Banking and Insurance Operations

Group Project

Credit Risk Modeling

Alberto di Salvo– E20191178

Beatriz Frazão– M20191149

Francisco Bettencourt – M20200350

Luiza Maia de Carvalho – M20201290

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

Table of Contents

Index of Figures

Index of Tables

1. Abstract

Due to the Global increase in competition, banks they seek to make able their services available to customers. Nowadays, Some other players in the financial sector lead by differentiation, while banks look for low cost leadership. That said, since the interest rates at the Index (Euribor) level are increasingly low and negative, it makes it impossible to obtain greater gains downstream, thus, it is increasingly evident that the levels of comparison, validation and acceptance for that a loan can be granted to a particular customer, need more scrutiny and better metrics. With a technologically evolved world, it is absolutely essential that entities exposed to risk, be able to assess with a higher degree of certainty, the possibility of a particular asset to devalue, through computational models and Machine Learning. Once these models are optimized and operational, the bank will be able to assign a certain probability of default to an individual, and through a selection method, affirm his suitability. This project intends to explain the purpose of using a risk model for the approval of loans to customers.

Keywords: Low-cost, Interest rates, metrics, Computational models and Machine learning, Probability of Default, Suitability.

2. Introduction

The purpose of this project is to use the knowledge gained in class and, along with the curse material design and implement a method to calculate an individual probability of default.

We will start with the exploratory data analysis, where we obtain a holistic understanding of all the data items and the assurance that the data is fit for modelling purposes.

Then we will use two different methodologies to calculate a probability of default of an individual and how we should use this information to know if they are suitable for the loan. The first methodology developed and applied is a binary logistic regression model, and then, we decide to create a decision tree, with the view to concluding which model will be more suitable for decision making. In the final chapter, the conclusions of our group project will be presented.

3. Exploratory data analysis

According to Coutinho and Miguel (2008), the exploratory data analysis is the study of data from all views and with all tools needed. With the objective to obtain all possible knowledge. In this project, we will use techniques to input missing values, remove outliers, calculate the correlation between the variables and identify their distribution.

Initially, the dataset provided was composed of 28 variables where 19 were factor, and 9 were numerical (figure 1).

A factor variable is a categorical variable that can be used in statistical models, once it has been appropriately transformed. It was quickly noted that some variables were not being read in correctly, so their type was changed accordingly, including converting some variables to date format.

```
Number of rows             1048575
Number of columns          28
_______________________
Column type frequency:
  factor                   19
  numeric                  9
```

Figure 1 – Dataset Information

The 28 different variables on the dataset are listed below (Table 1).

VARIABLES	VARIABLE TYPE
Id	Categorical
Loan_amnt	Numerical
Funded_amnt	Numerical
Funded_amnt_inv	Numerical
Term	Categorical
Int_rate	Numerical
Instalment	Numerical
Grade	Categorical
Emp_title	Categorical
Home_ownership	Categorical
Annual_inc	Numerical
verification_status	Categorical
Issue_d	Date
Emp_lenght	Categorical
Loan_status	Categorical
Purpose	Categorical
Addr_state	Categorical
Dti	Numerical
Delinq_2yrs	Numerical
Earliest_cr_line	Date
Inq_last_6mths	Numerical
Open_acc	Numerical
Pub_rec	Numerical
Revol_bal	Numerical
Revol_util	Numerical
Total_acc	Numerical
Out_prncp	Numerical
Total_pymnt	Numerical

Table 1 – Dataset Variables Description

3.1.1. Missing Values

Whilst analysing the data we identified that the variable “emp_title” was a factor variable with too many different values that were not crucial to be used in the predictive models. Therefore we decided to remove it.

The first step on the exploratory analysis was to identify and treat the missing values. In the dataset used, we identified many missing values. For numerical variables suche as “annual income”, “dti”, “revol_util”, “inq_last_6month” - we decided to replace the missing values with the series mean. It means the average of all values of each variable, excluding the observations with missing values for each set.

For categorical variables we could have replaced the missing values by the category with the highest frequency, but we decided to remove them instead.

