Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

CREDIT SCORING USING MACHINE LEARNING - CAUSAL INFERENCE AND FORECASTING

MAFALDA GOMES GASPAR

FEBRUARY - 2025

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
DISSERTATION

CREDIT SCORING USING MACHINE LEARNING - CAUSAL INFERENCE AND FORECASTING

MAFALDA GOMES GASPAR

SUPERVISION:
ADRIANA CORNEA-MADEIRA

FEBRUARY - 2025

GLOSSARY

AUC - Area Under the Receiver Operating Characteristic Curve

GAMLA - Generalized Additive Model with LASSO

KNN - K-Nearest Neighbors

KS - Kolmogorov-Smirnov

LASSO - Least Absolute Shrinkage and Selection Operator

OLS - Ordinary Least Squares

PLTR - Penalised Logistic Tree Regression

ROC - Receiver Operating Characteristics

RSS – Sum of Squared Residuals

ABSTRACT

Nowadays, the world is in constant changes and improvements, particularly in fields such as technology, with data and its applications becoming increasingly important, making it a fundamental aspect of the modern world. In this context, the present study integrates machine learning techniques with econometrics to analyze credit risk, focusing on both predicting defaults and understanding the key personal characteristics that drive default risk.

To achieve this, machine learning and econometric models were applied - Decision Tree, Generalized Additive Model with Least Absolute Shrinkage and Selection Operator, Gradient Boosting, K-Nearest Neighbors, Logistic Regression, Naïve Bayes and Random Forest - allowing for a comparative performance analysis. Additionally, the study discusses some advantages and limitations associated with the use of machine learning in credit scoring.

The results indicate that Gradient Boosting outperformed other methods, aligning with findings in the literature that highlight its effectiveness in handling imbalanced datasets and providing high accuracy in credit scoring assessment. Beyond performance comparison, the study also explores the key factors influencing default risk, particularly through the application of the Generalized Additive Model with Least Absolute Shrinkage and Selection Operator. The three most relevant variables identified are late payments, frequent missed payments, and high credit utilization.

This study underscores the benefits of adopting new technologies in credit risk management and, consequently, in daily life. Furthermore, it highlights areas for future research, such as integrating alternative data sources to enhance predictive power, improving explainability in model decisions, and developing techniques to strengthen data privacy, ensuring that sensitive information is protected during model training and application.

KEYWORDS: Credit Score, Default, Machine Learning, Models, Performance.

JEL CODES: C45; C53; C63; G17; G21; G32.

RESUMO

Atualmente, vivemos num mundo em constante mudança, onde a tecnologia desempenha um papel cada vez mais importante, e os dados e as suas aplicações têm vindo a tornar-se fundamentais no nosso dia a dia. Neste contexto, o presente estudo integra técnicas de *Machine Learning* com econometria, com o objetivo de analisar o risco de crédito, focando tanto na previsão de inadimplência quanto na identificação das principais características pessoais que influenciam esse risco.

Para tal, modelos de *Machine learning* e econométricos foram testados - *Decision Tree, Generalized additive Model with Least Absolute Shrinkage and Selection Operator, Gradient Boosting, K-Nearest Neighbors, Logistic Regression, Naïve Bayes and Random Forest* - permitindo uma análise mais detalhada do desempenho de cada modelo. Adicionalmente, o estudo aborda algumas vantagens e limitações do uso de *Machine Learning* na avaliação de crédito.

Os resultados obtidos indicam que o modelo *Gradient Boosting* foi o método com melhor desempenho, o que se alinha com as conclusões da literatura, que destaca esta técnica pela sua eficácia em lidar com dados com distribuição desproporcional e pela sua alta precisão na avaliação de crédito. Além da comparação de desempenho, o estudo também explora os principais fatores que influenciam o risco de inadimplência, particularmente por meio da aplicação do GAMLA. As três variáveis mais relevantes identificadas são os pagamentos em atraso, a frequência de faltas nos pagamentos e a alta utilização de crédito.

Este estudo menciona benefícios do uso de novas tecnologias na gestão de risco de crédito e, consequentemente, na vida quotidiana. Adicionalmente, destaca áreas para investigação futura, como a integração de fontes de dados alternativas, a melhoria da explicação das decisões dos modelos e o desenvolvimento de técnicas que reforcem a privacidade dos dados, assegurando que as informações sensíveis sejam protegidas durante o processo de treino e aplicação dos modelos.

Palavras-Chave**:** Risco de Crédito, Inadimplência, *Machine Learning*, Modelos, Desempenho

JEL : C45; C53; C63; G17; G21; G32.

TABLE OF CONTENTS

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to profoundly express my gratefulness to Professor Adriana Cornea-Madeira for allowing me to explore such an interesting topic and for all the help provided during these past months.

To my parents, who have always been supportive and ensured that, throughout my life, I had all the opportunities to grow both educationally and personally, thank you.

Last but not least, to all my friends who listened to my complaints and gave me motivation since the beginning of this chapter in my life, thank you for making me believe I was capable and for reminding me to never give up when facing challenges, since this is not the first and will not be the last.

INTRODUCTION

The importance of data and its applications has increased significantly in recent years, becoming fundamental pillars of the modern world. Data has never been more important than it is now, from personal financial management to business decision-making. This project arises from the need to explore this topic, combining two extremely relevant areas: credit risk analysis and machine learning, with the goal of integrating these two approaches in the study of credit default.

The study aims to apply machine learning techniques and econometric models to forecast default risk, with the goal of finding the most effective model for this prediction. At the same time, it seeks to analyze the personal characteristics that influence default risk, combining machine learning techniques with econometrics to better understand the factors driving default behavior.

The project includes a practical part, where a dataset from Kaggle called "*Give Me Some Credit*", focusing on credit defaults, is used to implement various machine learning techniques. These techniques are then compared to identify the most appropriate method for the dataset and to assess whether the results are consistent with the findings of existing studies on the topic.

Existing literature on credit default primarily focuses on improving risk prediction, but there is a gap in addressing the specific factors driving default behavior. While machine learning techniques have shown strong results in forecasting risk, understanding the underlying causes of default is an area that remains unexplored. This study aims to address this gap by applying models that combine machine learning with econometrics to identify and analyse the key variables that contribute to default risk, offering new insights into the drivers behind credit defaults.

Credit management and risk assessment impact both individuals and financial institutions. Machine learning is becoming a powerful tool in the field of behavior prediction and fraud prevention, where technology is becoming more and more important. Considering its contemporary relevance, the connection between machine learning and credit scoring emerges as an ideal topic for exploration.

After reviewing the literature, two questions were identified as interesting to guide this study:

- What are the most effective machine learning methods for predicting credit defaults?
- What are the key indicators that influence the risk of default?

These questions underscore the need to understand not only the general application of machine learning in the financial sector but also its practical use in real world scenarios. This work aims to contribute to a comprehensive analysis of these techniques, enhancing both the efficacy and the accuracy of default prediction.

This study is not just technical but also reflective, as it explores the impact of modern tools used on everyday decisions. By combining data science with a practical and relevant application in the financial context, this study aspires to make an interesting and valuable contribution.

Following this introduction, the work begins with the literature review in Section 1, where the concept of credit scoring and its importance in the financial context will be explored. Then, a discussion of advanced machine learning methods and traditional ones for analyzing credit risk. Furthermore, advantages of using machine learning will be mentioned throughout the text, along with a description of the challenges involved.

In Section 2, the methodology is described, including the techniques applied in this study and the models tested. Next, the practical part begins with a detailed description of the dataset used, including its characteristics and variables, as well as the preprocessing steps applied to prepare the dataset for modeling. In Section 3, the evaluation metrics chosen are explained and then the results obtained from each model are presented, analyzed and compared. The best machine learning method for predicting default in credit scoring, based on both the dataset used and existing literature will be presented.

Finally, the study concludes with a reflection on the findings gained throughout the project and future areas to explore.

1. LITERATURE REVIEW

This section aims to explore the fundamental concepts of credit scoring and machine learning, as well as describe their role in default prevention. It will address credit scoring, the use of machine learning in the financial sector and its applications in credit risk assessment. The advantages and limitations of using machine learning in credit scoring will be analyzed, followed by a discussion of the principal methods identified in recent studies. This approach provides a theoretical framework that supports the practical analysis conducted in this work.

1.1. Credit Scoring

Credit scoring is the process of evaluating the risk of a client meeting or failing to meet their financial obligations. As mentioned by Hand & Henley (1997), the principal objective of credit scoring is to classify the clients in two groups: the "good", representing those likely to meet their financial obligations, and the "bad", representing those, likely to default. To achieve this objective, a credit scorecard is used, it consists in some characteristics of the client that are assigned a numerical score equivalent to their risk level. This score is then compared to a predefined threshold to inform the decision-making about credit approval (Kennedy, 2013).

