debt consolidation, home improvement, other, renewable energy, vacation, wedding	
	3: credit card	
term_segm	1: 60 months	Loan term length
	2: 36 months	
inq_last_12m_segm	1: ≤ 1	Number of inquiries in the last 12 months
	2: 2-6	
	3: 7+	
	-1: MISSING	
fico_range_low_segm	1: ≤ 680	Borrower's FICO credit score range
	2: 685-700	
	3: 705-720	
	4: 725-790	
	5: ≥ 791	
inq_last_6mths_segm	1: 3+ inquiries	Number of inquiries in the last 6 months
	2: 2 inquiries	
	3: 1 inquiry	
	4: 0 inquiries	
verification_status_segm	1: Verified	Borrower's verification status
	2: Source Verified	
	3: Not Verified	
mths_since_recent_inq_segm	1: 0-2 months	Months since the borrower's most recent inquiry
	2: 3-10 months	
	3: 11+ months	
	-1: MISSING	
total_pymnt_round_segm	1: ≤ 4,999	Total payment rounded into different ranges
	2: 5,000-9,999	
	3: 10,000-14,999	
	4: ≥ 15,000	

Table 5.5.4.1: Neural Networks Variables

.

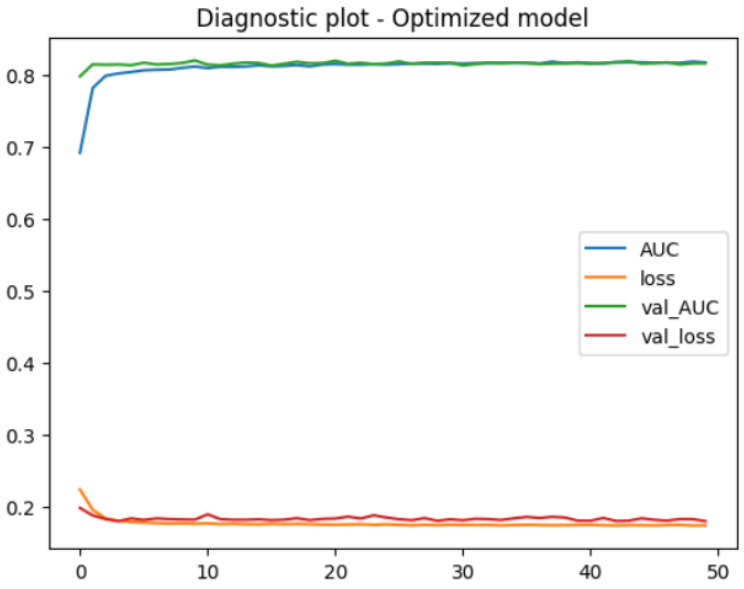


Figure 5.5.4.1: NN Diagnostic Plot

As can be seen from the diagnostic plot, the improvement of the model can be observed almost up to the 10th epoch, where AUC metric is close to its maximum value, and the corresponding loss curves have almost stabilized. This probably indicates that the data fed to the algorithm are not complex enough for this kind of problems, as the characteristics have been segmented -having also in mind the credit risk logic (hence that the patterns within the variables should be intuitively interpretable), as well as to prevent potential overfit of the algorithm.

