NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Machine Learning Approaches for Predicting Debt Consolidation Loan Approval

A Study Using Logistic Regression and Random Forest Models

Carla Isabel Pires Lopes

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Machine Learning Approaches for Predicting Debt Consolidation Loan Approval

A Study Using Logistic Regression and Random Forest Models

by

Carla Isabel Pires Lopes

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Supervised by

Bruno Jardim, PhD, NOVA Information Management School

November, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Porto, 30/11/2024

Carla Lopes

ABSTRACT

This study analyzes the performance of Logistic Regression and Random Forest models in predicting loan approvals, focusing on debt consolidation loans, a type of financial product that aggregates multiple debts into a single loan, aiming to reduce default and facilitate consumers' financial management. Although advantageous, predicting approval for this type of loan presents complex challenges due to the diversity of applicant profiles and the need for models that combine accuracy and interpretability. To address these issues, data preparation techniques were applied, resulting in significant improvements in the performance of the models. Random Forest stood out in all applied metrics, demonstrating its robustness and effectiveness in scenarios with complex and imbalanced data. Logistic Regression, on the other hand, although presenting slightly lower performance, proved to be a valuable option due to its simplicity and interpretability, making it ideal for applications in contexts that require regulatory transparency. This work reinforces that the choice of model should be guided by the specific objectives of the application, whether prioritizing predictive accuracy, as in the case of Random Forest, or interpretability, as in the case of Logistic Regression. The results demonstrate how predictive modeling can optimize loan approval processes, reduce financial risks and increase operational efficiency in financial institutions.

KEYWORDS

Machine Learning; Loan Approval; Logistic Regression; Random Forest

Sustainable Development Goals (SDG): https://sdgs.un.org/goals/goal1

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
DT	Decision Tree
GDP	Gross Domestic Product
LR	Logistic Regression
ML	Machine Learning
RF	Random Forest

1. INTRODUCTION

1.1. MOTIVATION

Economic growth is influenced by a multitude of factors, covering macroeconomic and microeconomic elements and is heavily dependent on increased production capacity achieved through technological and business advancement, which companies typically finance through loans, bonds, and new equity.

However, for this growth to be sustainable, a corresponding increase in demand is essential, driven primarily by household consumption (Guttmann & Plihon, 2010).

Household consumption is a decisive component of a country's Gross Domestic Product (GDP) and an essential indicator of economic health and consumer behavior. It represents the market value of all goods and services, including durable products (such as cars, washing machines, and home computers), purchased by households (World Bank, n.d.).

As for the Gross Domestic Product (GDP) indicator, it is a widely used economic metric that quantifies the total monetary or market value of all finished goods and services produced in an economy, over a specific period. Despite some controversy (Brynjolfsson & Collis, 2019) (3), it is still commonly used as a meter of a nation's economic well-being and the progress of its population, being a fundamental tool to guide policymakers, investors and companies in making strategic decisions.

In 2023, Portugal household consumption accounted for 62.58% of its GDP (The Global Economy, 2024), and in the European Union it accounted for 52.53% (Eurostat, 2023), making the consumer spending as the primary driver of economic demand in these regions.

Marketing strategies and continuous advertising play a significant role in household consumption by generating and addressing consumer needs, and thus, driving economic activity. However, the ultimate decision to purchase goods and services is often related to the availability of financial accessibility and economic conditions.

By offering mortgages, consumer loans and credit cards, financial institutions enable individuals to make purchases of significant value including homes, cars, consumer goods, home renovations, education and travel expenses, or just to respond to unexpected financial problems, that otherwise would be difficult or even impossible to overcome (Lansing, 2011).

In addition to the influence of consumer society, the recent economic downturn in the West, marked by inflation outpacing wage growth and increasing mortgages rates, reducing, consequently, the purchasing power, may increase the need for individuals to borrow more loans or use credit cards more frequently to cover essential or unexpected expenses.

This can result in over-indebtedness, where an individual may take out multiple loans to provide short-term financial relief, at the cost of high debt levels and greater financial stress later.

And this is where a special type of personal loan comes in: Debt consolidation is a product offered by many financial institutions with the goal of providing a financial solution

for consumers dealing with increasing debt and credit challenges and who struggle to keep up with their monthly installments.

Debt Consolidation aggregates most of the customer's consumer credits, even from different financial institutions, into a single loan often with a lower interest rate and a longer payment term, resulting in a simpler and smaller monthly installment.

Besides alleviating the borrower's financial struggle and reducing the risk of default, debt consolidation can potentially improve the customer's credit score. By consistently meeting repayments, the borrower demonstrates financial responsibility, which can lead to better loan terms and opportunities in the future.

However, the process of granting loans by financial institutions, especially debt consolidation loans, is often a bureaucratic and time-consuming process.

Initially, the applicant must submit various documents, such as identification, proof of income and residence, tax status, financial statements, statements of existing debts to be consolidated, and others. These documents are then reviewed to understand the borrower's creditworthiness and a risk assessment is conducted to determine the probability/capacity of loan repayment.

These analysis may involve various models, such as credit and behavior score models and fraud detection. However, human oversight is still required to interpret the results, especially in ambiguous situations, and to consider factors that the models may not fully detect.

1.2. RESEARCH GAP

Despite significant advances in the use of machine learning and predictive modeling for financial applications, there are still several gaps in the field of loan approval prediction, especially in the case of debt consolidation loans.

Most existing studies focus on standard consumer loan approval processes, not considering the specific complexities associated with the consolidation of multiple debts from different financial institutions. These types of loans involve additional factors such as the borrower's existing debt structure, their repayment behavior across multiple lenders, and their ability to manage a single consolidated payment. These unique challenges in predicting consolidation loan approval have not yet been fully explored in the literature.

Another important gap is in the comparative analysis of interpretable models, such as Logistic Regression, against more complex but powerful models, such as Random Forest, in the context of financial decision-making. While many studies emphasize predictive accuracy, few address the trade-off between model performance and interpretability, a crucial aspect in regulated industries such as the financial sector.

Finally, most existing research relies on static datasets and does not consider dynamic real-world factors such as changing economic conditions. These temporal aspects are particularly relevant for debt consolidation loans, as they often reflect borrower responses to financial distress. Addressing this gap requires incorporating temporal data and longitudinal

analyses to improve the robustness and scaling of predictive models in real-world applications.

This paper seeks to fill these gaps by comparing the performance of Logistic Regression and Random Forest models in predicting loan approval and the unique challenges associated with debt consolidation loans. In this way, it aims to contribute to the advancement of research in predictive modeling in the financial sector and provide practical insights to improve decision-making processes in financial institutions.

1.3. OBJECTIVES AND METHODOLOGY

This thesis aims to leverage data from a confidential Portuguese financial institution to develop a machine learning propensity model for credit approval. The primary objective is to create a model capable of predicting the likelihood of in-house customers being approved for a debt consolidation loan, enabling more precise targeting and improved decision-making in the credit approval process.

The rest of this thesis is organized as follows: **Literature Review** presents the theoretical background and previous research relevant to machine learning applications in credit risk assessment and loan approval processes. **Methodology** describes the steps taken to develop the propensity model, including data collection, preprocessing, feature engineering, model selection, implementation, and evaluation. **Results and Discussion** present and analyze the findings from the model implementation, discussing the performance of Logistic Regression and Random Forest in predicting loan approval, the importance of key variables, and the implications for improving decision-making and operational efficiency. **Conclusions** summarize the key findings and contributions of the research, emphasizing the advantages of leveraging machine learning for credit approval processes. **Limitations and Recommendations for Future Work** address the research's limitations and propose potential areas for improving the model, exploring other machine learning techniques, and extending the methodology to other products or areas within the financial institution.

2. LITERATURE REVIEW

2.1. CONSUMER CREDIT AND MACHINE LEARNING IN ECONOMIC GROWTH

Access to credit plays a crucial role in economic growth and household consumption. As highlighted by (Lansing, 2011), credit allows households to smooth consumption over time, invest in essential areas such as education and housing, and respond to unexpected financial shocks, thus contributing to economic stability. However, financial institutions face significant challenges in the loan approval process, especially in predicting defaults and maintaining a delicate balance between risk and return.

The integration of Machine Learning techniques offers transformative potential to address these challenges. (The Financial Stability Board, 2017) notes that ML applications in finance have improved areas such as credit scoring, fraud detection, and portfolio management, thereby increasing the efficiency and stability of financial systems.

As demonstrated by (Saini et al., 2023), ML models can process large volumes of data and identify complex patterns that traditional methods may miss, significantly improving the accuracy of credit risk predictions. Similarly, (Aniceto et al.2020) emphasized that ML models, such as Random Forest and AdaBoost, outperform traditional methods in predicting consumer credit, evidencing their ability to improve decision-making in financial institutions.

