UNIVERSITY OF THE PELOPONNESE & NCSR “DEMOCRITOS”
MSC PROGRAMME IN DATA SCIENCE

Sounding the Alarm!
AI Early Warning Systems for Loan Defaults

A Case Study of the Greek Lending Market

by

Anastasios Kanellopoulos

A thesis submitted in partial fulfillment
of the requirements for the MSc
in Data Science

Supervisor: Ilias Zavitsanos
Doctor

Co-supervisors: Konstantinos Bougatiotis,
Doctor,

Athens, 12 2024

Sounding the Alarm! AI Early Warning Systems for Loan Defaults

Anastasios Kanellopoulos

MSc. Thesis, MSc. Programme in Data Science

University of the Peloponnese & NCSR “Democritos”, 12 2024

UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOCRITOS"
MSC PROGRAMME IN DATA SCIENCE

Sounding the Alarm!
AI Early Warning Systems for Loan Defaults

A Case Study of the Greek Lending Market

by

Anastasios Kanellopoulos

A thesis submitted in partial fulfillment
of the requirements for the MSc
in Data Science

Supervisor:	Ilias Zavitsanos Doctor
Co-supervisors:	Konstantinos Bougatiotis, Doctor,

Approved by the examination committee on 12, 2024.

(Signature)	(Signature)	(Signature)
....	....	....
Ilias Zavitsanos Doctor	Anastasia Krithara Doctor	Spiros Skiadopoulos Doctor

Athens, 12 2024

UNIVERSITY OF THE PELOPONNESE & NCSR "DEMOCRITOS"
MSC PROGRAMME IN DATA SCIENCE

Declaration of Authorship

(1) I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

(2) I confirm that this thesis presented for the degree of Bachelor of Science in Informatics and Telecommunications, has

(i) been composed entirely by myself

(ii) been solely the result of my own work

(iii) not been submitted for any other degree or professional qualification

(3) I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or processional qualification except as specified.

(Signature)

....

Anastasios Kanellopoulos

Athens, 12 2024

Acknowledgments

I want to thank my supervisor Ilias Zavitsanos and co-supervisor Konstantinos Bougatiotis for providing guidance when needed, Qualco for providing the data, the members of the committee Anastasia Krithara and Spiros Skiadopoulos, my girlfriend Iro and all my friends for the patience and emotional support, and finally Ilias who provided valuable points and thoughtful discussions throughout this endeavor.

To my family.

Περίληψη

Σκοπός αυτής της εργασίας είναι διερεύνηση και εφαρμογή της μηχανικής μάθησης (ΜΛ) για την πρόβλεψη αθέτησης δανείων στην ελληνική αγορά δανεισμού. Χρησιμοποιώντας ένα μοναδικό σύνολο δεδομένων που παρέχεται από την Χυαλςο, έναν κορυφαίο πάροχο λύσεων διαχείρισης χρηματοοικονομικού κινδύνου, η έρευνα εστιάζει στην αναγνώριση δανειοληπτών που πληρούν τις τρέχουσες υποχρεώσεις πληρωμής τους, αλλά παρουσιάζουν υψηλό κίνδυνο αθέτησης στο μέλλον. Η μελέτη χρησιμοποιεί μια ολοκληρωμένη μεθοδολογία, που περιλαμβάνει την προεπεξεργασία δεδομένων, τη μηχανική χαρακτηριστικών και την εκπαίδευση και αξιολόγηση διαφόρων μοντέλων ΜΛ, συμπεριλαμβανομένης της λογιστικής παλινδρόμησης και προηγμένων αλγορίθμων όπως Ρανδομ Φορεστ, ΞΓΒοοστ, ΛιγητΓΒΜ και ἇτΒοοστ. Τα αποτελέσματα δείχνουν ότι οι μέθοδοι συνόλου ξεπερνούν το βασικό μοντέλο λογιστικής παλινδρόμησης, με το ἇτΒοοστ να επιτυγχάνει την υψηλότερη ακρίβεια. Η ανάλυση σημαντικότητας χαρακτηριστικών αποκαλύπτει βασικούς παράγοντες που οδηγούν σε κίνδυνο αθέτησης, συμπεριλαμβανομένου του ποσού των δόσεων του επόμενου μήνα, της κατάστασης της αίτησης και των ιστορικών προτύπων διακανονισμού. Τα ευρήματα συμβάλλουν σε μια βαθύτερη κατανόηση της πρόβλεψης αθέτησης δανείων στο ελληνικό πλαίσιο και πληροφορούν την ανάπτυξη πιο αποτελεσματικών στρατηγικών διαχείρισης πιστωτικού κινδύνου.

Abstract

The aim of this thesis is to apply the concept of machine learning (ML) to the problem of loan default prediction in the context of the Greek lending market. Using a unique data set from Qualco, a leading supplier of financial risk management software, the study aims at identifying the current payers who have a high propensity to default in the future. The research is carried out in a systematic manner whereby the study involves data preprocessing, feature engineering, and application of various ML models for training and evaluation including Logistic Regression and ensemble models such as Random Forest, XGBoost, LightGBM, and CatBoost. The experimental results show that ensemble methods outperform the baseline Logistic Regression model with CatBoost achieving the highest score (AUC). The results of feature importance analysis show that such factors as the sum of next-month installments, application status, and historical settlement patterns are the most influential in determining default risk. The results are useful in providing a better insight of the default risk assessment in the context of the Greek market and help in the creation of more efficient credit risk management mechanisms.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
EDA	Exploratory Data Analysis
ff	full-feature
fs	feature-selected
GA	Genetic Algorithms
GNN	Graph Neural Networks
KNN	K-Nearest Neighbors
L1	L1 Regularization
Lasso	Least Absolute Shrinkage and Selection Operator
LDP	Loan Default Prediction
LGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
ML	Machine Learning

NPL	Non-Performing Loan
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
P2P	Peer-to-Peer
RF	Random Forest
RFE	Recursive Feature Elimination
RNN	Recurrent Neural Networks
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Problem description

The banking and lending industry can be regarded as one of the most important branches of the economy. Banking institutions are the major lenders in an economy, providing money to the people and firms to meet their financial requirements. The scale of the lending business can be different within different countries, but it is quite relevant to the overall economy. Based on the data from World Bank, the total outstanding loans provided by banks across the globe as of 2019 was 102 trillion this is more than twice the global GDP [1]. Leverage and credit are the main engines that shape the economic development. Credit is the way to obtain money or any other product or service and to return it together with interest at some future date. Similar to that leverage is the process of utilizing other people's money or borrowed capital in order to invest in other projects with the aim of increasing on returns. Together, credit and leverage allow people and organizations to grow into new business lines and activities at faster rates or to undertake projects that would not be possible without debt financing [2]. For individuals, credit and leverage allowed people to buy a home or expenditures, increase which consumer in turn leads to economic activity since it leads to the creation of new jobs and increase in income Thus, credit and leverage are major forces that drive growth of the economy. Hence, it is crucial for the governments and other financial institutions to ensure that credit and leverage are made available to individuals and firms to boost growth and development of the economy. However, it is also crucial to consider the risks involved with credit and leverage to ensure that people and businesses are not over extending themselves and that the economy is balanced and sustainable in the long term for healthy economic cycle and achieve the soft landings and avoid the harsh recessions [3].

1.2 Motivation

The problem of loan defaults is one of the most important in the financial industry, especially in the Greek market, for a number of reasons that are interconnected and have an impact on the financial institutions and the overall economy. The Greek economy has gone through some challenges in the recent past such as the global financial crisis and the domestic economic condition which have led to increased non-performing loans [4]. As a result, it is imperative to discuss the financial consequences of loan defaults as far as the lending institutions are concerned. High incidences of loan defaults present a threat to profitability and the overall viability of banks and other related financial institutions [5]. For instance, Oghenekaro explains that is why it is important to make correct predictions of loan defaults for the financial institutions to make right decisions on who to give loans to especially due to the huge number of applicants they receive [6]. Also, the Greek financial profile has been characterized by the economic problems including high levels of unemployment and instability of the economy, which makes the estimation of loan defaults even more difficult [7]. The application of sophisticated methods of machine learning as the supporting analysis allows for improving the accuracy of default predictions [8].

1.3 Contributions

This paper's major innovation relates to the application of machine learning tools in credit risk management focusing on the use of ensemble learning techniques for loan default forecasting. Despite the ever increasing use of machine learning in financial risk management, there is still relatively limited use of ensemble methods. The institutions conduct transactions of billions of euros and even small increases in the accuracy of predictions can lead to reduction of losses and increased profitability.

Credit scoring is now accepted as an essential instrument for the growth of the economy and increasing financial inclusion. It enhances the availability of credit to individuals and medium enterprises while at the same time enhancing credit management. Over the last few years, credit scoring models have been used for more than just credit scoring. on They are used for setting various credit limits, managing customer relationships and identifying potential customers. These new applications demand enhanced computational power, availability of data and the need to enhance efficiency in the management of credit scoring techniques. Traditional statistical approaches have been augmented and, in some cases, outperformed by artificial neural networks, support vector machines, and other forms of AI and ML classifiers, including random forests and gradient boosters. It has also enhanced the types of data that can be incorporated in credit scoring models making the models easier to develop and more specific.

These innovations have the following advantages. 1. Enhancing the availability of credit to the broader population and businesses. 2. Improving model efficiency using sophisticated algorithms and performing better risk appraisals and finally, 3. Enhancing customer experience. Risk considerations such as data privacy, fairness, bias in historical data and model interpretability are some of the major challenges that impede wider adoption. These issues are even more acute in markets that have not fully developed or matured to have well defined regulations or best practices. Furthermore, in the emerging markets, credit scoring is still done manually or by using scorecards, judgment or regression models due to constraints in technology, talent and data infrastructure. Driven by these reasons this research, will try to apply ensemble learning methods in credit scoring with the goal to increase the accuracy of the predictions, decrease default rates and contribute to the stability of the economy. These methods can be very useful in the promotion of financial inclusion, risk management and the encouragement of sustainable economic development.

We suggest a machine learning pipeline for the Greek loan market, where ensemble learning methods are used for enhancing loan default risk prediction. This pipeline incorporates account level in combination with borrower level behavioral data, such as repayment patterns and payment delays, and demographics like age, employment sector, and geographic location, to develop a comprehensive ML dataset. The pipeline is designed to address key challenges in credit risk management, including imbalanced targets, by employing ensemble models, well optimized for tabular data capable of processing and handling complex relationships. Furthermore, it prioritizes explainability through tools like SHAP values, ensuring predictions are transparent and actionable, particularly in regulated financial environments. By focusing on the unique characteristics of the Greek market and emphasizing financial inclusion, this approach aligns with the broader goals of increasing credit availability, reducing default rates, and fostering economic stability. The pipeline's modular design also enables its adaptability to emerging markets, where constraints in technology and data infrastructure often limit the use of advanced credit scoring methods.

Chapter 2

Background

2.1 Key Terms

Loan Default:

This refers to the situation where a borrower fails to fulfill their debt obligations as specified in the loan agreement. This typically involves missing payments for a certain period, which can vary depending on the lender and the type of loan. In the context of your thesis, you might want to specify the exact criteria used to define a loan default within the Qualco dataset (e.g., number of missed payments, outstanding balance threshold).

Credit Risk: Credit risk is the potential for financial loss that a lender faces due to a borrower's failure to repay a loan. It is an inherent aspect of lending activities and encompasses various factors that can influence a borrower's ability or willingness to meet their debt obligations. These factors can include the borrower's credit history, financial stability, employment status, and even macroeconomic conditions. In your thesis, you're essentially developing models to assess and quantify credit risk, allowing lenders to make informed decisions about loan approvals and risk mitigation strategies.

Credit Scoring: Credit scoring is the process of using statistical models to evaluate the creditworthiness of an individual or business. These models produce a numerical score that represents the likelihood of a borrower defaulting on their debt obligations. Credit scoring models typically incorporate a variety of data points, such as payment history, debt levels, and demographic information, to assess the risk associated with extending credit to a particular borrower.

Predictive Modeling: Predictive modeling is the process of using statistical techniques and machine learning algorithms to create models that can predict future outcomes based on historical data. These models are trained on large datasets to identify patterns and relationships between variables, allowing them to make predictions on new, unseen data. In your research, you're using predictive modeling to develop models that can forecast the likelihood of loan defaults based on borrower characteristics and loan-related information.

Machine Learning: Machine learning is a branch of artificial intelligence (AI) that enables computer systems to learn from data without explicit programming. Instead of relying on rigid rules, machine learning algorithms identify patterns, make predictions, and improve their performance over time by analyzing and adapting to data. This learning process involves identifying relationships between input features (e.g., borrower characteristics) and output variables (e.g., loan default status), allowing the system to make accurate predictions on new, unseen data. In essence, machine learning empowers computers to perform tasks that traditionally require human intelligence, such as recognizing patterns, making decisions, and solving complex problems. This capability is particularly valuable in fields like finance, where large volumes of data and complex relationships can be effectively analyzed by machine learning algorithms to predict outcomes like loan defaults.

Supervised Learning: Supervised learning is a powerful paradigm within machine learning where algorithms learn from labeled data, meaning each data point in the training set is paired with its corresponding correct output or label. This approach mimics a guided learning process, where the algorithm is provided with examples and their known outcomes, allowing it to identify patterns and relationships between input features and the desired output. In the context of this thesis, supervised learning is employed to predict loan defaults. The algorithm is trained on a dataset where each borrower's information (input features) is linked to their actual loan repayment status (label), indicating whether they defaulted or not. By analyzing this labeled data, the algorithm learns to identify the characteristics and patterns that distinguish defaulters from non-defaulters. This learned knowledge is then used to predict the likelihood of default for new, unseen borrowers. The trained model analyzes the input features of a new borrower and assigns a probability of default based on the patterns it has learned from the labeled data. This predictive capability is crucial for credit risk assessment, enabling lenders to make informed decisions about loan approvals and risk mitigation strategies.

Classification: Classification, a fundamental task in supervised machine learning, aims to categorize data points into predefined classes based on their inherent characteristics, much like sorting objects into distinct bins based on their properties. In your research on loan default prediction, you are employing binary classification, where the objective is to assign each borrower to one of two classes: "default" or "no default." This process involves building a system that analyzes a borrower's profile and predicts which category they are more likely to fall into—the "likely to default" or "unlikely to default" bin. This categorization relies on identifying patterns and relationships from historical data, where the algorithm learns from a dataset of borrowers with known default outcomes. The key elements in this classification process include the predefined classes ("default" and "no default"), the features or attributes of the borrowers used for classification (such as age, income, and credit history), the classifier (the algorithm performing the classification), the training data (labeled data with known outcomes used to train the classifier), and the test data (unseen data used to evaluate the classifier's performance).

Ensemble Learning: Ensemble learning is a powerful machine learning paradigm where multiple individual models are strategically combined to create a more robust and accurate predictive model. Imagine it as a team of experts with diverse skills and perspectives coming together to solve a problem, where the collective intelligence of the team surpasses the capabilities of any individual member. In this approach, various base models, each with its own strengths and weaknesses, are trained on the data, and their predictions are aggregated to produce a final prediction. This aggregation process can involve averaging the predictions, weighted voting, or more complex techniques like stacking, where a meta-model learns to combine the predictions of the base models. The key advantage of ensemble learning lies in its ability to reduce bias and variance, leading to improved generalization and more reliable predictions. By leveraging the diversity of multiple models, ensemble learning can effectively capture complex patterns in the data and handle noisy or incomplete information, ultimately enhancing the accuracy and stability of the predictive model. This technique is particularly beneficial in challenging prediction tasks, such as loan default prediction, where the accuracy of the model can significantly impact financial decisions and risk management strategies.

Key Concepts in Model Evaluation: AUC-ROC F1-Score

Evaluating the performance of machine learning models is essential for understanding their predictive capabilities and selecting the most appropriate model for a given task. In the context of loan default prediction, two key evaluation metrics are commonly employed: AUC-ROC and F1-score. AUC-ROC, or the Area Under the Receiver Operating Characteristic Curve, measures the model's ability to distinguish between classes. It considers the model's performance across all possible classification thresholds, providing a comprehensive assessment of its discriminatory power. A higher AUC-ROC value indicates better classification performance, with a value of 1 representing perfect discrimination and 0.5 indicating random guessing. In contrast, the F1-score is a balanced measure of a model's accuracy that considers both precision and recall. Precision reflects the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall, also known as sensitivity, measures the proportion of correctly predicted positive instances out of all actual positive instances. The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both considerations. The F1-score is particularly useful when dealing with imbalanced datasets, where one class might be significantly more prevalent than the other.

Chapter 3

Related Work

3.1 General challenges addressed by the literature

This research, informed by the literature analysis, identified and analysed the diverse issues financial institutions encounter in assessing borrowers' loan default risk. The problems encompass data quality and availability, model selection and evaluation, borrower characteristics, and the impact of economic conditions. A significant challenge in predicting loan defaults is the quality and availability of data. Many studies emphasise the issue of excessive dimensionality and the class imbalance inherent in the bank loan datasets utilised for predictive training, which adversely affects numerous models' performance. information The pre-processing issues highlighted by Kou and Wang involve managing high-dimensional data and imbalanced classes, which might result in biassed predictions if inadequately addressed [9]. This is due to several financial organisations possessing restricted and occasionally erroneous data regarding borrowers and their repayment behaviours. Thus, models developed on limited or unrepresentative data may fail to generate dependable predictions regarding real-world settings, as demonstrated by the models evaluated in this paper [10]. A significant issue is the selection of a suitable model and its evaluation. Decision trees, random forests, and logistic regression models, among others, have been proven as effective tools for assessing loan default rates. Their performance is contingent upon various aspects, including the particular data set utilised and the attributes of the data. Zhou illustrates the application of various machine learning models in predicting loan defaults and notes that the efficacy of particular algorithms may fluctuate based on specific conditions [11]. This unpredictability presents a problem in selecting the most suitable model for a certain data set, a process that is sometimes complex and may necessitate considerable time and effort. The traits of borrowers also present a challenge in evaluating loan default risks. Income level, credit history, and loan purpose have been recognised as significant determinants of default rates. Serrano-Cinca et al. assert that these borrower characteristics are essential to integrate into the models to improve their accuracy [12]. The interconnections among these factors can be complex, and models that do not consider the subtleties of borrower behaviour may yield erroneous outcomes. Turiel and Aste discovered that dependence exclusively on conventional metrics, such as loan grade and interest rate, may fail to encompass the complete scope of borrower risk, resulting in potential biases in predictions [13]. Economic conditions further complicate the prediction of loan defaults. The research indicates that factors such as unemployment rates and economic development significantly affect borrowers' repayment behaviours. Devi and Radhika emphasise the necessity of considering these economic considerations when formulating models that can adjust to market conditions [14].

The incorporation of new technology, including artificial intelligence and big data analytics, offers both benefits and challenges. The technologies that enhance the prediction capabilities of models also present issues regarding interpretability and transparency. Wang elucidates that big data can improve loan default predictions; however, he also notes that these models tend to be intricate, complicating the stakeholders' understanding of the decision-making process [15]. The absence of transparency can undermine trust in automated systems, especially in the financial industry, where decisions may have substantial repercussions for both borrowers and lenders. Furthermore, the issue of mitigating information asymmetry in lending practices is a persistent topic in the literature. Numerous studies, like those conducted by Andrianova et al., underscore the challenges banks encounter in identifying opportunistic borrowers who may lack the will to repay loans [16]. This difficulty is exacerbated in markets with restricted access to dependable credit records, hindering lenders' ability to appropriately evaluate borrower risk.

