UNIVERISTY OF PIRAEUS - DEPARTMENT OF INFORMATICS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

MSC «Cybersecurity and Data Science»

ΠΜΣ «Κυβερνοασφάλεια και Επιστήμη Δεδομένων»

MSc Thesis

Μεταπτυχιακή Διατριβή

Thesis Title: Τίτλος Διατριβής:	**Credit Risk Analysis using Machine Learning Methods and Explainable AI** Ανάλυση πιστωτικού κινδύνου με χρήση μεθόδων μηχανικής μάθησης και εξηγήσιμης τεχνητής νοημοσύνης
Student's name-surname: Ονοματεπώνυμο Φοιτητή:	**Ailina Sopileidi** Αιλήνα Σωπηλείδη
Father's name: Πατρώνυμο:	**Vladimir** Βλαδιμίρ
Student's ID No: Αριθμός Μητρώου:	**ΜΠΚΕΔ21051**
Supervisor: Επιβλέπων:	**Dimitrios Apostolou, Professor** Δημήτριος Αποστόλου, Καθηγητής

July 2024/Ιούλιος 2024

Στην ολοκλήρωση της παρούσας μεταπτυχιακής διατριβής, ιδιαίτερα σημαντική ήταν η συμβολή του Διδάσκοντα του ΠΜΣ **κ. Ανδρέα Ζάρα**, που προσέφερε επιστημονική και συμβουλευτική καθοδήγηση σε όλα τα στάδια εκπόνησής της.

3-Member Examination Committee

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Dimitrios Apostolou **Professor**	**Aggelos Pikrakis** **Assistant Professor**	**Dionisios Sotiropoulos** **Assistant Professor**
Δημήτριος Αποστόλου Καθηγητής	Άγγελος Πικράκης Επίκουρος Καθηγητής	Διονύσιος Σωτηρόπουλος Επίκουρος Καθηγητής

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Αποστόλου Δημήτριο και τον Διδάσκοντα κ.Ζάρα Ανδρέα για όλη την καθοδήγηση, τις χρήσιμες συμβουλές και την γενικότερη αποτελεσματική μας συνεργασία καθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Επίσης, θέλω να ευχαριστήσω τα κοντινά μου άτομα, που με στήριξαν και με έμαθαν να μην τα παρατάω, στην οικογένειά μου και τον σύντροφο μου.

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στους καθοριστικούς παράγοντες της αθέτησης δανείων των πλατφορμών δανεισμού peer-to-peer και αναλύει αν μπορεί να προβλεφθεί η εξαρτημένη μεταβλητή “αθέτηση”. Τα σύνολα δεδομένων που χρησιμοποιούνται για αυτήν την έρευνα είναι 211.283 δάνεια της πλατφόρμας Bondora και 38 μεταβλητές. Η χρονική περίοδος που χρησιμοποιείται είναι από τον Φεβρουάριο του 2009 έως τον Οκτώβριο 2022 και λαμβάνονται υπόψη μόνο τα ολοκληρωμένα (εξοφλημένα ή αθετημένα) δάνεια. Αυτό θα γίνει χρησιμοποιώντας ένα μοντέλο δέντρου αποφάσεων για τη μέτρηση της πιθανότητας αθέτησης των δανείων.

Τα κύρια ευρήματα αυτής της έρευνας είναι ότι η ηλικία, η μηνιαία δόση πληρωμής, το επιτόκιο, το κεφάλαιο που πρέπει ακόμα να καταβληθεί από τον δανειολήπτη η διάρκεια του δανείου και η αναλογία χρέους προς εισόδημα σχετίζονται θετικά με την πιθανότητα αθέτησης. Αντίθετα, η μεταβλητή του συνολικού αριθμού προηγούμενων δανείων, το εισόδημα και η αξία προηγούμενων δανείων έχουν αρνητική συσχέτιση με τον κίνδυνο αθέτησης.
Επιπλέον, εκτελείται ένα μοντέλο δέντρου αποφάσεων και η σκορ κάρτα για την πρόβλεψη της πιθανότητας αθέτησης δανείου και στην συνέχεια εξετάζεται η προβλεπτική ικανότητα του μοντέλου δέντρου αποφάσεων, μέσω σύγκρισης με άλλα μοντέλα μηχανικής μάθησης όπως νευρονικά δίκτυα,το Random Forest & Decision Tree.

Με βάση τα μέτρα πρόβλεψης, το ποσοστό εσφαλμένης ταξινόμησης και την καμπύλη ROC, φαίνεται ότι το μοντέλο δέντρου αποφάσεων έχει σχετικά καλή προβλεπτική ικανότητα. Επιπρόσθετα, εξετάσαμε την ικανότητα των πελατών να αποπληρώσουν πιστωτικά δάνεια ταξινομώντας τους δανειολήπτες ως «υψηλού κινδύνου ή «χαμηλού κινδύνου» χρησιμοποιώντας σκορ κάρτα στην SAS.

Ο όρος «χαμηλού κινδύνου» δηλώνει ότι ο λήπτης του δανείου έχει αποδεκτή βαθμολογία και δεν υπάρχουν προβληματικά αρχεία πληρωμών. Από την άλλη πλευρά, η φράση «υψηλού κινδύνου» υποδηλώνει το αντίθετο, ότι ο αιτών έχει κακή πιστωτική βαθμολογία ή ότι υπήρχαν αρχεία για καθυστερημένες πληρωμές ή προηγούμενες αθετήσεις. Οι κάρτες βαθμολογίας είναι βασικές στη διαδικασία δανεισμού, καθώς ποσοτικοποιούν τον κίνδυνο που σχετίζεται με τους αιτούντες δάνειο. Μετατρέπουν διάφορα χαρακτηριστικά του δανειολήπτη, όπως το πιστωτικό ιστορικό, τα επίπεδα εισοδήματος και το καθεστώς απασχόλησης, σε μια αριθμητική βαθμολογία.

Αυτή η βαθμολογία υποδεικνύει την πιθανότητα αθέτησης του δανείου από έναν δανειολήπτη, βοηθώντας τους δανειστές να λαμβάνουν τεκμηριωμένες αποφάσεις που εξισορροπούν τον κίνδυνο και την ανταμοιβή.

Abstract

This thesis focuses on the determinants of loan defaults of peer-to-peer lending platforms and analyzes whether the dependent variable “default” can be predicted. The datasets used for this research are 211.283 Bondora platform loans and 38 variables. The time period used is from February 2009 until October 2022 and only completed (paid off or defaulted) loans are considered. P2P lending is the act of lending money to individuals or small and mid-size enterprises via online platforms that connects lenders and borrowers. One of the hot topics in this field is risk assessment of applicants. A P2P lending company, in order to make sure the client will be able to pay back the loan in agreed duration, assesses the risk of each applicant individually.

This will be done using a decision tree model to measure the default probability of loans.The main findings of this research are that age,PrincipalBalance,interest rate, loan duration, MonthlyPayment and Debt to Income are positively related to the probability of default. In contrast, the borrowers’ income, number of previous loans and value of previous loans have a negative correlation with default risk.

In addition, a decision tree model and scorecard is performed to predict the probability of loan default. The predictive ability of the decision tree model is examined, through comparison with other machine learning models as Neural Networks, Random Forest & Decision Trees.

Based on the predictive measures, the misclassification rate, the accuracy and the ROC curve, it appears that the Decision Tree model (Chi-Square) has relatively good predictive ability.

Additionally, we examined the ability of the customers in repaying credit loans by classifying the loan receivers as ‘high risk or ‘low risk’ using scorecard in SAS.

The term ‘low risk’ states that the loan receiver has an acceptable score and there has been no problematic payment records. On the other hand, the phrase ‘high risk’ suggests the opposite, that the applicant has a bad credit score or there were records for delayed payments or past defaults.

Scorecards are essential in the lending process as they quantify the risk associated with loan applicants. They transform various borrower attributes, such as credit history, income levels, and employment status, into a numerical score.

This score indicates the likelihood of a borrower defaulting on a loan, helping lenders make informed decisions that balance risk and reward.

Table of Contents

Περίληψη *4*

Abstract *5*

Introduction *8*

Literature Review *9*

Statistical Methods **9**

Machine Learning Methods **9**

Decision Trees 9

Random Forests and Neural Networks 9

Explainable AI (XAI) Techniques 9

Theoretical Background-Framework *11*

Credit Risk Management **11**

P2P Lending How works **11**

P2P Lending Platform **12**

P2P -Marketplace Lending vs Bank Lending **13**

Pros and Cons of P2P Lending **13**

Advantages 13

Disadvantages 14

Collection Agencies **14**

Problem Definition *14*

Models *16*

Logistic Regression **16**

Decision Trees **17**

Attribute Selection 18

Advantages of Decision Trees 19

Disadvantages of Decision Tree 19

Theory Behind Scorecard Development *21*

Scorecard **21**

Weight of Evidence (WoE) 21

Feature Selection using Information Value - IV 21

Binning of Variables 22

Model Fitting & Interpreting Results 22

Reject Inference 22

Methodology *23*

Data **23**

Dataset **23**

Data Preparation - Exploratory Data Analysis **24**

Descriptive Statistics - Understanding Features **26**

Collection Process **32**

Correlation **33**

Clustering **37**
Customer Segmentation **38**
Importance of AI Interpretability for Credit Risk **43**
Decision Tree and Interpretation **45**
Scorecard Creation **51**
Conclusion ***55***
Bibliography ***59***

Introduction

This thesis focuses on the critical issue of credit risk analysis for loans provided through peer-to-peer (P2P) lending platforms. P2P lending platforms enable direct lending transactions between individuals, bypassing traditional financial institutions. This model of lending introduces unique challenges and opportunities, particularly in the realm of risk assessment and management. P2P lending has grown significantly as an alternative to conventional banking, offering borrowers easier access to credit and providing lenders with potentially higher returns on their investments. However, this form of lending is not without its risks, the most significant of which is the risk of borrower default. Unlike traditional banks, P2P platforms do not have the same level of regulatory oversight and risk mitigation strategies, making the accurate assessment of credit risk even more crucial.

The aim of this thesis is to delve into the problem of credit risk on P2P lending platforms using advanced analytical techniques. Specifically, this research employs machine learning methods and explainable artificial intelligence (AI) to develop predictive models that can forecast the likelihood of loan defaults. This approach not only enhances the accuracy of risk predictions but also ensures that the models are interpretable and actionable for stakeholders. The primary objectives of this study are as follows:

Identify Key Determinants: To pinpoint the main factors that contribute to loan defaults on P2P lending platforms. This involves analyzing a wide range of borrower characteristics and loan attributes to determine their impact on default risk.

Develop Predictive Models: To create and validate machine learning models that can predict the probability of loan defaults with high accuracy. These models will be trained and tested using historical loan data from P2P platforms.

Enhance Model Interpretability: To apply explainable AI techniques that make the predictions of these models understandable to non-expert users. This includes visualizing the decision-making process of the models and identifying the most influential variables in predicting defaults.

By achieving these objectives, this thesis aims to provide valuable insights into the risk management practices of P2P lending platforms. The findings can help these platforms develop more effective strategies to mitigate credit risk, thereby protecting both lenders and borrowers. In summary, this research addresses a significant gap in the current understanding of credit risk management in P2P lending. By leveraging the power of machine learning and explainable AI, it seeks to enhance the accuracy and transparency of risk predictions, ultimately contributing to the stability and growth of the P2P lending industry.

Literature Review

The focus of this thesis is on loan service assessment and forecasting methods, particularly within the context of peer-to-peer (P2P) lending platforms. Therefore, this chapter should concentrate on reviewing existing scientific literature that deals with loan service assessment and forecasting methods, which are essential for understanding and improving credit risk analysis in P2P lending.Accurate loan service assessment is critical for evaluating borrower creditworthiness and predicting potential loan defaults. The literature encompasses a range of methods, from traditional statistical techniques to advanced machine learning models, each with its strengths and limitations.

Statistical Methods

Traditional statistical methods, such as logistic regression and linear discriminant analysis, have long been utilized for credit scoring and default prediction. Logistic regression, in particular, is popular due to its simplicity and ability to provide interpretable results. This method models the probability of default as a function of various borrower characteristics, offering insights into which factors significantly impact credit risk.

Machine Learning Methods

In recent years, machine learning methods have gained traction in the field of credit risk analysis. These methods are capable of handling large datasets and uncovering complex patterns that traditional statistical methods might miss. Among the most studied and applied techniques are decision trees, random forests, and neural networks.

Decision Trees

Decision tree models, including CHAID (Chi-Square Automatic Interaction Detector) and CART (Classification and Regression Trees), are favored for their transparency and ease of interpretation. These models work by splitting the dataset into branches based on the values of input variables, making the decision-making process easy to follow and understand.

Random Forests and Neural Networks

Random forests, an ensemble learning method that constructs multiple decision trees and aggregates their results, offer enhanced predictive performance compared to single decision trees. Neural networks, especially deep learning models, have shown high accuracy in predicting loan defaults by capturing non-linear relationships between variables. However, the complexity of neural networks can make them challenging to interpret .

Random Forests combine multiple decision trees to enhance predictive performance and reduce overfitting. These methods have shown significant improvements in predicting loan defaults by capturing complex interactions among variables (Müller & Guido, 2017) .

Explainable AI (XAI) Techniques

Given the complexity of advanced machine learning models, explainable AI (XAI) techniques have been developed to enhance model transparency. Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) help elucidate the contributions of individual features to the model's predictions. These techniques are crucial for gaining stakeholder trust and ensuring compliance with regulatory requirements (Tulio Ribeiro et al., 2016) .
SHAP values provide a unified measure of feature importance, indicating how much each feature contributes to the prediction, both positively and negatively. LIME approximates complex models with simpler, interpretable models locally around the prediction of interest, making it easier to understand why a particular decision was made.

Credit Risk Management in P2P Lending
Credit risk management in P2P lending platforms presents unique challenges due to the absence of traditional banking infrastructure and regulatory frameworks. Effective risk management strategies in this context rely heavily on advanced analytical methods to evaluate borrower creditworthiness and forecast loan performance.

Applications of Machine Learning in P2P Lending
Numerous studies have demonstrated the potential of machine learning models to improve default prediction accuracy in P2P lending. For example, decision tree models have been effective in identifying critical financial and demographic variables that influence default risk. Scorecards developed using logistic regression and other statistical techniques have been employed to categorize borrowers into risk tiers, facilitating more informed decision-making regarding loan approvals and interest rate determinations .

Emerging Trends and Future Directions
The integration of machine learning and XAI in credit risk analysis is an evolving field with significant potential. Emerging trends include the use of ensemble methods that combine various models to improve prediction robustness and the development of more sophisticated XAI techniques to enhance model transparency. Additionally, there is growing interest in using alternative data sources, such as social media activity and transaction data, to supplement traditional credit scoring models .

Conclusion
In conclusion, this literature review underscores the importance of robust loan service assessment and forecasting methods in the context of credit risk analysis for P2P lending platforms. The combination of machine learning models and explainable AI techniques offers a promising approach to enhancing the accuracy and transparency of credit risk assessments. This review lays the groundwork for the subsequent chapters of the thesis, which will focus on developing and evaluating predictive models for loan defaults in P2P lending, thereby contributing to more effective risk management practices in this growing sector.

Theoretical Background-Framework

Credit Risk Management

In banking, lending is the primary source of profit but also involves significant risks. Essentially, both lending and accepting deposits entail taking on risk. The fundamental nature of a bank is to seek profit while managing acceptable and measurable risks. Credit risk arises when customers are unable to repay their loans. As Murphy (2008) states, “Whenever a bank acquires an earning asset, it assumes the risk that the borrower will default, meaning not repaying the principal and interest on time.” Credit risk extends beyond loan products to other credit instruments, such as letters of credit and guarantees, where the bank agrees to act on behalf of a client if they fail to meet their business commitments, as well as in investment services or asset finance, where the bank lends out tangible assets like land, properties, and equipment (Murphy, 2008, pp. 203-204).

Credit analysis involves the assessment of credit files by bank officers to evaluate the repayment ability of current or prospective borrowers. This process addresses three key questions:

- What risks are present in the borrower’s business?
- What measures has the borrower taken to manage these risks?
- Have they been successful or not, and why?
- What actions can the bank take internally to mitigate potential losses when providing credit?

To answer these questions, credit analysts conduct both subjective and objective evaluations of credit risk. In subjective evaluations, banks often apply the five Cs of good credit as crucial criteria for assessing borrowers:

- **Character**: Reflects the borrower’s willingness to repay, their reputation in the industry, and their relationships with other lending institutions. Bank officers review historical transactions to identify any credit-related issues. For example, if a firm has consistently repaid its interest on time in the past, it is likely to maintain its reputation and fulfill its loan obligations for new loans.
- **Capital**: Refers to the borrower’s capital structure. Credit analysts examine the level of leverage by evaluating the balance of debt and equity used as financing sources.
- **Conditions**: Encompasses external factors that might affect the borrower’s financial situation and repayment ability. These include economic conditions and industry-related factors. For instance, the tobacco industry is highly sensitive to market changes, with demand fluctuating based on economic growth or downturns.
- **Capacity**: Focuses on the customer’s cash flow reports. Banks prefer to lend to firms with predictable, stable cash flows and alternative credit sources to ensure loan repayment.
- **Collateral**: Consists of assets pledged by the borrower to secure the loan. If the borrower fails to repay, the bank can sell these assets to recover part or all of the loss.

P2P Lending How works

Peer-to-peer (P2P) lending operates by connecting individuals who need loans with investors looking to earn returns on their investments. Borrowers submit loan requests to a P2P lending platform, where investors then compete to finance these loans in exchange for an interest rate. The P2P platform manages the entire process, including creditworthiness assessment, loan servicing, payments, and collections.

Initially, an investor creates a profile on the platform and transfers funds to be used for lending. Borrowers submit their financial information, which is used to assign a risk rating, determining the interest rate they will be charged. Investors can then review various loan offers and choose those that match their desired risk-reward ratio. Once loans are funded, borrowers start making interest payments as they repay their debt according to the agreed schedule. The P2P lending platform oversees the distribution of funds and the collection of loan repayments.

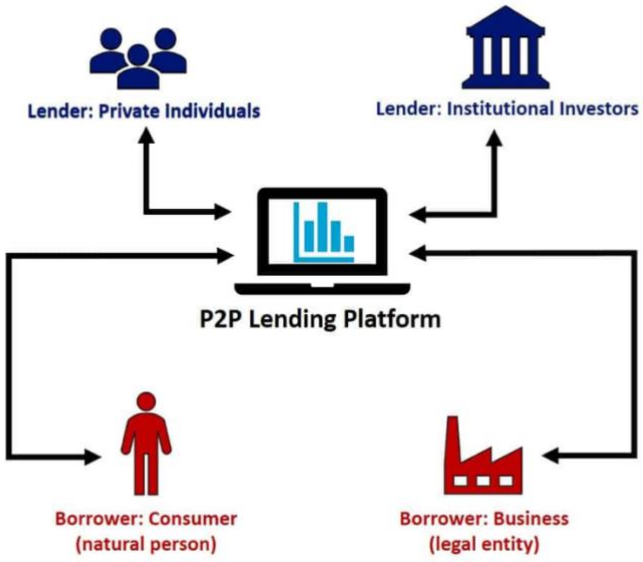

Figure 1: How Peer-to-Peer (P2P) Lending Works
taken from site p2pmarketdata.com

P2P Lending Platform

In P2P lending, loans are typically unsecured, meaning they lack physical collateral, and lenders seek higher returns to compensate for the increased financial risk. These decisions are made under conditions of information asymmetry, which often benefits borrowers. To make informed choices, lenders aim to minimize the risk of default for each loan and ensure that the returns adequately compensate for the associated risks.