3.1.2. Distribution

To understand the variable frequency distribution, we ploted the histogram of each numerical variable, first with the bin width of 10 and the second analysis with a bin width of 4. The histogram allowus to understand continuous variables and the inspection of distribution and skewness. The charts below are the variables with a normal distribution.

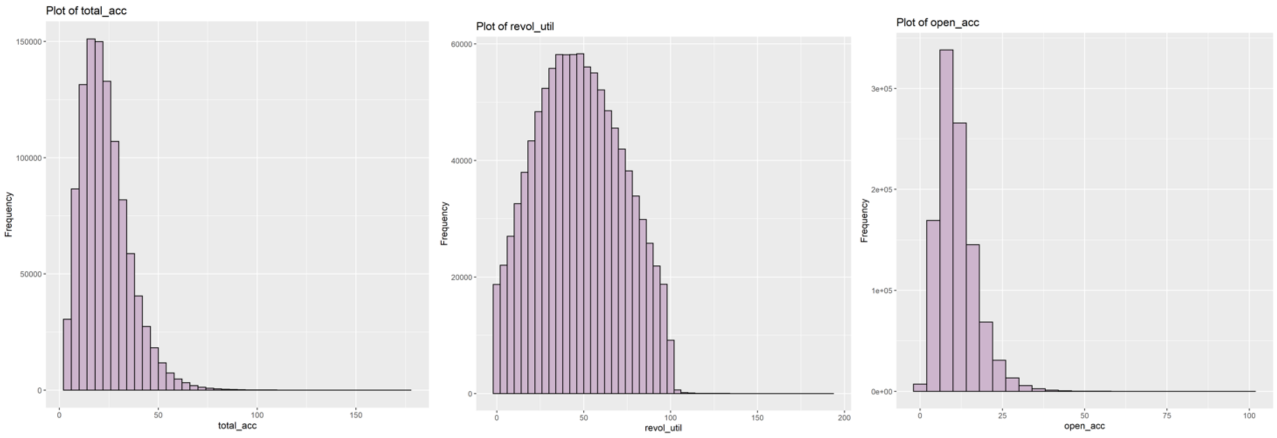


Figure 2 - Histogram most normal distributed variables

3.1.3. Outliers

To identify the outliers, we first split the dataset into numerical and categorical and removed the ID from the numerical dataset. Then we plot a boxplot of each variable to identify the values that were out with the boundaries for the boxplot, given by (Q1 – 1.5*IQR) and (Q3 + 1.5*IQR), where Q1 and Q3 represent quartile 1 and quartile 3, respectively, and IQR represents the inter quartile range (Q3-Q1).

The box plot shows that in this case the outliers are on high values instead of on low values. Therefore, we calculated the minimum value to be considered as an outlier to identify the numbers of rows with outliers.

The variables that have more than 5% of outliers were deleted. They were “delinq_2years”, “pub_rec”, and “revolt_bal”. for the other variables we removed the rows with outliers.