3.2 Datasets used in loan default prediction studies

The datasets employed in analysing loan default prediction also vary considerably based on the scope of the research, the kind of loans in question and the techniques used. When reviewing the literature it can be seen that the size of the dataset used can vary from few data sets to large data sets that include millions of data records.

Turiel and Aste used a large data set from LendingClub which includes 16 million rejected loan applications and 1. 6 million accepted loans to make about 15 million loans in total. After data cleaning the authors applied their analysis on 600000 fully paid loans and 150000 defaulted loans, which is about 15-20 of the total loans issued. The data set was divided by time where one set of loans was used to train the model while the other was set aside for testing purposes. The data set used in the study contained a large number of features such as borrower's characteristics and loan characteristics which facilitated the creation of a powerful logistic regression model for assessing the likelihood of loan default [13]. These large data sets are now becoming the norm especially when it comes to big data analysis and loan performance historical data from online lending platforms.

Therefore, some researchers have used smaller data sets. The data set considered in the study by Koutanaei et al. was obtained from the "Export Development Bank of Iran" over a two year period. It was based on legal customers and comprised an initial set of 1100 records with 59 distinct features [17]. Owusu et al. used a data set of defaulters of loans got from Kaggle which was unbalanced and had 890 thousand data points and 75 features [18] This shows that there may be a preference for using smaller data sets by researchers especially when dealing with certain aspects of loan default prediction such as the effect of data imbalance on the model.

Furthermore, the research conducted by Kohv and Lukason was based on corporate bank loan defaults, thus underlining the importance of variable domains in identifying defaults. The cross-sectional research was based on a longitudinal data set of 12,901 observations of an Estonian commercial bank's corporate loans with and without default. The data set included 12 variables of 3 types to monitor the different aspects of loan default prediction [19]. This indicates that perhaps the emphasis is more on the data being accurate and relevant rather than the size of the data.

Also, different research works have employed different types of data sets to test the effectiveness of the proposed models for LDP. For example, the Spanish Consumer Credits dataset used by Alonso and Carbó included 75,000 credit operations with 370 anonymized features of which the main focus was on point-in-time defaults without using temporal data [20]. Similarly, the Benchmark Credit Scoring dataset used by Rahmani et al. used borrowers and loan characteristics data that was made publicly available; the data was preprocessed using techniques such as Weight of Evidence to deal with class imbalance [21].

The Give Me Some Credit dataset was obtained from Kaggle and used by Bahnsen et al. and included variables that are specific to borrowers such as income and revolving balance; although it was rather small, it was also somewhat unbalanced [22]. Other datasets used their own specific data sets, for example, the Six-Bank dataset contained consumer tradelines and macroeconomic data for the duration of the model's development [23]. Furthermore, the Brazilian Bank dataset which has more than 711,000 records, and the Home Credit dataset which has rich loan-specific features shows the use of large data for modeling the default behavior [24]. Thus, these various datasets depict the differences in size, format, and feature richness, which allows the authors to apply the most suitable approaches to the problem of credit default risk management.

Finally, the research conducted by Barbaglia et al. used a data set of 12 million residential mortgages to analyze the behaviour of loan defaults across several European countries. This shows that large-scale analyses are possible and may produce conclusions about the behavior of defaults while taking into account the effects of borrowers' characteristics and regional economic factors [25].

3.2.1 Dataset Characteristics

The temporal features of the datasets are important as the loan performance can be vulnerable to economic conditions over time. Barbaglia et al. used a dataset of 12 million residential mortgages to analyze the default behavior of the borrowers over different economic cycles [25]. Thus, by integrating such factors as the loan origination date and the payment history, the models can be made more efficient in predicting trends and changes in the behavior of the borrowers.

The studies on loan default prediction can be broadly categorized into those that use structured data only and those that incorporate unstructured or multimodal data with increasing frequency, especially in the peer-to-peer lending and consumer behavior domains. The majority of the data sets for loan default prediction are structured which means that they are well-organized and have clearly defined variables that are numerical or nominal. The most common variables are demographics data of the borrowers (age, income, employment status), general information about the loan default). (amount, Some terms, research interest have rate), begun and to credit engage information in (credit the score, analysis history of of unstructured data particularly the textual data that is extracted from the loan applications or the borrower's statements [12] [26]. Netzer et al. focused on determining the signs of the loan default by analyzing the text from the loan applications; thus, the authors proved that textual data could enhance the predictive power [27]. Furthermore, Jiang et al. used soft information in the form of descriptive language in online peer-to-peer lending to improve the default risk assessment, the use of unstructured data with traditional predictive models [28].

An important point is the use of multimodal data that includes both, the structured and unstructured data as discussed in the current literature. Niu et al. studied the integration of the social network information with the conventional borrower information to improve the predictive models in the context of peer-to-peer lending [29]. This method depicts the growing trend of using diverse data to enhance the prediction of loan default.

The representativeness of the dataset used in the study is critical in order to extend the findings of the study to the general population. Shetty and Vincent draw attention to the need for more sensitive models because of the costly Type I errors that may occur when the data is not a good depiction of the target population. Consequently, the models developed in such situations may not perform well when applied in real life environments [30]. Also, the reliability of the data sources is an important importance aspect of of using data quality quality data as that discussed is earlier. collected Jumaa from et credible al. surveys highlighted and the well-established lending platforms to develop more accurate predictive models as done by the authors [31]. Data from a credible source is usually accurate and can enhance the overall quality and cohesiveness of the predictive analysis.

3.3 Preprocessing techniques

Pre-processing of data is very crucial in the application of machine learning (ML) for forecasting loan defaults to enhance the effectiveness of the prediction models. The literature survey reveals that there are numerous preprocessing techniques which are commonly used and some of them have been found to be effective in dealing with some of the challenges that are common in the data such as class imbalance, noise and feature selection.

The efficiency of these preprocessing techniques has been backed up by many research works. Its important to note that, the choice of machine learning algorithms is also very crucial in the construction of loan default prediction models. The recent developments have shown that the ensemble methods such as Random Forest, Gradient Boosting and XGBoost have outperformed traditional models like logistic regression and decision trees [32]. These algorithms take advantage of several models to increase the accuracy and robustness of the predictions. XGBoost is known to have a good data handling capacity and is able to learn complex features from the data [33]. Also, deep learning approaches have been used lately, especially for high-dimensional data, because they can automatically learn feature representations [31]. Thus, the complexity of these models requires a careful fine-tuning and optimization to avoid overtraining especially in unbalanced data sets.

3.3.1 Oversampling and Undersampling

A challenge that is common in loan default, is the challenge of data imbalance where there are many non-defaulting clients than defaulting clients. Wang et al. found out that oversampling especially the SMOTE technique is useful in increasing the efficiency of classification models especially where data is imbalanced as in this study [34]. Chen et al. used various sensitivity undersampling to reduce the large class while keeping important instances of the minority class and found it effective in enhancing the model performance [35]. These strategies are very vital in ensuring that the models do not biased towards the majority class as is the case with loan default datasets.

3.3.2 Normalization and Standardization

Normalization and standardization of data are also very common preprocessing methods. The integration of preprocessing techniques, including normalization and scaling has been identified to improve model performance by making sure that features are on the same scale during distance calculations in algorithms for instance KNN and SVM [9] [36]. Some of the commonly used normalization techniques include the Min-Max scaling and Z-score normalization which are used to transform the data into a form that is suitable for model training and improve the convergence rates and overall predictive accuracy [36]. Numerous studies have pointed out that scaling features are crucial in order to facilitate the proper functioning of the optimization algorithms employed in machine learning [37] [38]. The above techniques assist in standardizing the data to a certain scale thus reducing the effects of outliers while supporting the stability of the learning process

3.3.3 Handling Missing Values

Missing values are a common occurrence in real-world datasets and their handling is an important step in the data preprocessing process. Missing values and data cleansing improve the reliability of the predictions models by reducing the risks of biases and errors in the results [39]. Several approaches such as the imputation mean, methods median, namely mode imputation and deletion methods are used to deal with the missing data. The effectiveness of such measures can greatly influence the predictive power of such models. Babo and Beyene stated that missing data should be properly managed to ensure the quality of datasets and achieve accurate prediction [40].

3.3.4 Data Cleaning

Data cleaning refers to the process of identifying and correcting or removing wrong or incomplete data. Techniques such as outlier detection and removal, as well as the correction of erroneous entries, are commonly applied. According to Alreshidi et al., the lack of well-defined preprocessing protocols makes it difficult to compare the performance of various machine learning models across the studies [41] [42]. Hence, it is crucial to ensure that only high quality datasets are derived after cleaning the data carefully.

3.3.5 Feature Selection and Extraction

Feature selection and extraction is one of the most important preprocessing steps which is used for identifying the best set of criteria that can be used to predict loan defaults. The literature shows that proper feature selection can greatly enhance the performance of machine learning models [43]. Feature selection has become a popular topic in the loan default prediction domain. It also becomes more crucial for developing appropriate models for high dimensional data sets to avoid overtraining and increasing the expenses of computations.

One of the most commonly used feature selection methods is the filter method, which evaluates the relevance of features based on their intrinsic properties, independent of any machine learning algorithm. There are various techniques used in this category and they include correlation coefficients, Chi-square tests and information gain. For example, correlation coefficients enable us to determine which features are linearly related to the target feature and hence which features can be safely removed as they do not help the prediction task much [44]. The filter method is computationally efficient and is particularly useful in scenarios where the dataset has a large number of features, as it can quickly eliminate irrelevant features before applying more complex models [45].

Another prevalent approach is the wrapper method. The wrapper method assesses feature subsets using the performance of a learned machine learning model. This approach is even more computationally intensive than filter approaches but it often leads to better results because it also considers the interactions between features. Techniques such as recursive feature elimination (RFE) are commonly used in this category. RFE works by recursively removing the least important features based on the model's performance until the optimal feature subset is identified [44]. The wrapper method is most useful in situations where there are many complex interactions between features as it provides a better way of selecting models based on total model performance better capturing non-linear relationships. Approaches like Recursive Feature Elimination and tree-based feature importance measures have been employed to identify the most relevant features contributing to loan default predictions [46] [47]. Studies have also revealed that applying feature selection techniques can enhance the performance of algorithms for instance Random Forest and Gradient Boosting Decision Trees if only the most important variables are considered [48] [32].

Embedded methods combine the advantages of both filter and wrapper methods by incorporating feature selection into the model training process. Algorithms such as Lasso (L1 regularization) and decision tree-based models inherently perform feature selection during their training phase. For instance, Lasso regression penalizes the absolute magnitude of the coefficients, effectively driving some of them to zero. This results in a sparse model that retains only the most significant features [49]. Such methods reduce dimensionality while reducing the risk of overfitting, making them particularly effective for high-dimensional datasets.

In addition to these traditional methods, **metaheuristic approaches** such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) have gained traction in feature selection tasks. These algorithms mimic natural processes to explore the feature space and identify optimal subsets. For example, GA can be used in the following manner to evolve a population of feature subsets over generations and select the best ones based on the highest model performance [50] [51]. Such methods are particularly useful for large and very complex datasets since traditional approaches can be inadequate for determining the appropriate set of features due to the issue of high dimensionality and possible interactions between features.

Furthermore, there are also **hybrid strategies** that use combinations of different feature selection techniques. These methods take the benefits of a number of approaches to enhance the stability and accuracy of feature selection. For instance, a combination of filter and wrapper methods can be applied in the first place to reduce the number of features and then the selection is refined using a model-based approach [52]. This strategy may enhance performance, particularly in areas where the interactions among features are complex and diverse.

3.4 Using domain-specific features

Domain-specific characteristics, such as loan length and credit grade, play a key role in boosting the predictive accuracy of machine learning models used for loan default prediction. These attributes encapsulate critical information that indicates the financial behavior and risk profile of borrowers, hence greatly influencing the model's capacity to make accurate predictions.

Domain-specific features are customized to the distinctive attributes of the financial sector, offering insights that generic features would neglect. For example, the loan duration impacts the repayment plan and the possibility of default. In general, longer terms may increase the risk because the financial condition of the borrower may change over the course of the loan [25]. Likewise, credit grade is a composite measure of a borrower's creditworthiness that compiles various factors including past credit record, outstanding debt, and payment habits [53]. The use of these attributes enables models to utilize domain knowledge, potentially enhancing their efficacy in forecasting loan defaults.

Integrating domain-specific information helps both predicted accuracy and the

interpretability of machine learning models. Financial institutions frequently necessitate models that help elucidate their projections, particularly in critical situations such as loan approvals [54] [55]. By leveraging features that are well-understood in the financial environment, such as credit rating and loan term, stakeholders can better appreciate the logic behind model projections. This interpretability is essential for regulatory compliance and for sustaining trust between borrowers and lenders [55] [54].

A multitude of studies has evidenced the **efficacy of domain-specific features** in enhancing prediction accuracy. In a comparison examination, models employing domain-specific features attained superior accuracy rates and reduced error margins, highlighting the significance of these factors in elucidating borrower behavior nuances [11]. Furthermore, the inclusion of these variables has been found to increase the generalization of the model when trained on datasets that are different from the one used for training, thus making the model less likely to overfit to the specific training dataset [56].

The effectiveness of domain-specific features is often amplified when combined with advanced machine learning techniques. For instance, ensemble approaches like Random Forest and Gradient Boosting can efficiently use the information contained in these characteristics, leading to significant gains in predicting performance [57] [58]. Moreover, employing feature selection methods that emphasize domain-specific attributes might augment model accuracy by concentrating on the most pertinent predictors [58]. This collaboration between domain expertise and machine learning algorithms underscores the necessity of including expert inputs into the modeling process.

Despite the advantages of adopting domain-specific features, there are problems connected with their integration. A notable concern is the **risk of overfitting,** particularly when the dataset is limited or when the characteristics exhibit high correlation [59]. It is therefore important to undertake careful feature engineering and selection to manage this challenge. Furthermore, due to the fact that the nature of financial markets is constantly evolving, it makes features relevant from time to time, thus making it important to review the features periodically [25]. Therefore, although the concepts from the domain expertise are very useful, they have to be used with a lot of attention and precision.

The **usage of domain-specific features** in machine learning models for loan default prediction is likely to be refined further. Since more and more financial institutions are already adopting the use of advanced analytics, it will be important for the creation of more refined features that are capable of recognizing the trends as well as the behaviors of the borrowers [25]. Also, the use of real-time data and other types of data such as social media data or transaction data may provide better assessment of risk profile of borrowers [29]. This shows that there is the need for the data scientists and domain experts to work in tandem to ensure that the models are relevant and effective at all times.

3.5 Dataset Splitting methods

The following are the most prevalent methods used in order to split the data for assessing model performance in the context of machine learning for loan default. The most common methods include **train-test splits**, **k-fold cross-validation**, **temporal splits**, and **grouped k-fold or grouped splits**. Each method has its unique advantages and challenges i.e. how it handles the data and the implications for model evaluation.

3.5.1 Train-Test Split

This is the simplest and most commonly used method, where the dataset is divided into two subsets: It comprises of a training set that is used to train the model and a test set used to determine the performance of the model. In general, a rather conventional split is 70-30, where 70 of the data is used for training while the remaining 30 is for testing [60] [61]. Nevertheless, it is quite easy to understand and apply, and thus may cause high variance in model evaluation; the performance may vary significantly based on how the data is split.

3.5.2 K-Fold Cross-Validation

his method is especially useful when the data set is organized into classes or categories for example loans made to the same borrower or loans given by the same financial institution. Grouped k-fold makes a guarantee that the same group is not contained in the training set and the test set, which is important for avoiding correlation in the observations [62] [61]. This method reduces overfitting and is especially helpful in identifying the proper performance measure for models especially where the same entities may have related outcomes.

3.5.3 Grouped K-Fold or Grouped Splits

This method is particularly relevant when the data contains groups or clusters, such as loans issued to the same borrower or loans from the same financial institution. Grouped k-fold ensures that the same group does not appear in both the training and test sets, which is crucial for maintaining the independence of observations [63] [61]. This method helps to prevent overfitting and provides a more accurate assessment of model performance, especially in datasets where the same entities may have correlated outcomes.

3.5.4 Temporal Splits

In situations where time plays a significant role, temporal splits are especially useful. This method involves splitting the data based on time, ensuring that training data precedes testing data. Temporal splits enhance the assessment of time-dependent predictions since the data is split chronologically such that the model is trained on historical data and then evaluated on the future data, which is very important in financial applications like loan default prediction. By employing temporal splits, the model is assessed in a way that is akin to the real world, where the lenders are required to act based on the historical information about the borrowers without the access to future data [61] [25]. The main advantage of the temporal splits is that they do not allow data contamination which is a common issue that occurs when future information is included in the training set. This contamination can produce performance measures that are quite high and cannot be extended to new data samples. Since the data is arranged in chronological order, temporal splits preserve the integrity of the evaluation process by restricting the training data to be prior to the test data [61] [64].

3.6 Classification vs. Regression in Loan Default Prediction

3.6.1 Classification

In the context of loan default prediction, the problem is typically framed as a **classification problem** rather than a regression problem. This is mainly because of the nature of the outcome variable which in most cases is a binary one; the borrower is either a defaulter or a non-defaulter. This binary classification is in harmony with the aims of financial institutions to group the loan applicants into different risk groups in order to make proper decisions on lending. Mainly, it is to give a probability score which measures the possibility of default, which can then be used to make a binary decision on whether to approve the loan or not [12]. Most studies, such as those by Liu et al. and Lessmann et al., have emphasized the classification approach, where various machine learning algorithms, including logistic regression, decision trees, and support vector machines, are employed to predict the likelihood of loan defaults [65] [33]. The main goal is to assign a probability score that measures the likelihood of default, which can then be thresholded to make binary decisions regarding loan approvals.

3.6.2 Regressions

Classifying loan defaults is the most common objective of the models, although one can also use regressions to estimate the probability of default. Of all the models, logistic regression is the most often used and preferred method for its ease of interpretation. Since the coefficients produced by logistic regression are easily convertible into odds ratios, it is easy to see how changes in the predictor variables affect the odds of default. This interpretability allows stakeholders to understand how changes in predictor variables can affect the likelihood of default [66] [67]. This interpretability is a sought after feature in the financial sector, where decision-makers must justify their choices based on model outputs.

Another advantage is that regression models can effectively handle continuous predictor variables, such as income levels or loan amounts, which can provide nuanced insights into borrower behavior. For instance, studies have shown that larger loan amounts are associated with a higher probability of default, as indicated by a regression coefficient that reflects this relationship [68]. This capability allows for a more detailed analysis of the factors influencing loan defaults, which can inform risk management strategies.

Another advantage is that regression models can be useful when the focus is on the probability of default instead of merely labelling borrowers. Since regression models provide a continuous output, they can assist financial institutions to determine the risk level of different borrowers thereby facilitating appropriate lending decisions [15].

However, there are a number of potential pitfalls when applying regression models for measuring default risk. The major challenge that is likely to be encountered is the assumption of linearity of the relationship between the log-odds of the outcome and the predictor variables. However, if the relationship is non-linear then the logistic regression may not be able to explain the data well resulting to poor predictions [64] [69]. This limitation can be more evident in financial datasets where the relationships between variables may be complex and interconnected. Another problem is the multicollinearity of the explanatory variables, which increases the variance of the coefficients and makes it difficult to assess the influence of a particular variable on the probability of default [25]. This problem can lead to complications in the analysis of results and may require further data processing, for example, feature selection or dimensionality reduction to decrease the effects of multicollinearity.

Also, regression models may not be able to deal with imbalanced data set which is usually the case in loan default prediction where the number of defaulters is usually much lesser than the non-defaulters. In such cases, the regression models may produce estimates that are biased towards the majority class and may underestimate the risk that is posed by the smaller class [35]. Some of these include resampling or cost-sensitive learning and while applying these techniques, there is a need to complicate the modeling process.

