Loan Default Prediction Model

Thesis · January 2023

DOI: 10.13140/RG.2.2.22985.01126

1 author:

Platur Gashi
RIINVEST College
1 PUBLICATION **2** CITATIONS

SEE PROFILE

Management of Financial and Information Systems

Master Thesis

2022/2023

Platur Gashi

LOAN DEFAULT PREDICTION MODEL

Dr. Albin Ahmeti

January / 2023

ABSTRACT

Banks frequently face the challenge of loan defaults, which are an unavoidable issue. If loans are not repaid, banks experience financial losses. To minimize this problem, they aim to keep default rates as low as possible. In the credit risk management field, banks use Machine Learning (ML) techniques to build various models for predicting loan defaults. This helps them to avoid granting loans that have a high likelihood of defaulting. ML models help detect patterns in data, which is then used to categorize new records.

This paper presents the development of several models for predicting loan defaults using a variety of Machine Learning algorithms. Both individual and ensemble types of algorithms are used. The performance of these models was assessed through analysis and comparison to determine which model has the best results.

The data used in this paper is obtained from the well-known platform "Kaggle". According to the methodology, the data is initially explored and analyzed, followed by pre-processing to prepare it for modeling. The models are then trained using each algorithm separately, and finally their performance is assessed.

A major challenge encountered was the data imbalance in the target variable, which impacted the model performance. To overcome this challenge, the SMOTE method was used to pre-process the data. This approach increased the amount of data for the imbalanced value, resulting in a significant improvement in the performance of the algorithms.

The results demonstrate that ensemble algorithms outperform individual ones in predicting loan defaults. The top-performing algorithms were Boosted Decision Trees (Boosting) and Random Forest (Bagging).

Keywords: Prediction, Loan Default, Machine Learning, Algorithm, Ensemble, SMOTE.

DECLARATION OF ORIGINALITY

I declare that: (1) This thesis represents my original work, except in cases of citations and references and (2) This thesis has not previously been used as a paper or project in this college or in other universities/colleges/institutions.

ACKNOWLEDGMENTS

Firsly, I would like to express my sincere gratitude to my mentor Albin Ahmeti for his continuous support in the writing of the thesis and scientific research, for his patience, motivation and guidance shown during the writing of this paper. His support has been essential in the realization of this paper.

Also, I would like to thank all the staff of Riinvest College for the conditions offered during my studies. A special thanks also goes to all the professors who have been part of my studies, for their time and expertise, which have influenced the quality of this paper.

An incomparable thanks goes also to my family, who have supported and helped me throughout my studies.

Finally, I would also like to thank my fiancee for her incomparable support, cooperation and motivation throughout my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ML – Machine Learning

NN – Neural Network

SVM – Support Vector Machine

SMOTE – Synthetic Minority Oversampling Technique

CBK – Central Bank of Kosova

NPL – Non-Performing Loan

CRISP-DM – Cross Industry Standard Process for Data Mining.

1 INTRODUCTION

Loans have played a significant role in the global economy, but managing them can be complex. One of the main challenges associated with loans is the risk of default. The authors define default risk as "*Obligor fails to service debt obligations due to borrower specific or market-specific factors*" (Bandyopadhyay, 2016). Loan default is also defined as " *Default is the failure to make required interest or principal repayments on a debt, whether that debt is a loan or a security. Individuals, businesses, and even countries can default on their debt obligations"* (Chen, 2022).

Loans that default are also called as non-performing loans (NPL). According to CBK annual report, the NPL ratio in Kosovo is low and stable. The ratio of non-performing loans and provision coverage (protection against NPL with reserve funds) is presented below:

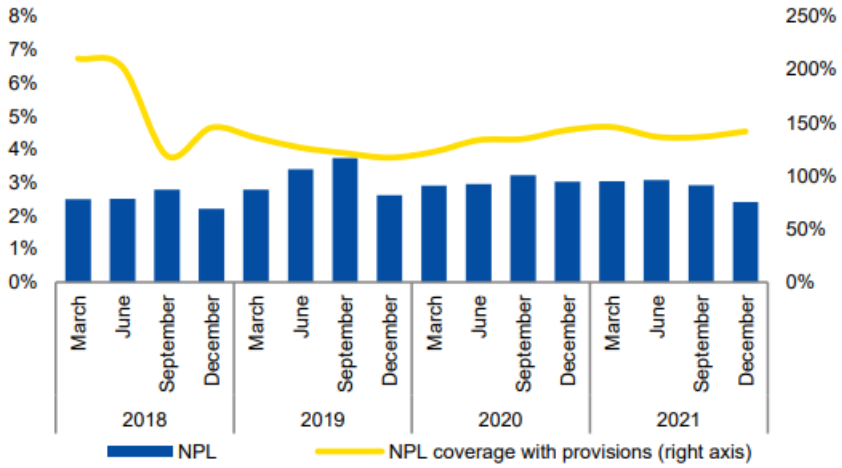

Figure 1.1 Non-Performing Loans in Kosovo. Source: (Central Bank of Kosovo, 2021).

So, according to the year-end report (Central Bank of Kosovo, 2021), the NPL ratio in Kosovo at the end of 2021 was 2.4%, marking a decrease from the previous year, which was 3%.

When compared to the region, according to the annual report (Worldbank, 2022), Kosovo has the lowest NPL ratio in the region (Western Balkans). Kosovo is well below the global average of 5.4% for the year 2021, ranking the 13th globally with the lowest NPL ratio in the world (according to existing data for 2021).

The table below displays the NPL ratios in the region:

Table 1.1 Non-Performing Loans Ratio in the Region

State	2020	2021
Albania	7.65	5.39
Bosnia and Herzegovina	6.12	N/A
Montenegro	5.87	6.83
North Macedonia	3.26	N/A
Kosovo*	2.47	2.13
Serbia	N/A	N/A

** Kosovo's NPL levels differ slightly from the CBK report, since these measurements are based from the World Bank.*

According to the same source, (Worldbank, 2022), below are presented the top 15 countries with the lowest NPL ratio in the world:

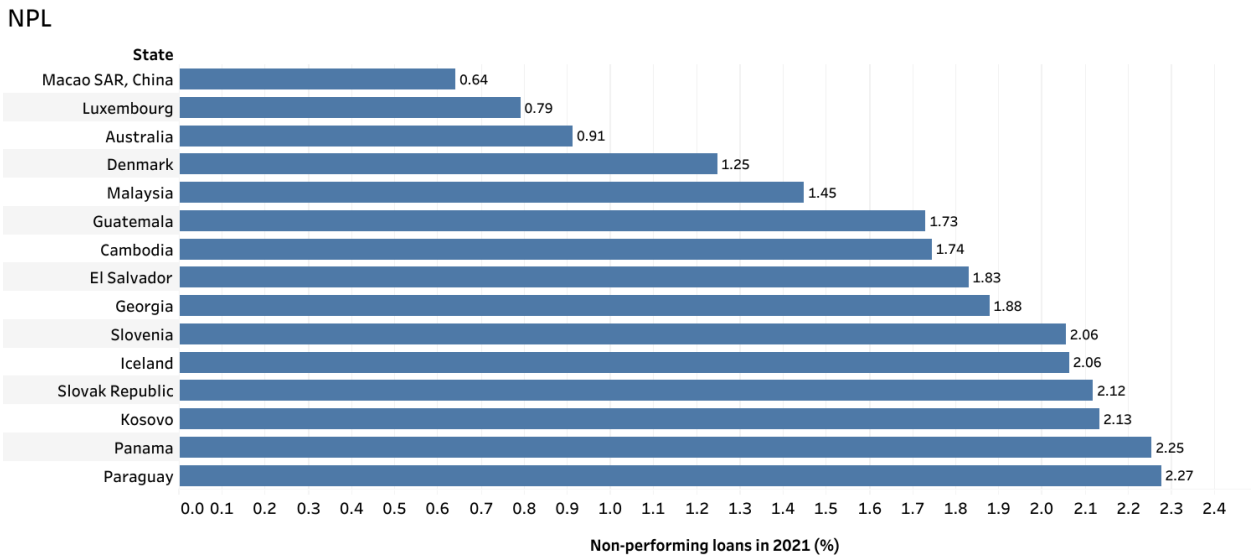

Figure 1.2 Top 15 countries with the lowest NPL ratio.

2 FUNDAMENTALS

This chapter outlines key methods and concepts utilized in this study, starting with the data processing methods applied. It then covers the ML algorithms used in the paper, as well as those utilized by referenced authors. Finally, the performance estimators are explained to provide an understanding of how the model is evaluated.

2.1 Data Processing Methods

2.1.1 Missing Values

Missing values in data can pose challenges for some algorithms, affecting their performance. There are several methods to handle missing data, such as removing entire variables or rows if they contain numerous missing values. For numeric data types, replacing missing values with the mean is a viable option, and for categorical data types, the mode can be used as a replacement. The use of mean and mode substitutions was also employed in previous works of (Tariq, et al., 2019) and (Victor & Raheem, 2021). The most useful ways for dealing with missing values are:

- Substitution by mean
- Substitution by median
- Substitution by mode
- Row deletion (when there are too many missing values)
- Variable deletion (when there are many missing values)

2.1.2 Data Normalization

This method is used for continuous numerical variables so that the data is more suitable for some algorithms. Normalizing the data involves the transformation the data in order thar they appear similar across all records, particularly when the rank among them is very high. "*Normalization*

prevents one attribute dominating the distance results because of large values." (Kotu & Deshpande, 2019).

2.1.3 Feature Selection – Filter Based Feature Selection

The Filter Based Feature Selection method is a popular technique for identifying the variables that have the most impact on the target variable. "*The main aim of these techniques is to remove irrelevant or redundant features from the dataset*" (Yildirim, 2015). According to this author, there are several techniques for feature selection, such as:

- Information Gain
- Relief
- Principal Component Analysis (PCA)
- Correlation Based Feature Selection

2.1.4 Imbalanced Data. SMOTE Operator

"*SMOTE is a machine learning technique that solves problems that occur when using an imbalanced data set*" (Korstanje, 2021). In many cases, we occur data where the target variable is imbalanced. For example, if the target variable has two values (YES and NO), and one value appears in 90% of cases while the other in only 10%, then we can say that this dataset is imbalanced. In these cases, a naive algorithm which would classify based of the dominant values, then it would be 90% correct in our case. According to (Korstanje, 2021), when using the SMOTE method, the prediction performance for the minority class in the dataset is increased. What the SMOTE operator does, is that it creates new records of the minor value, and thus balances the values in the target variable.

2.2 Other Important Concepts in ML

In this section, some additional key concepts related to ML are explained which are used and referenced throughout the paper.

Supervised & Unsupervised: This concept refers to the approach used for training data in the modeling process. In the case when the training is Supervised, the target variable is present in the dataset. Thus, the model is trained based on the historical data based on the values that are present in the target variable.
On the other hand, in unsupervised learning, the target variable is not present in the dataset. Instead, the model is trained by discovering patterns and relationships among the data, which then helps in predicting the target variable.

Overfitting vs Underfitting: According to (Nautiyal, 2022)**,** The concept of Overfitting and Underfitting refers to the performance of a model on the training and test data. Overfitting occurs when the model fits the training data too closely and performs well on it, but performs poorly on the test data. This can happen when the model is overly complex, trained too extensively, or has a large amount of data.
On the other hand, Underfitting occurs when the model performs poorly on both the training and test data, this is usually a result of a simple model, insufficient data, or poorly processed data.

Correlation: The correlation coefficient is a statistical measurement of the linear relationship between two numerical variables, indicating the extent to which one variable changes in response to a change in the other. The correlation coefficient can range from -1 to 1.