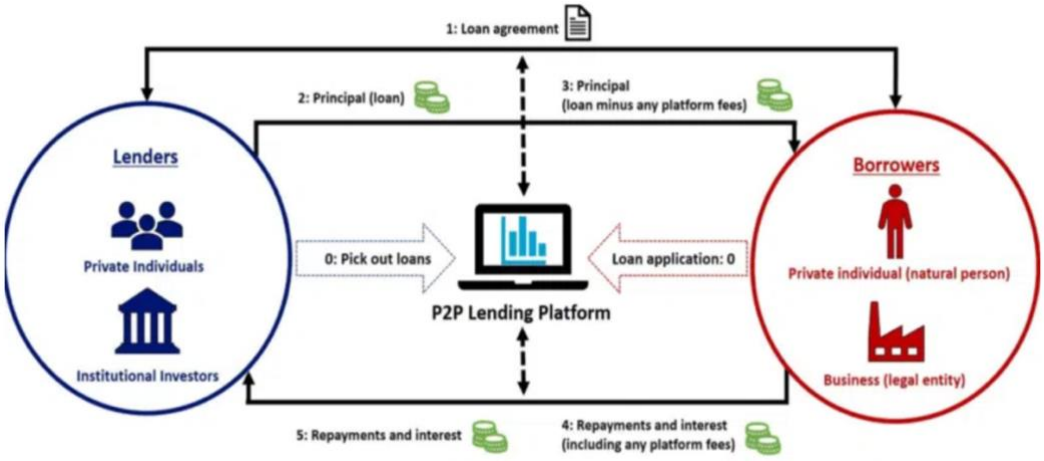

Figure 2 : Overview of Peer-to-peer lending framework
taken from site p2pmarketdata.com

Peer-to-peer lending encompasses various models and types, including personal loans, business loans, student loans, and mortgage financing, with most platforms specializing in one or two borrower types. Personal loans are the most common, usually unsecured, ranging from $1,000 to $25,000, with repayment terms of up to five years. Business loans offered on P2P sites typically range from $50,000 to $500,000 with flexible repayment options.

Two of the most popular P2P lending platforms are Lending Club and Prosper:

- **Lending Club**: This platform facilitates personal and small business loans in the USA. Investors can achieve returns of up to 3.65% from their savings account.
- **Prosper**: Another leading platform, Prosper offers unsecured personal loans of up to $50,000 with competitive interest rates for American citizens. Investors can purchase notes backed by these loans with minimal fees and a low minimum investment requirement.

P2P -Marketplace Lending vs Bank Lending

Consumers primarily utilize P2P loans for various purposes such as debt consolidation, credit card debt repayments, and home improvements. Despite the different purposes, marketplace lending generally mirrors the behavior of non-revolving unsecured borrowing from traditional banking institutions. A notable example is LendingClub, which in its early stages operated much like a technology firm, allocating significant marketing resources to platforms like Facebook and leveraging word-of-mouth promotions. Unlike traditional banks, marketplace lending leverages technological innovations to enhance efficiency and derive deeper insights into borrower behaviors.

There are several key distinctions between P2P lending and traditional bank lending:
a) Income Verification: Marketplace lenders verify a borrower's income by obtaining direct access to the borrower's bank accounts. This allows for a more accurate analysis of cash inflows and repayment ability, whereas traditional banks rely on voluntary disclosures, which can be easily manipulated or falsified.
b) Loan Pricing: The pricing of loans in marketplace lending is typically high enough to cover potential losses from accrued interest. Marketplace lending platforms do not bear principal losses, making them less conservative compared to banks, which are more cautious in granting loans since they have more at stake.
c) Application Process: The application process for a P2P loan is significantly faster than that of a bank loan. A P2P loan application can be completed in as little as 30 minutes, whereas a bank may take several hours or even days to process a loan application. This expedited process reduces costs associated with paperwork and time consumption in traditional banking.
d) Securitization and Risk Diversification: An important aspect of financial innovation in P2P lending includes the securitization of loans and the use of public risk diversification. The P2P loan process involves two main components: loan origination and loan securitization. In contrast, banks typically retain the risk on their balance sheets unless they engage in portfolio restructuring or hedging activities.
e) Interest Rates: The interest rates from marketplace lending platforms and banks differ in their attributes. Marketplace lending platforms often offer competitive rates that reflect the innovative underwriting frameworks and risk management practices they employ.

Pros and Cons of P2P Lending

Advantages

- **Anonymity and Versatility:** Investors can view loan requests without personally identifying the borrower. P2P loans can be used for a wide range of purposes, including substituting for second mortgages, home equity credit lines, or traditional bank loans.
- **Speed and Convenience**: The entire process, from application to receipt of funds, can be completed within two to three days. Traditional loans, by comparison, can take weeks or even months to finalize.
- **Competitive Interest Rates**: P2P loans generally offer competitive interest rates and fixed monthly payments. The application process does not affect credit scores, and credit requirements are often less stringent than those of traditional lenders.

Disadvantages

- **Default Risk and Collections**: Lenders face significant risks if borrowers default. They must manage the collection process unless they engage a facilitator company or collection agency, both of which charge fees. To mitigate potential losses, lenders should diversify their investments by making multiple small loans rather than a few large ones. Consulting a financial advisor is recommended to manage these risks effectively.
- **Increased Borrower Risk**: P2P loans can be riskier for borrowers. A 2017 report from the Federal Reserve Bank of Cleveland found that borrowers increased their credit card balances by 34% after obtaining P2P loans. Although the report was later retracted due to methodological issues, it highlights potential risks. However, P2P loans are generally safer than payday loans or credit card cash advances.

Collection Agencies

Collection agencies are hired by lenders or creditors to recover overdue funds or defaulted accounts. Typically, creditors engage collection agencies after multiple failed attempts to collect receivables. Debt-collection activities may be outsourced to third-party agencies, or managed by an internal department or subsidiary.
There are two main types of third-party debt collectors:

- Agencies Collecting on Behalf of Creditors: These agencies pursue payments in exchange for a percentage of the collected amount, typically ranging from 25% to 50%.
- Debt Buyers: These entities purchase debt from creditors for a fraction of its value, reflecting the low likelihood of collection.

Some examples of collection agencies include Cepal, Do Value, and Interu.

- doValue Greece: An independent loan and real estate management company and a leader in the Greek market. Part of the doValue SpA Group, it manages portfolios exceeding €162 billion, serving as a growth hub in Southeast Europe.
- Cepal: Specializes in providing servicing management for loans and credit receivables, offering mutually beneficial solutions for both clients and borrowers.

Problem Definition

The primary problem is both a prediction and classification problem. Specifically, it aims to forecast the likelihood of loan defaults and categorize borrowers into different risk levels based on their probability of default. This classification aids in making informed lending decisions and effectively managing the risk portfolio.

Variables Involved:
The analysis involves several key variables, which can be broadly categorized into borrower characteristics, loan attributes, and external economic factors. Examples of these variables include:

Borrower Characteristics:

- **Age:** Age can be an indicator of life stability and financial responsibility.
- **Employment Status:** The type and stability of employment can affect the ability to repay loans.
- **Income Level:** Higher income levels generally suggest a greater ability to repay.
- **Credit History:** Past credit behavior is a strong predictor of future repayment behavior.
- **Loan Attributes:**
- **Loan Amount:** The principal amount borrowed can impact repayment ability.
- **Loan Term:** The duration of the loan affects the repayment schedule and the total amount of interest paid.
- **Interest Rate:** The cost of borrowing money, which impacts the total repayment amount.
- **Repayment Schedule:** The frequency and amount of repayments.

External Economic Factors:

- **Economic Growth Indicators:** Metrics such as GDP growth can influence borrowers' ability to repay loans.
- **Inflation Rate:** Inflation affects purchasing power and can impact borrowers' financial stability.
- **Market Interest Rates:** Fluctuations in market interest rates can influence borrowers' repayment ability and the attractiveness of loans.

These variables include both continuous variables (e.g., income level, loan amount) and categorical variables (e.g., employment status, credit history).

Robustness of model

In predictive modelling there is a problem in robustness and ensuring the robustness of machine learning (ML) models is essential for maintaining their reliable performance, especially in critical applications such as credit risk assessment. Robustness means that a model can continue to make stable and accurate predictions despite the unpredictable and changing nature of real-world data.

- **Stochasticity in Variables:**

  Stochasticity in variables is a crucial factor in credit risk assessment for peer-to-peer (P2P) lending on the Bondora platform. This concept refers to the inherent randomness and unpredictability found in key variables such as borrower characteristics, loan attributes, and external economic conditions. For instance, borrower behavior can be affected by unforeseen personal events like sudden job loss or medical emergencies, making it challenging to predict. Additionally, external economic factors, such as market interest rates, inflation, and economic growth, can vary due to macroeconomic conditions and policy changes, adding further unpredictability. These random variations highlight the necessity of employing robust modeling techniques that can manage and account for these uncertainties. The stochastic nature of these variables underscores the importance of continuous monitoring and dynamic updating of models to maintain their accuracy and relevance, thereby supporting effective risk management and decision-making in the P2P lending environment on Bondora.

- **Validity of models overtime**

  The validity of ML models pertains to their ongoing accuracy and relevance as data changes over time. Models can become outdated due to shifts in data distributions, changes in business logic, or other external factors. Regular updates and retraining help models stay relevant and valid. Monitoring for data drift and implementing strong preproduction and postproduction practices are vital for maintaining model validity. Balancing advanced techniques with model interpretability ensures that models not only perform well but are also understandable and trustworthy to stakeholders. Long-term validity involves creating adaptive systems that can self-correct and manage dynamic environments. By focusing on both robustness and validity, ML systems can provide reliable, accurate, and trustworthy predictions, which are crucial for applications like credit risk assessment in peer-to-peer lending platforms.

Models

This section explains the models used to create the credit model in this thesis. A discussion of advantages and disadvantages is also included. Logistic regression is included because this is a popular model within credit scoring.Decision trees is simple machine learning model which is easy to explain.

Logistic Regression

In the banking sector, logistic regression is the most widely utilized method for developing credit models. This approach allows for estimating the relationship between a set of independent variables X and a binary dependent variable Y. In the context of credit scoring, logistic regression is particularly useful for predicting whether a borrower will default on a loan or not.Logistic regression expresses the probability that the dependent variable Y equals one, given the features X.

The equation is:

$$\Pr(y = 1|X_1, \ldots., X_k) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}} = p$$

where:

- Pr(y=1|$X_1, \ldots, X_k$) Probability that the outcome y equals 1 given the predictors X_1, X_k.
- e: The base of the natural logarithm, approximately equal to 2.71828.
- β_0: The intercept term of the logistic regression model.
- $\beta_1, \ldots., \beta_k$:The coefficients of the predictor variables $X_1, \ldots., X_k$ respectively.
- $X_1, \ldots, X_k$: The predictor variables.
- p: The probability that the outcome y equals 1.

Breakdown of the Components:

Probability Notation:
Pr(y=1|X_1,...,X_k): This denotes the conditional probability that the outcome y is 1, given the predictors X_1,…, X_k.
Exponential Function:
$e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}$: This represents the exponential function applied to the linear combination of the intercept and predictor variables. The exponential function ensures that the result is always positive.
Logistic Function:
$\frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}$: This fraction is the logistic function, which maps any real-valued number into the range (0, 1). This is crucial for modeling probabilities, as probabilities must lie within this range.
Outcome Probability:
p: The result of the logistic function, which gives the probability that y is 1, given the predictors.
Interpretation:
Intercept (β_0): This is the expected log-odds of the outcome when all predictor variables are zero.
Coefficients ($\beta_1, \ldots., \beta_k$): These represent the change in the log-odds of the outcome for a one-unit change in the predictor variables. A positive coefficient indicates that as the predictor increases, the log-odds of the outcome (and hence the probability) increase.
Practical Use:
In the context of credit risk analysis on the Bondora platform:
Predictor Variables (X_1,…, X_k): These could include borrower characteristics (age, employment status, income level, credit history), loan attributes (loan amount, loan term, interest rate), and external economic factors (economic growth indicators, inflation rate, market interest rates).

Outcome Variable (y**)**: This is whether the borrower defaults (1) or does not default (0).
By fitting this logistic regression model to the data, we can estimate the probability of default for new loan applicants based on their characteristics, allowing Bondora to make informed lending decisions.

Logistic regression operates under five main assumptions:

- The target variable is categorical.
- The error terms are independent.
- The predictors are uncorrelated.
- There is a linear relationship between the log-odds of the target variable being one and the independent variables.
- The variables are relevant.

One of the main advantages of using logistic regression in credit scoring is its relative accuracy and interpretability. It allows for clear explanations of why a particular classification was made, which is beneficial in credit scoring. Additionally, the coefficients from logistic regression can be transformed into a scorecard format.
However, logistic regression has its limitations. It may not capture complex patterns and relationships as effectively as some other machine learning models, especially when dealing with complex and noisy data. Moreover, logistic regression requires extensive data preprocessing, as it cannot handle missing data, and non-linear effects of some variables need to be transformed.

Decision Trees

A decision tree is a machine learning algorithm designed to classify a target variable by iteratively asking a series of questions derived from the training data (Müller and Guido, 2017). The following illustration provides an example of a decision tree.

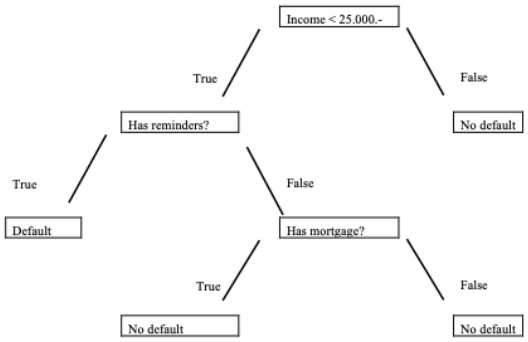

Figure 3 : Example of Decision Tree

The construction of a decision tree begins with the algorithm examining different features to identify the best test, which, in machine learning terms, is the question that best separates the data (Müller and Guido, 2017). For instance, a test might involve splitting the dataset based on whether age > 30 years. The algorithm selects the best test, which becomes the top node of the tree. It then continues to separate the data by posing additional questions until each leaf node contains only one class. If there are no restrictions on the depth of the tree, it might grow very complex. Constraints on the number of questions can result in some leaves not being pure, containing both positive and negative outcomes. When predicting borrower behavior, the predicted outcome is determined by the majority class within the leaf where the borrower falls.

Attribute Selection

Different decision tree algorithms use various methods to choose the attribute for splitting a node. The most common methods include:

Entropy

Entropy measures the level of disorder or randomness in a system. In decision trees, it helps determine how homogeneous a node is after a split.
Lower entropy indicates a purer node.
The formula of Entropy is :

$$E(S)=-p_{(+)}\log p_{(+)}-p_{(-)}\log p_{(-)}$$

where:
E(S): Entropy of the set S
p(+): Probability of the positive class in the subset S
p(−): Probability of the negative class in the subset S
log: Logarithm base 2 (commonly used in information theory)

If the dataset is completely homogeneous (all instances belong to the same class), the entropy is 0, indicating no disorder.
If the dataset is evenly split between the positive and negative classes, the entropy is 1, indicating maximum disorder. By calculating the entropy before and after a split, we can measure how much uncertainty is reduced, which helps in selecting the best attribute to split on. For example, if a node contains only one class, its entropy is zero, indicating no randomness or disorder.

Information Gain

Information gain is calculated by subtracting the entropy of an attribute from the total entropy. The goal is to maximize this difference.
Formula:

$$\text{Information Gain}=\text{Entropy(S)}-\sum\left(\frac{|S_v|}{|S|}\times Entropy(S_v)\right)$$

Where:

- Entropy(S): Entropy of the original dataset S.
- |S|: Total number of instances in the original dataset S.
- S_v: Subset of S created by splitting on a particular attribute v.
- $|S_v|$: Number of instances in the subset S_v.
- Entropy(S_v) Entropy of the subset S_v.

Information Gain measures the reduction in entropy achieved by the split. A higher information gain indicates a more effective split, leading to purer child nodes. This method is widely used in decision trees, such as the ID3 algorithm.

Gain Impurity

Gini Impurity measures the likelihood of a random sample being incorrectly classified if it were randomly labeled according to the distribution of labels in the dataset. It ranges from 0 to 0.5, with lower values being preferable.

Formula:

$$\text{Gini Impurity: } 1 - \sum(p_i)^2$$

Where:

- p_i is the probability of class i.

Gini Impurity is used in algorithms like CART (Classification and Regression Trees). A node with lower Gini Impurity indicates a purer node. It is computationally simpler than entropy and is preferred in some implementations due to its efficiency.

Chi-Square

The Chi-Square test assesses the relationship between two categorical variables. A high Chi-Square value indicates that the attribute significantly affects the target variable.
Formula:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- Oi is the observed frequency of class ii.
- Ei is the expected frequency of class ii.

Chi-Square is used to test the independence of two variables. In the context of decision trees, it helps determine whether an attribute is a good candidate for splitting by checking if there is a significant difference between observed and expected frequencies.

ANOVA

ANOVA is used in regression problems to compare the variances within groups to those between groups. A higher F-value suggests a better split.
ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. In decision trees, ANOVA helps to identify attributes that have the greatest effect on the target variable.

Advantages of Decision Trees

Interpretability:
Decision trees are highly intuitive, easy to understand, and interpret. The rules implemented by decision trees can be displayed in a flowchart-like manner, making it simple to explain model predictions to stakeholders. This transparency boosts stakeholder confidence and provides detailed insights into the decision-making process.
Less Data Preparation:
Decision trees require minimal data preprocessing. Unlike other algorithms, decision trees do not need data normalization, standardization, or extensive handling of missing values, making them a 'go-to' algorithm for many data scientists.
Non-Parametric:
Decision trees do not rely on any assumptions about the data distribution. This non-parametric nature allows them to handle a variety of data types and structures effectively.
Versatility:
Decision trees are versatile and can be used for various tasks beyond standard predictions, such as data exploration and quality assessment. They are applicable to both regression and classification problems and can also be adapted for segmentation tasks.

Disadvantages of Decision Tree

Overfitting:
Decision trees can easily overfit the training data due to their high variance, creating overly complex decision rules. To mitigate overfitting, parameters can be tuned to control the splitting process, or pruning can be performed. However, these methods may have limited effectiveness. Setting a maximum depth for the tree can reduce overfitting, increasing test data accuracy at the expense of training data accuracy.