Initially, credit scoring models were developed as quantitative methods to support decisions about whether to grant credit or not. Although it is still used nowadays, the evolution of markets, the increasing complexity of financial systems, and the growth in the volume of available data have driven the development of more precise models. It can also be seen as a classification task, where clients are categorized based on observed attributes (Min & Lee, 2008) like credit history, their financial condition, macroeconomic factors, all of them contributing to assessing the probability of default (Bello, 2023).

Advances in data technologies and big data have introduced new approaches to credit scoring. Data from social networks, for example, provides alternative information about client profile, particularly for underrepresented groups or clients with limited financial histories. Big data techniques allow the processing of large volumes of structured data, semi-structured and unstructured data, facilitating the extraction of relevant insights for the banking sector (Tounsi et al., 2017).

1.2. Machine Learning

Machine learning is a subfield of Artificial Intelligence that creates algorithms capable of learning from analyzed data, allowing for predictions and decisions to be made without the need of programming. Janiesch et al. (2021) note that it automates the creation of analytic models to detect patterns and relationships in large volumes of data, enabling systems to improve their performance over time based on the experience they accumulate.

Machine Learning enables the extraction of valuable insights from complex data, in tasks like classification, regression and clustering. For example, in financial environments, it helps identify hidden patters and predict future behaviors, which improves accuracy and efficacy (Bello, 2023). Notable advancements have been made, allowing more robust and comprehensive analyses, along with the capacity to handle vast amounts of information, which is crucial to make rapid and informed decisions. Although there have been significant advancements in credit scoring, it also presents challenges.

1.3. Methods Used in Credit Scoring

1.3.1. Traditional Methods

The traditional methods most cited in credit scoring literature are valued for their simplicity, interpretability, and regulatory compliance. The methods are as follows:

Logistic Regression – It is considered the industry standard for risk assessment models. Studies by Li & Hand (2002) and Lee & Chen (2005) confirm the predominance of this method in risk assessment due to its simplicity and consistent performance. However, it is important to note that its limited capacity for modeling non-linear relationships reduces its effectiveness in more complex datasets.

Linear Discriminant Analysis (LDA)– According to West (2000), this was the first model employed in credit scoring, being a simple parametric statistical technique. Mentioned by Hand & Henley (1997) and Reichert et al. (1983) as a frequent method, it is used when data exhibits normality and homogeneity of variance.

Decision Trees – A method that recursively splits data into branches to improve classification accuracy, known for being intuitive and easy to interpret, it has been studied as a benchmark in credit scoring applications by Yap et al. (2011) and Kao et al. (2012).

1.3.2. Advanced Machine Learning Methods

Advanced methods commonly discussed in credit scoring literature handle complex, non-linear relationships and large datasets, enhancing the performance while making the model more challenging to interpret. These methods include:

Random Forest – Known for being robust and accurate when dealing with complex datasets, including those with outliers (Liu et al., 2021). It generates multiple versions of the model during training, reducing variance and improving reliability (Bello, 2023).

Gradient Boosting – Highlighted in benchmarks due to its high precision and ability to handle imbalanced data (Bello, 2023; Liu et al., 2021). It enhances the performance by focusing on instances that were previously misclassified.

Neural Networks – Discussed in studies such as Malhotra & Malhotra (2002) and Min & Lee (2008), this model is widely used in credit risk assessment due to its ability of modelling complex non-linear relationships. However, its lack of interpretability, limit its application in a regulatory context (Chen et al., 2024).

Dumitrescu et al. (2022) suggest integrating logistic regression with machine learning models to improve both accuracy and interpretability. In their work, they introduced the Penalised Logistic Tree Regression (PLTR), which outperformed traditional logistic regression in measures such as the Area Under the Receiver Operating Characteristic Curve (AUC) and the Proportion of Correct Classification. PLTR also demonstrated competitive performance with Random Forest while offering better interpretability due to its simpler decision rules.

Advances in credit scoring techniques not only enhance the decision-making processes regarding credit but also expand access to credit for historically neglected groups, while, at the same time, increasing efficiency and reducing risks for financial institutions (Tounsi et al., 2017).

1.4. Challenges of Machine Learning in Credit Scoring

There are several limitations when applying machine learning in credit scoring that must be considered:

I. Imbalanced Data – This is common in credit scoring since the number of defaulters is much smaller than that of non-defaulters. This imbalance negatively impacts the performance of machine learning models, as they tend to favor the majority class, reducing the effectiveness of recognizing minority cases (Chen et al., 2024; Tounsi et al., 2017);
II. Feature Selection - An essential pre-processing step, but often slow and complicate. As the dataset expands, it becomes challenging to identify and generate relevant features (Tounsi et al., 2017);
III. Interpretability - Many algorithms, such as Ensemble methods, are frequently considered as "black-boxes" , complicating the comprehension of the decision-making process for both clients and regulators (Dumitrescu et al., 2022). This raised concerns regarding transparency in Artificial Intelligence models, which are critical in the credit industry (Valdrighi et al., 2024);
IV. Privacy and Data Security – Financial institutions manage confidential information, such as personal details and transaction records. Keeping the data secure and safe presents a challenge, due to the risks of cyberattacks and data breaches (Bello, 2023); and,
V. Noisy and Non-linearity Data – Often, data contains errors, missing values and outliers, compromising the reliability of the models. Pre-processing and robustness cleaning techniques are required to address these problems (Tounsi et al., 2017).

Bello (2023) emphasizes that the dynamic nature of financial markets requires regular updates to machine learning models. These models must be adapted to answer the constant changes, requiring continuous monitoring and robust revalidation practices. These limitations highlight the importance of responsible and organized strategies to implement machine learning in credit scoring.

2. METHODOLOGY

The aim of this research is to perform a comparison of forecasting performance between different machine learning methods and traditional econometric models for credit scoring. For this, the training dataset from the "*Give Me Some Credit*" Kaggle competition will be used. The dataset will be split into training and testing subsets, allowing for the application of different machine learning models and then the performance of the models will be compared using several evaluation metrics. This comparative analysis aims to identify which models are the most effective for credit scoring based on their forecasting ability.

The Generalized Additive Model with Least Absolute Shrinkage and Selection Operator (GAMLA) was chosen for this study due to its ability to combine the predictive power of machine learning with the interpretability of econometric models. Introduced by Flachaire, Hacheme, Hué, and Laurent (2022), GAMLA builds upon the Generalized Additive Model (Hastie and Tibshirani, 1990), incorporating the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) for variable selection. This enhancement allows the model to effectively handle non-linear components and interaction effects, while reducing the risk of overfitting.

According to Flachaire et al. (2022), the GAMLA is a partially linear regression model designed to capture both linear and non-linear effects within a dataset. This approach enhances traditional regression models by incorporating flexible, non-parametric components while retaining the interpretability of linear models. The model is mathematically represented as:

$$y = \mathrm{Z}\gamma + \sum_{j=1}^{p} g_j\left(X_j\right) + \varepsilon \quad (1)$$

where

- y represents the dependent variable,
- Z is a vector of explanatory variables that enter the model linearly,
- γ as the corresponding parameter vector,
- $g_j(X_j)$ denotes smooth, non-parametric functions applied to the explanatory variables X_j, capturing complex non-linear relationships, and,
- ε represents the error term.

The linear component ($Z\gamma$) ensures interpretability, as it allows for straightforward computation of marginal effects, while the non-linear functions ($g_j(X_j)$) offer flexibility in capturing complex patterns in the data.

To ensure robustness and improve estimation accuracy, in their work, Flachaire et al. (2022) propose a variable selection process that integrates the LASSO or Autometrics. Their method builds on the double residual approach (Robinson, 1988), which involves estimating residuals in two steps: first, extracting non-linear components using Generalized Additive Models, and then applying LASSO to select the most relevant linear and non-linear predictors. The LASSO is formulated to minimize the following objective function:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (2)$$

where

- y represents the dependent variable,
- β_0 is the intercept term,
- β_j denotes the regression coefficients,
- x_{ij} represents the j-th predictor for the i-th observation .
- λ is a regularization parameter that controls the amount of shrinkage applied to the coefficients, and,
- p denotes de number of predictors.

This sequential process ensures that linear terms are correctly estimated even in the presence of non-linear dependencies, enhancing both model accuracy and interpretability. By combining machine learning techniques with econometric principles, GAMLA strikes a balance between flexibility, predictive power, and interpretability, making it particularly well-suited for applications where both non-linear patterns and interaction effects play a crucial role.