- When the correlation is close to the value 1 or -1, indicates a strong correlation between the two variables.
- When the correlation is close to the value 0.5 or -0.5, indicates an average correlation between the two variables.
- When the correlation is close to 0, indicates that the two variables are not related at all.

It's important to note that a high correlation does not imply a cause-and-effect relationship between the variables (causation).

2.3 ML Algorithms

To build a model that predicts loan defaults, it needs to be trained using algorithms. In this study, various algorithms are used and analyzed to determine which one is the best performing one. The algorithms used in this research are:

- Decision Tree
- Logistic Regression
- Artificial Neural Network
- SVM
- Ensemble Learning

2.3.1 Decision Tree

"*Decision trees (also known as classification trees) are probably one of the most intuitive and frequently used data science techniques*" (Kotu & Deshpande, 2019). This algorithm is used for classification problems, i.e., when the target variable is divided into classes (e.g., 0 and 1).
Decision Tree works by separating into different nodes, where each node represents a variable. The variable that affects the prediction the most is placed at the beginning and then continuing with the other variables below which are of less important (hierarchy). The tree arrows show the values that a particular record takes. The last part of the tree is the leaf, which represents the value of the prediction, which shows in which class the record will be predicted. Below is a visual representation of a decision tree:

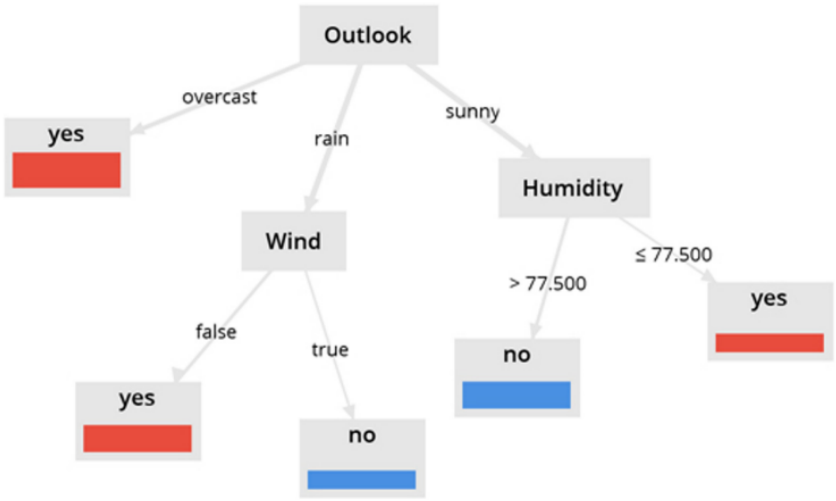

Figure 2.1 Visual Representation of the Decision Tree. Source: (Kotu & Deshpande, 2019)

To explain the output, when a new case/record must be predicted, it goes through the tree and ends at a certain leaf which shows the value that is predicted for that case.

2.3.2 Logistic Regression

This algorithm is a statistical method that predicts the probability of each class. Typically, Logistic Regression is used for binary classification problems, where the target variable has only two possible classes (e.g., 0 and 1). Each record is assigned a predicted class (e.g., 0 or 1) and the algorithm displays the probability of that prediction (e.g., an 80% chance that the class is 1). "*The Logistic Regression models were recognized as the most appropriate models in deciding to grant credit to individuals and regarded as the industry standard in credit scoring model development*" (Victor & Raheem, 2021).

Below is shown the visual representation of the Logistic Regression function:

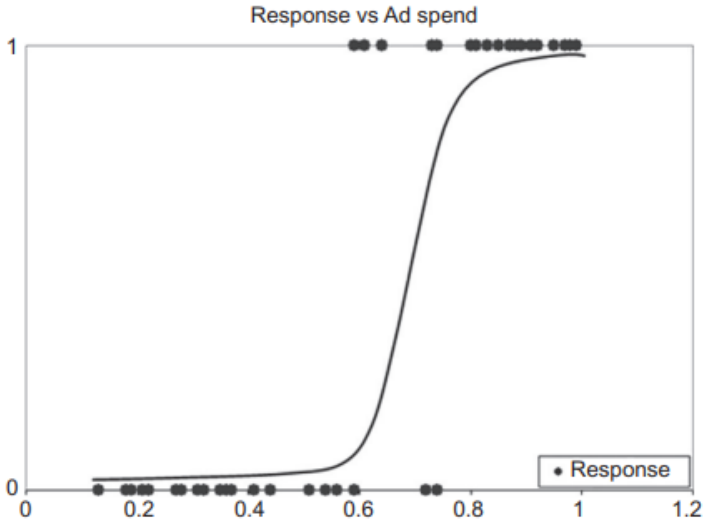

Figure 2.2 Visual Presentation of Logistic Regression. Source: *(Kotu & Deshpande, 2019)*

From the illustration, it is evident that the records can only assume two specific values, for instance 0 and 1. The bottom axis displays all values as "0" and the top axis displays all values as "1".

2.3.3 Neural Network

This algorithm is more advanced than the classical algorithms. "*The NN is a computational and mathematical model inspired by the biological nervous system*" (Kotu & Deshpande, 2019). "*A neural network consists of a set of neurons that are connected together. A neuron takes a set of numeric values as input and maps them to a single output value. At its core, a neuron is simply a multi-input linear-regression function.*" (Kelleher & Tierney, 2018).

So, its basic function is the linear regression formula which is shown in figure 2.3, but then these values are weighted through different nodes and finally revealed as output.

$$Y = 1 + 2X_1 + 3X_2 + 4X_3$$

Figure 2.3 Linear Regression Formula. Source*:* (Kotu & Deshpande, 2019)

The typology of the NN model is visually presented below:

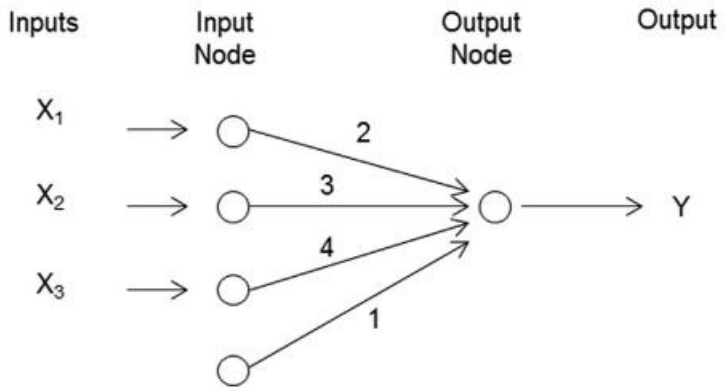

Figure 2.4 Typology of the NN model. Source: *(Kotu & Deshpande, 2019)*

Therefore, this algorithm is highly complex and demands a significant amount of computing power due to the numerous calculations it performs to reach the final result.

2.3.4 SVM

Another algorithm used for classification problems, is the Support Vector Machine (SVM). "*It works on the principle of fitting a boundary to a region of points that are all alike (that is, belong to one class)*" (Kotu & Deshpande, 2019). So, the model is trained on some historical data and based on similarities it is divided into regions, thus when the new records comes it goes to the region where the features are similar. The functioning and division of classes according to SVM is presented visually below:

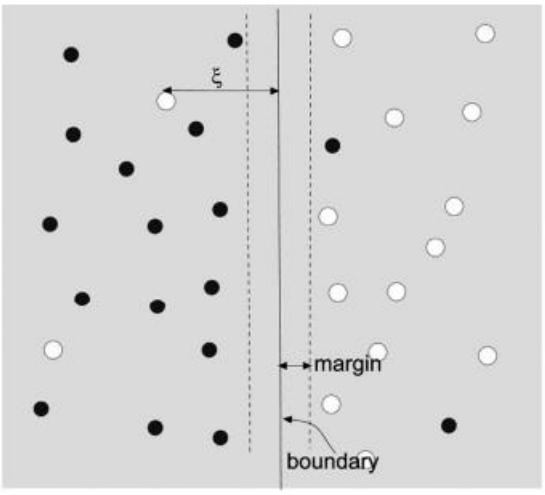

Figure 2.5 Division of class regions according to SVM. Source: *(Kotu & Deshpande, 2019)*

According to the figure, it can be seen that the classes are divided into two regions. The separation is done in such a way, that the similar calss records are in one region, but also allows records of the other class to be part of the wrong region. This is allowed since it is impossible to have 100% correct separation, and this algorithm makes the separations in such way, that the error is as low and comprehensive as possible.

SVM requires large computing power and belongs to the group of advanced algorithms. The advantages of this algorithm are that it is very flexible and resistant to overfitting.

2.3.5 Naïve Bayes

This is a probabilistic algorithm that is based on the statistical "Bayes" theorem. It has the name naïve, because it works on the basis or assumption that each variable is independent of each other.

According to (Kotu & Deshpande, 2019), the main principle of this algorithm is that it uses the probability theorem, where it calculates the probability of each variable as independent. So, it measures the probability of each input and then according to the highest probability it predicts the output value.

This algorithm is used for classification problems. In the real world it is not used much, since the variables are usually related to each other, while this algorithm treats them as independent.

2.3.6 Ensemble Learning

Recently, more advanced ML methods have emerged for classification problems. While a single algorithm can be affected perform poorly in certain situations, ensemble learners typically show improved performance. "*Ensemble techniques are the methods that use multiple learning algorithms or models to produce one optimal predictive model. The model produced has better performance than the base learners taken alone*" (Khandelwal, 2021).

Therefore, ensemble learners involve training the dataset multiple times either with the same algorithm or different algorithms, and then making a final prediction based on a majority vote. For instance, if the model is trained 100 times and 60 times the outcome is "1" and 40 times the outcome is "0", then the final prediction will be "1". A visual representation of the functioning of an ensemble learner is provided below:

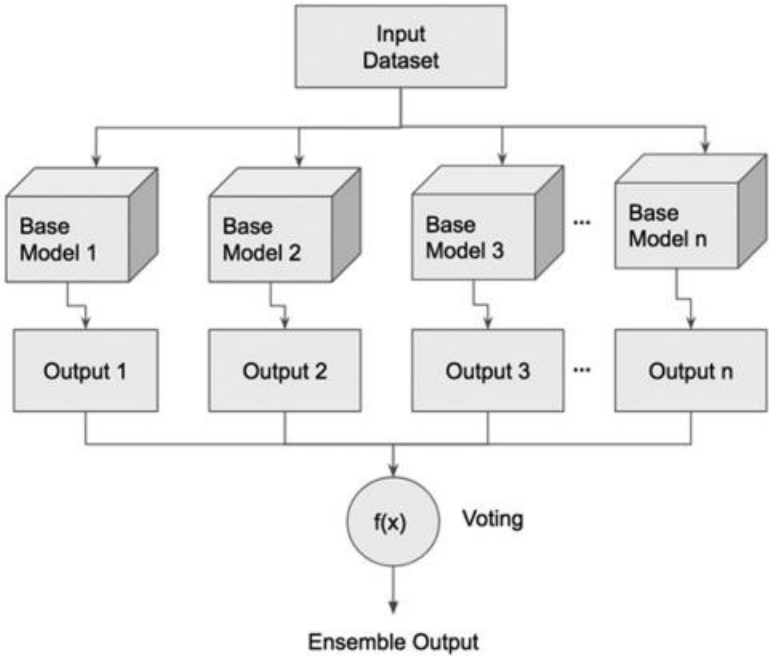

Figure 2.6 Ensemble Learning Model. Source: (Kotu & Deshpande, 2019)

There are 3 types of ensemble learning techniques:

1. **Bagging:** in the Bagging technique, multiple subsets of the main dataset are created and then each of these subsets is trained using a single algorithm. The final prediction is determined by taking the majority vote from the predictions made by all the trained models. An example of such a technique is the Random Forest algorithm, which is based on the fundamental Decision Tree algorithm.
2. **Boosting:** it's similar to the bagging method, but with a different approach to creating the sample datasets. The training sets are generated one by one, with each subsequent set aiming to improve the errors of the previous model. The first set of data is trained, then a new set is created and trained, and this process continues. The training set evolves for each model, focusing on the records that have the lowest accuracy and worst performance, by retraining them.
3. **Stacking:** this method involves using a single training dataset that is trained with various algorithms, such as Decision Tree, SVM, Neural Network, etc. The final predictive model is selected based on the results of the votes from all the different algorithms applied to the same training dataset.