Instability:
Decision trees are sensitive to small changes in the training data, which can drastically alter the tree's structure. This high variance makes them unstable and susceptible to significant changes based on minor data variations (Hastie, Tibshirani, and Friedman, 2009).
Optimization:
The decision tree algorithm is considered greedy because it looks for the purest node at each level without considering future splits. This heuristic approach enhances interpretability but does not guarantee a globally optimal result. Variables with significant influence can dominate the process, leading to potential issues with data leakage. These problems can be mitigated by using an ensemble of decision trees, though this comes at the cost of reduced interpretability.
Bias: Decision trees can be biased if some classes dominate. Techniques such as balancing the dataset or using cost-sensitive learning can mitigate this issue**.**

Theory Behind Scorecard Development

Scorecard

All credit lending institutions, including banks, use sophisticated credit models that incorporate information from loan applications—such as salary, credit commitments, and past loan performance—to determine a credit score for an applicant or an existing customer. This score predicts the likelihood that a lender will be repaid on time if they issue a loan or credit card to the applicant.One widely used type of credit model is the credit scorecard. Its popularity stems from its relative ease of interpretation for customers and its established development process, which has been in use for several decades.
Typically, the target variable in these models is binary, indicating whether a customer is performing (e.g., paying on time) or has defaulted (e.g., more than 90 days late on payment).

Weight of Evidence (WoE) Transformation
To enhance predictive accuracy, all independent variables (e.g., age, income) are transformed using the Weight of Evidence (WoE) method. WoE measures the strength of grouping for differentiating between good and bad risks based on the proportion of good applicants to bad applicants within each group. This method seeks to establish a monotonic relationship between the independent variables and the target variable.
WoE and Information Value (IV) are crucial for feature engineering and selection in credit scoring. WoE assesses the predictive power of an independent variable concerning the target variable, differentiating between good and bad customers. IV ranks features based on their relative importance.

Weight of Evidence (WoE)

The formula to calculate WoE is as follow:

$$WoE = \ln\left(\frac{\%\ of\ good\ customers}{\%\ of\ bad\ customers}\right)$$

- % of good customers: Percentage of clients with a lower risk of default.
- % of bad customers: Percentage of clients with a higher risk of default.
- ln: Natural logarithm.

A positive WoE indicates a higher proportion of good customers compared to bad customers, whereas a negative WoE indicates the opposite. WoE is a powerful tool in predictive modeling, capturing the non-linear relationships between independent variables and the target variable, and is widely used in applications like credit scoring, fraud detection, and customer segmentation.
Steps for WoE Feature Engineering:

1. Calculate WoE for each unique value (bin) of a categorical variable.
2. Bin continuous variables into discrete bins based on distribution and unique observations, using techniques like pd.cut (fine classing).
3. Calculate WoE for each derived bin of the continuous variable.
4. Combine bins following coarse classing rules.

Feature Selection using Information Value - IV

IV helps rank features based on their importance. It's good practice to perform feature selection to determine if all features are necessary for the model, often eliminating weak features for a simpler, more effective model.

- IV < 0.02: Useless for predictions.
- 0.02 ≤ IV < 0.1: Weak predictive power.
- 0.1 ≤ IV < 0.3: Medium predictive power.
- 0.3 ≤ IV < 0.5: Strong predictive power.
- IV ≥ 0.5: Potentially too good to be true, requiring further investigation.

Binning of Variables

Binning involves grouping numerical predictors into categories to simplify and enhance interpretation. This process, performed during exploratory data analysis (EDA), helps create logical separations, especially for continuous variables. For example, age can be divided into several groups, and WoE is calculated for each group to find a trend. Non-monotonic trends require adjusting the bins to establish a logical grouping for further analysis.

Model Fitting & Interpreting Results

Logistic regression is a common technique used for credit scoring. It models the scoring function and estimates client creditworthiness, with the regression coefficients used to scale the scorecard. The scorecard conforms to a specific score range, and the final credit score is a linear function of the predictor variables.
For each attribute, multiplying its WoE by its regression coefficient gives the score points. An applicant's total score is proportional to the logarithm of the predicted bad/good odds.

Reject Inference

Application scorecards may have selection bias if based only on accepted applications with known performance. Rejected customers are excluded due to unknown performance, creating bias. Reject Inference (RI) addresses this by inferring the unknown performance of rejected applicants, ensuring both accepted and rejected populations are included in the modeling process.

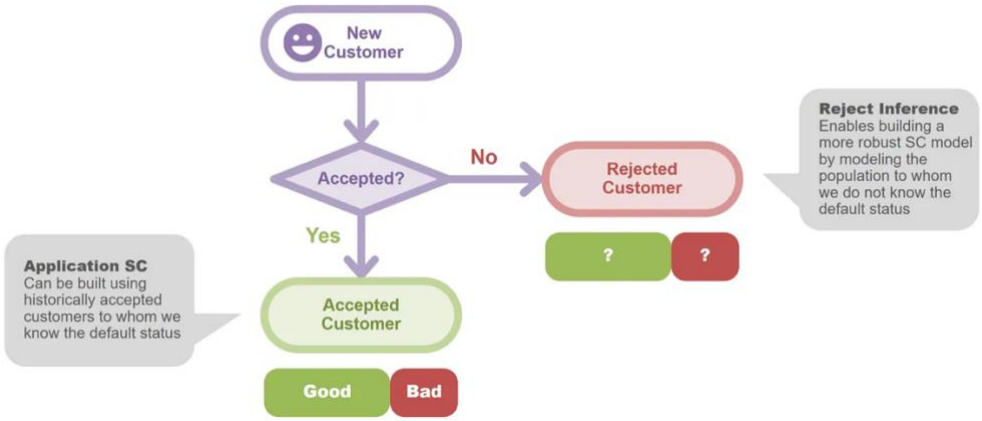

Figure 4 : Accepted and rejected populations

Methodology

Data

Bondora operates a direct marketplace lending platform, adhering to the traditional or "pure" peer-to-peer (P2P) lending model. This platform functions as an online marketplace where borrowers can apply for loans, and investors can fund these loans.
Bondora Capital OU, based in Estonia, manages the platform. The company has been granted a credit license by Estonia's Financial Supervisory Authority (FI). Through Bondora, investors can provide funding to individuals by investing in personal loans. Additionally, individuals seeking financing can apply for funding on the platform.

Dataset

The dataset utilized in this research was retrieved from Bondora.com on November 6, 2022. It contains comprehensive information on 317,059 loan applications, encompassing 112 variables, and spans the period from February 2009 to October 2022. The borrowers represented in the dataset are from four different countries, with the majority being residents of Estonia. Other significant borrower populations come from Finland and Spain, contributing to the dataset's diversity.

To ensure the relevance and quality of the analysis, several variables were excluded from the dataset. These exclusions were based on their lack of relevance to the research objectives and insufficient observations. Variables that did not significantly contribute to the predictive power of the model or had too many missing values were removed to streamline the analysis and enhance model performance.

The dataset is publicly accessible on Bondora's website, which underscores the transparency and openness of Bondora's operations.

Status	Frequency	Percentage
Repaid	122812	38,73%
Current	105723	33,92%
Late	88524	27,92%
Total	317059	100,00%

Table 1: Initial status records of dataset

For this research, the data was specifically filtered to focus on loans that have either been fully repaid or have defaulted. This filtering process resulted in a refined dataset of 211,283 loan applications. Loans that are currently in progress (with a status of "current") were excluded from the analysis. The reason for this exclusion is that these loans have not yet reached maturity, and their final status (repaid or defaulted) is still undetermined, making them unsuitable for a conclusive analysis.

The primary focus of this research is to predict loan defaults. To this end, the default variable, which serves as the target variable in the predictive models, was constructed by combining the status and default date variables. This combined approach ensures a precise and accurate definition of defaults, enhancing the model's ability to differentiate between successful and unsuccessful loan repayments.

The extensive dataset from Bondora provides a robust foundation for analyzing and predicting credit risk in peer-to-peer lending. By focusing on relevant variables and excluding incomplete or irrelevant data, the research aims to develop a reliable and accurate predictive model. This model will help in understanding the factors that influence loan repayment behaviors and defaults, ultimately contributing to better risk management practices in the P2P lending industry.

Status	Frequency	Percentage
Repaid	106861	50,58%
Late	104422	49,42%
Total	211283	100,00%

Table 2: Filtered status records of dataset

For the clustering step of this research the data set consists of 104.422 loans over the period of 02/ 2009 – 10/2022, the dataset includes 38 variables.

Variable Definition

Variable Name	Variable Description	Variable Scale	Variable Role
Age	The age of the borrower when signing the loan application	interval	input
Amount	Amount the borrower received on the Primary Market.	interval	Rejected
AmountOfPreviousLoansBeforeLoan	Value of previous loans	interval	Rejected
AppliedAmount	The amount borrower applied for originally	interval	input
BidsApi	The amount of investment offers made via Api	interval	input
BidsManual	The amount of investment offers made manually	interval	input
BidsPortfolioManager	The amount of investment offers made by Portfolio Managers	interval	input
Country	County of the borrower	Nominal	input
CreditScoreEsMicroL	A score that is specifically designed for risk classifying subprime borrowers	Nominal	input
DebtToIncome	Ratio of borrower's monthly gross income that goes toward paying loans	interval	input
DefaultStatus	Indicator variable, 1 if loan was defaulted, and 0 otherwise.	Binary	Target
Education	1 Primary education 2 Basic education 3 Vocational education 4 Secondary education 5 Higher education	Nominal	input
EmploymentDurationCurrentEmploye	Employment time with the current employer	Nominal	input
EmploymentStatus	1 Unemployed 2 Partially employed 3 Fully employed 4 Self-employed 5 Entrepreneur 6 Retiree	Nominal	input
ExistingLiabilities	Borrower's number of existing liabilities	interval	input
FreeCash	Discretionary income after monthly liabilities	interval	input
Gender	0 Male 1 Woman 2 Undefined	Binary	input
HomeOwnershipType	0 Homeless 1 Owner 2 Living with parents 3 Tenant, pre-furnished property 4 Tenant, unfurnished property 5	Nominal	input
IncomeTotal	Borrower's total income	interval	input
Interest	Maximum interest rate accepted in the loan application	interval	input
InterestAndPenaltyPaymentsMade	Unpaid interest and penalties	interval	input
LanguageCode	1 Estonian 2 English 3 Russian 4 Finnish 5 German 6 Spanish 9 Slovakian	Nominal	Rejected
LiabilitiesTotal	Total monthly liabilities	Interval	input
LoanDuration	Current loan duration in months	Interval	input
MaritalStatus	1 Married 2 Cohabitant 3 Single 4 Divorced 5 Widow	Nominal	input
MonthlyPayment	Estimated amount the borrower has to pay every month	Interval	input
NewCreditCustomer	Prior credit history of customer in Bondora 0 Customer had at least 3 months of credit history in Bondora 1 No prior credit history in Bondora	Binary	input
NoOfPreviousLoansBeforeLoan	Number of previous loans	Interval	input
OccupationArea	1 Other 2 Mining 3 Processing 4 Energy 5 Utilities 6 Construction 7 Retail and wholesale	Nominal	input
PartyId	A unique ID given to the borrower	Nominal	Rejected
PreviousEarlyRepaymentsCountBefo	How many times the borrower had repaid early	Nominal	input
PrincipalBalance	Principal that still needs to be paid by the borrower	Interval	input
PrincipalPaymentsMade	Note owner received loan transfers principal amount	Interval	input
Rating	Bondora Rating issued by the Rating model	Nominal	input
RefinanceLiabilities	The total amount of liabilities after refinancing	Interval	input
Restructured	The original maturity date of the loan has been increased by more than 60 days	Binary	input
UseOfLoan	0 Loan consolidation 1 Real estate 2 Home improvement 3 Business 4 Education 5 Travel 6 Vehicle	Nominal	input
VerificationType	Method used for loan application data verification 0 Not set 1 Income unverified 2 Income unverified	Nominal	input

Table 3: Variable Description

Data Preparation - Exploratory Data Analysis

Data cleansing is a crucial step in machine learning and an essential component of developing a robust machine learning model. It involves ensuring that the dataset is free from irrelevant or incorrect information.

Main Steps of the Data Cleaning Process:

1. **Removing Irrelevant Variables (Columns)**:
   - Eliminate columns that do not contribute to the predictive power of the model.
2. **Handling Null Values**:
   - Remove or replace missing values to ensure algorithms can be implemented without errors. For instance, the variable "Rating" had null values which were replaced with the most frequent value, "F".
3. **Converting Categorical Values to Numerical Values**:
   - Transform categorical data into a numerical format that can be used by machine learning algorithms.

Handling Null Values

To implement machine learning algorithms effectively, it is essential to handle all null values in the dataset. If null values are present, the algorithms may fail to function correctly. Therefore, all rows and columns with null values were either removed or the missing values were replaced with the most frequent value. For example, null values in the "Rating" variable were replaced with the most common value, "F".

Dropping Irrelevant Variables
In the dataset, 74 out of 112 columns were removed because they were not relevant to the analysis. For instance:

- Columns such as "Loan ID", "Loan Number", "UserName", and "DateOfBirth" (since "Age" is already included) were excluded because they are mainly for identification purposes and do not contribute to default prediction.
- Columns related to specific income sources, such as "IncomeFromPrincipalEmployer", "IncomeFromPension", "IncomeFromFamilyAllowance", "IncomeFromSocialWelfare", "IncomeFromLeavePay", "IncomeFromChildSupport", and "IncomeOther", were removed since the "IncomeTotal" variable is already present.
- The "Amount" column was dropped due to its 94% correlation with "AppliedAmount", which is excessively high.

The primary objective of selecting attributes is to prevent overfitting and improve model performance in terms of cost and time efficiency. Additionally, the relevance of each variable's impact on the outcome was considered during the selection process. This ensures that the final dataset is streamlined and optimized for better predictive accuracy.

Variable conversion
Additionally, we checked the data type of all features and concluded to change many variables from numeric to categorical. Some of the variables that the type changed shown below :

- Verification Type
- Language Code
- Gender
- Use of Loan
- Education
- Marital Status
- EmployementStatus
- OccupationArea

Creating Target Variable

Status variable is the variable which help us in creating target variable. The reason for not making Status as target variable is that it has three unique values **current, Late and repaid**. There is no default feature but there is a feature **default date** which tells us when the borrower has defaulted means on which date the borrower defaulted. So, we will be combining **Status** and **Default date** features for creating target variable.The reason we cannot simply treat Late as default because it also has some records in which actual status is Late but the user has never defaulted i.e., default date is null. So we will filter out all the current status records because they are not matured yet they are current loans.

Then, we will create new target variable based on the DefaultDate attribute in which 0 will be assigned when default date is null means borrower has never defaulted while 1 in case default date is present. It means those borrowers having default date belong to defaulted class in our target attribute.

Descriptive Statistics - Understanding Features

Important metrics have had some of their statistical characteristics examined to improve the overall view of the data.
The statistical measures are extracted and summoned in the table below.

	Mean	Min	Max
Age	39.5	18	77
Amount	2.549	6.39	15.948
Applied Amount	2.673	10	15.948
Income Total	1.982	0	1.012.019
Interest	32	2	264
Liabilities Total	465	0	12.400.000
Loan Duration	47	1	120
Debt to Income	4.8	0	198
Principal Balance	959	-472	12.557
Monthly Payment	111	0	2.368

Table 4: Statistical Report of variables

After removing observations that have missing values for one of the critical variables for this research 211.283 observations remain.

From the frequency diagram of observations of the dependent qualitative variable “DefaultStatus”, Status ,it can be seen that almost half of the Bondora platform loans are in default. Specifically, 49.4% of the loans are in default and the remaining 50.6 % are fully repaid.

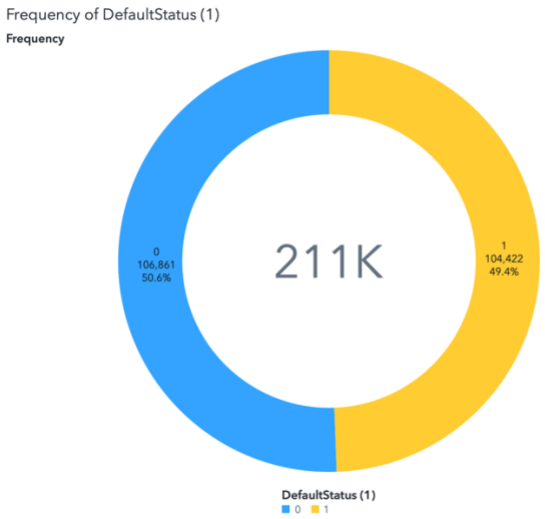

Figure 5: Distribution of DefaultStatus

Dataset included observations of men, women or undefined borrowers. As about the distribution of borrowers 59.2% were male, 34.2% percent female and 6.7% undefined.

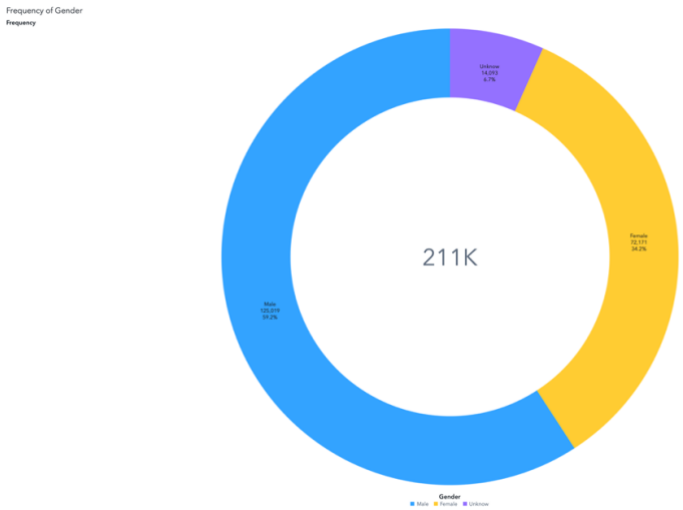

Figure 6: Gender Distribution of Borrowers

Regarding the residency of the borrowers, the majority, over 52.4%, are residents of Estonia. Approximately 13% of the borrowers reside in Spain, while 34% are from Finland. A very small proportion, less than 1%, fall into the "Other" category, representing residents of countries outside the main groups.

