NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Different Approaches of Machine Learning Models in Credit Risk

A Case Study on Default on Credit Cards

Eduardo Barreto Sulz Gonsalves

Dissertation presented as partial requirement for obtaining the Master's degree of Statistics and Information Management specialized in Risk Management and Analysis

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

Different Approaches of Machine Learning Models in Credit Risk

by

Eduardo Barreto Sulz Gonsalves

Dissertation presented as partial requirement for obtaining the Master's degree of Statistics and Information Management specialized in Risk Management and Analysis

Advisor: Bruno Miguel Pinto Damásio

November 2022

DEDICATION

Dedico este trabalho principalmente a minha avó Ju que sempre me apoiou, especialmente nos meus estudos. Se não fosse por ela, nada disso seria possível. Também gostaria de dedicar essa tese aos meus pais, Nilo e Andrea que sempre me incentivaram a querer mais e não ficar na minha zona de conforto. Agredeço aos meus grandes amigos irmãos Alfredo e Ravi que me fizeram companhia nessa etapa da vida e ainda fazem! E gostaria de agradecer em especial a minha namorada Carolina que me ajudou de forma incomensurável com essa tese me motivando e me ajudando nos momentos em que eu pensei que não iria conseguir.

ACKNOWLEDGEMENTS

During the writing of this thesis I received a great amount of help and assistence and therefore I am very thankful.

I would like to thank in special my thesis supervisor, Professor Bruno Damásio, who accepted to be my supervisor under te most complicated condition possible. His insightful ideas and suggestions allowed this thesis to be concluded with proper high level standards. Also thank him for his readyness and availability to help when was necessary. This thesis was a result of four years of study with lots of mishaps and complications because of the pandemic but fortunatelly, this cicle was completed thanks to the help of Professor Bruno Damásio. Thank you.

ABSTRACT

Credit scoring is a very important process for banks. It allows the credit analysts to calculate the probability of a client defaulting a payment on a specific time horizon. This process helps the bank to manage their assets, preparing themselves ahead of time for possible defaults and also in the decision-making process of conceding or denying a loan to a new client.

There are several different machine learning classifiers that can be used to calculate the probability of default. Studies shown that there is no specific model that can be used as the best one for all circumstances, each model will depend on the dataset.

In this study, six different machine learning models are applied on datasets to classify and predict clients more likely to commit credit default. The models compared in this study were chosen based on the most frequently used techniques in this field and because of the lack of studies comparing these six models in specific, namely Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, k-NN and Naïve Bayes. The goal of this comparison is to identity if there is a model that constantly outperforms the others.

Three datasets are used. The first one is the German Credit Data with socioeconomic information from the clients requesting for a loan. The second one is the Credit Card Default Dataset with historic information about previous payments of credit cards invoice from clients, both datasets are from UCI repositorium. The last dataset is about credit concession with sociodemographic information about the clients obtained from Kaggle.

To compare the models AUC is the main common metric used, followed by confusion matrix. After analysis, the random forest model presents the higher AUC for all datasets, other models vary their position on the ranking depending on the dataset. Finally, decision tree presented a bad AUC since it does not calculate probabilities but had one of the best accuracies of all models for two of the three datasets.

KEYWORDS: Credit Scoring; Logistic Regression; Random Forest; Decision Tree; k-NN; SVM; Naïve Bayes.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Networks
AUC	Area Under the Curve
CDS	Credit Default Swap
DLR	Default Loss Rate
DT	Decision Tree
EAD	Exposure at Default
ECOA	Equal Credit Opportunity Act
GDPR	General Data Protection Regulation
IAS39	International Accounting Standards 39
IFRS9	International Financial Reporting Standard 9
k-NN	k Nearest Neighbors
LR	Logistic Regression
MARS	Multivariate Adaptive Regression Spline
NB	Naive Bayes
OR	Odds Ratio
PC	Principal Component
PCA	Principal Component Analysis
PD	Probability at Default
RF	Random Forest
ROC	Receiver operating characteristic
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
UCI	University of California-Irvine

1. INTRODUCTION

In the past 100 years, some of the largest world recessions took place, as it was the Wall Street Crash in 1929 and the Great Recession in 2008 . Boyle (2021) debates whether the process of growth-recession-recovery is inevitable or not in this model. He points out that cycle ends up occurring but the reason for it is debatable. One of the reasons could be related to the psychological factor of people in it. It also raises the possibility of economic shocks, such as wars and epidemics.

This cyclical tendency of growth-recession-recovery can be seen in the crash of US stock market in 1929 and more recently the 2008 crisis which affected most of the countries. As a response to the 2008 crisis, more restrict regulations were established with IFRS9 and Basel III, having their focus on the new rules for capital adequacy and market liquidity. (Porretta, Letizia & Santoboni, 2020) performed a comparison between the old IAS 39 and the new IRFS. One of the changes in this regulation was the consideration of Expected Credit Losses (ECL) in the reports, something that was not carried out in the previous regulation, IAS 39. In IRFS 9, the calculation of impairment considering the ECL allowed banks to produce financial statements considering losses that already occurred and expected future losses, ensuring them more preparedness for eventual situations.

No matter the cause for the recession, they are likely to happen. Thus, financial institutions must be ready to handle the situation. Whether it is linked to the stock market or any other financial instruments, there are some approaches that can be done to properly handle the risk and the unexpected losses. The new regulations also require financial institutions to have more specific calculations of metrics such as estimate the probability of default for capital requirements (Basel III) or for point-in-time estimates of expected credit losses (IFRS9) (Dumitrescu, Hué & Hurlin, 2021).

International banks and other big financial institutions have been using statistical models to estimate these values but there is a big discussion about what is the best statistical model/method to use to estimate these metrics. This is crucial for the financial institutions not only requirement matters but also to have more knowledge about their risks and be able to manage it properly.

For example, using the definition of Hull (Hull, 2019) the Credit Default Swap can be compared to an insurance contract that pays off if a particular company or country defaults. The company or country is known as the reference entity. The buyer of credit protection pays an insurance premium, known as the CDS spread, to the seller of protection for the life of the contract or until the reference entity defaults. A bank could flag their clients requesting credit into three categories: payer, likely to pay and unlikely to pay based on the probabilities of the statistical models. With these probabilities the bank could decide whether they provide the loan to the client, or not. Also, the bank could offer the loan but get a Credit Default Swap financial instrument.

In this study, several statistical models and machine learning models are applied in three different datasets and asses the performance of the models in each of them. The first one is the South German Credit data set available in the machine learning repository of the University of California – Irvine (UCI). Secondly there is the UCI Credit Card dataset available also on UCI repository with information about clients from a bank in Taiwan and their credit card payments from April 2005 to September 2005. The third and last dataset is available on Kaggle and is composed of synthetic data for educational purpose.

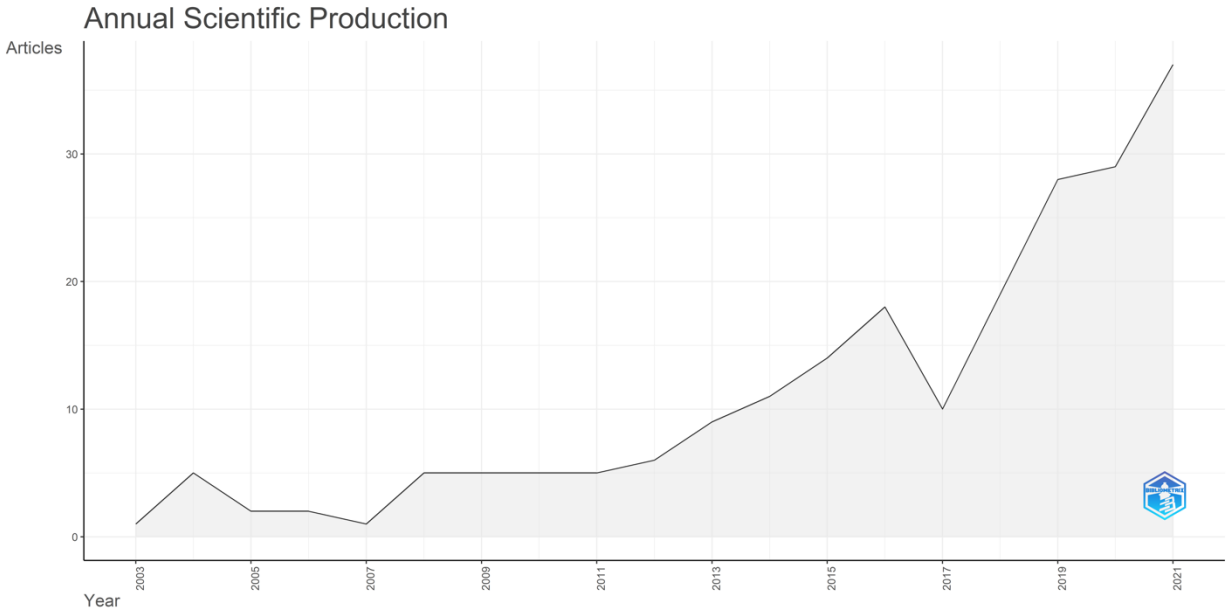
The first dataset used contains 1000 rows of credit concession to anonymous bank clients. It also contains socioeconomic information about those clients and a binary variable that indicates whether they paid all installments on time or if they delayed or default one or more payments (Grömping, 2019). The second dataset has over 29.000 rows and 24 columns with information about client's previous payments of credit card invoices and personal information (Yeh & Lien, 2009; Lichman, 2013). The last dataset has more than 32.000 rows and 12 columns with synthetic data about mortgage applications. All data and outputs were generated using Python and R (Tse, 2020).

The reminder of this article is organized as follows. Section 2 presents the literature review on the topic; Section 3 describes the methodology of the machine learning models used in this study along other required techniques; Section 4 explores the datasets used and their compositions and the procedures that were applied in this study to obtain the results presented here; Section 5 indicates the exploratory analysis held in each one of the datasets and the data preparation required to use the models; Section 6 is where the model fitting is conducted and the results are presented; Section 7 concludes this study and presents some recommendations for future works.

2. LITERATURE REVIEW

In the early 2000s the development of credit scoring studies had developed even further with more advanced techniques and evaluation criteria as the Receiver Operating Characteristic (ROC) and GINI coefficient. Also, the increase computational capabilities allowed more complex models to be applied, which improved the study and development of credit scoring field, as presented in Figure 1.

Figure 1: Number of credit scoring articles published by year



The data from Figure 1 comes from all articles published on Web of Science between 2003 and 2021 with the keyword “credit scoring” and the topics were filtered by the themes related to machine learning and statistical analysis, namely Artificial Intelligence & Machine Learning, Economics, Statistical Models and Management.

The application of credit scoring allows banks and lenders to reduce their risks and provide larger numbers of loans with the same capital requirement. Therefore, they are able to lend money to more clients, resulting in the earning of more interest rates from installments, a scenario that in the past was not possible.

2.1. Subjective versus Credit Scoring Systems

Overall, as cited in (Crook, 1996), there are two techniques that can be used when approving a credit/loan: loan officer’s subjective assessment and credit scoring. The logic behind credit scoring is that the characteristics of a potential client, such as age, work, income, and others, are analyzed and compared with these features from past clients who have received a loan already. If the potential client has the same attributes as a past client that received the loan and defaulted, the new one is flagged as a potential defaulter and the loan is rejected. On the other hand, if their characteristics matches with a past client that did not default, then they are not flagged, and the credit is provided. On the

contrary, in the judgmental system, the loan officer must individually analyze each client's profile and decide whether the loan will be allowed or not.

When choosing mathematical and statistical methods instead of subjective methods, the process of classification is done much faster and the feature importance for separation of good and bad clients is much clear to the loan officers. It is important to note, however, that there are some critics to credit scoring regarding the assumption of the data models for the techniques utilized (Abdou & Pointon, 2011).

2.2. Benefits and Criticism of Credit Scoring

One of the benefits of credit scoring is that it requires less variables to provide a good estimate. Importantly, only the significant statistical variables are considered in the models. By applying the subjective analysis, loan officers have more difficulty to identify which variables are more strongly related to the repayment variable, thus impairing the process of conducting a variable reduction (Crook, 1996).

Other consideration to consider is that credit scoring is a mathematical modelling that will generate the similar results regardless of the analyst. Moreover, the same data can be quickly reanalyzed by statisticians and credit analysts using roughly the same model. The classifications might not be the same if two credit analysts apply a judgmental assessment of the same data, which would also be much more time-consuming (Crook, 1996).

Chandler and Coffman (1979) states that the empirical evaluation of credit worthiness of a client is superior to judgmental analysis due to its ability of forecast results, high capacity of data analysis, and for management reports, which points out the main features of a bad and a good client. On the other hand, they highlight that judgmental analysis is more likely to identify cases that are truly exceptions from experience and probably would be misclassified by the model.

One of the criticisms to credit scoring is that since they are defined by mathematical rules, the results are very static and with a rigid cut-off point to decide whether a client is a good or bad one. The change in the value of a variable, such as the client changing its job or moving to another neighborhood could change the classification from the model (Al Amari, 2002). Another limitation relies on the use of historical data. Unless there is a constant update on the dataset and weights are applied, the estimates of the model would be dependent on outdated information, which results it less accurate classifications.

Another problem raised in the literature is the dichotomous outcome. There are several outputs that can occur. For example, there are different scenarios such as the client can: pay the installment on time; delay a few days; delay bigger for periods, but still pay; or do not pay at all. The problem is that all these outcomes can't be compared. They can be inputted and analyzed, but only two at a time (Heffernan, 2005).