2.4 Model Evaluation

To determine the effectiveness and performance of the predictive model, it must be evaluated using various metrics or estimators. The evaluation is carried out using the following metrics, whose explanations are based on two sources: (Zhu, et al., 2019) and (Kotu & Deshpande, 2019) .

Performance Matrix:

- ***True Positives (TP):*** if the positive class (e.g., class "0", which indicates that the loan has not defaulted) is predicted correctly.
- ***True Negatives (TN):*** if the negative class is predicted correctly.
- ***False Positives (FP):*** if the positive class is not predicted correctly.
- ***False Negatives (FN):*** if the negative class is not predicted correctly.

For example, in this case the positive class is "0" (the loan has been paid and has not defaulted) and the negative class is "1" (the loan has not been paid and has defaulted).

Precision (P): shows how accurately the positive cases are predicted in relation to the total predicted positive class. P = TP / (TP + FP).

Recall (R): shows how correctly the positive cases are predicted in relation to the total number of real positive class. R = TP / (TP + FN)

Accuracy (A): shows how accurately both classes are predicted compared to all records. It serves as a good measure when the dataset is balanced. A = TP + TN / (TP + FP + TN + FN)

F1-Score: is the harmonic mean of Precision and Recall. In cases where the dataset is imbalanced, it serves as the best measure. F1 Score = 2 ((P x R) / (P + R)).

ROC curve: is a consolidated metric that shows True Positive Rate and False Positive Rate which serve to show the performance of a model in a visual format.

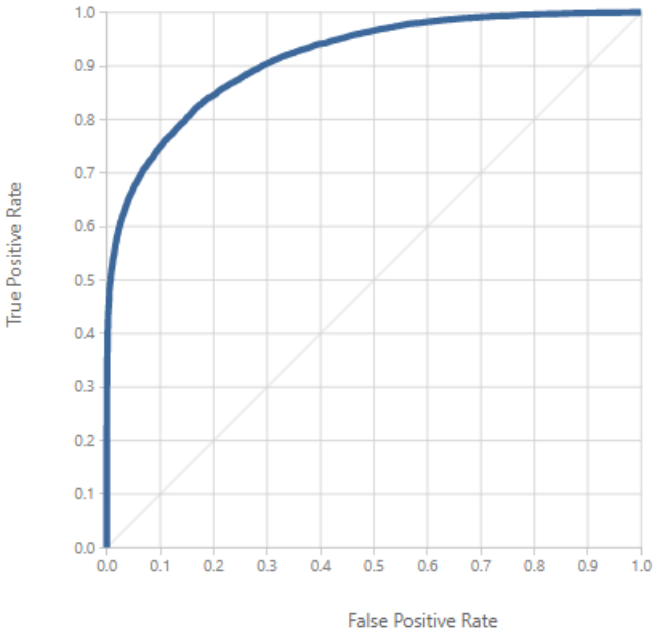

Figure 2.7 ROC Curve

Worth mentioning, that all the estimators explained above can take values from 0 to 1. The performances are higher when they are closer to 1.

3 LITERATURE REVIEW

The issue of loan defaults has recently gained significant attention in both financial institutions and the scientific community. With the growth of data and the advancement of machine learning, numerous authors have studied the topic by constructing various models and evaluating their performance.

3.1 Related Works

- In the paper of (Tariq, et al., 2019), it is provided a comprehensive examination of the ML process and the creation of a predictive model for loan defaults, using the SEMMA (Sample, Explore, Modify, Model and Assess) methodology. In this paper, the data is obtained from the well-known platform, Kaggle. The first stage is sampling, or in other words the selection of the data from the initial dataset, with which the model will be trained. It is worth mentioning that this stage is optional. The second phase is exploration, where the data which the work will be done is understood and explored for patterns. The dataset contains about 1K records and 13 attributes in this paper. The type of attributes was explored, where 4 of them are continuous numerical and the others are categorical. In this phase, there were made statistical analysis, various visualizations such as histogram, scatter plot, box plot, distribution of variables in tabular and visual format such as bar chart, etc. The 3rd phase involves data pre-processing, where the data is processed and adapted for modeling. In this paper, the transformation of the data type was done at first. All categorical data has been converted to numeric ones, in order that some algorithms could perform better. Then, the missing values were treated by replacing them with the mean method. In the modeling phase, 3 different algorithms were used: Decision Tree, Logistic Regression and Neural Network. The last stage is the performance evaluation. This phase was implemented in parallel with the modeling phase, where different operators were used, to see which ones show the best performance. The ROC curve, Accuracy Sensitivity, Specificity and Error were taken as performance evaluation measures. According to the Accuracy metric, the best performing algorithm was Neural Network with 83%, leaving

behind Logistic Regression with 81% and Decision Tree with 80%. However, in this paper, the algorithm with the best performance was chosen Logistic Regression, since it was the best in predicting "True Negatives" or negative values (how accurate is the model for predicting loans that default), which is the most important issue.

- The paper (Alomari, 2017), also explains the methodology and phases involved in predicting loan defaults. The data used in the study was obtained from "Lending Club", a publicly accessible online lending platform. This dataset contains information about customer characteristics and loan repayment history. To begin, the dataset was firstly explored to understand what data would be used and what the dependent variable was. It is worth mentioning, that the dependent variable in this paper and other papers reviewed was the variable indicating whether the loan had defaulted or not, which typically has two values (0 and 1 representing whether the loan has defaulted or not). Afterwards, descriptive statistics analysis was performed on the dataset, analyzing the central tendencies and data distribution of the main numerical variables through various visualizations such as histograms, bar charts, and scatter plots. Next, the data was pre-processed to make it suitable for the chosen ML algorithms. In this study, the dataset of 188K records was reduced to 40K records and from 52 variables to 20 variables through various methods such as removing records and variables with a large number of missing data, replacing missing data, and removing irrelevant variables. Additionally, the "Filter Attribute Selection" method was used to eliminate variables that had little or no effect on the dependent variable, allowing the models to run faster without sacrificing performance. After the data was prepared, ML algorithms were applied and the model was trained. Several algorithms were used in this study, including Naive Bayes, Decision Tree, kNN, 1R, Neural Network, and ensemble models like Random Forest. Upon evaluating the performance of the models, Random Forest was found to be the most successful with a performance of 71.75% as measured by Accuracy.

- Another paper, (Zhu, et al., 2019), deals with the same problem and uses the same dataset from the "Lending Club" platform, where data was taken for the first quarter of 2019. In this paper, an initial exploration of the dataset is also done to understand what data will be

worked with. Unlike the (Alomari, 2017) study, this paper doesn't give much attention to the exploration process. After this phase, the data preparation begins, where the original dataset with 102 variables is reduced to 15. The main method used for selecting the most important variables is a similar operator to "Filter Attribute Selection" called "Recursive Feature Elimination". This operator works based on the logic of removing all variables that have a low correlation with the dependent variable. In addition to this operator, many other data preparation methods were used, such as replacing missing values with the mean, transforming the values of the dependent variable into only 2 values (0 and 1) because there were more value types, etc. After the data is processed, different ML models were applied, such as Decision Tree, SVM, Logistic Regression, and Random Forest as an ensemble model. What sets this study apart, is that the SMOTE method was used together with the Random Forest algorithm. This method addresses the issue of data imbalance. In the initial dataset, the dependent variable was distributed with 99% successful loans and only 1% that have defaulted. In such cases, models usually don't give the desired accuracy due to the data imbalance. After training each model, the highest accuracy was achieved by the Random Forest algorithm along with the SMOTE method with an "Accuracy" of 98% and an F1 Score of 98%.

- The issue of data imbalance regarding the dependent variable was addressed in (Zhou & Wang, 2012), which focuses mainly on "Random Forest" algorithm. The data for this study was obtained from the Kaggle platform, a customer dataset with loan defaults. The dataset contains around 150K records and 12 attributes. There is no explenation of how the data processing/transformation was carried out. The algorithm used in this work is "Random Forest," where the authors develop an improved Random Forest algorithm (Algorithm 2) in addition to the normal Random Forest (Algorithm 1). This algorithm is an ensemble model that works by creating many Decision Trees and distributing the data among these trees based on the sample, and the prediction/training is made based on the majority vote. Meanwhile, the improved algorithm developed by the authors provides that the votes collected from the Decision Trees are weighted based on their importance. Furthermore, the SMOTE and Samp-size methods are used to address the data imbalance issue. The SMOTE method increases the number of records for smaller data by creating new records

based on existing data. On the other hand, the Samp-size method is based on the stratified sampling that favors smaller data during the model training. Then, these models were trained and the performance was evaluated for both Random Forest algorithms, as well as for the SVM, KNN, and Decision Tree algorithms. The results show that the proposed/ developed Random Forest has the highest accuracy of 93.66% compared to the normal Random Forest with 93.58%. To assess the complete model performance, it is also necessary to evaluate the minor class, i.e., True Negatives, to see how well these cases have performed. The Balanced Accuracy was used as a measure of this aspect. For addressing this issue, the SMOTE and Samp-size methods were used, and they were very successful in significantly increasing the Balanced Accuracy. There was no significant difference in terms of which method was more accurate, as they were very similar.

- Also, in the study of (Iain & Mues, 2012), the issue of loan default is addressed by taking into consideration that the dependent variable is imbalanced. Five different datasets, collected from various state institutions, were analyzed and modeled in this research. Initially, all the algorithms that are used are explained. Then, the datasets are explored and the data is described. Next, the authors test how different algorithms perform on these datasets by changing the data balance of the dependent variable. Initially, the dataset was divided into 70/30 balance between 2 values up to a balance of 99/1 of the target variable values. After applying different models, it was observed that the performance of algorithms varies depending on the balance of the dependent variable. Performance in this work is measured through the AUC metric. According to the results, it is observed that in cases where the dependent variable was imbalanced, Random Forest and Gradient Boosting performed the best. Meanwhile, when the dependent variable was balanced at a 70/30 level, the SVM algorithm performed the best. It should be noted that 10 different algorithms were analyzed and tested during this study.

- The paper of (Victor & Raheem, 2021), is based on the "Genetic Algorithm" method, which selects the most important variables impacting the dependent variable during the building of the ML model. The problem of loan defaults is also addressed, based on a dataset obtained from the Kaggle platform on the Lending Club company dataset. The original

dataset consists of approximately 2.2 million records and 151 variables resulting from 2007 to 2018. In this study, only the data from the years 2015 and 2016, totaling 855K records and the same number of variables, were used as a sample. The methodology used in this study starts with data sampling, continuing with exploration and processing, followed by division into training and test data, and finally, the building and evaluating of the model. Initially, the data is explored to gain a better understanding. The exploration in this study focuses mainly on the description of each variable. After this step, the data is processed to be ready for ML model training. Data processing of this study includes: removing variables with more than 30% missing data, replacing missing data with the mean for numerical data and mode for categorical data, transforming categorical data into numerical data, removing outliers, and normalizing the data. The "Genetic Algorithm" method is used as the main method for selecting input variables in the model. This method is based on Darwin's theory of evolution, which after many calculations selects the variables that have the most impact on the dependent/target variable. Logistic Regression, Random Forest, and SVM algorithms were used for predictive modeling. Accuracy, AUC, MCC, Precision, Recall, and F1 score were used as model performance metrics. In most of these metrics, the most successful algorithm was Random Forest, achieving 82% F1 score, leaving behind Logistic Regression at 72% and SVM at 71%.