2.2. MACHINE LEARNING IN LOAN PREDICTION

Machine Learning has significantly improved decision-making processes in the financial sector, especially in the context of credit evaluation and loan eligibility. ML models enable financial institutions to analyze applicant data more efficiently and accurately, improving eligibility predictions.

(Shinde et al., 2022) investigated the application of ML techniques, specifically Logistic Regression and Random Forest, to assess the eligibility of loan applicants. Using variable engineering, they developed new predictors from existing data, enabling the models to identify patterns more effectively and generate more accurate predictions. Key variables such as customer credit history, disposable income (after payment of installments), total income of the applicant and co-applicant (where applicable), and equated monthly installments (EMI) were identified as critical determinants of loan approval. The results indicated an accuracy of 82% for Logistic Regression and 79% for Random Forest, highlighting the potential of ML models to simplify the loan approval process. However, the study had limitations, such as the use of a relatively small dataset, with only 600 samples, and the lack of consideration of dynamic factors, such as changes in economic conditions or behavioral analyses over time.

Similarly, (Sheikh et al., 2020) focused exclusively on the use of logistic regression as a machine learning model to predict and optimize loan approval decisions and their safety based on the applicant characteristics. Using a dataset of 1500 applicants, the authors incorporated numerical and categorical variables, such as income, credit history, loan amount, and property location, to predict approval outcomes.

The study emphasized the importance of exploratory analysis, data preprocessing and feature engineering to achieve optimal results. The model achieved an accuracy of 81.1%, and the analysis showed that applicants with poor credit histories were more likely to be rejected, while those with higher incomes and smaller loan applications were more likely to be approved. Interestingly, demographic variables such as gender and marital status were found to be irrelevant to loan outcomes. While the study also used assessment metrics such as precision, recall, and F1-score, the specific results of these measures were not detailed.

(Saini et al., 2023) conducted a comparative analysis of several ML classification algorithms for predicting loan approval. The study evaluated models such as Random Forest, Logistic Regression, K-Nearest Neighbors, and Support Vector Classifier using a dataset that included information such as income, credit history, loan amount, and socioeconomic data. The results showed that the Random Forest model outperformed the others, achieving an accuracy of 98.04% and an ROC score of 0.97, highlighting the potential of ML techniques to improve loan approval processes.

Similarly, (Viswanatha et al., 2023) explored the application of ML algorithms such as Random Forest, Naive Bayes, Decision Tree, and K-Nearest Neighbors to predict loan approval status. The research demonstrated that the Naive Bayes algorithm achieved the highest accuracy at 83.73%, highlighting the effectiveness of ML models in processing complex data and improving predictive results in banking operations.

(Fekadu et al., 2022) analyzed various machine learning models, such as Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and XGBoost, to predict non-performing loans based on data from an Ethiopian bank. Their study identified applicant age, employment duration, and total income as the most critical factors in credit risk assessment, surpassing the significance of collateral-related attributes.

(Biecek et al. 2021) conducted a comparative analysis of predictive models, including logistic regression and advanced artificial intelligence (AI) algorithms, within the credit scoring domain. Their findings revealed that tree-based models consistently delivered superior predictive performance. The study also highlighted the critical role of interpretability in machine learning, presenting techniques that unravel the complexity of opaque models. These methods aim to make AI algorithms more understandable and usable for credit risk professionals. The research also emphasized the importance of striking a balance between model complexity and transparency needed to foster trust and understanding in predictive systems.

These studies exemplify the critical role of machine learning in modernizing loan approval systems by providing financial institutions with robust tools to assess creditworthiness with greater accuracy. Through the use of ML algorithms, credit institutions can improve their decision-making processes and reduce default rates.

3. METHODOLOGY

In this study, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was adopted as the foundational framework for developing a Machine Learning model. CRISP-DM was selected due to its structured and iterative approach, which facilitates alignment between technical analysis and business needs (Shearer, 2000).

Widely recognized in the industry, CRISP-DM consists of six phases—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—that provide a comprehensive roadmap for managing data-driven projects (Chapman et al., 2000).

The choice of CRISP-DM reflects its adaptability and proven effectiveness in handling complex data projects, especially within financial contexts. Its flexibility allows for iterative refinement of models and revisitation of previous phases as new insights emerge, ensuring a robust and reliable process. This systematic approach supports the development of models tailored to meet specific business challenges, while maintaining a focus on delivering actionable outcomes (Chapman et al., 2000).

This adaptability is particularly valuable in projects with complex data and evolving business requirements, such as the propensity model analysis in this study. To operationalize the CRISP-DM methodology, a project flow to this study was designed, as illustrated in Figure 1, to guide the analysis. This flow was structured to align with the six phases of CRISP-DM, but with adaptations tailored to the development of a propensity model for identifying the most viable customers for debt consolidation loan approvals.

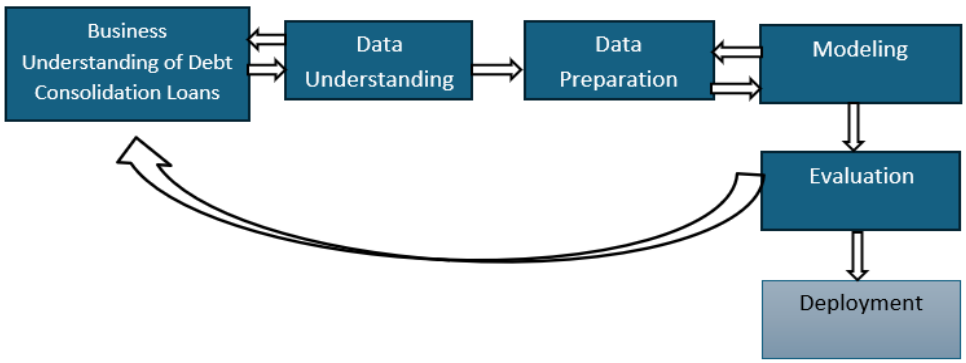


Figure 1 – Crisp-DM methodology applied to the study.

Each upcoming subsection in this chapter will detail the steps, tools, and techniques employed within the Project Flow, highlighting its alignment with the CRISP-DM framework. This structured and customized approach ensures a clear pathway from defining the business challenge to generating meaningful insights, enabling improved decision-making and operational efficiency.

3.1. BUSINESS UNDERSTANDING

The Business Understanding of this thesis focuses on the growing need for financial institutions to improve loan approval processes, ensuring faster, more accurate decisions that are aligned with the risk profile of applicants. With the increase in data complexity and the evolution of economic conditions, traditional methods have proven insufficient to deal with these challenges.

The use of machine learning emerges as a promising solution, allowing the analysis of large volumes of data and the identification of complex patterns that help predict defaults and improve credit allocation. The ultimate goal is to develop robust models that not only optimize approval rates but also reduce the risks associated with defaults, promoting greater operational efficiency and financial inclusion.

3.2. DATA UNDERSTANDING

In this section, we introduce the key elements and details regarding the dataset used in this study. This contextualization is essential for the reader to understand the relevance and characteristics of the dataset and the alignment with the study's primary objectives.

The data selection criteria not only considered these objectives but also prioritized the quality and volume required to effectively implement the proposed machine learning models. Nonetheless, it is also important to note that some potentially interesting variables were either unavailable or inaccessible and, therefore, could not be included in this research.

The dataset is confidential since it was obtained through a segmentation of debt consolidation applications done by in-house customers of the organization. Prior to the analysis it was also anonymized to ensure compliance with data protection regulations and privacy policies.

The period analyzed spans from the third quarter of 2022 to the second quarter of 2024. This interval allows us to obtain a greater volume of data and eventually consider changes in economic conditions that have occurred during this time.