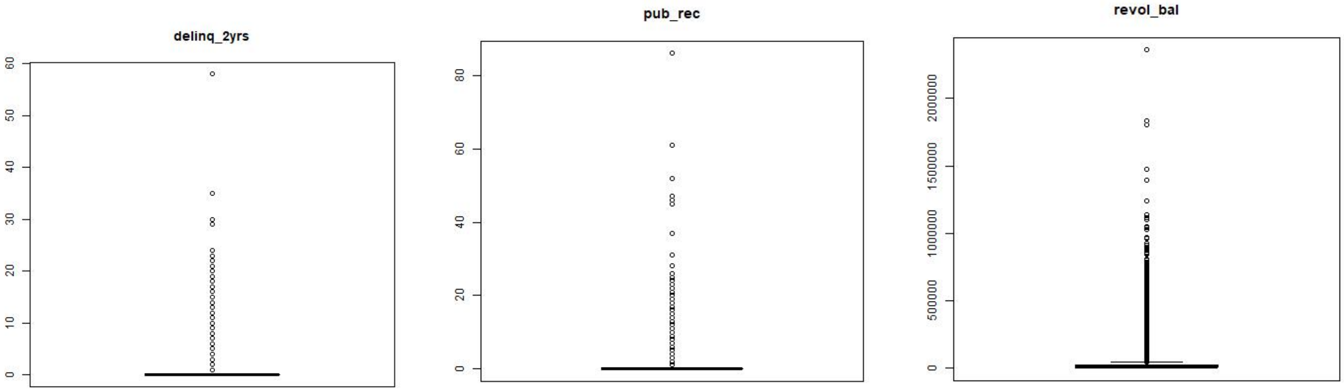


Figure 3 – Boxplot with more than 5% outliers

3.1.4. Correlation

After treating the missing values and removing the outliers, we analyzed the correlation between the variables to understand how they are linearly related to each other. To calculate the correlation is necessary to have quantitative variable. Thus, we used only the numerical variable on this step.

The heatmap below shows the degree of correlation. Correlation is a measure of association between any two numeric variables and it varies from -1 to 1, where both -1 and 1 represent perfect correlation, albeit -1 represents inverse correlation. In this case, we don't have a lot of dark blues squares, because there we don't have many negative correlated variables. We can see that the "loan_amnt", "instalment", "funded_amnt", "funded_amnt_inv" are highly positively correlated to each other (over 0.94 correlation factor). Hence, we decide to keep for the predictive model, in order to prevent multicollinearity, only one of those variables.

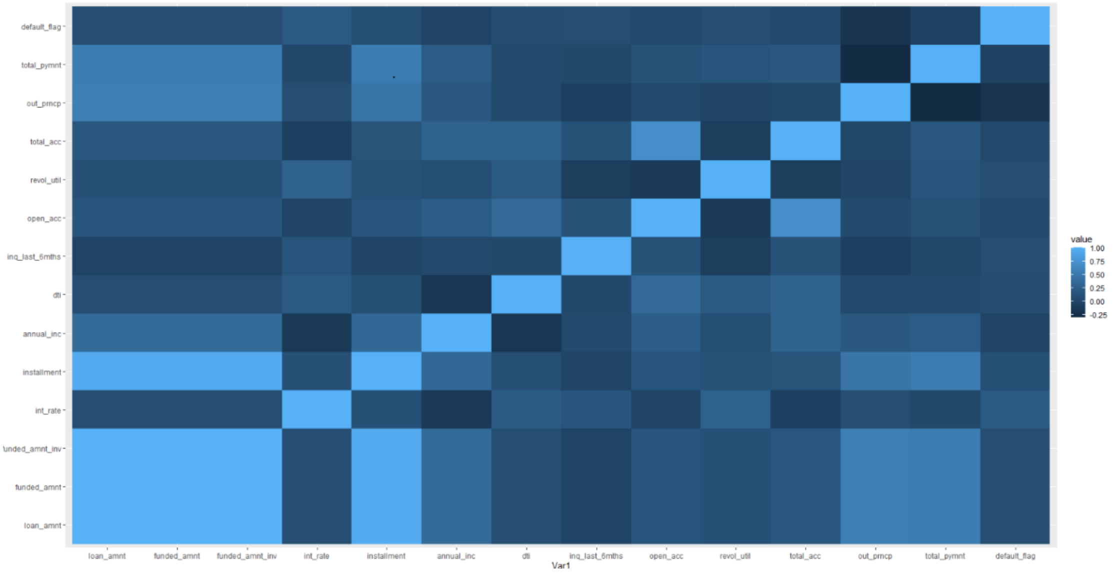


Figure 4-Correlation heatmap

3.1.5. Feature Engineering

The last step in the exploratory analysis was to create more variables to make the predictive model more robust. Therefore we used some variables to create others.

The first was the effort rate by dividing the instalment by the annual income. The second was the ratio between the number of open accounts and the total number of accounts. Then, the percentage of the loan that still left to be paid, and also, a new variable that allows to understand in which month the loan was issued.