2.3. Key Factors of Credit Scoring

There have also been discussions about key factors that impact in the credit scoring such as number of variables, sample size, dataset split and cut-off point. In some papers, the variables used in the study were as little as 3 to 20 variables (Pendharkar, 2005; Jo *et al*., 1997), while other studies have used

thousands as in Bellotti and Crook, 2009. Regarding sample size, the number of entries can vary from a few dozen (Dutta *et al.*, 1994) to thousands of observations (Hsieh, 2004). On the topic of validation method, Paliwal and Kumar, 2009, summarized a few studies and noticed that the main validation method used was the training-validation split, the second most used was the n-fold cross-validation and a few used different methods as three subsets (training, validation and testing), jack-knifing, and other methods.

The work from West *et al.* also raised the issue of which metric to use as comparison. Moreover, the author discusses that feature selection is a key for the performance of the models and there are several ways to conduct it, considering different metrics (West & Bhattacharya, 2016). Another factor that can impact the metrics from the models is the ratio of response events on the dataset. As presented by Brown *et al.,* different ratios on the target variable in the training can result in different models having better performance with the same dataset (Brown & Mues, 2012). In their work they use a technique called Synthetic Minority Oversampling Technique (SMOTE), developed by Chawla *et al.*, to properly generate datasets with different 0's and 1's ratios, but with same characteristics from the original dataset (Chawla *et al.*, 2002).

Based on the discussion raised on the articles mentioned, there is no clear evidence of a proper way to develop the perfect model. There is no standard rule for sample size, number of variables, validation method and cut-off point (Al Amari, 2002). All these parameters will depend on the model that is being developed and its circumstances. Regarding the technique, there is no specific method/model that consistently provide the best classification, but most of studies identified that Random Forests (RF), Neural Networks (NN), XGBoost (GB) and SVM usually provide better results than other methods and these models should be used as benchmark for other models (Henrique *et al.*, 2019; Gunnarsson *et al*., 2021). Nevertheless, Baesens *et al*., 2003, points out that other simpler models also provide good performance, and in most cases, the difference between these complex models and simpler models are not statistically significative. The Table 1 presents the different methods used in several articles and reaffirm that there is no "one model fits all". Noteworthy, depending on the dataset and the different steps applied in the process, each study has a different better model.

Table 1: A comparison of percentage correctly classified from published articles

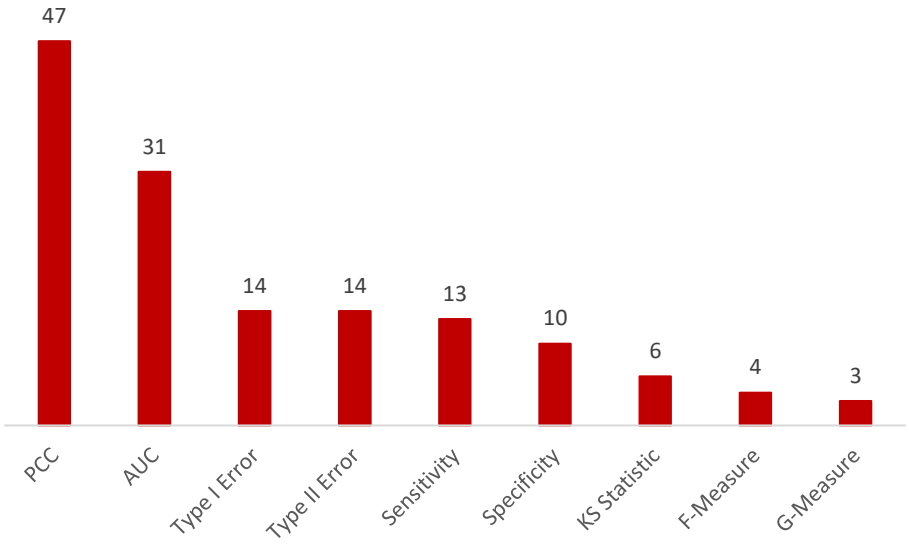
Classifier	Kumari & Mishra (2018)	Bao et al. (2019)	Moscatelli et al. (2020)	Arora & Kaur (2020)	Ashofteh & Bravo (2021)
LR	85,92	93,4	79,6		**67,0**
RF	89,07	93,6	**81,4**	57,6	62,0
LDA			79,6		
Penalized LR			79,6		
XGB		92,7	81,3		
DT	85,92	90,4			
ANN		93,7			
k-NN	87,29	88,9		**61,1**	
SVM	73,56	**94,1**		50,00	50,0
NB	**90,61**			60,1	
MLP	77,46				
RBF	84,03				

Other interesting study carried on this topic evaluates the performance of three different methods, Logistic Regression (LR), Credit Scorecard model and Decision Tree (DT) (Yap, Ong & Husain, 2011). Despite been simpler than Neural Networks (NN), Support Vector Machine (SVM) and other models, Yap *et al.* states that once the high computing capability of data mining software is currently widely available, business institutions are allowed to explore those tools to analyze and gather important information from their large customer database. This was not possible in the recent past, however with the technological development, computers can perform the calculations of new and more complex methods. In fact, some studies have been using and applying more complex models, as random forests and gradient boosting (Wang *et al*., 2012; Chen *et al.*,2019). On the other hand, Yap *et al*. conclude that sophisticated techniques such as Multivariate Adaptive Regression Splines (MARS), NNs, and SVM have not shown significant improvements in classification accuracy (Yap, Ong & Husain, 2011). Furthermore, most of the time regulators prohibit financial institutions of using complex AI models to calculate metrics as probability at default (PD), default loss rate (DLR) and exposure at default (EAD) due to their lack of interpretability. This is the reason why most of practical examples of machine learning models' usage in credit risk are more interpretable models as logistic regression and decision trees (Yan & Lin, 2019).

2.4. Performance Evaluation Method for Credit Scoring

Regarding the performance evaluation criteria, there are several different methods that can be used, such as, confusion matrix, mean-square error (MSE), root-mean-square error (RMSE), mean absolute error (MAE), the ROC curve and other criteria. Figure 2 is based on the results from (Dastile *et al.,* 2020), presents the most frequent methods in several different studies.

Figure 2: Frequently used performance evaluation measures



The most common performance evaluation criteria is PCC (Percentage Correctly Classified) obtained from the confusion matrix. One of the advantages of this evaluation criteria is that it easily allows the analyst to spot the Error Type I (false positive) and Error Type II (false negative). Also, it permits the calculation of other metrics, such as F1-Score and recall. The downside with this method is that it does not take into account the costs of Error Type I and II. It is much less severe for a bank to reject a loan to a client that would not default than to provide a loan to a client that will default (Baesens *et al*.,

2003). In the first case, the cost is not earning the potential interest from the contract. In the second case, the cost is not earning the interest from the contract and also losing the principal, though.

More recently ROC Curve have been used as an alternative to confusion matrix. The ROC Curve is a two-dimensional plot that presents in the Y axis, the proportion of bad cases classified as bad, i.e. sensitivity, and in the X axis the proportion of good cases classified as bad, i.e. the complementary of specificity. In other words, we can say that sensitivity is one minus Error Type II and specificity is one minus Error Type I. The ROC Curve presents the overall performance of the model for all possible cut-off points allowing the analyst to choose the desired cut-off to balance off Erros Type I and II.

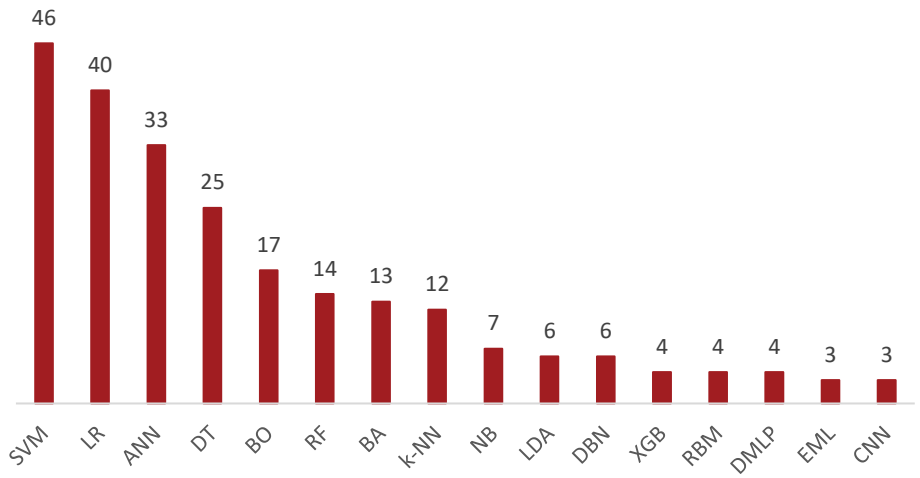
2.5 Study reference and motivations

Credit scoring is a key analysis to banks because allows them to calculate the probability of a default happen given the circumstances This knowledge is also used to calculate ECLs and help the bank to manage their assets and prepare themselves for eventual defaults.

In the past few years, the advancement of computational capabilities allowed the development and implementation of new complex machine learning models, increasing the range of options available to create a model. Considering the possible transformations and manipulations that can be done in a dataset, combined with possible number of models to be applied, there are countless possibilities of developing a model until finding the best one.

In (Dastile *et al.,* 2020), 74 articles between 2010 and 2018 are listed along the machine learning models used in each one of them. The Figure 3 presents the summarization of that table.

Figure 3: Numbers of articles that used each classifier



Legend: LR (Logistic Regression), NB (Naïve Bayes), LDA (Linear Discriminant Analysis), XGB (XGBoost), EML (Extreme Learning Machines), k-NN (k-Nearest Neighbor), SVM (Support Vector Machine), ANN (Artificial Neural Network), BA (Bagging), BO (Boosting), RF (Random Forest), RBM (Restricted Boltzmann Machine), DBN (Deep Belief Network), DMLP (Deep Multi-Layer Perceptron), and CNN (Convolutional Neural Network).

In this study, 6 out of the top 10 most used models between 2010 and 2018 in credit risk studies were chosen to be compared. The goal is to apply them in three distinct datasets with different structures, number of observations and number of variables and identify if there is a model that constantly

outperforms the others. In the table from (Dastile *et al.,* 2020), it is also possible to identify that none of the studies analyzed used and compared this set of classifiers, namely Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, k-NN and Naïve Bayes.

3. METHODOLOGY

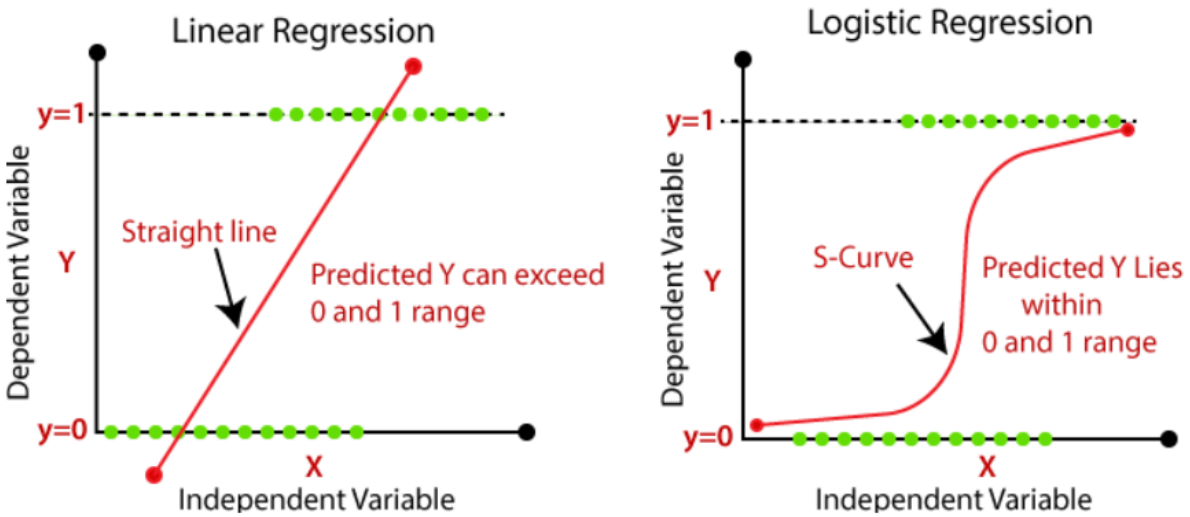
In this chapter we present the different methods used to compare the data sets and the methods to analyze it.

3.1 Logistic Regression

The logistic regression is a special case of the linear regression because in this case, the response variable is categorical and essentially dichotomous (but there are ways for estimation when the number of categories is higher than two). The logistic regression is considered to be a parametric approach because it's function and definition can be parameterized by a finite number of parameters (Wasserman , 2004). Therefore, the technique consists in a model, with n observations that relates a set of p explanatory variables $X_1, X_2, \dots, X_p$ to a response variable Y that assumes only two possible states, usually 0 or 1, that in this study represents a client not defaulting in any moment in the payment of instalments and defaulting the payment. The logistic model also allows the direct estimation of probability of occurrence or not of the event ($Y = 0$ and $Y = 1$, respectively).