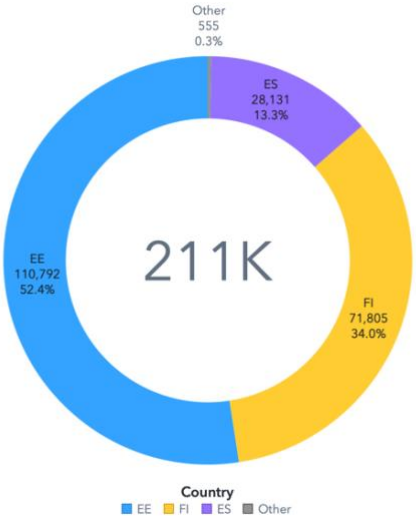

Figure 7: Residency of the borrowers

The figure below illustrates the purposes for which loans were utilized by applicants. The majority of applicants, accounting for 83.1%, did not specify the purpose of their loan, categorizing it as "Not Set". Conversely, 4.5% of borrowers either used the loan for unspecified purposes or did not disclose the loan purpose in their applications. Approximately 4.2% of the loans were designated for home improvement.
The remaining 8.2% of loans were used for various other reasons. Specifically, a smaller proportion of loans were allocated for purposes such as loan consolidation, real estate, vehicle purchase, business investments, travel, education, and health.

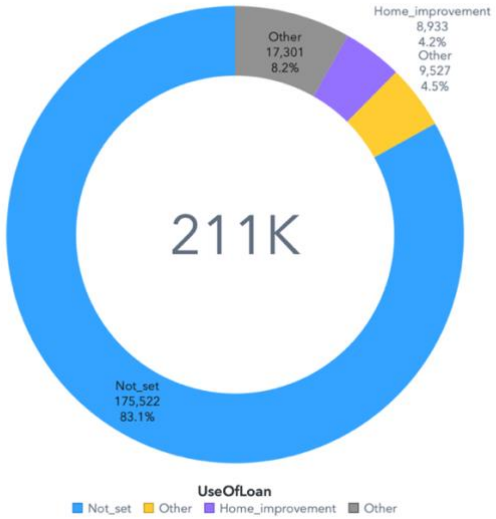

Figure 8: Use of loan

The age of borrowers ranges from 18 to 70 years. The age distribution is somewhat skewed, with borrowers over 20 years old taking out more loans compared to those under 20. As age increases, the number of loans decreases. Most borrowers fall within the age range of 24 to 38 years, with an average age of nearly 38 years.

For further analysis, the ages were grouped into the following categories:

- Young Adults (18-30 years old)
- Middle-Aged Adults (30-45 years old)
- Older Adults (45-70 years old)

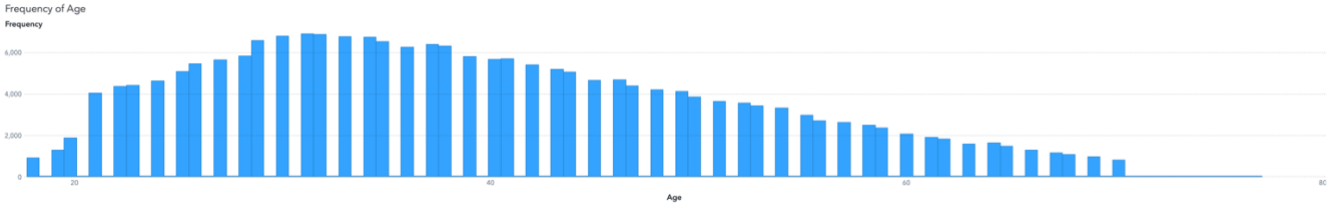

Figure 9: Distribution of Borrowers Age

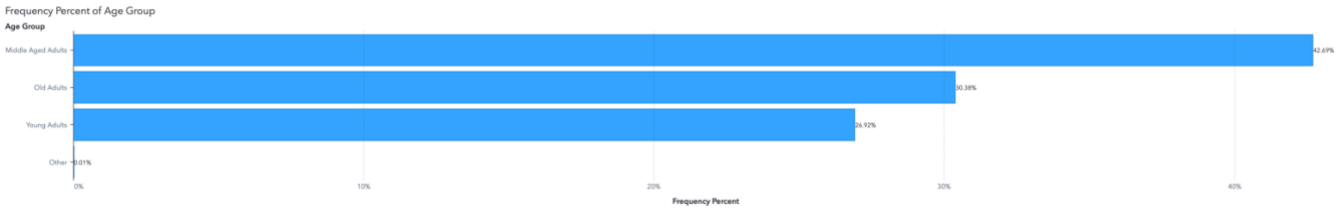

Figure 10: Distribution of age groups

The majority of borrowers belong to the Middle-Aged Adults group, as illustrated in the diagram below.

Borrowers with Secondary education have the most observations, then those with Vocational education. The least defaulted borrowers are with the education Not Present, Basic & Primary.

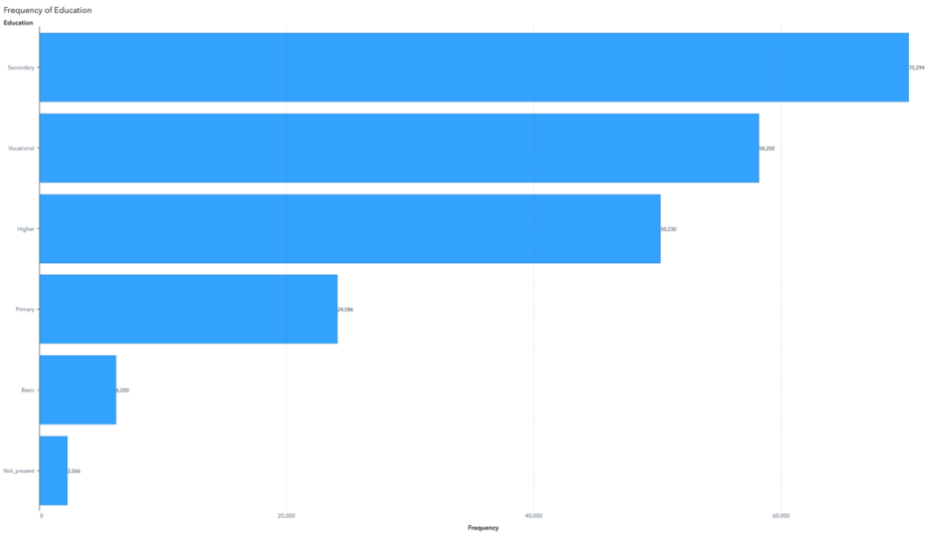

Figure 11:Education of Borrowers

The borrowers in the dataset are divided into eight credit groups: AA, A, B, C, D, E, F, and HR, where the risk level increases from AA to HR. The majority of borrowers fall into the risk category C, with 51,959 borrowers. The number of borrowers in risk categories E and F are 32,953 and 28,428, respectively.
While categories A and AA represent the lowest risk, the high number of default loans within these categories indicates that risk classification alone cannot be solely relied upon when investing in loans. This suggests that additional factors should be considered to make informed investment decisions.

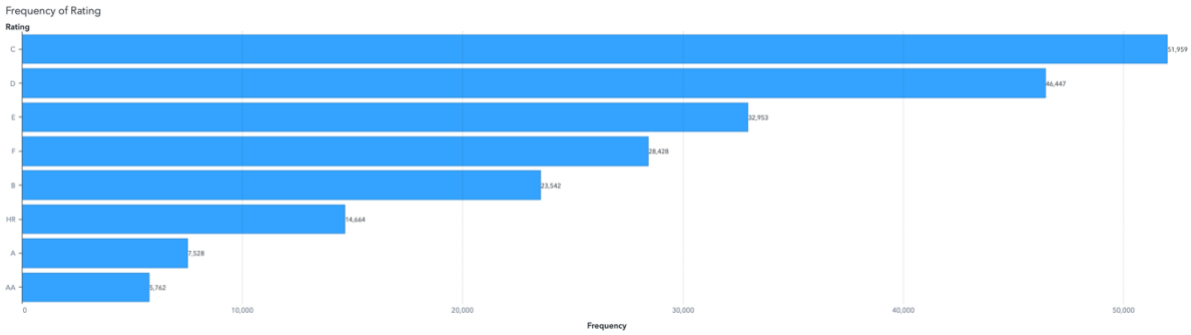

Figure 12: Risk Categories of Borrowers

The maximum amount of loan applied by the borrowers is 15.948 Euros, while the average amount applied is 2.125 Euros. A borrower, in average, is funded 2.125 euros, the lowest funded amount is 6.39 Euros and the highest is 15.948 Euros. The duration of the loans ranges from 1 month to 120 months.

	Minimum	Maximum	Average
Applied Amount	**10**	**15.948**	**2.125**
Funded Amount	**6,39**	**15.948**	**2.125**
Duration	**1**	**120**	**60**

Table 5: Loan Description of Borrowers

Most of the borrowers are between 24 and 44-year-old with an average age of almost 40 years old.

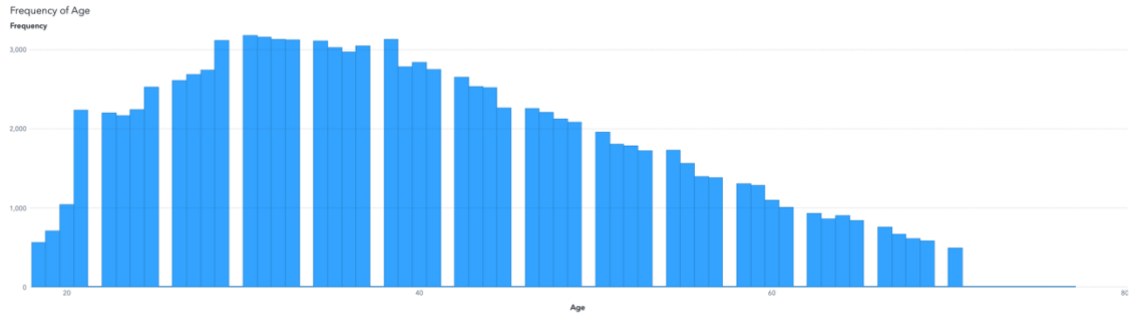

Figure 13: Age of Defaulters

The distribution of borrowers of only default loans is 59.9% were male, 29.2% percent female and 10.4% undefined.

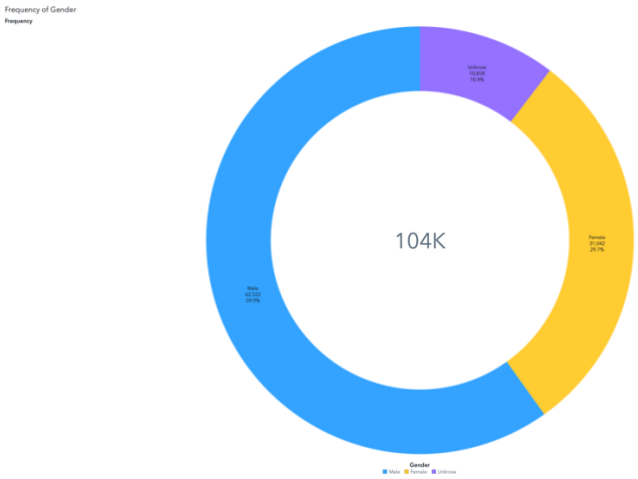

Figure 14: Gender of Defaulters

We can observe that borrowers having no clear Marital Status are defaulted the most, then those with single marital status.On the other hand, the least observations are with marital status widow.

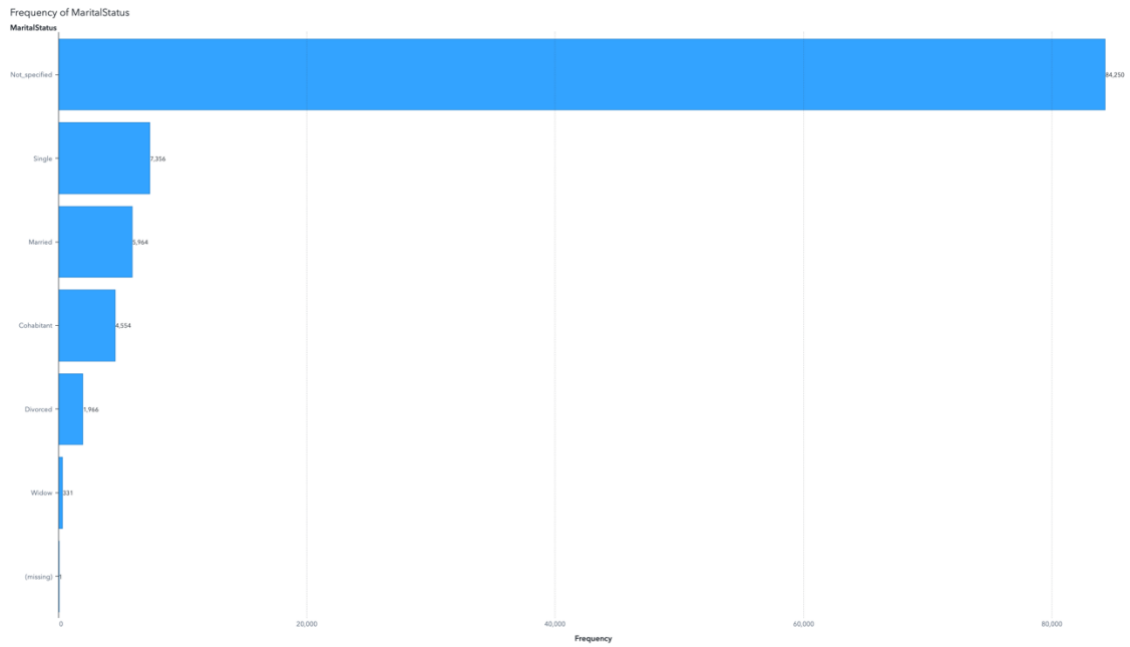

Figure 15: Marital Status of Defaulters

Collection Process

The collections process necessitates significant and ongoing interaction with the client. This interaction begins with a meticulous analysis of the client's financial situation and continues through consistent, timely, and frequent contact for the duration of the loan. Effective collections require offering clients payment alternatives that are appropriate and timely for each specific situation. Additionally, all collections activities should be meticulously recorded to facilitate continuous monitoring, follow-up, and control of client compliance with the negotiated agreements. The following steps provide a detailed outline of typical collections activities, followed by a flowchart illustrating the collections process:

1. **Case Analysis**:
Conduct an in-depth analysis to understand who the client is, what their financial situation entails, the original conditions of the loan, and why the loan has become past-due. This step involves considering internal data as well as external sources such as credit bureaus and bad-debtor lists to gain a comprehensive understanding of the client's financial behavior and history.
2. **Client Contact**:
Initiate contact with the client to gather more information, understand their current circumstances, and verify their location. Review any previous actions taken to ensure all communications and efforts are well-coordinated and informed.
3. **Assessment**:
Identify the core problem causing the current delinquency. Determine the type of client being dealt with, whether they are generally reliable but currently facing difficulties, or if they have a history of financial mismanagement. This assessment helps tailor the subsequent actions and solutions offered.
4. **Suggesting Alternatives**:
Develop and propose potential solutions to address the delinquency. The goal is to communicate the benefits of timely payments clearly and effectively, fostering a positive payment culture and encouraging the client to commit to a feasible repayment plan. This might include restructuring the loan or offering temporary relief measures.
5. **Securing Payment Commitments**:
Negotiate effectively to ensure clear and concrete payment commitments from the client. This involves specifying when, where, how, and how much the client will pay. It is essential to understand the client's prioritization of bills, especially in cases of over-indebtedness or reduced income, to ensure that loan repayment is given adequate importance.
6. **Compliance with Payment Commitments**:
Monitor the client's adherence to the agreed-upon payment schedule. The objective is to demonstrate consistency and follow-through throughout the collections process. Collections staff must not rely solely on the client's apparent goodwill and positive attitude but must actively ensure that payment commitments are being met.
7. **Recording Collections Activities**:
Ensure all collections activities are documented in a detailed and coordinated manner. This documentation is crucial for maintaining continuity and coherence in the collections process, enabling any staff member who takes over the case to have a complete understanding of all actions taken and communications made.
8. **Follow-up on the Case**:
Continuously stay informed about the client's situation and the specific collections activities applied to the case. Regular follow-ups ensure that the collections strategy remains relevant and effective in light of any changes in the client's circumstances.
9. **Intensification of Collections Activities**:
As a last resort, escalate collections activities to secure the loan repayment as promptly as possible. This may involve assessing the client's assets and considering legal action if necessary. At this stage, the primary objective is to recover the outstanding loan amount, even if it results in losing the client.

This structured and detailed approach to the collections process ensures that every step is systematically executed, fostering a professional and supportive interaction with the client while focusing on recovering the outstanding loan amount. The comprehensive recording of activities and regular follow-ups contribute to a consistent and effective collections strategy, ultimately enhancing the likelihood of successful loan recovery.
This expanded explanation adds more details and nuances to each step of the collections process, providing a thorough understanding of the various aspects involved in managing and recovering past-due loans.

Correlation

During the preliminary analysis, a thorough correlation test was conducted to evaluate the relationships between the numeric variables in the dataset. This test was essential to identify any pairs of variables that exhibit a high degree of correlation, as this could indicate redundancy and potentially lead to multicollinearity issues in the predictive model. By calculating the correlation coefficients, we aimed to determine how closely related the numeric variables are to each other.Understanding these relationships is crucial because variables that are too similar can have a comparable effect on the outcome variable, thereby diminishing the model's accuracy and interpretability. High correlation between predictor variables means they contribute overlapping information, which can complicate the model and reduce its effectiveness.The calculated correlations, specifically contrasted with the DefaultStatus variable, are presented below. This analysis helps in identifying which variables have significant relationships with the DefaultStatus, guiding further steps in feature selection and model refinement. By focusing on variables with unique and significant contributions to predicting loan defaults, we aim to enhance the robustness and predictive power of the model.

This detailed correlation analysis is a foundational step in ensuring that the predictive model is both efficient and effective, reducing redundancy and focusing on the most impactful variables.

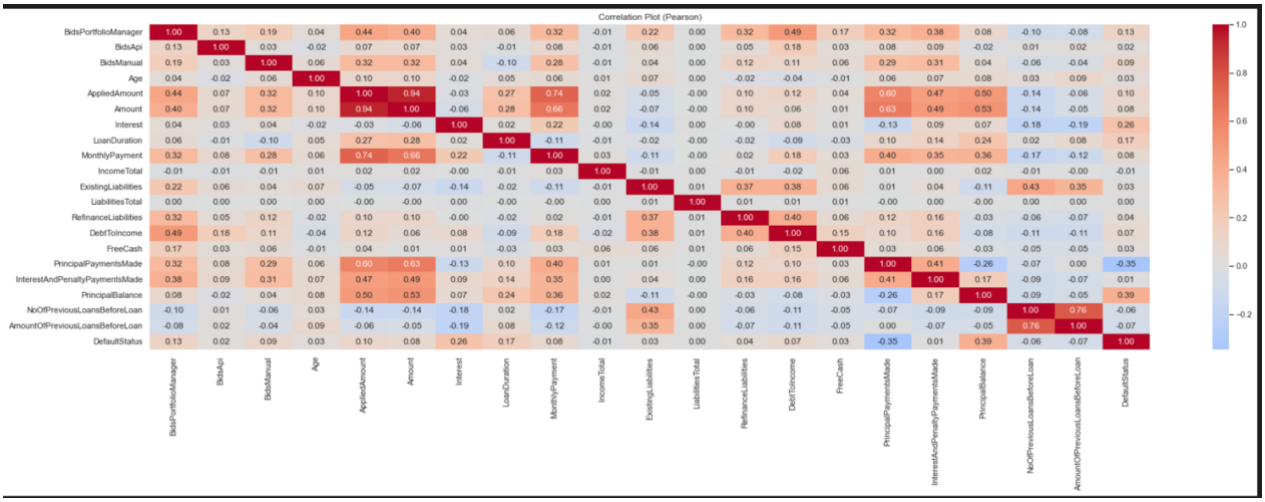

Table 6: Correlation of the numeric variables

We took a sample of variables and checked their correlation with variable DefaultStatus.