The dataset, as summarized in Figure 2, is a sample of 8,820 records and 26 numerical features, which are best described in Table 1. It captures information related to debt consolidation applications and applicant profiles, such as loan characteristics (numbers 2 to 7), the applicant's financial status and creditworthiness (numbers 8 to 14 and 22 to 25), demographic information (numbers 15 to 21), and the economic context at the time of the application submission (number 26).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8820 entries, 0 to 8819
Data columns (total 26 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   ID                           8820 non-null   float64
 1   loan_status                  8820 non-null   float64
 2   application_date             8820 non-null   float64
 3   Loan_amount                  8820 non-null   float64
 4   total_amount_charge_client   8820 non-null   float64
 5   installment_amount           8820 non-null   float64
 6   loan_term                    8820 non-null   float64
 7   net_salary_1                 8820 non-null   float64
 8   net_salary_2                 8820 non-null   float64
 9   Total_other_income           8820 non-null   float64
 10  net_monthly_income           8820 non-null   float64
 11  total_amount_consumer_loans  8820 non-null   float64
 12  total_amnt_credit_cards      8820 non-null   float64
 13  Total_extra_installments     8820 non-null   float64
 14  marital_status_single        8820 non-null   float64
 15  marital_status_married       8820 non-null   float64
 16  number_dependents            8820 non-null   float64
 17  education_lower              8820 non-null   float64
 18  education_higher             8820 non-null   float64
 19  age                          8820 non-null   float64
 20  sex                          8820 non-null   float64
 21  housing_status_WithFamily    8820 non-null   float64
 22  housing_status_Rent          8820 non-null   float64
 23  housing_status_Own           8820 non-null   float64
 24  overdue_credit               8820 non-null   float64
 25  inflation_rate               8820 non-null   float64
dtypes: float64(26)
```

Figure 2 - Dataset overview.

Table 1- Description of the features presented in the dataset.

#	Feature	Description
1	**ID**	Unique identifier assigned to each loan application
2	**loan_status**	Binary variable indicating the application status: 0 = Loan rejected / 1 = Loan approved
3	**application_date**	The date the application was submitted (format: YYYYMM)
4	**loan_amount**	Total loan amount requested by the borrower
5	**total_amount_ charge_client**	Total cost of the loan to the borrower, includes the total amount requested plus the total interest
6	**installment_amount**	Monthly installment amount
7	**loan_term**	Loan term chosen by the applicant in months
8	**net_salary_1**	The net salary of the primary proponent
9	**net_salary_2**	The net salary of the secondary proponent, if existing
10	**Total_other_income**	The total amount of other monthly incomes, if available

11	**net_monthly_income**	The total net amount of the applicant's income, combining all income sources available
12	**total_amount_consumer_loans**	The total outstanding balance on consumer loans
13	**total_amnt_credit_cards**	The total outstanding balance on credit cards
14	**Total_extra_installments**	The total amount of the remaining monthly installments that will be left out of the credit consolidation
15	**marital_status_single**	Binary feature indicating the marital status: 0 = the applicant is not single / 1 = the applicant is single/divorced/widowed
16	**marital_status_married**	Binary feature indicating the marital status: 0 = the applicant is not married / 1 = the applicant is married
17	**number_dependents**	The number of dependents the applicant financially supports
18	**education_lower**	Binary feature indicating the educational background of the applicant: 0 = the applicant does not have a lower education level / 1 = the applicant has a lower education level
19	**education_higher**	Binary feature indicating the educational background of the applicant: 0 = the applicant does not have a higher education / 1 = the applicant has a higher education
20	**age**	The applicant's age
21	**sex**	Binary feature indicating the applicant's gender: 0 = Female / 1 = Male
22	**housing_status_WithFamily**	Binary feature indicating the housing status of the borrower: 0 = the borrower does not live in his relatives' house / 1 = the borrower lives in his relatives' house
23	**housing_status_Rent**	Binary feature indicating the housing status of the borrower:

		0 = the borrower is not renting a house / 1 = the borrower lives in a rented property
24	**housing_status_Own**	Binary feature indicating the housing status of the borrower: 0 = the borrower does not own his residence/ 1 = the borrower owns his residence
25	**overdue_credit**	Binary feature indicating the borrower's default status by the Portuguese external regulatory authority, at the time of the loan application: 0 = the applicant was not flagged as a defaulter / 1 = the applicant was flagged as being in default
26	**inflation_rate**	Portuguese inflation rate at the date of the application creation date

These 26 variables have non-null values and are of numerical type (as seen in Figure 2). Nine variables were previously categorical in nature but have been encoded into binary values for this dataset, this will be further addressed in the next subchapter, Data Preparation.