- In another study (Bayraci & Susuz, 2019)**,** the issue of loan default prediction was addressed. The authors used two datasets, collected from commercial banks in Turkey, one with 80K records and the other with approximately 500K. Both datasets had a significant number of variables, 64 and 60 respectively. The authors briefly described the data preparation process, without going into much detail. Then, they explained the training of various algorithms. An interesting aspect of the paper is that the datasets were divided into two categories, one showing the client's loan application data and the other showing the loan performance data. In the dataset containing loan application data, the "Deep Neural Network" algorithm achieved the highest accuracy at 85.7%, surpassing the "Decision Tree" at 82.3%, "Logistic Regression" at 78%, "SVM" at 77.9% and "Naive Bayes" at 75.2%. On the other hand, in the dataset with loan performance data, the SVM algorithm showed the best results with 78.1%, followed by the Deep Neural Network at 77.9%.

- The issue of loan default prediction has not only been explored in scientific papers, but also in master's level papers. One such paper is (Adelabu, 2021). This study has taken the data from the "UCI Machine Learning Repository" website, which includes around 1,000 records with 20 variables. An exploration of the data was carried out initially, where each variable was analyzed through visualizations and compared with the dependent variable "Default". The paper does not provide much information about the data processing phase, only stating that the data was normalized prior to applying the algorithms. The author used three algorithms for building the model: Random Forest, Logistic Regression, and Decision Tree. The paper also outlines the performance metrics like Accuracy, Precision, Recall, and F1 score in advance. Based on the author, the F1 score is the best metric for evaluating the models, as it takes into account the accuracy across all predicted classes. After the models were trained, the results were compared, and Random Forest resulted as the best algorithm with an F1 score of 87%, followed by Logistic Regression with 85% and Decision Tree with 82%.

- Data pre-processing/transformation is one of the main challenges in building an ML model as it directly impacts the training and performance of the model. One aspect of data processing is handling with missing values. This issue has been thoroughly addressed in the master's thesis by (Martin, 2015), which focuses on missing values in city data (Open City Data). The data was gathered from various web pages of different cities by extracting it and uploading it into a database. During this process, data processing was done by removing duplicates, transforming it into tabular form, cleaning/removing records and variables (when a record or variable had more than 90% of missing values). After these steps, there were about 683K records remaining, with 46% of them being with missing values. To handle the missing values problem, the author used kNN, Linear Regression, and Decision Tree algorithms. These algorithms were used in two stages. The first stage was "Building Complete Subsets" or creating smaller datasets based on predicting missing values through variables that have the most correlation with the dependent variable, which was the missing value in this case. The second approach, Principal Component Regression, predicted missing values by reducing the size of variables and keeping the most important

ones that impact missing values. After applying these algorithms, it was determined that the first approach was more accurate but performed poorly when there were many variables with missing data. On the other hand, the second approach was less accurate compared to the first but was more suitable for forecasting when there were numerous missing variables.

3.2 Conclusion

During the literature review, it was determined that the methodology for predicting loan defaults consists of exploring, processing, modeling, and evaluating data. This approach is commonly seen in most works and will be further discussed in the Methodology chapter.

It was also noted that when the data is imbalanced, using the SMOTE method can well tackle the issue. Additionally, "Filter Attribute Selection" methods are crucial in both data processing and modeling, particularly when there is a high number of variables.

Moreover, the best performing algorithms for predicting loan defaults were also determined. Table 3.1 presents the algorithms with the highest performance as evaluated by the authors of the papers reviewed, who primarily used Accuracy and F1-Score metrics.

Table 3.1 ML Top performing algorithms according to referenced papers

Related Work	Best ML Algorithm
Tariq, et al., 2019	Neural Network and Logistic Regression
Alomari, 2017	Random Forest
Zhu, et al., 2019	Random Forest
Iain & Mues, 2012	SVM and Random Forest
Victor & Raheem, 2021	Random Forest
Bayraci & Susuz, 2019	Neural Network and SVM
Adelabu, 2021	Random Forest

4 PROBLEM STATEMENT

Loan defaults can result in significant financial losses for banks. The costs associated with loan defaults include financial loss, increased provisioning costs, liquidity risk, staff management costs etc. The importance of managing loan defaults cannot be overstated and must be properly managed and mitigated.

During the loan approval process, evaluating the client is a crucial step to minimize the risk of loan default or non-payment. Traditional methods of evaluating clients involve manual analysis such as calculating income and expenses, checking credit history, and assessing collateral status. However, these methods are time-consuming and prone to errors due to the reliance on human labor.

Recently, many financial institutions have started to automate the loan evaluation process by using machine learning. This new approach is more accurate, takes less time, and its use is rapidly increasing. From the problem's perspective, this type of modeling falls under classification problems in machine learning. According to (Kotu & Deshpande, 2019), in the classification or class prediction problem, the goal is to classify or predict records into two or more classes based on the information gathered from the predictors or independent variables. The main objective of this approach is to create an accurate and suitable model for predicting loan defaults while taking into account a certain amount of historic data.

In this study, the records will be predicted based on two classes. Class “0” indicates that the borrower has made timely loan repayments, while class “1” signifies that the borrower has defaulted on the loan. The prediction of loan defaults will be modeled based on the other variables part of the dataset, also known as the independent variables.

The issue of predicting loan defaults has been the subject of research and studies by various authors. According to (Alomari, 2017), the main problems of modeling loan defaults are data pre-processing and handling of the large number of variables. In his study, numerous data processing

techniques are explained, as well as the "Filter Feature Selection" method, which selects the most important variables, thus reducing the number of variables used.

On the other hand, as per (Zhu, et al., 2019), one of the main problems of predicting loan defaults is the data imbalance on the target variable. As a consequence of this problem, the accuracy of the algorithms is not at satisfactory levels. The main method that tackles this problem is SMOTE operator, which is also used in this study.

4.1 Research Purpose

The objective of this paper is to develop predictive models using ML in order to identify future loan defaults based on past loan data. The models will be developed by training different algorithms on the historical loan data. The performance of the models will be assessed through various data pre-processing techniques and multiple ML algorithms.

4.2 Research Objectives

The objectives of this work are stated below:

- Understanding of ML modeling and different algorithms that provide solutions to the problem of this paper.
- Data selection and exploration.
- Data Pre-Processing techniques
- Predictive modeling and performance comparison throu different algorithms, including ensemble models.
- Achieving the highest accuracy of the ML model.

4.3 Research Questions

The research questions of this paper are listed below:

- How to create a predictive ML model on loan defaults?

- How to handle the data imbalance problem?
- How to achieve the highest model performance/ accuracy?
- Which algorithm works best for predicting loan defaults?

4.4 Research Hypotheses

This study in the context of our problem will test the following hypotheses:

1. H_0: The performance of the predictive model for loan defaults is impacted by the SMOTE method.
2. H_0: There is no significant difference in the performance of the predictive model for loan defaults between individual and ensemble algorithms.

4.5 Research Structure

The rest of the paper is structured as follows:

- Chapter 2, *Fundamentals*, outlines all the concepts and techniques used in the study, covering topics like data processing methods, machine learning algorithms, and performance evaluation.
- Chapter 3, *Literature Review*, reviews previous works related to the problem.
- Chapter 4, *Methodology*, explains the data source, the methodology for data processing and modeling, and the tools used in this paper.
- Chapter 5 and 6, *Analysis and Results*, are the key part of the paper, where the whole work and ML modeling is explained, starting from the data exploration to the comparison of the achieved results.
- Chapter 7, *Conclusions and Recommendations*, summarizes the work and provides final conclusions, and includes suggestions for future research.

5 METHODOLOGY

5.1 Modeling Methodology

The ML modeling methodology used in this paper is CRISP-DM (Cross Industry Standard Process for Data Mining). "*The CRISP-DM process is the most widely adopted framework for developing data science solutions*" (Kotu & Deshpande, 2019). There are also other methodologies used for ML modeling such as "DMAIC" or "SEMMA", which have been discussed in the literature review, for this study, the CRISP-DM methodology has been chosen to be used. A visual representation of this methodology is presented below:

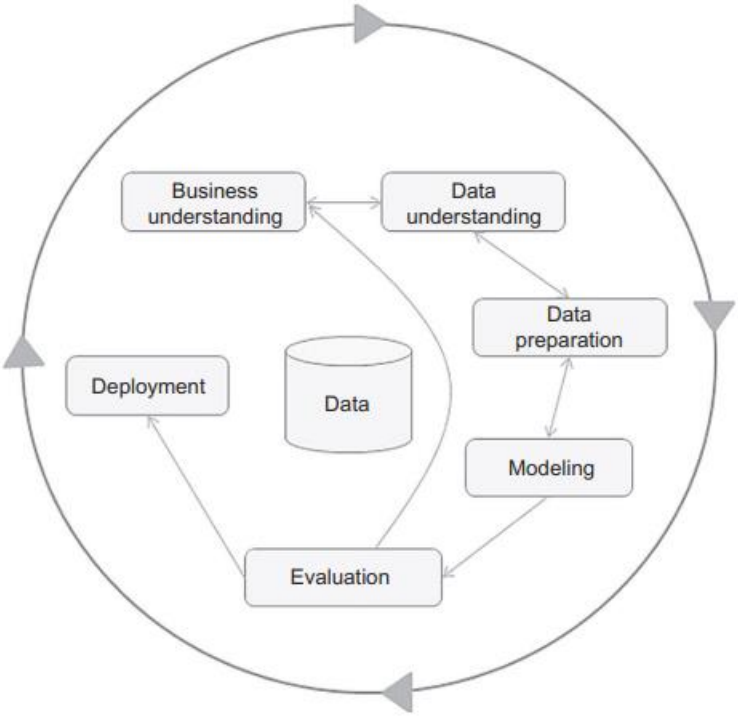

Figure 5.1 CRISP DM methodology. Source: (Kotu & Deshpande, 2019)

The following steps should be followed according to the CRISP-DM methodology:

1. Business understanding: In other words, understanding the problem.
2. Data understanding: Understanding each variable and exploring them.
3. Data preparation: Preparing/ Pre-Processing the data for modeling.

4. Modeling: Creating the model, this phase can be returned to phase 3 repeatedly until the desired results are achieved.
5. Evaluation: is the problem solved, is the model performing satisfactorily.
6. Deployment: Its use on the market.

5.2 Data Source

The data used for this study are secondary and were obtained from the well-known Kaggle platform (Buji, 2021), which is widely used worldwide to provide different datasets for research and exercises. The data used in this paper are also attached in the appendices section. Data from the Kaggle platform have been used in other studies, including (Victor & Raheem, 2021), (Zhou & Wang, 2012) and (Tariq, et al., 2019), which are part of the literature review.

The dataset used in this study contains 87,500 records and 30 variables. Among the variables, 18 are numerical and 12 are categorical. The dependent/ target variable in modeling is "Default", which is binary with values 0 and 1. A value of 0 represents a client who has paid off their loan, while a value of 1 indicates a loan default. The Kaggle platform, from which the dataset was obtained, provides detailed descriptions of all variables.