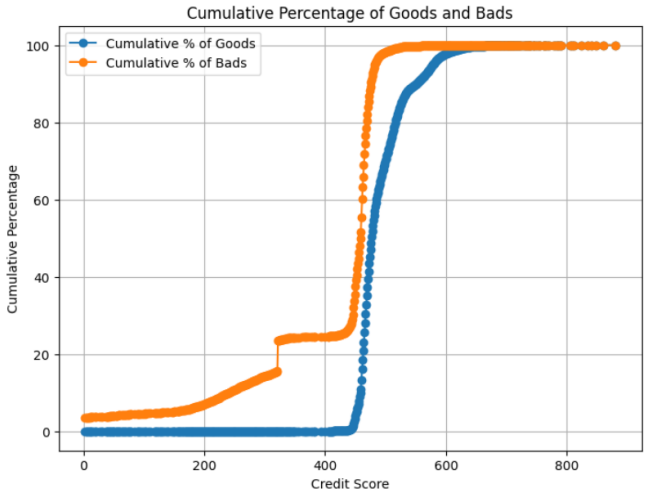


Figure 5.5.4.2: KS - NN – Training Set

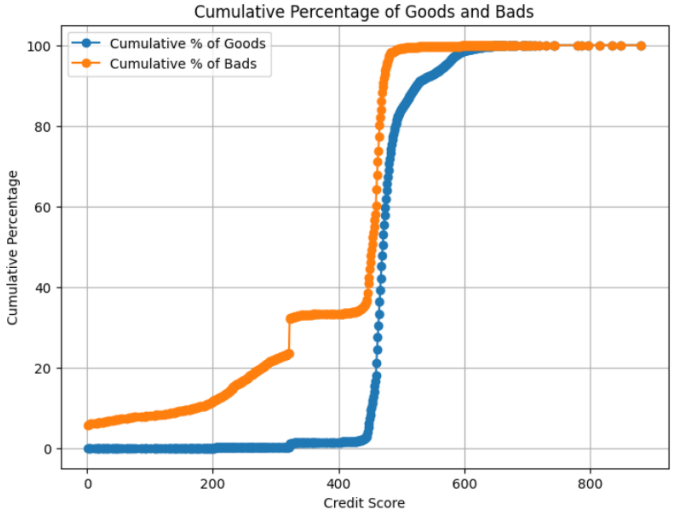


Figure 5.5.4.3: KS – NN – Test Set

The KS statistic is quite high in both the training and test sets (46.46% vs 44.02% respectively), indicating that the model generalizes well on unseen data. However, it can be seen from the corresponding good-bad distributions, that the model fails to discriminate the two populations in the middle scores. Specifically, it seems that the model is not discriminating well between the scores of around 420-500, with the cumulative distributions abruptly increasing in this scoring interval.