A minor adjustment was made in the variable selection process, instead of using LASSO, Adaptive LASSO was applied. This modification was made based on testing both methods, and Adaptive LASSO outperformed standard LASSO in terms of model performance. The key difference between them lies in the penalization process: while LASSO applies the same penalty across all variables, Adaptive LASSO adjusts the

penalties based on the variables' importance, with larger penalties for less relevant variables and smaller penalties for more relevant ones.

To enhance the model's ability to identify and analyze the most important factors influencing default risk, Ordinary Least Squares (OLS) is incorporated into the GAMLA method, allowing for statistical inference on the parameters of interest. This approach can be justified by the Frisch-Waugh-Lovell (1933) theorem, which demonstrates that the coefficients obtained from a full regression are identical to those obtained from a regression on the residualized variables, where the influence of the control variables has been controlled for. The equivalence between these two procedures extends to the residuals and, under certain assumptions, to the estimated standard errors, ensuring the validity of the inference.

The other methods that were tested to forecast are: Decision Tree, Gradient Boosting, K-Nearest Neighbors (KNN), Logistic Regression, Naïve Bayes and Random Forest. As the primary focus is to understand the GAMLA model, which has been addressed above, the steps followed are now presented in Table 1.

Table 1 - GAMLA Steps

Steps	Description
1	Define the binary variable y_i (1 for default, 0 for non-default) and the interest variables $x_{j,i}$ and control variables $z_{s,i}$
2	Split the data into training and testing subsets
3	Estimate, using OLS on the training sample: $y_i = f(z_i) + u_i$, (3) and save the residuals $\hat{u}_i = y_i - \hat{f}(z_i)$ (4)
4	For each variable $x_{j,I}$ estimate by OLS: $x_{j,i} = g(z_i) + v_{j,I}$ (5) and save the residuals $\hat{v}_{i,j} = x_{j,i} - \hat{g}(z_i)$ (6)
5	Combine the residuals into a vector: $v_i = (v_{1,i}, \dots, v_{k,i})'$
6	Use Adaptive LASSO to perform variable selection in the regression: $\hat{u}_i = \delta v_i + \varepsilon$ (7)
7	Based on the residuals selected in the previous step, determine the subset of variables $x_{j,i}$ that were selected. Denote this subset as x_i^*
8	On the training sample, estimate by OLS: $y_i = \gamma x_i^* + bz_i + \varepsilon$ (8)
9	Make out-of-sample predictions.

The steps outlined in the table describe a procedure that combines variable selection with predictive modelling. In this approach, the functions f and g, represent the systematic effect of control variables on the target variable and the interest variables, respectively. These functions can be estimated using machine learning techniques. Methods such as LASSO provide a flexible approach to model complex relationships in the data while preventing overfitting. The key feature of this methodology is the use of Adaptive LASSO regression in step 6, which performs automatic variable selection by identifying the most relevant explanatory variables from the residualized interest variables. In step 8, the final model allows for statistical inference on the coefficient γ, representing the effect of the selected variables on the probability of default. Standard t-

tests can then be applied to assess the significance of these variables, offering valuable insight into the primary drivers of credit default.

Now, a brief description of the other methods employed is provided bellow. The information is primarily sourced from James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013), in their work *An Introduction to Statistical Learning*.

Decision Trees – A method used for both regression and classification tasks. In the case of classification trees, the objective is to predict a qualitative response by assigning each observation to the most common class within the terminal node. The tree is built through recursive binary splitting, where the predictor space is iteratively divided to create distinct regions that better separate the classes. A new observation is classified based on the majority class in the corresponding terminal node. Decision Trees can handle both continuous and categorical predictors, assigning categorical values to specific branches when necessary. However, they are prone to overfitting, making them highly sensitive to the training data. To address this issue and enhance generalization performance, techniques such as pruning and cross-validation are often employed.

Gradient Boosting – It trains decision trees sequentially, with each new tree correcting the errors of the previous ones. Instead of fitting multiple models in parallel, gradient boosting builds its model gradually by focusing on residuals errors left by previous trees. The method employs regularization techniques such as shrinkage, to reduce the impact of each tree via a small parameter (λ), preventing overfitting.

However, gradient boosting can become complex and difficult to interpret, particularly as the number of trees increases. While it handles imbalanced data effectively, it can also be sensitive to noisy data and may overfit if not carefully tuned. The method requires fine-tuning parameters such as the number of trees, tree depth, and learning rate, and cross-validation is used to determine the optimal configuration.

KNN – Simple and intuitive classifier that assigns labels to new instances based on the majority class of its closest neighbors in the feature space. Given a positive integer K and a test observation x_0 , the KNN classifier first identifies the K closest points in the training data to x_0, typically using a distance metric such as Euclidean distance. Once these distances are calculated, the algorithm estimates the conditional probability for each class by taking the fraction of points in the neighborhood that belongs to each

class. The final step of KNN is to classify the test observation x_0 to the class with the highest estimated probability. This makes KNN a powerful method for problems where the relationship between features and labels is complex and non-linear.

KNN does not require an explicit training phase, instead it stores the entire training dataset and classifies new data points based on their proximity to existing data during prediction. However, this approach can be inefficient, as it requires significant memory for large datasets and becomes slower with high-dimensional data due to the need to calculate the distance to every point in the training set.

Logistic Regression – This method models the relationship between a binary outcome and one or more predictor variables by estimating a linear combination of the predictors. The output from this linear combination is then transformed using the logistic function, to ensure that the result is a probability between 0 and 1. Each predictor variable has an associated coefficient, such as β_0 and β_1, that represents the change in the log-odds of the outcome for a one-unit change in the predictor. These coefficients are estimated by maximizing the likelihood of the observed data, known as maximum likelihood estimation.

Logistic Regression is particularly useful for understanding the influence of individual predictors on the likelihood of an outcome. The model's simplicity and interpretability make it an attractive choice, especially in applications where the relationship between predictors and the outcome is approximately linear.

Naïve Bayes – A classifier based on Bayes' theorem that calculates the posterior probability of a class given an observation and its features. The model first estimates the prior probability of each class and the likelihood of the features, assuming that the features are conditionally independent given the class (Rish, 2011). For each class, the algorithm multiplies the probabilities of each feature given the class, then applies Bayes' theorem to produce the posterior probability. The class with the highest posterior probability is selected as the predicted class for the observation.

Naïve Bayes is computationally efficient due to its simplicity, requiring only the calculation and multiplication of probabilities rather than learning complex relationships between features.

Random Forest –This method builds multiple decision trees using bootstrapped training samples, introducing randomness at each split to enhance predictive performance. At each node, rather than considering all available predictors, a random subset of m predictors (typically $m \approx \sqrt{p}$, where p is the total number of predictors) is selected as candidates for the split. This approach prevents any single predictor from dominating the tree-building process and ensures that different trees capture diverse patterns in the data. By forcing the trees to consider different subsets of predictors, random forests help reduce the correlation between individual trees.

2.1.Data

2.1.1. Dataset Analysis

2.1.1.1. Dataset Description

The dataset used in this study is from the *Give Me Some Credit* competition hosted on Kaggle in 2011. This competition challenges participants to improve existing techniques in credit scoring by forecasting the probability that an individual will incur financial distress in the next two years. The dataset has already been used in other studies, such as *Research on personal credit evaluation based on machine learning algorithm* by Liu, Wang, and Han (2021) and *Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects* by Dumitrescu et al. (2022).

There are 11 columns and 150.000 rows in the dataset. The first column represents the target variable, indicating whether a borrower is expected to experience financial distress within the next two years. The explanatory variables, the remaining 10 columns, provide a range of various financial indicators that capture and reflect the borrower's credit behavior, debt history, and current financial obligations, offering a comprehensive view of their financial health. Understanding and analyzing these variables is essential for constructing a credit scoring predictive model.

2.1.1.2. Variables Description

Table 2 provides a description of each variable presented in the dataset.

Table 2 - Description of Each Variable

Variable Name	Description	Type
SeriousDIqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N (1 = Yes, 0 = No)
RevolvingUtilizationOf-UnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	Percentage
Age	Age of borrower in years	Integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years	Integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthy gross income	Percentage
MonthlyIncome	Monthly income	Real
NumberOfOpenCredit-LinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	Integer
NumberOfTimes90Days-Late	Number of times borrower has been 90 days or more past due	Integer
NumberRealEstateLoans-OrLines	Number of mortgage and real estate loans including home equity lines of credit	Integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years	Integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc)	Integer

Source: Extracted from Kaggle - *Give Me Some Credit* dataset

2.1.1.3. Variables Analysis

In this section, the distribution of the variables will be analyzed, supported by visualization of histograms, which can be found in Figure 4 in the Appendix.