5.3 Data Understanding and Preparation

To gain a deeper understanding of the dataset, the data will be explored first. This involves reading and understanding the definitions of each variable and its respective values, and assessing the data quality by identifying the missing values. Variables will be analyzed using statistical methods. Data visualization will be used to analyze the distributions, correlations, and other variable characteristics. In order to optimize the model's performance, the data will be pre-processed. This involves transforming, cleaning, normalizing the data and other preparation methods, which is a critical step that directly impacts the performance of the ML model.

5.4 Algorithms Used for Modeling

In this paper multiple classification algorithms will be used, trained and compared to determine which one performs the best for predicting loan defaults. Based on the literature review and the works cited in this paper, the most commonly used algorithms for this problem are presented below:

Table 5.1 Count of algorithms used from referenced papers

Algorithm	Count of Usage
Decision Tree	7
Random Forest - Ensemble	6
SVM	4
Logistic Regression	4
KNN	3
Neural Network	3
Naïve Bayes	2

Based on the most commonly used algorithms found in the literature review, the following classification algorithms have been chosen to be used in this paper: **Decision Tree, Logistic Regression, Support Vector Machines (SVM), Neural Network and Naïve Bayes**. In addition, for ensemble models, **Decision Tree (Boosting), Random Forest (Bagging)** and a combination **(Stacking)** of the aforementioned algorithms **(Logistic Regression, SVM and Neural Network)** will be used.

5.5 Tools Used

The following tools were utilized in this work, along with their corresponding purposes:

- **Excel:** used for data processing and creation of diverse tables.
- **Tableau:** utilized for data visualization via various types of graphs.
- **AzureML:** used for data processing, training the predictive models, and evaluating their performance.

- **Python:** utilized for data visualization and analysis, creating and evaluating predictive models. Random Forest and Decision Tree algorithms were modeled in Python since they are not supported by AzureML.

6 ANALYSIS

6.1 Data Exploration

To develop an effective model, it is crucial to have a thorough understanding of the data that will be used. This can be achieved by exploring the data, which is the initial phase according to the CRISP-DM methodology.

As previously mentioned, the data used in this study was obtained from the Kaggle platform (Buji, 2021) and includes information on customers who have either paid or failed to pay their loans.

The dataset consists of 87,500 records and 30 attributes, with the "Default" variable being the focus of prediction in this study. The "Default" variable can take two values, 0 and 1, where 0 indicates that the client has paid the loan and 1 indicates that the client has failed to pay the loan. A table with more information on each variable is presented below:

Table 6.1 Data Dictionary

Variable	Data Type	Description
ID	Numerical	Unique ID
Asst_Reg:	Numerical	Value of all the assets registered under the borrower's name
GG Grade	Categorical	Credit Scoring Rate
Experience	Numerical	Total year of work experience of the borrower
Validation	Categorical	Validation status of the borrower (of application)
Yearly Income	Numerical	Total yearly income of the borrower
Home Status	Categorical	Borrower living status
Unpaid 2 years	Numerical	No. of times the Borrower has defaulted in last two years
Already Defaulted	Numerical	Number of other loans the borrower was default
Designation	Categorical	Designation of Borrower
Debt to Income	Numerical	Debt to Income ratio
Postal Code	Numerical	Postal code of borrower
Lend Amount	Numerical	Total funded amount to borrower
Deprecatory Records	Numerical	An entry that may be considered negative by lenders because it indicates risk and hurts your ability to qualify for credit or other services
Interest Charged	Numerical	Interest charged on total amount
Usage Rate	Numerical	Processing Charges on the Loan Amount
Inquiries	Numerical	Inquiries in Last 6 Months

Present Balance	Numerical	Current balance in the borrower account
Gross Collection	Numerical	The gross amount payable by way of Settlement or judgment in respect of the Claims, excluding any costs
Sub GG Grade	Categorical	Sub Credit Scoring Rate
File Status	Categorical	Status of the loan file
State	Categorical	State to which borrower lives
Account Open	Numerical	Total number of open accounts in the name of Borrower
Total Unpaid CL	Numerical	Unpaid dues on all the other loans
Duration	Numerical	Loan Maturity
Unpaid Amount	Numerical	Unpaid balance on the credit card
Reason	Categorical	Loan Reason
Claim Type	Categorical	Borrowers Claim Type (Individual or Joint)
Due Fee	Numerical	Charges incurred if the payment on loan amount is delayed
Default	Categorical	Target Variable. Has the loan been paid

To better understand the variables and data of this dataset, we will visualize them. " *Vision is one of the most powerful senses in the human body. As such, it is intimately connected with cognitive thinking* " (Few, 2006).

6.1.1 Target Variable

The variable/attribute "Default" represents whether or not the customer has paid or defaulted on loan. This variable is the target or dependent variable that will be predicted in the model. The following graph (pie chart) provides a visual representation of the distribution of data for this variable.

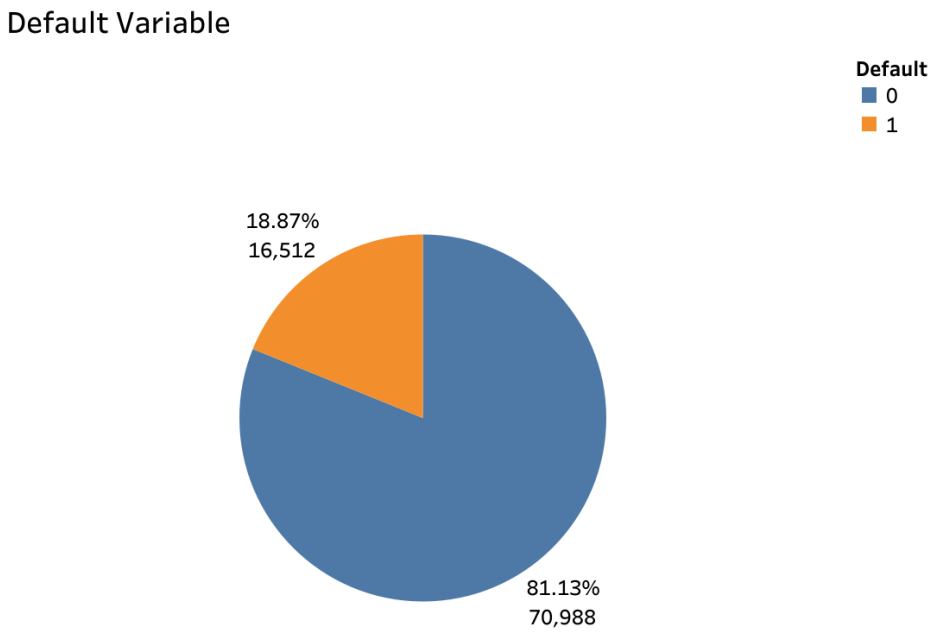

Figure 6.1 Pie Chart. Target Variable – Default

Based on the chart, it is evident that out of the total 87,000 records, 70,988 (81%) are classified as 0 and 16,512 (19%) are classified as 1. This indicates that the majority of customers has paid the loan and did not default. This distribution is known as imbalanced, which could cause modeling issues and low performance levels for the imbalanced value. In the upcoming phases, this issue will be tackled.

6.1.2 Correlation Matrix

To determine the variables that have a greater correlation with the dependent variable 'Default', a correlation matrix is used to measure the strength of the relationship between variables. This matrix only applies to numerical variables and illustrates how much one variable is correlated to another. The following figure represents the correlation matrix for all numerical variables:

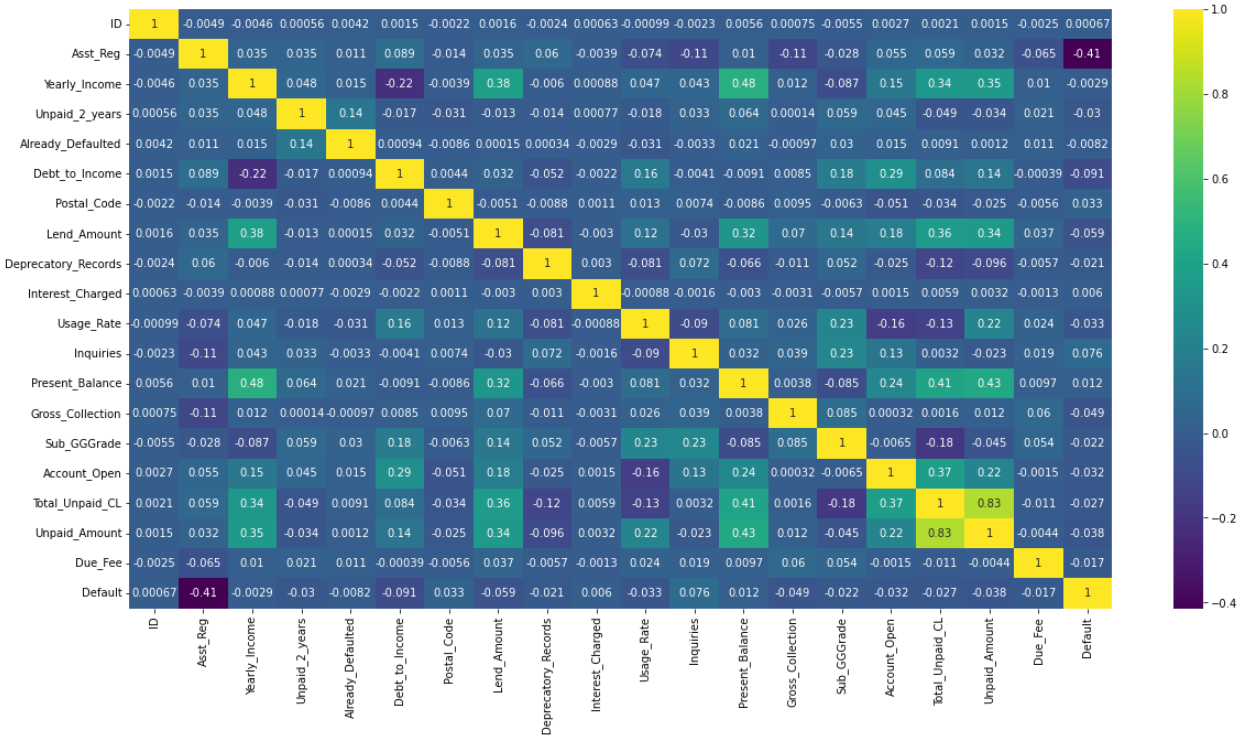

Figure 6.2 Correlation Matrix

In order to interpret the correlation matrix, we should start by focusing on the Default variable, which is located at the bottom right of the matrix. Based on the numbers (ranging from 0 to 1) and colors, we can observe the degree of correlation between each variable with the Default variable. In this matrix, we can see that the variable Asst_Reg has the highest correlation of 41% with the Default variable, followed by the variable Debt_to_Income with a correlation of 9%, Lend_Amount with approximately 6%, and so on.