Finally, traditional regression models can only provide useful information and may not be able to identify all the patterns in the data that other more advanced models like the ensemble models or neural networks can. Hence, relying on regression models alone can restrict the predictive capacity of the loan default prediction systems to some extent [70]. Thus, even though regression models can be used for identifying the default probabilities, these methods are supplemented with more elaborate machine learning approaches to improve the overall predictive power.

3.7 Machine Learning Models

In the domain of loan default prediction, many ML models have been used for the prediction of loan default, each having its own advantages and disadvantages. The model choice varies with the nature of the data set as well as the needs of the prediction problem. The choice of model often depends on the specific characteristics of the dataset and the requirements of the prediction task. As the field continues to evolve, the integration of novel approaches and hybrid models is likely to produce further improvements in predictive performance. This section gives an overview of the most common models, and distinguish them as linear, non-linear, tree-based, ensemble, gradient boosting, neural networks, hybrid models, and novel approaches.

3.7.1 Linear Approaches

Logistic regression is a widely used method for credit scoring and loan default prediction and often serves as a benchmark for other machine learning methods. Although logistic regression has some benefits such as simplicity, interpretability, and the capability of giving the probability of an event occurring, it has some drawbacks such as; it assumes linearity, cannot capture complex feature interactions and is prone to overfitting. Nevertheless, the logistic regression model is one of the most effective in practice and is still relevant today, and sometimes only a slight improvement in the predictive accuracy is observed, when using more complex models, for example, random forest or XGBoost. Logistic regression models can be tuned with feature selection and regularization methods and their performance quantified by various metrics. Also, calibration and explainability techniques can improve the effectiveness of the model in real life. Hence, logistic regression is one of the basic techniques used in credit scoring and loan default risk assessment. [71] [72] [73].

3.7.2 Non-Linear Approaches

Support Vector Machines (SVM): While the basic SVM model works for linearly separable data, real-world data often exhibits non-linear relationships. SVMs overcome this by employing kernel functions which takes the data in to a higher dimensional space where the data can be linearly separated [74]. The sources mention several kernel functions, for instance, linear kernels that give the best results for linearly separable datasets, polynomial kernels that are commonly used in credit scoring and classification and radial basis function kernels which is used for non-linear data [74]. SVM model has several hyperparameters which have a big impact on the performance of the model, including the regularization parameter and kernel parameters. These parameters have to be tuned to optimize the performance of the model [56] [74].

The sources also reveal how SVMs have been employed in credit scoring and other pertinent domains including credit scoring algorithms, hybrid credit score models, clustered SVMs, credit scoring optimization, feature selection, and reject inference [56] [74]. Based on comparative studies, it has been observed that SVMs produce better results than other models, including the ensemble methods though deep learning models have also been found to produce good results. Some of the benefits of SVMs include the ability to deal with noisy data and work in high dimensional spaces; however, they are costly from the computational point of view and are black-box models [75]. There are techniques for extracting rules and explainability methods that can be used to increase the interpretability of SVM models. Although

SVMs have certain advantages, they are limited by their high computational costs and semi-black-box approach; therefore, the sources also stress on the importance of parameter tuning as well as the development of methods that would increase the interpretability of SVM models [75].

3.7.3 Tree-Based Approaches

Decision Trees: Decision trees are intuitive models that split data based on feature values to make predictions [76]. Decision trees have been applied to credit scoring and loan default prediction but they have a poor performance for measuring the exact probability of default. Several results indicate that decision trees are one of the most effective classification algorithms for credit scoring, but they can provide low accurate estimates of default probabilities. However, decision trees can serve as foundational blocks for more complex ensemble models, and use their interpretability to make them useful for understanding the factors contributing to loan defaults [17]. Overall, research indicates that while decision trees have their place in credit risk assessment, other models or ensemble methods may be more effective for accurately predicting loan defaults [77].

Random Forest: Random Forest (RF), an ensemble of decision trees, has gained popularity for its robustness and accuracy. Oghenekaro implemented a Random Forest algorithm to predict loan defaults, achieving high accuracy rates [6]. The model's capability to work with big data and the feature importance ranking that it provides makes it efficient. The authors also show that Random Forests is an extremely useful ML model that can be used effectively for detecting loan defaults. RF models are types of ensemble learning algorithms which are based on decision trees and are capable of providing more detailed information regarding the probabilities of default, outperforming the conventional credit scoring models such as logistic regression. RF models are ensemble learning methods based on decision trees that can provide more granular insights by estimating default probabilities, outperforming traditional credit scoring models like logistic regression. The literature has identified several benefits of RF, such as their capability to deal with big data, their non-linearity and non-parametric nature, and their ability to combine bagging and boosting techniques for improving the predictive performance [56]. Research findings reveal that RF models provide high levels of accuracy in their predictions and outperform other machine learning algorithms as well as the conventional credit scoring models [20]. However, interpretability is an issue of concern as are the other sources emphasized on the predictive capability and the flexibility of Random Forests making it effective for credit risk assessment and loan default prediction [69].

3.7.4 Ensemble Approaches

XGBoost: XGBoost is an extended version of gradient boosting algorithm that is now used in many ML competitions due to its efficient performance. For instance, Nguyen stated that XGBoost provided a correct prediction for more than 80 of the loan defaults, thus proving its efficiency and efficacy in dealing with large data sets such as the ones used in this study [78]. However, the current literature has some inconsistency when reporting the performance of XGBoost in loan default prediction. Due to its performance, XGBoost has been known to produce good results in loan default prediction models, outperforming other models such as logistic regression, decision trees among others; however, other sources claim that its dominance is not that clear. Certain research revealed that XGBoost provided results that were as accurate as other classifiers such as Random Forest [32]. However, some sources have pointed out that there are other gradient boosting algorithms that may outperform XGBoost depending on the specific dataset and problem setting such as LightGBM [33].

CatBoost: CatBoost is a gradient boosting algorithm which was proposed by Dorogush et al. It is specifically built to work well with categorical data without the need for any special handling or preprocessing. It implements ordered boosting which aids in minimizing overfitting and enhances the model's ability to make predictions [79] [80]. This makes CatBoost particularly suitable for datasets with a significant number of categorical variables, which are common in financial applications. Xia et al. utilized CatBoost in their study on predicting loan defaults in peer-to-peer lending, integrating both hard and soft information into their credit scoring model [81]. The algorithm's ability to manage categorical data directly contributed to improved predictive accuracy, demonstrating its effectiveness in this domain. When compared with other boosting algorithms, CatBoost has been seen to provide comparable results. According to Guo and Zhou, CatBoost outperformed the traditional models in terms of the accuracy and stability in personal loan default prediction [82]. The efficiency of the algorithm to deal with categorical features without much preprocessing has been a major benefit which has resulted into better model performance [83]. The available literature provides a mixed perspective on the effectiveness of CatBoost for predicting loan defaults. Some studies highlight CatBoost's strengths, including its ability to handle heterogeneous data and its success in blended models. However, some sources have pointed out that the performance of CatBoost can be situational and it does not always beat other algorithms such as LightGBM and XGBoost [84]. It can be seen that the performance of CatBoost relies on parameters such as the dataset, features used, and the hyperparameter tuning [83]. Future research directly comparing CatBoost with other gradient boosting models and the study of the effectiveness of feature engineering in CatBoost for loan default prediction.

LightGBM: LightGBM, developed by Ke et al., is another gradient boosting framework that is optimized for speed and efficiency. It uses a histogram-based approach to bin continuous features, which significantly reduces memory usage and increases training speed [85] [86]. LightGBM is particularly effective for large datasets, making it a popular choice for financial applications. LightGBM has been widely applied in loan default prediction studies. LightGBM LightGBM is has especially been effective extensively for used big in data, loan and default hence prediction it and is the frequently following applied are in some financial of the works that have applied it in their research. For instance, Ma et al. applied LightGBM in LightGBM in their study on P2P lending, and they found out that LightGBM performed well with high-dimensional data and produced higher accuracy than the traditional models. The study also reveals that LightGBM is superior to XGBoost especially when Multi-Observation data cleaning technique is employed [33]. The structure of the algorithm makes it suitable for dealing with large sets of features, which is important in financial models. LightGBM has been frequently used in the experiments on loan default prediction and showed great results. Dong states that LightGBM gave an AUC of 0.73 in an experiment done using a lender dataset from the Tianchi Platform, thus proving its efficiency in predicting loan defaults. The analysis indicates that LightGBM is efficient and accurate and thus can be useful to financial institutions in the management of loan risks [73].

3.7.5 Hybrid and Novel Approaches

Here we will discuss neural networks, their potential and current usage in the field of machine learning as well as the difficulties and issues related to the application of neural networks in credit scoring and financial risk assessment. Some of the research presented in the sources show how neural networks can be applied in credit scoring and other similar tasks. For example, one study revealed that a neural network model provided a better results than decision tree and logistic regression models in terms of the prediction of credit data [87].

The sources also recognise the capability of neural networks to learn non-linear relationships in data set, making them a good candidate for credit risk scoring models where the credit worthiness of an applicant is a function of several variables [88]. However, the same sources that report these positive outcomes, also discuss some of the drawbacks of neural networks. A significant disadvantage is the opacity of the neural networks for which a common term is 'black box'. The complex internal workings of neural networks make it difficult to understand how they arrive at their predictions, which poses a challenge in credit scoring where explainability is crucial for regulatory compliance and building trust with customers. [89] Overall, the sources present a nuanced view of neural networks, recognizing their potential in credit scoring and related domains while acknowledging their limitations and the ongoing research aimed at addressing them [90] [91]. The above sources also emphasize on the fact that there is a challenge and opportunity between predictive analysis and interpretable models, and different approaches have been put in place to achieve the right balance. Neural networks are most probably set to become even more important in credit scoring and risk appraisal as data complexity is expected to increase in the future, nevertheless, the issue of transparency and interpretability of the results will have to be addressed [90].

CNNs have been used in loan default prediction especially when other types of data such as textual data from the loan applications are used. Zhu et al. **combined CNN with LightGBM** and achieved better results than the classical models, which proves the effectiveness of the deep learning methods in this problem [64].

Hybrid models that combine multiple algorithms have been developed to leverage the strengths of different techniques. For instance, the combination of Random Forest and XGBoost has been tried as techniques to improve the prediction. According to Zeng the combined model was superior to the lone models indicating that the integrated approach is effective [92].

Graph Neural Networks (GNN) have received significant attention lately. One of the sources discusses a new model of **Dynamic Multilayer Graph Neural Network (DYMGNN)** that uses GNNs along with Recurrent Neural Networks for credit risk evaluation with borrowers' connections as factors affecting risk propagation. DYMGNN develops a series of multilevel network snapshots in which each level signifies a different kind of borrower connection, including geographical, mortgage, and other characteristics. The model implements GNNs to learn the structural relations between nodes and, therefore, represent borrower structure and features while RNNs are employed to capture temporal dynamics of borrowers' behaviour. There is also an attention layer used in order to distinguish the more relevant time steps. The experiments evidence that the proposed DYMGNN model is capable of forecasting loan default, and thus outperforms other baseline models, as it is able to capture the dynamics and the multiple layers of the network for the analysis of credit risk dynamics, as opposed to other models that only consider characteristics of the individual borrower [93].

3.8 Explainability in Credit Risk

Explainability in Financial Models: The concept of Explainability in Financial Models: There are several reasons as to why explainability is crucial for financial models especially in loan default prediction. First, financial institutions are in a position whereby they make decisions that can have a great impact on the borrowers' lives for example through loan approval or rejection. Hence, it becomes imperative that the stakeholders of the model are able to understand how the model was able to make such predictions so that the decisions made are appropriate and fair [94]. Second, it also assists in determining biases in the model which is crucial in enhancing fairness and equality in lending institutions. The models which are not clear sometimes become a tool of bias and thus result in unfair lending practices [94].

Improving Stakeholder Trust and Regulatory Compliance: This enhances the stakeholder's trust since it helps to understand how the models work [95]. When the stakeholders, including the regulators, borrowers and lenders understand how the decisions are made, they are more likely to accept the decisions made. This trust is very important in order to build a strong and sustainable business relationships between the financial institutions and their customers [96]. Another advantage of the explainable models is that the institutions can meet the legal requirements since the models provide clear records of how the decisions are made thus minimizing legal consequences [96].

3.8.1 SHAP and LIME

SHAP (SHapley Additive exPlanations) and **LIME (Local Interpretable Model-agnostic Explanations)** are two popular techniques for explaining machine learning model predictions.

Computational Cost: SHAP is generally more computationally intensive than LIME, especially when using the Shapley value calculations, which can require significant resources for large datasets [97] [98]. Of the two, SHAP is often computationally more expensive than LIME especially when using the Shapley value calculations which can be very demanding in terms of resources for large data sets [99]. Nevertheless, the disadvantage is that LIME's explanations are not always as precise or as stable as those produced by SHAP.

Effectiveness: SHAP offers a single way of computing feature importance using cooperative game theory and therefore provides more coherent and stable explanations for different instances [54]. LIME is efficient but the explanations it generates are not always fixed and can be rather arbitrary depending on the local environment of the instance being explained, thus making the results less constant [74]. Therefore, both of the methods are useful, but since SHAP is more robust and has a stronger theoretical background, it is more often applied [97].

3.8.2 Rule-Based Methods

Among these, decision trees have been widely used for the explanation of model's outcomes although they have certain drawbacks, especially when used with today's sophisticated models. Such rules are capable of exaggerating some aspects of the data while at the same time underplaying others. It may not capture the complex correlations between features that other models with multiple features such as ensemble models or neural networks can. This in turn results in over-simplification which may lead to incorrect understanding of the model's behavior [100]. The more complex the model the more difficult it is to create clear and understandable rule-based explanations. For example, a decision tree may grow very large with many branches which may not yield much information [100]. The rule-based methods are not always robust and might not work well with different data sets or different conditions. They are often specific to a particular data set and this may serve to restrict their use to a number of situations [100].

3.8.3 Emerging Trends in Explainability

New Explainability Techniques for Deep Learning Models: As deep learning models become more prevalent in loan default prediction, new explainability techniques are being explored. Techniques such **Integrated Gradients** are being developed to visualize which parts of the input data contribute most to the model's predictions [101] [102]. Such methods help to demystify the "black box" nature of deep learning models, making them more interpretable for stakeholders. To solve the problem of deep learning explainability, Hayashi refers to using rule extraction techniques to extract symbolic rules from trained neural networks, allowing users to gain insights into the model's decision-making process [89].

3.9 Common evaluation metrics

Commonly Used Metrics: In credit risk studies, several evaluation metrics are commonly used to assess model performance, particularly in the context of loan default prediction, these most prevalent metrics include:

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): This metric measures the effectiveness of the model in differentiating between the default and non-default clients. AUC value is between 0 and 1 and higher the value better is the model performance. The ROC curve looks like this: An AUC of 0.5 suggests no discrimination, while an AUC of 1.0 indicates perfect discrimination [103] [104].

Precision-Recall (PR) Curve: This metric is more relevant in the case of class-imbalanced datasets as it emphasis on the performance of the model for the positive class i.e. defaults. Precision is a measure of the proportion of true positive predictions to all all positive predictions made by the model while recall or sensitivity is a measure of the proportion of actual positives that are correctly identified as such. The F1 score which is the harmonic mean of precision and recall is also used quite often to compute the average of the two scores [103] [104].

Accuracy: This metric is more relevant in the case of class-imbalanced datasets as it emphasis on the performance of the model for the positive class i.e. defaults. Precision is a measure of the proportion of true positive predictions to all all positive predictions made by the model while recall or sensitivity is a measure of the proportion of actual positives that are correctly identified as such. The F1 score which is the harmonic mean of precision and recall is also used quite often to compute the average of the two scores [103] [104].

Confusion Matrix: This gives a detailed information of true positive, true negative, false positive and false negative which enable one to analyze the performance of the model accurately [103] [104].

Influence of Imbalanced Datasets: When the dataset is skewed and there are many more instances of non-defaults than defaults, it can affect the selection of evaluation metrics. Thus, accuracy metrics may over-estimate the performance of a model since it allows the model to learn the majority class easily. Hence, AUC-ROC and Precision-Recall scores are used more often, since they give a better picture of the models' ability to classify the defaults [103] [104].

3.10 Hyper-parameter selection techniques

Hyperparameter Tuning Methods: Hyperparameter tuning is crucial for optimizing machine learning models. Common methods found in the loan default literature include:

Grid Search: This method exhaustively searches through a specified subset of hyperparameters. Although it is comprehensive, it can be computationally expensive and time-consuming, especially when there are many parameters to choose from [105] [106].

Random Search: This method samples the hyperparameter space stochastically and while it may be less expensive than grid search, it can be more so if some hyperparameters are more important to the model performance than others. Random search has been proved to give results that are as good as those from grid search but with much reduced computational costs [106].

Genetic Algorithms: These are optimization algorithms which are based on the natural selection process. They can be used for hyper-parameter optimization, especially when there are a lot of parameters to tune. Nevertheless, they can be slower and may need more resources than the grid search or random search [105] [107].

Multi-Objective Optimization Techniques: Multi-objective optimization techniques are very vital in model selection since they enable one to optimize for several objectives at the same time such as accuracy and model simplicity. These techniques can be useful when identifying the set of optimal solutions (Pareto front) of the problem, which is very important in credit risk assessment models as both accuracy and interpretability of the models are crucial [108].

Chapter 4

Methodology

4.1 Problem Definition

This thesis focuses on a major problem of loan default risk assessment in the context of the Greek lending industry. In particular, the study uses Qualco's data to develop and test the models that can assess the probability of the borrowers who are currently paying their instalments steadily, yet are most likely to default within the specified period. This is because this form of early prediction can help in preventing loss of money to lenders and hence help in the stability of the Greek economy which has a long way to go in dealing with non-performing loans. In order to improve the current state of credit risk assessment and to support the development of financial inclusion in the region, this research proposes the application of sophisticated machine learning approaches and the focus on the characteristics of the Greek market.

4.2 Research Questions

This research aims to address critical questions regarding the application of machine learning for loan default prediction within the Greek lending market. Specifically, the study seeks to answer the following:

Can ensemble models outperform standard models in the industry, like Logistic Regression?

This question aims at determining if ensemble learning algorithms that are known to produce highly accurate results across a wide range of machine learning tasks can outperform a more conventional logistic regression model for loan default risk assessment. The following state that logistic regression is the conventional model used in credit scoring and scoring, but it has a number of drawbacks including the inability to identify nonlinear relationships. Some of the ensemble methods that have been employed include AdaBoost and Random Forest which have been seen to outperform logistic regression in aspects such as accuracy. Nevertheless, there is no one method that dominates all others and is deemed to provide the best results; it all depends on the particular problem at hand and the particular dataset used. The sources confirm that logistic regression is the commonly applied model in credit scoring, but it has some constraints in dealing with nonlinear relationships. The sources confirm that logistic regression is the commonly applied model in credit scoring, but it has some constraints in dealing with nonlinear relationships. The sources confirm that logistic regression is the commonly applied model in credit scoring, but it has some constraints in dealing with nonlinear relationships [109] [54] [88]. Lastly, a newer Gradient Boosting Trees model, LGBM, has not been utilized in the context of loan default risk assessment and therefore we plan to use it in this research as well [110]

How do different feature sets and feature engineering techniques impact model performance?

This research question focuses on the importance of feature engineering in enhancing the effectiveness of the machine learning algorithms especially in loan default risk modeling. Feature engineering is the process of selecting, constructing, and sometimes even generating new features from the raw data with the aim of enhancing the performance of any machine learning algorithm. The selection of features as well as the techniques applied to create them can greatly affect the model's capability of identifying correlations and coming up with proper predictions.