The target variable in this study is “Loan_status”, which is a binary variable where 1 represents an approved debt consolidation loan and 0 represents a rejected one.

As shown in the Figure 3, the dataset has around 85.7% more rejected loan applications than approved ones. This can lead to a bias when applying ML models, since it may tend to benefit the majority class, by predicting it with more accuracy than with the minority class (the approved loans).

Because of this imbalance, the minority class may experience a low recall, in which the model identifies approved loans as rejected loans and, in a real-world scenario, that translates into a loss of profits and possible reputational damage to the organization, driving away reliable customers and potentially harming the company’s credibility.

It is, therefore, important to technically address this issue further ahead, also on the Data Preparation subchapter.

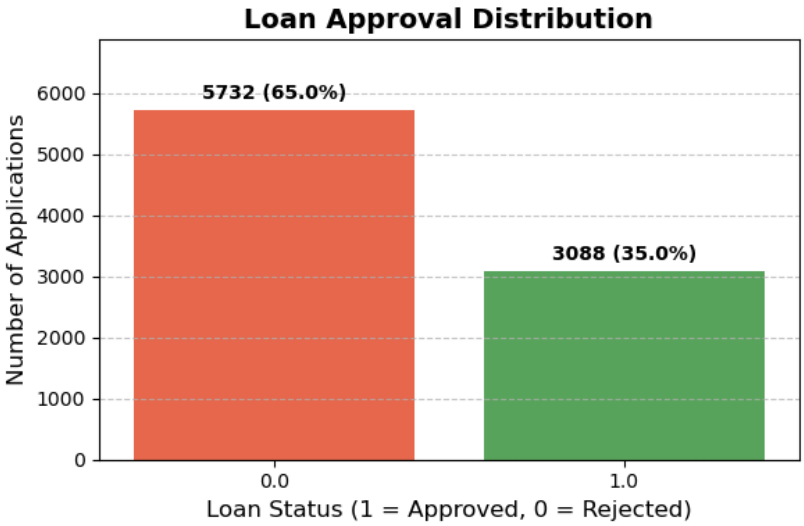
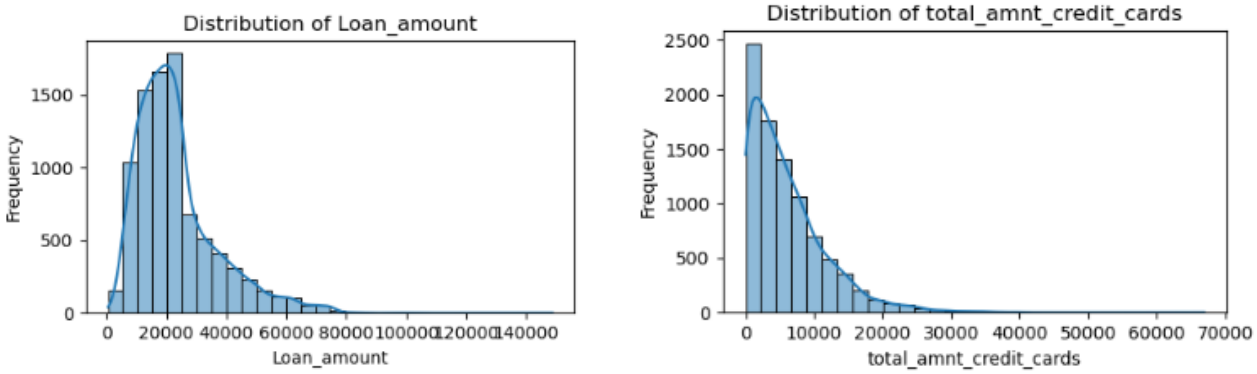


Figure 3 - Class distribution of Loan Status

Afterwards, an exploration of the distribution of the features was carried out, as can be seen in the example in Figure 4. "Loan_amount", "total_amount_consumer_loans" and total_amont_credit_cards" have a right-skewed distribution, which may require previous transformation before being used in Logistic Regression.

These histograms also reveal potential outliers, as they show a long tail at the higher end. A careful evaluation will be carried out on Data Preparation phase to understand if these outliers are important or should be removed.



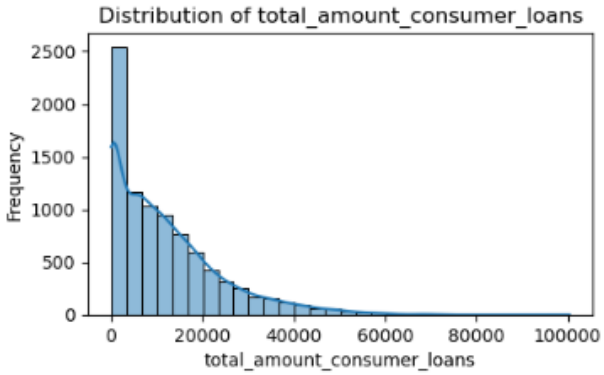


Figure 4 - Distributions of Loan and Debt-Related Financial Features

An analysis of multicollinearity between independent variables is of great importance, especially before applying a Logistic Regression model.

Highly correlated variables make it difficult to identify each variable's contribution to the model's interpretability and its utility for decision-making. They also increase standard error, by making coefficients appear statistically redundant even when they may be important.

To analyze the existence of multicollinearity between variables, a correlation matrix heatmap that visually represents the Pearson correlation coefficients between the variables of the dataset can be seen in Figure 5.

It is possible to verify, for instance, that the variables "Loan_amount", "total_amount_charge_client" and "installment_amount" are highly correlated, as well as "net_salary_1" with "net_monthly_income".

In relation to the target variable, "loan_status" shows weak correlations with the other variables, indicating that these variables, when analyzed individually, may not be strong enough to predict loan approval outcomes. Thus, this further highlights the interest in using the Random Forest algorithm. This technique is appropriate to capture non-linear relationships and multivariate interactions, as opposed to Logistic Regression, that may require feature engineering to achieve better prediction results.

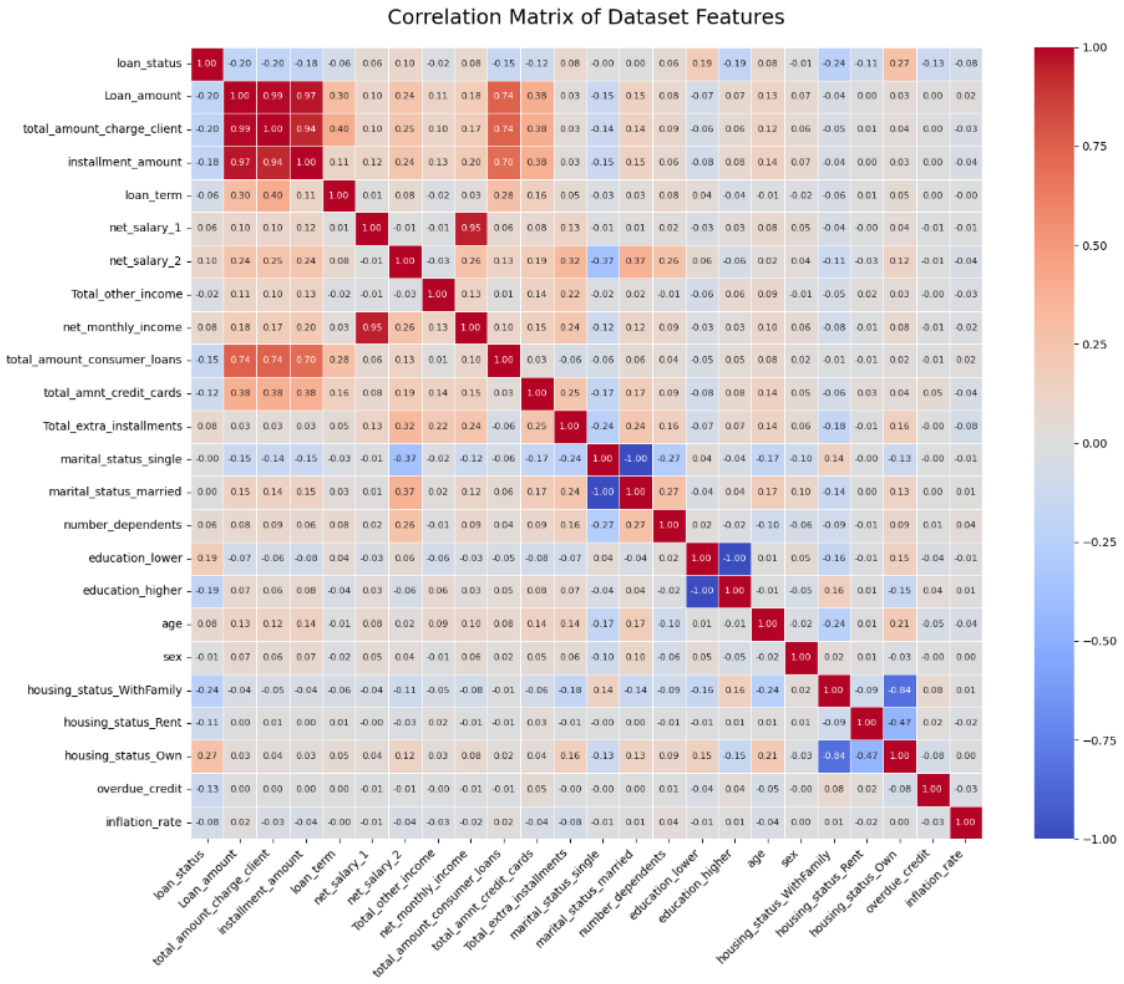


Figure 5 - Correlation Matrix of Dataset Features

3.3. DATA PREPARATION

The data preparation process is an essential step to address the gaps and flaws identified on the previous subchapter, Data Understanding.

It involves cleaning, transforming and organizing the data into a final format that is more suitable and optimized for implementing the models in this study, Logistic Regression and Random Forest.

This phase began with the collection of raw data on debt consolidation applications from the organization's database. SAS Enterprise Guide was used for data extraction and to do preliminary preparation such as initial filtering and formatting. Subsequently, Python was utilized for more advanced data processing, including cleaning, transformation, exploratory analysis, feature engineering and model preparation.

Initially, the nature of the customer was filtered, so to include only in-house customers.

Subsequently, the status of loan processes was assessed, and it was decided to apply the Binary Encoding method. This feature was originally categorical and included various statuses such as: “Approved”, “Rejected”, “Cancelled”, “Withdrawal” and others not deemed relevant for the main objectives of this research.

Only the options “Approved” and “Rejected” were considered, which we later transformed into a numeric binary variable, “loan_status”, with values 0 and 1, as previously described in Table 1.

This method is important to align the data with the study objectives on predicting loan approval, exclude irrelevant categories and to be used more effectively by the ML algorithms. As Matteucci et al. (2023) highlighted, Binary Encoding is a reliable and efficient method for preprocessing categorical variables, particularly for binary classification tasks.

Regarding the “age” feature, the customer’s age at the time of the application submission was determined by calculating the difference between the application submission date and their date of birth.

The “overdue_credit” was also created as a numeric binary feature that encompasses the default information of the applicant at the time of the application’s submission. This data was collected from a Portuguese external regulatory agency, Banco de Portugal, that centralizes this kind of information.