The "SeriousDlqin2yrs" variable shows that the vast majority of data values are 0, suggesting that almost none of the observations exhibit serious delinquency within 2 years. Out of 150,000 rows, 139,974 have a value of 0, while only 10,026 are predicted to incur in financial distress, indicating a low delinquency rate.

The "RevolvingUtilizationOfUnsecuredLines" variable presents a distribution with a high concentration of low values, suggesting that most individuals use a small proportion of their unsecured credit lines. However, there are some higher values, indicating that a few people might be using a larger portion of their credit lines. The "Age" distribution appears roughly normal, peaking at approximately 50 years of age, revealing that more than half of the participants in the sample are middle-aged.

For "NumberOfTimes30-59DaysPastDueNotWorse," almost all values are 0, implying that only a few individuals have a history of 30-59 days of delay. A similar pattern is seen in "NumberOfTimes60-89DaysPastDueNotWorse" and "NumberOf-Times90DaysLate", where 0 is the most frequent, suggesting that late payments are rare among the participants.

The "DebtRatio" variable, which represents the ratio of monthly debt payments and living costs to monthly gross income, shows that most individuals have small debt ratios, with fewer participants having higher values, suggesting some variation in debt management. The "MonthlyIncome" variable has a prevalence of lower monthly income values (represented as $\chi * 10^6$), suggesting that most people in the sample report lower income levels. However, when converting these values, it becomes evident that some entries represent high income levels.

The "NumberOfOpenCreditLinesAndLoans" variable indicates that while most of the sample has a typical range of open credit lines and loans, there is variation, with some individuals having more active credit lines. The "NumberRealEstateLoansOrLines" variable shows that most people either have none or a small number of real estate loans or lines. Lastly, "NumberOfDependents" reveals that most individuals have few or no dependents, with variation in the sample.

2.1.2. Data Preprocessing

2.1.2.1. Filling Missing Values

In the dataset, two columns contained missing values: one referring to the individuals' monthly income and the other to the number of dependents. To address this issue, two techniques were applied.

For the monthly income, there were 29731 missing values. It was determined that the mean monthly income would be calculated and used to fill the blank space.

For the variable regarding the number of dependents, there were 3924 missing values. It was assumed that these blank entries were not filled because the individuals did not have dependents, so these spaces were replaced with a value of 0.

2.1.2.2. Outliers

After addressing the missing values, the dataset was analyzed for outlier identification, with the objective of identifying patterns and potential issues within the data. The visual method employed was box plot, Figure 5 in the Appendix.

The "SeriousDLin2yrs" distribution is highly imbalanced, with most observations being 0, indicating a low distress rate overall. There is a strong presence of high-value outliers in "RevolvingUtilizationOfUnsecuredLines", suggesting that some individuals are using a large portion of their credit lines, potentially exceeding their credit limits. To address this, values above 100% were limited at 100% (value = 1), restricting the extreme ones. The "Age" distribution, although generally balanced, shows some anomalies, such as entries of 0 years (impossible) and some values over 100 years old, which is uncommon. These outliers were addressed by setting values below 18 years to 18, and values above 100 to 90.

The variables "NumberOfTimes30-59DaysPastDueNotWorse", "NumberOf-Times60-89DaysPastDueNotWorse" and "NumberOfTimes90DaysLate" all show similar patterns. Most values are 0, indicating that late payments are uncommon. However, higher values suggest that a few individuals have a history of delayed payments, with a peculiar absence of values between 14 and 95. This gap raises the possibility of data entry errors, particularly for the values 96 and 98, which appear

unexpectedly and may be incorrect. To handle these, the values 96 and 98 were replaced with NaN, and then imputed with the median value for each variable.

The "DebtRatio" variable has significant outliers, indicating that some individuals are extremely over-indebted. These outliers were limited at 200% (value = 2). For "MonthlyIncome," high-income outliers stand out, suggesting that the recorded values may be unusually high, pointing to possible data entry errors or a unique subset of participants. To handle these, values exceeding the 99th percentile were capped at that percentile. Similiarly, the outliers present in the variable "NumberOfOpenCredit-LinesAndLoans" indicate that some individuals have an unusually high number of active credit lines, deviating from the norm. Also the outliers in "NumberRealEstate-LoansOrLines" suggest that a few individuals have multiple real estate lines, therefore, the same approach applied to "MonthlyIncome" was applied to these two variables.

Finally, outliers in "NumberOfDependents" shows some participants with a high number of dependents, which may not be typical, but no changes were made to this variable.

2.1.3. Heatmap

In the figures below (Figures 1 and 2), the Point-biserial and Pearson correlation heatmaps are presented. The first figure illustrates the relationship between the binary variable and the continuous ones, while the second figure shows the correlations among the continuous variables in the dataset. Both visualizations enable analysis and highlight key insights for understanding the data and constructing predictive models.

The color grading in the heatmap reflects the strength of the correlations, with dark red indicating a strong positive correlation while blue represents a strong negative correlation. Values near zero are shown in neutral colors, suggesting little or no linear relation between the analyzed variables. Correlation values range from -1 to 1, where 1 indicates a perfect positive correlation, -1 a perfect negative correlation and 0 means no linear correlation.

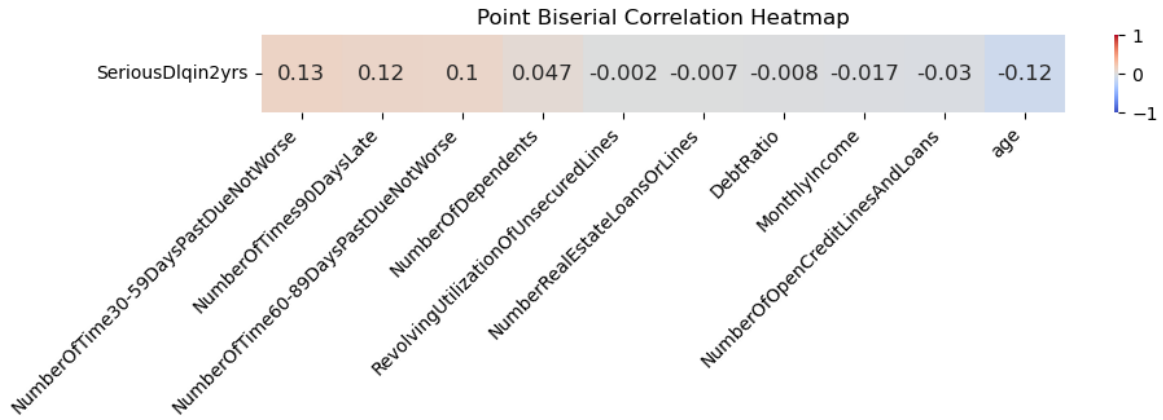

Figure 1 - Point Biserial Correlation Heatmap

In the Point-Biserial correlation heatmap (Figure 1), the highest positive correlations with the target variable are observed in the late payment variables: NumberOfTime30-59DaysPastDueNotWorse (0.13), NumberOfTimes90DaysLate (0.12) and NumberOf-Time60-89DaysPastDueNotWorse (0.10). This suggests that clients with late payment histories are more likely to become delinquent. Additionally, age shows a moderate negative correlation (-0.12), indicating that younger individuals may be more prone to financial distress.

Figure 2 - Pearson Correlation Heatmap

The Pearson correlation heatmap reveals several key patterns among the continuous variables. First, there is a very strong positive correlation between the three late payment variables - NumberOfTime30-59DaysPastDueNotWorse, NumberOfTimes90DaysLate and NumberOfTime60-89DaysPastDueNotWorse - with correlation coefficients above 0.98. This confirms that payment delays are highly interrelated and clients who delay their payments once are very likely to do it again.

A moderate positive correlation (0.43) is found between NumberOfOpenCreditLinesAndLoans and NumberRealEstateLoansOrLines, suggesting that individuals with more active credit lines also tend to hold more real estate credit products.

The remaining variables show weak or no linear correlations with each other. Variables such as RevolvingUtilizationOfUnsecuredLines, DebtRatio and MonthlyIncome have correlations values near zero with the other variables in the dataset. However, Age shows a slight negative correlation with several variables, the strongest being with NumberOfDependents (-0.22), which may suggest that older individuals are less likely to have financial dependents and tend to be more financially stable.

2.1.4. New Variables

After analyzing the heatmap, new variables were created to enrich the dataset and be able to do a more comprehensive analysis.