What are the most significant factors influencing loan default risk based on the Qualco data?

This research question aims at establishing the most important factors that are responsible for loan default in the Greek lending market based on the dataset from Qualco. Through the analysis of the relative importance of various borrower characteristics this study wants to identify the most important factors that are related to loan default risk in the context of the Greek market.

4.3 Proposed Methodology

This research employs a rigorous and comprehensive methodology to investigate the application of machine learning for loan default prediction within the Greek lending market, utilizing the Qualco dataset. The methodology is divided into several key stages, beginning with data preprocessing and feature engineering. The Qualco data undergoes a thorough cleaning process to handle missing values, convert data types, and address inconsistencies. Feature engineering techniques are then applied to create new informative features from the existing data, enhancing the dataset's predictive power [32]. These techniques include extracting temporal features from dates, encoding categorical variables, and potentially creating interaction terms.

After the data preparation, several machine learning models are used for loan default predictions. Some of them include Logistic Regression which is one of the most commonly used statistical models especially in binary classification and thus used as a baseline for other models; Ensemble methods such as Random Forest, XGBoost, LightGBM, and CatBoost, which are known to be very accurate and are capable of dealing with data that has complex patterns are then used [111] [112]. These models are trained on the processed Qualco data through supervised learning where the algorithms learn how to make associations between the characteristics of the borrower and the outcomes of the loans.

In order to enhance the performance of each model, hyperparameter tuning is done to seek the best hyperparameter settings for each algorithm. Performance measurent with AUC-ROC is used to assess the performance of the model after which, cross-validation methods are used in order to make the evaluation more robust and less dependent on the data set chosen [113].

The performance of different models is compared using cross validation and statistical tests to determine if any model demonstrates statistically significant outperformance. This analysis involves Friedman tests and then pairwise comparisons with Nemenyi tests, helping identify the most effective model for predicting loan defaults in the Greek market. To understand the key drivers of loan default risk, feature importance analysis is conducted. This includes examining feature importance scores from the models and utilizing SHAP (SHapley Additive exPlanations) values to quantify the contribution of each feature to the prediction. SHAP values provide a consistent and locally accurate measure of feature importance, even for complex ensemble models [114].

This systematic approach is devised to respond to the research questions stated earlier, to know if ensemble models are better than logistic regression, if any certain Gradient Boosting Tree model is significantly better than the others, how the different feature sets or feature engineering techniques affect the model performance and what are the most important factors that define loan default risk using the Qualco data. Through answering these questions through training, evaluating, comparing of models and analyzing feature importance, this research is expected to offer useful information regarding application of machine learning for loan default prediction in the Greek lending market. The findings will inform the development of more effective credit risk management strategies, leading to more informed lending decisions and greater financial stability within the region [25].

Chapter 5

Experiments

5.1 Initial data exploration processing

In this section, we provide a detailed overview of the four datasets used in the loan default prediction and collections management analysis. Each dataset—**actions**, **accounts**, **customers**, and **transactions**—provides a different perspective on the lifecycle of loan accounts and their related activities. Together, these datasets enable the development of predictive models to assess default risk and evaluate recovery strategies. Below, each dataset is described in terms of its key features and their relevance to the analysis.**Actions Dataset**

The **actions dataset** captures the various actions taken on loan accounts, including interventions by the lender or third-party agencies to manage delinquent loans. It records the specific activities and strategies applied to resolve loan issues, making it a valuable resource for understanding the progression of loan recovery efforts.

Key features include:

- **TRAN_FLAG_MATCH_PROMISE**: This indicator tracks whether the borrower has fulfilled a promise to pay. In the context of a collection, promises to pay are key behavioral signals that reflect a borrower's intent and ability to meet their financial obligations. Failure to fulfill promises may suggest a heightened risk of default.

- **TRAL_STEP_ENTRY_DATE**: The date the loan or account entered a specific step in the **Non-Performing Loan (NPL) management strategy**. Tracking this entry date allows the lender to monitor how long an account has been in each recovery phase, providing insights into the efficiency of the NPL process.

- **TRAL_STRATEGY_STEP**: The specific strategy or step applied to the account within the NPL process (e.g., legal action, restructuring). This field is critical for understanding which recovery methods are being applied to which loans and how effective they are in preventing defaults.

The **actions dataset** is central to understanding the **management of delinquent loans**. It allows us to evaluate the effectiveness of different recovery strategies, track borrower behavior, and analyze the impact of interventions on the likelihood of loan recovery or default.**Accounts Dataset**

The **accounts dataset** focuses on the financial structure and status of the loan accounts. It includes information about the outstanding balances, arrears, collateral, and any guarantees associated with the loans. This dataset is crucial for evaluating the financial risk tied to each loan and assessing recovery potential in case of default.

Key features include:

- **ACCL_AMT_INSTALMENTS_FUTURE**: This column reflects the total value of future installments, indicating the remaining debt. Accounts with larger remaining balances may present a higher risk of default, especially if the borrower is already in arrears.

- **ACCL_COLLATER_AL_OR_RE_MARKET_VALUE**: The market value of the collateral (e.g., real estate) backing the loan. Loans with higher collateral values are considered less risky because they provide security for the lender in case of default, allowing for asset recovery.

- **ACCL_DELINQ_STRATEGY**: The NPL strategy applied to the account, such as restructuring or legal action. Understanding the strategies applied to delinquent loans helps in assessing the likelihood of recovery and the timeframes for resolution.

The **accounts dataset** provides a comprehensive view of the **financial health** of each loan. By examining loan balances, collateral values, and delinquency strategies, we can build a robust risk profile for each account and predict which loans are more likely to default or recover.**Customers Dataset**

The **customers dataset** contains demographic, geographic, and behavioral information on the individuals holding the loan accounts. This dataset is vital for understanding borrower characteristics and how these traits influence loan performance and default risk.

Key features include:

- **CUSL_BIRTHDATE**: The date of birth of the customer, allowing us to compute their age. Age can be a significant factor in loan performance, as different age groups may have varying levels of financial stability and credit history.
- **CUSL_GEOGRAPHICAL_REGION**: The geographic region of the customer. Regional economic conditions can influence default rates, as borrowers from regions with higher unemployment or lower economic activity may face greater difficulty in meeting their financial obligations.
- **CUSL_TOTAL_ARREARS**: The total amount in arrears across all accounts. Customers with high arrears are at a greater risk of default, and this metric serves as a direct indicator of financial distress.

This dataset enables a deeper understanding of the **borrower's profile** and allows for segmentation based on demographic and behavioral factors. By incorporating this data into predictive models, we can improve the accuracy of default predictions and develop targeted strategies for different borrower segments.**Transactions Dataset**

The **transactions dataset** captures all financial events associated with loan accounts, such as payments, fees, and adjustments. Each transaction provides a record of the financial activity related to the loan, enabling a detailed analysis of payment behavior and arrears management.

Key features include:

- **TRAN_AMOUNT**: The monetary value of each transaction, representing payments, interest charges, or penalties. Analyzing transaction patterns can reveal trends in borrower behavior, such as consistent payments or periods of missed payments, which are important indicators of default risk.
- **TRAL_DEBT_AMOUNT**: The outstanding debt at the time of the transaction. Higher debt amounts, especially when combined with missed payments, suggest a higher risk of default.
- **TRAL_CURRENT_BUCKET**: The current delinquency status of the account, categorized into buckets based on the number of days past due (e.g., 30, 60, 90 days). Movement between buckets over time is a strong indicator of the borrower's risk of default, with higher buckets representing more severe delinquency.

The **transactions dataset** offers a time-based view of the **financial lifecycle** of each loan. By analyzing how payments and arrears evolve over time, we can gain critical insights into borrower behavior, the progression of delinquency, and the likelihood of default.

The integration of the **actions**, **accounts**, **customers**, and **transactions** datasets provides a comprehensive, multi-dimensional view of loan performance and collections management. Each dataset contributes unique and valuable information that, when combined, offers a holistic understanding of borrower behavior, financial status, and recovery strategies. Specifically, the datasets encompass:

Behavioral data: This includes the tracking of customer engagement, fulfillment of promises to pay, and responses to collections efforts. These indicators are crucial for understanding borrower intent and predicting the likelihood of future payments or defaults.

Financial data: The datasets capture critical financial metrics, such as outstanding loan balances, amounts in arrears, payment histories, and the collateral backing each loan. These variables provide the foundation for assessing the financial health of loan accounts and estimating default risk.

Demographic data: Borrower characteristics, including age, geographic location, and other personal attributes, offer valuable insights into how different segments of borrowers behave and their associated risk levels. Demographic factors often play a key role in shaping repayment behavior and can inform more targeted risk models.

Strategic data: This data details the various actions taken by lenders and debt collection agencies, including Non-Performing Loan (NPL) management strategies and recovery steps. These strategies are critical for understanding the effectiveness of different interventions in recovering delinquent loans.

By leveraging these datasets collectively, predictive models can be developed to assess **default risk** with greater accuracy and identify the **most effective recovery strategies** for delinquent loans.

Dataset Merging

The integration of multiple datasets is a crucial part of this research, ensuring that the different data sources are aligned correctly and consistently across the analysis. We worked with four datasets: **Accounts**, **Customers**, **Transactions**, and **Actions**. Each dataset contains unique identifiers for different entities, and the process of renaming and mapping these identifiers was essential to ensure consistency and avoid confusion.

The original names for the identifiers in each dataset were not uniform, which could potentially lead to errors during merging. For instance, the **Transactions** dataset originally used ID to refer to the transaction identifier, TRAN_ACCT_CODE to represent the account, and TRAN_CUST_CODE_OWNER to refer to the customer. Similarly, the **Actions** dataset used ID for action identifiers, ACTION_ACCT_CODE_CONCE for the account, and ACTION_CUST_CODE_CONCERNED for the customer involved in an action. In contrast, the **Accounts** and **Customers** datasets used ID to refer to account and customer identifiers, respectively.

To ensure clarity and to make the datasets compatible for merging, we renamed these key columns as follows:

- In the **Transactions** dataset:
  - ID was renamed to TRANSACTION_ID,
  - TRAN_ACCT_CODE to ACCOUNT_ID, and
  - TRAN_CUST_CODE_OWNER to CUSTOMER_ID.

  In the **Actions** dataset:
- - ID was renamed to ACTIONID,
  - ACTION_ACCT_CODE_CONCERNED to ACCOUNT_ID, and
  - ACTION_CUST_CODE_CONCERNED to CUSTOMER_ID.
- In the **Accounts** and **Customers** datasets, the column ID was renamed to ACCOUNT_ID and CUSTOMER_ID, respectively.

This renaming process ensured that when we merged the data, the identifiers that were used to represent the same entities were consistent throughout the different data sets, and thus there was no confusion and fewer chances of errors. For example, the **ACCOUNT_ID** in the **Accounts**, **Transactions**, and **Actions** datasets consistently referred to the same account, facilitating a smooth merging process.

As part of our exploratory data analysis (EDA), we validated this approach by checking the integrity of these mappings. Specifically, we ensured that each **Action** and **Transaction** correctly matched an existing **Account**. Through this validation process, we confirmed that all datasets were aligned as expected, with no unexpected mismatches between the keys used for merging. This step was crucial in ensuring that the datasets could be used reliably for subsequent analysis and modeling.

5.1.1 Choice Between Account-Level and Customer-Level Aggregation

In the context of this research, one of the key decisions involved choosing whether to aggregate the data at the **account** or **customer** level. Both approaches have different implications for the analysis, modeling, and interpretation of results, and the choice between them depends on the specific goals of the study and the nature of the data.Account-Level Aggregation

Aggregating at the account level means treating each account as a distinct entity. In this approach, the features and transactions associated with each account are aggregated and analyzed independently of other accounts belonging to the same customer. This method allows the model to capture variability between accounts, even when they are linked to the same customer. For instance, a customer might have both high-risk and low-risk accounts, and aggregating at the account level would allow the model to capture the specific risk profile of each individual account.

The account-level aggregation is particularly useful when the focus is on understanding the risk associated with **individual accounts**. For example, in loan default prediction, different accounts held by the same customer may have different characteristics—such as loan amounts, repayment schedules, or interest rates—that could affect the likelihood of default. By aggregating and modeling at the account level, we ensure that these distinct features are considered, allowing the model to make more granular predictions for each account.Customer-Level Aggregation

On the other hand, the customer-level aggregation means that all the accounts associated with the same customer are grouped together. This approach combines all the information for all of the customer's accounts into one, thus providing a consolidated view of the customer's financial activities. Features such as total outstanding balance, total number of accounts, and cumulative repayment history are aggregated across all accounts linked to a single customer.

Customer level aggregation is appropriate when the general objective is to examine the **customer's financial condition or risk**. This approach can be useful in the identification of patterns that are inherent in the customer rather than the risk that is associated with the specific accounts. For instance if a customer has several accounts, a model trained on customer level data will take into account the overall financial behavior of the customer as opposed to the dynamics of the individual accounts.

5.1.2 Exploratory Data in Application Status and Related Features

In the accounts dataset, certain features exhibited missing values, prompting a deeper analysis to understand their implications. The variable **Status of the last submitted application within the observed period (ACCLAPPLSTATUS)**, which captures the application status of accounts, had missing values for specific accounts across a range of monthly snapshots, indicated by the **Snapshot number (SNAPNUM)** feature. **Snapshot number (SNAPNUM)** represents a snapshot identifier, where each value denotes a specific month. For this analysis, the focus was on accounts with missing values for **Status of the last submitted application within the observed period (ACCLAPPLSTATUS)** in snapshots numbered from 38 to 48, which represent the months following the initial observation month (i.e., February 2019 to December 2019, after January 2019 which is **Snapshot number (SNAPNUM)** 37).

To identify these accounts, two conditions were applied: one to filter accounts with missing values in **Status of the last submitted application within the observed period (ACCLAPPLSTATUS)**, and another to focus on the specific **Snapshot number (SNAPNUM)** range of 38 to 48. The resulting subset consisted of accounts that had at least one missing value for **Status of the last submitted application within the observed period (ACCLAPPLSTATUS)** in months beyond the initial observation month. Upon further inspection, it was found that these accounts consistently lacked valid entries for **Status of the last submitted application within the observed period (ACCLAPPLSTATUS)** throughout the observation period. This consistent absence of data suggests that these accounts did not have any ongoing applications or interactions with the lender during this timeframe. All the completely empty accounts and their descriptions can be found in the appendix.

In order to check the data completeness for these accounts the list of all the accounts which met the criteria was also selected for further review. The examination was made on features which are related to payment amounts, installment, payment and collateral since these are crucial measures of an account's financial transactions and position. The filtered columns were selected which contained words such as principal, installment, payment, amount and collateral.

After reviewing these features it was observed that the selected account had no values apart from the mandatory (non-null) values in all these columns. This is further supported by the missing values in Status of the last submitted application within the observed period (**ACCL_APPL_STATUS**) and the lack of data in some of the financial variables such as payment, installment and collateral update which suggests that the account had no major activities in the course of the study period. This observation was made across the other accounts with missing application status value, this meant that these accounts may have had little or no financial transactions in the course of the study period.

These findings are important as they show that there is a group of accounts which may not provide useful information to the model due to the lack of financial transactions. Such accounts can either be removed from the model in a bid to enhance data quality or set aside for review. The following missing values are very important to handle in order to make the predictive model more stable and reliable as improper or inconsistent data may lead to poor performance of the model and the predictions [39].

5.1.3 Target Variable Creation Process

The creation of the target variable for the predictive model included several steps beginning with data preparation and identifying the right columns. The target variable is used to show that an account is likely to become a risk or move to a riskier state in the next observation period. Further, here's a detailed account of how the target variable was developed.

First, additional columns were created to represent the status and application bucket for the following month. Specifically, for each account, the **Status of the last submitted application (ACCL_APPL_STATUS)** and **Bucket of the latest application (ACCL_APPLICATION_BUCKET)** were shifted by one month using the .shift(-1) method, effectively creating two new columns (**ACCL_APPL_STATUS+1** and **ACCL_APPLICATION_BUCKET+1**). These columns provide insight into the status and application grouping for the subsequent month, which is crucial for defining the target outcome.

Next, only accounts labeled as "Running" in **ACCLAPPLSTATUS** were considered for further analysis. These represent active accounts that are currently in good standing. The rows corresponding to these active accounts were then processed to handle missing values: any missing value in **ACCL_APPL_STATUS+1** was filled with "Running," and similarly, any missing value in **ACCL_APPLICATION_BUCKET+1** was filled using the current month's value. This step ensured continuity in the status and application bucket information and prevented the target creation process from being affected by missing values.

Once the data for active accounts was prepared, the target variable was created based on two conditions:

1. **Application Bucket Change**: If the **ACCL_APPLICATION_BUCKET+1** (next month's application bucket) was greater than the current month's bucket, it indicated a worsening situation, suggesting an increased likelihood of default. Therefore, the target was set to 1.

2. **Status Change**: If the **ACCL_APPL_STATUS+1** (next month's application status) was not in the list of acceptable statuses (such as 'Running', 'Fulfilled', 'Partially Fulfilled', or 'Out of Collection'), it indicated a negative outcome, and the target was also set to 1.

For each account, if either of these conditions was met, the target variable was set to 1, representing an increased risk of default or a negative transition. Otherwise, the target was set to 0, indicating no significant risk detected. It is important to note that even if an account is initially assigned a target of 1 due to a worsening condition, it remains valid for monitoring in subsequent periods, especially if it starts making payments again and the target shifts back to 0. This continuous monitoring allows for capturing any future defaults. This target variable provides a binary indicator of whether an account is likely to transition into an undesirable state in the next period.

After defining the target, a comparison of the original dataset and the filtered dataset of active accounts was made. The original dataset consisted of over 4 million entries, with 243 features, occupying approximately 7.4 GB of memory. In contrast, the filtered dataset of active accounts, which included the additional target column and next-month information, contained around 158,305 entries, with 246 features, and required significantly less memory (297.1 MB). This reduction in size was a result of focusing solely on active accounts and the necessary columns for predicting future states.

In summary, the target variable was created to indicate potential risk, based on changes in application status and application bucket between consecutive months. This approach ensures that the model can effectively identify accounts that are at risk of defaulting or entering a negative financial state.

5.2 Cleaning

5.2.1 Datetime Features

After creating the target variable, it was crucial to clean the data and ensure that each column had the correct data type. This process involved converting relevant columns to appropriate formats, particularly for date and datetime features.

First, the columns identified as containing date or datetime information were converted to the correct format. The features categorized as **Date features (MT_DATE)** or **Datetime features (MTD_ATETIME)** were extracted from the dataset metadata. These columns were spread across different entities, including **Account**, **Action**, **Customer**, and **Transaction**. Ensuring these columns were correctly formatted helped standardize the temporal information, allowing for more accurate time-based analyses.

During the conversion process, a loop iterated through all the date features, attempting to convert each column to a datetime format using pd.todatetime(). For most of these columns, the conversion was successful; however, some columns could not be converted due to problematic values that were outside the bounds of valid datetime values. Specifically, certain columns contained unrealistic dates, such as

2333-03-15, **9999-12-31**, and **3030-05-30**, which caused conversion errors.

To address these issues, the problematic dates were handled individually. For example:

For the column **Next installment date (ACCH_DATE_INSTALMENT_NEXT)**, records containing the date **2333-03-15** were identified, and the problematic values were replaced using a custom-generated date range. This date range was created using pd.daterange() with a monthly frequency from January 2019 to December 2019, ensuring that the corrected values were within a valid and relevant timeframe.