The “net_monthly_income” was calculated based on the sum of the net salaries of the first and second proponent when applicable (“net_salary_1”, “net_salary_2”) and also the “total_other_income” that the customer might also have and can be considered to the customer’s financial effort rate. This feature might give a better understanding of the customer’s global income.

As for the “housing_status_WithFamily”, “housing_status_Rent” and “housing_status_Own”, they were transformed by the application of the method One-Hot Encoding, from one categorical variable to the current three numerical binary variables available. The marital status and education features were also transformed with the same approach. As Hancock et al. (2020) described, this technique is common and straightforward to use as a first step to transform categorical data into a more suitable format for ML algorithms, such as numerical formats.

3.3.1. HANDLING CLASS IMBALANCE ON THE TARGET VARIABLE “LOAN_STATUS”

As previously illustrated in Figure 3, the distribution of the target variable “loan_status” is imbalanced. If this is not properly addressed, the ML models applied may become biased to the majority class, the rejected loans, resulting in worse performance when predicting the minority class.

For this reason, the Synthetic Minority Oversampling Technique (SMOTE) was used as part of the data preparation process. This algorithm implements an oversampling strategy to rebalance the original training set by generating synthetic examples, instead of replicating them from the minority class (Fernández et al., 2018).

It involves identifying the k-nearest neighbors for each minority class instance, selecting a random neighbour, and then generating synthetic data by interpolating between the original occurrence and the selected neighbour. This method is repeated until the dataset is balanced, ensuring equal representation for both classes and, in this case, a more reliable prediction of loan approvals, which is essential for the objectives of this research. The outcome can be seen of Figure 6.

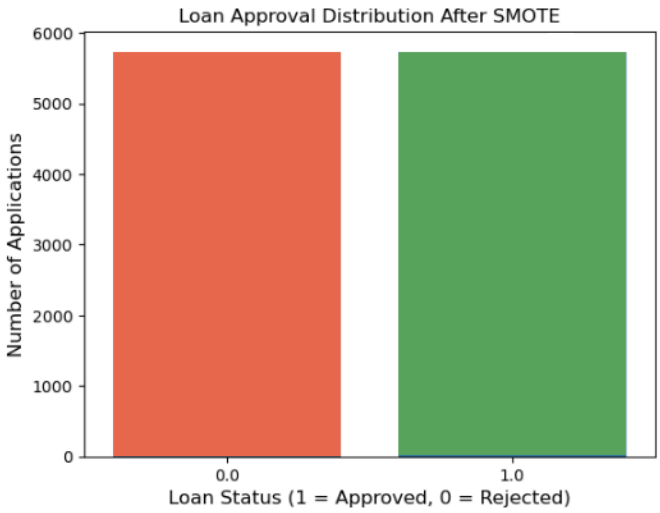


Figure 6 - Loan Approval Distribution after SMOTE

3.3.2. OUTLIERS

After identifying outliers in the features, it was decided to do the outlier capping technique to 9 features. With this, outliers are replaced with the closest acceptable value. The percentile range chosen was the 1st and 99th percentiles.

This technique has the advantage of improving data quality while preserving all rows in the dataset, as it minimizes the influence of these extreme values, particularly in linear models such as Logistic Regression. An example of the application of outlier capping can be seen on Figure 7.

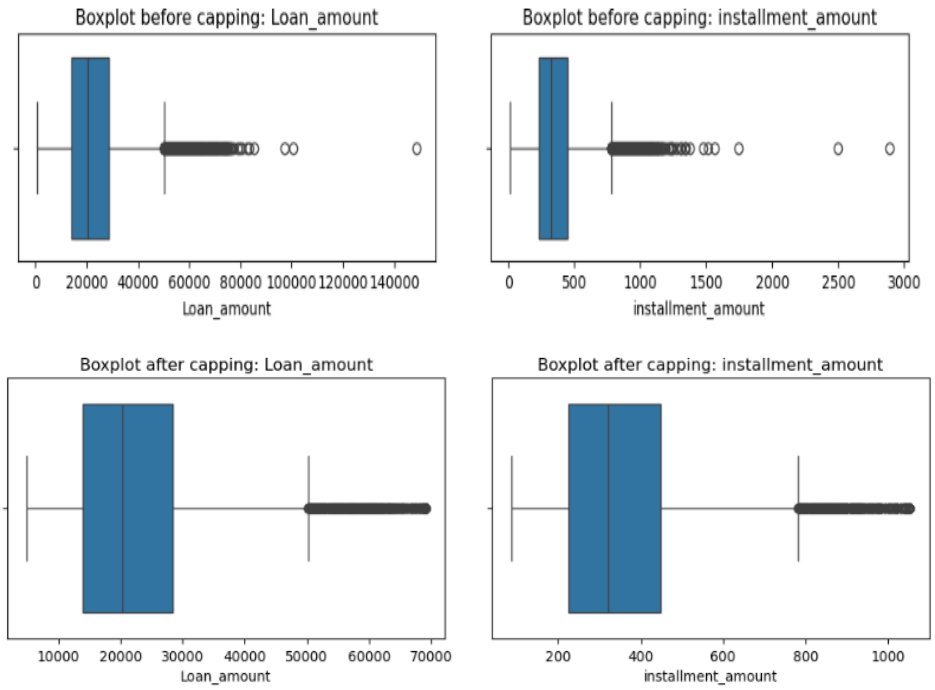


Figure 7 - Example of Outlier capping on “Loan_amount” and “Installment_amount” Distributions

3.3.3. HANDLING HIGH CORRELATION

As described in the previous chapter of Data Preparation, Figure 5, multicollinearity can hide each variable's individual contribution to the model, reducing interpretability. It can also increase standard errors, by making some coefficients seem unnecessary even if they are relevant to the model.

Therefore, “total_amount_charge_client”, “installment_amount” and “net_salary_1” were dropped, as it can be observed in Figure 8.

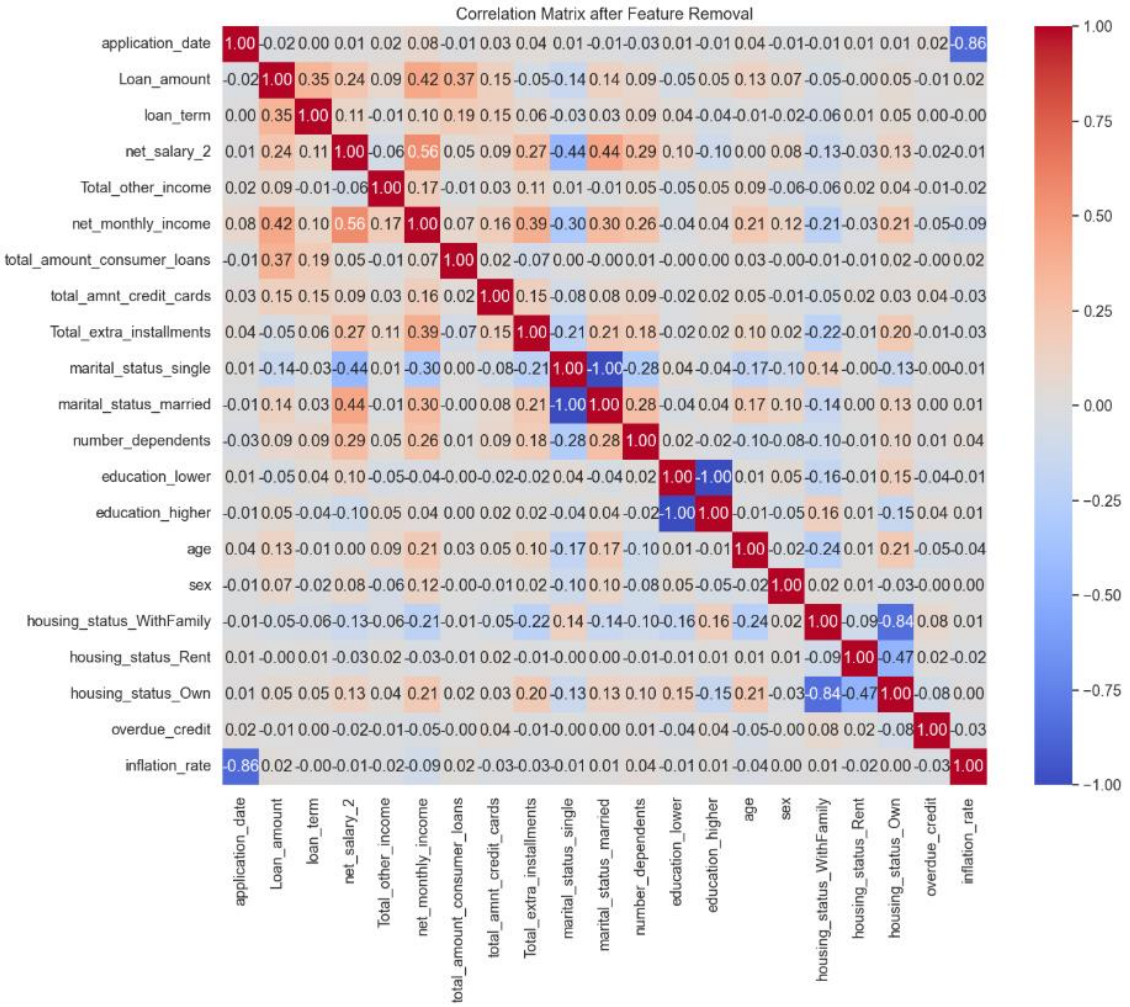


Figure 8 - Correlation heatmap after Feature Removal

3.3.4. SKEWNESS

Addressing skewness, as partially observed in Figure 4, is essential for Logistic Regression because it can distort the linear relationship between the predictor variables and the target variable, potentially resulting in inaccurate predictions.

As for Random Forest, it is less significant, but it is still beneficial for improving feature significance calculations.

For this reason, a logarithm transformation was applied to reduce skewness of the features, by normalizing the feature distributions, an example can be observed in Figure 9.

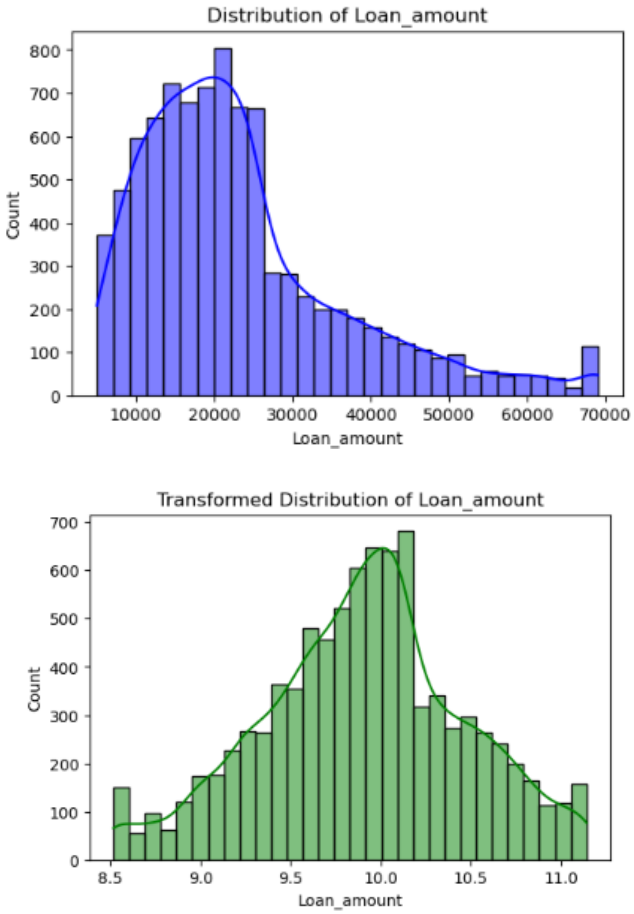


Figure 9 - “Loan_amount” distribution before and after transformation.

3.3.5. SCALING

Scaling was applied to standardize the numerical variables of the dataset, examples can be seen in Figure 10. This step is important because of the difference in scale between the variables.

By using the StandardScaler, a known standardization process, all numerical features were transformed to have a mean of 0 and a standard deviation of 1.

This process centers and scales each feature evenly, ensuring comparability and eliminating any biases caused by differences in their original scales. This prevents any single feature from disproportionately influencing the machine learning models like Logistic Regression, which is sensitive to feature scales.

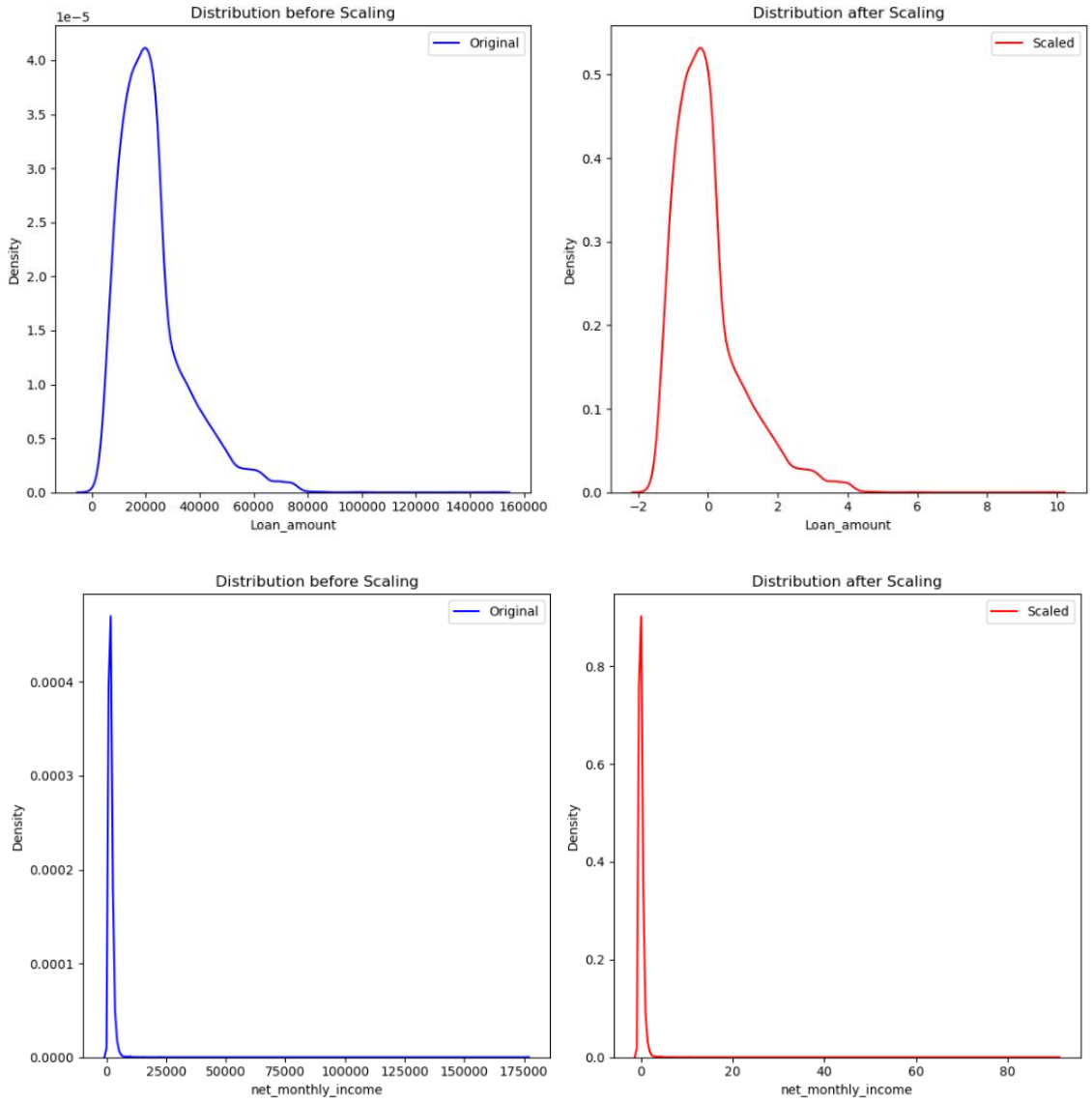


Figure 10 - Comparison of “loan_amount” and “net_monthly_income” Distributions Before and After Scaling

3.4. MODELING

The choice of Logistic Regression and Random Forest models to be used in this study came from the relevance to predict loan approval outcomes, as previously summarized in the Literature Review, as well as the advantages that both models bring, whether due to their simplicity and ease of interpretation or ease of applicability.

Logistic Regression is a popular method for binary classification problems, such as loan approval prediction. It models the relationship between a binary dependent variable and one or more independent variables by estimating probabilities using a logistic function, ensuring that the predicted probabilities fall between 0 and 1 (James et al., 2013).