6.1.3 Asst Reg Variable

The variable "Asst_Reg", which means Assets Registered, represents the borrower's assets or property under his name. Asst_Reg is a continuous numeric variable that has the largest impact of all continuous numeric variables in the dataset on Default. Descriptive statistics and visualizations of its data distribution are presented below:

Table 6.2 Descriptive Statistics of Asst Reg

Descriptive Statistics	Value
Mean	3,798,914
Median	4,132,011
Min	24,847
Max	7,351,847
Standard Deviation	2,289,038
Unique Values	83,966
Missing Values	-

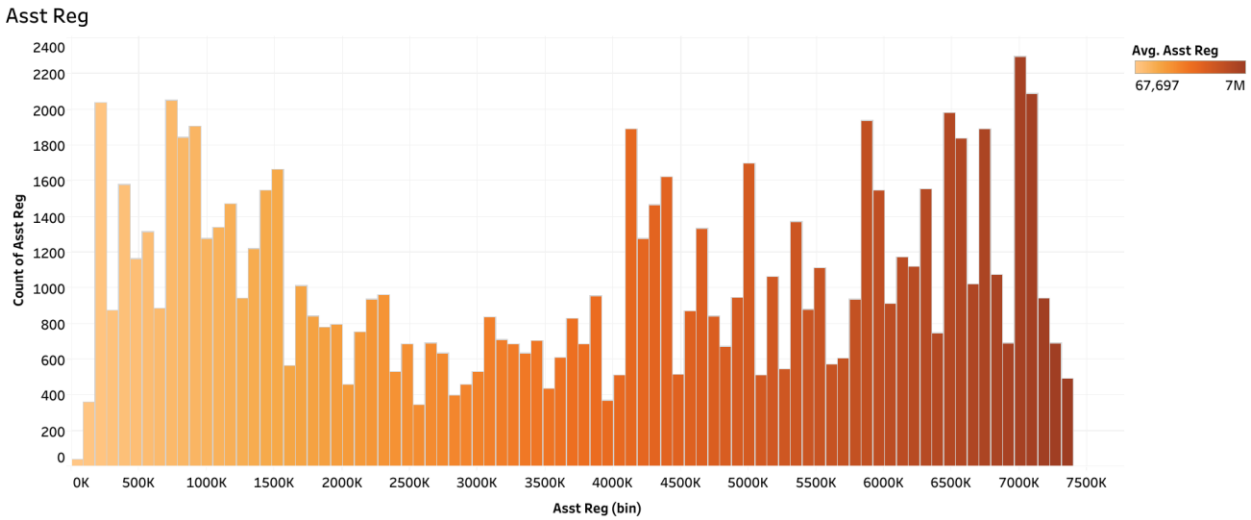

Figure 6.3 Histogram. Asst_Reg Data Distribution

According the visualization above, the data is symmetrically distributed from 24,000 to 7 million USD asset value.

To see how Asst_Reg affects the dependent variable Default, another visualization was produced. The figure 6.4 shows that, as the value of Asst_Reg increases, the proportion of customers who default on their loans decreases. In other words, there is a negative correlation between Asst_Reg and Default. These observations suggest that Asst_Reg is an important predictor of loan default and should be considered in the model building process.

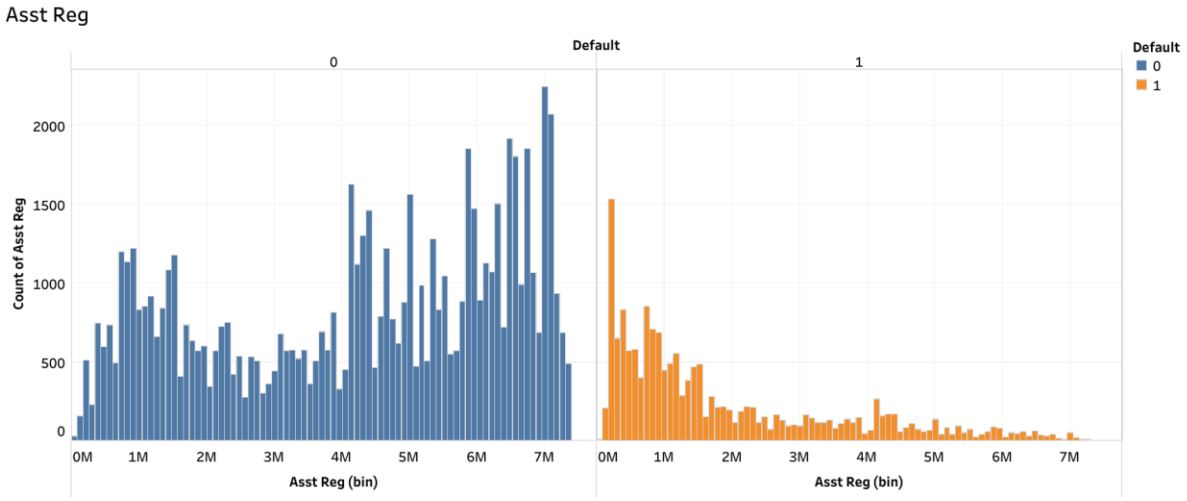

Figure 6.4 Histogram. Data distribution of the Asst Reg variable according to the Default variable

6.1.4 Lend Amount Variable

The variable "Lend Amount" refers to the amount of money that the customer has borrowed. It is a continuous numeric variable and has a significant correlation with the Default variable amongst other variables. Descriptive statistics and visual distributions of this variable are displayed below:

Table 6.3 Descriptive Statistics of Lend Amount

Descriptive Statistics	Value
Mean	25,921
Median	23,940
Min	1,710
Max	59,850
Standard Deviation	14,434
Unique Values	1,298
Missing Values	-

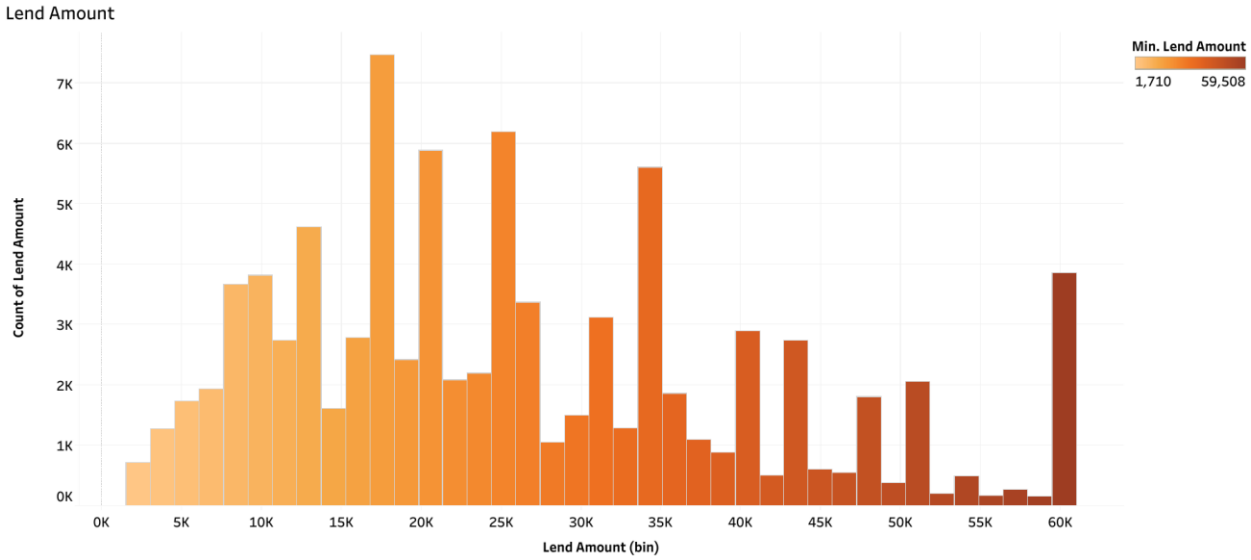

Figure 6.5 Histogram. Lend Amount Data Distribution

The data appears to be symmetrically distributed with an average of around 25K, as can be seen. Another way to demonstrate this is by illustrating the effect of Lend Amount on the Default variable, as shown in the figure below:

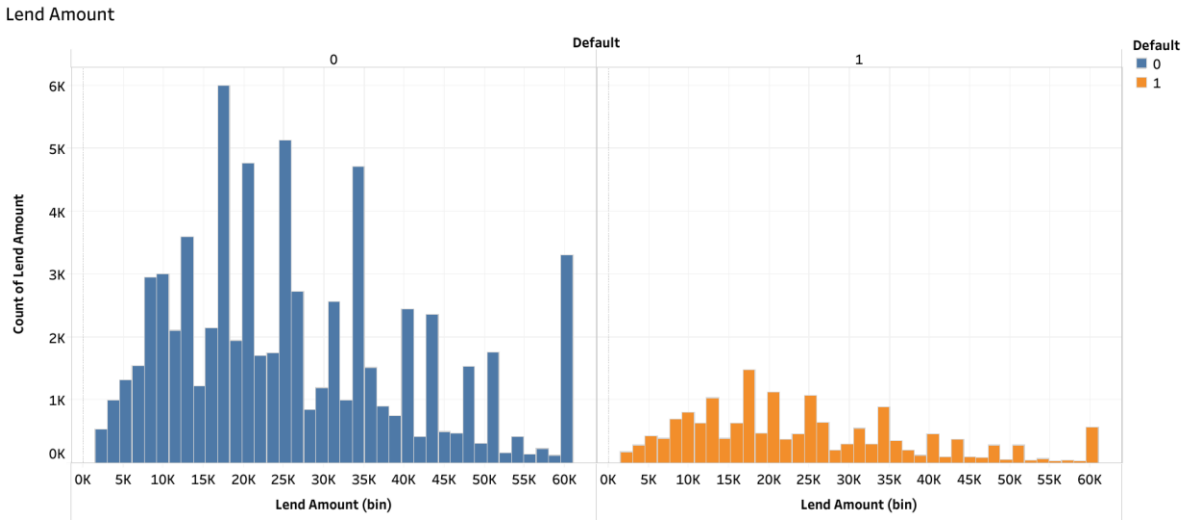

Figure 6.6 Histogram. Data distribution of the Lend Amount variable according to the Default variable

According to this visualization, no significant impact of the Lend Amount variable on the Default variable can be observed.

6.1.5 Reason Variable

This variable represents the purpose for which the loan was granted to the customer. It is a nominal categorical variable with 13 different categories indicating the reason for borrowing. In this dataset it is explored the number of loans issued based on loan purpose. The following visualization provides more information on this exploration:

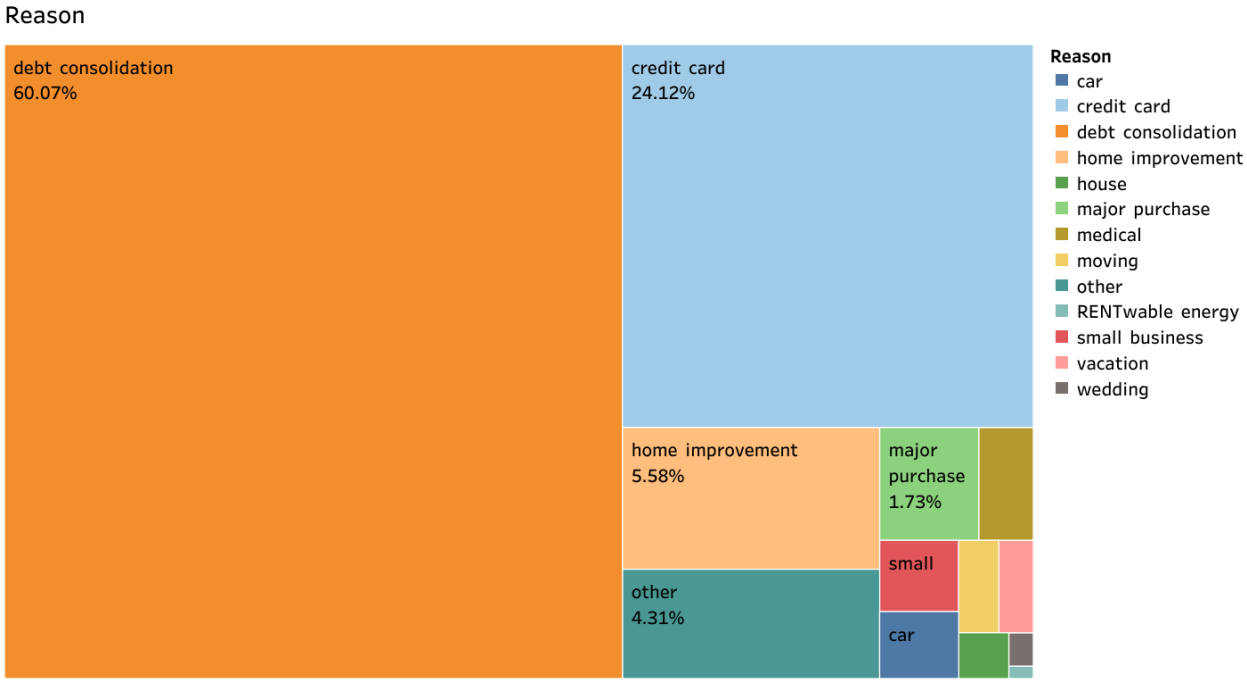

Figure 6.7 Tree maps. Analysis of the Number of Loans Issued Based on the Reason Variable

Based on the presented visualization, it can be observed that the most frequent cause for borrowing a loan was debt consolidation, accounting for approximately 60% of the loans. Credit card was the second most common reason, with a proportion of 24%, and so on.