For **Expiry date (ACCH_DATE_EXP)**, entries with the date **9999-12-31** were replaced with a placeholder date, **0000-00-00**, to indicate missing or invalid data.

Similarly, for **Starting date of denormalized application (ACCLDENORMAPLL-STARTINGDATE)**, the erroneous date **3030-05-30** was replaced with **0000-00-00**.

These replacements enabled to successfully convert these columns into datetime format and to handle the remaining problems by forcing those entries to be NaT (Not a Timestamp) using the errors= ‘coerce’ parameter. This ensured that all valid dates were properly converted while all the other values were identified as missing.

Following the data type corrections, an assessment of the datasets revealed the number of datetime columns present in each entity. The **accounts** dataset had 43 datetime columns, while **transactions**, **actions**, and **customers** datasets had 0, 5, and 8 datetime columns, respectively. This was very important in cleaning the data and thus enable the data to be analyzed using different temporal features. This was done to make sure that all date related columns were correctly formatted for the next stages of modeling.

5.2.2 Categorical Features

The analysis of categorical features across different datasets was essential to understand the structure and diversity of values, as well as the implications for predictive modeling. The main datasets included **customers**, **transactions**, **actions**, and

accounts, each containing multiple categorical columns that exhibited unique patterns in the distribution of values and missing data.

The datasets contain multiple categorical variables that represent various aspects of customer characteristics, account behavior, transactions, and actions taken. In order to get a deeper insight into the data we conducted an extensive analysis of these categorical columns to determine the unique values for each column, the occurrence of each value and the missing value rate. This enabled us to capture the features heterogeneity and to determine whether such features are relevant to the prediction of loan default.

The analysis revealed the following key insights:

Categorical Features Diversity: Across the four datasets, we observed a wide range of unique values for categorical variables. Some columns, like CUSH_CATEG_TYPE and TRAL_STRATEGY_STEP, had a significant number of unique values (593 and 116, respectively). This suggests a high degree of granularity in describing the customers or transaction activities, which might require feature grouping or encoding techniques to make them suitable for machine learning models. Additionally, the analysis showed that certain features had a small set of dominant categories, such as CUSL_HAS_CONTACT_ATTEMPT_EVER, which had 96.22 of its values as ”Yes”. Such observations imply potential challenges in handling class imbalances during modeling.

High Frequency of Missing Values: Some categorical columns showed high percentages of missing data. For instance, CUSL_COUNTRY had a missing value ratio of 99.77. This level of missingness creates questions as to the effectiveness of these features and whether they can be effectively used without first having to undergo some form of imputation or transformation. Therefore, such features with very high levels of missing values may be removed if they are deemed to contribute little to the model.

Columns with Only One or Few Unique Values: Several features had only one or a very small number of unique values, such as TRAN_STATUS_ACCEPTANCE, which had ”Accepted” as its sole value across the dataset. Such columns were deemed non-informative for the predictive modeling task, as they provide no variance and thus were excluded from further analysis.

Potential for Feature Grouping: For columns like CUSL_PROFESSION_GROUPING, which had 39 distinct values, it was noted that certain categories represented only a small proportion of the data. Grouping such infrequent categories into broader classifications can enhance model generalization while avoiding overfitting. In cases like ACTL_DEPARTMENT, which had a considerable proportion of missing values and a large number of distinct values, effective grouping or imputation strategies are required to improve the model's performance and manage data sparsity effectively.

5.2.3 Analysis of Categorical features

The categorical analysis that was supplied was a foundation for the understanding of the data distribution as well as sparsity of the data when multiple data sets were integrated. Such features with a large number of possible values or severely skewed distributions may need to be transformed or binned to avoid being dominant in the model. Such features with very low sparsity and very few categories may be questioned they on may their not predictive add power much as to the model or even pollute it.

These insights were very helpful in deciding which features to use, which data to quality modify, and and make which sure to that remove the in predictive order model to is enhance stable the and able to accommodate a large number of categorical variables. The next thing that could be done is that dimensionality reduction techniques like feature binning or PCA could be used to manage high cardinality features such as ACCH_CATEG_TYPE and TRAL_STRATEGYS_TEP. Also, the imbalanced features can be balanced using sampling or weighting methods to improve the effectiveness of the final model.

Customers Dataset

The customers dataset contains **30** object-type columns, including categorical features such as **gender** (CUSH_GENDER), **geographical region** (CUSL_GEOGRAPHICAL_REG **profession grouping** (CUSL_PROFESSION_GROUPING), and **status law** (CUSL_STATUS_LAW The analysis revealed the following:

High Concentration of Values: Several columns had a high concentration of values within a few unique categories. For instance, **gender** had **65** labeled as "M" and **35** as "F". Similarly, **profession grouping** had **40.45** in "ΛΟΙΠΑ ΕΠΑΓΓΕΛΜΑΤΑ" (Other Professions) and **15.77** as "Non-working".

Imbalance and Sparsity: Certain categorical columns exhibited severe imbalance, with the majority of observations concentrated in a single category. For example, **CUSL_REFUSE** had **98.16** values as "No", indicating that the feature may not provide significant discriminatory power for modeling purposes. Additionally, **CUSH_STATUS_LITIGATION** had **42** unique values, but **66.69** of the observations were under **"Complaint - Notice"**.

Promising Columns:

CUSL_POSITION: Represents the customer's position or type (e.g., individual or legal entity). Different positions may imply varying financial stability.

CUSH_PROFESSION: Professions provide insight into income stability, which directly impacts repayment capacity.

CUSL_ADMINISTRATION_STATUS: Changes in administration status could correlate with increased default risk.

CUSL_HAS_PAYMENT: Indicates whether a payment has been made, which is a direct indicator of financial engagement.**Transactions Dataset**

The transactions dataset contained **11** object-type columns, primarily detailing transaction types and states such as **application type** (TRAL_APPL_TYPE), **bucket** (TRAL_CURRENT_BUCKET), and **strategy step** (TRAL_STRATEGY_STEP). The following observations were made:

Diverse Categories: TRAL_APPL_TYPE had **20** unique values, with categories such as "Settlement", "Out of Mandate", and others with Greek labels indicating different types of transactions or arrangements. Only a few categories, like "Settlement", made up a significant portion of the data.

Sparsity in Strategy Steps: TRAL_STRATEGY_STEP had **116** unique values, indicating high diversity in how transactions were managed. Notably, **20** unique values accounted for **85.15** of all observations, suggesting that the remainder was relatively sparse.

Promising Columns:

TRAL_CURRENT_BUCKET: Tracks repayment status, with movement indicating worsening or improving risk levels.

TRAN_AMOUNT: Provides insight into financial activity and ability to manage debt.

TRAL_APPL_TYPE: Different types of applications (e.g., settlements) are related to financial distress.

TRAL_TYPE_DESC: Describes transaction type, potentially reflecting changes in behavior.**Actions Dataset**

The actions dataset contained **14** object-type columns with various categorical features, including **source of activity** (ACTL_SOURCEOF_ACTIVITY), **acceptance status** (ACTL_ACCEPT_ANCESTATUS), and **department** (ACTL_DEPARTMENT). Key insights included:

High Acceptance Rate: ACTL_ACCEPTANCE_STATUS had **89.05** of actions labeled as "Accepted" and **10.95** as "Internal", reflecting an operational consistency in processing actions.

Departments with Sparse Data: **Departments** like "IT" and "Quant Collections" constituted a majority of entries, while other departments had minimal representation, with features such as "Small Business" having a near-zero ratio.

Promising Columns:

ACTL_ACCEPT_ANCES_TATUS: Acceptance status can indicate successful interactions or actions related to repayment.

ACTL_ACTIVITY_TYPE: Different activities may imply different risk levels, such as settlements or legal actions.

ACTL_DEPARTMENT: Departments involved in customer interactions may suggest varying levels of customer engagement or financial distress.

ACTL_SOURCE_OF_ACTIVITY: Differentiates between internal or external sources, providing insight into customer interactions.**Active Accounts Dataset**

The accounts dataset contained **246** columns in total, including several object-type columns related to **application status** (ACCL_APPL_STATUS), **geographical distribution** (ACCH_CATEG_TYPE), and **litigation status** (ACCH_STATUS_LITIGATION). The main observations were:

Large Number of Categories: Features such as ACCHCATEGTYPE had **593** unique values, requiring binning or dimensionality reduction strategies for effective use in predictive modeling.

High Sparsity in Litigation Status: ACCH_STATUS_LITIGATION showed **46.09** missing values, and over **66.69** of non-missing values were in a single category, raising questions about the feature's utility for prediction.

Promising Columns:

ACCH_APPLICATION_BUCKET: Represents the current application bucket status and reflects changes in customer status.

ACCH_DEBT_AMOUNT: Amount of debt directly impacts repayment ability, making it a strong predictor.

ACCH_LAST_PAYMENT_AMOUNT: Recent payment amount provides a direct indicator of financial health.

ACCH_COLLATERAL_VALUE: Collateral value impacts the perceived security of the loan.

ACCH_MAX_CURRENT_BUCKET: Historical risk data indicating the worst past condition of the account.

5.2.4 Handling Missing Values

The process of filling missing values was carried out in order to contribute to the completion of the data sets and to prepare them for modelling without the accumulation of biases. The fillvalues() function was used in order to fill missing values for columns and the main approach was to use forward fill (ffill()) and backward-filling (bfill()). The function iterated over each group of records associated with a unique ID, selecting only the columns of the specified type (e.g., object type for categorical features). This enabled the filling of missing values in a structured manner based on the groups, thus maintaining the coherence of the data.

This approach also served to maintain the temporal coherency of each account as a side effect of the missing data imputation [115]. There were some columns which had a very high missing value and cannot be filled appropriately were identified for removal in the subsequent modeling process.

5.2.5 Handling Categorical Variables and Feature Binning

The handling of categorical variables involved careful consideration of category frequency, binning of low-frequency categories, and the creation of dummy variables to enable effective use of these features in predictive modeling. Feature binning and the creation of dummy variables were used in order to deal with high cardinality features and thus make the model focus on the most important aspects of the data. This process was very important in order to make sure that categorical variables model were in incorporated a into way the that would help the predictive power of the model without complicating it or increasing the chances of over-fitting [116]. Considering the fact that there are many categorical features in the datasets, which can be related to customers or transactions, a systematic approach was taken to transform these variables into a machine learning ready format.

The function **createcategoricals()** played a significant role in managing categorical variables and creating dummy variables for binned categories. The primary goal was to convert categorical variables into a form that could be easily interpreted by machine learning models while addressing the challenges posed by high-cardinality features.Key Steps in Handling Categorical Variables

Category Selection and Binning The **createcategoricals()** function included a parameter called catstokeep, which specified which categories should be retained. This approach allowed the retention of only the most frequent or relevant categories, while less common categories were grouped under a new category labeled 'Other'. This binning strategy was particularly useful for features with many unique values, as it helped reduce the overall number of categories, thereby simplifying the dataset.

Grouping Categories into 'Other': The process of assigning categories to the 'Other' category was based on the frequency of category appearance for each categorical feature. In particular, less frequent categories were clustered under the 'Other' category in order to decrease the number of categories and thus make the model simpler the and 'Other' less category likely was to based overfit. on The the exact frequency cut-off distribution for of categorizing value which for category each to feature put and in this was done in a bid to identify the This most approach important ensured categories that to the keep final while dataset categorising had the a others more as reasonable 'Other'. cardinality which in turn contributed towards enhanced generality and less computational time.

Dummy Variables Creation: To convert the selected categories into a numerical format, **pd.getdummies()** was used within the function. This method converted categorical values into binary (0/1) columns, which represent the presence or absence of each category. Dummy variables are particularly useful for machine learning models that require numerical input, as they enable the use of categorical data without introducing ordinal relationships between categories that do not naturally exist.

Feature Reduction and Dimensionality Management After creating dummy variables, the original categorical columns were dropped (dropcol=True), which helped reduce the dimensionality of the dataset. This reduction was beneficial for machine learning models, especially those that can be negatively affected by high-cardinality categorical features. By replacing the original high-cardinality categorical variables with a smaller number of meaningful binary features, the dataset became more manageable and suitable for modeling, with fewer risks of overfitting.

Handling High-Cardinality Features High-cardinality categorical features for example ACCH_CATEG_TYPE posed certain challenges because they had a number large of unique values. These features were examined to identify which categories could be kept and which required being classified as 'Other'. For instance, the features with hundreds of unique values were binned in order to decrease the number of categories to a reasonable number while still trying to preserve the information. This not only made the model simpler but also increased the ability of the model to generalise by identifying the most important categories.

Feature Engineering of Datetime Features

The subsequent step involved extensive feature engineering on the datetime columns across the datasets to enhance the predictive power of the model. This involved extracting meaningful components from datetime features to capture the temporal patterns underlying customer behavior. Feature engineering plays a critical role in adding value to raw data by creating new, informative features that are more predictive for modeling purposes.

Extraction of Date Components

The datetime features had a lot of information that could be useful for modeling behavioral data over time. In order to extract different components of date and time, extractdatecomponents( ) function was used for each of the datetime columns. This way we were able to extend the datasets with new features which could help to understand the dynamics of customer behaviour in terms of time series.

1. **Identification of Datetime Features**: The datetime columns were identified from each dataset, which included transaction dates, account updates, and application submission dates. These temporal attributes were pivotal in capturing the dynamics of customer behavior across different stages of their financial lifecycle.

2. **Extraction of Temporal Components**:

Year, Month, and Day: Extracting the year, month, and day components allowed us to decompose the datetime features into granular levels. The year component, while not adding variability in this dataset, could indicate longer-term patterns in more extended datasets. The month and day components provided seasonality, capturing cyclical patterns that could indicate specific times of increased financial stress or repayment.

Day of the Week: This component captured differences in customer behavior across weekdays and weekends. Customers might be more likely to engage with loan repayment activities on weekdays when they have access to their financial services, compared to weekends when financial transactions might decrease.

Month Start and Month End Indicators: Binary features indicating whether a given date was the start or end of a month were included to identify points in time when customer financial behaviors often change, such as salary receipt dates, billing periods, and other regular financial activities.

The newly engineered temporal features were added to each dataset to enhance their descriptive power. Dropping the original datetime columns was a strategic choice aimed at:

Reducing Dimensionality: High-dimensional datasets can lead to increased computational burden and risks of overfitting. By retaining only the engineered features, we reduced the number of input variables while preserving meaningful temporal information.

Preventing Redundancy: Retaining both the original datetime and the extracted temporal components could have resulted in correlated features, which may negatively affect certain machine learning models by violating their assumptions of independence between features.

5.2.6 Creation of Final Dataset

The process of transforming the datasets into a unified and concise view that is both predictive and computationally efficient involved several key steps. This was one of the most important transformations which involved identifying accounts that had more than one transaction per month and aggregating the data accordingly.Handling Accounts with Multiple Transactions per Month

The transactions dataset was initially analyzed to identify accounts that had more than one transaction within a single monthly snapshot (SNAPNUM). Such instances were important because they indicated increased account activity, which could be a strong predictor of either heightened financial risk or engagement. Specifically:

1. **Identifying High Activity Accounts**: Accounts with more than one transaction per month were flagged for deeper analysis. This identification was done by grouping the dataset by ACCOUNTID and SNAPNUM and calculating the size of

each group. Accounts with multiple transactions were filtered for further inspection.

2. **Aggregation of Multiple Transactions**: For accounts with multiple transactions in a single month, an aggregation function (listifmultiple) was used to consolidate these transactions into a single representation for that period. This involved aggregating financial amounts, identifying the earliest and latest transaction dates, and retaining crucial temporal information about the activities.

3. **Data Consolidation and Simplification**: The resulting aggregated dataset ensured that each account had only one entry per SNAPNUM, thereby reducing redundancy while preserving meaningful transactional activity. This transformation helped maintain the temporal integrity of the dataset while also reducing its complexity.Categorical Feature Reduction and Encoding

To further optimize the dataset, categorical features with high cardinality were reduced in complexity. For example, variables like TRAL_APPL_TYPE and TRAL_STRATEGY_STE had numerous unique categories, many of which were sparse. To simplify these features:

Category Grouping: Only the most frequent categories were retained, while less common categories were grouped under a label called "other." This grouping reduced the overall cardinality, making the data more manageable and preventing overfitting in subsequent modeling steps.

One-Hot Encoding: The most relevant categorical variables were transformed using one-hot encoding. This process created binary features for each category, allowing machine learning models to easily interpret these categorical values without assuming an ordinal relationship between them.Aggregation Across Datasets

The final stage of the dataset creation involved merging the four key datasets—accounts, transactions, customers, and actions—into a single, comprehensive dataset. This merging process was done using a left join strategy, which ensured that all active accounts were retained while details from other datasets were added where available. Specific aggregation functions were applied:

Transaction-Level Aggregation: Key metrics like the total transaction amount per month, minimum outstanding debt, and the count of transactions were aggregated to provide a summarized view of account activity.

Action-Level Aggregation: Actions data were aggregated to reflect the number of actions taken per account per month, along with the type and status of actions (e.g., settlements, requests). This provided a clear picture of how accounts were managed over time.Handling Missing Values

Handling missing values was a critical part of the dataset creation process to ensure data integrity and consistency:

1. **Check for Columns with Complete Missingness**: After merging, columns that contained only missing values were identified and subsequently dropped. This step helped reduce computational overhead by eliminating features that provided no meaningful information.

2. **Preservation of Useful Data**: For columns with some missing values but also containing meaningful data, imputation strategies or retaining NaN values as indicators were considered. This allowed the model to use the presence of missing data itself as a predictor, especially when missing financial data might suggest non-engagement or financial distress.Resulting Dataset

The merged and transformed dataset comprised of 352 features and includes a mix of continuous, categorical, and datetime features:

Continuous Features (174): These included financial metrics such as loan amounts, debt values, and payment information, stored across multiple data types (float16, float32, and float64) to optimize both precision and memory usage.

Integer Features (7): These features represented count-based metrics, encoded categorical values, and other identifiers, providing numerical insight into account characteristics and aggregation.

Categorical Features (134): Representing customer classifications, product types, and other non-numerical variables, these features were optimized using the category dtype, allowing for significant memory savings and faster computations.

Datetime Features (37): The datetime columns, represented as datetime64[ns] and datetime64[s], captured different temporal granularities for various events, such as transaction dates, step entry dates, and installment schedules. These features were essential for understanding the temporal sequence and frequency of account events, allowing the model to identify patterns related to default risk.

The final dataset, after all these transformations, was reduced to a memory footprint of around 240 MB, ensuring that it remained computationally efficient without sacrificing the richness of information. This efficient dataset structure is highly conducive to the subsequent machine learning modeling phase, where both interpretability and computational feasibility are key considerations.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step that provides a deep understanding of the data and uncovers relationships between variables, potential outliers, and patterns that may affect the predictive model. In this study, EDA was conducted to ensure that the engineered and merged dataset was ready for modeling and to identify any further transformations that might be required. several distribution and count plots were used along with descriptive statistics to uncover important details in the data.

The target variable, indicating whether an account was at risk of default or not, was found to be imbalanced, with 64.17 of accounts marked as non-default (target=0) and 35.83 flagged as potentially risky (target=1). This class imbalance is critical to acknowledge as it can significantly affect model performance. Models trained on imbalanced data might be biased towards the majority class, resulting in poor predictive accuracy for the minority class (i.e., accounts at risk of default). To address this, techniques such as oversampling, undersampling, or using cost-sensitive algorithms may be required during the modeling phase.