	Age	Applied Amount	Amount	Interest	Loan Duration	Income Total	Principal Balance	Debt to Income	DefaultStatus
Age	100%	10%	10%	-2%	5%	1%	8%	-4%	3%
Applied Amount	10%	100%	94%	-3%	27%	2%	50%	12%	10%
Amount	10%	94%	100%	-6%	28%	2%	53%	6%	8%
Interest	-2%	-3%	-6%	100%	2%	0%	7%	8%	26%

Loan Duration	5%	27%	28%	2%	100%	-1%	24%	-9%	17%
Income Total	1%	2%	2%	0%	-1%	100%	2%	-2%	-1%
Principal Balance	8%	50%	53%	7%	24%	2%	100%	-8%	39%
Debt to Income	-4%	12%	6%	8%	-9%	-2%	-8%	100%	7%
Default Status	3%	**10%**	8%	**26%**	**17%**	-1%	39%	7%	100%

Table 7: Correlation of Selected variables

Variable	Value
DefaultStatus	1.000000
PrincipalBalance	0.387825
Interest	0.258632
LoanDuration	0.217185
BidsPortfolioManager	0.126639
AppliedAmount	0.092439
BidsManual	0.091419
Amount	0.079062
MonthlyPayment	0.078794
DebtToIncome	0.072576
RefinanceLiabilities	0.036566
Age	0.038004
FreeCash	0.028267
ExistingLiabilities	0.026021
BidsApi	0.016947
InterestAndPenaltyPaymentsMade	0.006156
LiabilitiesTotal	0.002768
IncomeTotal	-0.014341
NoOfPreviousLoansBeforeLoan	-0.054557
AmountOfPreviousLoansBeforeLoan	-0.069758
PrincipalPaymentsMade	-0.346348

Table 8: Correlation in contrast to variable default

By checking the correlation of variables with the DefaultStatus, it was observed that the variables **interest rate ,loan duration and AppliedAmount** had the most correlation with the Default Status applicants. The variable PrincipalBalance although it has high correlation it is not clearly understood its effect and how it can be measured.
Our finding is that the critical values are AppliedAmount, interest rate and loan duration.
When variables are correlated with the target variable (in this case, DefaultStatus) in a loan dataset, it means there is a statistical association between these variables and the likelihood of a loan default. Here's what this association might imply for different types of variables:

Positive Correlation

Variables with a positive correlation to DefaultStatus: As the value of these variables increases, the likelihood of a loan default also increases. Example: If DebtToIncome has a positive correlation with DefaultStatus, it implies that higher debt-to-income ratios are associated with a higher probability of loan defaults.

Negative Correlation

Variables with a negative correlation to DefaultStatus: As the value of these variables increases, the likelihood of a loan default decreases. Example: If IncomeTotal has a negative correlation with DefaultStatus, it suggests that higher total income is associated with a lower probability of loan defaults.

As next step we checked for Outliers the input variables in order to proceed to risk applicants-borrower segmentation.

The upper whisker of AppliedAmount is 8.854.

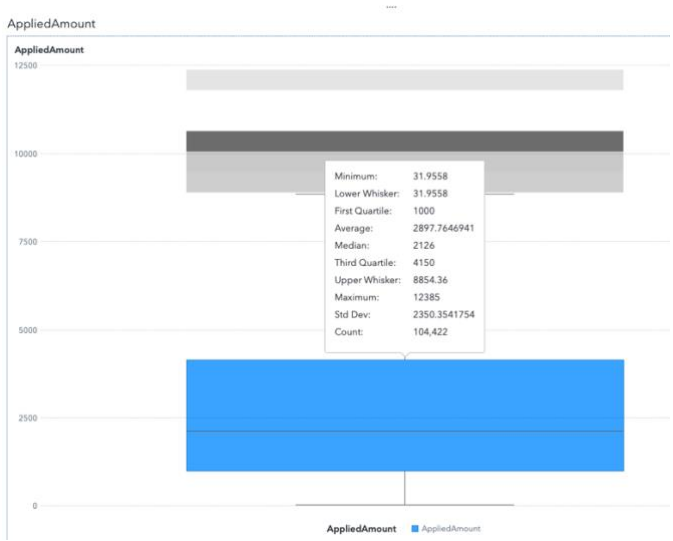

Figure 16: Analysis of AppliedAmount

The upper whisker of LoanDuration is 96.

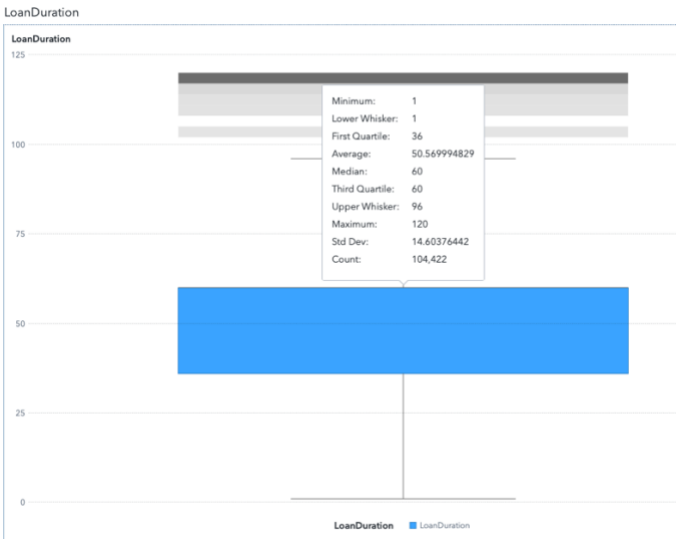

Figure 17:Analysis of LoanDuration

The upper whisker of Interest is 77.94.
After identifying the outliers, we putted limits at the variables that would be used at the clustering.

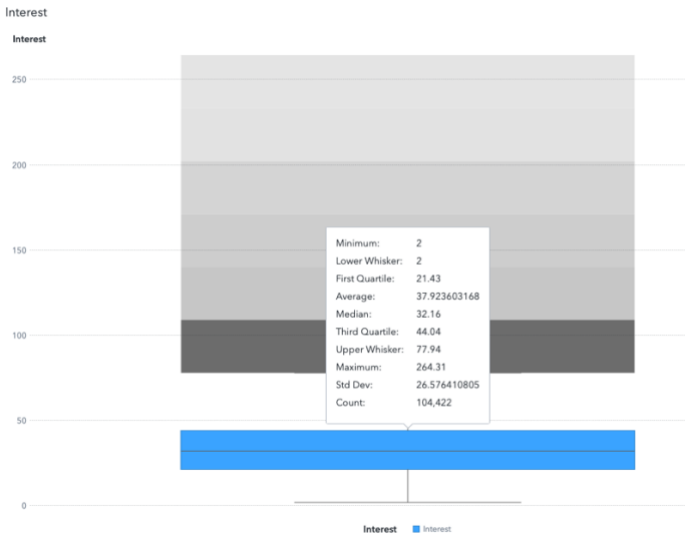

Figure 18: Analysis of Interest

Clustering

As part of our effort to identify key features with high predictive power for the model, we performed variable clustering. This analysis aims to uncover patterns within the population, grouping similar data points together while highlighting differences between these groups. We utilized the K-means clustering method, a widely used approach for such tasks. The objective of K-means clustering is to partition the data into K clusters, where K is predetermined based on the analysis requirements. Each cluster is represented by its mean (or centroid). This method allows us to define the desired number of clusters in advance, facilitating a structured approach to identifying and analyzing distinct patterns within the data.

By clustering variables, we aim to enhance the model by focusing on features that exhibit significant predictive capabilities, ensuring a more robust and accurate predictive model.

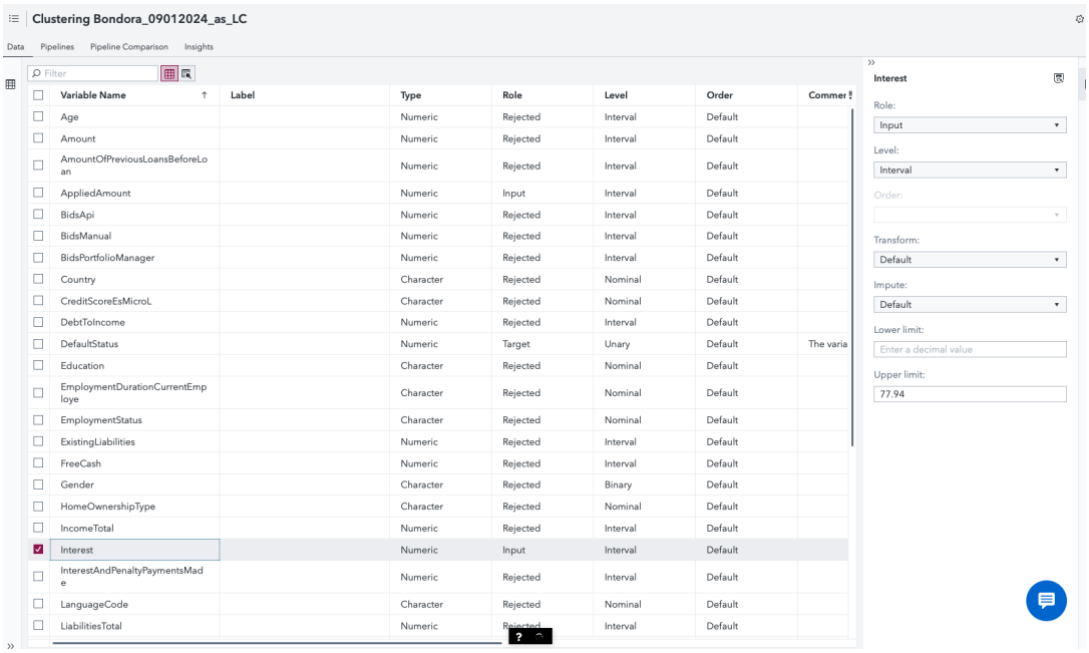

Figure 19: Key input variables for clustering

For the clustering we used 4 variables, the variable DefaultStatus as a Target and as input the variables Interest, Applied Amount & LoanDuration.After running the cluster analysis, 4 clusters are derived based on SAS's gap statistic.

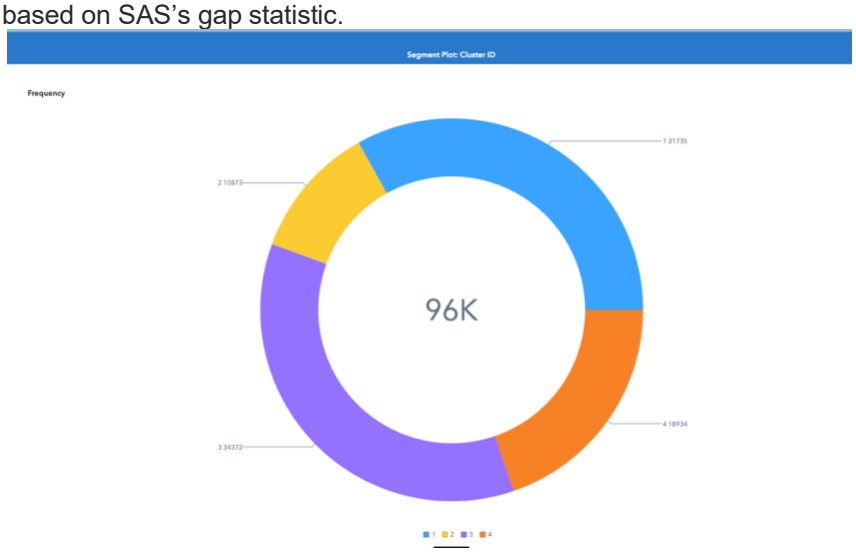

Figure 20:Cluster ID

The size of each cluster is 19.7%(orange), 35.8% (purple),11.3%(yellow) and 33.1% (blue) respectively.

Customer Segmentation

As part of segmentation process we divided the borrowers into smaller, more specific groups based on the input variables as was mentioned at the clustering step (Interest , applied amount & loan duration).

Cluster ID	Cluster Segment	Frequency	AppliedAmount	Interest_Average	LoanDuration_Average
3	**Challenging Debtors**	34,372	€4,317.54	2,660%	56.1
4	**High Value Debtors**	18,934	€833.59	2,629%	53.6
1	**Financially Challenged Debtors**	31,735	€2,471.97	5,163%	50.1
2	**High Potential Debtors**	10,873	€1,134.09	2,122%	26.9
Total		**95,914**	**€2,658.27**	**3,421%**	**50.3**

Table 9: Segments of Clusters

At the above image you can see the average of the input variables and the clusters that were created and categorized.

Financially Challenged Debtor:
These customers may have lower applied amounts, higher interest rates, and shorter loan durations compared to other clusters.
They may have difficulty meeting financial obligations, leading to smaller loan amounts and potentially higher interest rates due to perceived credit risk.
Loan durations may be shorter due to the need for quicker repayment or inability to secure longer-term financing.

High Value Debtor:
Customers in this cluster likely have higher applied amounts, lower interest rates, and longer loan durations.
They may be considered low-risk borrowers with higher credit scores or strong financial backgrounds, allowing them to qualify for larger loan amounts at lower interest rates.
Longer loan durations may indicate a lower urgency for repayment or the ability to secure longer-term financing at favorable terms.

Challenging Debtor:
These customers may exhibit a mix of characteristics, with moderate to high applied amounts, moderate interest rates, and varying loan durations.
They may have credit profiles that pose some risk to lenders, resulting in moderate loan amounts and interest rates.
Loan durations could vary based on individual financial circumstances and risk profiles.

High Potential Debtor:
Customers in this cluster might have moderate to high applied amounts, competitive interest rates, and medium durations.
They may represent opportunities for growth or expansion in lending, with potential for larger loan amounts and favorable terms.
Longer loan durations may indicate a strategic approach to supporting customers with growth potential while mitigating risk.
Overall, the behavior of each cluster is influenced by a combination of factors such as credit risk, financial stability, borrowing capacity, and repayment ability. Understanding these

differences can help lenders tailor their products and services to meet the diverse needs of customers across different segments. Additionally, ongoing monitoring and analysis of customer behavior within each cluster can inform decision-making and risk management strategies.

To **differentiate between the "High Potential Debtor" and "Challenging Debtor"** clusters based on the input variables of applied amount, interest rate, and loan duration, you can analyze the characteristics and behaviors associated with each cluster. Here's how you can approach it:
Applied Amount:

High Potential Debtors: Likely to have moderate to high applied amounts, indicating a need for substantial financing to support growth or investment opportunities.
Challenging Debtors: May have varying applied amounts, possibly moderate, reflecting different financial circumstances and borrowing needs.

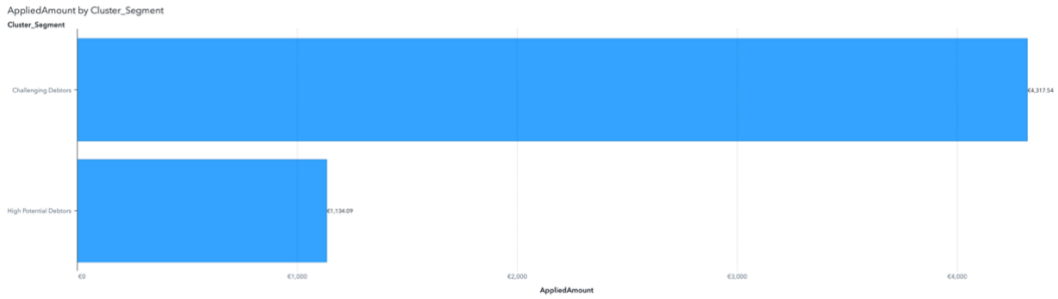

Figure 21:Analysis of AppliedAmount per Cluster Segment

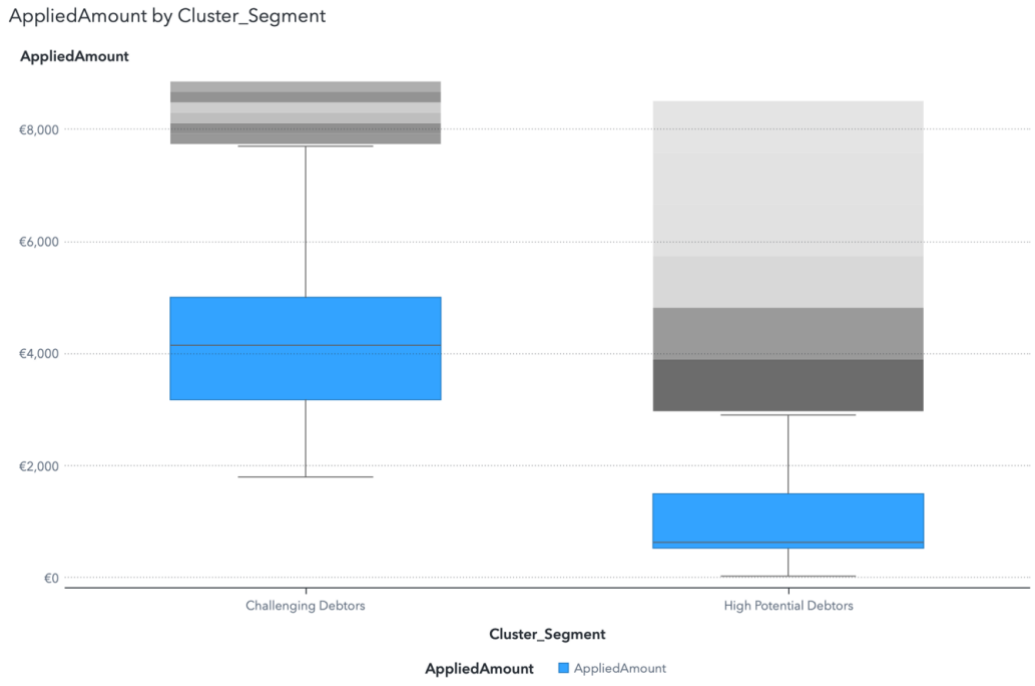

Figure 22: Technical Analysis of AppliedAmount per Cluster Segment

Interest Rate:
High Potential Debtors: Expected to have competitive or favorable interest rates, reflecting their lower perceived credit risk and potential for future growth or profitability.
Challenging Debtors: Might have moderate interest rates, reflecting their moderate risk profile and possibly some credit challenges.