6.1.6 GG Grade Variable

The variable GG Grade, also known as credit score, is assigned to borrowers based on their creditworthiness. It ranges from 1 to 7, with each level representing a different level of creditworthiness. Therefore, this variable is an ordinal categorical variable that indicates the level of evaluation. The following chart presents the distribution of this variable and its impact on the Default variable.

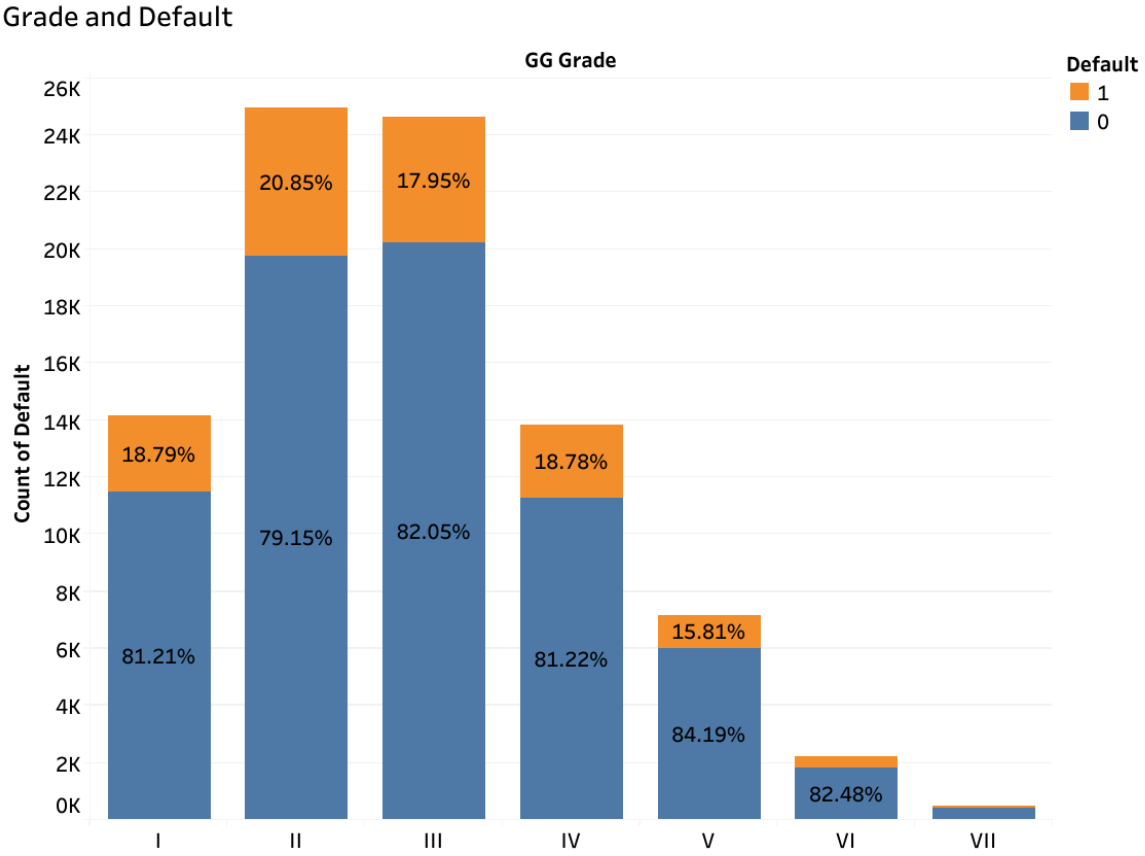

Figure 6.8 Stacked Bar. Data distribution of the GG Grade variable according to the Default variable

Based on the graph, it can be observed that the majority of customers have a grade of 2 or 3. However, there is no significant impact of the GG Grade variable on the Default variable, as the distribution remains consistent across all grades.

6.1.7 Home Status Variable

The Home Status variable denotes the borrower's type of residence, with the values Mortgage, Own, and Rent indicating whether the borrower lives in a mortgaged property, owns the property in their name, or lives on rent. This variable is a nominal categorical variable and the following graph provides further information on this variable and its effect on the Default variable.

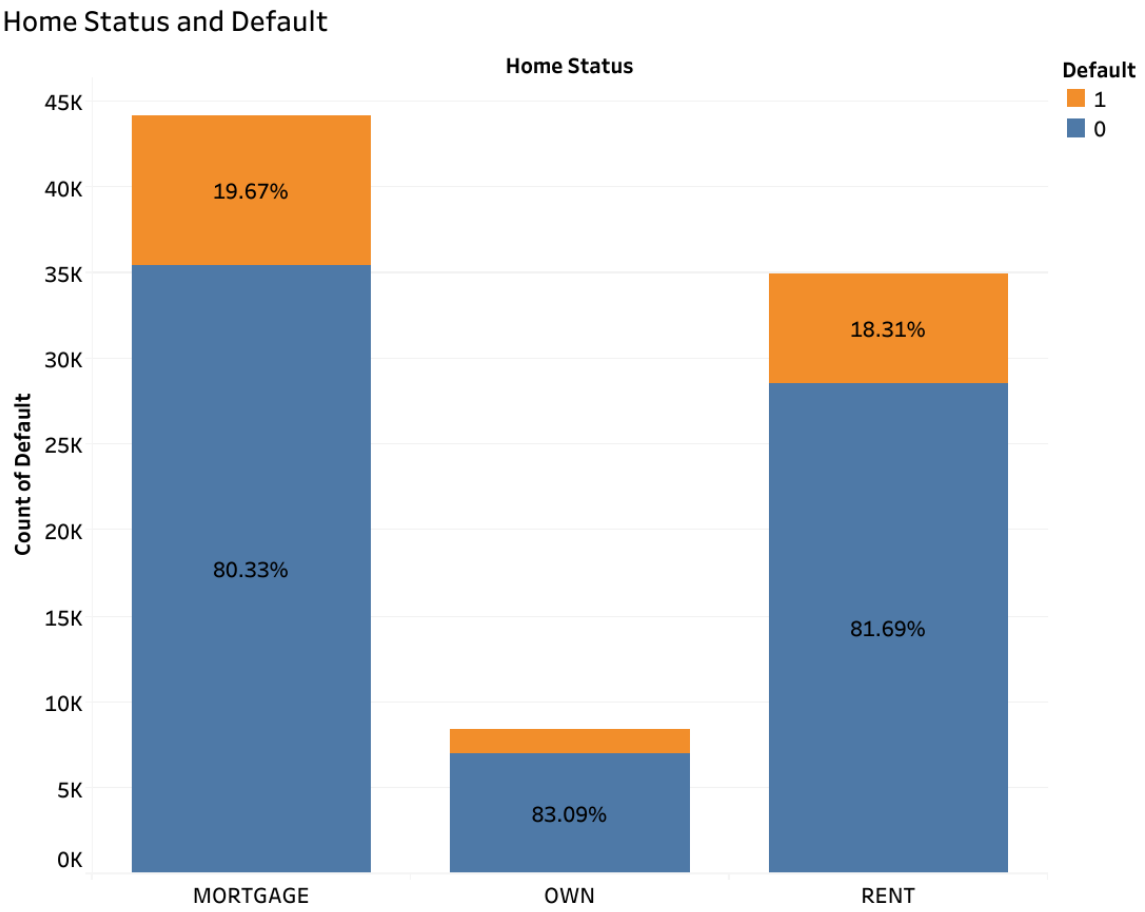

Figure 6.9 Stacked Bars. Data distribution of the Home Status variable according to the Default variable

Based on the graph presented above, it can be observed that the majority of borrowers either have their residence under mortgage or live on rent. However, regarding the loan defaults, it is distributed almost equally among each category.

6.2 Data Preparation

After exploring and understanding the data in the second stage of the CRISP DM methodology, the third stage involves data preparation/ pre-processing. This stage involves various techniques to prepare the data for modeling and ensure its adaptability to different algorithms.
Processing techniques include replacing the missing data, transforming variable types, removing unnecessary variables, normalizing data, selecting the most relevant variables for modeling and splitting the dataset. In this paper, most of these techniques were applied to improve the model's success rate and will be explained separately.

6.2.1 Data Transformation

In order to ensure compatibility with certain algorithms, all string variables were converted to numeric in the data pre-processing phase. For example, categorical variables such as Home Status, which had three values (Mortgage, Own, and Rent), were transformed into numerical values (1, 2, and 3). Several variables underwent this transformation, including GG Grade, Experience, Validation, Home Status, File Status, Duration, Reason, and Claim Type.

Additionally, the Postal Code and Unpaid Amount variables required a separate transformation as they were originally numeric but contained N/A values and were reported as strings. Therefore, the N/A values were removed, and the variables were automatically converted to numeric values.

6.2.2 Missing Values

As previously mentioned, some variables on the dataset have had missing data. Specifically, the Yearly Income, Debt to Income, Total Unpaid CL, and Unpaid Amount variables have had missing values. These variables are of a continuous numerical type, and on average, the missing data was present in about 5% on these variables. To address this, the missing values were replaced with the corresponding mean, since they were numeric type.

6.2.3 Data Normalization

To improve the performance of certain algorithms during the training phase, the data normalization technique was used, specifically on continuous numeric variables. The following variables were normalized during modeling: Asst Reg, Debt to Income, Land Amount, and Total Unpaid CL.

6.2.4 Feature Selection

Two variables were initially removed from the model due to their adverse effects on model training. The Designation variable was removed because it had too many unique values (around 40K), while the State variable was removed because it was the same as the Postal Code variable.

To make the model as efficient as possible, various techniques were used to select the most important input variables. The "Filter Based Feature Selection" technique was used to select the 15 most important input variables, which had the greatest impact on the output variable or the predicted value. There are different methods for selecting the most important variables, but the Pearson Correlation method was used in this paper.

The technique of "Filter Based Feature Selection" was also used in the study by (Alomari, 2017). Moreover, in the paper by (Zhu, et al., 2019), the technique of selecting the most important input variables was used with the Pearson Correlation method.