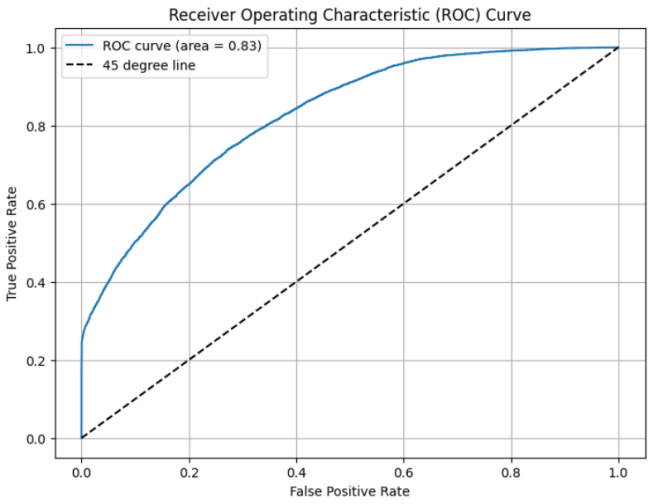


Figure 5.5.4.4: ROC Curve – NN – Training Set

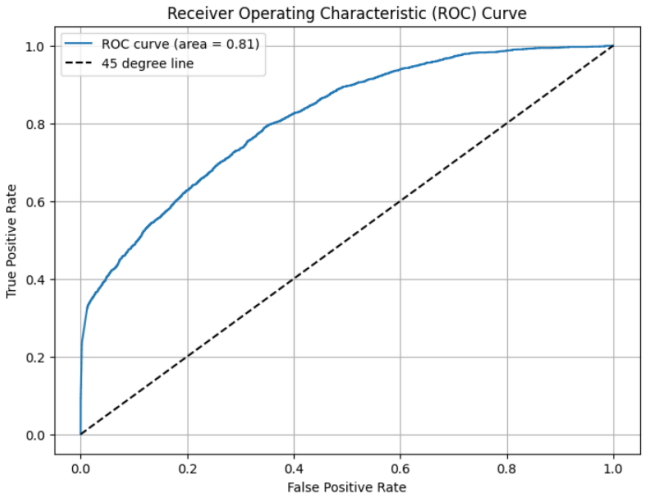


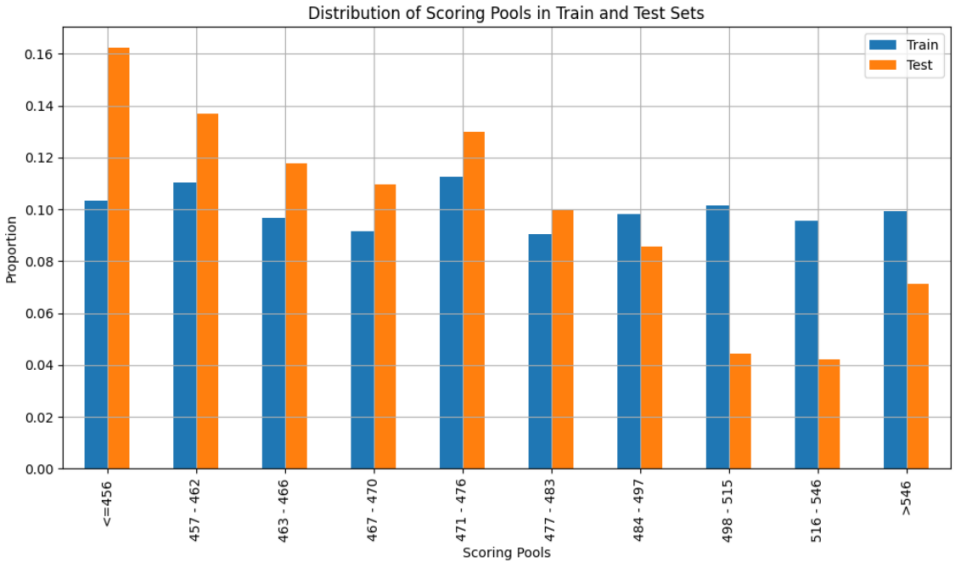
Figure 5.5.4.5: ROC Curve - NN – Test Set

The AUC metric (0.83 training vs 0.81 test set) indicates a minimal overfit, with the generalization on unseen data remaining in high levels.

```
The Kolmogorov-Smirnov statistic on the TRAINING data is: 46.46
AUC metric on the TRAINING data is: 0.83
Gini metric on the TRAINING data is: 0.65

The Kolmogorov-Smirnov statistic on the TEST data is: 44.02
AUC metric on the TEST data is: 0.81
Gini metric on the TEST data is: 0.63
```

Figure 5.5.4.6: Output Metrics – NN – Training Vs Test Set




```
The PSI statistic between Training and Test sets is: 0.145
Moderate shift in the population (PSI = 0.145)
```

Figure 5.5.4.7: Credit Score PSI – NN – Training Vs Test Set

Again, same as in all the previous model implementations, the PSI statistic of 0.145 indicates a moderate shift between the training and test set distributions.