The logistic regression is a special case from the linear regression because linear regressions have as output any number while the logistic regression can only have an output between 0 and 1, since it is predicting a probability. To apply this transformation the logit, a link function, is used to create a linear relation with the explanatory variables. The logit transformation, also known as log odds, is what allows the interpretation of parameters in variables in a probability point of view (MacKenzie et al., 2018). The Figure 4 illustrate the different outputs from a linear and logistic regression.

Figure 4: Support of linear regression vs. support of logistic regression



From: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

To maximize the probability of classifying a real defaulter as defaulter and a non-defaulter as a non-defaulter, the logistic regression uses the Maximum Likelihood Estimation the iteratively change the coefficients in order to maximize the logit function. The result of the regression formula, in terms of log-odds can be given by the following function:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (1)$$

Where p is pre probability of a default happen, β are the parameters respective to each X_i variable and k is the number of independent variables. Working the math around to isolate the probability of a default happening, the result is the subsequent function:

$$p = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \quad (2)$$

Since the outcome of a logistic regression is a probability and not a specific value of 0 or 1, a boundary is established to classify that probability. For example, a threshold of 0.5 could be used and probabilities above this values are classified as 1 and below this value are classified as 0.

One of the great advantages of the logistic regression model is that its outputs are very simple to understand and using the odds ratio, they have a direct interpretation. Another interesting property of this method is the capacity to estimate punctual probabilities of each individual.

3.1.2 ROC Curve

The ROC curve is a technique used to assess the quality of the developed model. The ROC curve is the chart of sensitivity (true positives) versus the false positive. These values are calculated from the cross table of a real values versus predicted values.

The sensibility is the proportion of true positives, it reflects the capacity of the model in correctly classify true events when they are actually true. The calculus is done by dividing the true positives by the total of positives. The specificity is the proportion of true negatives, it reflects the ability of the model in predicting the absence of the condition when the condition in fact doesn't exist. Its calculation is done by dividing the actual true negatives by the real total of negatives

The ROC curve has a characteristic line from the position (0,0) to the point (1,1). Curves in the upper part of the chart, above this line have, have a good predictive power while curves bellow this curve are worse than a random models. The ROC curve is measured as the area under the curve and the higher the value, the better is the predictive model.

3.2 Support Vector Machine (SVM)

SVM is a statistical classification method proposed in 1998 (Vapnik, 1998). Considering a set of two variables $\{x_i, y_i\}$, where x_i is the explanatory variable and y_i is the dichotomous variable. Boundaries are calculated from the most extreme points of each y_i group in such way that it is equally distant from those points. Based on this new threshold, the model decides as which y_i group the new x_i input should be classified as. The optimal hyper-plane could be written as follows:

$$\sum_{i=1}^{n} w_i x_i + b = 0 \quad (3)$$

Where W is the normal to hyper-planes and b is a scalar. The maximization of the distance between the extreme points to set the boundaries are usually obtained through the Lagrange multipliers and using linear, polynomial, Gaussian or sigmoidal separations. In other words, the training data is transformed into a higher dimension in which a linear separation is done using a hyper-plane (Henrique *et al.*, 2019)

3.3 Decision Tree (DT)

Classification and Regression Trees is a simple, yet a robust model (Breiman *et al.*, 1984). It consists in the creation of a tree-based structure that classifies the individuals in a series of if-then conditions. Using historical data, the model identifies the variables and its threshold that best splits correctly the response variable. After that split is done the process repeats itself with the next best explanatory variable until an optimal point where the split of the data no longer improves the accuracy.

3.4 Random Forest (RF)

Random Forest is an ensemble method. Ensemble methods are built using base methods and aggregating them into a more robust and complex model. There are two groups of ensemble methods: bagging and boosting. In the bagging method the base models are set in parallel and in the boosting they are sequential. There are two crucial steps in the bagging method, the bootstrap sampling with replacement to assure independence of the different training samples and the aggregation of the base models created. Li and Chen, 2020, provided an interesting explanatory scheme to this process that is presented in the Figure 5.

Figure 5: The algorithm of Bagging

Input: Dataset $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$;
Base learning algorithm $\mathcal{L}$;
Number of base learners m.

Process:
For $j - 1,2, \dots, m$:
$S_j = booststrap(S)$; % Generate a bootstrap sample from S
$h_j = \mathcal{L}(S_j)$ % Train a base learner h_j from the bootstrap sample
end.

Output: $H(x) = mode(h_1(x), \dots, h_m(x))$ % For classification studies

Source: Li & Chen, 2020.

In RF this process happens using DTs as the base model. Several randomly generated trees are created to each training set generated in the first step. After that, when a new observation enters and has to be classified, it is applied to all base trees created and compared the output in each one of them. The output that occurs the most will be the output from the RF model.

3.5 k-Nearest Neighbors (k-NN)

As cited by Baesens *et al.*, 2003, in k-NN method the distance between all points is calculated using statistical methods as Euclidian Distance, more commonly, and Mahalanobis Distance. The Euclidian Distance calculation can be done with the following formula:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left|\left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|\right| = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{\frac{1}{2}} \quad (4)$$

After the results, the response variable of the k nearest neighbors of the new point is taken into account to classify the new point as a good or bad client, in the example of credit scoring. One common step in this process is to assume an odd number for k to avoid ties.

3.6 Naïve Bayes

Another simple method that often performs very well is the Naïve Bayes. This method is based on the Bayes Theorem and Conditional Probability Theory. As presented by Baesens *et al.*, 2003, in this method the probabilities of occurrences in each variable are calculated. Then, using prior probabilities the probability of a certain circumstance happens (explanatory variable) given a specific output from response variable is calculated, i.e., posterior probability.

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})} \quad (5)$$

Following the assumption behind Naïve Bayes classifier that all explanatory variables are independent from each other the following can be done:

$$p(y|\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|y) \quad (6)$$

After this step the probabilities of the y_1 and y_2, the two possible outcomes from the response variable, are normalized and the highest value is assigned as the probability of that event for that observation.

3.7 Multicollinearity

Multicollinearity is defined as when two or more independent variables present a high correlation value between them. As pointed in (Rahayu *et al.*, 2017), the effect of multicollinearity in regression analysis is the estimation of week parameters for the model that do not represent the real nature of that independent variable. In order to identify the existence of multicollinearity in the dataset the variance inflation can be calculated and is denoted by the following formula:

$$VIF = \frac{1}{1 - R^2} \quad (7)$$

Where R^2 represents the unadjusted coefficient of determination for regressing the i^{th} independent variable.

The VIF is calculated for each predictor in the model and the calculation of its square root results in the number of times that the standard deviation from this variable is inflated because of the other independent variables in the model. For instance, if a variable has a VIF of 16, it means that its standard deviation is 4 times larger than if the variable had no correlation with the other variables. In (Daoud, 2017), a table to interpret VIF is presented as the following:

Table 2: VIF value interpretation

VIF value	Interpretation
VIF = 1	Not correlated
1 < VIF ≤ 5	Moderately correlated
VIF > 5	Highly correlated

To handle with multicollinearity, it is possible to apply Principal Component Analysis (PCA) to create new variables uncorrelated using the independent variables. This is method is presented in depth in the next topic.

3.8 Principal Components Analysis (PCA)

Principal Components Analysis is a multivariate technique that is used to reduce the number of dimensions and remove correlation between independent variables. As explicated in (), when the main component is derived from a population of multivariate normal random vector $X = (X_1, X_2, \dots, X_p)$ and vector average $\mu = (\mu_1, \mu_2, \dots, \mu_p)$ and covariance matrix Σ with root characteristic (eigenvalue) that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ obtained a linear combination of the main components is as:

$$Y_1 = e_1'X = e_{11}'X_1 + e_{21}'X_2 + \cdots + e_{p1}'X_p \quad (8)$$

$$Y_2 = e_2'X = e_{12}'X_1 + e_{22}'X_2 + \cdots + e_{p2}'X_p \quad (9)$$

…

$$Y_p = e_p'X = e_{1p}'X_1 + e_{2p}'X_2 + \cdots + e_{pp}'X_p \quad (10)$$

Then $Var(\mathrm{Y}_i) = e_i'\sum e_i$ and $Cov(\mathrm{Y}_i, \mathrm{Y}_k) = e_i'\sum e_i$where $k = 1, 2, \dots, p$.

Requirements to form the main component is a linear combination of variable X in order to have maximum variant is to select a feature vector (eigen vector) that $e = (e_1, e_2, \dots, e_p)$ such that $Var(\mathrm{Y}_i)$ = maximum $e_i'\sum e_i$ and$e_i'e_i = 1$. Therefore, a principal component can be denoted as a linear combination that maximizes $e_i'X\ Var(e_i'X)$, provided $e_i'e_k = 1$ and $Cov(e_i'e_k) = 0$ for $k < 1$.

The final result of the PCA are new variables called principal components independent from each other.

4. DATA & PROCEDURE

4.1 Data

To compare the quality and ability of the models in making good classifications, three different data sets regarding credit approval will be used. The first data is the German Credit Data with data from 1000 customers with 21 variables being most of them categorical variables. The goal of this data set is to classify the credit worthiness of the customers and train a model able to detect potential bad clients in the future. The proportion of clients that paid the debt and the ones who did not is 70% and 30% respectively. The table below presents the variable names, description, and type of data for this dataset.

Table 3: Variable description for German Credit Data

Variable name	Content	Data Type
status	status of the debtor's checking account with the bank	categorical
duration	credit duration in months	quantitative
credit_history	history of compliance with previous or concurrent credit contracts	categorical
purpose	purpose for which the credit is needed	categorical
amount	credit amount in DM	quantitative; result of monotonic transformation
savings	debtor's savings	categorical
employment_duration	duration of debtor's employment with current employer	ordinal; discretized quantitative
installment_rate	credit installments as a percentage of debtor's disposable income	ordinal; discretized quantitative
personal_status_sex	combined information on sex and marital status; categorical; sex cannot be recovered from the variable, because male singles and female non-singles are coded with the same code (2); female widows cannot be easily classified, because the code table does not list them in any of the female categories	categorical
other_debtors	Is there another debtor or a guarantor for the credit?	categorical
present_residence	length of time (in years) the debtor lives in the present residence	ordinal; discretized quantitative
property	the debtor's most valuable property, i.e. the highest possible code is used. Code 2 is used, if codes 3 or 4 are not applicable and there is a car or any other relevant property that does not fall under variable classification.	ordinal
age	age in years	quantitative

other_installment_plans	installment plans from providers other than the credit-giving bank	categorical
housing	type of housing the debtor lives in	categorical
number_credits	number of credits including the current one the debtor has (or had) at this bank	ordinal, discretized quantitative
job	quality of debtor's job	ordinal
people_liable	number of persons who financially depend on the debtor (i.e., are entitled to maintenance)	binary, discretized quantitative
telephone	Is there a telephone landline registered on the debtor's name?	binary
foreign_worker	Is the debtor a foreign worker?	binary
credit_risk	Has the credit contract been complied with (good) or not (bad) ?	binary (response variable)

The second dataset consists of socioeconomic data and the history of payments of credit cards from clients of a bank in Taiwan and the goal is to estimate the probability of default for each client for the next month. This data set is composed of 25 variables being all of them numerical variables and 30000 rows. The proportion of clients that paid the debt and the ones who didn't is 78% and 22% respectively. The table below presents the variable names, description, and type of data for this dataset.

Table 4: Variable description for Credit Card Default dataset

Variable name	Description	Data Type
LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit	quantitative
SEX	Gender	categorical
EDUCATION	Highest Education level from creditor	categorical
MARRIAGE	Marital status	
AGE	Creditor's age (years)	quantitative
PAY_SEP	Count how many months delayed was the bill for September paid	quantitative
PAY_AUG	Count how many months delayed was the bill for August paid	quantitative
PAY_JUL	Count how many months delayed was the bill for July paid	quantitative
PAY_JUN	Count how many months delayed was the bill for June paid	quantitative
PAY_MAY	Count how many months delayed was the bill for May paid	quantitative
PAY_APR	Count how many months delayed was the bill for April paid	quantitative
BILL_AMT_SEP	Amount of bill statement in September (NT dollar)	quantitative
BILL_AMT_AUG	Amount of bill statement in August (NT dollar)	quantitative
BILL_AMT_JUL	Amount of bill statement in July (NT dollar)	quantitative
BILL_AMT_JUN	Amount of bill statement in June (NT dollar)	quantitative
BILL_AMT_MAY	Amount of bill statement in May (NT dollar)	quantitative
BILL_AMT_APR	Amount of bill statement in April (NT dollar)	quantitative
PAY_AMT_SEP	Amount of previous payment September (NT dollar)	quantitative
PAY_AMT_AUG	Amount of previous payment August (NT dollar)	quantitative
PAY_AMT_JUL	Amount of previous payment July (NT dollar)	quantitative
PAY_AMT_JUN	Amount of previous payment June (NT dollar)	quantitative

PAY_AMT_MAY	Amount of previous payment May (NT dollar)	quantitative
PAY_AMT_APR	Amount of previous payment April (NT dollar)	quantitative
DEFAULT	Indicates if a default happened or not	categorical

The third data set is the Kaggle dataset, a didactic dataset with information about mortgage applications of made-up clients and the goal is to estimate correctly what applicants are more likely to repay the debt and who is not. It has over 32.000 rows and 12 variables with 7 numeric and 5 categorical. The proportion of clients that paid the debt and the ones who didn't is 78% and 22% respectively. The table below presents the variable names, description, and type of data for this dataset.