Additionally, the impact of each variable on the target variable is represented by coefficients, providing valuable insights into how specific factors, such as income or loan amount, affect the probability of loan approval. This model also serves as a reliable benchmark model for comparing more complex algorithms (Hosmer & Lemeshow, 2000; Kuhn & Johnson, 2013).

The Logistic Regression model in this study was configured with a balanced class weighting to address class imbalance in the dataset, ensuring that identical importance is given to both approved and rejected loans.

A fixed random seed (with value of 42) was used to guarantee that results remained consistent throughout multiple runs. The model was trained using the fit method, in which the model learns from the input data "X_train" and its target labels "Y_train", and thus, optimizing coefficients to predict binary results.

Its performance is evaluated on the test dataset (X_test) by using various common metrics, such as precision, recall, F1-score and AUC-ROC score for both classes (approved or rejected loans).

Random Forest, on the other hand, is an ensemble learning model that can address some limitations of the Logistic Regression model. This model does not practically require transformations while it captures non-linear relationships between features and the target variable.

Plus, its ability to handle imbalanced datasets, a common challenge in loan approval prediction, makes it particularly effective in financial applications (Breiman, 2001; Liaw & Wiener, 2002; Chen, Liaw, & Breiman, 2004).

The Random Forest model in this work was configured by using the "RandomForestClassifier", from the Scikit-learn library in Python, with a fixed random seed, again to guarantee that the results remain the same throughout various runs.

This model constructs multiple decision trees during the training and combines their predictions, relying on the majority vote, to make a final decision, and in this scenario, to classify for loan approval.

As with Logistic Regression, this model's performance is evaluated on the test dataset (X_test), with metrics such as precision, recall, F1-score and AUC-ROC score.

3.5. EVALUATION

The performance metrics used to evaluate both models included precision, recall, F1-score and AUC-ROC.

Precision can be defined as a measure of how many positive predictions were correct, calculated as the ratio of true positives to the total predicted positives, as shown in Equation (1).

For loan approval predictions, this metric is important to minimize false positives (when an unqualified loan application is incorrectly approved), as it can lead to potential financial losses for the company, if the applicant defaults.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

Recall is a ratio that identifies the proportion of correctly predicted positives out of all actual positive instances. It captures all actual positive instances, therefore identifying all eligible applications, its formula can be seen below, in Equation (2)

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

The F1-score combines both precision and recall into a single metric, balancing their trade-offs and thus, ensuring the model approves loans with accuracy and completeness. It is calculated as the harmonic mean of these two metrics and is particularly useful when the dataset is imbalanced towards one class.

If F1-score is close to value 1, it demonstrates that the model is effective in making reliable predictions for the positive class. On the other hand, if the metric is closer to value 0, then the model used struggles to make predictions for the positive class with accuracy and completeness.

AUC-ROC is used for classification models to assess how well they distinguish between positive and negative classes.

The ROC curve (Receiver Operating Characteristic Curve) illustrates graphically the trade-off between the Recall (true positive rate) and the false positive rate at various thresholds. The AUC represents the Area Under the Curve, which summarizes this performance into a single value varying from 0 to 1.

An AUC close to 1 represents near perfect classification, while closer to 0 indicates poor performance, with the model often making incorrect predictions.

3.5.1. MODELS' PERFORMANCE

After training both models, Logistic Regression and Random Forest, an evaluation on the outcomes was performed.

For the Logistic Regression model, it achieved a precision of 71% in terms of accuracy for the predicted approved loans (class = 1). By analyzing the recall metric, 79% of the actual approved loans were correctly identified.

The F1-score marked 75% in the positive class, reflecting a good balance between identifying true approved loans and minimizing false approvals.

The overall accuracy of this model was 73%, while the AUC-ROC score was 79.8%, indicating it performs well in distinguishing between approved and non-approved loans, though its performance might be improved.

Regarding the Random Forest model, it achieved identical values for precision, recall and F1-score, with 80% for class 1 (approved loans), meaning 80% accuracy in predicting approved loans, in identifying them successfully and the model's robustness.

The overall accuracy was also 80% and the AUC-ROC was 88.9%, indicating the model has a strong predictive capability for loan approval predictions.

The visualization of these findings can be seen on both Figure 11.