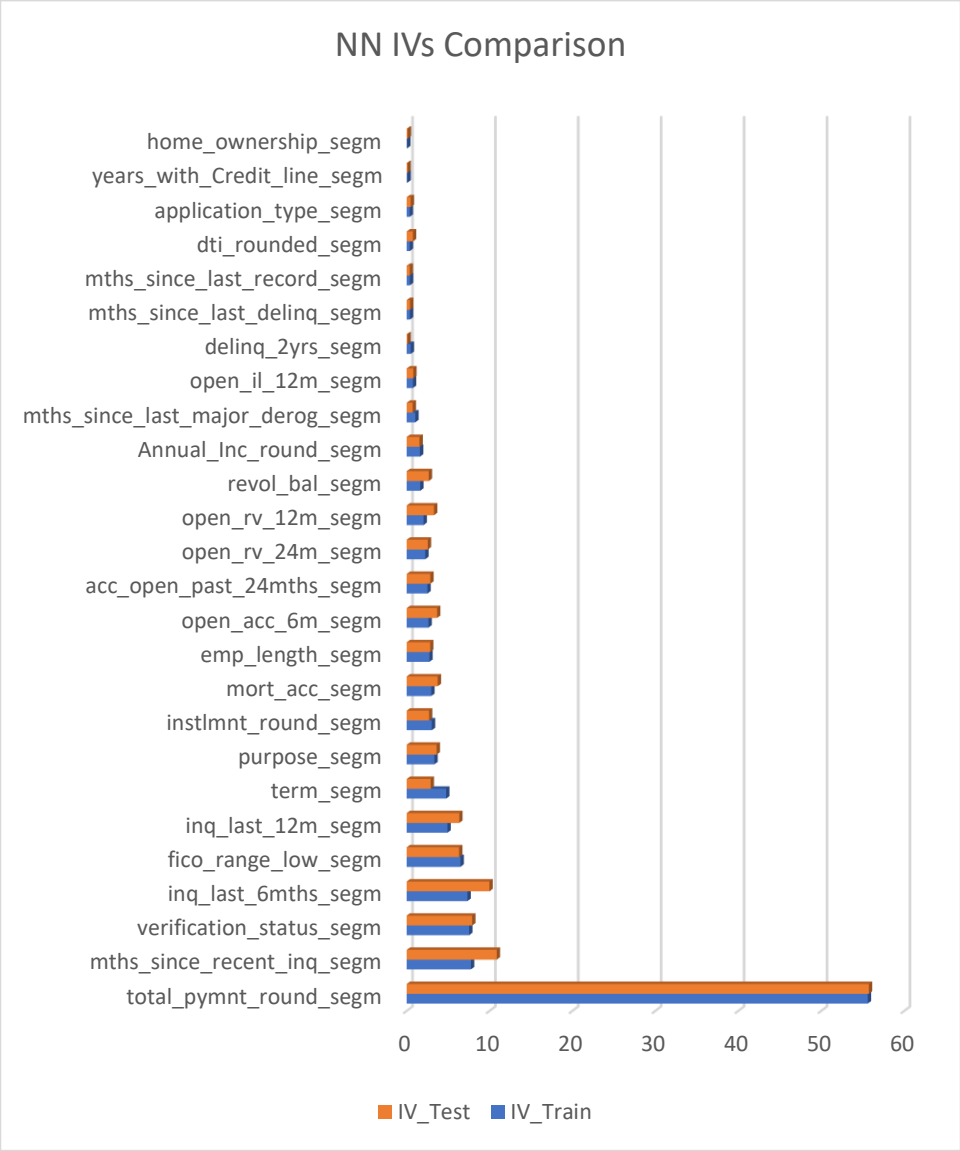


Figure 5.5.4.8: IV Comparison – NN - Training Vs Test Set

No significant shifts are observed in the predictive power of the characteristics used by the model, indicating that these characteristics continue to capture the patterns that were observed on the training set on unseen data. However, it should be mentioned that a slight increase in the predictive power is observed in the characteristics of "inq_last_6_mths_segm" and "mths_since_recent_inq_segm", as also discussed in the previous implementations.

5.5.5 Model Comparison – Performance Metrics

Below are presented the separation metrics, as well as the corresponding PSI tests:

Model	KS Training	KS - Test	Gini - Training	Gini - Test	AUC - Training	AUC - Test	PSI
LR	46.86%	43.40%	0.66	0.63	0.83	0.81	0.16

RF	48.74%	42.39%	0.68	0.62	0.84	0.81	0.15
GB	49.23%	44.15%	0.68	0.63	0.84	0.82	0.15
NN	46.46%	44.02%	0.65	0.63	0.83	0.81	0.14

Table 5.5.5.1: Model Performance Comparison

In evaluating the model performance metrics, it is evident that the segmentation of the variables significantly contributed to preventing overfitting across all models. This is reflected in the relatively close performance between the training and test sets for the various algorithms. For instance, the LR model demonstrated a KS statistic of 46.86% on the training set and 43.40% on the test set, alongside Gini coefficients and AUC scores that show minimal drop-off from training to testing. This indicates strong generalization capabilities.