After the execution of this operator in the study, 15 variables (excluding ID) were chosen as participants in the modeling, and solely through them, the "Default" variable will be predicted. The variables that were selected are displayed below:

Table 6.4 Input Variables in the Model

Variable	Data Type	Description
ID	Numerical	Unique ID
Asst_Reg:	Numerical	Value of all the assets registered under the borrower's name
GG Grade	Categorical	Credit Scoring Rate
Validation	Categorical	Validation status of the borrower (of application)
Home Status	Categorical	Borrower living status
Unpaid 2 years	Numerical	No. of times the Borrower has defaulted in last two years
Debt to Income	Numerical	Debt to Income ratio
Lend Amount	Numerical	Total funded amount to borrower
Deprecatory Records	Numerical	An entry that may be considered negative by lenders because it indicates risk and hurts your ability to qualify for credit or other services
Gross Collection	Numerical	The gross amount payable by way of Settlement or judgment in respect of the Claims, excluding any costs
Sub GG Grade	Categorical	Sub Credit Scoring Rate
File Status	Categorical	Status of the loan file
Account Open	Numerical	Total number of open accounts in the name of Borrower
Duration	Numerical	Loan Maturity
Unpaid Amount	Numerical	Unpaid balance on the credit card
Reason	Categorical	Loan Reason

6.2.5 SMOTE Operator

To tackle the issue of dataset imbalance, specifically in the target variable, it was used the SMOTE method. The dataset had an imbalanced distribution between 0 and 1, with 81% and 19% respectively. As discussed in the literature review and fundamentals chapter this problem is common and can hinder model performance, especially when predicting the value of 1 (loan defaults), which in our case is the minor value. So, the model training focus is to accurately predict the loan defaults, since it is crucial to the problem statement of this study.

Two studies referenced in this study, (Zhou & Wang, 2012) and (Zhu, et al., 2019), also faced this problem while predicting the Default variable and used the SMOTE method to balance the dataset and improve model performance, which has resulted in success.

In the data preparation phase, it was applied the SMOTE method to increase the value of 1 (which was 19%) by 250%. Following this technique, the balance in the dataset became 55% for the value 0 and 45% for the value 1, with the number of records increasing from 87.5K to 128.8K for the default value. This led to a significant improvement in model performance, which we will discuss in the following chapter.

6.2.6 Dataset Split

As part of the data pre-processing, it is necessary to divide the data into two groups: the first group contains data that will be used to train the model, while the second group will be used to test and evaluate the model. These two groups are commonly referred to as train/test data. In this study, the data has been divided into 75% for training and 25% for testing. Although there is no specific guideline on the exact ratio for data division in the literature, it is generally recommended to have a larger training group than testing group.

7 RESULTS

The final stages in the ML methodology involve modeling and evaluating the model's performance. As previously mentioned, this paper uses various algorithms such as Logistic Regression, Decision Tree, SVM, Neural Network, and ensemble models of Bagging, Boosting, and Stacking, which were trained and modeled by using AzureML and Python tools, and the whole work is shown in the appendices section.

To measure the predictive model's performance, the F1-Score was chosen as the primary performance measure. This estimator was chosen due to the dataset's imbalanced nature, with a minority of value 1 indicating loan defaults. Thus, predicting correctly the values when loan defaults is more important than predicting those which do not default. Moreover, F1-Score is more commonly used in such cases as it represents the average performance of each class.

The model's performance is evaluated in two parts. The first part assesses the model's performance based on data pre-processing methods, including the SMOTE method, to observe the data preperation effect on the model's performance. This phase is modeled only by the Logistic Regression algorithm. The second part evaluates the performance of each algorithm seperately after the data is prepared.
The performance evaluation based on data pre-processing methods by using the Logistic Regression algorithm is presented below:

Table 7.1 Performance evaluation of the ML model according to data pre-processing methods

Logistic Regression				
Data Pre-Processing	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Raw Dataset	83.2%	59.0%	28.2%	38.1%
Data Transforming (Categorical to Numeric)	83.4%	61.1%	27.4%	37.8%
Replacing the Missing Values	83.5%	60.8%	28.2%	38.5%
Selecting the Columns - Filter Based Feature Selection	83.6%	62.5%	26.1%	36.8%
Data Normalization	83.6%	62.7%	26.2%	37.0%
SMOTE	76.9%	74.3%	75.0%	74.6%

The table presented above illustrates that the SMOTE method has had the most significant impact on the model's performance. Specifically, the F1-Score improved by almost two times as a result of this method. However, the other phases did not have a considerable effect on the model's performance, although they are still crucial for other algorithms performance and the overall ML process.

The remaining part of the performance evaluation focuses on assessing each algorithm's performance separately on the prepared dataset. The following table shows the performance of each algorithm used for modeling:

Table 7.2 Performance evaluation of ML models according to different algorithms

Individual	ML Algorithm	Accuracy	Precision	Recall	F1-Score
	Logistic Regression	76.9%	74.3%	75.0%	74.6%
	Decision Tree	75.3%	75.2%	75.5%	75.2%
	SVM	75.9%	73.0%	74.5%	73.7%
	Neural Network	80.3%	95.5%	59.2%	73.1%
	Naïve Bayes	46.5%	67.5%	51.5%	34.6%
Ensemble	**ML Algorithm**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
	Bagging - Random Forest	83.1%	83.1%	82.6%	82.8%
	Boosting - Decision Tree	84.9%	85.5%	80.2%	82.8%
	Stacking - LR, SVM, NN	73.7%	76.3%	71.8%	74.0%

The results presented above indicate that the ensemble algorithms outperformed the individual ones. In terms of individual algorithms, Neural Network achieved the best performance according to the Accuracy estimator, while Decision Tree performed the best according to the F1-Score evaluator.

Moreover, the Boosting method using Decision Tree performed best among the ensemble algorithms according to the Accuracy estimator, whereas both Boosting and Bagging (Random Forest) performed similarly according to the F1-Score evaluator.

In terms of the hypotheses presented in the paper, the first hypothesis (H_0) cannot be rejected as the SMOTE method was found to have a significant impact on the model's performance. Meanwhile, the second hypothesis (H_0) was rejected, and H_A was supported, as the ensemble models outperformed the individual ones.

8 CONCLUSIONS AND RECOMMENDATIONS

To manage and reduce credit risk in banks, it is crucial to increase the accuracy of loan assessment by predicting loan defaults through ML modeling. This paper aimed to explore various ML techniques and different algorithms for predicting loan defaults and evaluating their performance. The results indicate that the SMOTE method significantly impacts model performance and ensemble algorithms outperform individual ones, with Boosted Decision Tree (Boosting) and Random Forest (Bagging) being the best performers.

The study has important implications for banks, highlighting the potential of ML in credit risk assessment. However, it's important to note that model performance depends on the type of data and its quality.

It is recommended that further studies with more complex datasets can be conducted. Moreover, studies using other methods, such as reducing data to the majority variable (under-sampling), should be conducted.

Overall, this study highlights the value and importance of data and its modeling through ML in credit risk management and loan default prediction.

APPENDICES

- **Appendix 1:** The data used in this paper: https://www.kaggle.com/datasets/marcbuji/loan-default-prediction

- **Appendix 2:** Data visualization in Tableau: https://public.tableau.com/views/TemaeDiplomes1/GradeandDefault?:language=en-US&:display_count=n&:origin=viz_share_link

- **Appendix 3:** Model development for Loan Default Prediction in Python (Random Forest, Decision Tree and Naïve Bayes): https://github.com/platurgashi/loan-default-prediction

- **Appendix 4:** Data preparation and model development for Loan Default Prediction in Azure ML (Logistic Regression, Neural Network, SVM and Boosted Decision Tree): https://gallery.cortanaintelligence.com/Experiment/Master-Thesis-Loan-Default-Prediction-3

- **Appendix 5:** Data preparation and model development for Loan Default Prediction in Azure ML (Stacking method by using LR, SVM, NN): https://gallery.cortanaintelligence.com/Experiment/Master-Thesis-Loan-Prediction-with-Ensemble-Learning-3

- **Appendix 5.1:** This appendix provides additional information on the construction of the model using ensemble learning, specifically the Stacking method. Its implementation is more complex, and therefore requires further explanation.

  The dataset is processed in the same way as explained in the paper's data processing section. The model is initially trained with each algorithm separately (SVM, Logistic Regression, Neural Network). Based on the performance evaluation, Logistic Regression algorithm showed the highest performance, and thus was given a weight of 0.4, while the other two algorithms were given 0.3 each (total 1).

  Afterwards, the results from the "Score Model" operator are processed and re-named according to each algorithm. The variables from the “Score Model” were selected through the "Select Column" operator. The variables that were selected are: ID, Default, Label (prediction), and Prob (probability of that prediction) for each model.

  These variables are then joined using the Join operator, which joins the data based on the ID variable as a reference. Afterwards, through SQL, these data are calculated using the weighting mentioned earlier and evaluated. The final result is determined using the principle of votes, where logically, if two algorithms predicted "1" and one predicted "0", the final result was 1.
  Finally, the script operator from Python is used to provide the final performance results.

REFERENCES

Adelabu, B., 2021. *Credit Risk: Assessing Defaultability through Machine Learning Algorithms.* Senegal: African Institute for Mathematical Science Senegal.

Alomari, Z., 2017. Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications. *New Zeland Journal of Computer-Human Interaction,* p. 22.

Bandyopadhyay, A., 2016. *Managing Portfolio Credit Risk in Banks.* 1st ed. Delhi: Cambridge University Press.

Bayraci, S. & Susuz, O., 2019. A Deep Neural Network (DNN) based classification model in application to loan default prediction. pp. 76-84.

Buji, M., 2021. *Kaggle.* [Online]
Available at: https://www.kaggle.com/datasets/marcbuji/loan-default-prediction
[Accessed 27 November 2021].

Central Bank of Kosovo, 2021. *Annual Report 2021,* Prishtinë: Central Bank of Kosovo.

Chen, J., 2022. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/d/default2.asp
[Accessed 11 December 2022].

Few, S., 2006. *Information Dashboard Design: The Effective Visual Communication of Data.* 1st ed. Sebastopol: O'Reilly Media, Inc..

Iain, B. & Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Elsevier,* pp. 3447-3453.

Kelleher, J. & Tierney, B., 2018. *Data Science.* 1st ed. Cambridge: The MIT Press Essential Knowledge Series.

Khandelwal, Y., 2021. *Analytics Vidhya.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/
[Accessed 19 December 2022].

Korstanje, J., 2021. *Towards Data Science.* [Online]
Available at: https://towardsdatascience.com/smote-fdce2f605729#:~:text=SMOTE%20is%20a%20machine%20learning,with%20this%20type%20of%20data.
[Accessed 07 January 2023].

Kotu, V. & Deshpande, B., 2019. *Data Science Concepts and Practice.* 2nd ed. Cambridge: Morgan Kaufmann.

Martin, C., 2015. *Predicting Missing Values in the Context of an Open City Data Pipeline.* Vienna: Vienna University of Economics and Business.

Mohita, N., 2022. *Nakuri Learning.* [Online]
Available at: https://www.naukri.com/learning/articles/handling-missing-data-mean-median-mode/
[Accessed 15 January 2023].

Nautiyal, D., 2022. *geeksforgeeks.* [Online]
Available at: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/
[Accessed 16 January 2023].

Tariq, H., Sohail, A., Aslam, U. & Batcha, N., 2019. Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA). *American Scientific Publishers,* Volume 16, pp. 3489-3503.

Victor, L. & Raheem, M., 2021. Loan Default Prediction Using Genetic Algorithm: A Study Within Peer-To-Peer Lending Communities. *International Journal of Innovative Science and Research Technology,* VI(3), pp. 1195-1205.

Worldbank, 2022. *The World Bank.* [Online]
Available at:
https://data.worldbank.org/indicator/FB.AST.NPER.ZS?end=2021&locations=XK&start=2010&view=chart
[Accessed 11 December 2022].

Yildirim, P., 2015. Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. *International Journal of Machine Learning and Computing,* V(4), pp. 258-263.

Zhou, L. & Wang, H., 2012. Loan Default Prediction on Large Imbalanced Data Using Random Forest. *TELKOMNIKA Indonesian Journal of Electrical Engineering,* X(6), pp. 1519-1525.

Zhu, L. et al., 2019. A study on predicting loan default based on the random forest algorithm. *Elsevier,* pp. 504-513.