The exploratory analysis enables the dataset to be ready for the predictive model. After all the action taken on the dataset, we built the final credit risk data model as described below, although not all variables were used in the model, as it will be explained in the modelling section.

```
                        Values
Name                       credit_risk_data_model
Number of rows             755577
Number of columns          25
_______________________
Column type frequency:
  Date                     2
  factor                   7
  numeric                  16
________________________
Group variables            None
```

Figure 5 - Dataset model Prepared

4. Modelling

After concluding the exploratory data analysis portion of the project, we now have a holistic understanding of all the data items and the assurance that the data is fit for modelling purposes. Therefore, we are ready to understand how we can, by using two different methodologies, calculate the probability of default of an individual, and how we should, then, use this information to inform if an individual is suitable or not, for a loan.

There are several data preparation steps and considerations to be had that are common to both methodologies and will therefore be outlined in this section before delving into the detail of each of the methodologies used.

A default flag was generated and was made up of a series of 0's and 1's (one value per individual) that indicated if an individual had (1) or had not (0) defaulted on their loan. The goal of applying both methods was to understand the set and weight of different characteristics (in the form of independent variables) that would indicate an individual's probability of Default.

When generating the default flag, it was immediately noted that the dataset was heavily biased towards individuals that had not defaulted on their loans (Table 2). Therefore, it was decided to use the entire dataset in the modelling process, to ensure that the information pertaining to the individuals that had defaulted was not diluted even further, causing an excessively overfitted final model.

Non-Default	Default
89%	11%

Table 2-Percentage of defaults vs non-defaults in the entire dataset

As with any modelling process where sufficient data exists, it is necessary to split the dataset into 3 subsets – train (60%), validation (20%) and test (20%) – since will assume the random number ("65748") that will work as our random seed, providing data integrity once that for every time we run the code, our model will never assume some rows from the training set, into the test set, that would compromise the entire decision.

By splitting our dataset into three subsets, we can quantify how overfitted (i.e., only able to explain data that has been "seen") is the final model. In practical terms, the training dataset is used to adjust the regression model. The validation set is used to calculate cut-off points, ROC curves and preliminary accuracy metrics and the test set is used to confirm the model on data never previously used. When calculating accuracy measures on the test dataset, we can evaluate the model performance vs its performance on the validation set. If there are significant discrepancies between the model performance across the different sets, there is a big possibility that the model is overfitted and not suitable to be used to calculate the probability of default for new individuals.

The first methodology developed and applied was a binary logistic regression model. Binary logistic regression is a statistical model that uses a logistic function to model a binary dependent variable, based on one or more predictive (or independent) variables.

In this type of model, the logarithm of the odds (Figure 6) for the value of interest (always labelled "1") is a linear combination of one or more independent variables.

$$Odds = \frac{\mathcal{P}(Y=1)}{\mathcal{P}(Y=0)} = \frac{p}{1-p}$$

Figure 6 - Odds calculation of a logistic regression

When using a logistic regression model, the independent variables can take any value in the continuous or binary space, and the output is always confined to represent a probability between 0 and 1 (Figure 7).

$$y = \frac{1}{1+e^{\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\ldots+\alpha_n x_n}}$$

Figure 7 - Logistic regression model with multiple covariables.

When adjusting a logistic regression model, three models are typically created to ensure the goodness of fit and validity of the final model (in this project also called step model): the null model, the saturated model and a model that typically sits in between the null and the saturated models called the reduced model. The null model consists of modelling the probability of default without including any predictors, i.e., it is model that assigns the same value of the probability of default to every individual, just based on the calculated intercept. The saturated model, on the other hand, takes into account every single predictor variable available within the model formula and offers a linear combination of all predictors to calculate the probability of default. The reduced model can be seen as an improved version of the saturated model, where redundant predictors are eliminated from the equation, and the coefficients of the remaining predictors are adjusted and optimized. More often than not, depending on the methodology used for model selection, the reduced model is the model with the lowest AIC (Akaike Information Criterion), which deals with the trade-off between goodness of fit and the simplicity of the model.