Some of the variables created were the Debt to Income Ratio, which measures the proportion of debt relative to income, providing insights into financial burden; the Total Late Payments, consolidating the three categories of delays (30-59 days, 60-89 days and 90+ days) into a single metric. Additionally, a binary variable was created to signal the individuals with two or more late payments exceeding 90 days. The Log of Monthly Income was also calculated since it is normally used to normalize the distribution of the income variable which is typically skewed. To capture combined effects, the interaction of the debt ratio and high credit utilization was developed referred to as the Debt Utilization Interaction.

Some dummy variables were also generated: Debt Ratios, were categorized into four groups - Low, Moderate, High and Very High; the Number of Dependents, divided

in - Up to 2 and Above 2; lastly, Monthly Income that was divided into quartiles, creating categories for different income levels.

These new variables were created, incorporating both mathematical transformations and categorical groupings, with the aim of capturing more complex relationships and improving the quality of analyses for the methods that were going to be used next.

2.1.5. Model Application

After processing the dataset with the steps outlined so far, it was time to apply the selected methods.

The first model implemented was the GAMLA and as it is shown in the Table I, the first step involved defining the binary (y), interest (x) and control (z) variables. For the binary variable, "SeriousDLin2yrs" was chosen, as it represents whether an individual is predicted to default (1) or nor (0).

Three personal characteristics were chosen as control variables – Age_squared, Age, NumberOfDependents. Age and age squared were included to capture non-linear effects, as age can have a complex relationship with default risk. Younger individuals may face greater financial challenges compared to older ones. The number of dependents was included to reflect the financial burden of supporting additional household members, directly impacting financial obligations and default risk. All remaining variables were used as the interest ones.

Next, the dataset was split into training and test sets, with 75% used as training data and the remaining 25% reserved for testing across all models.

Following the implementation of the GAMLA model, the other methods were applied to compare and validate the robustness of the findings. In these models, the data is only divided into χ and γ, with the binary variable remaining the dependent one, and the previously defined control variables included in χ.

3. RESULTS

This section presents the results of the analysis conducted in this study, starting with the causal inference analysis of the GAMLA model, which includes two results specific

to this model. Following that, the general performance metrics across all models will be discussed.

3.1. Casual Inference: GAMLA Model Results

3.1.1. Variables Selection

By applying the GAMLA model, the most relevant variables were selected through Adaptive LASSO were identified. These selected variables are as follows: NumberOfTime30-59DaysPast-DueNotWorse, NumberOfTime60-89DaysPastDueNot-Worse, NumberOfTimes90-DaysLate, Multiple_severe_delays, High_utilization, IncomeQuartile_Up-To25%, IncomeQuartile_UpTo50%, Income-Quartile_UpTo75%, IncomeQuarti-le_Above75%, DebtRatio_High, DebtRatio_Low, DebtRatio_Moderate, Debt-Ratio_VeryHigh, Real_estate_loans_ratio, NumberRealEstateLoansOrLines, NumberOfOpenCreditLinesAndLoans, Income_dependents, Log_monthly_income.

3.1.2. Statistical Measures of the GAMLA Model Variables

In Table 3, the control variables and those selected by the Adaptive Lasso are displayed. This table will be used to identify and conclude the three most relevant and two irrelevant variables. The selection of the most relevant variables is based on the statistical inference performed in step 8 of Table 1 - GAMLA Step, where the significance of the selected variables is tested using standard t-tests, providing a robust interpretation of their impact on the probability of default.

- Estimate Coefficient - A positive value increases the probability of default and a negative value decreases it;
- Standard Error - A small value indicates that the coefficient estimation is reliable, while a large value suggests uncertainty;
- T-statistics – Tests if the coefficient is statistically different from zero. Higher absolute values indicate greater significance. Large values (positive/negative) show that the variable has a real effect, while small values suggest it may not be relevant;
- P-value - If $p < 0.05$, the variable significantly impacts default; if $p > 0.05$, there is no significant effect and it may be irrelevant for the model; and,

- Confidence interval - Shows where the coefficient lies with 95% confidence. If the interval does not include zero, the variable is significative, otherwise it does not have real effect.

Table 3 - Statistical Measures of Model Coefficients and Variables

Variables	Estim.	Std.Err.	t	P > |t|	[0.025]	[0.975]
Intercept	0.007	0.008	0.850	0.395	-0.009	0.024
NumberOfTime30-59DaysPastDueNotWorse	0.044	0.001	41.262	<0.001	0.042	0.046
NumberOfTime60-89DaysPastDueNotWorse	-0.008	0.002	-4.137	<0.001	-0.011	-0.004
NumberOfTimes90Days-Late	-0.036	0.002	-19.823	<0.001	-0.040	-0.032
multiple_severe_delays	0.449	0.007	74.646	<0.001	0.486	0.512
high_utilization	0.100	0.002	59.842	<0.001	0.096	0.103
IncomeQuartile_Up to 25%	0.019	0.002	4.157	<0.001	0.015	0.023
IncomeQuartile_Up to 50%	0.010	0.002	4.157	<0.001	0.005	0.015
IncomeQuartile_Up to 75%	-0.013	0.003	-4.017	<0.001	-0.019	-0.06
IncomeQuartile_Above 75%	-0.009	0.003	-3.271	0.001	-0.015	-0.04
DebtRatio_High	0.021	0.005	4.242	<0.001	0.012	0.031
DebtRatio_Low	0.012	0.005	2.617	0.009	0.003	0.021
DebtRatio_Moderate	0.005	0.005	0.943	0.346	-0.005	0.014
DebtRatio_Very High	0.033	0.005	6.777	<0.001	0.023	0.042
real_estate_loans_ratio	-0.069	0.013	-5.388	<0.001	-0.094	-0.044
NumberRealEstateLoans-OrLines	0.009	0.001	6.876	<0.001	0.007	0.012

Variables	Estim.	Std.Err.	t	P > |t|	[0.025]	[0.975]
NumberOfOpenCreditLines-AndLoans	0.000	0.000	0.991	0.332	0.000	0.001
income_dependents	0.000	0.000	-0.281	0.779	0.000	0.000
log_monthly_income	0.006	0.001	7.961	<0.001	0.005	0.008
age_squared	0.000	0.000	2.806	0.005	0.000	0.000
age	-0.002	0.000	-5.300	<0.001	-0.002	-0.001
NumberOfDependents	0.004	0.001	5.492	<0.001	0.002	0.005

Starting with the three most relevant variables: “multiple_severe_delays”, the “high_utilization” of credit and “NumberOfTime30-59DaysPastDueNotWorse”. It is seen that the impact of multiple delays is very significative in the risk of default, as it has a coefficient of 0.449, which means the more delays the higher the chance of default. Furthermore, the t-statistics is large, 74.646, indicating that the effect is significant. As the confidence interval [0.486, 0.512] does not include zero, it is possible to affirm that this variable is relevant in the model.

Regarding the utilization of credit, the p-value shows a statistically significant relationship with default. The coefficient of 0.100 suggests that the more credit is used, the higher the chance of default. The confidence interval [0.096, 0.103] confirms that this variable has a positive effect on the model.

The third relevant variable, the number of times paid between 30 to 59 after the supposed, has a coefficient of 0.044, indicating that the more often a customer has these past due, the higher the changes of defaulting. The t-statistic value of 41.262 reinforces the significance of this variable. The p-value confirms the relevance of the variable and the confidence interval [0.042, 0.046] once again does not include zero, supporting its importance in the model.

For the irrelevant variables, “NumberOfopenCreditLinesAndLoans” and “income_dependents”, both have a coefficient of 0.000, and the standard error is also 0.000. Additionally, both p-values are greater than 0.05, which suggests that the relation

between these variable and the risk of default is not strong enough to be considered significant. Finally, the confidence intervals for both include 0.000, reenforcing the idea that these variables have no effect on default.

Based on these results, it can be concluded that three relevant factors when predicting default are payment delays, the frequency of these delays and credit utilization. These factors have the greatest impact on the likelihood of default, as indicated by their high coefficients, elevated t-statistics, and statistically significant p-values in this method and study.

3.2. Forecasting and Model Evaluation

In this section, the most common metrics used to evaluate the performance of a classifier will be presented and described to do a better evaluation and comparison between the models used.

Recalling that in this work, the binary classifier is represented as follows: 0 for non-defaulter and 1 for defaulter. The metrics applied in the study for all models tested are:

- Confusion Matrix and the metrics derived – Accuracy, Precision, Recall, F-score;
- Receiver Operating Characteristic (ROC) and the Area Under the Receiver Operating Characteristic Curve (AUC);
- Brier Score, Gini Index and Kolmogorov-Smirnov test (KS statistic); and,

This section will begin with an explanation of the measure, followed by the presentation of the results obtained using that measure.