Table 5: Variable description for Kaggle Dataset

Variable name	Description	Data Type
age	Age	quantitative
income	Annual Income	quantitative
house	Home ownership	categorical
emp_dur	Employment length (in years)	quantitative
loan_intent	Loan intent	categorical
loan_grade	Loan grade	categorical
loan_amnt	Loan amount	quantitative
loan_int_rate	Interest rate	quantitative
credit_risk	Loan status (0 is non default 1 is default)	categorical
loan_percent_income	Percent income	quantitative
historical_default	Historical default	categorical
cred_hist_length	Credit history length	quantitative

The data sets were obtained from the public repository of Machine Learning from University of California Irvine (UCI) and from public Kaggle competitions. The different proportions of categorical and numerical variables in each data set along with the sample size and different ratios of 1s and 0s in the target variable provide a wide range of variety to compare how each model works in each situation and what is the best overall method. The table bellow summarizes some descriptive metrics from each dataset listed above.

Table 6: Overview of each dataset

Title	German Credit Data	Credit Card Default	Kaggle Dataset
Source	UCI	UCI	Kaggle
Number of records	**1000**	**30000**	**32581**
Default records	300	6636	7108
Non-default records	700	23364	25473
% of default	30%	22%	22%
Number of features	**20**	**24**	**11**
Numeric features	7	21	7
Categorical features	13	3	4

4.2 Procedure

The goal of this study is to have an in-depth comparison of some statistical methods in different situations and assess what is the best overall model. In order to do so, an exploratory analysis will be conducted to all three data sets. The objective with this analysis is to clean the data set, identify wrong values and filter the data for the modeling part.

As some of the models do not accept missing values, an imputation process was conducted using the median value of that variable where there was no data. Regarding the distribution of data in categorical variables, if a category had too few observations in it, it was aggregated to a similar category and renamed. In some variables numeric variables where the distribution was highly skewed or had values in different scales from other variables from the dataset, it was transformed to avoid the inflation of parameter of the models.

After that, six different methods will be applied to the data for the modeling. The first one will be the logistic regression which is one of the most common methods in this field. Also is a parametric method and easy to interpret the outputs.

The second method to be applied is the random forest, a more complex method but still widely used in the field. It is based on the decision tree method, which is simple to understand, but the random forest itself is not as simple as the decision tree method.

The third model is the decision tree, a rule-based method very simple to apply, yet very good performing. The fourth model was naïve bayes, a method based on the Bayes Theorem of conditional probability. It considers the values on the data on training set as the prior probability and calculates the probabilities of an event happening based on that prior.

Support Vector Machine (SVM) and k-Nearest Neighbors (k-NN) are powerful methods that uses more complex math to classify individuals. The SVM model apply transformations to the data bringing it to higher dimension spaces and find optimal cutoff points to split the data. The k-NN uses the classification of the nearest neighbors to a data point and classifies that data point as the most frequent classification among those neighbors.

After all the methods have been applied to all data sources, metrics and statistics will be used to measure the performance of each model in each data sets in order to understand what model performed the best in the overall.

It is important to emphasize that the method used for the comparison of the models must take in account the same measure for all different models, such as ratio of true positives and true negatives.

5. EXPLORATORY ANALYSIS

The first dataset to be analyzed was the South German Credit Data. A database with 1000 rows, 20 explanatory variables, most of them categorical and the response variable. This dataset is already very polished and doesn't have any missing values.

Figure 6: Credit Risk proportion and frequency

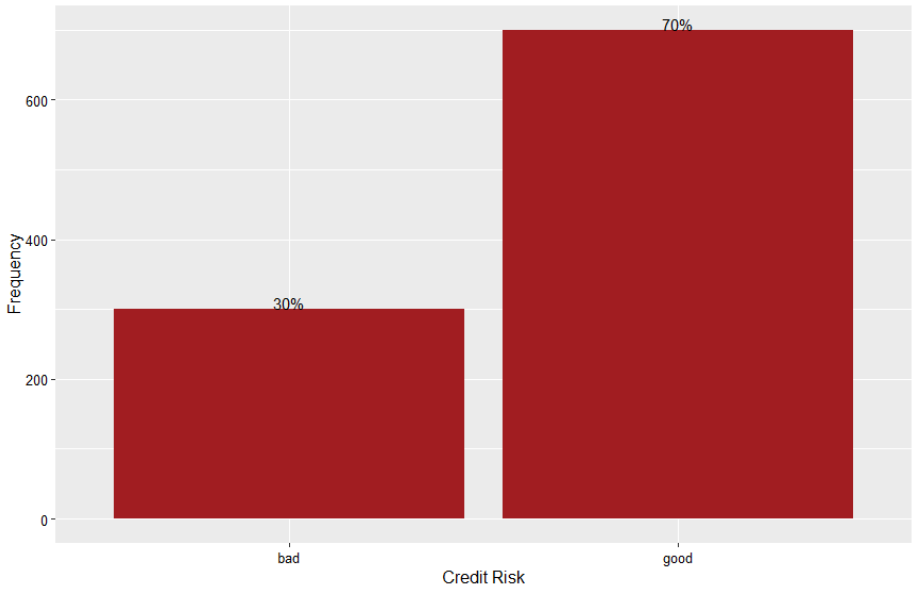
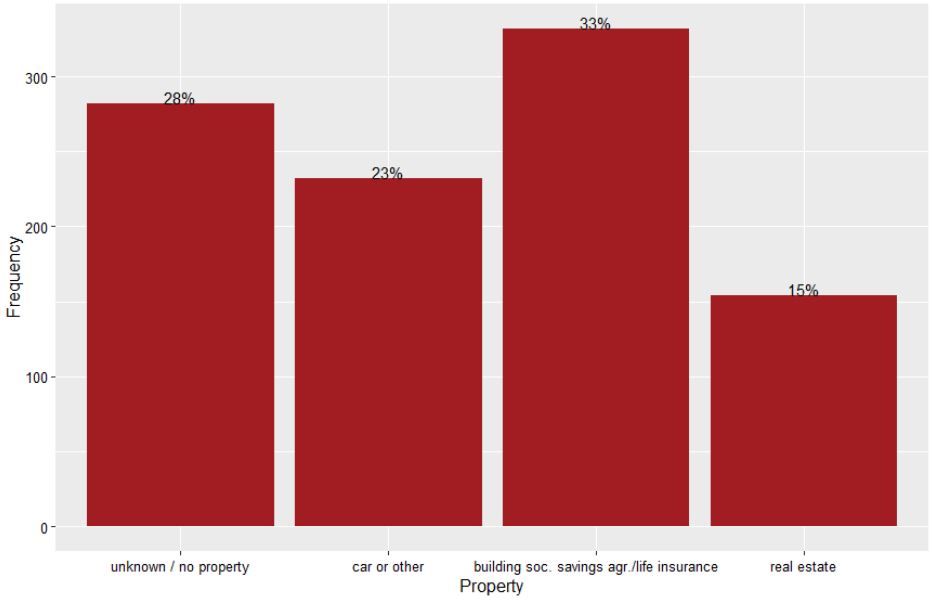


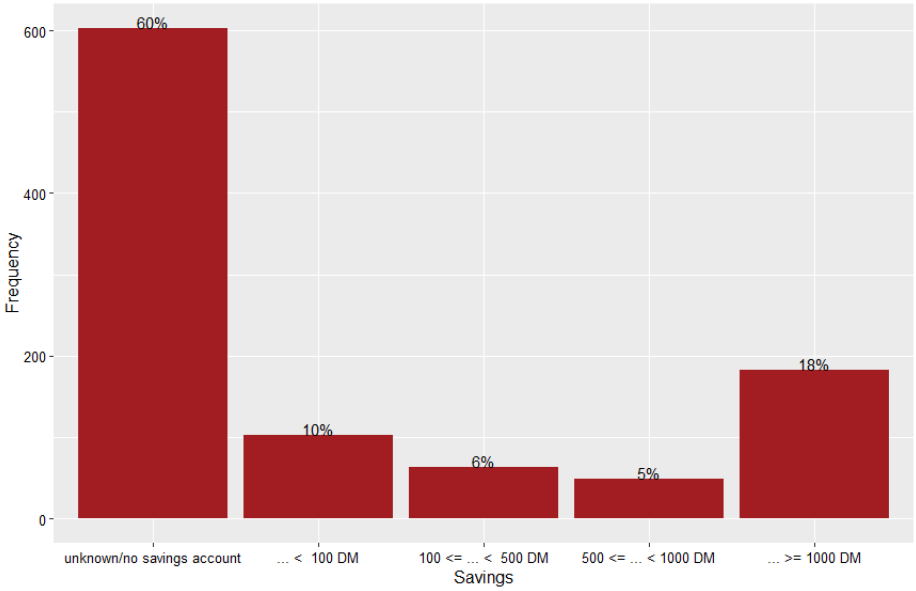
Figure 6 presents the proportion of good and bad creditors in our dataset. Therefore, we have 300 bad clients that default their credit at a certain point and 700 good clients that never default their credit. Some other variables are presented in the charts bellow.

Figure 7: Property proportion and frequency



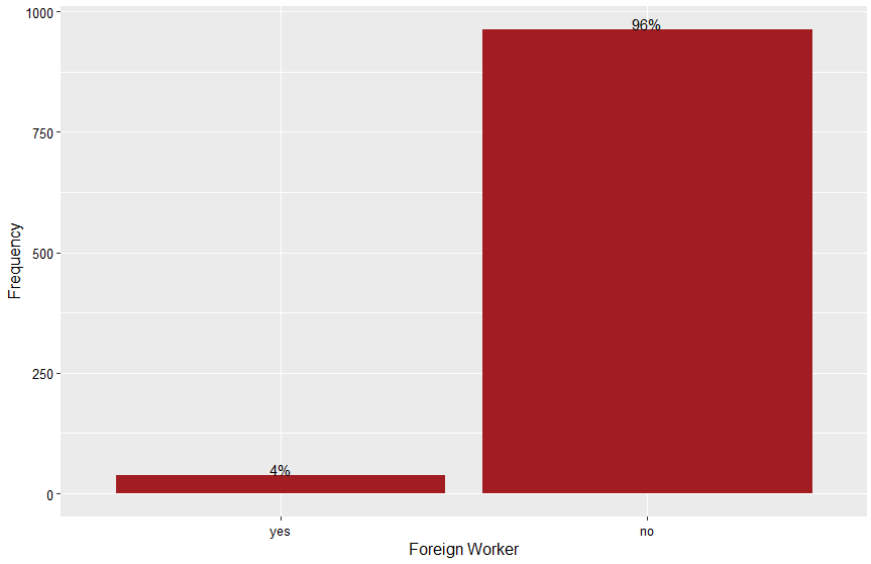
Some variables already have a good balance on the proportions on its categories, as is the case of Figure 7. All categories of the variable have a close number of observations or a sufficient amount for the models to be applied.

Figure 8: Savings proportion and frequency



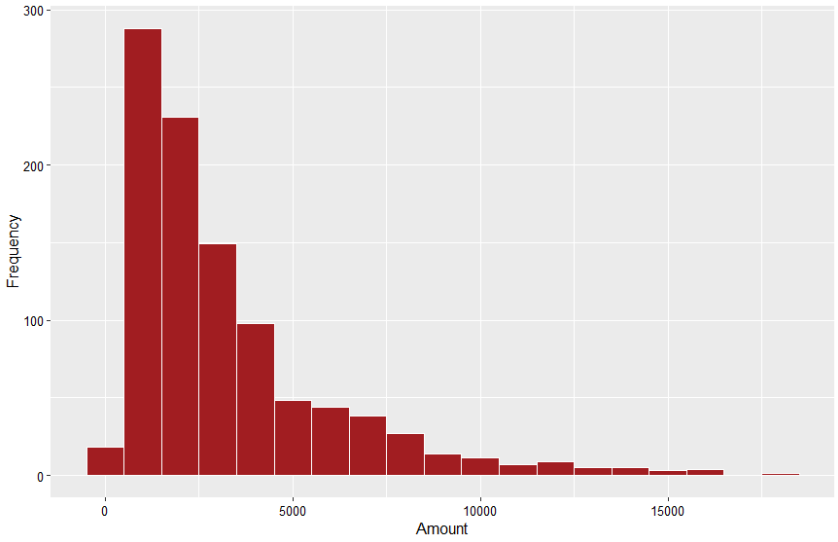
Other variables have different proportions along their categories, but it is possible to mix categories as it's the case of ordinal variables and create new categories with higher proportions. This is the case of Figure 8. The variables "purpose", "savings", "employment_duration", "number_credits", "personal_status_sex" and "other_installment _plans" had some of their categories aggregated into a more representative new category.