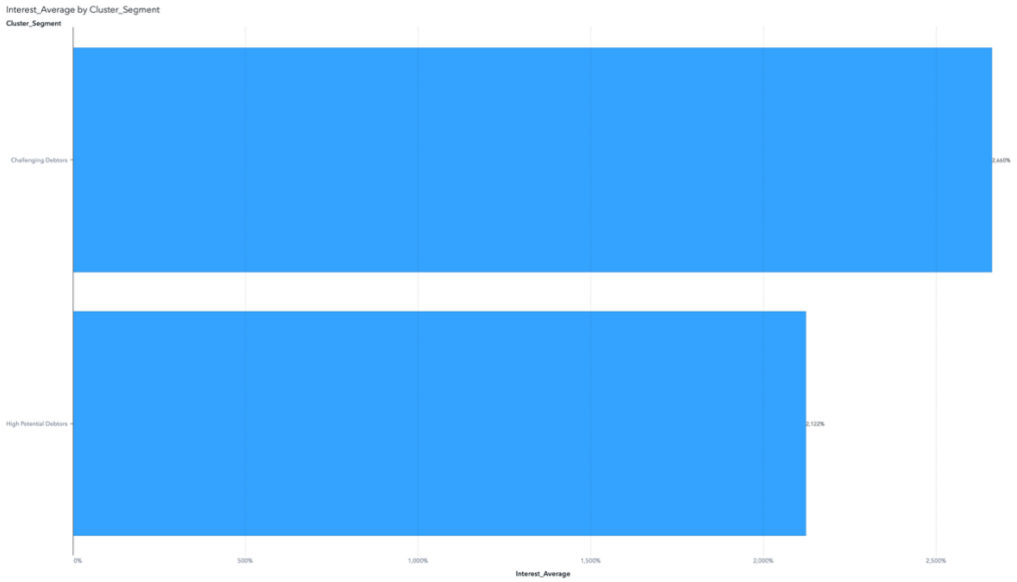

Figure 23: Analysis of Average Interest per Cluster Segment

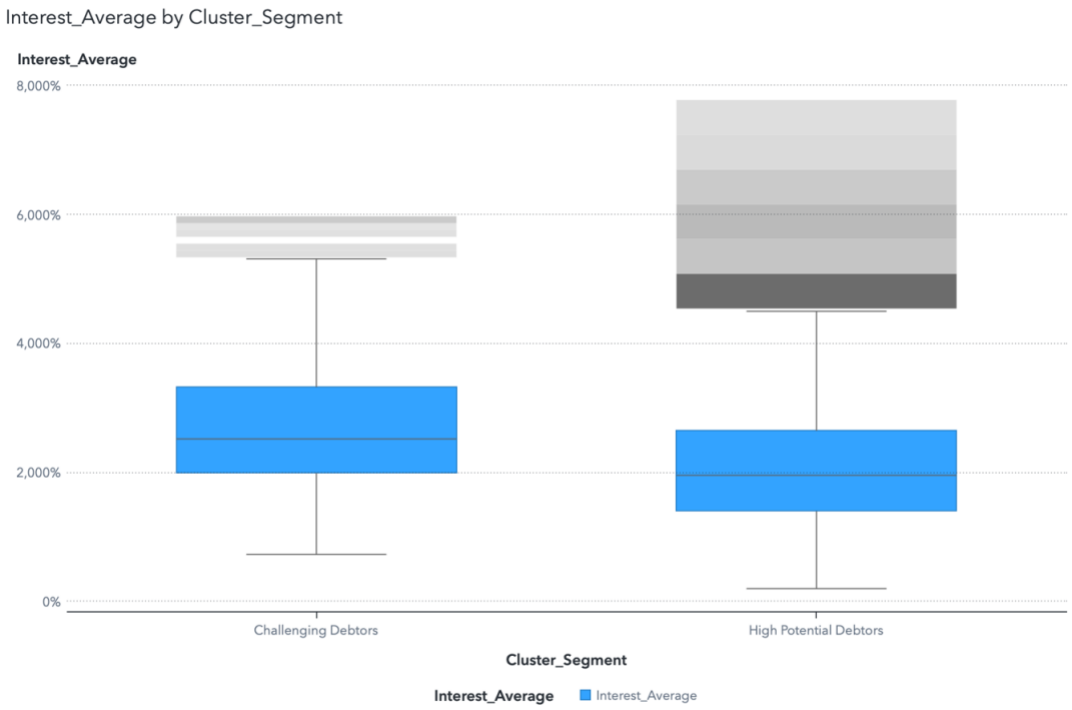

Figure 24: Technical Analysis of Average Interest per Cluster Segment

Loan Duration:
High Potential Debtors: Likely to have medium to long loan durations, suggesting a strategic approach to financing and investment with longer-term objectives.
Challenging Debtors: May have varying loan durations, reflecting different financial situations and repayment capabilities, with durations ranging from short to medium-term.

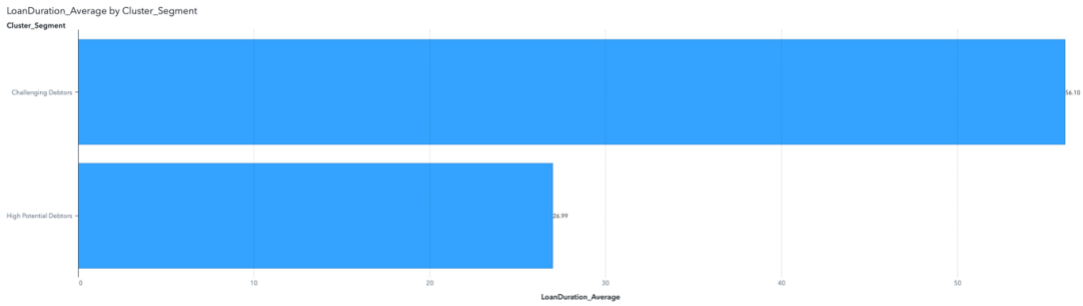

Figure 25: Analysis of Average Loan Duration per Cluster Segment

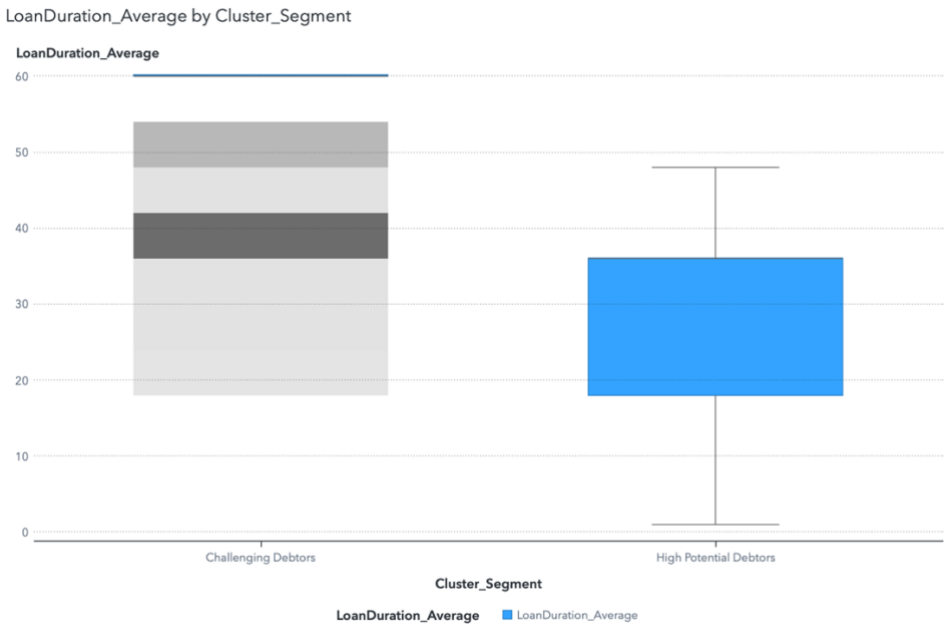

Figure 26: Technical Analysis of Average Loan Duration per Cluster Segment

Credit Risk Profile:
High Potential Debtors: Expected to have relatively lower credit risk profiles, potentially with higher credit scores, stable incomes, and stronger financial backgrounds.
Challenging Debtors: May exhibit moderate credit risk profiles, with varying credit scores, income levels, and financial stability, indicating some risk to lenders.
Financial Stability and Growth Potential:
High Potential Debtors: Likely to demonstrate financial stability and growth potential, with opportunities for expansion, investment, or revenue generation.
Challenging Debtors: May have more limited financial stability and growth prospects, facing challenges or constraints that could impact their ability to borrow or repay debt.

Several factors can influence the behavior of clusters of debtors, including:

- Credit Risk Profiles

Customers with different credit risk profiles may behave differently. Factors such as credit history, credit scores, and debt-to-income ratios can influence a debtor's behavior in terms of borrowing habits, repayment patterns, and willingness to take on additional debt.

- Financial Stability

Debtors' financial stability, including income levels, employment status, and savings, can impact their behavior. Those with higher incomes and more stable employment may exhibit different borrowing and repayment behaviors compared to those with lower incomes or less stable employment.

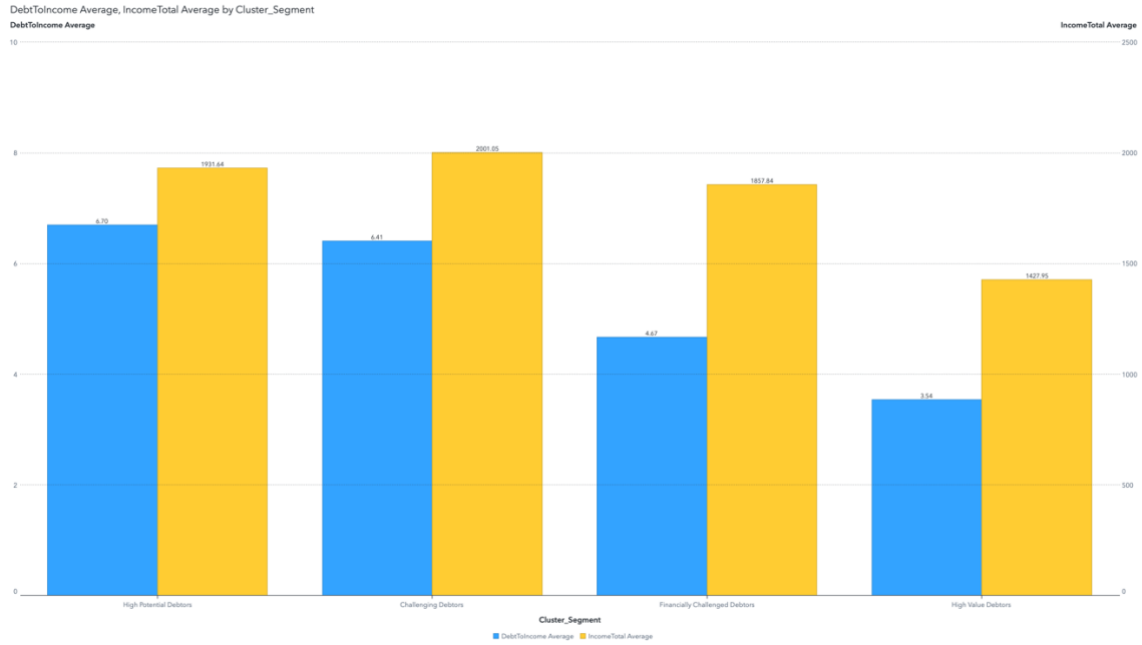

Figure 27:Analysis of variables per cluster

At the above diagram, the Life stage and demographics, such as age, marital status, and household composition, can influence debtors' behavior. For example, younger debtors may have different financial priorities and risk tolerances compared to older debtors

Market Conditions:
Economic conditions, interest rates, and market trends can influence debtors' behavior. Changes in interest rates, inflation, or unemployment rates can impact borrowing costs, repayment abilities, and overall financial decisions.

Importance of AI Interpretability for Credit Risk

Artificial Intelligence (AI) is increasingly being employed in credit risk assessment due to its superior predictive capabilities compared to traditional methods. However, this shift introduces challenges, particularly regarding the interpretability of AI models. The complexity of these models can obscure the rationale behind their predictions, raising concerns about transparency, bias, and regulatory compliance. Given the significant impact of credit decisions on individuals and businesses, model interpretability is crucial. A prediction model must not only be accurate but also interpretable to ensure that stakeholders, including applicants and regulatory bodies, understand and trust the decisions being made.

In our thesis, we chose to utilize decision trees and logistic regression (scorecard) models instead of more complex methods like neural networks, support vector machines (SVMs), and other black-box algorithms. This decision was guided by the necessity of maintaining model interpretability in the context of credit scoring, where applicants have the right to understand why they were accepted or rejected for a loan. Deep learning and other sophisticated AI algorithms can often lack transparency, making their decision-making processes difficult to interpret. These models typically provide an output without revealing the underlying rationale for their decisions. Explainable AI (XAI) aims to address this issue by making AI's actions understandable and analyzable by humans, presenting insights in various formats such as visualizations, mathematical equations, or natural language explanations. There are two main approaches to explainable AI: one focuses on model interpretability during the modeling process, and the other emphasizes post hoc analysis. The first approach involves building models that inherently provide insight into the relationships they have learned. However, this often results in simpler models with potentially lower accuracy. In contrast, the post hoc analysis approach prioritizes achieving the best possible accuracy and then modifying the model to enhance interpretability.

SHAP (SHapley Additive exPlanations) is a technique used to measure the contribution of each variable to the predicted outcome compared to the average prediction. These contributions are known as Shapley values. The paper "Interpretable Machine Learning for Imbalanced Credit Scoring Datasets" explores various black-box machine learning methods used for credit scoring, particularly in the context of imbalanced datasets. These methods are termed "black-box" due to their complexity and the difficulty in understanding their internal workings and decision-making processes. While they offer high predictive accuracy, their lack of transparency poses challenges, especially in regulated industries like finance. The paper suggests using logistic regression (a simpler, interpretable model) as a baseline to establish theoretical results and applying SHAP or LIME to interpret logistic regression in the context of class imbalance. This approach aims to understand how interpretability methods (SHAP or LIME) perform with simpler models, evaluate their reliability despite class imbalance, and compare these findings to more complex black-box models.

LIME (Local Interpretable Model-agnostic Explanations) and SHAP both provide local explanations for feature contributions. SHAP, with its foundation in game theory, ensures fair distribution of predictions among features and offers global model interpretations consistent with local explanations. SHAP provides contrastive explanations by comparing predictions with average predictions, which LIME cannot do. SHAP is more suitable for applications requiring comprehensive local and global interpretations based on a solid theoretical foundation. While LIME lacks SHAP's theoretical foundation and can be less stable, it is time-efficient and offers human-friendly explanations. Tree SHAP is relatively faster for tree-based models, but other SHAP algorithms like Kernel SHAP can be slow and impractical for large datasets. LIME is more appropriate when time is limited, or access to data is restricted.

The paper "Why Should I Trust You? Explaining the Predictions of Any Classifier" by Marco Tulio Ribeiro et al. addresses the challenge of interpreting predictions made by complex black-box models compared to more interpretable models like decision trees and logistic regression. The authors introduce the LIME framework, which generates local surrogate models around individual predictions to approximate the black-box model's behavior locally. This approach aims to bridge the interpretability gap, providing insights into why specific predictions were made.

Laurent Dupont and Olivier Cliche's paper "Governance of Artificial Intelligence in Finance" discusses the governance implications of AI in finance, emphasizing interpretability's

role in ensuring responsible and effective deployment of AI technologies. The increasing use of AI in financial institutions promises enhanced efficiency, risk management, and decision-making capabilities, but often involves complex black-box algorithms. The paper highlights the need for transparency and interpretability to meet regulatory standards and stakeholder expectations in the financial sector.

The paper "Using Deep Learning and Explainable AI to Predict and Explain Loan Defaults" explores the application of advanced machine learning techniques, specifically deep learning, in predicting loan defaults. It emphasizes integrating explainable AI methods to enhance model transparency. The authors discuss using deep learning models to capture complex data patterns and the necessity of employing XAI techniques like SHAP and LIME to provide insights into model predictions. This transparency is crucial for regulators, loan officers, and borrowers to understand and trust the model's decisions.

Deep learning models, known for their ability to capture intricate patterns in data, are utilized due to their capability to handle large amounts of data and extract high-level features automatically. However, recognizing the inherent opacity of these models, the paper emphasizes the integration of XAI techniques. Explainable AI methods like SHAP or LIME are employed to provide insights into how the model arrives at its predictions, enhancing transparency and trust in the model’s decisions. The authors highlight the benefits of using deep learning coupled with XAI, including improved predictive accuracy and enhanced model transparency. They also address challenges such as model complexity, interpretability trade-offs, and regulatory compliance, emphasizing the need for effective governance and validation frameworks to ensure the reliability of AI-driven loan prediction systems.

Decision Tree and Interpretation

After preprocessing for this study, 211.283 observations were used and dataset was split into two parts: Training and Validating dataset. Below is available the summary of observations per dataset.

Samples	No. of Dataset	Percentage (%)
Training	147.898	70
Validating	63.385	30
Total	211.283	100

Table 10:Summary of Dataset

We ran multiple decision trees by changing at the splitting options the class target criterion and had selected the Reduced Error at pruning options to find the model with the lowest misclassification rate.
Based on the multiple trials as shown below the most suitable model is the Chi-Square with enabled Bonferroni , as it has a better misclassification rate and it can be explained.
We choose Chi-square model because in contrast to Gini & Chi-Square without Bonferroni it has less variables of Importance & Leaves.

Following this , the chosen model had many leaves (333) and 28 variables of importance. Our goal is to learn which are the variables of importance and what is their meaning.

The Decision Trees, like all other algorithms, have hyperparameters which they have to choose the best ones based on an optimization process. SAS has the Autotuning process that optimizes with genetic algorithms, but it doesn't work at educational version of SAS therefore we do some tests manually with a limited range. Our goal is not to find the lowest misclassification rate but to be interpretable.

Champion	Name	Algorithm	Misclassification Rate
Champion	Dt - chi square with bonferroni	Decision Tree	0.1017
	Neural Network	Neural Network	0.4942
	Dt - chaid with bonferroni	Decision Tree	0.3090
	Dt - chaid without bonferroni	Decision Tree	0.3082
	Logistic Regression	Logistic Regression	0.1799
	Dt - information gain ratio	Decision Tree	0.1489
	Forest	Forest	0.1236
	Dt - chi square without bonferroni	Decision Tree	0.1024
	Dt - gini	Decision Tree	0.1023
	Dt - entropy	Decision Tree	0.1017

Table 11: ML Algorithms Misclassification rate

The Chi-square model with Bonferroni correction enabled has proven to be the most accurate for predicting loan risk. On the other hand, the Neural Network and Chi-square Automatic Interaction Detection (CHAID) models with Bonferroni correction are the least accurate among the models evaluated.

In general, a machine learning model with an accuracy rate between 0.8 and 0.9, or higher, is considered to achieve an ideal and realistic level of accuracy.