While being the simplest model among the evaluated ones, LR stands out for its high interpretability and robust generalization. Despite its simplicity, LRs' performance in terms of AUC and Gini on both training and test sets is competitive with more complex models such as RF, GB and NN. The respective AUC scores for the training and test sets in LR are 0.83 and 0.81, which are only marginally lower than those of the other models, underscoring LRs effectiveness without the complexity of "black-box" models.

It is important to mention (once again) that the segmentation of characteristics played a pivotal role in stabilizing the model predictions and avoiding overfitting. For example, variables like total_pymnt_round_segm, verification_status_segm, and fico_range_low_segm have high IVs, indicating robust predictive power, with no shifts in the predictive power being observed when compared to the test set -except in the LR implementation (as discussed to the corresponding analysis).

The PSI values indicate that the credit scores of the test set were more concentrated in the lower score ranges, probably attributed to potential over-penalization of certain population segments. That means that some certain variables may over-penalize some specific characteristics of the customer. Variables with augmented IVs between training and test sets, such as mths_since_recent_inq_segm and inq_last_6mths_segm, suggest improved separation power in the test set, but may also imply that these predictors are disproportionately impacting specific segments. This underscores the need for a comprehensive characteristic analysis to understand why lower scores are prevalent in the test set across all the implemented models -as to ensure balanced PD assessment. Therefore, it's crucial to perform a detailed comparison of the training and test characteristics to identify and address any underlying biases, ensuring that the scoring models remain fair and accurate across different populations. However, apart from the characteristic analysis, another thing that could cause the shift of the score towards the lower scoring pools, could also be attributed to the fact that the customers of the test set are indeed riskier, and thus are correctly scored in lower scoring pools by the corresponding models.

To conclude, all the tested models demonstrated strong separation statistics, indicating high discriminatory power. The discrimination statistics (i.e. KS, Gini, AUC) indicate that both the LR and NN models exhibited minimal overfitting to the training set, while the GB and RF models showed slightly more overfitting.

Between the LR and NN models, LR is more suitable for this type of modeling, primarily because of its high interpretability. While Neural Networks are often considered “black-box” models, making them harder to interpret, LR provides clear insights into the relationship between variables and the predicted outcome. Although the LR model produced satisfactory discrimination statistics, there is room for improvement, particularly in its performance on hard-to-classify cases, as shown in the middle score ranges of the cumulative distributions for "goods" and "bads." Despite this, the combination of high discriminatory power and interpretability makes LR the better choice for this analysis.

CHAPTER 6

Conclusion

This thesis analyzed the critical role of PD estimation in financial institutions. Initially, it detailed the risks and challenges faced by financial institutions, highlighting the fundamental importance of accurate PD estimation. A comprehensive literature review was conducted, covering both traditional statistical models and recent advancements in ML, along with industry-standard performance metrics such as KS, Gini, and AUC.

The theoretical background of LR, RF, GB, and NN was thoroughly discussed. In the fifth chapter, these models were implemented using real data from LendingClub.com, each in order to create a corresponding application scorecard. All tested models achieved high separation statistics, indicating strong discriminatory power. Specifically, the discriminatory power was measured using KS, Gini, and AUC metrics.

The analysis revealed that LR and NN exhibited minimal overfitting to the training set, whereas GB and RF displayed slightly more overfitting. Between LR and NN, LR emerged as more suitable for such modeling purposes due to its high interpretability, whereas NN are regarded as “black-box” models.

Additionally, the IV criterion and PSI test were utilized for the segmentation analysis and monitoring of the models. These metrics provided further insights into the stability and predictive power of the models over-time.

In summary, this thesis demonstrated that statistical ML methods can be effectively applied to estimate the PD. The results underscore the importance of selecting appropriate models based on the specific requirements of credit risk assessment, as well as the need for continuous optimization and evaluation to achieve the best possible outcomes.