Figure 9: Foreign worker proportion and frequency



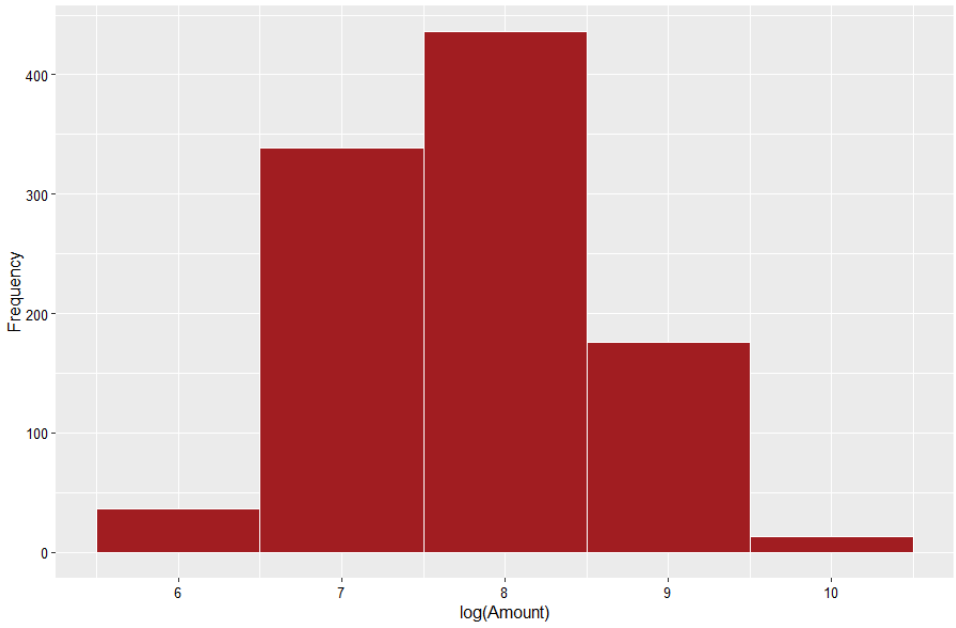
Variables as Foreign Worker presented on Figure 9, present a small quantity of categories and very low frequency of observation in one of the categories. Therefore, this already indicates that this may not be a good classifier for the model. The remaining variables are presented in the appendix.

Figure 10: Histogram of Amount



Numeric variables with extreme values, very asymmetrical or in a scale very different from other numeric values in the dataset need to be transformed, as is the case of the Amount in Figure 10. Either to level the scale of the values from that variable in comparison with the other variables or because some machine learning methods require the numeric variables to follow the gaussian distribution to perform properly. In this case the log was applied and now on the variable Amount should be read as log of Amount. It is presented on the figure bellow.

Figure 11: Histogram of log(Amount)



The same transformations were applied to the variables Duration and Age.

To better understand the relationship of the explanatory variable and the response variable, the statistical hypothesis testing was conducted on categorical and numerical explanatory variables, respectively. The Table 7 presents the statistics and results of the tests.

Table 7: Test of independence between response variable and explanatory variables

Variable	P-value	Significance
status	**< 0,0001**	**Significant**
duration	**< 0,0001**	**Significant**
credit_history	**< 0,0001**	**Significant**
purpose	**< 0,0001**	**Significant**
amount	**0,0059**	**Significant**
savings	**< 0,0001**	**Significant**
employment_duration	**< 0,0001**	**Significant**
personal_status_sex	**0,0222**	**Significant**
other_debtors	**0,0361**	**Significant**
property	**< 0,0001**	**Significant**
age	**0,0003**	**Significant**
other_installment_plans	**0,0016**	**Significant**
housing	**< 0,0001**	**Significant**
foreign_worker	**0,0158**	**Significant**
installment_rate	0,14	Not significant
telephone	0,2789	Not significant
present_residence	0,8616	Not significant
number_credits	0,4451	Not significant
job	0,5966	Not significant
people_liable	1	Not significant

From the table above we can see that 6 out of the 20 explanatory variables did not have enough evidence to reject the null hypothesis of the chi-squared test that states that the tested variable and the response variable are independent from each other. This can be verified by looking at the proportions in each category of the explanatory variable against the response variable.

Table 8: Cross table of Credit Risk and Status (%)

Status		no checking account	... < 0 DM	0<= ... < 200 DM	... >= 200 DM / salary for at least 1 year
Credit Risk	bad	45%	35%	5%	15%
	good	20%	23%	7%	50%

Table 9: Cross table of Credit Risk and Job (%)

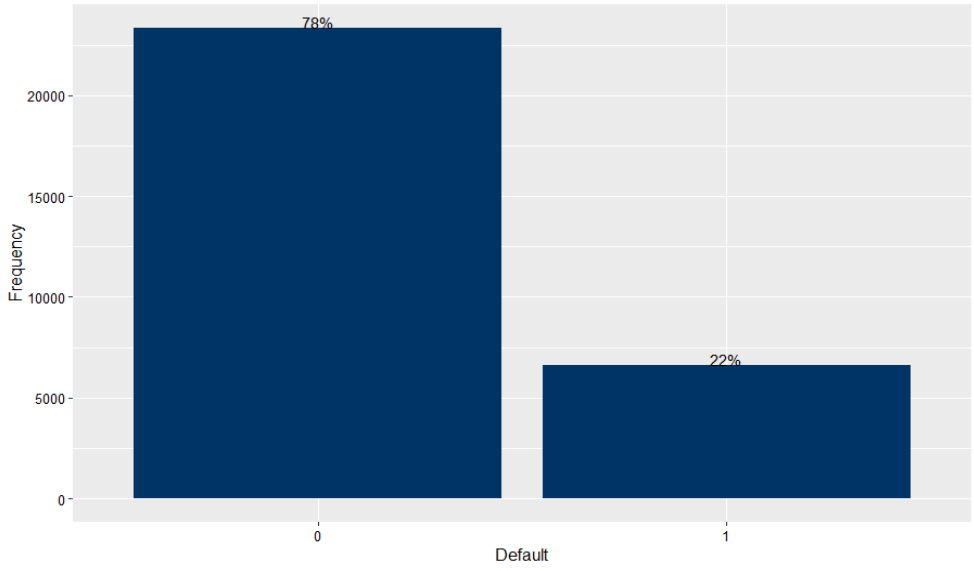
Job		unemployed/unskilled - non-resident	unskilled - resident	skilled employee/official	manager/self-empl./highly qualif. employee
Credit Risk	bad	2%	19%	62%	17%
	good	2%	21%	63%	14%

To better analyze this, note that the percentage is related to the marginal distribution of credit risk. This means that we are dividing the 100% of bad creditors along the Status and Job categories in each table. The same happens for the good creditors. Looking at Table 8, we can see that the proportions of good creditors and bad creditor along the Status categories are different. This indicates that exits a relationship between the variables. While in Table 9 we can see approximately the same distribution for bad and good creditors. This indicates that the variable Job is not a good variable indicate if the person is a good or bad creditor as we have seen with the Chi-Squared test.

Finally, the correlations of all variables were tested against each other using Anova, Caramer's V or Pearson's correlation to identify the high correlations between the explanatory variables and avoid the multicollinearity. None of the explanatory variables presented a high correlation between themselves, only with the response variable which is the desired scenario.

The second dataset to be analyzed was the UCI Credit Card Default. A database with 30000 rows, 23 explanatory variables, most of them numerical, and the response variable. This dataset is not as clean as the South German Credit Data. There was some data cleansing required to fix typo errors and other rows where removed. The proportion of good and bad creditors in the dataset can be seen in the next figure.

Figure 12: DEFAULT proportion and frequency



Different from the other dataset, in the Default Credit Data most of the variables are numerical, so there is no need for aggregation. The problem that arises here is other. There are many extreme values that we can see in almost all distributions of the numerical response variables. Therefore, a way to handle them must be found. In the boxplot below we can clearly spot some outliers.

Figure 13: Boxplot of Credit Amount

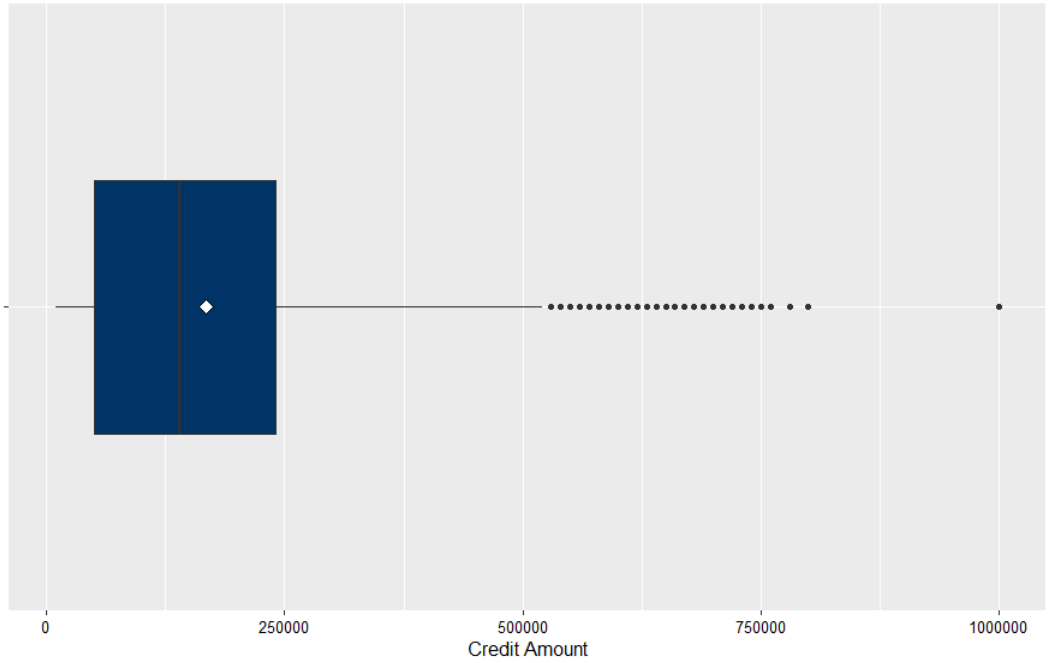
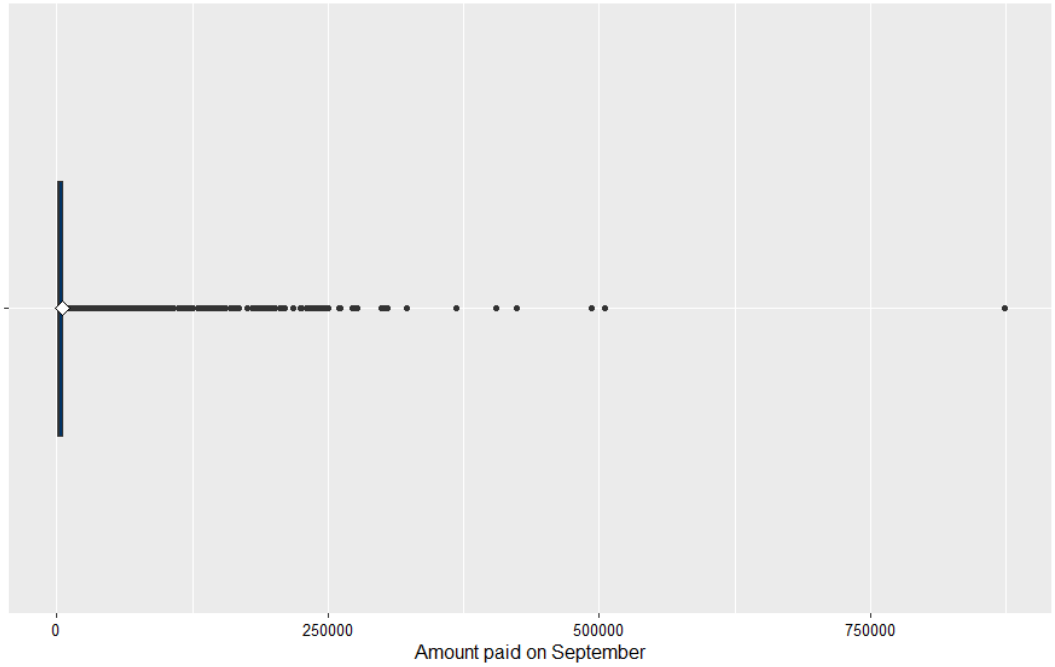


Figure 14: Boxplot of Amount paid in September



In the Figure 13 we can spot some black dots indicating outliers, but they are not so extreme and as far as we expected from the distribution. In Figure 14 the situation is different, here the interquartile size, the size of the "box" is barely visible, and the extreme values are much further away from the distribution. Using these values in the model could lead to wrong estimates of the parameters of the model. Therefore, a treatment for these values is required and there are tree approaches when handling outliers

The first one is the trimming, i.e., removing the outliers from the data, reducing the number of observations. The advantage of this option is that is easy to implement and fast, the problem is that removes information, and we will not be able to properly forecast and predict accurately the outcome of the model for extreme values.

The second one is capping, also called quantile base flooring, i.e., limiting the values of the outliers. If the outliers exceed a certain threshold, for example the 90th percentile, then they are set to a certain value. And the same thing for the lower limit, for example if they are bellow 10th percentile they are set to other value.

The third option is the median/mean imputation, i.e., replace the outliers by the median or mean. The median is recommended since the mean is influenced by extreme values.

After treating the outliers from all columns, the correlation analysis between all variables was conducted in order to identify the possibility of multicollinearity. Since the data in this dataset represent past payments of debt, the explanatory columns might have a high correlation. The Table 10 presents some correlations between explanatory variables that stood out.

Table 10: Correlation between explanatory variables

Variable 1	Variable 2	Correlation
BILL_AMT_SEP	PAY_AMT_APR	0,41
BILL_AMT_AUG	PAY_AMT_APR	0,43
BILL_AMT_JUL	PAY_AMT_APR	0,46
BILL_AMT_JUN	PAY_AMT_APR	0,48
BILL_AMT_MAY	PAY_AMT_APR	0,51
BILL_AMT_APR	PAY_AMT_APR	0,53
PAY_AMT_SEP	PAY_AMT_APR	0,46
PAY_AMT_AUG	PAY_AMT_APR	0,49
PAY_AMT_JUL	PAY_AMT_APR	0,51
PAY_AMT_JUN	PAY_AMT_APR	0,55
PAY_AMT_MAY	PAY_AMT_APR	0,55

One approach to handle the issue of multicollinearity is to aggregate the variables with high correlation between each other into a new variable that represents them all. This technique is known as Principal Components Analysis (PCA). The PCA technique was applied separately to each group of variables that were mainly related: PAY, BILL_AMT and PAY_AMT. The tables bellow presents the principal components of each analysis.