Application Status (ACCL_APPL_STATUS): A detailed analysis of application statuses revealed that most accounts were categorized as ”Running” (indicating ongoing activity with no immediate issues), while smaller proportions fell into statuses such as ”Not Fulfilled,” ”Fulfilled,” ”Review,” etc. Notably, certain statuses like ”Review,” ”For Approval,” and ”Quality Control” represented accounts undergoing some level of scrutiny, potentially indicating early warning signs of financial distress.

Application Bucket (ACCL_APPLICATION_BUCKET): The analysis of the application bucket distribution highlighted a decreasing frequency of accounts as the bucket value increased. Lower bucket values (e.g., 0-6) were predominant, representing less risky accounts, whereas higher values (7 and above) were rare but indicative of elevated risk levels. This feature is crucial as it directly correlates with the account's risk profile and will be important in differentiating the risk tiers during modeling.

5.2.7 Data Leakage Considerations

Data leakage poses a substantial risk in predictive modelling tasks, potentially resulting in too optimistic performance evaluations and models that fail to generalise effectively to novel data. In this case, data leaking denotes the utilisation of information during model training that would not be accessible at the time of prediction [74].

When handling numerous datasets connected by account or customer identifiers, it is crucial to exercise caution about the aggregation of data and the division into training and test sets. If data from several accounts of the same customer is included in both the training and test sets, the model may unintentionally acquire patterns unique to the customer instead of the account, thereby ”leaking” information from the test set into the training phase [117].

If data is aggregated at the account level, numerous accounts belonging to the same customer may be divided between the training and test sets. In some instances, the model may obtain data regarding customer behaviour via the training accounts, therefore indirectly affecting predictions for the test accounts. This would result in exaggerated performance measures during evaluation, as the model is utilising information that would not ordinarily be accessible in a real-world context [117].

Conversely, aggregating the data at the customer level would alleviate this issue, as all accounts linked to a consumer would be consolidated. This approach may also hide the differences in risk profiles of specific accounts of the same consumer. A client can have both high-risk and low-risk accounts and when the data is aggregated at the customer level, it may become difficult to identify these differences properly. In order to avoid data leakage and at the same time keep the account-level data detailed, much attention was given to how the data sets were split for modelling and assessment. There was a major aim that ensured that the accounts of the same customer were always assigned to the same set regardless of whether the data was used for training or testing. Thus, the model was inhibited from acquiring knowledge about test set accounts via common consumer behaviour in the training set.

5.3 Data Preparation and Machine Learning Pipeline

5.3.1 Overview

This chapter describes the systematic approach undertaken to prepare, process, and analyze the unified dataset created for predicting loan defaults. The following methodology is proposed to address the challenges of class imbalance, high cardinality, and data leakage to develop robust and interpretable predictive models. The approach taken involves first conducting further data cleaning, then performing feature engineering, splitting of the dataset into training and testing sets, training the model, tuning the hyper-parameters and evaluating the model. All the steps are carefully planned to ensure that diverse features are used effectively without compromising the computational time and to avoid biases.

5.4 4. Dataset Processing and Model Development Methodology

5.5 4.1 Dataset Preparation and Feature Engineering

The utilization of data in machine learning applications is a critical process that must consider various aspects that could greatly influence the model's performance. This section states the procedures followed in converting the raw financial data into a form that is suitable for predictive modeling while at the same time ensuring that data quality is well preserved although in a computation-friendly manner.

5.5.1 4.1.1 Data Type Optimization

The appropriate choice of data types is one of the critical factors that have to be taken into consideration in the machine learning pipelines as it affects both the model performance and the computational costs. In the context of financial data, there are two conflicting factors that need to be considered: The level of precision that is needed for the calculations depends on the memory that is available especially when working with big data sets with many numerical features.

In the current study, the float-type columns were first saved as float16 which offers a range of about ±65,504 with reduced Bit precision. Although this data type helps in conserving memory, it is inadequate for financial calculations as it lacks the desired level of precision. According to the literature, float32 is accurate enough for most, if not all, financial applications while using the least amount of memory as needed for storing the calculations data [118]. Consequently, all float16 columns were converted to float32, ensuring sufficient precision for financial calculations while maintaining reasonable memory consumption.

Integer columns, originally stored as int8 (-128 to 127 range), were converted to int32. This decision was driven by the need to accommodate larger ranges for count-based features and identifiers, which are common in financial datasets. The int32 data type, supporting values from -2,147,483,648 to 2,147,483,647, was employed because it is more efficient in terms of memory as compared to int64.

5.5.2 4.1.2 Temporal Feature Engineering

Temporal feature engineering plays an important role in default prediction because the traditional models used in default prediction are temporal in nature given that the payment behaviours and default trends are sequential in nature. The strategy that was used in the processing of datetime features was developed based on prior research in machine learning and personal knowledge in time series based problems. The datetime processing pipeline was initiated with the purpose of identifying temporal components help that in would describing various temporal characteristics. The extractdatecomponents function was created in order to obtain several types of temporal variables from each datetime column. The choice of temporal features was made based on the the research findings in of the field of behavioral finance and credit risk modeling. The monthly indicators are especially useful in the given case since they are closely linked to the salaries received by employees and the repayments of loans.

The inclusion of day-of-week and weekend indicators is added with the hope of capturing the operational cycle of banking systems and customer behavior patterns. Transaction and payment activities typically show distinct patterns between weekdays and weekends, reflecting both banking operations and customer preferences. Furthermore, quarterly features capture seasonal patterns in financial behavior in general. A difficulty in the temporal feature engineering was how to deal with wrong dates. There were some invalid date fields in the dataset, for instance, expiry date could be ‘9999-12-31’ while other datetime fields could have values such as ‘2333-03-15’. Such values need to be dealt with carefully so that they do not cause problems in the modeling pipe. Instead of removing these records which may lead to loss of information, the following strategy was undertaken to deal with the invalid dates:

1. Future dates beyond reasonable bounds (e.g., ’9999-12-31’) were replaced with a standardized null date representation (’0000-00-00’). This approach maintains data integrity while clearly identifying these records for subsequent analysis.

2. The dates which were in a possible but wrong order, for example ‘2333-03-15’ was changed to be within the study period. This decision was made on the basis of the fact that these dates could have been clipped or misrecorded rather than being actual future dates.

5.5.3 4.1.3 Categorical Feature Processing

Handling categorical data in machine learning poses certain issues which are not found in other types of data especially in financial datasets where categorical variables are usually informative. The strategy applied for the processing of categorical variables was based on the current best practices in machine learning and on the needs of the financial data analysis. A two-stage processing pipeline was applied to manage categorical processing.

This initial processing phase addressed several key considerations in categorical data preparation. The decision to add a ”missing” category rather than implementing more complex imputation strategies was based on research and anecdotal evidence suggesting that missingness patterns in data often carry meaningful information about customer data. This approach preserves the potential predictive power of missing values while maintaining the interpretability of the features.

The second phase of categorical processing involved applying the cardinality-based approach to feature encoding. This approach is based on the fact that there are different encoding techniques that are suitable for different levels of cardinality and this has been discussed in the machine learning literature [116].

For high-cardinality features, those features were classified as high cardinality if they had over 255 unique values, and for such features target encoding was employed. This threshold was set based on the models that would be used in this research later on. Target encoding is more effective and precise way of dealing with high cardinality features while attempting to preserve their predictive value. In order to avoid the problem of target leakage, the encoding was done with the help of cross-validation, that is the target information from the validation data did not influence the encoding of training data.

Low cardinality features, the features which contain 255 or less unique values were one hot encoded with some restrictions to avoid the problem of dimensionality. There was a restriction on the number of categories to be five with the other categories falling under the category ‘other’ which is common in similar cases [119]. When it comes to binary features, one category is usually dropped in order to avert the condition of multicollinearity which can be destructive to some machine learning algorithms.

5.6 4.2 Missing Value Treatment

The treatment of missing values represents a critical challenge in machine learning, particularly in the context of loan default prediction where missingness often carries meaningful information about customer behavior. The implemented methodology for handling missing values was developed based on extensive research in financial machine learning and the specific characteristics of our dataset.

5.6.1 4.2.1 Numerical Feature Imputation

The approach to handling missing values in numerical features was guided by both statistical considerations and domain knowledge. A sophisticated imputation pipeline was implemented.

The selection of median imputation for numerical features was based on several key considerations. First, financial data typically contains outliers that can significantly skew mean values, making them unsuitable for imputation. Median imputation is less sensitive to outliers and produce less noisy result than mean imputation especially when working with financial datasets that usually have many extreme values which are also meaningful. **The decision to include missing indicators** was driven by research and anecdotal evidence showing that missingness patterns in financial data often carry predictive information gain. These binary indicators enable the model to learn from both the imputed values and the missing data pattern as well. This is especially important in the loan default prediction problem where missing financial information may depict that the customer is not engaged or is financially insolvent.

5.6.2 4.2.2 Categorical Feature Imputation

The problem of missing data in categorical variables called for a more refined approach that involved the cardinality of the variables and their relationship with the target variable. The implementation of target encoding for categorical features with missing values represents a departure from simpler approaches like mode imputation or the creation of a ”missing” category. Target encoding handles these issues in a quite natural way as well as coping with missing values during the encoding process. The encoding was done with cross-validation so that there is no data contamination, this means that the encoding of categories and the target variable was done with only the training data. [120].

5.7 4.3 Data Splitting Strategy

Developing an efficient data splitting strategy was very important in order to conduct proper model assessment while at the same time preserving the temporal characteristics of the financial data. The implemented approach combines temporal splitting with group-based validation to address the specific challenges of loan default prediction.

5.7.1 4.3.1 Temporal Splitting Implementation

The primary split of the dataset was performed based on temporal ordering, recognizing the non-stationary nature of financial data. The implementation followed a structured approach. The implementation followed a structured approach. The decision to use 60 of the temporal sequence for training set, 20 for validation and the remaining 20 for test set is quite normal in time series forecasting and financial forecasting models. This split ratio offers enough history data for the model to learn complicated patterns as well as enough data for validation and testing. Research proves that small training sets cause performance issues in financial applications especially in identifying events that occur infrequently such as loan defaults.

5.7.2 4.3.2 Group-Based Validation

To prevent data leakage from account characteristics and ensure a robust model evaluation, a group-based splitting strategy was implemented within the training set. This methodology ensures that all records belonging to the same account remain together during the splitting process. The importance of this approach is supported by research showing that traditional random splitting can lead to optimistic performance estimates in financial modeling due to information leakage between related accounts [121].

5.8 4.4 Feature Selection Methodology

The implementation of feature selection in financial machine learning requires careful consideration of both statistical significance and domain relevance. This study employed a comprehensive approach combining multiple feature selection methods to ensure robust identification of predictive features while maintaining interpretability.

5.8.1 4.4.1 Recursive Feature Elimination with Cross-Validation

The first feature selection methodology implemented was Recursive Feature Elimination with Cross-Validation (RFECV), which was used for the feature subset selection and the ability to evaluate the model performance on the cross-validation data. The implementation used multiple base estimators in order to evaluate feature importance in the best possible manner.

The choice of step (step = 100) is a reasonable compromise between computational time and the ability to discriminate between features during the selection process. Study suggests that smaller step sizes, produce more accurate feature selection, but the additional computational power needed to arrive at the finer step sizes is often wastage as regards enhancement of model performance [52].

The implementation involved the use of several base estimators such as Logistic Regression, Random Forest, XGBoost, LightGBM and CatBoost. This is because feature selection is done using multiple algorithms and features that are selected by many of the algorithms are preferred over others. Features selected consistently across multiple algorithms were given higher consideration, following research demonstrating the effectiveness of ensemble feature selection methods in financial applications [122] [83] [65].

5.8.2 4.4.2 Mutual Information Analysis

Complementing the RFECV approach, the mutual information analysis was applied in order to model non-linear dependency between features and the target. This method is especially useful for financial datasets where the relationships between variables are not necessarily linear: Implementation of mutualinfoclassif for feature scoring with the threshold of feature selection being 0.08. This threshold was set in light of the empirical findings and the exploratory analysis done.

5.8.3 4.4.3 Variance Threshold Selection

The last stage of the feature selection pipeline used is variance filtering with standardization: The choice of 0. 2 as the variance threshold was done after analyzing those the which distribution have of the features potential in of the being datasets predictive used. and This which threshold have eliminates enough features variation. which The are standardization almost is constant applied while before keeping the variance thresholding to ensure that the variance threshold is defined on the same scale for all features.

5.9 4.5 Model Development and Optimization

The analysis of existing models and the creation of new models' implications and optimization were done in a very structured manner starting from the creation of the baseline model and ending with the hyperparameter optimization.

5.9.1 4.5.1 Baseline Model Implementation

The development and optimization of predictive models began with establishing a baseline Logistic Regression (LR) model, selected for its simplicity and extensive use in credit scoring applications [123]. Logistic Regression was chosen because it is one of the most commonly used models in the financial sector as it is easy to interpret and very reliable in risk assessment models. Also, it is computationally lightweight and hence suitable for initial data exploration as well as for comparing its performance with other more sophisticated classifiers.

The Logistic Regression model was implemented with a detailed preprocessing pipeline in order to deal with the multivariate character of the data set. The preparation of the data included filling the missing numerical values by the target median, and categorical scaling variables of encoding features based by on standardization. [124]. Additionally, the pipeline included mutual information based feature selection to select the 50 best features. The final logistic regression model was trained with a class weight balance for handling imbalanced data distribution in the dataset.

The model was assessed by 10-fold stratified cross-validation with grouping according to the ACCOUNTID to avoid splitting of data from the same account. This approach preserved the temporal characteristics of the data and reduced the possibility of data contamination. The performance of the model was evaluated by the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), a widely used measure for the binary classification problems particularly in risk modeling contexts [54].

There were two ways of calculating AUC-ROC score: according to the predicted class labels and predicted probabilities. In the case of class labels, the basic Logistic Regression model gave an AUC-ROC score of 0.80. However, this approach only looks at the predictions made at a certain threshold mostly 0.5 and does not take into account the certainty of the predictions and does not provide much information about the models performance. On the other hand, the second method used predicted probabilities, which gave a better AUC-ROC score of 0.84. This method allowed for a better assessment of the models' capacity to sort observations according to their probability of being defaulted. Since there is a greater concern in differentiating between different categories of risk, the second method was selected for all the AUC-ROC calculations thereafter. [125].

5.9.2 4.5.2 Hyperparameter Optimization Strategy

The optimization of model hyperparameters was implemented using Optuna, employing a Bayesian optimization approach. The hyperparameter search spaces were carefully defined based on both theoretical considerations and empirical research. Extensive parameters with a wide space were employed in order to allow for the optimization to be done in the best possible way. In addition, the optimization process incorporated a median pruning strategy, which meant that the optimization algorithm could terminate unfit trials early and so effectively search through the hyperparameter space without wasting computational time. This pruning strategy was very helpful in reducing computation time without negatively impacting the optimization process in any way. A number of machine learning models were tested during the optimization process these include CatBoost, Random Forest, XGBoost, LightGBM, HistGradientBoosting, and ExtraTrees Classifiers. Each model's hyperparameter space was specifically tailored to its architecture and the characteristics of financial default prediction. The optimization process was done with a cross-validation Stratified Group K Fold to ensure that account groups were not split between folds and that the target classes were well represented in each fold.

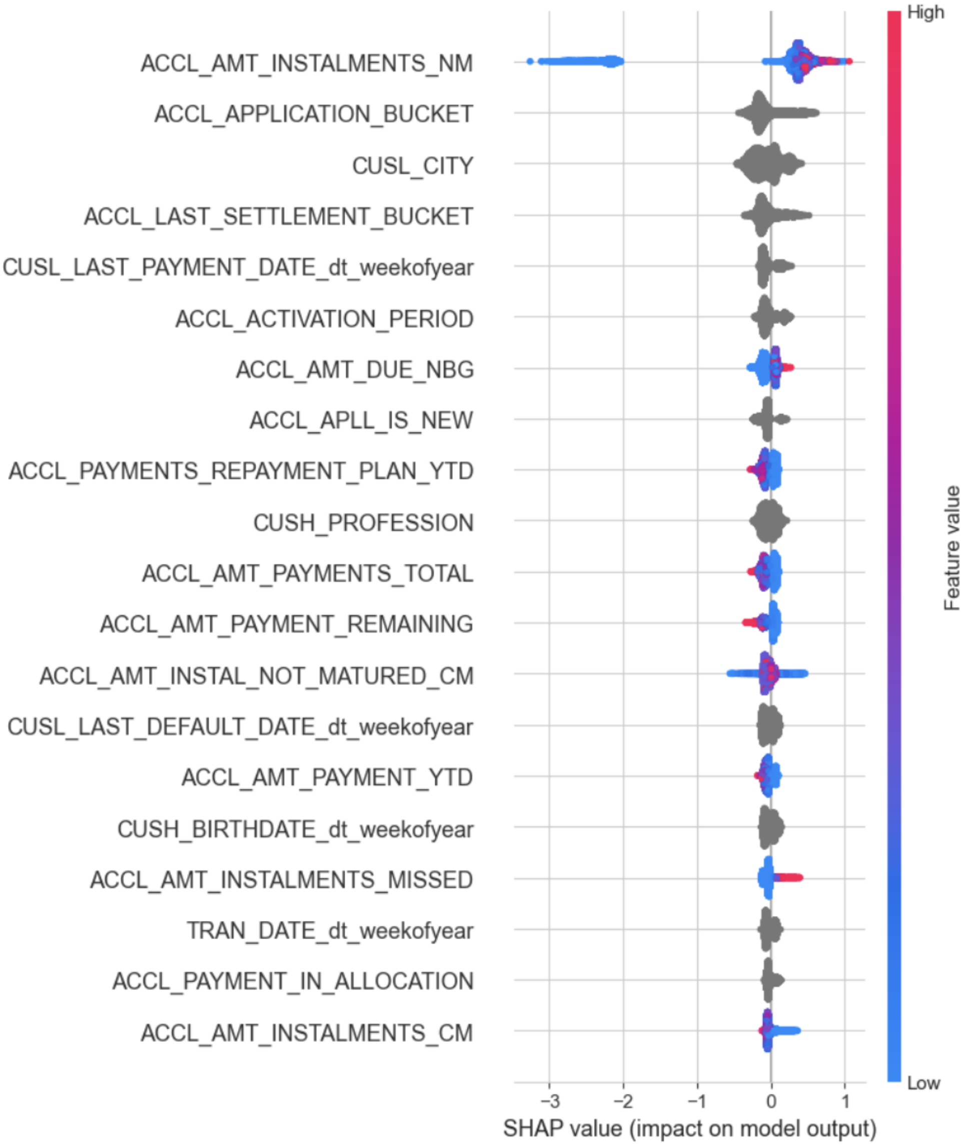
5.10 Findings

The evaluation proceeded in a stepwise fashion, guided by the overarching research questions and the principle of fair comparisons within relevant model groups. Performance was measured using the Area Under the ROC Curve (AUC) and validated through ten-fold cross-validation. Statistical significance was assessed using the Friedman test, followed by appropriate post-hoc Nemenyi tests to determine which models differed meaningfully from one another.