Appendix

The dataset was downloaded from the following link: https://www.kaggle.com/datasets/wordsforthewise/lending-club

The dictionary of the dataset can be found on the following link: https://github.com/dosei1/Lending-Club-Loan-Data/blob/master/LCDataDictionary.csv

Literature

Greek

[1] Μπούτσικας Μ. (2023). Διαχείριση Κινδύνων. Σημειώσεις Μεταπτυχιακού Μαθήματος του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς.

[2] Saunders Cornett (2017). Διοίκηση Χρηματοπιστωτικών Ιδρυμάτων και Διαχείριση Κινδύνων.

[3] Ατζουλάτος Άγγελος (2011). Κυβερνήσεις Χρηματαγορές και Μακροοικονομία.

[4] Sapountzoglou G., Pentotis C. (2017). Banking Economics.

[5] Iliopoulos, G. (2022). Generalized Linear Models, Notes for the postgraduate course in the Master of Science in Applied Statistics, Department of Statistics and Insurance Science, University of Piraeus.

[6] Politis K. (2022). Generalized Linear Models, Notes for the postgraduate course in the Master of Science in Applied Statistics, Department of Statistics and Insurance Science, University of Piraeus.

[7] Koutras M., Boutsikas M. (2011). Statistics II, Notes for undergraduate course, Department of Statistics and Insurance Science, University of Piraeus.

[8] Bersimis S. (2021). Statistical Machine Learning, Notes for the postgraduate course in the Master of Science in Applied Statistics, Department of Statistics and Insurance Science, University of Piraeus.

[9] Iliopoulos, G. (2023). Statistical Machine Learning, Notes for the postgraduate course in the Master of Science in Applied Statistics, Department of Statistics and Insurance Science, University of Piraeus.

[10] Tasoulis, S. (2021). Statistical Machine Learning - Introduction to Neural Networks, Notes for the postgraduate course in the Master of Science in Applied Statistics, Department of Statistics and Insurance Science, University of Piraeus.

Foreign

[11] Altman E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance*, **23**, 189-209.

[12] Markowitz Harry. (1952). Portfolio Selection, *The Journal of Finance*, **Vol 7**, 77-91.

[13] Markowitz Harry. (1959). *Portfolio Selection: Efficient Diversification of Investments*, Yale University Press, New Haven.

[14] Sommerville, R.A., Taffler, R.J., 1995. Banker judgement versus formal forecasting models: The case of country risk assessment. *Journal of Banking and Finance, 281-297*.

[15] Daniel Martin. (1977). Early Warning of Bank Failure – A Logistic Regression Approach, *Federal Reserve Bank of New York*, New York.

[16] Altman et al. (1977). ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations. *Journal of Banking and Finance*, **Vol 1**, 29-54

[17]Oldrich Alfons Vasicek (1984). Credit Valuation, *KMV, LLC (KMV)*, San Fransisco, California, USA.

[18]Robert A. Eisenbeis. (1978). Problems in applying discriminant analysis in credit scoring models, *Journal of Banking & Finance*, **Vol 2**, 205-219.

[19]Myers, Greeley and Siera, Steven. (1980). Development and Validation of Discriminant Analysis Models for Student Loan Defaultees and Non-Defaultees", *Journal of Student Financial Aid*, **Vol. 10**.

[20]Mona J. Gardner, Dixie L. Mills. (1989). Evaluating the Likelihood of Default on Delinquent Loans, *Financial Management*, **Vol. 18**, 55-63.

[21]Lawrence et al. (1992), An analysis of default risk in mobile home credit, *Journal of Banking & Finance*, **Vol. 16**, 299-312.

[22]William Greene. (1998). Sample Selection in Credit-Scoring models, *Japan and the World Economy*, **Vol. 10**, 299-316.

[23]Gabriel Sabato. (2008). Solving Sample Selection Bias in Credit Scoring: The

Reject Inference, Bank of Scotland.