Rank	Name of Models	Accuracy
1	Dt - Chi square with enabled Bonferroni	0.8983
2	Dt - Entropy	0.8983
3	Dt - Gini	0.8977
4	Dt – Chi square without Bonferroni	0.8976
5	Random Forest	0.8764
6	Logistic Regression	0.8201
7	Dt – Chaid without Bonferroni	0.6918
8	Dt – Chaid with bonferroni	0.6910
9	Neaural Network	0.5058

Table 12: ML Algorithms Accuracy Rate

The Area Under the Curve (AUC) is one of the most widely used ranking methods and is considered one of the best single-number metrics for evaluating machine learning algorithms.Essentially, AUC measures a classifier's ability to avoid false classifications. It is derived from the Receiver Operating Characteristic (ROC) curve, with values ranging from 0 to 1. Higher AUC values indicate better prediction performance and fewer misclassifications. AUC is particularly useful for ranking different machine learning algorithms and is advantageous when dealing with imbalanced data. Unlike accuracy, which can be strongly biased towards the majority class, AUC provides a more balanced evaluation of model performance in such scenarios.

Champion	Name	Algorithm	ROC Separation
Champion	Dt - chi square with bonferroni	Decision Tree	0.7963
	Dt - entropy	Decision Tree	0.7961
	Dt - gini	Decision Tree	0.7951
	Dt - chi square without bonferroni	Decision Tree	0.7949
	Forest	Forest	0.7522
	Dt - information gain ratio	Decision Tree	0.7019
	Logistic Regression	Logistic Regression	0.6400
	Dt - chaid without bonferroni	Decision Tree	0.3840
	Dt - chaid with bonferroni	Decision Tree	0.3824
	Neural Network	Neural Network	0.0000

Table 13: ROC Separation of ML Algorithms

We need to make multiple trials by changing the maximum depth and minimum leaf size.

Trial	Max Depth	Min Leaf Size	Num of Important Variables	Num Leaves	Misclassification Rate
1	10	5	28	302	0,1013
2	10	10	21	234	0,1013
3	10	15	21	228	0,1012
4	10	20	21	223	0,1016
5	9	5	25	221	0,1127
6	9	10	22	189	0,1126
7	9	15	19	168	0,1127
8	9	20	18	158	0,1128
9	8	5	15	90	0,1191
10	8	10	20	119	0,1190
11	8	15	14	91	0,1190
12	8	20	14	89	0,1191
13	7	5	18	75	0,1335
14	7	10	18	70	0,1335
15	7	15	17	69	0,1334
16	7	20	17	69	0,1334
17	6	5	8	23	0,1429
18	6	10	8	23	0,1429
19	6	15	8	23	0,1429
20	6	20	8	23	0,1429
21	5	5	7	16	0,1539
22	5	10	7	16	0,1539
23	5	15	7	16	0,1539
24	5	20	7	16	0,1539

Table 14: Trials of finding the optimal decision tree

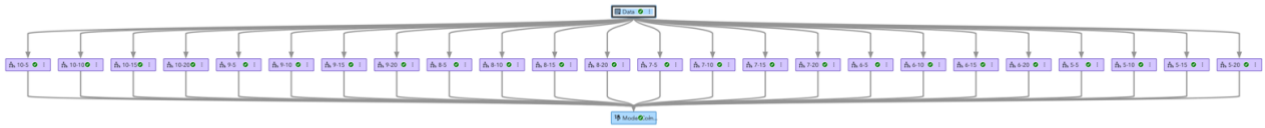

Figure 28: Summary of trials

By making the trials we concluded that the optimal tree is with max depth 6 & minimum leaf size 5.The optimal tree has enough number of important variables and leaves to analyze and classify the good and bad clients.
Furthermore, this decision tree has the same important variables and does not include the variable – IncomeTotal which has negative correlation with variable default.Last but not least the selected decision tree can be easier explained ,trusted by management and business users and be applicable.
In particular, the below variables of the model provide information as to which variables hold greater importance in achieving good performance for our model.

Variable Name	Validation Importance	Relative Importance
PrincipalBalance	9,749.1339	1
PrincipalPaymentsMade	3,095.6585	0.3175
BidsPortfolioManager	1,336.2013	0.1371
Interest	1,257.4446	0.1290
InterestAndPenaltyPaymentsMade	966.7779	0.0992
AppliedAmount	783.0623	0.0803
BidsManual	331.1825	0.0340
Restructured	208.2126	0.0214

Table 15: Variable importance values used in the model evaluation

Below you can find the full decision tree that I choose with max depth 6 & minimum leaf size 5.

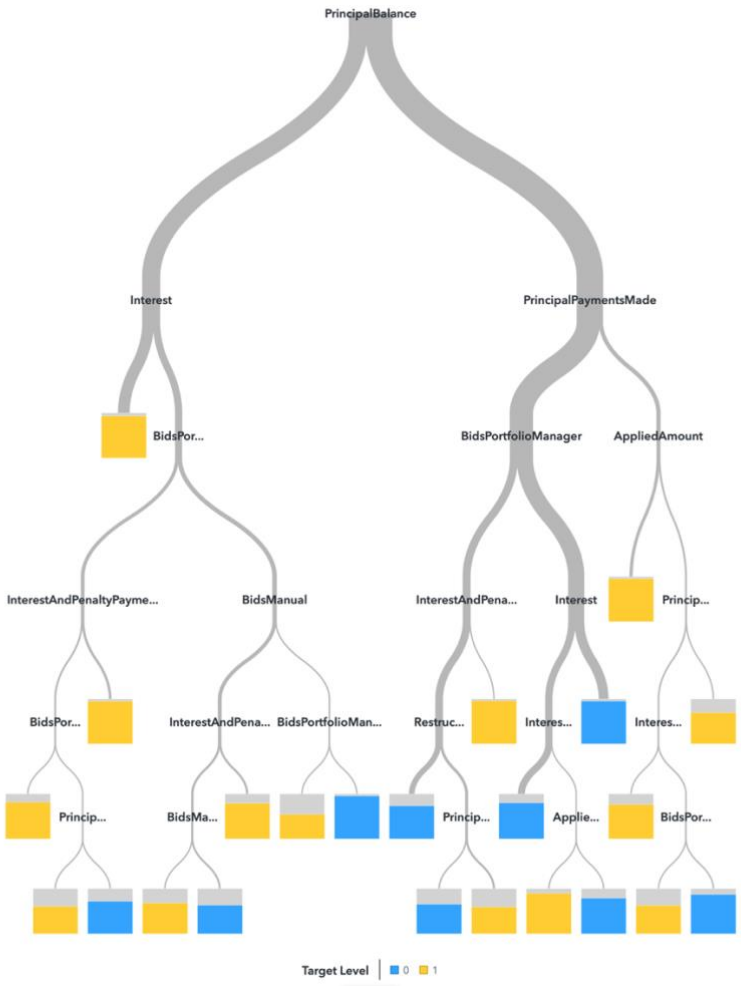

Figure 29: The complete selected decision tree

Furthermore, I am attaching a piece of the decision tree and explain it.

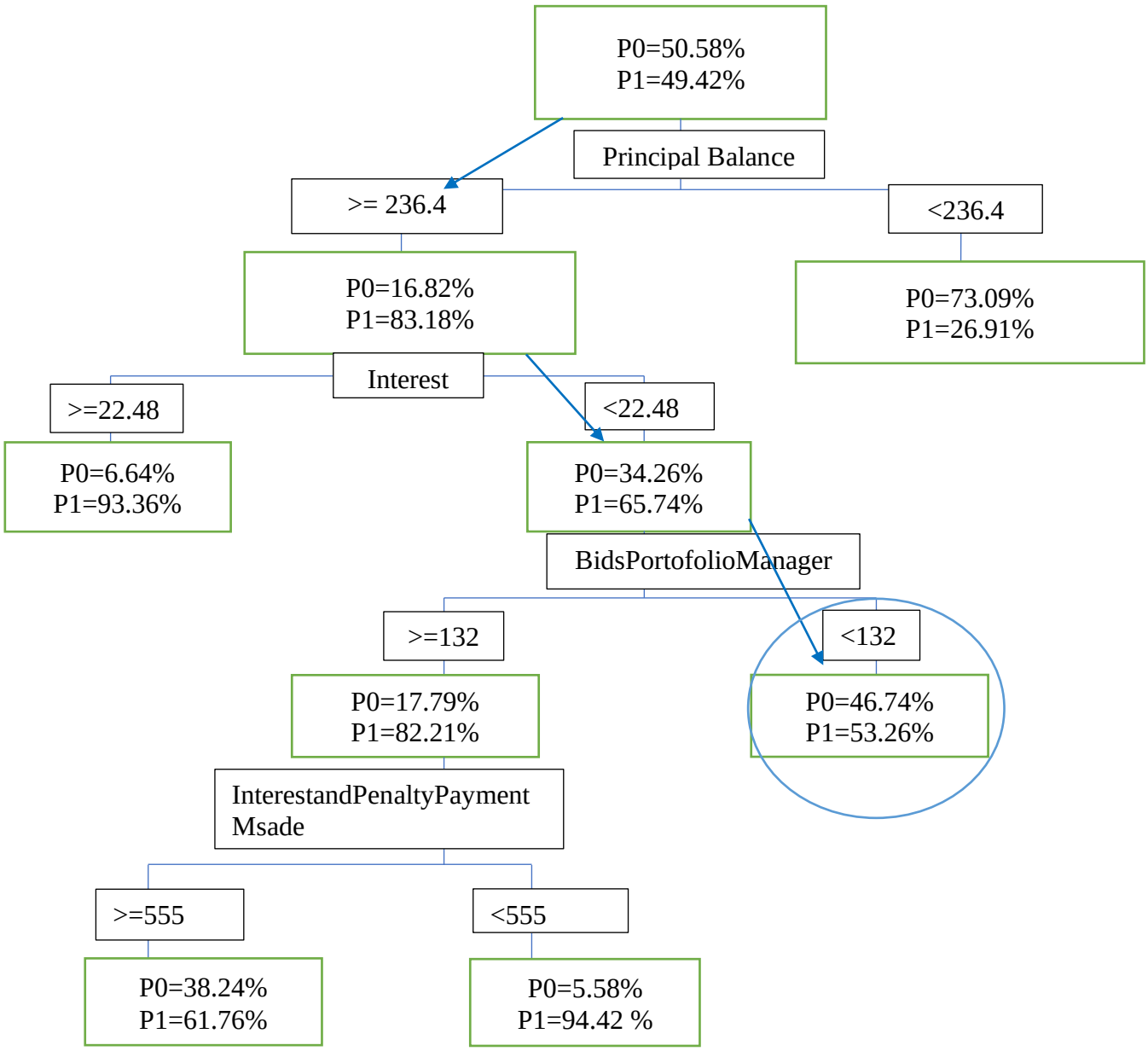

Figure 30: Part of parameterized Decision Tree

If PrincipalBalance is greater than 236.4 ,interest rate less than 22.48 , BidsPortofolioManager is less than 132 then there is a possibility of 53.26% to be a bad client while a 46.74% of being a good client.A training data table can contain a large number of predictor variables. Variable importance is an indication of which predictor variables are most useful for predicting the response variable.The variable importance is very useful to us as it is a very strong driver in your model to predict loan default.
The important variables, ranked in the order of importance (from most to least important) are: PrincipalBalance,PrincipalPaymentsMade,BidsPortfolioManager,Interest,InterestAndPenaltyPaymentsMade,AppliedAmount,BidsManual,Restructured.

Variable	Relative Importance
PrincipalBalance	1
PrincipalPaymentsMade	0.31
BidsPortfolioManager	0.13
Interest	0.12
InterestAndPenaltyPaymentsMade	0.09
AppliedAmount	0.08
BidsManual	0.03
Restructured	0.02

Figure 31: Importance of variables

Scorecard Creation

A scorecard is constructed from multiple characteristics, each comprising various attributes. For instance, the loan term is considered a characteristic. Each attribute within these characteristics is assigned a specific number of scorecard points. These points are statistically determined to differentiate risk, taking into account the predictive power of the characteristic variables, their correlations, and relevant business considerations. This method ensures that the scorecard effectively identifies and quantifies the risk associated with different borrower profiles.

For creating an effective scorecard, it is crucial to utilize the Weight of Evidence (WoE) and Information Value (IV) calculations. Using the Interactive Grouping node in SAS Viya, all independent variables (such as debt-to-income ratio, annual income, and interest rate) are transformed using the WoE method. This method aims to find a monotonic relationship between the input features and the target variable by binning each feature and assigning a weight to each bin. This process not only helps in understanding the contribution of each variable to the outcome but also standardizes the variables to enhance the model's predictive power.
The WoE transformation process in SAS Viya is automated, simplifying the task of handling large datasets with numerous variables. Additionally, it calculates the Information Value (IV) for each row. IV measures the predictive power of independent variables, which is crucial for feature selection. Generally, strong features have an IV between 0.3 and 0.5, indicating high predictive power. Features with an IV ranging from 0.1 to 0.3 have medium predictive power, while those with an IV between 0.02 and 0.1 have weak predictive power. Features with an IV below 0.02 are considered not useful for prediction. An IV above 0.5 is considered suspicious, potentially indicating data issues or overfitting.

The importance of using WoE and IV in scorecard development cannot be overstated. WoE helps in transforming categorical variables into a continuous scale that reflects the strength of the relationship between the independent variables and the target variable. This transformation facilitates the creation of a robust predictive model by highlighting variables that contribute significantly to predicting loan defaults. IV further aids in ranking these variables based on their predictive power, ensuring that only the most relevant features are included in the final model.
By integrating these statistical methods, the scorecard becomes a powerful tool for assessing credit risk. It provides a clear and quantifiable measure of risk associated with each applicant, helping financial institutions make informed lending decisions. The process of transforming variables using WoE and evaluating them using IV also enhances the interpretability of the model, making it easier for stakeholders to understand and trust the predictions.

By building the scorecard and running the interactive grouping node with its default settings and the scorecard node we received as a result 10 variables with power of predictability, which are the following:

- PrincipalBalance
- PrincipalPaymentsMade
- BidsPortfolioManager
- Interest
- BidsManual
- Restructured
- AppliedAmount
- Rating
- LoanDuration
- InterestAndPenaltyPaymentsMa

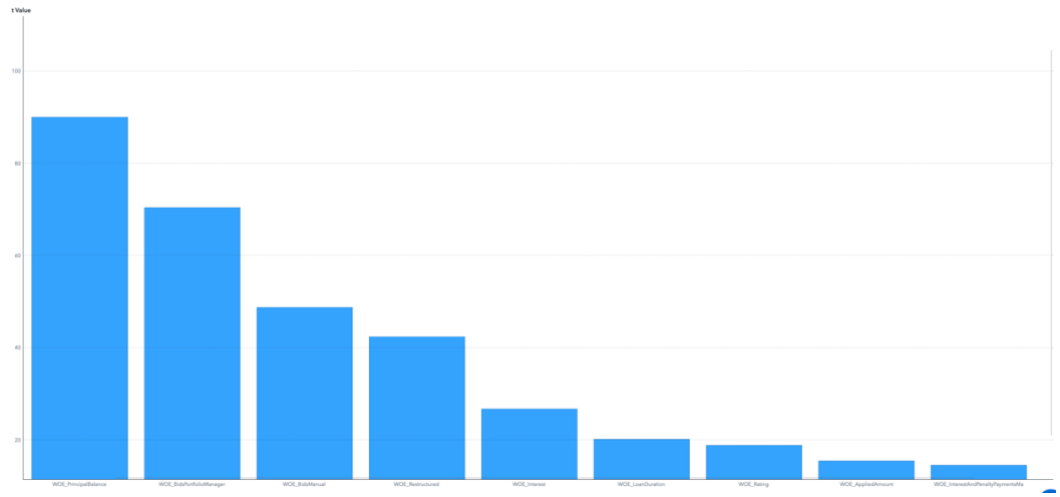

Figure 32:t-Value Chart of variables

The features highlighted in the t-Value chart are the crucial variables used by our logistic regression model for classification. These features have been identified as having significant predictive power in determining the likelihood of loan repayment. In our model, all the features have a negative sign, indicating that an increase in the value of each feature is associated with an increased probability that the data point represents a good customer.

In practical terms, a higher score correlates with a greater likelihood that the customer will repay the loan. This means that customers with higher scores are considered lower risk and thus more creditworthy. The model effectively identifies these low-risk customers by analyzing the impact of various features on the likelihood of repayment. For example, higher scores in certain features might indicate a stronger financial position or more stable employment, both of which contribute positively to the customer's ability to repay the loan.

The interpretability of the logistic regression model is particularly valuable in the credit scoring process. Stakeholders, including loan officers and regulatory bodies, can understand how different features influence the final score and make informed decisions based on these insights. This transparency also ensures that customers can be given clear explanations for their credit ratings, which can help maintain trust and compliance with regulatory standards.
A comprehensive understanding of these key features allows the financial institution to better manage credit risk. By accurately identifying and approving low-risk borrowers, the institution can minimize defaults and enhance the overall health of its loan portfolio. Additionally, the ability to quantify the impact of each feature on the likelihood of repayment provides a robust framework for continuous improvement of the credit scoring model.

Selecting a single customer at random from the validation dataset with a total score of 201 points , the total score results from:

- 10 points from Applied Amount
- 15 points from BidsManual
- 44 points from BidsPortfolioManager
- 22 points from interest
- 13 points from InterestAndPenaltyPaymentsMa
- 24 points from LoanDuration
- 26 points from PrincipalBalance
- 17 points from PrincipalPaymentsMade
- 15 points from Rating
- 15 points from Restructured

SCORECARD_P...	SCR_AppliedAmount	SCR_BidsManual	SCR_BidsPortfolioManager	SCR_Interest	SCR_InterestAndPenaltyPaymentsMa	SCR_LoanDurat...	SCR_PrincipalB...	SCR_PrincipalPaymentsMade	SCR_Rating	SCR_Restructured
201	10	15	44	22	13	24	26	17	15	15

Figure 33: Points per variable of customer

Following, we review with more detail the scorecard points of each variable.