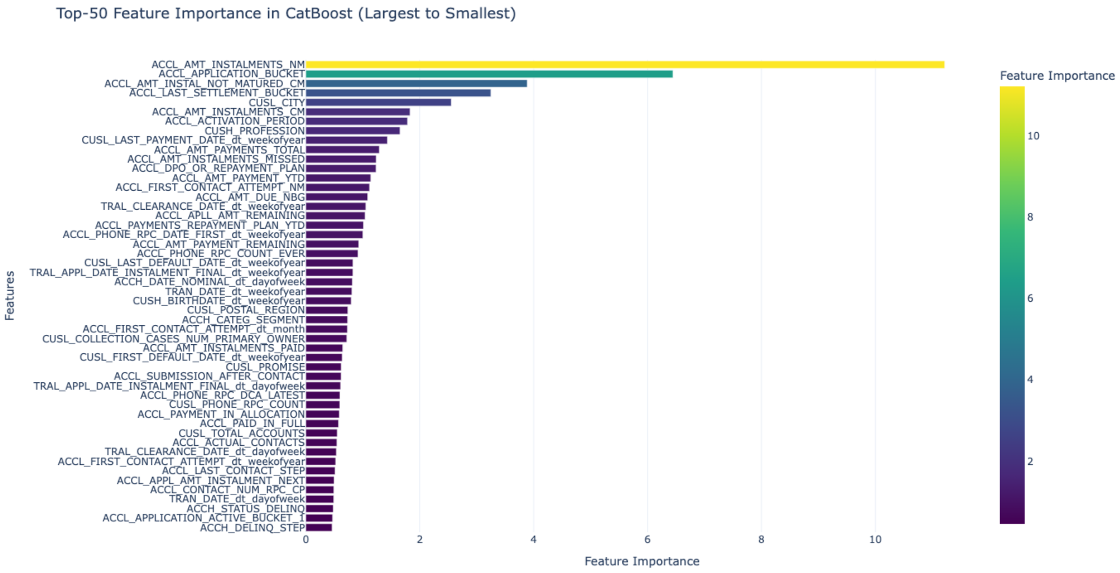
In the first stage (Q1), the analysis considered the baseline Logistic Regression model alongside the full-feature (ff) ensemble models. The Friedman test indicated a significant difference in performance among these models (X2 = 46.171, $p < 0.001$). Post-hoc comparisons confirmed that Logistic Regression's mean AUC (0.8418) was statistically outperformed by several of the ff ensemble methods. Notably, ff_CatBoostClassifier (mean AUC = 0.9507), ff_LGBM, and ff_XGB all emerged as substantially stronger predictors of loan default risk than the baseline. This finding supports the initial expectation that more complex, non-linear ensemble techniques can surpass traditional linear approaches in capturing the intricacies of credit risk data.

Subsequently, the second stage of analysis (Q2) bifurcated into two separate inquiries. For Q2_1, the focus centered exclusively on the ff ensemble models, thereby removing the baseline and allowing a fair comparison within this advanced class of algorithms. The Friedman test once again revealed a significant difference (X2 = 34.640, $p < 0.001$). While all ff ensemble models exhibited strong predictive capabilities, the post-hoc analysis highlighted that certain gradient boosting approaches, particularly ff_CatBoostClassifier, tended to achieve the highest AUC scores, although differences among top performers (ff_CatBoostClassifier, ff_LGBM, and ff_XGB) were subtler and, in some cases, not statistically important.

For Q2_2, the analysis turned to the feature-selected (fs) ensemble models to examine whether reduced feature subsets could maintain competitive performance. The Friedman test revealed a statistically significant difference (X2 = 19.840, p = 0.001) among these fs models. While no fs model matched the peak performance observed in the ff setting, certain fs models—such as fs_XGB and fs_LGBM—still achieved respectable AUC scores (approximately 0.85), underscoring that even a more parsimonious feature set can yield effective predictions. The post-hoc results indicated that differences among fs models were present, though generally less pronounced than those observed in the ff scenario. This can be attributed to limited available features compared to the ff models leading to less interactions and features that could potentially increase the predictive power of the models.



Beyond raw predictive performance, Q3 addressed explainability: understanding which features drive the model's decisions. SHAP (SHapley Additive exPlanations) analyses on the final CatBoost model identified a set of top predictive features. Among these, **ACCL_AMT_INSTALMENTS_NM**, representing the amount of next-month instalments (denormalized), stood out as the most influential. The importance of this feature aligns with the intuition that upcoming financial obligations shape default risk expectations—borrowers with substantial next-month instalment amounts may face greater payment pressures, thereby increasing their likelihood of defaulting.



Other key predictors included **ACCL_APPLICATION_BUCKET**, which categorizes the borrower's latest application state within the observation period, and **ACCL_AMT_INSTAL_NOT_MATUREDBQ**, reflecting still-maturing obligations that have not yet come due. Such variables capture nuanced facets of a borrower's credit lifecycle and indicate that both current loan structuring and anticipated future payment schedules strongly inform risk. Additionally, **ACCLL_AS_TSETTLEMENT** the bucket of the last settlement reached, highlights the importance of historical renegotiations or settlement patterns in predicting future payment behavior. Variables such as **CUSL_CITY** (customer's city) offer context—regional economic conditions, local market stability, or employment patterns—which may subtly influence a borrower's repayment capacity.

Collectively, this diverse feature set allowed CatBoost to identify complex, nonlinear patterns effectively and along with the SHAP results reveal that factors linked to how and when customers make (or miss) payments, how their loans evolve over time, and how their applications progress through various states provide crucial signals for predicting default risk. Such insights are instrumental in validating the model's predictions and offer actionable interventions.

5.11 Discussion

Finally, in Q4, the analysis considered both ff and fs ensemble models together to ascertain whether reducing features systematically impacted performance. The Friedman test demonstrated a significant overall difference (X2 = 83.040, $p < 0.001$). Post-hoc tests revealed that ff models, as a group, generally outperformed fs models, confirming that the richer, full-feature space contributed to stronger predictive performance. Among the ff models, CatBoost, LightGBM, and XGBoost consistently ranked near the top, whereas fs models, while competent, were clearly disadvantaged in direct comparisons against their ff counterparts.

The results collectively affirm the hypothesis that ensemble method exhibit meaningful differences compared to the logistic regression's performance and relative to one another, depending always on the feature space employed. The initial baseline comparison (Q1) has shown that traditional Logistic Regression, while widely used and interpretable, does not match the predictive power of advanced ensemble models like CatBoost, LightGBM, and XGBoost especially when trained with a comprehensive feature set. This is in consistent with the current research that states that linear models are inadequate for capturing the complex non-linear relationships while ensemble models are efficient in capturing the dynamics of credit risk [126].

Within the family of ff ensembles (Q2_1), the best-performing models exhibited subtle distinctions. The consistently high AUC values obtained by CatBoost, LightGBM, and XGBoost reinforce the idea that complex gradient boosting techniques can be highly effective for default prediction, especially when provided with a rich feature set. However, a statistical analysis of difference by performing pairwise comparisons of the top models showed that there was no statistically significant difference between these models, indicating that any of these state-of-the-art ensemble methods can be useful in credit risk management and other applications and that other factors such as training time and model complexity should be the main factors in determining use which in model most to cases.

Examining the fs ensembles (Q2_2) highlighted the value and potential limitations of dimensionality reduction. However, when compared to the ff setting, the predictive performance of many fs models was still fairly good. This indicates that even though feature engineering is advisable, it is feasible to develop good if not decent models from some of the features' subsets.

The final analysis (Q4), integrating both ff and fs ensembles, reinforced the conclusion that a richer feature representation yields more robust predictive models. The large differences in performance between ff and fs models show how important it is to properly engineer features for the credit risk models. This is because ensemble methods' strength lies in the diversity of the features used hence any reduction in feature space must be done while considering the possibility of eliminating vital features.

In sum, these findings offer practical insights for financial institutions and model developers. First, moving beyond logistic regression to advanced ensemble algorithms can seem to be able to substantially improve the early detection of borrowers at risk. Second, while feature selection has the potential to enhance the modeling process, it should be applied cautiously to avoid sacrificing predictive accuracy in the preocess. Third, among the considered ensembles, gradient boosting methods, particularly when applied to a comprehensive set of features, emerge as most promising candidates for creating more effective credit risk strategies in the Greek lending market.

Crucially, the interpretability analysis (Q3) moves the conversation beyond predictive metrics. The identification of key predictive features, ranging from imminent instalment amounts to settlement bucket, provides clarity on why the model rates certain borrowers as high-risk. This is not hust a technical study in explainability, it can really guide strategic decision-making since lenders may design early intervention strategies for customers identified as at risk due to high upcoming instalment burdens or consider revised settlement terms for borrowers repeatedly appearing in late-settlement buckets. For example, understanding the prominence of location-related variables might lead to regional policy adjustments or targeted outreach in areas associated with higher default trends.

Hence, it is evident that there is an need to incorporate powerful predictive techniques with the capacity of explaining themselves. Thus, it is possible to identify which characteristics – taken from a large set of features, including credit score, application status, payments plan, and demographic information – influence the default risk predictions, and financial institutions will not only increase their prediction precision but also achieve a deeper level of interpretation. This approach enables them to use the advanced ensembles in practice with confidence while at the same time being able to understand the mechanism of how credit risk is affected and therefore be in a position to make better, fair and more prudent lending policies.

While this study provides valuable insights into loan default prediction within the Greek lending market, it is essential to acknowledge certain limitations.

First, the nature of the data used in this study in the form of snapshots aggregated on a monthly basis may not capture the detailed time-series loan repayment behaviour data. This aggregation may hide important short term trends or seasonality that could be used to make better predictions. Future research could also attempt to use more comprehensive temporal data and thus, employ specific time-series models including recurrent neural networks or time-based transformers so as to capture the dynamics of the data in a better manner.

Secondly, although the analysis considered the predictive potential of missing values, it was not possible to determine the implications of missingness of different features clearly since there was not much information provided in the documentation. More understanding of the missing data, especially on key financial variables, may also help to improve the performance and the interpretability of the model.

Finally, the initial data structure presented challenges. The presence of four separate datasets with inconsistent identifier naming conventions required careful preprocessing and mapping to ensure data integrity. Clearer documentation and more standardized data saving practices would streamline future analyses and reduce the risk of errors.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

This thesis has aimed at identifying the main issue of the loan default prediction in the Greek context lending of market the with the help of the Qualco's data set. The research designed and conducted in this paper involves data cleaning and preprocessing, feature creation, machine learning model training, and evaluation to generate findings that are meaningful for financial institutions and the Greek economy.

Key Findings

The research has successfully addressed the core research questions, providing clear answers based on the empirical analysis:

- **Ensemble models significantly outperform the industry-standard Logistic Regression model.** The study also established this by showing that ensemble methods such as CatBoost, LightGBM, and XGBoost produced higher AUC than the baseline Logistic Regression model. This is because ensemble methods provide a better way of representing the non-linear relationships that exist in credit risk data.

- **Feature engineering and selection play a crucial role in model performance.** Although the model with all features provided better results, some hand-picked feature sets also provided reasonable performance. This also confirms the importance of proper feature selection as well as the possibility to create efficient models that will have high predictive power while at the same time being easy to compute and not too complicated.

- **The most significant factors influencing loan default risk include upcoming installment amounts, application status, and historical settlement patterns.** SHAP analysis of the most important features showed that the features that describe the payment plans, the application process, and the previous settlement of the borrowers were the most correlated with the default risk. This is a very useful information for the lenders to determine the borrowers risk and come up with measures to manage them.

Implications for the Greek Lending Market

The findings of this research have broader implications for the Greek lending market which has been an issue with non-performing loans. From this case, they learn how the advanced machine learning models can help in determining the likelihood of loan defaults hence enhancing the credit risk management system. This paper also shows how risk management especially in identifying high risk defaulters help in avoiding adverse impacts such as financial losses and hence contributing to stability of the Greek economy.

Practical Applications for Financial Institutions

The practical applications of this research are significant. Financial institutions can leverage the insights and methodologies presented here to:

- **Enhance credit scoring models:** For instance, since ensemble methods are used in combination with the most important features, lenders can improve the accuracy of their credit scoring models and make better lending decisions.
- **Develop early intervention strategies:** Based on the amount of next installment and the application status of high risk borrowers, it is possible to design interventions that can be made to avoid.
- **Optimize resource allocation:** It is thus possible to effectively manage risk of default by channeling efforts in collection, especially on accounts that are

likely to default while not wasting effort on accounts that are low risk

6.2 Future Work

This work can be enhanced and developed in a number of fruitful manners with emphasis on the research methods and its practical applications. Thus, how to tackle class inequality is one of the most pressing issues that should receive more consideration. Imbalanced datasets represent a significant challenge in credit scoring, as instances of default constitute merely a small portion of the entire dataset.

Methods such as synthetic data generation, including SMOTE variants or generative adversarial networks (GANs), may be utilized to balance the dataset. Another limitation is that the current dataset lacks sufficient and varied data which can be included to increase the size of the dataset. Social media and network data can be very useful in understanding the borrower's behaviour and their financial decisions. This would help in understanding the macroeconomic factors that affect the environment that the loans are being given, thus enabling the model to take into consideration the overall political and economic conditions of the region, country or the continent and thus being able to make better generalizations. The suggested changes will make the model more stable and will help to determine the factors that affect the loan default risk on both micro and macro levels.

Although AUC-ROC was the main measure to assess the performance of the models in this study, using only AUC-ROC can be misleading as it does not take into account other important properties of the models such as calibration, interpretability and the costs of misclassifications. The work should also go further to extend the evaluation framework to include other performance measures that will offer a better understanding of how the models deal with imbalanced data sets which is typical in credit risk assessment. Calibration curves could also be used to check on the certainty of the predicted probabilities which are very crucial in financial decisions since overconfidence in the predictions can cause poor decision making. There is also a need to investigate how the models can be used in the future by studying the relationship between sensitivity (recall) and specificity at various threshold values since this might help in understanding the impact of the models in a more nuanced way. For example, in the case of loan want default, to the focus institutions on may high recall to find as many borrowers at risk as possible even if it leads to more false positives. These trade-offs can be very useful in order to fine-tune the models in view of various business goals.Another area that can be explored in order to come up with new way of representing categorical and textual data. For instance, recordings of phone conversations between the borrowers and collection agencies can be combined with other features and thus create a comprehensive data set. Advanced encoding approaches, such embedding layers in deep learning models, autoencoders, or tree-based methods such as random tree embeddings, can be employed to obtain compact and significant representations of high-dimensional categorical data.

The following statistical validation methods such as bootstrapping and confidence interval estimates can be used in order to increase the robustness of the model performance assessment. Also, there are various techniques for sampling which may enhance the reliability of the assessment procedures and especially in the case of imbalanced datasets. The explainability of machine learning models is an important aspect in the context of credit management. Since specific requirements exist in the field for the transparency and accountability of the decision-making process, developing surrogates as interpretable proxies for complex machine learning models will help in understanding the predictions made by the models for the stakeholders. It is important for stakeholders to recognize and mitigate bias in order to ensure that the implementation of machine learning models is ethical and non-biased.

Subsequent research can focus on assessing any bias in the model especially using sensitive data such as age, gender or location. Thus, implementing the bias detection methods together with fairness-oriented algorithms, it is possible to create the forecasts that are both reliable and free from prejudice. Ultimately, the exploration of more elaborate techniques, including Graph Neural Networks (GNNs), represents a great potential. Graph Neural Networks (GNNs), which have demonstrated promising results in recent studies, can model the relationships between borrowers, lenders, and other factors as a graph structure. This method might provide connections and associations that may not be possible to capture well by the tabular data alone. This presents a completely new and powerful way of looking at the loan default prediction problem. All of these potential directions can possibility lift the current methodology and produce even greater results.

Appendix: Pairwise Comparison Results

.1 Comparison 1: Logistic Regression vs. Full-Feature Models

	b_logisticRegression	ff_CatBoostClassifier	ff_HistGradientBoosting	ff_LGBM	ff_RandomForest	ff_XGB
b_logisticRegression	1.000000	0.000000	0.260953	0.000412	0.706032	0.000142
ff_CatBoostClassifier	0.000000	1.000000	0.004537	0.629168	0.000244	0.777005
ff_HistGradientBoosting	0.260953	0.004537	1.000000	0.324066	0.979939	0.205990
ff_LGBM	0.000412	0.629168	0.324066	1.000000	0.065990	0.999894
ff_RandomForest	0.706032	0.000244	0.979939	0.065990	1.000000	0.033452
ff_XGB	0.000142	0.777005	0.205990	0.999894	0.033452	1.000000

Pairwise comparisons between Logistic Regression and full-feature models.

.2 Comparison 2: Full-Feature Models

	ff_CatBoostClassifier	ff_HistGradientBoosting	ff_LGBM	ff_RandomForest	ff_XGB
ff_CatBoostClassifier	1.000000	0.000214	0.351295	0.000004	0.526101
ff_HistGradientBoosting	0.000214	1.000000	0.114112	0.915320	0.055868
ff_LGBM	0.351295	0.114112	1.000000	0.010067	0.998603
ff_RandomForest	0.000004	0.915320	0.010067	1.000000	0.003730
ff_XGB	0.526101	0.055868	0.998603	0.003730	1.000000

Pairwise comparisons among full-feature models.

.3 Comparison 3: Feature-Selected Models

	fs_CatBoost	fs_HistGradientBoosting	fs_LGBM	fs_RandomForest	fs_XGB
fs_CatBoost	1.000000	0.789935	0.003730	0.002198	0.157115
fs_HistGradientBoosting	0.789935	1.000000	0.114112	0.080831	0.789935
fs_LGBM	0.003730	0.114112	1.000000	0.999910	0.708013
fs_RandomForest	0.002198	0.080831	0.999910	1.000000	0.618449
fs_XGB	0.157115	0.789935	0.708013	0.618449	1.000000

Pairwise comparisons among feature-selected models.

.4 Comparison 4: All Models

	ff_CatBoostClassifier	ff_HistGradientBoosting	ff_LGBM	ff_RandomForest	ff_XGB	fs_CatBoost	fs_HistGradientBoosting	fs_LGBM	fs_RandomForest	fs_XGB
ff_CatBoostClassifier	1.000000	0.445798	0.994324	0.190715	0.998421	0.003606	0.000284	0.000000	0.000000	0.000015
ff_HistGradientBoosting	0.445798	1.000000	0.963100	0.999990	0.926647	0.796635	0.395743	0.014370	0.011045	0.111001
ff_LGBM	0.994324	0.963100	1.000000	0.796635	1.000000	0.091175	0.014370	0.000069	0.000048	0.001455

Pairwise comparisons among all models.

thesis.bib

Bibliography

[1] *World bank annual report 2021 appendixes*, Sep. 2021.

[2] A. R. Khaki and A. Akın, “Factors affecting the capital structure: New evidence from gcc countries”, *Centre of Sociological Research, Ternopil, Ukraine*, vol. 13, no. 1, pp. 9–27, Mar. 2020.

[3] J. Geanakoplos, “Solving the present crisis and managing the leverage cycle”, *RELX Group (Netherlands)*, vol. undefined, no. undefined, Jan. 2010.

[4] V. Giannopoulos, “The effectiveness of artificial credit scoring models in predicting npls using micro accounting data”, *OMICS Publishing Group*, vol. 07, no. 04, Jan. 2018.

[5] D. T. Priyankara and E. A. G. Sumanasiri, “Determinants of microfinance loan default: An empirical investigation in sri lanka”, *Sciencedomain International*, vol. undefined, no. undefined, pp. 1–13, Aug. 2019.

[6] L. U. Oghenekaro and M. C. Chimela, “Design and implementation of a loan default prediction system using random forest algorithm”, vol. 22, no. 3, pp. 137–144, Jan. 2024.

[7] D. P. Louzis, A. T. Vouldis, and V. L. Metaxas, “Macroeconomic and bank-specific determinants of non-performing loans in greece: A comparative study of mortgage, business and consumer loan portfolios”, *Elsevier BV*, vol. 36, no. 4, pp. 1012–1027, Oct. 2011.

[8] F. Barboza, H. Kimura, and E. I. Altman, “Machine learning models and bankruptcy prediction”, *Elsevier BV*, vol. 83, no. undefined, pp. 405–417, Apr. 2017.