3.2.1. Confusion Matrix

In binary classification problems, the confusion matrix is commonly used as it represents the algorithm's predictions versus the actual values, a concept frequently discussed in different studies (e.g., Albanesi & Vamossy, 2019; Luque, Carrasco, Martín, & de Las Heras, 2019).

Confusion matrix is a 2 x 2 matrix with two classes, positive and negative and it has four categories:

- True Positives – Instances that are truly positive and are correctly identified as positive by the model;
- False Positives – Instances that are negative but are incorrectly classified as positive by the model;
- True Negatives – Instances that are truly negative and are correctly identified as negative by the model; and,
- False Negatives – Instances that are positive but are incorrectly classified as negative by the model.

Regarding these values, various metrics are defined to enhance the model evaluation. Some of these metrics are:

I. Accuracy – Measures the proportion of correct predictions. In cases of imbalanced datasets, accuracy may not be reliable

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population}$$

II. Precision – Calculates the proportion of true positives out of all predicted positive values

$$Precision = \frac{\sum True\ Positive}{\sum Test\ outcome\ Positive}$$

III. Recall – Measures the proportion of actual positives cases correctly identified by the model

$$Recall = \frac{\sum True\ Positive}{\sum Actual\ Positive}$$

IV. F-score – A metric that combines precision and recall, calculated as their harmonic mean.

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

In Figure 2 presented below, the confusion matrix for all the models employed is shown, and the values of each model will be analyzed next.

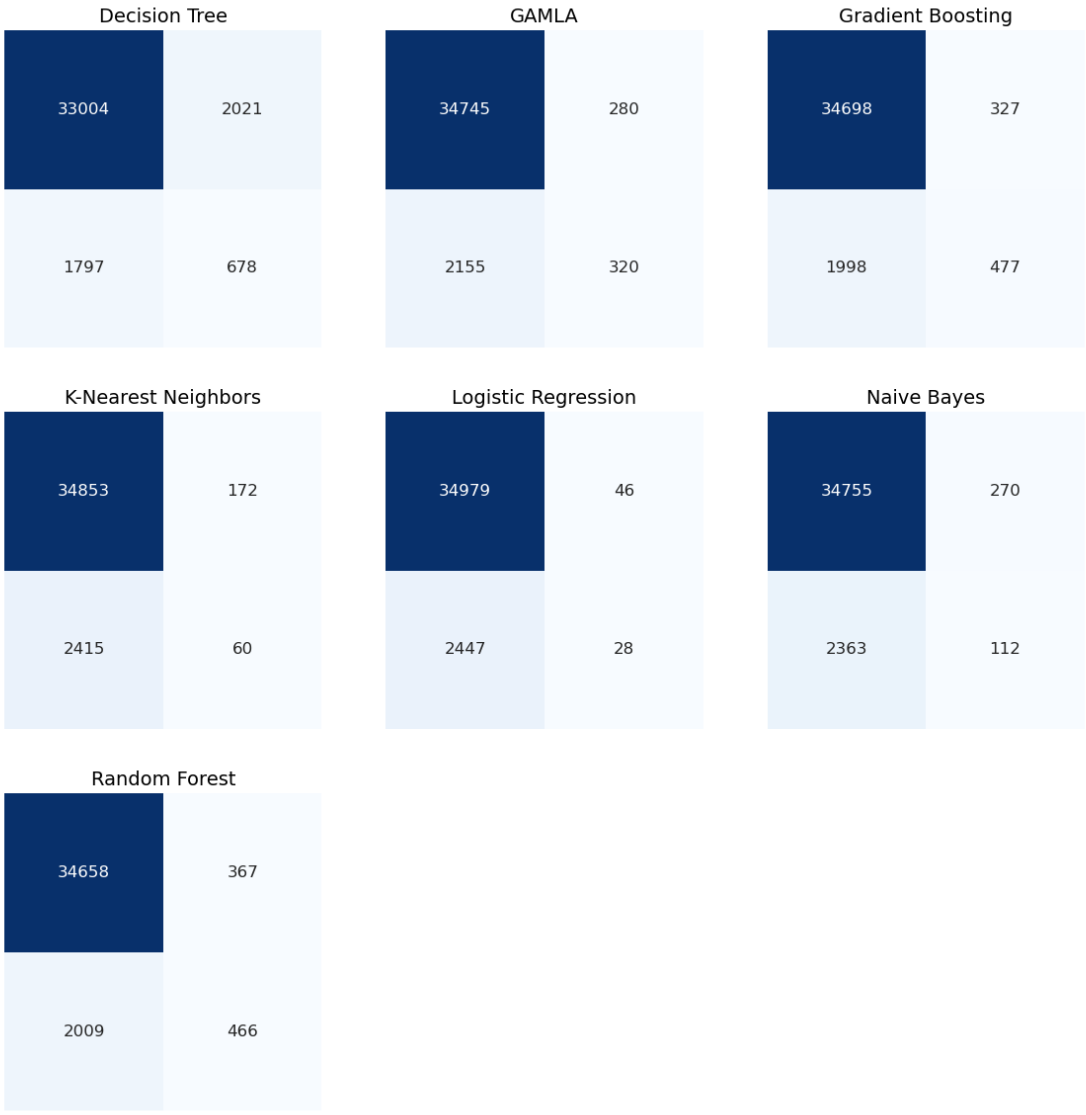

Figure 3 - Confusion Matrices of the models

The Decision Tree model shows a good quantity of true positives, but it also forecasts a large number of false positives. True negatives are reasonable (678), indicating that the model is able to capture significant part of the negative cases. However, the high false positive rate suggests that the model have difficulties distinguishing between negative and positive classes. About the predictions made by the GAMLA model, it is showed a high number of true positive and a low number of false positives. However, the true negatives (320) are small, suggesting that the model struggles to correctly forecast negative cases.

Following to other model tested, the Gradient Boosting had a good performance in detecting both positive and negatives, with 477 true negatives. It also keeps false positives and false negatives low, suggesting that the model is consistent across both classes. This model seems to offer a balanced approach in its prediction. The KNN although it has a high number of true positives, the model only produced 60 true negatives, which indicates that it fails to classify positive cases correctly.

The Logistic Regression also performed well in predicting true positives, but it has a low number of true negatives (28), this indicates that the model struggles to distinguish between negatives and positive classes. The Naïve Bayes model presents a balanced performance with 112 true negatives, which is better than the results obtained from Logistic Regression. It identifies both positive and negative cases, though it still faces challenges.

Lastly, Random forest is a strong model for capturing both positives and negatives, with 466 true negatives. This suggests that the model can accurately distinguish between both classes, being reliable when doing predictions. If only the results obtained from the confusion matrices are considered, the model with the best performance is the Gradient Boosting, showing the best balance between predicting positives and negatives.

Now, the metrics that can be analyzed using the values that were just reviewed will be presented in Table 4.

Table 4 - Accuracy, Precision, Recall and F-score results

Model	Accuracy	Precision	Recall	F-score
Decision Tree	0.898	0.251	0.274	0.262
GAMLA	0.935	0.533	0.129	0.208
Gradient Boosting	0.938	0.593	0.193	0.291
K-Nearest Neighbors	0.931	0.259	0.024	0.044
Logistic Regression	0.934	0.378	0.011	0.022
Naïve Bayes	0.930	0.293	0.045	0.078
Random Forest	0.937	0.559	0.188	0.282

The Decision Tree has a reasonable accuracy (89.8%), but both the precision and the recall are low, which indicates that, although the model is capable of correctly identifying a large portion of predictions, it fails to differentiate between positive and negative cases, resulting in a low F-score (26.2%). The GAMLA presents a good accuracy (93.5%) but has a low recall (12.9%). Even with the good precision, the low recall means the model fails to identify a significant portion of the positive cases, as reflected in the low F-score (20.8%).

The Gradient Boosting model has an excellent accuracy (93.8%) and a good precision, but the recall remains low (19.3%), indicating that the model misses a significant portion of true positives. The F-score reflects this imbalance, being reasonable but not great (29.1%). KNN also has a good accuracy (93.1%) but a very low recall (2.4%), resulting in an extremely low F-score (4.4%).

The Logistic Regression model presents a good accuracy (93.4%) with a precision and recall of 37.8% and 1.1%, respectively. Once again, the model fails to detect positive cases and makes imprecise predictions when it does, leading to a similarly low F-score is (2.2%). The next model tested, the Naïve Bayes, the accuracy is still good, like the other models, but the recall remains small (4.5%), resulting in a low F-score (7.8%).