The final output of adjusting a logistic regression model via the reduced model method is a combination of predictors and their associated coefficient that will provide a probability, between 0 and 1, of default for each individual, via the equation presented in Figure 7.

When comparing the null, the saturated and the reduced model, the following values for AIC were recorded:

Saturated	Reduced	Null
127356	127351	306068

Figure 8 - AIC table for all models adjusted

So, as we may conclude the better and most suitable model, is the one who contains lower information loss, which means lower AIC and that is our reduced model.

We can also look at the log-likelihood to understand the goodness of fit of our models' coefficients. Log-likelihood values cannot be used alone as an index of fit because they are a function of sample size but can be used to compare the fit of different coefficients. Because we want to maximize the log-likelihood, the higher value is better. In this case, it looks like the saturated model has better coefficient fit, however the reduced model still has lower AIC so it's our chosen model.

Saturated model	Reduced model	Null model
-63719.27	-63728.74	-153032.8

Table 3 - Log-likelihood for the saturated, reduced and null models

By performing a Wald Test, similar to the T-test on linear regression, we can evaluate the statistical significance of each coefficient in the overall effect over the target variable, and it is given by the squared value of the regression coefficient, divided by the variance of the coefficients (Figure 9).

$$W_j = \frac{\beta_j^2}{var(\beta_j)}$$

Figure 9 - Wald Test formula

If the p-value is lower than 0.05, the overall effect of the variable is statistically significant. For our chosen model, i.e. the reduced model, the Wald test is significant meaning that there is evidence to say that the coefficients of the reduced model are statistically significant.

Chi-Squared	DF	P(> X2)
27051.0	3	0

Figure 10-Wald test results

Once the step model has been validated via the statistical tests performed, we can use the predictors and respective coefficients to calculate the probability of Default on all individuals in the validation set, using the formula outlined in Figure

7. This probability can then be compared to the actual observed value of the default flag, and a ROC (Receiver Operating Characteristic) curve can be drawn (Figure 11). The ROC curve offers a view into how the values predicted by the model compare to the actual observed models, by analyzing the false positive and the true positive rate.

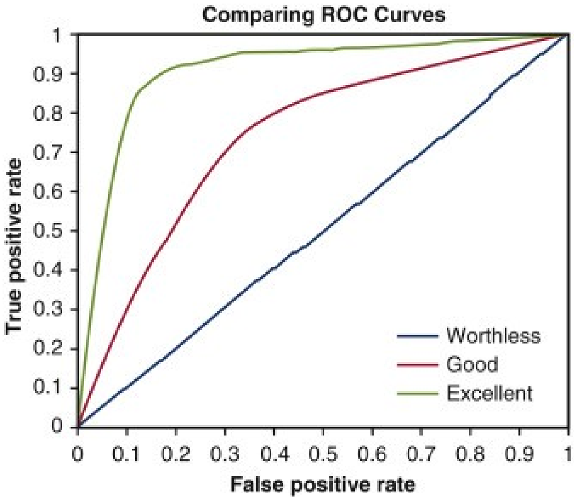


Figure 11 - Different ROC curves

However, calculating a probability of default is not enough, as we do not yet have an understanding on where to “cut-off” to say with some degree of certainty that an individual is or is not going to default. This “cut-off” point can be chosen in a number of ways, such as randomly choosing the value 0.5 and stating that every probability above 0.5 is considered to be default and vice-versa. However, as we are seeking maximum efficiency and statistical optimization, we are going to use Youden’s Index (also called Youden’s J statistic - Figure 12) to calculate the optimal cut-off, maximum tangent line, one that minimizes the false negative and the false positive rate.

$$J = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} + \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} - 1$$

Figure 12 - Youden’s J statistic

For our final reduced model, the ROC curve is shown in Figure 13. Furthermore, the optimal cut-off is deemed to be 0.1674 and the Area Under the Curve is 0.91.