[9] G. Kou, Y. Peng, and C. Lü, “Mcdm approach to evaluating bank loan default models”, *Vilnius Gediminas Technical University*, vol. 20, no. 2, pp. 292–311, Jun. 2014.

[10] H. Kiefer and T. Mayock, “Why do models that predict failure fail?”, *RELX Group (Netherlands)*, Jan. 2020.

[11] S. Chen, Z. Guo, and X. Zhao, “Predicting mortgage early delinquency with machine learning methods”, *Elsevier BV*, vol. 290, no. 1, pp. 358–372, Aug. 2020.

[12] C. Serrano-Cinca, B. G. Nieto, and L. López-Palacios, “Determinants of default in p2p lending”, *Public Library of Science*, vol. 10, no. 10, e0139427–e0139427, Oct. 2015.

[13] J. Turiel and T. Aste, “Peer-to-peer loan acceptance and default prediction with artificial intelligence”, *Royal Society*, vol. 7, no. 6, pp. 191 649–191 649, Jun. 2020.

[14] S. Devi and Y. Radhika, “A survey on machine learning and statistical techniques in bankruptcy prediction”, vol. 8, no. 2, pp. 133–139, Apr. 2018.

[15] Q. Wang, “Research on the method of predicting consumer financial loan default based on the big data model”, *Wiley*, vol. 2022, no. undefined, pp. 1–9, Mar. 2022.

[16] S. Andrianova, B. H. Baltagi, P. Demetriades, and D. Fielding, “Why do african banks lend so little?”, *Wiley*, vol. 77, no. 3, pp. 339–359, Jun. 2014.

[17] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, “A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring”, *Elsevier BV*, vol. 27, no. undefined, pp. 11–23, Jul. 2015.

[18] E. Owusu, R. Quainoo, S. Mensah, and J. K. Appati, “A deep learning approach for loan default prediction using imbalanced dataset”, *IGI Global*, vol. 19, no. 1, pp. 1–16, Mar. 2023.

[19] K. Kohv and O. Lukason, “What best predicts corporate bank loan defaults? an analysis of three different variable domains”, *Multidisciplinary Digital Publishing Institute*, vol. 9, no. 2, pp. 29–29, Jan. 2021.

[20] A. Alonso and J. M. Carbó, “Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation”, *RELX Group (Netherlands)*, vol. undefined, no. undefined, Jan. 2021.

[21] R. Rahmani, M. Parola, and M. G. C. A. Cimino, *A machine learning workflow to address credit default prediction*, Jan. 2024.

[22] A. C. Bahnsen, D. Aouada, and B. Ottersten, “Example-dependent cost-sensitive logistic regression for credit scoring”, vol. undefined, no. undefined, pp. 263–269, Dec. 2014.

[23] F. Butaru, Q. Chen, B. J. Clark, S. Das, A. W. Lo, and A. R. Siddique, “Risk and risk management in the credit card industry”, *Elsevier BV*, vol. 72, pp. 218–239, Aug. 2016.

[24] T. Sun and M. A. Vasarhelyi, “Predicting credit card delinquencies: An application of deep neural networks”, *Wiley*, vol. 25, no. 4, pp. 174–189, Aug. 2018.

[25] L. Barbaglia, S. Manzan, and E. Tosetti, “Forecasting loan default in europe with machine learning”, *Oxford University Press*, vol. 21, no. 2, pp. 569–596, Apr. 2021.

[26] P. M. Addo, D. Guégan, and B. K. Hassani, “Credit risk analysis using machine and deep learning models”, *Multidisciplinary Digital Publishing Institute*, vol. 6, no. 2, pp. 38–38, Apr. 2018.

[27] O. Netzer, A. Lemaire, and M. Herzenstein, “When words sweat: Identifying signals for loan default in the text of loan applications”, *SAGE Publishing*, vol. 56, no. 6, pp. 960–980, Aug. 2019.

[28] C. Jiang, Z. Wang, R. Wang, and Y. Ding, “Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending”, *Springer Science+Business Media*, vol. 266, no. 1, pp. 511–529, Oct. 2017.

[29] B. Niu, J. Ren, and X. Li, “Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending”, *Multidisciplinary Digital Publishing Institute*, vol. 10, no. 12, pp. 397–397, Dec. 2019.

[30] S. H. Shetty and T. N. Vincent, “Corporate default prediction model: Evidence from the indian industrial sector”, *SAGE Publishing*, vol. 28, no. 3, pp. 344–360, Aug. 2021.

[31] M. Jumaa, M. Saqib, and A. Attar, “Improving credit risk assessment through deep learning-based consumer loan default prediction model”, *Hasan Dincer*, vol. 12, no. 1, pp. 85–92, Jun. 2023.

[32] W. Wu, “Machine learning approaches to predict loan default”, *Scientific Research Publishing*, vol. 14, no. 5, pp. 157–164, Jan. 2022.

[33] X. Ma, J. Sha, D. Wang, Y. Yu, Y. Qian, and X. Niu, “Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning”, *Elsevier BV*, vol. 31, no. undefined, pp. 24–39, Aug. 2018.

[34] S. Wang, H. Luo, S. Huang, *et al.*, “Counterfactual-based minority oversampling for imbalanced classification”, *Elsevier BV*, vol. 122, pp. 106 024–106 024, Mar. 2023.

[35] Y. Chen, J. Zhang, and W. W. Y. Ng, “Loan default prediction using diversified sensitivity undersampling”, vol. 2014, no. undefined, pp. 240–245, Jul. 2018.

[36] A. A. Egwa, “Default prediction for loan lenders using machine learning algorithms”, vol. 5, no. 1, pp. 1–12, Dec. 2022.

[37] “Loan default prediction in microfinance group lending with machine learning”, vol. undefined, no. undefined, Jan. 2023.

[38] G. EROL, B. Uzbaş, C. Yücelbaş, and Ş. Yücelbaş, “Analyzing the effect of data preprocessing techniques using machine learning algorithms on the diagnosis of covid-19”, *Wiley*, vol. 34, no. 28, Oct. 2022.

[39] H. Nugroho and K. Surendro, “Missing data problem in predictive analytics”, vol. 3, no. undefined, pp. 95–100, Feb. 2019.

[40] S. B. Babo and A. M. Beyene, *Bank loan classification of imbalanced dataset using machine learning approach*, Mar. 2023.

[41] I. Alreshidi, I. Moulitsas, and K. Jenkins, “Miscellaneous eeg preprocessing and machine learning for pilots’ mental states classification: Implications”, vol. undefined, no. undefined, pp. 29–39, Oct. 2022.

[42] V. Padimi, V. S. .., and D. D. Ningombam, “Applying machine learning techniques to maximize the performance of loan default prediction”, vol. undefined, no. undefined, pp. 44–56, Jan. 2022.

[43] Y. Xue, “Towards personal credit default prediction method based on data mining”, vol. undefined, no. undefined, Jan. 2023.

[44] *Scikit-learn user guide - feature selection*, Jan. 2024.

[45] J. Li, K. Cheng, S. Wang, *et al.*, “Feature selection”, *Association for Computing Machinery*, vol. 50, no. 6, pp. 1–45, Dec. 2017.

[46] H. J. Bature, D. D. Wisdom, T. T. Dufuwa, and I. O. Ayetuoma, “Credit default prediction system using machine learning”, vol. 3, no. 1, pp. 51–60, Jun. 2023.

[47] A.-H. Chang, L.-K. Yang, R.-H. Tsaih, and S.-K. Lin, “Machine learning and artificial neural networks to construct p2p lending credit-scoring model: A case using lending club data”, *AIMS Press*, vol. 6, no. 2, pp. 303–325, Jan. 2022.

[48] S. A. Akanmu and A. R. Gilal, “A boosted decision tree model for predicting loan default in p2p lending communities”, vol. 9, no. 1, pp. 1257–1261, Oct. 2019.

[49] R. Tibshirani, “Regression shrinkage and selection via the lasso”, *Oxford University Press*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[50] N. Abd-Alsabour, *A review on evolutionary feature selection*, Oct. 2014.

[51] O. Almomani, “A feature selection model for network intrusion detection system based on pso, gwo, ffa and ga algorithms”, *Multidisciplinary Digital Publishing Institute*, vol. 12, no. 6, pp. 1046–1046, Jun. 2020.

[52] G. Roffo, *Report: Feature selection techniques for classification.* Jul. 2016.

[53] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu, “Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending”, *Taylor Francis*, vol. 47, no. 1, pp. 54–70, Oct. 2014.

[54] V. Moscato, A. Picariello, and G. Sperlí, “A benchmark of machine learning approaches for credit score prediction”, *Elsevier BV*, vol. 165, no. undefined, pp. 113 986–113 986, Sep. 2020.

[55] M. S. Park, H. Son, C. Hyun, and H. J. Hwang, “Explainability of machine learning models for bankruptcy prediction”, *Institute of Electrical and Electronics Engineers*, vol. 9, no. undefined, pp. 124 887–124 899, Jan. 2021.

[56] M. Malekipirbazari and V. Aksakallı, “Risk assessment in social lending via random forests”, *Elsevier BV*, vol. 42, no. 10, pp. 4621–4631, Feb. 2015.

[57] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: From early developments to recent advancements”, *Taylor Francis*, vol. 2, no. 1, pp. 602–609, Sep. 2014.

[58] M. Shahhosseini and G. Hu, *Improved weighted random forest for classification problems*, Jan. 2021.

[59] J. Tang, S. Alelyani, and H. Liu, *Feature selection for classification: A review*, Jul. 2014.

[60] R. Srivastava, “Extrapolation of loan default using predictive analytics: A case of business analysis”, vol. 23, no. undefined, pp. 37–37, Jan. 2022.

[61] “Scikit-learn, “1.5: Cross-validation: Evaluating estimator performance,””, vol. undefined, no. undefined, Jan. 2024.

[62] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif, and S. Islam, “Application of machine learning in credit risk assessment: A prelude to smart banking”, vol. undefined, no. undefined, pp. 2023–2028, Oct. 2019.

[63] G. Varoquaux and O. Colliot, *Evaluating machine learning models and their diagnostic value*, Jan. 2023.

[64] Q. Zhu, W. Ding, M.-S. Xiang, M. Hu, and N. Zhang, “Loan default prediction based on convolutional neural network and lightgbm”, *IGI Global*, vol. 19, no. 1, pp. 1–16, Dec. 2022.

[65] Y. Peng, G. Wang, G. Kou, and Y. Shi, “An empirical study of classification algorithm evaluation for financial risk prediction”, *Elsevier BV*, vol. 11, no. 2, pp. 2906–2915, Dec. 2010.

[66] S. Mor, R. Aneja, S. Madan, and S. Gupta, “Artificial intelligence and loan default: The case of commercial banks in india”, *Wiley*, vol. 31, no. 6, pp. 571–580, Oct. 2022.

[67] H. Xiao, “The impact of macro-economic environment on probability of non-performing loans in financial institutions”, vol. 39, no. undefined, pp. 16–20, Feb. 2023.

[68] S. Basri and N. Kumar, “Non-performance of financial contracts in agricultural lending”, *Emerald Publishing Limited*, vol. 76, no. 3, pp. 362–377, Sep. 2016.

[69] Y. Zhou, “Loan default prediction based on machine learning methods”, vol. undefined, no. undefined, Jan. 2023.

[70] “Scikit-learn, “1.5: Linear models,””, vol. undefined, no. undefined, Jan. 2024.

[71] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, “Consumer credit risk: Individual probability estimates using machine learning”, *Elsevier BV*, vol. 40, no. 13, pp. 5125–5131, Mar. 2013.

[72] E. Hahami and D. Piper, “A meta-analysis evaluating the performance of machine learning models on probability of loan default”, *rScroll*, vol. 11, no. 2, May 2022.

[73] X. Dong, “Loan default prediction based on machine learning (lightgbm model)”, vol. 25, no. undefined, pp. 457–468, Aug. 2022.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, *“why should i trust you?” explaining the predictions of any classifier*, Jan. 2016.

[75] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: A review", *Scientific Research Publishing*, vol. 8, no. 4, pp. 341–357, Jan. 2020.

[76] A. Criminisi, J. Shotton, and E. Konukoglu, *Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning*, Jan. 2011.

[77] Q. Zhang, "Loan risk prediction model based on random forest", vol. 5, no. 1, pp. 216–222, Apr. 2023.

[78] L. Nguyen, M. Ahsan, and J. Haider, "Reimagining peer-to-peer lending sustainability: Unveiling predictive insights with innovative machine learning approaches for loan default anticipation", vol. 3, no. 1, pp. 184–215, Mar. 2024.

[79] A. V. Dorogush, V. Ershov, and A. Gulin, *Catboost: Gradient boosting with categorical features support*, Jan. 2018.

[80] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features", vol. 31, pp. 6639–6649, Dec. 2018.

[81] Y. Xia, L. He, Y. Li, N. Liu, and Y. Ding, "Predicting loan default in peer-to-peer lending using narrative data", *Wiley*, vol. 39, no. 2, pp. 260–280, Jul. 2019.

[82] W. Guo and Z. Z. Zhou, "A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction", *Wiley*, vol. 41, no. 6, pp. 1248–1313, Mar. 2022.

[83] J. Hancock and T. M. Khoshgoftaar, "Catboost for big data: An interdisciplinary review", *Springer Science+Business Media*, vol. 7, no. 1, Nov. 2020.

[84] E. A. Daoud, "Comparison between xgboost, lightgbm and catboost using a home credit dataset", vol. 13, no. 1, pp. 6–10, Jan. 2019.

[85] *Lightgbm's documentation*, Feb. 2023.

[86] G. Ke, Q. Meng, T. Finley, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree", Dec. 2017.

[87] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis", *Elsevier BV*, vol. 37, no. 1, pp. 534–545, May 2009.

[88] C.-F. Wu, S.-C. Huang, C.-C. Chiou, and Y.-M. Wang, *A predictive intelligence system of credit scoring based on deep multiple kernel learning*, Jul. 2021.

[89] Y. Hayashi, "Application of a rule extraction algorithm family based on the re-rx algorithm to financial credit risk assessment from a pareto optimal perspective", *Elsevier BV*, vol. 3, no. undefined, pp. 32–42, Jan. 2016.

[90] C. Modarres, M. Ibrahim, M. Louie, and J. Paisley, *Towards explainable deep learning for credit lending: A case study.* Nov. 2018.

[91] B. S. Trinkle and A. A. Baldwin, "Research opportunities for neural networks: The case for credit", *Wiley*, vol. 23, no. 3, pp. 240–254, May 2016.

[92] L.-L. Zeng, J. Sun, and Y. Zhou, "Auto loan default prediction based on stacking model", vol. undefined, no. undefined, pp. 286–292, Jan. 2023.

[93] S. Zandi, K. Korangi, M. Óskarsdóttir, C. Mues, and C. Bravo, *Attention-based dynamic multilayer graph neural networks for loan default prediction*, Jun. 2024.

[94] M. Dudík, W. Chen, S. Barocas, *et al.*, "Assessing and mitigating unfairness in credit models with the fairlearn toolkit", Sep. 2020.

[95] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models", *KeAi*, vol. 6, no. 3, pp. 123–133, May 2023.

[96] O. X. Kuiper, M. van den Berg, J. van der Burgt, and S. Leijnen, *Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities*, Jan. 2022.

[97] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, Jan. 2017.

[98] B. H. Misheva, A. Hirsa, J. Osterrieder, O. Kulkarni, and S. F. Lin, “Explainable ai in credit risk management”, *RELX Group (Netherlands)*, Jan. 2021.

[99] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, *Fooling lime and shap: Adversarial attacks on post hoc explanation methods*, Nov. 2019.

[100] W. Yang, Y.-C. Wei, H. Wei, *et al.*, “Survey on explainable ai: From approaches, limitations and applications aspects”, *Springer Nature*, vol. 3, no. 3, pp. 161–188, Aug. 2023.

[101] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, Jan. 2017.

[102] W. Samek, T. Wiegand, and K.-R. Müller, *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*, Jan. 2017.

[103] M. Hossin and S. M.N, *A review on evaluation metrics for data classification evaluations*, Mar. 2015.

[104] *Scikit-learn developers, ”model evaluation”*, Oct. 2024.

[105] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, *Algorithms for hyper-parameter optimization*, Dec. 2011.

[106] *Scikit-learn developers, ”model-selection,” scikit-learn: Machine learning in python, 2024*, Jan. 2024.

[107] D. Ari and B. B. Alagöz, “Dehypgpols: A genetic programming with evolutionary hyperparameter optimization and its application for stock market trend prediction”, *Springer Science+Business Media*, vol. 27, no. 5, pp. 2553–2574, Oct. 2022.

[108] R. de Sousa Maximiano, V. A. de Santiago Júnior, and E. H. Shiguemori, “On the benefits of automated tuning of hyper-parameters: An experiment related to temperature prediction on uav computers”, vol. undefined, no. undefined, pp. 509–520, Nov. 2022.

[109] N. L. Torrent, G. Visani, and E. Bagli, *Psd2 explainable ai model for credit scoring*, Jan. 2020.

[110] “Scikit-learn developers, ”1.11. gradient boosted trees,” scikit-learn user guide”, vol. undefined, no. undefined, Jan. 2024.

[111] P. Hall, *On the art and science of machine learning explanations*, Jan. 2018.

[112] S. B. Jabeur, C. Gharib, S. Mefteh-Wali, and W. B. Arfi, “Catboost model and artificial intelligence techniques for corporate failure prediction”, *Elsevier BV*, vol. 166, no. undefined, pp. 120 658–120 658, Feb. 2021.

[113] S. Acharya, *Comparative analysis of classification accuracy for xgboost, lightgbm, catboost, h2o, and classifium*, Oct. 2021.

[114] A. Gramegna and P. Giudici, “Shap and lime: An evaluation of discriminative power in credit risk”, *Frontiers Media*, vol. 4, Sep. 2021.

[115] A. T. Jebb, S. Parrigon, and S. E. Woo, “Exploratory data analysis as a foundation of inductive research”, *Elsevier BV*, vol. 27, no. 2, pp. 265–276, Aug. 2016.

[116] D. Micci-Barreca, “A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems”, *Association for Computing Machinery*, vol. 3, no. 1, pp. 27–32, Jul. 2001.

[117] L. Sasse, E. Nicolaisen-Sobesky, J. Dukart, *et al.*, *On leakage in machine learning pipelines*, Jan. 2023.

[118] Oct. 2004.

[119] A. V. Dorogush, V. Ershov, and A. Gulin, *Catboost: Gradient boosting with categorical features support*, Jan. 2018.

[120] undefined, *Targetencoder*, Jan. 2007.

[121] *Cross-validation: Evaluating estimator performance*, Jan. 2007.

[122] G. Wang, J. Ma, and S. Yang, “An improved boosting based on feature selection for corporate bankruptcy prediction”, *Elsevier BV*, vol. 41, no. 5, pp. 2353–2361, Sep. 2013.

[123] K. J. Leonard, “Credit-scoring models for the evaluation of small-business loan applications”, *Oxford University Press*, vol. 4, no. 1, pp. 89–95, Jan. 1992.

[124] L. Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, “The choice of scaling technique matters for classification performance”, *Elsevier BV*, vol. 133, pp. 109 924–109 924, Dec. 2022.

[125] R. J. Irwin and T. Irwin, “Appraising credit ratings: Does the cap fit better than the roc?”, *Wiley*, vol. 18, no. 4, pp. 396–408, Aug. 2013.

[126] M. Zhu, Y. Zhang, Y. Gong, K. Xing, Y. Xu, and J. Song, “Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble”, May 2024.