Finally, Random Forest performed well in terms of accuracy and precision, but its recall (18.8%) and F-score (28.2%) are still relatively low, though higher than the other models analyzed.

Once again, if only the results obtained from these previous metrics are considered, the model with the best performance is the Gradient Boosting, due to its strong accuracy and reasonable recall and F-score results.

3.2.2. Receiver Operating Characteristic Curve

The ROC is a graphical tool used in binary classification to evaluate a model's performance in distinguishing between two classes. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. A perfect classifier is in (0,1) and the points closer to the top left corner indicate better performance (Kennedy, 2013).

The AUC curve is a widely used metric for evaluating the performance of the ROC curve by calculating the area under it. AUC values range from 0.5 (random performance) to 1 (perfect). AUC reflects the probability that a classifier will correctly classify a randomly chosen positive instance over a negative one, and is especially useful for unbalanced datasets (Rodriguez & Rodriguez, 2006; Bradley, 1997)

In the figure below, figure 3, the AUC of all the models tested is represented.

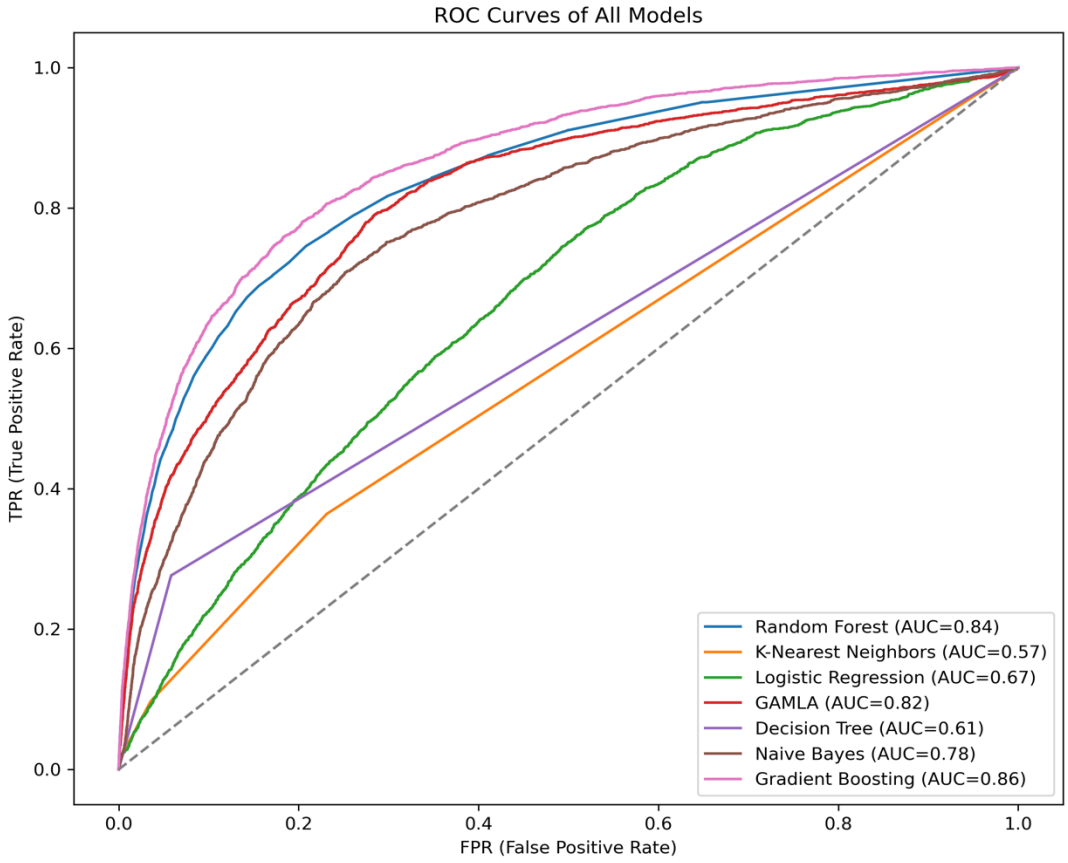

Figure 4 - ROC Curve of All Models

As shown, the best performing model, with the highest AUC (0.86) is Gradient Boosting, indicating its superior ability to separate positive and negative classes. Following this model, the next highest performers, with competitive scores, are Random Forest and GAMLA, with AUCs of 0.84 and 0.82, respectively.

About the other models tested, Naïve Bayes ranks fourth in AUC, with a score of 0.78, followed by Logistic Regression, where the AUC is 0.67.

The models with the lowest AUC performances are Decision Tree and KNN, with results of 0.61 and 0.57, respectively.

3.2.3. Receiver Brier Score, Gini Index and KS Statistic

The Brier score evaluates the classifier calibration, focusing on the accuracy of its predicted probabilities. Introduced by Brier in 1950, it is computed as the mean squared error between the predicted probabilities and the actual outcomes. Lower values indicate better calibration, quantifying how close predicted probabilities are to actual results.

The area under the ROC Curve is used to calculate the Gini index, which measures a classifier's discriminatory power. Higher values indicate better class discrimination and it is computed as twice the area between the ROC curve and the random classification line. Dumitrescu et al. (2022) highlighted the importance of focusing on thresholds in the lower tail of the distribution for credit applications, where the Partial Gini is useful. However, the Partial Gini was not used in this work, as all metrics were evaluated with a fixed threshold of 0.5 for consistency.

KS Statistic evaluates a classifier's ability to separate two classes, measuring the maximum distance between the cumulative distributions of the predicted scores for the positive and negative classes (Thomas et al., 2002). A higher KS value indicates a bigger separation between classes, which is essential in credit scoring to distinguish between default and non-default loans. The values for these three metrics are presented in the following Table 5.

Table 5 - Brier, Gini and KS Score

Model	Brier	Gini	KS
Decision Tree	0.262	0.218	0.218
GAMLA	0.261	0.631	0.506
Gradient Boosting	0.259	0.728	0.576
K-Nearest Neighbors	0.262	0.143	0.133
Logistic Regression	0.265	0.337	0.252
Naive Bayes	0.252	0.564	0.457
Random Forest	0.260	0.682	0.538

The Decision Tree has a Brier score of 0.262, indicating an acceptable performance in terms of predicted probabilities. However, its Gini and KS are relatively low (0.218 both), reflecting the model's limited ability to distinguish between positive and negative classes, impacting its efficacy. The GAMLA achieves a Brier score approximately the same obtained as the Decision Tree, but it presents higher Gini and KS values (0.631 and 0.506, respectively), these values indicates that this model has better performance when predicting class separation.

With a Brier score of 0.259, the Gradient Boosting stands out because of having superior performance both in Gini (0.728) and KS (0.576), making it an effective model at distinguishing positive and negative classes. The KNN returns to low Gini (0.143) and KS (0.133) values, making the model less efficient compared to the other models.

Logistic Regression exhibits a slightly higher Brier score (0.265) than the other models, but Gini and KS are not among the best (0.337 and 0.252, respectively). Naïve Bayes achieves a similar Brier score to the other models and performs well in terms of Gini and KS values (0.564 and 0.457, respectively), making it a strong model. As seen previously, Random Forest performs well with a competitive Brier score and has very good results in Gini, 0.682, and KS, 0.538, confirming its reliability.

Finally, Random Forest performs well with a competitive Brier score and has very good results in Gini, 0.682, and KS, 0.538, confirming its reliability. As with the earlier

metrics, Gradient Boosting emerges as the model with the best performance, particularly in this forecast.

3.3. Optimal Model Performance

After doing this comprehensive analysis of all the indicators between various models, it becomes evident that no single metric is sufficient to assess overall performance. Instead, a comprehensive evaluation of all metrics is essential for identifying the model with the best performance. Among the models tested, Gradient boosting consistently demonstrated strong results.

Other models, such as Random Forest and GAMLA also showed competitive performance, especially in metrics like Gini index and KS. Naïve bayes also delivered reasonable results, ranking fourth in the AUC and achieving a good Gini Index, positioning it above some of the weaker models.

In comparison, models such as Decision Tree and KNN exhibited lower values in multiple metrics, including recall and F-score, reflecting challenges in handling the dataset effectively. Similarly, Logistic Regression displayed lower predictive power, with the lowest recall score.

To conclude, Gradient Boosting emerges as the most robust and versatile model for the dataset used in this study.

CONCLUSION

This study encountered several challenges, particularly in the practical part. The dataset's imbalance initially led to a low number of default predictions when applying the methods. Various techniques, such as oversampling and undersampling, were tested to address this issue. The methods were implemented with minimal adjustments to simulate real-world scenarios, including the presence of outliers, such as individuals with high salaries or a large number of dependents. Given the initial bias towards the majority class (non-defaults), additional modifications were made, such as outlier treatment and the creation of new variables. These adjustments resulted in more robust and representative results.