Table 11: Principal Components variability explained for PAY variables

Importance of components	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1,95	0,97	0,70	0,58	0,50	0,44
Proportion of Variance	63%	16%	8%	6%	4%	3%
Cumulative Proportion	63%	79%	87%	93%	97%	100%

Table 12: Principal Components variability explained for BILL_AMT variables

Importance of components	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2,32	0,57	0,35	0,27	0,23	0,22
Proportion of Variance	90%	5%	2%	1%	1%	1%
Cumulative Proportion	90%	95%	97%	98%	99%	100%

Table 13: Principal Components variability explained for PAY_AMT variables

Importance of components	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1,67	0,84	0,80	0,79	0,78	0,78
Proportion of Variance	47%	12%	11%	11%	10%	10%
Cumulative Proportion	47%	58%	69%	80%	90%	100%

The table above present the principal components (PC) created for each PCA analysis. The proportion of variance represents the percentage of the variance explained by that PC alone. The sum of all PCs is equal to all the variance explained by the variables used to create the PC. This way, in the PCA for PAY variables the first three PCs explain 87% of all variance, we are using half the amount of variables to explain nearly 90% of the existing variance on those six initial variables.

For the second PCA Analysis using BILL_AMT variables, the PCs selected were PC1 and PC2 representing 95% of all variability. For the PCA of PAY_AMT variables the four first PCs were selected resulting in 80% of the variability in those PCs.

To interpret the PCs selected to keep we can look at their eigenvectors looking for very positive or very negative values. The eigenvector represents by which coefficient each variable was multiplied by to result in the PC. The table below present the eigenvectors for the selected PCs for BILL_AMT PCA.

Table 14: Eigenvectors of BILL_AMT PCA

Variables	PC1	PC2
BILL_AMT_SEP	0,40	0,54
BILL_AMT_AUG	0,41	0,43
BILL_AMT_JUL	0,42	0,16
BILL_AMT_JUN	0,42	-0,18
BILL_AMT_MAY	0,41	-0,43
BILL_AMT_APR	0,40	-0,53

In the first eigenvector we can see that it's giving basically the same weight to all variables. The eigenvector for the second PC it's clearer to see that there is an ascending trend going from a negative coefficient from older BILL_AMTs and increasing as the time passes. This PC could represent the increase of BILL_AMTs along the time.

As it was done with the first data set, to better understand the relationship of the explanatory variable and the response variable, the statistical hypothesis testing of chi-squared and Mann-Whitney were conducted on categorical and numerical explanatory variables, respectively. This time, most of the

tests conducted were of Mann-Whitney since there is a prevalence of numerical explanatory variables in this dataset. The table below presents the statistics and results of the tests.

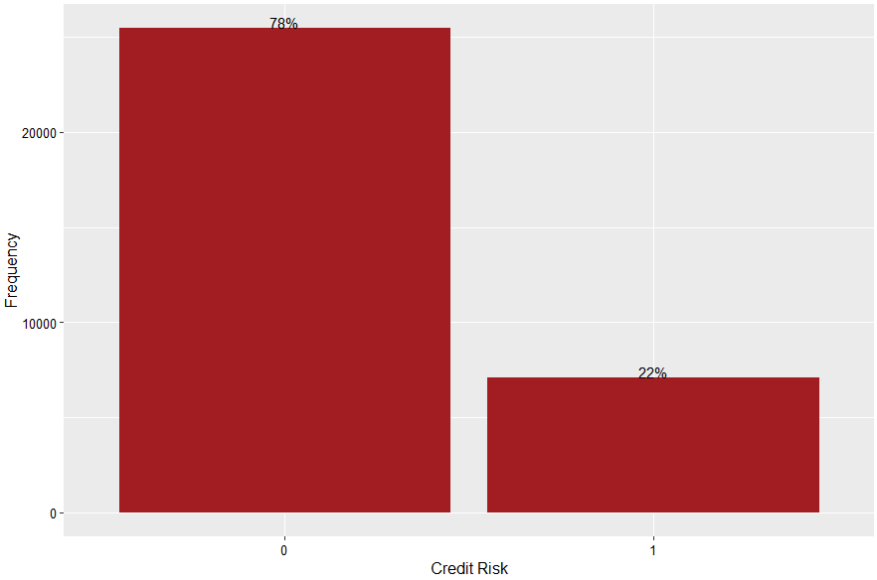
Table 15: Statistical tests outputs for all explanatory variables

Variable	P-value	Significance
LIMIT_BAL	**< 0,0001**	**Significant**
SEX	**< 0,0001**	**Significant**
EDUCATION	**< 0,0001**	**Significant**
MARRIAGE	**< 0,0001**	**Significant**
PAY_SEP	**< 0,0001**	**Significant**
PAY_AUG	**< 0,0001**	**Significant**
PAY_JUL	**< 0,0001**	**Significant**
PAY_JUN	**< 0,0001**	**Significant**
PAY_MAY	**< 0,0001**	**Significant**
PAY_APR	**< 0,0001**	**Significant**
BILL_AMT_SEP	**< 0,0001**	**Significant**
BILL_AMT_AUG	**0,0071**	**Significant**
BILL_AMT_JUL	**0,0282**	**Significant**
PAY_AMT_SEP	**< 0,0001**	**Significant**
PAY_AMT_AUG	**< 0,0001**	**Significant**
PAY_AMT_JUL	**< 0,0001**	**Significant**
PAY_AMT_JUN	**< 0,0001**	**Significant**
PAY_AMT_MAY	**< 0,0001**	**Significant**
PAY_AMT_APR	**< 0,0001**	**Significant**
AGE	0,3725	Not significant
BILL_AMT_JUN	0,1478	Not significant
BILL_AMT_MAY	0,2354	Not significant
BILL_AMT_APR	0,9895	Not significant

Looking at the table above we could see that most of the explanatory variables have a relationship with the response variable. The only exceptions are the age of the client and the bill amount for the months of June, May and April. It is interesting to notice that the information of the most recent month about the bill amount is relevant and has a relationship with the response variable but as the months go by, this information is not so relevant anymore. We can notice that looking at the p-value and test statistic of this set of variables (bill amount), these values decrease as the months passes.

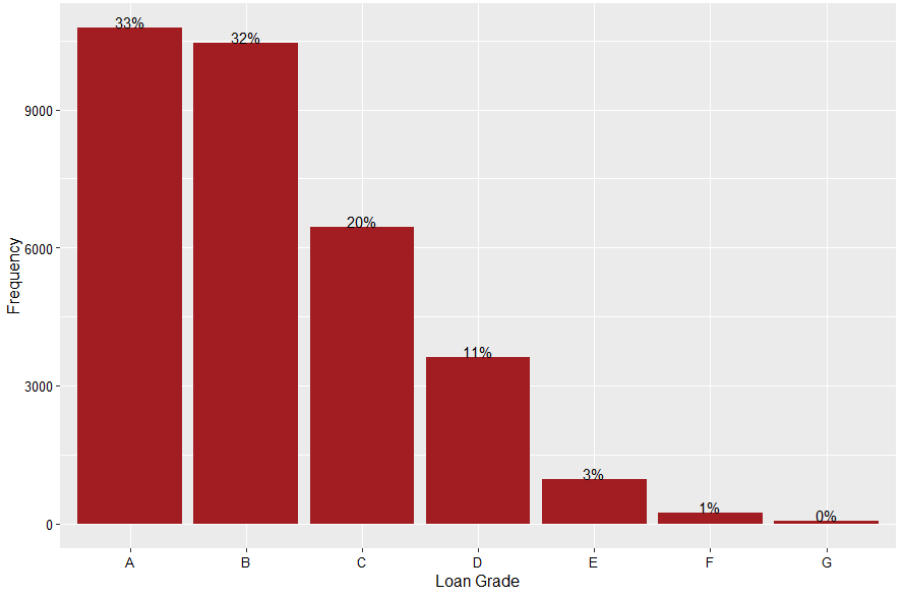
The last dataset to be analyzed was the Credit Risk dataset with over 32.000 rows and 11 explanatory variables, being 4 of them categorical and 7 numerical. The figures below present some descriptive information about some variables.

Figure 15: Distribution of good and bad clients



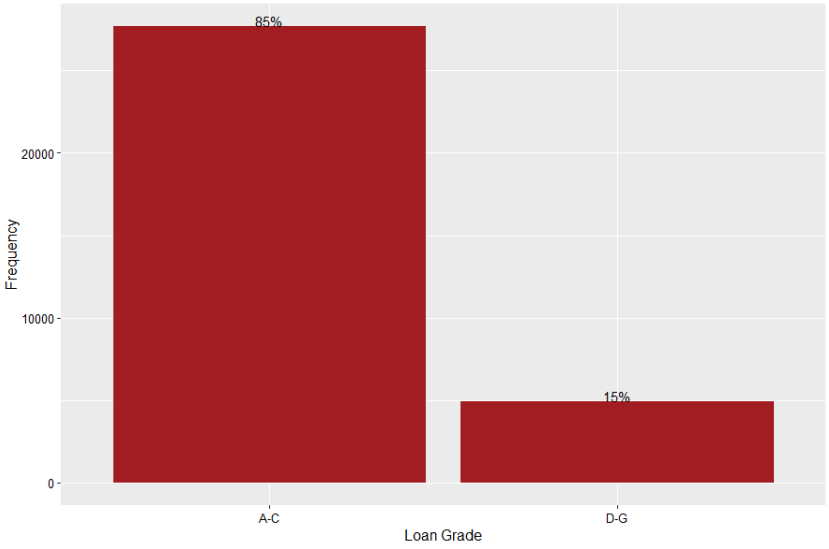
In figure above we can see the unbalanced proportion of good and bad clients being 78% good clients and 22% considered bad client. In this context the bad client is the one who didn't pay his debts and the goo is the one who met his obligations with the bank.

Figure 16: Distribution of loan grades



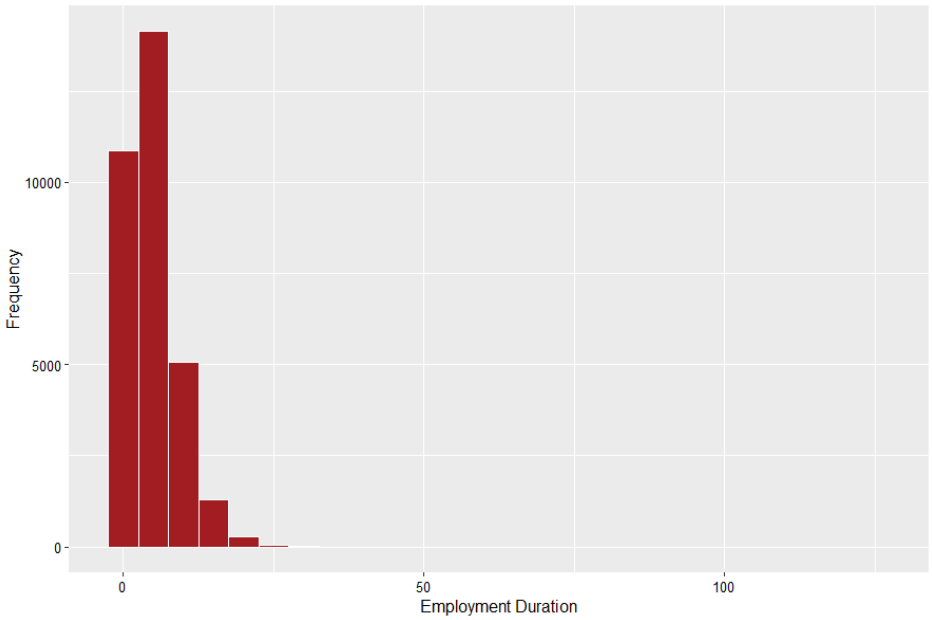
In Figure 16, the letters represent the grades given to each client. The closer to A the grade, more reliable is the client in paying his debts, the closer to G, worse is the clint in paying his debts. The bar chart emphasizes the imbalance between the categories. As discussed before, variables with very few observations could lead to the overestimation or underestimation of a parameter in the models. For this reason, the categories were grouped into new groups with bigger frequency as presented below.

Figure 17: Distribution of loan grades grouped



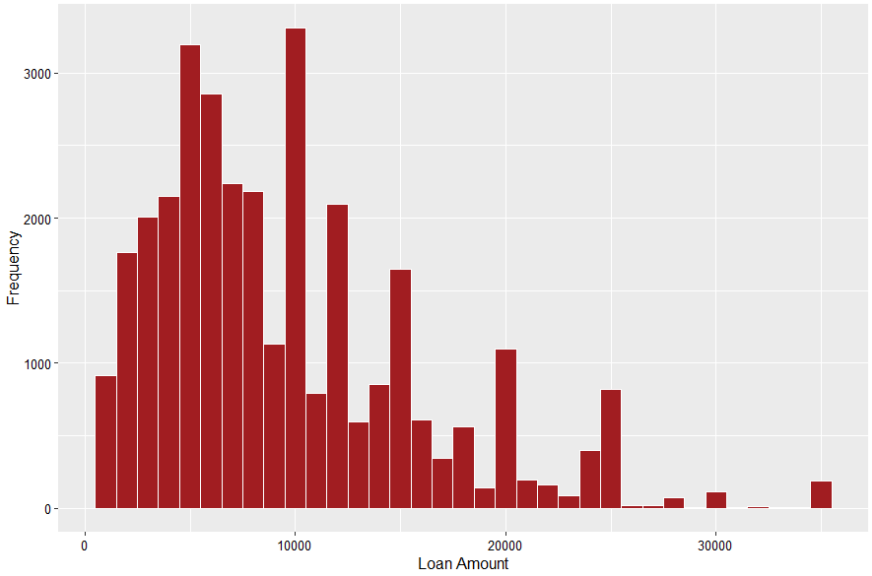
The category D-G still has only 15% of the overall data but it's already better than the previous distribution. The category C could not be aggregated to D, E, F and G because they presented a different behavior in relation to the response variables compared to these categories. Clients in category A-C are more likely to pay their debts while clients in categories D-G are less likely.

Figure 18: Histogram of employment duration in years



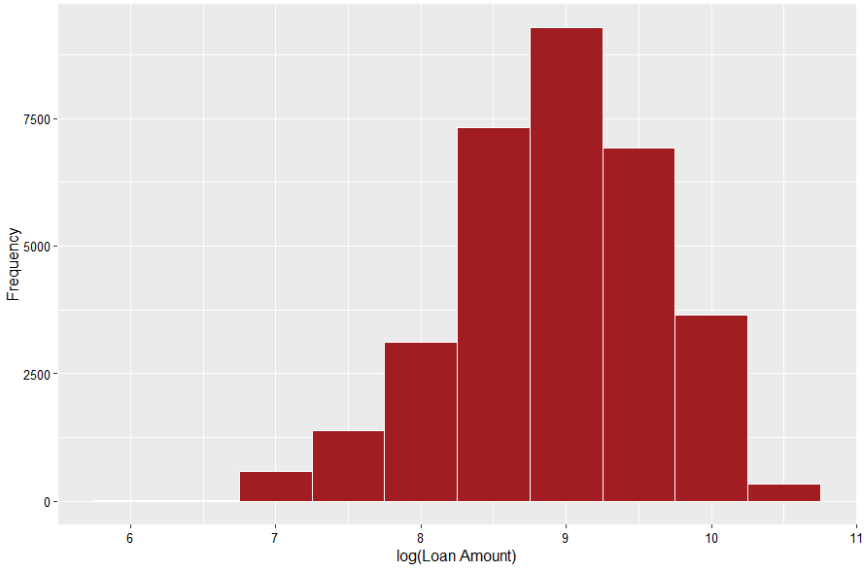
For the employment duration we can see that there are some extreme values that doesn't make sense as clients that have an employment duration of over 100 years. To fix this, all values above 60 years of employment duration were replaced by the average as discussed before.

Figure 19: Histogram of loan amount



In the histogram for the loan amount, we can see the skewed distribution and the difference of scale compared to variables as employment duration and age. As seen on the treatment of the South German Credit dataset, the best approach is to apply a transformation to these types of variables. Along with loan amount, income was also applied the log transformation. The table below presents the histogram for the log of loan amount, now more symmetrical and in the same scale as other variables.

Figure 20: Histogram of log loan amount



To better understand the relation between the variables the table below was made with information about the variable name, the statistical test done to verify association, the statistic of the test, the p-value and the interpretation of the result.

Table 16: Statistical tests outputs for all explanatory variables

Variable	P-value	Significance
age	**< 0,0001**	**Significant**
income	**< 0,0001**	**Significant**
house	**< 0,0001**	**Significant**
emp_dur	**< 0,0001**	**Significant**
loan_intent	**< 0,0001**	**Significant**
loan_grade	**< 0,0001**	**Significant**
loan_amnt	**< 0,0001**	**Significant**
loan_int_rate	**< 0,0001**	**Significant**
loan_percent_income	**< 0,0001**	**Significant**
historical_default	**< 0,0001**	**Significant**
cred_hist_length	**< 0,0001**	**Significant**

All variables presented a strong correlation with the response variable.

6. MODEL FITTING AND RESULTS

For the credit score analysis, the models used in this work were selected based on the interpretability. There is a huge discussion about machine learning models regarding their interpretability. Models considered interpretable where knowledge about how the explanatory variables impact the response variables and the intensity of it are considered white-box models. In the other hand, models where there is no clear understanding of the process and how each explanatory variable impact the response variable are considered black-box models.

It is not a general rule, but usually black-box models tend to perform slightly better than white-box models. But depending on the application of the model, knowing the behavior of the response variables in relation to the explanatory variable is desirable. For instance, in credit scoring it's not only interesting to know whether a client is classified as a good or bad payer but also understand why the client was classified as that because of the "right to explanation" regulation introduced by the General Data Protection Regulation (GDPR) and the Equal Credit Opportunity Act (ECOA) (Farrar & Glauber, 1967; Alin, 2010).

The models used in the datasets are a mix of white-box and black-box models. Although the black-box models used in this work do not provide an interpretability as the white-box models used here, they still provide some interpretability regarding the most important variables for forecasting the outcome of the response variable. The white-box models used in the work were Logistic Regression (LR), Decision Tree (DT) and Naïve Bayes (NB). The black-box models used in this here were Random Forest (RF), k Nearest Neighbors (k-NN) and Support Vector Machine (SVM).

The same process was followed by all models when possible, data preparation (one hot encoding for models that required only numeric variables as SVM), split the data into train and test sets (the seed was always reset to the same when splitting the data to guarantee that the data was in all train and test sets were always the same), feature selection (backward and forward methods for the logistic regression for example), build the model and fine tuning using GridSearch when possible to get the best possible parameters that maximize the correct predictions.

There are several metrics that can be used to compare the models as accuracy for instance. Accuracy measures the total amount of right predictions over total entries. In cases of unbalanced data, especially in extreme unbalanced data, accuracy may not be the best measure to compare. In cancer studies for example, the patients that have cancer represent a very small fraction of all patients. If 99% of the patients don't have cancer and 1% have, if we are trying to maximize accuracy with our model, the best option for the model would be to classify all patients as not having cancer, because it would be right in 99% of cases, but this is not acceptable.

Other used measure is the precision score, this metric measure that proportion of positive cases that were correctly predicted. The recall score is the complementary of precision score, it measures the proportion of negative cases that were correctly predicted. This metrics are more useful for unbalanced data because they measure individually the performance of the model in predicting positive and negative outcomes, respectively. Another option is the F1-Score, this metric combines the precision and recall int an harmonic mean between those values.

The problem that arises is that all this metrics are subject to change since they are calculated from the cross-table of predicted and actual values for the response variable. For models that generates probabilities the predicted value is generated from a cutoff point the is defined by the user. We can always use the best cutoff point that maximizes the number of correct guesses but usually the specialist of the field may want to use other value that makes more sense depending on the Type of Error that it's being tried to be avoided.

In the case of the patients with cancer, it's better to recommend to a patient to do the exam to verify if he/she has cancer and don't have than to not recommend to a patient do the exam and they have. For financial institutions they may be comfortable in changing their cutoff limit and provide credit to another 50 clients where 45 of them were classified as good payers and 5 of them as bad.

For this reason, we won't use none of these metrics to compare the model's performance between them, but we will use the area under the curve (AUC) from ROC curve. This metric gives a value for the overall performance of the model, but in any case, in the comparison tables accuracy, recall and precision will also be presented for informative purpose.

The first models to be compared will be the models generated from the South German Credit data. The table below presents the AUC, accuracy, recall, precision and confusion matrix for each one of the models.

Table 17: Model metrics comparison for South German Credit dataset

Model	AUC	Accuracy	Recall	Precision
Random Forest	0,795	0,760	0,923	0,333
Naïve Bayes	0,785	0,748	0,912	0,319
SVM	0,765	0,764	0,873	0,478
Logistic Regression	0,749	0,748	0,901	0,348
Decision Tree	0,742	0,736	0,818	0,522
k-NN	0,730	0,704	0,917	0,145

Based on the table above, we notice that the random forest presented the highest AUC and Accuracy for this dataset followed by Naïve Bayes, with results very close to the RF model, and SVM in this order. In the bottom of the table, we can see that the k-NN model had the worst performance for this dataset. We can notice also that as recall gets higher values, precision tends to shrink and vice-versa. The Decision Tree presented the highest precision value and the lowest recall value. In the other hand, k-NN presented the second highest recall and the lowest precision. Models with the highest AUC usually have a good balance of recall and precision as we can see for the Random Forest and Naïve Bayes.

Table 18: Feature Importance for each model for South German Credit dataset

Feature Importance	Random Forest	Naïve Bayes	SVM	Logistic Regression	Decision Tree
1º	status	status	status	status	status
2º	duration	duration	duration	duration	duration
3º	amount	credit history	savings	purpose	amount
4º	credit history	savings	age	savings	savings

5º	savings	property	purpose	personal status sex	purpose
6º	purpose	age	credit history	credit history	age
7º	installment rate	employment duration	personal status sex	installment rate	credit history
8º	personal status sex	personal status sex	foreign worker	age	housing
9º	other debtors	installment rate	housing	other installment plans	present residence
10º	housing	other installment plans	other debtors	housing	job

The Table 18 presents data about feature importance for each of the models. Note that there is no column for k-NN because there is no direct way to calculate feature importance for it. The feature importance helps to understand what are the variables that impact the most in the accuracy of the model if they were removed. Looking at the table we can notice that for all models the variables status and duration were the most important variables in predicting if the client is a good or a bad payer. Other variables that call our attention is savings that is on the top 5 for all models. Other very consistent variables for the models were purpose, credit history and age.

To better understand these results, the odds ratio for the coefficients of the logistic regression model are presented in the table below.

Table 19: Outputs from Logistic Regression model for South German Credit dataset

Variables/Categories	Estimate	OR	P-Value	Significance
(Intercept)	-3,83	0,02	0,03	Significant
statusno checking account	1,56	4,77	0,00	Significant
status... < 0 DM	1,25	3,50	0,00	Significant
status0<= ... < 200 DM	1,04	2,83	0,01	Significant
duration	0,95	2,59	0,00	Significant
credit_historydelay in paying off in the past	1,59	4,90	0,00	Significant
credit_historycritical account/other credits elsewhere	0,69	2,00	0,14	Not Significant
credit_historyno credits taken/all credits paid back duly	0,49	1,63	0,04	Significant
credit_historyexisting credits paid back duly till now	0,54	1,71	0,15	Not Significant
purposeothers	1,62	5,07	0,17	Not Significant
purposecar (new)	-0,01	0,99	1,00	Not Significant
purposecar (used)	0,69	1,99	0,56	Not Significant
purposefurniture/equipment	0,58	1,78	0,63	Not Significant

purposerepairs/domestic equipment	1,43	4,20	0,23	Not Significant
purposebusiness/retraining	0,97	2,65	0,42	Not Significant
savingsunknown/no savings account	0,88	2,42	0,00	Significant
savings... < 100 DM	0,37	1,45	0,31	Not Significant
installment_rate25 <= ... < 35	0,10	1,11	0,77	Not Significant
installment_rate20 <= ... < 25	0,34	1,40	0,37	Not Significant
installment_rate< 20	0,85	2,34	0,01	Significant
personal_status_sexmale : married/widowed or female : single	-0,69	0,50	0,00	Significant
age	-0,76	0,47	0,03	Significant
other_installment_plansbank/stores	0,53	1,69	0,04	Significant
housingrent	-0,46	0,63	0,07	Significant
housingown	0,53	1,69	0,17	Not Significant

Based on the table above and in the columns of odds ratio (OR) we can interpret that clients that have delayed in paying off in the past have 390% more chance to be classified as bad payers when compared with clients that never delayed payments in the past. For the duration variable we can interpret from the table that for each increase in 10% in the duration increases the chance of being a bad client in 9.5%. Regarding the savings, clients with no savings account have 142% more chance to be a bad payer then client with savings account with more than 100 DM.

The next models to be compared are the models originated from the UCI Credit Card Default data. The table below presents the AUC, accuracy, recall, precision and confusion matrix for each one of the models.

Table 20: Model metrics comparison for UCI Credit Card Default dataset

Model	AUC	Accuracy	Recall	Precision
Random Forest	0,760	0,76	0,941	0,370
Logistic Regression	0,759	0,748	0,917	0,423
k-NN	0,746	0,704	0,946	0,321
Naïve Bayes	0,743	0,748	0,874	0,485
SVM	0,740	0,764	0,911	0,440
Decision Tree	0,687	0,804	0,939	0,358

Based on Table 20 we can see that as for the South German Credit data, the model that performed the best was Random Forest with an AUC of 0.78. The next best models were logistic regression and k-NN. The Decision Tree had the worst AUC value but the best Accuracy. This happens because the predictions of DTs are not continuous, they don't provide a probability. So, their output for the ROC curve are only the specific points of each node. Therefore, their AUC is underestimated, but the model presented a good accuracy overall. The models with best precision were naïve bayes, SVM and logistic

regression with 48.5%, 44% and 42.3% correct classifications of positive cases, respectively. Note that the cutoff to decide whether the client is a good or bad payer is changeable and this will affect the values of accuracy, recall and precision. In the next table we can see the most important variables for each model

Table 21: Feature Importance for each model for UCI Credit Card Default dataset

Feature Importance	Random Forest	Naïve Bayes	SVM	Logistic Regression	Decision Tree
1º	pay_PC1	pay_PC1	pay_PC3	pay_PC1	pay_PC1
2º	pay_amt_PC1	pay_amt_PC1	pay_PC1	pay_PC3	pay_PC3
3º	bill_amt_PC1	LIMIT_BAL	pay_PC2	pay_PC2	pay_PC2
4º	bill_amt_PC2	pay_PC2	pay_amt_PC2	pay_amt_PC1	
5º	pay_amt_PC2	pay_PC3	pay_amt_PC1	bill_amt_PC1	
6º	pay_amt_PC3	bill_amt_PC2	pay_amt_PC4	LIMIT_BAL	
7º	pay_amt_PC4	pay_amt_PC2	pay_amt_PC3	MARRIAGE	
8º	AGE	EDUCATION	bill_amt_PC2	bill_amt_PC2	
9º	pay_PC2	SEX	pay_amt_PC1	SEX	
10º	LIMIT_BAL	MARRIAGE	AGE	pay_amt_PC3	

We can see on Table 21 that pay_PC1 is the most important feature for all models, except SVM, and the second most important to SVM, also on the topmost important we have pay_PC2, pay_PC3, pay_amt_PC1 and pay_amt_PC2. These variables are the ones that helps the most in differentiating good clients from bad clients. To better understand what each principal component represents, the tables below present the respective eigenvectors for each principal component.

Table 22: Eigenvectors of principal components for PAY PCA

Variables	PC1	PC2	PC3
PAY_SEP	-0,37	-0,53	-0,50
PAY_AUG	-0,40	-0,50	-0,02
PAY_JUL	-0,42	-0,15	0,64
PAY_JUN	-0,44	0,23	0,36
PAY_MAY	-0,43	0,42	-0,16
PAY_APR	-0,39	0,46	-0,43

Table 23: Eigenvectors of principal components for PAY AMOUNT PCA

Variables	PC1	PC2
PAY_AMT_SEP	0,41	-0,53
PAY_AMT_AUG	0,41	-0,46
PAY_AMT_JUL	0,42	-0,19
PAY_AMT_JUN	0,40	0,35
PAY_AMT_MAY	0,41	0,43
PAY_AMT_APR	0,40	0,41

Each value of the eigenvector is the coefficient which the corresponding variable was multiplied and then added up to give the PC1 value. We can see that for both PAY AMOUNT PCA and PAY PCA the first component is a general overview of all variables, having all variables multiplied by a similar coefficient. In the second principal component for both PCAs we can see that there is a sequential logic involved, having the oldest months a positive coefficient that decreases until an equally negative coefficient for the most recent month.

The last dataset to be analyzed was the Credit Risk dataset from Kaggle. This dataset is composed of non-practical data about credit default. In this data the same six models were applied, and their results are presented in the table below.

Table 24: Model metrics comparison for Credit Risk dataset

Model	AUC	Accuracy	Recall	Precision
Random Forest	0,938	0,934	0,995	0,719
k-NN	0,899	0,903	0,984	0,624
Logistic Regression	0,876	0,863	0,958	0,532
SVM	0,870	0,900	0,985	0,604
Naïve Bayes	0,869	0,853	0,915	0,636
Decision Tree	0,858	0,921	0,996	0,662

Based on Table 24 we can see that the model that performed the best was random forest with 0.938 of area under the curve, followed by k-NN with 0.899, and logistic regression with 0.876, respectively. It's interesting to notice that although the decision tree presented the worst value for AUC, it had the second highest accuracy of all models, losing only for random forest. This happens because decision tree models do not provide a probability in their calculations. This model is a classifier that returns 0s and 1s. therefore the calculation of the AUC is not properly done since there are not enough points to estimate it.

It is important to notice that the values of AUC and accuracy for these models are much higher when compared with the same models applied to the other datasets. This happens because this is a synthetical dataset created for educative purpose only. So, the developer of the dataset might have created the variables to suit the classification. In real problem data, such results are unlikely to be seen.

Table 25: Feature Importance for each model for Credit Risk dataset

Feature Importance	Random Forest	Naïve Bayes	SVM	Logistic Regression	Decision Tree
1º	house	loan percent income	income	loan percent income	loan percent income
2º	loan grade	loan interest rate	loan grade	loan grade	loan grade
3º	loan percent income	income	loan percent income	loan amount	house

4º	loan intent	loan grade	employment duration	house	loan interest rate
5º	employment duration	house	age	loan interest rate	income
6º	income	historical default	loan amount	loan intent	loan intent
7º	loan interest rate	employment duration	loan interest rate	employment duration	employment duration
8º	loan amount	loan amount	credit history length		loan amount
9º	age	loan intent	loan intent		
10º	credit history length	age	historical default		

The Table 25 provides the feature importance for each model, it's noticed that loan percent income, loan grade, income and house are the most import variables to classify if a client is a good or a bad client. Variables as credit history length and historical default were not so important for the classification. Using the odds ratio from logistic regression we can compare the different categories of variables and see how the odds of been classified as bad client increases or decreases when we change that value in that variable. The table below presents the odds ratio for the Credit Risk dataset.

Table 26: Outputs from Logistic Regression model for Credit Risk dataset

Variables/Categories	Estimate	OR	P-Value	Significance
(Intercept)	2,51	12,32	0,00	Significant
houseMORTGAGE	1,87	6,47	0,00	Significant
houseRENT/OTHER	2,60	13,42	0,00	Significant
emp_dur	-0,01	0,99	0,09	Significant
loan_intentDEBTCONSOLIDATION	0,86	2,36	0,00	Significant
loan_intentHOMEIMPROVEMENT	0,95	2,58	0,00	Significant
loan_intentMEDICAL	0,67	1,96	0,00	Significant
loan_intentPERSONAL	0,24	1,27	0,00	Significant
loan_intentVENTURE	-0,24	0,78	0,00	Significant
loan_gradeD-G	2,23	9,31	0,00	Significant
loan_amnt	-1,24	0,29	0,00	Significant
loan_int_rate	0,12	1,13	0,00	Significant
loan_percent_income	0,15	1,16	0,00	Significant

Looking at the table above we can take some interesting insights. Considering two clients with the exact same characteristics but one with rented house and the other owning the house, the clients that

have a rented house have 13 times more chance to be classified as bad clients than the ones with own house. In the same point of view, clients with loan grade of D to G have their chance of being a bad client increased by 800% compared with clients with loan grade from A to C. Finally, for each 1% increase in the percent of the income that the loan represents, the client has an increase of their odds of 16% in being a bad client. That said, clients with loans that have installments that represents 40% of their monthly income have 160% more chance of having a default then clients with installments that represents 30% of their monthly income.

7. CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORKS

In this study, six different machine learning techniques were applied to three different datasets to classify clients into committing a default or not. The goal was to compare their performance and asses the better overall model. The six models used were Logistic Regression, Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, and k Nearest Neighbors. The dataset which these models were applied varied in their sizes, types of variables and proportion of response variable. In all datasets transformations and data preparation was held to avoid multicollinearity, underestimating or overestimating coefficients, and to deal with outliers.

To compare the models the AUC was the metric used since the Accuracy, F1-Score, Recall and Precision are metrics that change depending on the cutoff point although the optimal cutoff point was always used to present the confusion matrix. From the results obtained, for all the three datasets, the random forest model was the one to present the better AUC values. From the remaining models, logistic regression was the one that presented the second-best AUC values overall. It's interesting to note that the decision tree model presented the worst AUC values in two of the datasets and the second worst in the other dataset but had one of the highest accuracies in two of the datasets. This happens because decision trees generate binary outputs instead of probabilities, therefore, the calculation for the AUC is not so accurate and continuous as the models that provides probabilities.

Therefore, the random forest presented to be a good model option in the overall for allowing interpretation of the output and the variables in the model, but decision tree is an even simpler model with easier understanding that presented good results in the accuracy when applying the test data to it.

For future studies, there are several recommendations that can be considered. The first one would be to implement and add to the comparison other models that were not analyzed in this study such as neural networks and gradient boosting. Other interesting option that can be done is to apply balancing techniques to the data before applying the models. Imbalanced can be an issue for some models as k-NN. Finally, new datasets could be added to increase the datasets compared and provide a more in-depth result.

8. BIBLIOGRAPHY

Abdou A.H., & Pointon J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intell. Sys. Acc. Fin. Mgmt. 18*, 59–88.

Al Amari A. (2002). The credit evaluation process and the role of credit scoring: a case study of Qatar. PhD thesis, University College Dublin.

Alin A. (2010). Multicollinearity. *WIREs Computational Statistics, 2*(3), 370-374.

Arora, N. & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing Journal* 86 105936.

Ashofteh, A. & Bravo, J. M. (2021). A conservative approach for online credit scoring. *Expert Systems With Applications* 176, 114835.

Baesens B., Gestel T.V., Viaene S., Stepanova M., Suykens J., Vanthienen J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society 54*(6), 627–635.

Bao, W., Lianju, N., Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems With Applications* 128, 301–315.

Bellotti T. & Crook J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications 36*(2), 3302–3308.

Berkson, J. (1944). Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc. 39(*227), 357–365.

Boyle M. (2020). Are Economic Recessions Inevitable?. *Investopedia.* Accessed 10 July 2022, <https://www.investopedia.com/ask/answers/032015/are-economic-recessions-inevitable.asp>.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA.

Brown I., & Mues C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453.

Chandler G.G., Coffman J.Y. (1979). A comparative analysis of empirical vs. judgemental credit evaluation. *The Journal of Retail Banking 1*(2), 15–26.

Chawla N. V., Bowyer K. W., Hall L. O., & Kegelmeyer W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Crook J.N. (1996). Credit scoring: an overview. Working paper series No. 96/13, British Association, Festival of Science, University of Birmingham, The University of Edinburgh.

Chen C., Yokoyama S., Yamashita T., & Kawamura H. (2019). Application of XGBoost to Credit Scoring. *IPSJ SIG Technical Reports, 11*, ICS-194.

Dastile, X., Celik, T., Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91, 106263.

Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics.: Conference Series 949 012009.*

Dua, D. & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Dumitrescu E. I., Hué S., & Hurlin C. (2021). Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds.

Dutta S., Shekhar S., Wong W.Y. (1994). Decision support in non-conservative domains: generalization with neural networks. *Decision Support Systems 11*(5), 527–544.

Farrar D. E. & Glauber R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics, 49*(1), 92-107.

Grömping U. (2019). South German Credit Data: Correcting a Widely Used Data Set. Report 4/2019, Reports in Mathematics, Physics and Chemistry, Department II, Beuth University of Applied Sciences Berlin.

Gunnarsson B. R., vanden Broucke S., Baesens B., Óskarsdóttir M., Lemahieu W. (2021). Deep learning for credit scoring: do or don't*?. Eur J Oper Res, 295*(1), 292-305.

Henrique B.M., Sobreiro V.A., Kimura H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications, 124*, 226-251.

Heffernan S. (2005). Modern Banking. John Wiley and Sons, Ltd: Chichester.

Hull J. C. (2017). Options, futures and other derivatives. *Pearson*. 10th Edition.

Hsieh N-C. (2004). An integrated data mining and behavioral scoring model for analysing bank customers. *Expert Systems with Applications 27*(4), 623–633.

Jo H., Han I., Lee H. (1997). Bankruptcy prediction using case-based reasoning, neural network and discriminant analysis. *Expert Systems with Applications 13*(2), 97–108.

Kumari, P. & Mishra, S. P. (2019). Analysis of Credit Card Fraud Detection Using Fusion Classifiers. Computational Intelligence in Data Mining. *Advances in Intelligent Systems and Computing* 711.

Li, Y., & Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics 8(*10),1756.

Lichman M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Louzada F., Ara A., Fernandes G. B. (2016). Classification methods applied to credit scoring: systematic review and overall comparison*. Surv Oper Res Manag Sci, 21*(2), 117-134.

Moscatelli, M., Parlapiano, F., Narizzano, S., Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications* 161, 113567.

Paliwal M., & Kumar U.A. (2009). Neural networks and statistical techniques: a review of applications. *Expert Systems with Applications 36(*1), 2–17.

Pendharkar P.C. (2005). A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers and Operations Research 32*(10), 2561–2582.

Porretta P., Letizia A., & Santoboni F. (2020). Credit risk management in bank: Impacts of IFRS 9 and Basel 3. *Risk Governance and Control: Financial Markets & Institutions*, *10*(2), 29-44.

Tse L. (2020). Credit Risk Dataset. *Kaggle*. Accessed 15 April 2021, <https://www.kaggle.com/datasets/laotse/credit-risk-dataset?select=credit_risk_dataset.csv>

Vapnik, V. (1998). Statistical Learning Theory.

Wang G., Ma J., Huang L., & Xu K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61-68.

West J., & Bhattacharya M. (2016). An investigation on experimental issues in financial fraud mining. In *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)* (1796-1801). IEEE.

Yan H., & Lin S. (2019). New Trend in Fintech: Research on Artificial Intelligence Model Interpretability in Financial Fields. *Open Journal of Applied Sciences*, *9*(10), 761.

Yap B. W., Ong S. H., & Husain N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274-13283.

Yeh I. C., & Lien C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473-2480.