Scorecard

		Scorecard Points
AppliedAmount	AppliedAmount < 530	20
	530 <= AppliedAmount < 2125, _MISSING_	10
	2125 <= AppliedAmount < 3632	5
	3632 <= AppliedAmount < 4250	13
	4250 <= AppliedAmount	4
BidsManual	BidsManual < 5	27
	5 <= BidsManual < 8	3
	8 <= BidsManual < 49	26
	49 <= BidsManual < 101	15
	101 <= BidsManual, _MISSING_	0
BidsPortfolioManager	BidsPortfolioManager < 30	44
	30 <= BidsPortfolioManager < 119	15
	119 <= BidsPortfolioManager < 314	2
	314 <= BidsPortfolioManager < 1305, _MISSING_	-5
	1305 <= BidsPortfolioManager	-11
Interest	Interest < 18.83	16
	18.83 <= Interest < 21.63	22
	21.63 <= Interest < 31	8
	31 <= Interest < 46.97, _MISSING_	3
	46.97 <= Interest	-7
InterestAndPenaltyPaymentsMade	InterestAndPenaltyPaymentsMade < 6.76	1
	6.76 <= InterestAndPenaltyPaymentsMade < 37.95	16
	37.95 <= InterestAndPenaltyPaymentsMade < 58.54	13
	58.54 <= InterestAndPenaltyPaymentsMade < 177.51	10
	177.51 <= InterestAndPenaltyPaymentsMade, _MISSING_	8
LoanDuration	LoanDuration < 18	24
	18 <= LoanDuration < 36	18
	36 <= LoanDuration < 48	6
	48 <= LoanDuration < 60	6
	60 <= LoanDuration, _MISSING_	7
PrincipalBalance	PrincipalBalance < 236.4, _MISSING_	26
	236.4 <= PrincipalBalance < 1517.54	-20
	1517.54 <= PrincipalBalance < 2691.3	-27
	2691.3 <= PrincipalBalance < 4086.83	-11
	4086.83 <= PrincipalBalance	-27
PrincipalPaymentsMade	PrincipalPaymentsMade < 60.94	-55
	60.94 <= PrincipalPaymentsMade < 510.69	-29
	510.69 <= PrincipalPaymentsMade < 531	82
	531 <= PrincipalPaymentsMade < 3189, _MISSING_	17
	3189 <= PrincipalPaymentsMade	60
Rating	HR	-4
	F	0
	E	4
	D	10
	A, AA, B, C, _MISSING_, _UNKNOWN_	15
Restructured	YES	-5
	NO, _MISSING_, _UNKNOWN_	15

Figure 34: Scorecard points per variable

For this scorecard, a cutoff score of 200 points has been established. This cutoff serves as a benchmark for distinguishing between good and bad credit risks. The scoring system is meticulously scaled so that a total score of 200 points corresponds to good/bad odds of 50 to 1. In other words, at a score of 200, the likelihood of a borrower being a good credit risk is 50 times higher than being a bad risk. Additionally, an incremental increase of 20 points in the score results in a doubling of the good/bad odds, meaning that with each 20-point increase, the borrower is twice as likely to be a good credit risk.
It is crucial to highlight that the selection of the cutoff score does not influence the inherent predictive strength of the scorecard. The scorecard's effectiveness remains consistent regardless of the specific cutoff value chosen. This scorecard is generated based on the specified values using SAS Viya, a powerful tool for advanced analytics and machine learning.

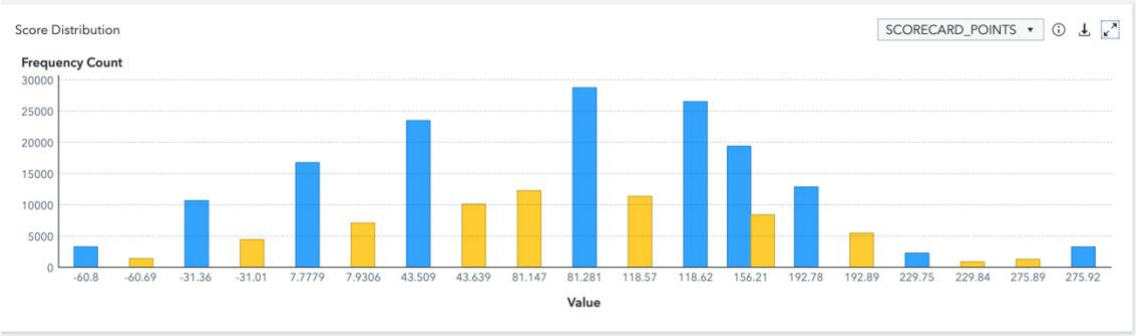

Figure 35: Distribution of scores at the scorecard

The distribution of the scorecard scores shows a mean score of 81, indicating the central tendency of the credit scores within the dataset. These scores are instrumental in categorizing consumers into distinct tiers based on their creditworthiness. By analyzing the score distribution, customers can be segregated into various risk bands, facilitating better credit risk management and decision-making.

Specifically, loan applicants are divided into four distinct risk bands:

- **Low Risk**: The top 20% of loan applicants, who fall into this category, have the highest likelihood of repaying their loans. These borrowers are considered the most creditworthy and pose the least risk to lenders.
- **Medium Risk**: The next 20% of applicants fall into the medium risk category. While these borrowers are slightly riskier than the low-risk group, they still represent a relatively stable credit risk.
- **High Risk**: The subsequent 20% of applicants are classified as high risk. These borrowers have a higher probability of default compared to the medium and low-risk groups, necessitating more cautious lending practices.
- **Extremely High Risk**: The bottom 20% of applicants fall into this category, representing the highest risk of default. These borrowers are the least creditworthy and pose the greatest risk to lenders, requiring stringent credit evaluation and monitoring.

This segmentation allows lenders to tailor their lending strategies according to the risk profile of each group, enhancing overall risk management. By differentiating between these risk bands, lenders can implement more effective credit policies, offer appropriate loan products, and take proactive measures to mitigate potential defaults.

Conclusion

This study aimed to evaluate and predict credit risk in the context of peer-to-peer (P2P) lending platforms using various machine learning methods. The primary objective was to classify borrowers based on their likelihood of defaulting on loans. This section discusses the obtained results, analyzes the limitations and advantages of the methods used, compares these methods, and relates the findings to similar published results.
The analysis utilized a decision tree model, logistic regression, and other machine learning algorithms. The decision tree model (Chi-Square with Bonferroni enabled) proved to be the most accurate for predicting loan risk. ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve) metrics assessed model performance. A higher AUC value signifies better performance in distinguishing between default and non-default cases.

Key findings include:

- Principal Balance, Interest Rate, and Loan Duration were the most influential variables in predicting loan defaults.
- The decision tree model with Chi-Square and Bonferroni adjustments had the highest accuracy and lowest misclassification rate.
- Logistic regression, though less accurate than the decision tree, provided valuable insights due to its interpretability.
- Neural networks, while potentially more accurate, suffered from interpretability issues, making them less suitable where understanding the decision-making process is crucial.

Comparison of Methods

1. Decision Tree (Chi-Square with Bonferroni):
   - Advantages: High accuracy, easy to interpret, provides clear decision rules.
   - Disadvantages: Can overfit if not properly pruned, may become complex with large datasets.
2. Logistic Regression:
   - Advantages: High interpretability, straightforward to implement, good for baseline comparison.
   - Disadvantages: Lower accuracy compared to decision trees, may not capture complex patterns in data.
3. Neural Networks:
   - Advantages: High potential accuracy, capable of handling complex relationships in data.
   - Disadvantages: Poor interpretability, requires extensive computational resources, overfitting can be an issue.
4. Random Forests:
   - Advantages: Reduces overfitting by averaging multiple decision trees, handles large datasets well.
   - Disadvantages: Lower interpretability than single decision trees, computationally intensive.
5. Support Vector Machines (SVMs):
   - Advantages: Effective in high-dimensional spaces, robust to overfitting in certain conditions.
   - Disadvantages: Difficult to interpret, requires careful tuning of parameters, computationally demanding.

Comparison with Similar Published Results

The findings align with several key studies in the literature. For instance:

- Müller & Guido (2017) discussed the advantages of using decision trees for their simplicity and interpretability, supporting this study's findings.
- Tulio Ribeiro et al. (2016) and Chen et al. (2023) addresses the challenge of interpreting predictions made by complex "black-box" machine learning models compared to more interpretable models like decision trees and logistic regression. The authors introduce the Local Interpretable Model-agnostic Explanations (LIME) framework as a solution to this challenge. LIME generates local surrogate models around individual predictions to approximate how the black-box model behaves locally. This approach aims to provide insights into why a particular prediction was made, thereby bridging the interpretability gap between complex models and simpler, more transparent models. The authors conclude that while complex models like deep neural networks and ensemble methods offer superior predictive performance, their lack of transparency poses challenges for deployment in sensitive domains such as healthcare and finance. LIME represents a significant step toward making these models more interpretable and trustworthy by providing explanations that are understandable to users and stakeholders.
- Netzer, Lemaire, and Herzenstein (2019) explored the use of text data(NLP techniques) from loan applications to predict defaults, highlighting the potential of integrating different data types for improved accuracy.
- Sulejmani(2021)the study highlights the considerable potential of advanced machine learning, especially neural networks, in predicting loan defaults in peer-to-peer lending, demonstrating superior accuracy. However, it also points out the interpretability issues of these complex models. Incorporating explainable AI techniques like SHAP and LIME addresses these concerns, ensuring transparency and regulatory compliance. The research advocates for future exploration of hybrid models that blend high accuracy with interpretability, thereby enhancing credit risk assessment in the financial sector.
- Koskimaki,M., (2021) uses a variety of machine learning models, including logistic regression, decision trees, and random forests, to predict loan defaults. Koskimaki finds that while decision trees provide a good balance of accuracy and interpretability, random forests offer higher predictive performance. The study emphasizes the importance of incorporating country-specific variables to improve the accuracy of default predictions. Both studies highlight the utility of decision trees for their interpretability and reasonable accuracy in predicting loan defaults.
- Driesse(2022) study's findings highlight the promise of hybrid models that integrate deep learning with traditional machine learning methods to improve credit risk prediction in P2P lending. These models effectively balance high predictive accuracy with interpretability, meeting both practical and regulatory needs.
- Zhang and Xie (2021) examined the use of deep learning and explainable AI in predicting loan defaults, finding that while deep learning models are powerful, their lack of transparency is a significant drawback compared to more interpretable methods.
- Byanjankar (2015) study on credit risk analysis using machine learning methods both aim to improve the accuracy of credit risk prediction in P2P lending. Byanjankar's focus on neural networks demonstrates their superior predictive power due to their ability to capture complex patterns. However our employs decision trees and logistic regression, highlighting their interpretability and ease of use. Both papers recognize the advantages and limitations of their chosen methods. Byanjankar (2015) points out the high accuracy of neural networks but also notes the interpretability issues, which can be a significant drawback in financial applications where transparency is critical. The attached paper discusses the trade-offs between accuracy and interpretability, choosing decision trees and logistic regression for their balance of both. It highlights the importance of stakeholder trust and regulatory compliance, which are supported by the clear and understandable nature of these models.
- Dupont and Cliche (2020) emphasized the governance challenges associated with AI in finance, particularly the need for transparency and interpretability to meet regulatory requirements and maintain stakeholder trust.a

These comparisons underscore that while complex models may offer higher accuracy, the interpretability provided by simpler models like decision trees and logistic regression is invaluable in the financial sector. This is consistent with broader literature emphasizing the balance between accuracy and transparency in credit risk modeling.

The study demonstrated that decision trees, particularly those with Chi-Square and Bonferroni adjustments, are highly effective for predicting loan defaults in P2P lending. These models strike a balance between accuracy and interpretability, making them suitable for credit risk analysis. Logistic regression serves as a useful baseline model, offering insights despite its lower accuracy. The integration of explainable AI techniques, such as SHAP and LIME, could further enhance the transparency of more complex models, although simpler models already provide significant value in this context.

A critical aspect of this study was the analysis of the relative importance of various variables. The results indicated that financial characteristics, such as PrincipalBalance,Interest and AppliedAmount, were significantly influential in predicting default risk .This insight is crucial for lenders as it directs focus towards the most impactful factors when assessing loan applications.

The Scorecard and Decision Tree models share several key variables that significantly influence and predict credit risk. These common variables include PrincipalBalance, Interest, AppliedAmount, BidsPortfolioManager, PrincipalPaymentsMade, Restructured, and BidManual. These shared variables are essential as they offer insights into the borrower's financial status and behavior, thereby enhancing the models' predictive accuracy.

In contrast, the Scorecard model also incorporates additional variables that are not included in the Decision Tree model. These additional variables are Country, IncomeTotal, Rating, LoanDuration, and MaritalStatus. The inclusion of these variables provides a more comprehensive view of the borrower's profile. For example, Country adds geographical context, IncomeTotal reflects the borrower's overall financial capacity, Rating assesses creditworthiness, LoanDuration indicates the loan term, and MaritalStatus provides demographic context. By incorporating these extra variables, the Scorecard model aims to enhance the detail and precision of credit risk predictions, offering lenders a more robust decision-making tool.

Moreover, the scorecard developed using SAS further classified borrowers into 'high risk' and 'low risk' categories based on their credit scores and payment histories. This classification system is instrumental for lenders, enabling them to make informed decisions about loan approvals, effectively balancing the potential risks and rewards associated with each borrower. By distinguishing between high and low-risk borrowers, lenders can optimize their lending strategies, enhance portfolio performance, and mitigate potential losses.

In the thesis we made an effort to develop a scorecard which has many advantages as :

- Objective Evaluation

  Credit scorecards provide a reliable and uniform method to evaluate credit risk, reducing dependence on subjective opinions and ensuring fair decision-making.
- Reducing Credit Losses

  Identify high-risk borrowers to effectively manage credit losses and apply targeted strategies to minimize potential losses.
- Informed Lending Decisions

  Equip lenders with essential information to make sound decisions, enabling adjustments to credit terms, interest rates, and loan amounts based on risk profiles.
- Managing Credit Exposure

  Effectively control credit risk exposure through the use of credit scorecards, maintaining a balanced and healthy loan portfolio.
- Promoting Responsible Lending

  Encourage responsible lending practices by using scorecards to extend credit to deserving applicants while minimizing default risks.
- Consistent Evaluation

  Ensure uniform assessment of applicants with a standardized credit scorecard methodology, regardless of the lending institution's size or type.
- Insightful Data Analysis

Utilize historical data through scorecards to gain valuable insights into borrower behaviors and credit trends, continuously refining the lending process.

The results contribute to the broader field of credit risk analysis by validating the use of decision trees in P2P lending and highlighting the importance of model interpretability in financial applications. Future research could explore the integration of additional data sources and advanced interpretability techniques to further improve model performance and transparency.

Bibliography

- Alkhyeli K., 2023. Explainable AI for Credit Risk Assessment, Khalifa University.
- Anderson, R., 2007. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press.
- Basel Committee on Banking Supervision., 2005. Studies on the validation of internal rating systems. Bank for International Settlements.
- Bora, U., 2021. Master thesis: Credit Risk Analysis using Machine Learning Methods and Explainable AI. University of Bergen. Retrieved from https://bora.uib.no/bora-xmlui/bitstream/handle/11250/2762661/Master-thesis-June-2021.pdf?sequence=1&isAllowed=y
- Bradley, A. P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159.
- Byanjankar, A., 2015. Predicting Credit Risk Levels in Online Peer to Peer Lending Using Neural Network, Åbo Akademi University.
- Cepal. (n.d.). Our Services. Retrieved from https://www.cepal.gr
- Chen, Y., Calabrese, R., & Martin-Barragan, B., 2023. Interpretable machine learning for imbalanced credit scoring datasets.
- Christensen M., 2023. What is Peer-to-Peer (P2P) Lending? How it works. Retrieved from https://p2pmarketdata.com/articles/p2p-lending-explained/#types-of-peer-to-peer-p2p-lending-websites
- Crook, J. N., Edelman, D. B., & Thomas, L. C., 2007. Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465.
- Das D., 2021. Credit Risk Modeling using Machine Learning Approach Part 1. Retrieved from https://thirdeyedata.ai/credit-risk-modeling-using-machine-learning-approach-part-1/
- Deloitte., (N.D.) ,2013. The Analytics Advantage. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-analytics-advantage-report-061913.pdf
- doValue Greece. (n.d.). About Us. Retrieved from https://www.dovaluegreece.gr
- Driesse M.B., 2022. Peer-to-Peer Credit Forecasting a Deep Hybrid Learning Approach, Erasmus University Rotterdam.
- Dupont, L., & Cliche, O., 2020. Governance of Artificial Intelligence in Finance. Journal of Financial Transformation, 52, 45-56.
- Fay B., 2023. Peer Lending. Retrieved from https://www.debt.org/credit/solutions/peer-lending/
- Federal Reserve Bank of Cleveland.,2017. Peer-to-Peer Lending and Credit Card Use.
- Gavin M., 2019. The Importance of Business Analytics. Retrieved from https://online.hbs.edu/blog/post/importance-of-business-analytics
- Gavira M., 2018. LinkedIn. How Netflix Uses AI and Data to Conquer the World. Retrieved from https://www.linkedin.com/pulse/how-netflix-uses-ai-data-conquer-world-mario-gavira/
- Ghosh, A., 2012. Managing Credit Risk in Banking: A Practitioner's Guide. Wiley Finance Series. John Wiley & Sons.
- Hovdenakk A., 2021. June, Machine Learning vs logistic regression in credit scoring: A trade-off between accuracy and interpretability, University of Bergen.
- Hastie, T., Tibshirani, R., & Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Kagan J., 2024. Collection Agency: Definition, How It Works, and Regulations. Retrieved from https://www.investopedia.com/terms/c/collectionagency.asp
- Kapil A., 2022. Decision Tree Algorithm in Machine Learning: Advantages, Disadvantages, and Limitations. Retrieved from https://www.analytixlabs.co.in/blog/decision-tree-algorithm/
- Koskimaki, M., 2021. Default Prediction in peer-to-peer lending and country comparison, LUT School of Business of Management, 52-55.
- LendingClub. (n.d.). About LendingClub. Retrieved from https://www.lendingclub.com
- Müller, A. C., & Guido, S., 2017. Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.

- Netzer, O., Lemaire, A., & Herzenstein, M., 2019. When words sweat: Identifying signals for loan default in the text of loan applications. Journal of Marketing Research, 56(6), 998-1011.
- Nguyen, L., 2017. Credit risk control for loan products in commercial banks. Case: Bank for Investment and Development of Vietnam. University of Economics, Ho Chi Minh City.
- Oracle. (N.D.). What is AI? Retrieved from https://www.oracle.com/in/artificial-intelligence/what-is-ai/
- Qualtrics, 2022. Risk Management Analytics. Retrieved from https://www.qualtrics.com/blog/risk-management-analytics/
- Rachman M., 2023. Credit Scoring Prediction. Retrieved from https://medium.com/@m.ariefrachmaann/credit-scoring-prediction-acd1f576fd86
- Ramaswamy S., 2017. How Companies Are Already Using AI. Retrieved from https://hbr.org/2017/04/how-companies-are-already-using-ai
- SAS Institute Inc., 2021. SAS Viya 4: Machine Learning. SAS Institute Inc.
- Siddiqi N.,2006,Credit Risk Scorecards : Developing and Implementing Intelligent Credit Scoring, SAS Institute Inc.
- Sokolova, M., Japkowicz, N., & Szpakowicz, S., 2009. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In AI 2006: Advances in Artificial Intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.
- Stedman C., 2023. Business Intelligence (BI). Retrieved from https://www.techtarget.com/searchbusinessanalytics/definition/business-intelligence-BI
- Sulejmani S., 2021. Using Deep Learning and Explainable AI to Predict and Explain Loan Defaults, ZHAW School of Management and Law.
- Tulio Ribeiro, M., Singh, S., & Guestrin, C., 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.
- Wang, S., 2020. Default Risks in Marketplace Lending. Kent State University College.
- Zhang, H., & Xie, H., 2021. Using Deep Learning and Explainable AI to Predict and Explain Loan Defaults. Journal of Finance and Data Science, 7(3), 142-158.