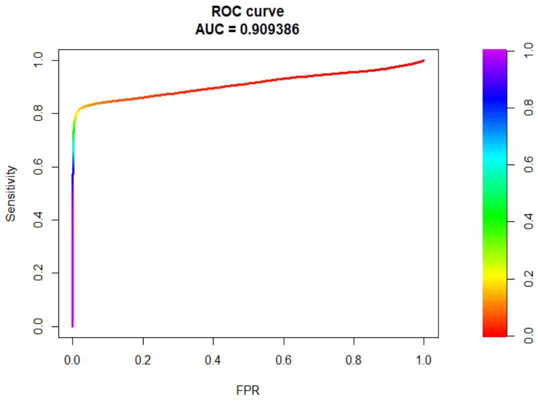
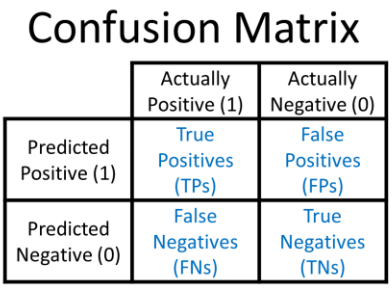


Figure 13- ROC Curve and AUC

Predictors	Value
optimal_cutpoint	0.1674
youden	0.7969
accuracy	0.9619
sensitivity	0.8179
specificity	0.979

Figure 14-Predictors table

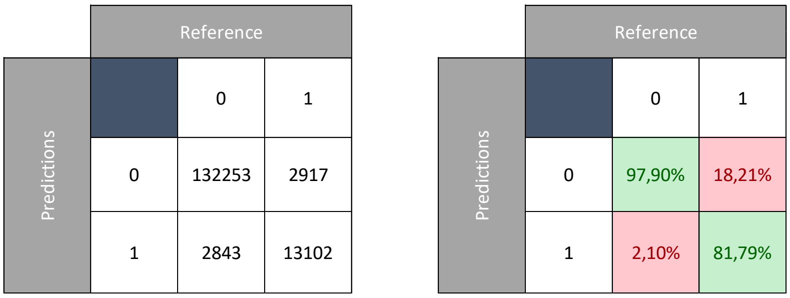
Once an optimal cut-off is calculated, a confusion matrix can be generated, from which false and true positive and negative rates, as well as accuracy, can be calculated.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 15-Confusion Matrix rules



		Reference	
Predictions		0	1
	0	132253	2917
	1	2843	13102

		Reference	
Predictions		0	1
	0	97,90%	18,21%
	1	2,10%	81,79%

Figure 16-Logistic Regression Confusion Matrix

As we are able to assume, an FNR with a high value such as 18,21% means that even using the best cut-off point for our model we would still concede 2917 loans that will be defaulted.

Regarding another approach, we decided to create a decision tree in order to conclude which model will be more suitable for our decision making.

A decision tree provides a highly effective structure within which can lay out options and investigate the possible outcomes of choosing those options.

They also help to form a balanced picture of the risks and rewards associated with each possible course of action. Basically, every time a decision comes to mind, a decision tree previews two possible outcomes (binary) and for each one is entailed a probability. After, if we multiply all the probabilities by the outcome that they generate, we should come out with a probability for every possible outcome, and then allocate the individuals, by their covariables values, to the same outcome, and then, decide which will be the ones who are not suitable for a loan.

Regarding our training data, we obtain the following Decision tree:

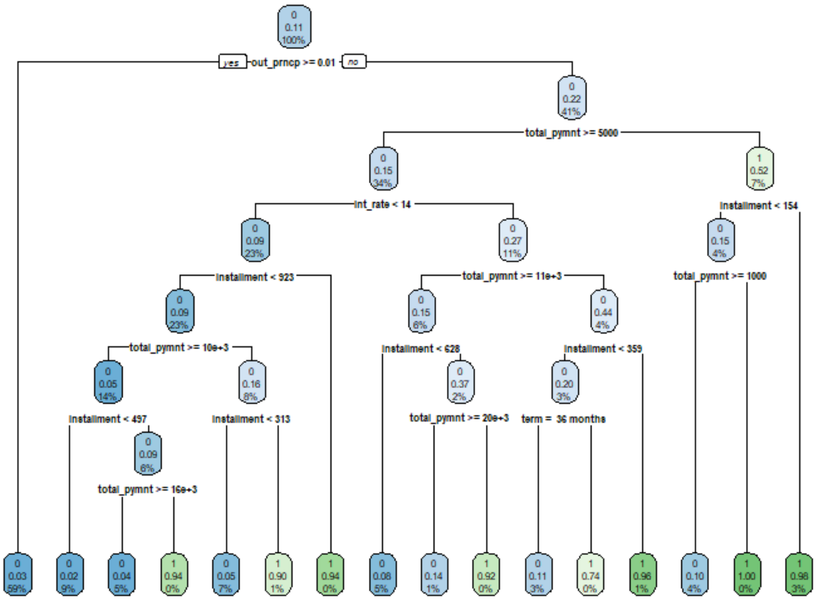
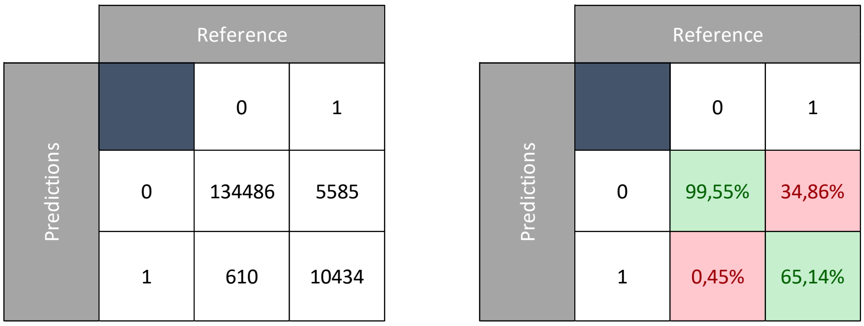


Figure 17-Decision Tree

As it is shown in Figure 17 if we assume the first branch, the model verifies for each individual if the value of out_prncp larger than 0.01, if it is we will automatically receive that probability default value that is 0.03%. When an individual does not meet demanded value, he should go through a new process of confirmation, until we demand a certain condition, and with that receive is a probability of default.

Once an optimal cut-off and the other main predictors are calculated, a confusion matrix can be generated, from which false and true positive and negative rates, as well as accuracy, can be calculated, all this for Validation data set.

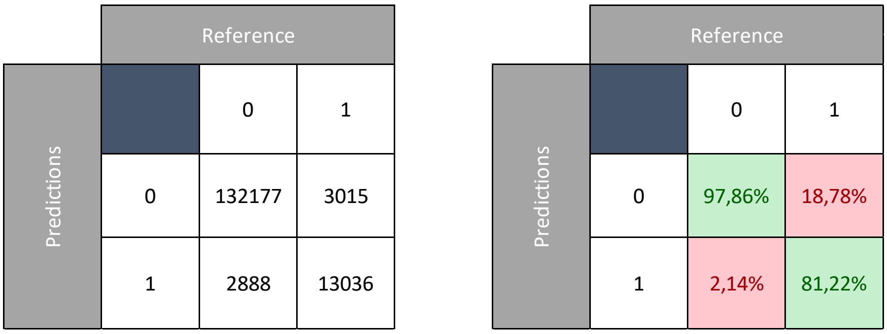


		Reference	
		0	1
Predictions	0	134486	5585
	1	610	10434

		Reference	
		0	1
Predictions	0	99,55%	34,86%
	1	0,45%	65,14%

Figure 18-Confusion Matrix Decision tree

Now that we already adjust the two models with the training data, calculate all predictors with the validation set, is now time, to understand how both models work with never seen data, test data set.
For the logistic regression model, we have the following Confusion Matrix:

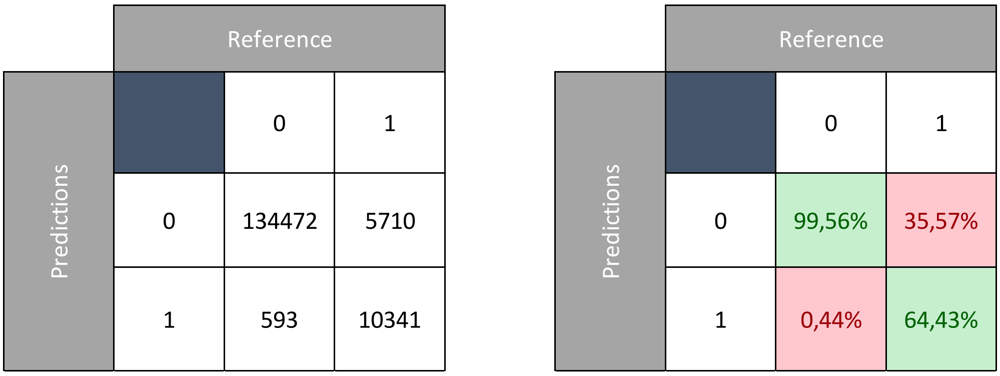


		Reference	
		0	1
Predictions	0	132177	3015
	1	2888	13036

		Reference	
		0	1
Predictions	0	97,86%	18,78%
	1	2,14%	81,22%

Figure 19-Logistic regression model, test confusion matrix

For the Decision tree model, we have the following Confusion Matrix:



		Reference	
		0	1
Predictions	0	134472	5710
	1	593	10341

		Reference	
		0	1
Predictions	0	99,56%	35,57%
	1	0,44%	64,43%

Figure 20-Decision Tree, test confusion matrix

5. Conclusion

The proposed work arises from the necessity of modelling the probability of default. We tried to develop different models which allowed us to find the probability of Default, and so the possible insolvency of some customers, in the context of credit risk management. We began, as usual, with the exploratory analysis: in this section, we have visualised the data set, and it is variables.

After the visualization and correction of the variables, we were in front of a double object represented by the selection of the variables to keep and to drop for the purpose of the model. After the detection of the worthiness of each variable, we treated the missing values, as them may be very important for the outcome of our model. In this regard, the missing values are usually replaced with the mean of the respective variable. Besides missing values, we decided to manage outliers too. For this purpose, we settle down a level of percentage (5%) beyond which any observation was not counted in the data set that was to be taken as input for the model. So after all these assumptions for the model, we developed a logistic regression for the variable" Default status". The main reason for which we have chosen the logistic regression is embodied in the dichotomous aspect of the" default" variable and in the dependence on the other variables(input data set). So after modelling several approaches, as shown before, we analyzed the differences between the confusion matrices resulting from the different methods. In any approach, we discovered a high level of False Negative Rate: since any "FNR" represent an unexpected loss for the bank, our goal would be to reduce as much as possible this eventuality. In this regard we could have been in front of a trade-off between False-positive and False Negative; every "FP" is also a loss in the balance of the bank since it does not let the new business increase. However, analyzing the rate of both, it seemed to us more appropriate to manage the FN not considering the FP instead. It is deemed that the logistic regression model is the most appropriated for this project as it provides the smallest FPR of both models. However, future work would be needed to try and improve the final model, in order to ensure maximum financial gains. Countless techniques and methodologies could be used, however there are three that we predict would yield positive results:

1- Deal with the bias in the training data towards non-default records.
2- Test the interaction between variables when adjusting the model.
3- Optimize the cut-off point in order to minimize FNR as opposed to the current trade off between FN and FP.

Bibliography

Cohen, Jacob and Cohen, P. (2002) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Routledg

Hosmer, David W. and Lemeshow, S. (2000) Applied Logistic Regression (2^{nd} ed.). John Wiley and Sons, Inc

Gomes, J. (2012) Notas de apoio à disciplina de modelos estatísticos.

Coutinho, Cileda and Miguel, Maria Inez. (2008) "Análise exploratória de dados: um estudo diagnóstico sobre concepções de professores." PUC-SP São Paulo. Brazil.

Filip Schouwenaars. (2018) Additional round of cleanup. https://www.rdocumentation.org/