```
Logistic Regression Metrics
------------------------------
Classification Report:
              precision    recall  f1-score   support

         0.0       0.76      0.67      0.72      1720
         1.0       0.71      0.79      0.75      1720

    accuracy                           0.73      3440
   macro avg       0.74      0.73      0.73      3440
weighted avg       0.74      0.73      0.73      3440

AUC-ROC: 0.7982713628988642

Random Forest Metrics
------------------------------
Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.80      0.80      1720
         1.0       0.80      0.80      0.80      1720

    accuracy                           0.80      3440
   macro avg       0.80      0.80      0.80      3440
weighted avg       0.80      0.80      0.80      3440

AUC-ROC: 0.889351845592212
```

Figure 11 - Classification report of models Logistic Regression and Random Forest

5. RESULTS AND DISCUSSION

This chapter provides a comprehensive assessment of the models used for loan approval prediction, in terms of performance, strengths and limitations.

The outcomes are analyzed not only from a statistical standpoint but also in the context of business purposes, since the models were selected for a comparative analysis rather than exclusively to determine which one gives a more predictive result.

5.1. LOGISTIC REGRESSION MODEL

The Logistic Regression model's metrics provide valuable insights in predicting loan approvals. The AUC-ROC is important to evaluate the model's performance in prioritizing the positive class (approved loans), across various thresholds. Its value of 79.8% demonstrates the model's capability to distinguish between the two classes, approved and rejected loans, and effectively reducing misclassification between them.

Having a precision of 71% for approved loans, the model ensures that most loans predicted as approved are accurate, accounting for the risk of false positives, which is of most importance for financial institutions in general, not interesting in having default borrowers.

Considering the recall value of 79%, it shows that the model can successfully identify actual approved loans, reducing, therefore, the probability of false negatives. This is relevant for the company, so it does not miss profit opportunities, and it does not produce customer dissatisfaction.

The F1-score of 75% for the positive class shows a balance between precision and recall, which is essential to maintain model reliability.

An accuracy of 73% and consistent macro averages of 74% across metrics indicate an interesting performance even with dataset imbalances.

Overall, the Logistic Regression model provides effective predictions and, due to its simplicity, it is a highly scalable option for low-computational environments, where speed and simplicity are keys, or when interpretation is the main objective.

Nevertheless, these advantages come with the drawback of potentially reduced performance compared to more complex models.

5.2. RANDOM FOREST MODEL

With an AUC-ROC value of 88.9% and an 80% in precision, recall and F1-score for the positive class, the Random Forest model outperformed the Logistic Regression across all metrics.

The performance comparison graph, as shown in Figure 12, clearly illustrates the superiority of Random Forest over Logistic Regression in all evaluated and already discussed metrics.

Even though both models do perform well, the difference in AUC-ROC emphasizes Random Forest's advantage in capturing complex relationships in the data and guaranteeing efficiency in decision-making for loan approvals.

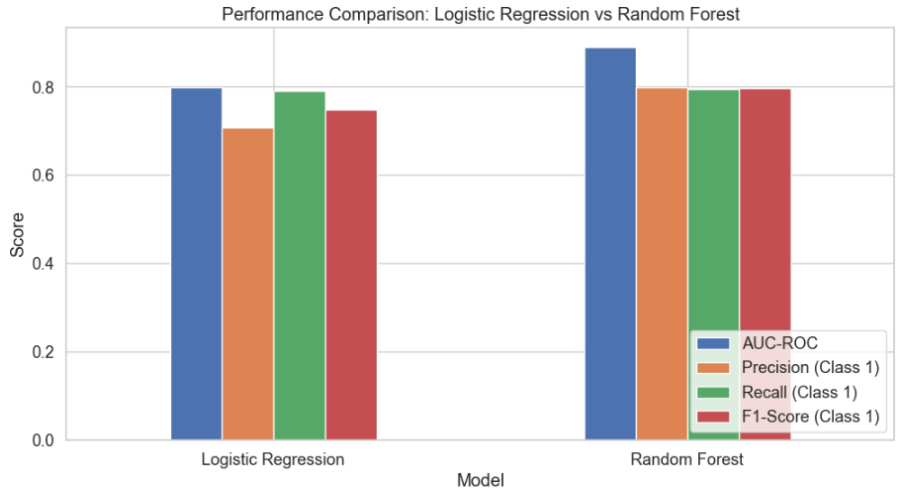


Figure 12 - Model Performance Comparison between Logistic Regression and Random Forest

In addition, an analysis of the importance of variables in the Random Forest model was performed and demonstrated that features such as "net_monthly_income", "total_amount_consumer_loans", "loan_amount" and "total_extra_installments" are the most influential in predicting loan approval, as shown in Figure 13.

These variables highlight the model's priority on factors related to customers' financial capacity. On the opposite, variables such as "housing_status_Rent", "total_other_income" and "marital_status_married" were less relevant.

This information is important to understand how this model makes decisions, providing transparency even in a "black box" algorithm.

In addition, the most important variables identified can be prioritized by financial institutions to refine their credit approval systems, minimizing risks and optimizing processes. Nevertheless, it is worth noting that this analysis is relative to the dataset used and should be interpreted with caution, although the relevance of these variables is in line with previous studies on credit prediction.

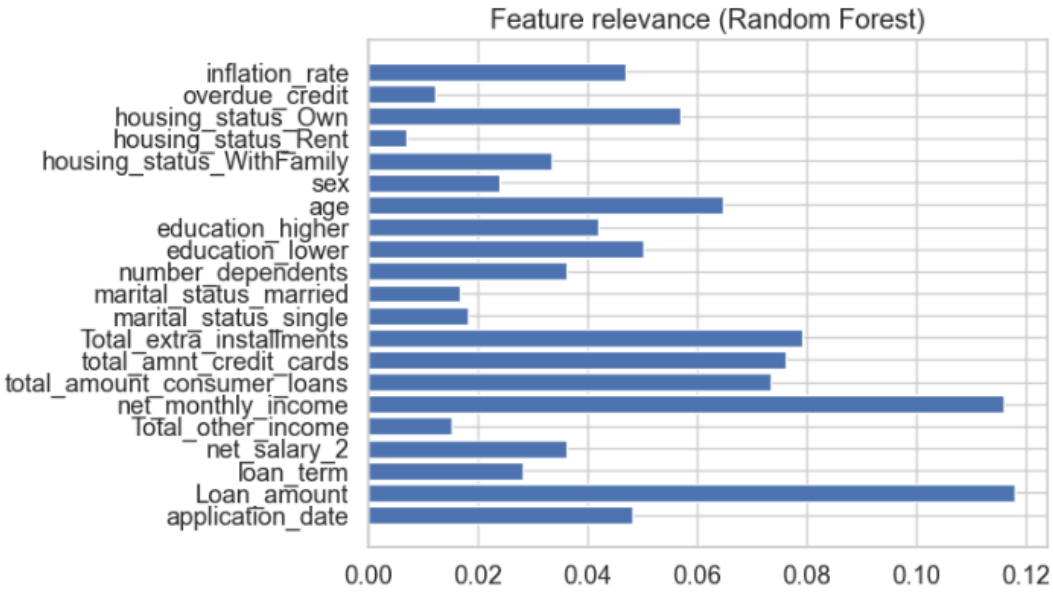


Figure 13 - Feature Importance in Random Forest

In summary, the results imply that Logistic Regression is an effective model for binary classifications, such as loan approval prediction, due to being an easy-to-understand model. It provides insights related to the features contributions to the model, which can be useful to understand the main factors contributing to the predictions.

In this study, this algorithm had moderate precision and recall scores, indicating more limitations in capturing non-linear relationships, in contrast with the Random Forest algorithm.

On the other hand, Random Forest's superiority performance proves its capacity to handle complex interactions and therefore, find better patterns in the data.

This is extremely important in financial contexts where loan approval decisions involve multiple interdependent factors.

Its higher AUC-ROC score demonstrates its advantage in class separation and the results in precision and recall confirm reliable predictions for both approved and rejected loans.

The data preparation steps, such as handling class imbalance, removing multicollinear features, transforming skewed data and scaling, positively impacted both models' performance, but had a more evident effect on Logistic Regression, due to the nature of the algorithm being sensitive to extreme values and non-linearities.

Random Forest, by comparison, is less affected by multicollinear features and skewed data because of its algorithm nature and the splitting procedures of decision tree.

This highlights the importance of preprocessing the data to achieve reliable outcomes. Additionally, the dataset balancing through SMOTE contributed to a propoer performance across classes, guaranteeing that predictions for the minority class (approved loans) were not disproportionately ignored.

Overall, the Random Forest algorithm stands out as the ideal model for loan approval prediction and balanced results.

Nevertheless, Logistic Regression continues to be of great value because of its interpretability, making a suitable choice when understanding model decisions is critical, especially for key users in an organization, or when speed, rather than accuracy, is a more relevant priority.

It is, therefore, important to choose the appropriate model based on the objectives of the organization and its application, depending on which factor is more determinant, if accuracy or interpretation.

These outcomes highlight the complementary strengths of both models in addressing the challenges and complexities of loan approval prediction, depending on specific criteria such as accuracy, interpretation or even operational requirements.

6. CONCLUSIONS

This study examined and compared the performance of Logistic Regression and Random Forest models for predicting loan approval, focusing on both their predictive capabilities and their practical relevance for organizations.

The results showed that Random Forest outperformed Logistic Regression in all metrics, such as accuracy, precision, recall, F1-score and AUC-ROC, standing out as the model with the greatest predictive power.

However, Logistic Regression proved to be a valuable algorithm due to its simplicity and interpretability, which are essential factors in scenarios where understanding the model's decisions or response speed is critical for stakeholders.