As demonstrated in this study, Gradient Boosting was the best performing method for the dataset used in this project. This aligns with the findings from the literature review, where Gradient Boosting is highlighted as a method that excels in credit scoring assessment due to the high accuracy and robust ability to handle imbalanced data.

The results also revealed that payment delays, the frequency of these delays, and credit utilization are the most relevant factors when predicting credit default. These factors had the most significant impact on the likelihood of default.

Through this study it is evident the significant role of machine learning and its contribution in real world, in this case, in credit scoring. It is also clear that machine learning will continue to be a subject of extensive research to improve its accuracy and become even more useful in daily life. As mentioned by Bello (2023), future advancements in credit risk assessment can benefit from the integration of other data sources, such as social media activity, as these sources can provide additional insights that traditional data may overlook. Techniques like Natural Language Processing are becoming very useful for analyzing textual data from social media and extract insights into the borrower's behavior and sentiment (Hohnen et al., 2021).

Developing techniques that are able to provide clear explanations of models' decisions is essential and remains a critical area for future research. It is also important to improve the interactions between humans and Artificial Intelligence systems, making easier for users to query and better understand the decisions made by the models . Bello (2023) also highlighted that an important area to explore involves the ethical problems in machine learning models. As mentioned before, managing sensitive information remains a challenge so there is the need to develop techniques that enhance data privacy, such as avoiding data sharing during model training.

REFERENCES

Albanesi, S., & Vamossy, D. F. (2019). *Predicting consumer default: A deep learning approach* (No. w26165). National Bureau of Economic Research.

Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1), 41-59.

Bello O.A. (2023) *Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis*, International Journal of Management Technology, Vol.10, No 1, pp.109-133

Bradley, A. P. (1997). *The use of the area under the ROC curve in the evaluation of machine learning algorithms*. Pattern recognition, 30(7), 1145-1159.

Carlin, J. B., & Doyle, L. W. (2001). *Statistics for clinicians: 4: Basic concepts of statistical reasoning: hypothesis tests and the t-test.* J Paediatr Child Health, 37(1), 72-77.

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). *Interpretable machine learning for imbalanced credit scoring datasets.* European Journal of Operational Research, 312(1), 357-372.

Credit Fusion, W. C. (2011). Give me some credit. Kaggle. https://kaggle.com/competitions/GiveMeSomeCredit

West, D. (2000). *Neural network credit scoring models.* Computers & operations research, 27(11-12), 1131-1152.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). *Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects*. European Journal of Operational Research, 297(3), 1178-1192.

Flachaire, E., Hacheme, G., Hué, S., & Laurent, S. (2022). *GAM (L) A: An econometric model for interpretable Machine Learning*. arXiv preprint arXiv:2203.11691.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, No. 1).

Janiesch, C., Zschech, P., & Heinrich, K. (2021*). Machine learning and deep learning*. Electronic Markets, 31(3), 685-695.

Kataria, A., & Singh, M. D. (2013). *A review of data classification using k-nearest neighbour algorithm*. International Journal of Emerging Technology and Advanced Engineering, 3(6), 354-360.

Kennedy, K. (2013). *Credit scoring using machine learning*. Doctoral thesis. Technological University Dublin. doi:10.21427/D7NC7J.

Liu, S., Wang, R., & Han, Y. (2021). *Research on personal credit evaluation based on machine learning algorithm*. In 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT) (pp. 48-52). IEEE.

Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). *The impact of class imbalance in classification performance metrics based on the binary confusion matrix.* Pattern Recognition, 91, 216-231.

Min, J. H., & Lee, Y. C. (2008). *A practical approach to credit scoring*. Expert Systems with Applications, 35(4), 1762-1770.

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Rodriguez, A., & Rodriguez, P. N. (2006). *Understanding and predicting sovereign debt rescheduling: A comparison of the areas under receiver operating characteristic curves.* Journal of Forecasting, 25(7), 459-479.

Thomas, Lyn & Edelman, David & Crook, Jonathan. (2002). *Credit Scoring and its Applications.* Society for industrial and Applied Mathematics.

Tounsi, Y., Hassouni, L., & Anoun, H. (2017). *Credit scoring in the age of big data–A state-of-the-art.* International Journal of Computer Science and Information Security (IJCSIS), 15(7), 134-145.

Valdrighi, G., Ribeiro, A. M., Pereira, J. S., Guardieiro, V., Hendricks, A., Garcia, J. D. N., ... & Raimundo, M. M. (2024). *Best Practices for Responsible Machine Learning in Credit Scoring*. arXiv preprint arXiv:2409.20536.

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). *Comparing two SVM models through different metrics based on the confusion matrix*. Computers & Operations Research, 152, 106131.

APPENDICES

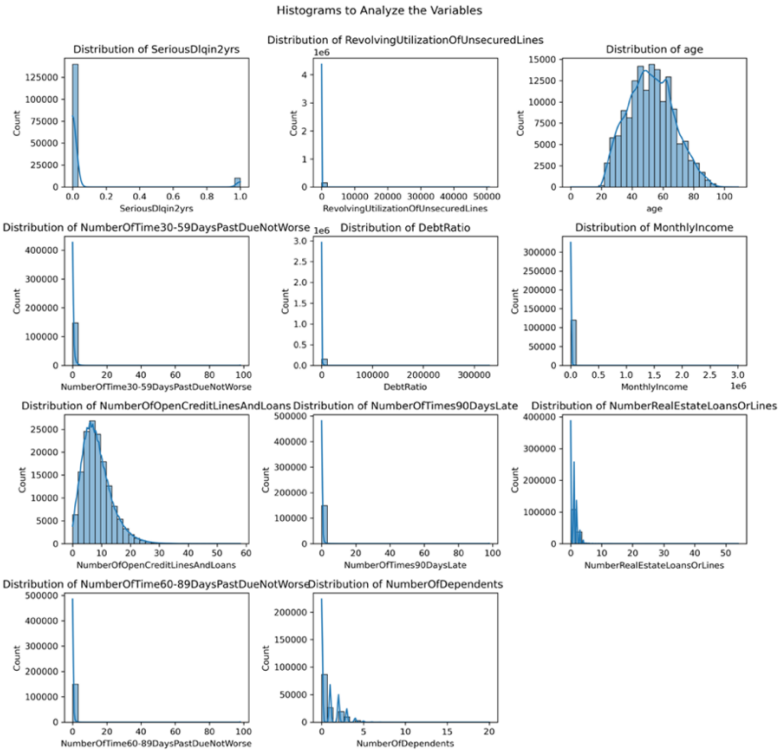

Figure 5 - Histograms to Analyze the Variables

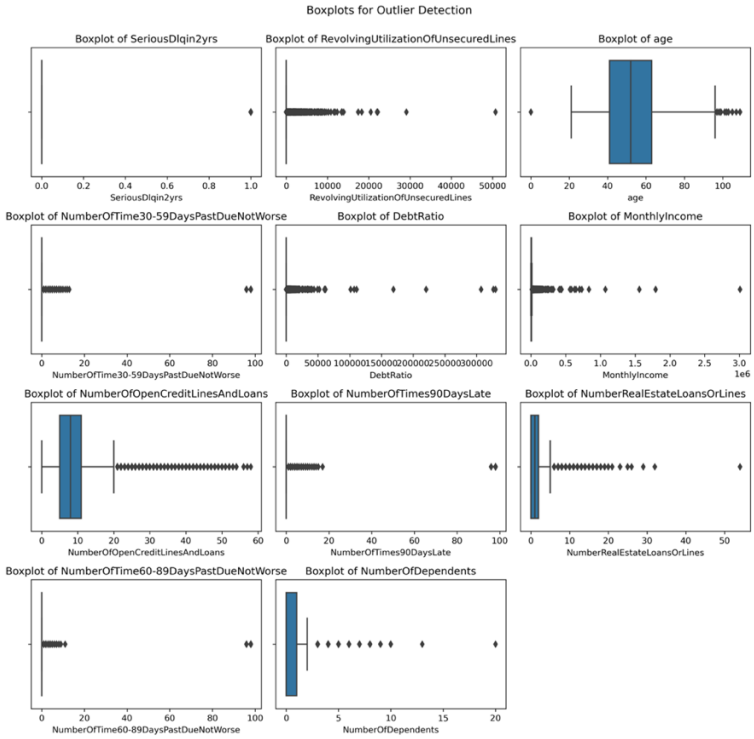

Figure 6 - Boxplots for Outliers Detection