[24]Adrien Ehrhardt, Christophe Biernacki, Vincent Vandewalle, Philippe Heinrich, S_ebastien Beben. (2021). Reject Inference Methods in Credit Scoring.

[25]Amos Taiwo Odeleye, (2019). Developing a Credit Risk Model Using SAS®, *TD Bank*.

[26]Douglas L Weed, (2006). Weight of Evidence: A Review of Concept and Methods, *Risk Analysis*, **Vol. 25**, 1545-57

[27]Bruce Lund, David Brotherton. (2013). Information Value Statistic, *Magnify Analytics Solutions*, Detroit, MI.

[28]Zheng Yang, Yue Wang, Yu Bai, Xin Zhang. (2004). Measuring Scorecard Performance, *Colledge of Economics*, China.

[29]Joseph L. Gastwirth. (1972). The Estimation Of The Lorenz Curve And Gini Index, The Review of Economics And Statistics, **Vol. 54**, pp 306-316.

[30]Andrew P. Bradley, (1997). The Use Of The Area Under The ROC Curve In The Evaluation of Machine Learning Algorithms, *Pattern Recognition*, **Vol. 30**, pp 1145-1159.

[31]Bilal Yurdakul, Joshua Naranjo, (2021). Statistical Properties of Population Stability Index, USAA.

[32]Tam K.Y., Kiang M., (1992). Neural Network Models and the Prediction of Bank Bankruptcy, *Decision Sciences*, **Vol. 23**, pp 926-947.

[33]Banu Yobas, Jonathan Crook, (2000). Credit scoring using and evolutionary technique, *IMA Journal of Mathematics Applied in Business and Industry*, **Vol. 11**, pp 111-125.

[34]Loris Nanni, Alessandra Lumini. (2009). An experimental comparison of ensemble of classifiers, *Science Direct*, **Vol. 36**, 3028–3033

[35]Stefan Lessmanna, Bart Baesens, Hsin-Vonn Seowd, Lyn C. Thomas, (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research*, **Vol. 247**, pp 124-136.

[36]Pedro G. Fonseca, Hugo D. Lopes. (2017). Calibration of Machine Learning Classifiers.

[37]Büşra Alma Çallı, Erman Coşkun, (2021). A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors.

[38]Jerome H. Friedman, Robert Tibshirani, Trevor Hastie (2009). The Elements of Statistical Learning, second edition.

[39]Alexey Natekin, Alois Knoll, (2013). Gradient Boosting Machines, A Tutorial.

[40]Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, Weinan E. (2020). Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning.

[41]Camilla Calì, Maria Longobardi, (2015). Some mathematical properties of the ROC curve and their applications.

[42]Petro Liashchynskyi, Pavlo Liashchynskyi, (2019). Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS.

[43]Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, Si-Hao Deng, (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization.

[44]Brown, Iain. (2014). Developing Credit Risk Models Using SAS® Enterprise MinerTM and SAS/STAT®: Theory and Applications. Cary, NC: SAS Institute Inc.

Web Pages

[45]PWC - Audit Services - IFRS9 - Impairment Stages - Significant Increase In Credit Risk.pdf

[46]PWC - EBA Credit Risk - Default of Definition

[47]Κατευθυντήρια γραμμή (ΕΕ) 2020/978 της Ευρωπαϊκής Κεντρικής Τράπεζας της 25ης Ιουνίου 2020 σχετικά με την άσκηση της διακριτικής ευχέρειας του άρθρου 178 παράγραφος 2 στοιχείο δ) του κανονισμού (ΕΕ) αριθ. 575/2013 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου από τις εθνικές αρμόδιες αρχές όσον αφορά το όριο βάσει του οποίου εκτιμάται το ουσιώδες των καθυστερημένων πιστωτικών υποχρεώσεων των λιγότερο σημαντικών ιδρυμάτων (ΕΚΤ/2020/32)

[48]ECB - What makes a Bank significant?

[49]The ABCs of Basel I, II & III