The data preparation process, including the treatment of multicollinearity and the transformation of skewed data, played an important role in improving the performance of these models. As previously seen, Logistic Regression, in particular, benefited from these processes due to its dependence on well-structured data to produce satisfactory results.

On the other hand, Random Forest demonstrated greater strength to data imperfections, highlighting its reliability for dealing with complex and imbalanced data sets.

This research contributes to the field of predictive modeling in financial decision-making by providing a structured evaluation of these two models, Logistic Regression and Random Forest, in the context of loan approval systems, specifically in debt consolidation loans.

It highlights how machine learning approaches can improve decision accuracy, reduce financial risks for the company, and optimize operational processes.

In conclusion, this study emphasizes that the choice of predictive model must be aligned with the specific objectives required, whether these are in terms of accuracy of results or interpretation and speed of results.

It also points out the relevance and importance of these algorithms as complementary forces, demonstrating the potential of machine learning to transform financial systems and address the challenges arising by this sector’s ever-growing and continuously evolving demands.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The main limitations of this work include some factors that influenced the results and the applicability of the models.

First, the dataset used has limitations, since it may not fully represent the diversity of profiles of applicants for debt consolidation loans, since it only deals with in-house customers of the organization, with verifiable information on the customer profile. Besides this, not all variables could be included in the dataset, such as information on the number of debts and specific credit entities, which could have been a relevant addition to the models.

Variability in demographic characteristics, financial behaviors, and regional differences may reduce the scalability of the findings. In addition, the class imbalance in the dataset, even with the application of techniques such as SMOTE, may still have influenced the performance of the models, possibly exaggerating their effectiveness in identifying less frequent cases, such as the positive class, loan approval.

Another important point is the limitation of the scope of the models analyzed. The dissertation focused exclusively on the comparison between Logistic Regression and Random Forest, leaving aside more advanced algorithms, such as Gradient Boosting Machines (XGBoost, LightGBM, or CatBoost) or deep learning techniques, which could offer greater precision or additional insights. In the case of Logistic Regression, its reliance on linear relationships between variables and the target variable may have limited its performance, even with multicollinearity addressed.

Incorporating interpretability tools, such as Counterfactual Explanations, which shows “what if” scenarios and explain how changes in the input variables affect predictions may be interesting to explore, as it could help to understand why some loans were denied and what variables could improve their odds.

Regarding Random Forest, the model provided strong results, but its interpretability is limited compared to Logistic Regression, which may hold back its application in regulatory or some decision-making contexts, where transparency can be crucial.

Finally, this study did not perform a cost-benefit analysis, which could quantify the financial impacts of false positives (approval of risky loans) and false negatives (rejection of eligible applicants), providing a more comprehensive practical perspective. Furthermore, the models were not tested in real-world scenarios, where operational challenges may influence the results.

These limitations open opportunities for future work that could address these gaps and improve the models for real-world applications.

BIBLIOGRAPHICAL REFERENCES

Aniceto, M. C., Barboza, F., & Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. Future Business Journal, 6(1). https://doi.org/10.1186/s43093-020-00041-w

Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., Sztachelski, J., & Wojewnik, P. (2021). Enabling machine learning algorithms for credit scoring—Explainable artificial intelligence (XAI) methods for clear understanding complex predictive models. arXiv preprint arXiv:2104.06735. https://doi.org/10.48550/arXiv.2104.06735

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brynjolfsson, E., & Collis, A. (2019). How should we measure the digital economy. Harvard business review, 97(6), 140-148.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS Inc. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley.

ElMasry, M. H. A. M. T. (2019). Machine learning approach for credit score analysis: a case study of predicting mortgage loan defaults (Master's thesis, Universidade NOVA de Lisboa (Portugal)).

Eurostat. (2023). Household consumption by purpose. Retrieved September 30, 2024, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Household_consumption_by_purpose

Fekadu, R., Getachew, A., Tadele, Y., Ali, N., & Goytom, I. (2022). Machine learning models evaluation and feature importance analysis on NPL dataset. arXiv preprint arXiv:2209.09638. https://doi.org/10.48550/arXiv.2209.09638

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192

Guttmann, R., & Plihon, D. (2010). Consumer debt and financial fragility. International Review of Applied Economics, 24(3), 269–283. https://doi.org/10.1080/02692171003701420

Hancock, J.T., Khoshgoftaar, T.M. Survey on categorical data for neural networks. J Big Data 7, 28 (2020). https://doi.org/10.1186/s40537-020-00305-w

Hosmer, D. W., Jr., & Lemeshow, S. (2000). Applied Logistic Regression. New York, NY: Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York, NY: Springer

Lansing, K. J. (2011). Consumers and the economy, part I: Household credit and personal saving. *FRBSF Economic Letter*, *1*, 1-5.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.

Matteucci, F., Arzamasov, V., & Boehm, K. (2023). A benchmark of categorical encoders for binary classification. arXiv preprint arXiv:2307.09191. Retrieved from [https://arxiv.org/abs/2307.09191]

Saini, P. S., Bhatnagar, A., & Rani, L. (2023). Loan approval prediction using machine learning: A comparative analysis of classification algorithms. 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 1821–1826. https://doi.org/10.1109/ICACITE57410.2023.10182799

Shearer, C. (2020). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13–22.

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). An approach for prediction of loan approval using machine learning algorithm. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), 490–494. IEEE.

Shinde, A., Patil, Y., Kotian, I., Shinde, A., & Gulwani, R. (2022). Loan prediction system using machine learning. In ITM Web of Conferences (Vol. 44, p. 03019). EDP Sciences.

The Global Economy. (2024). Household consumption, percent of GDP - Country rankings https://www.theglobaleconomy.com/rankings/household_consumption/European-union/

The World Bank. (n.d.) Metadata Glossary. Retrieved May 22, 2024, from https://databank.worldbank.org/metadataglossary/world-development-indicators/series/NE.CON.PRVT.PC.KD

Viswanatha, V., Ramachandra, A. C., Vishwas, K. N., & Adithya, G. (2023). Prediction of loan approval in banks using machine learning approach. International Journal of Engineering

and Management Research, 13(4). https://ijemr.vandanapublications.com/index.php/j/article/view/1318

Yildiz, A. (2023). Determining the factors for individual credit approval by applying logistic regression and hierarchical logistic regression. International Journal of Management Studies and Social Science Research, 5(6), 58–67. https://doi.org/10.56293/IJMSSSR.2023.4705

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa