Credit Risk Modelling

Diogo Rafael Dias Coreixas

Area: Industrial Engineering and Management, Instituto Superior Técnico

December, 2020

Setting a credit limit to companies is a major source of concern in financial risk management. The methods that calculate the credit limit have been developed as a response to the market requirements and changes. A good and appropriate practice of setting credit limits has a huge impact on financial institutions savings, once this results in a reduction of the potential companies which can default in their financial obligations. The present research aims to build reliable models by using different methods, namely statistical and AI technologies, which automatically define the credit limit of the companies, according to their financial data. Before developing the models, it was necessary to process the data, in order to minimize the negative effects of its inherent aspects, which may impair the research. Taking into account the methods used in this study, the main objective was to find out which one can develop the most accurate model. Considering the final results, the model that has revealed the best predictive performance was obtained by an AI technique. On the other hand, Multiplicative Models left strong evidences of being the analytical method that calculates the credit limit for companies, despite the developed model was not reliable.

Keywords: *Credit limit, Multiple Linear Regression, Multiplicative Models, Multilayer Perceptron, Radial Basis Function.*

1. Introduction

The credit concept has several meanings in the financial world. Despite of this concept is generally associated only to banks or credit institutions, by lending money, there also may be an exchange of goods and/or services in exchange for a deferred payment. In this way, credit can be described as a contractual agreement in which a borrower receives something of value, and agrees to repay it at a later date to the lender (Kenton, 2019).

The big problem for lenders of granting credits, is the possibility of borrowers not meeting their obligations, and default in the repayment of the credit, and this is one of the main concerns for financial institutions (lenders) related with financial risk management. In order to prevent the occurrence of this phenomenon, it's necessary to define the maximum economic value that a credit provider should provide to a borrower, considering its capacity to repay the loan, by analyzing its financial data. By setting credits, sellers provide to buyers the opportunity to buy goods or services and delay its payment, attracting costumers that are not able to pay at the moment of exchange of goods or services, by taking costumer's default risk, which represents the risk of the costumer not fulfilling the payment agreement (Lou & Wang, 2017).

In this way, the credit limit of a company must be seriously studied, once if a credit is denied, a potential profitable customer may end up in a competitor company. Considering these factors, it's necessary to evaluate the scenarios of a client default in its obligations, and other of losing a profitable client by denning a credit, when making decisions about granting a credit.

Over the years, given the importance of setting credit limits, there are several corporations that have been studying the best model to adopt, aiming to find the most appropriate credit limit, in order to ensure that borrowers have conditions to repay the credit (plus taxes or interests, depending on the lender), minimizing the losses of the lender. To build these models are used various techniques, among them there are statistical tools, feeding a predictive model with companies past financial data, seeking to define the effect or relationship between each variable (which consists

in a piece of financial information) and the credit limit, choosing which one fits better to include in the model. (Bazzi & Chamlal, 2015).

On the other hand, Artificial Intelligence (AI) methods can also be applied for the purpose of setting the most suitable credit limit, by using past financial data of companies, and through this, the computer can define the credit limit. Artificial Neural Networks (ANN) techniques (which is a branch of AI) allow to create models with high accuracy rates. This research uses two types of neural networks techniques: Multilayer Perceptron (MLP) and Radial Basis Function (RBF) (Méric, 2018).

2. Theoretical Background

2.1. Multiple Linear Regression

Multiple Linear Regression seeks to modelling the linear relationship between the dependent variable (the one which is intended to predict) and the independent variables, and these last ones explain the variation of the dependent variable. This technique has been widely applied in credit limit studies, in order to search which financial indicators or data are relevant when setting when setting a credit (Abu Bakar & Mohd Tahir, 2009). The general MLR model describes the relationship between k independent variables, X_j, and dependent variable, $\widehat{Y}_i$, as in the following equation:

$$\widehat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mathcal{E}_i \quad (1)$$

There are $p=(k+1)$ parameters β_j, $j=0,1,\ldots, k$, these values are regression coefficients, representing the effect on the dependent variable $\widehat{Y}_i$, of changing one unit in each explanatory variable in the model. $\mathcal{E}_i$ is the random error associated to each group of data, representing the possible difference between the observed and the predicted values.

In order to reach the model that fits better according the data, it was used de Least-Square method, which aims to minimize the sum of the residuals squared, being the residual the difference between the observed and the predicted value (A. Marill, 2004). Being n the number of observations and e_i the residual of each observation, the Sum of Squares Error/Residuals (SS_E), which is the squared differences between the observed and the predicted value, is calculated through:

$$SS_E = \sum_{i=1}^{n} e_i^2 \quad (2)$$

There are several methods to assess the suitability of the model and R^2 (coefficient of determination) appears as one of the most widely used statistical tool, assessing the goodness of fit of the model. This coefficient estimates the proportion of the dependent variable that's explained through the regression of all predictors in the model (Renaud & Victoria-Feser, 2010).

2.2. Multiplicative Models

Logarithmic transformations are often recommended for skewed data, such as monetary measures. This technique generally has the effect of spreading out clumps of data and bringing together spread-out data. There are several bases, and in the present research is considered the decimal logarithm of base 10, once there is a big range in term of credit limit, or financial data. If the range of the dependent variable were smaller it can be used a lower logarithm base, in order to avoid a loss of resolution, which is what happens when are considered higher bases than what is desirable. For example, the credit limit ranges between 0 and millions of euros, which is extremely wide, so a higher logarithm base should be applied (Osborne, 2002).

In order to handle situations where occur a non-linear relationship between explanatory variable and dependent variable in a regression model, one of the most common way is to logarithmically transforming variables. Using this transformation allows to preserve the linear model, even if the relationship between variables is non-linear (Benoit, 2011). By applying a logarithmic transformation on equation 1, and considering the logarithmic properties, the model can be described as:

$$\widehat{Y}_i = \beta_0 {X_1}^{\beta_1} {X_2}^{\beta_2} \ldots {X_k}^{\beta_k} \quad (3)$$

In this way, using the MM approach, independent variables have a multiplicative relationship with the dependent variable instead of the usual additive relationship of the MLR.

Performing a linear analysis with logarithmic transformation, the regression coefficients β_j, can be analyzed in a different way than in the traditional method. Applying this methodology, the coefficients are considered as elasticities of each independent variable included in the model. The elasticity gives the expected percentage of change in the dependent variable $\widehat{Y}_i$, when changing a certain dependent variable X_k. So, when increasing Z in X_k, $\widehat{Y}_i$ will suffer an increase of $\beta_k Z$ (Benoit, 2011). The methodology used to develop a model by MM method is the same that used in MLR method, assessing the same parameters.

2.3. Artificial Neural Network

Artificial Neural Networks are inserted in the Deep Learning models, and have the objective to study regressions and classifications (Dotai, 2015). The units which have the role of processing information are called neurons, and each neural network is composed by several neurons. Each neuron is connected with the others, and the connection between units are synapses which are really just weighted values, being that each connection has an assigned weight, and the higher the number, the greater influence one unit has on another (Marr, 2018).

In each network there are three layers, the first one is the input layer, where the data is inserted in the network, the second is the hidden layer, which is responsible for data processing, as a human brain, and at last the output layer which is the final results of the study. Another important concept is the transfer/activation function, which converts the input signals to output signals through mathematical equations in each neuron. In other words, this function decides whether a neuron should be activated or not by calculating the weighted sum. Without transfer function the model would not be able to recognize patterns, and learn from the training sample (Singh Chauhan, 2019).

The MLP is the first AI method considered in this research, and one of the most ANN used techniques in the credit limit study, and over the years has been tested in several studies. In this technique there are a number of arbitrary hidden layers, where the higher the number of layers, the more complex the system will be (Kang, 2017). For the adjustment of the parameters, it's used the backpropagation algorithm, which basically, initializes the model with random weights values, and through an error measurement these values are refined, once the values change towards a reduction of the of the overall error of the network. With backpropagation, the weights adjustment of the inputs is simplified, given that this technique allows the performance of backward passes which attempt to minimize the difference between the real and the predicted values. The conservation of the transfer function variance it's a desirable property, once this allows information to flow well upward and downward in the network (Nicholson, 2017).

The other artificial neural network that is tested is RBF, and in this network there is only one hidden layer, ensuring a faster learning speed. This feedforward method can be described as: *"Each hidden input represents a particular point in input space, and its output, or activation, for a given instance depends on the distance between its point and the instance- which is just another point. Intuitively, the closer these two points, the stronger the activation"* (Witten & Pal, 2017).

Several RBF have been studied, but the most widely used is the Gaussian type, once it's not only suitable in generalizing a global mapping, but also in refining local features without changing the already learned mapping (Sadeghkhani, Ketabi, & Feuillet, 2012). The output of RBF is dependent on three parameters: the input vector, the center and the width of the respective neuron. The input vector is defined for having one dimension for each explanatory variable, and the center of the neuron corresponds to a point with as many dimensions as the input vector. In this way, the similarity of these two vectors must be analyzed through the distance between them. Finally, the spread controls the smoothness of the drop seen in the function, which evaluates the distance between input and center vectors, for greater distances (Militký, 2011).

3. Input data Collection, Analysis and Treatment

3.1. Input data collection process

In order to perform the present research, it's necessary a large amount of financial data related with thousands of companies. The data used in this study was obtained from one of the biggest financial database in the world. This database is widely used by private institutions, governments and financial worldwide, seeking to assess the credit risk, having access to financial information about others companies, that could be a future business partner.

In the data are included companies headquartered in Spain or Portugal, containing several financial ratios and information, risk class and main business sector. It's essential to access to the maximum financial data possible, aiming to determine the influence of each variable on the credit limit. One less variable, could mean the loss of an important piece for the credit limit calculation. In this way, the greater the number of the variables, the greater the probability of reaching a model capable of calculating the credit limit. The initial data set, without any kind of treatment, displays about 21700 corporations, with 43 financial indicators.

3.2. Input data analysis

The input data has a diversified set of information about each company, covering different areas of the financial sector, which

includes: raw financial data, structural indicators, operational indicators, profitability indicators, and qualitative indicators (such as risk class, main business sector, among others). The status of the corporation is mentioned, having six hypotheses: active, bankruptcy, dissolved, in liquidation, inactive or status unknown. Generally, only active corporations have an associated credit limit, which means that this type of companies are the main focus of this research.

In order to perform a good research, it's required a good understanding of the data, given that without this it's impossible to assess whether the results make sense. In this way, it's crucial to understand which variables are more likely to be important to build the model that define the credit limit.

It should be noted that risk class classification of the companies has suffered an alteration, since this variable is translated by letters. So, instead of using letters, it was transformed to numbers, where the best class is represented by a 10 and the worst class by a 1.

3.3. Removal of invalid cases

Before performing any type of study, it's important to analyze the data set, searching for cases that may jeopardize all research. Were excluded from the present study, cases that are included on the following scenarios:

- Cases that didn't have an associated credit limit, due to mathematical errors, conversion errors between softwares, among others;
- Companies with an associated credit limit of 0 were also excluded from this analysis, given that these cases may jeopardize the statistical models, because it's the minimum threshold of the monetary unit, and there is no differentiation between cases that may have completely different financial data;
- Companies which have the following main business sectors were not considered in this research: Public administration and defense, Compulsory Social Security, Education and Activities of Extraterritorial Organizations and Bodies. These cases represent entities that are financed by governments, and these types of companies don't usually become insolvent because they have State aids, something that doesn't happen with private companies.

3.4. Correlation analysis

Correlation analysis is a statistical method which aims to evaluate the strength of the relationship between two quantitative variables. The higher the correlation between two variables, the stronger is the relation between these two, while a weak correlation indicates that there is no relationship with each other (Franzese & Iuliano, 2019). The introduction of correlated variables may affect the model, once it's difficult for the model to estimate the relationship between each independent variable and the dependent one independently because the independent variables tend to act in accordance with each other. So, if two correlated variables are inserted in the model, at least one of them aren't adding value to the model, so must be excluded (Rogers & Boyd Enders, 2013).

The simplest and fastest technique to analyze the presence of correlated variables, is to analyze the variance inflation factor (VIF) for each variable. If there are signals of collinearity, this measure can detect the degree of the multicollinearity between the variable in question and the remaining independent variables. The VIF may be calculated by the following expression:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad (4)$$

Where R_j^2 is the multiple correlation coefficient, which gives the proportion of variance in the independent variable j associated with the remaining independent variables. Values of VIF exceeding 10 are often source of concern, indicating multicollinearity. A VIF value of 10 implies a R_j^2 equal to 0,9, which means that 90% of the variability of the variable j is explained by the remaining independent variables.

In order to solve collinearity problems, it was adopted the following methodology: the variables were removed iteratively until no VIF values were over 8. By adopting a limit of 8, a comfortable margin is being given for variables that have similar behaviors but don't depend on others, once it's normal for certain financial indicators to behave in a similar way but not to depend on one another, such as revenue and assets of the company. By applying this methodology were removed 14 variables from the research. After the removal of these variables, no VIF value is above the established limit of 8, mitigating correlation problems present in the initial data set.

3.5. Sampling Method

The data set must be divided mainly in two groups: training and test samples. Where the training group is responsible for adjusting the models' parameters to make them as accurate as

possible, while test set aims to evaluate impartially the model fit on the training dataset, and it's only used when the model is completely trained. Regarding the cases, these two samples must be equal for all approaches and methods performed in this research, in order to guarantee the same conditions for all techniques.

To ensure a fair assessment and comparison, it's also necessary to diversify the two samples, to cover all type of cases. In this way, it's mandatory to include several classes of cases in these samples, seeking to create balanced samples. For instance, if the training sample contains mainly cases of companies with an associated credit limit higher than 100 thousand euros, and the test sample only includes companies with an associated credit limit lower than 100 thousand euros, the assessment of this method would be a bit pointless once samples are not balanced.

The main limitation when defining the two types of samples is the possibility of some cases may be only available for certain models, in accordance with the variables which are included in these models due to missing values present in data set. The cases available depend largely on the variables that are considered in the respective model, once if the model includes a certain independent variable, all cases that don't have a valid value for this variable are not considered.

Despite this limitation, it was defined about 15% of the data set as the test sample, and the remaining 85% the training sample, being each sample well balanced, covering all classes of cases in the same way.

4. Model Building

4.1. Multiple Linear Regression

One of the most important prerequisites before applying MLR study is to have a detailed knowledge of each variable that may enter in the model. In addition to this, it's necessary to have a notion about the potential relationship between the dependent variable and the explanatory ones, such as if certain variable has a negative or positive correlation with the credit limit. In this way, it were defined all the signals of correlations between the dependent variable and the independent variables.

Another important aspect that should be previously defined is the approach to develop models by using MLR method. The adopted methodology is based on an iterative trial/error process, through the insertion and/or removal of variables from the model. In each iteration, at least one variable is inserted in the model, and excluded or insignificant variables of the previous iteration are removed. Aiming to facilitate the analysis of the variables it was defined the Stepwise method selection, which means that variables that don't contribute to credit limit definition are excluded from the model. On the other hand, the significance of the variables is analyzed through the t-value of each variable included in the model. The basic principle of this process is to assess the R^2 after each iteration, and check whether all correlation signals are correct.

In order to perform a robust analysis, it was necessary to develop several models, being the R^2 of each one the main indicator of the quality of the adjustment of the model with the credit limit. If a model shows evidence of adjustment to the model which calculates the credit limit, it was considered that it must have an R^2 value of at least 75%.

One very important aspect when applying the Stepwise selection method is the stepping method criteria. Variables can be entered or removed from the model depending on the significance of the F value, which is an arbitrary value, depending on the specifications defined for the inclusion or not of variables in the model (Pope & Webster, 1972). The entry and removal value of F value were the defined ones by default in IBM SPSS Statistics 25, 3,84 and 2,71, respectively.

It were developed several MLR models, but two (MLR-1 and MLR-2) stood out from the others, despite the low R^2 values.

	MLR-1	MLR-2
R^2_{Adj}	65.5%	72.3%
F	1 601	1 329
N	5 895	3 562

Table 1- Statistical results of MLR models.

As it can be seen in table 1, none of the developed models contain an R^2 value greater than 75%, having no clear evidence that this method fits into the calculation on the credit limit. It should also be noted that MLR-1 contains more observations than MLR-2 model due to the fact that the last includes more independent variables, which results in a decrease of observations caused by missing values. By analyzing the F value of the both models it becomes evident that these two models have similar values, which means that there is no model which has variables that are more significant than variables present in the other model in the credit limit calculation.

Considering table 2 which describes the codification of the variables:

Variables	Code
Probability of default (x_1)	x_1
Current Assets (x_2)	x_2
Shareholders' funds (x_3)	x_3
Operating Revenue (x_4)	x_4
Operating P/L [=EBIT] (x_5)	x_5
Profit margin (x_6)	x_6
Solvency ratio (x_7)	x_7
Long term debt (x_8)	x_8
Other non-current liabilities (x_9)	x_9

Table 2- Codification of the variables included in MLR models.

The equations of the models are described by the expressions below:

$$MLR1(X) = 536\,474 - 260\,608x_1 + 2{,}5x_2 + 3.4x_3 + 18.7x_5 + 6\,529x_6 - 2.9x_8 - x_9 \quad (5)$$

$$MLR2(X) = 26251 + 2.4x_2 + 0.4x_3 + 0.5x_4 + 15.8x_5 + 5\,207x_6 + 2\,458x_7 - 0.9x_8 \quad (6)$$

The remaining variables available in the set were not included in both models, once don't showed evidences of being important for credit limit calculation. By analyzing the results of the models and considering the properties of the MLR study there is one major conclusion that has to be made. This regression has an additive character, so doesn't make much sense to sum different ratios, once this type of information doesn't take into consideration the dimension of the company.

Lastly, it was performed a residual analysis in order to check whether the four assumptions are verified when applying a linear study are met: be approximately normally distributed, have a mean of 0 and a constant variance, and be independent of one another. By checking these assumptions it was concluded that residuals from MLR models don't meet the first two conditions. In this way, MLR models can´t be validated, once these ones don´t satisfy all conditions, which may indicate a poor adjustment to the credit limit calculation.

4.2. Multiplicative Models

The Multiplicative Models was the second and last statistical method implemented in this research. Regarding MM, the adopted approach to find the best model to calculate the credit limit and the requirements are the same than the ones used in the previous method. The big difference between MM and MLR is the interpretation of the results, once the studies are conducted in the same way by using IBM SPSS Statistics 25.

While in MLR, the coefficients of the regressions relate to the parameters of the linear regression, in MM these coefficients are interpreted as elasticities. However, the requisites are the same when analyzing each model, where each variable included in the model must have an associated elasticity that respects the respective signal of correlation, which are equal as correlation signal of in MLR method.

In this way it were developed two MM models (MM-1 and MM-2) that have shown strong evidences of fit with the model that calculates the credit limit, being that one of the major concerns is the inclusion of variables in the model that may drastically decrease the sample of cases available to develop the model due to missing values present in the data.

	MM-1	MM-2
R^2_{Adj}	93.1%	92.2%
F	10 558	28 156
N	3 913	7 099

Table 3- Statistical results of MM models.

As it can be seen in table 3, the values of R^2_{Adj} in both models are quite promising, once the threshold of 75% for this coefficient is largely exceeded. On the other hand, MM-2 has a F value much higher than MM-1, which may mean the presence of more significant variables in the model.

The difference between these two models is the exclusion of the variable "Solvency ratio" from the model, which consequently led to the exclusion of the variable "Operating P/L [=EBIT]". This ratio it was excluded because it contains several missing values, jeopardizing the sample responsible for model's development. As it can be verified, MM-1 model (which contains this variable), has much less observations than MM-2.

Taking into account the codification listed in table 2, these two models are described by the following expressions:

$$MM1(X) = 0{,}755 \times x_2^{0{,}368} \times x_3^{0{,}184} \times x_4^{0{,}324} \times x_5^{0{,}063} \times x_7^{0{,}495} \quad (7)$$

$$MM2(X) = 1{,}652 \times x_2^{0{,}108} \times x_3^{0{,}675} \times x_4^{0{,}182} \quad (8)$$

The big advantage of this method when compared to MLR is the possibility of including ratios and raw financial data in the same model, given that makes sense to multiply these two types

of financial information, however, doesn't make any sense to sum them up.

Given that this method is performed by using a Multiple Linear Regression procedure but with data transformation, it's also important to perform a residual analysis, to check whether the four assumptions about residuals are verified, as explained in MLR method. In this method all models were validated once all of them show evidences of respecting the four conditions.

4.3. Multilayer Perceptron

When applying an AI method, the big concern is to adjust all the configurations and options, to be programmed in the context of the study to be conducted. If these conditions are not meet the study may be impaired, giving completely unsatisfactory results. The stopping rule that was considered was the maximum number of steps without a decrease in error, 8 in specific.

One very important aspect that must be considered in MLP is the type of training, which determines how the network processes the records. In this step it was considered the Online training once this type of training has the desired characteristics regarding the conditions of this research. The Online training, besides having a better performance when the study includes several inputs, and guarantees constant case-by-case updating, Online training type also provides a chance to reuse the available data which represents a big advantage compared with the others training types, once the same observation may be adjusted more than one time, improving the adjustment of the model (IBM, 2019).

Regarding the architecture of the MLP, there are several options to select. First, the number of hidden layers, which in the software only can be one or two. As it is initially not known which of these options is the best, it's important to test both through the development of two models (one with one hidden layer, and other with two), and in the end assess which one guarantees the best accuracy. The number of units in each hidden layer is set automatically by the software, which means that the software calculates the ideal number of units for each layer.

Beyond the number of hidden layers, it's necessary to define the activation function in these layers and there are two options: Hyperbolic tangent or Sigmoid. Once again, there is no certainty of the best option, which leads to the development of two models, each with one option. About the output layer it's only necessary to define the activation function, being that this choice it's simpler comparing with the hidden layers. As it known, when the aim is to represent a model which describes a regression (as it happen in credit limit calculation) the most adequate activation function is the Linear/Identity function, being the most used function, which creates an output signal proportional to the input multiplied by the weights for each neuron (Sharma, 2017)

Model	Number of hidden layers	Activation function in hidden layer
MLP-1HT	1	Hyperbolic tangent
MLP-1S	1	Sigmoid
MLP-2HT	2	Hyperbolic tangent
MLP-2S	2	Sigmoid

Table 4- MLP models' details.

Considering all the configurations and options mentioned, all the conditions for the development of models are met, resulting in the models described in table 5.

	MLP Models	
	One-layer models	Two-layers models
Number of units (input layer)	24	
Number of units (hidden layers)	9	9 (1st layer) / 7 (2nd layer)
Training Sample	2505	
Testing Sample	438	

Table 5- Summary of the four MLP models.

The number of observations available to develop the model is a cause of concern, because, as it can be seen in table 5, there are only about 3000 observations to train and test the model. When applying an AI method, it's important to take as much data as possible in order to "teach" the model well. Due to this fact, the analysis of the importance of each variable in the model arises as a method to reduce the number of variables included in the study, being that variables with a low importance in the model can be excluded from the study. Therefore, the criterion was defined that variables with a normalized importance below 10% in these 4 models must be removed from the study. In this way, by setting a minimum threshold for normalized importance of 10%, variables which have an insignificant impact will be removed from the generated model.

It were removed 8 independent variables, and is expected that the sample for model development

increase, once the higher the number of variables inserted in the study the lower the number of observations is, due to the missing values present in the data.

As a result of the implementation of this measure is necessary to develop again more 4 updated MLP models, and analyze whether the option of privileging the number of observations by removing independent variables is beneficial for the accuracy of the model.

	Updated MLP Models	
	One-layer models	Two-layers models
Number of units (input layer)	16	
Number of units (hidden layers)	8	8 (1st layer) / 6 (2nd layer)
Training Sample	5369	
Testing Sample	925	

Table 6- Summary of the four updated MLP models.

With the removal of the 8 independent variables, it's notorious the increase of the observations when compared to preceding models. These updated models contain over double the observations of previous models, which is a big difference.

4.4. Radial Basis Function

As made with MLP method, when applying an RBF model, it's necessary to define several running settings. All the configurations were the same in both AI methods, such as stopping method criterion, among others.

Regarding the architecture of the RBF model, it's necessary to define the activation function for the hidden layer. There are two options for this setting in IBM SPSS Statistics 25: Normalized radial basis function and Ordinary radial basis function. As it happened with the MLP method, there is no certainty about which of these is the best option, so two models will be developed, one with each option, to test both scenarios, and find out which one ensures a better accuracy.

Model	Number of hidden units	Activation function in hidden layer
RBF-N	Set automatically	Normalized RBF
RBF-O	Set automatically	Ordinary RBF

Table 7- RBF models' details.

Given that all the customizable settings were defined, the models were created, producing the results described in table 8.

Once these models will have the same number of observations than MLP models, was produced more two updated models, where in these models don't consider the 8 independent variables that were excluded in updated MLP models, in order to increase the number of available observations, due to missing values, since there are no variables with a normalized importance of less than 10% in both root models (being the root models those which consider all variables).

The next table (table 8) describes the characteristics of all RBF models, the root and the updated models.

	RBF Models		
	RBF-N and RBF-O	RBF-UPD-N	RBF-UPD-O
Number of units (input layer)	24	16	
Number of units (hidden layers)	10	9	10
Training Sample	2 505	5 369	
Testing Sample	438	925	

Table 8- Summary of all RBF models.

After the removal of eight independent variables, the sample number of observations more than doubled, as it happened in updated MLP models. As obvious, the number of units in the input layer also has decreased, once in root models are considered 24 independent variables and in updated models are considered 16.

5. Comparison between models

Aiming to find out which developed model is the most accurate it was performed a comparison among all models considering only the testing sample (the sample that was not used to build the models) representing 15% of the total data set, as already mentioned in section 3.5.

Initially, it were made comparisons between models which were developed by the same method, once the larger the number of models included in the comparison, the smaller the comparison sample size, Due to the presence of missing values in the data set, some observations are not available for all models, depending on the variables included in these models. It's important to determine the most accurate model through clear evidences, by analyzing the R^2 (being the most important indicator) and the error rate of each model. So, in each comparison between models

developed by the same methods are only considered observations that all models have in common.

	Statistics	
Models	Sample size	R^2
MLR-1	**616**	**83.6**
MLR-2	616	75.8
MM-1	**690**	**94.1**
MM-2	690	91.2
MLP-1HT	438	92.8
MLP-1S	438	92.2
MLP-2HT	438	89.0
MLP-2S	438	95.7
MLP-UPD-1HT	438	90.5
MLP-UPD-1S	438	94.9
MLP-UPD-2HT	438	90.8
MLP-UPD-2S	**438**	**95.4**
RBF-N	**438**	**55.8**
RBF-O	438	27.4
RBF-UPD-N	438	13.8
RBF-UPD-O	438	27.6

Table 9- R^2 and sample size of the comparison for each model.

As shown in table 9, the models that stand out the most are those developed by MM and MLP methods. The models in bold are those considered to be the most accurate, considering models developed by the same method. It should be noted that MLP-UPD-2S was considered more accurate than MLP-2S despite having a lower R^2, since the difference between the two values is minimal, and the first has better error rates.

Lastly, a final comparison was performed considering only the four models highlighted in the previous table, where each model presents the maximum number of observations, not only the ones that all have in common, as made in the last comparison.

	Statistics	
Models	Sample size	R^2
MLR-1	1 019	72
MM-1	690	94.1
MLP-UPD-2S	925	95.4
RBF-N	438	55.8

Table 10- Models' R^2 and sample size for the final comparison.

By analyzing the previous table (table 10), it can be concluded through a direct comparison that RBF-N and MLR-1 models have worse values for all parameters compared to MM-1 and MLP-UPD-2S models, which means that these two models can be out of equation for the most accurate model over all methods. Since the MLP-UPD-2S model has a slightly higher R^2 than the MM-1, having a higher number of observations, keeping the most accurate model selection criterion verified in the previous considerations, the MLP-UPD-2S was considered the most accurate model.

6. Conclusions

This research allowed the comparison between statistical and AI methods, adding value in the literate of the credit limit modelling.

The first phase of the practical part of the present study was based on the treatment of the data used in the study. This phase is crucial to ensure a credible and fair research, given than the data are the foundation of the study.

In alignment with the academic literature of this thesis, an MLP model was considered the most accurate model developed in this research, but MM-1 model left good evidences about the credit limit definition, which may be the analytical method to calculate it. So, MM and MLP methods stood out as the most suitable for defining the credit limit, while in the opposite direction appear the MLR and RBF methods, which are not suitable for calculating the credit limit.

Regarding limitations faced in the course of the research, the most influential should be highlighted: the presence of missing values in the data set. This occurrence has a great impact on the performance of the study, since it influences more than one aspect of this research. During the development of the models, special attention had to be paid to the trade-off between the variables considered in each model and the respective available observations. Given that there were variables in the database that were more susceptible to having many missing values, many of these variables may have been poorly exploited, as they reduce the available observations which are essential for model adjustment. The impact of this limitation is aggravated in the AI methods, because in these methods all variables should be considered, unlike statistical methods, where it is chosen which variables should enter in the model.

Another fact that should also be noted is that it's very likely that is needed more than one model to calculate the credit limit of companies, and a regression for each class of companies may be required, being that each class of companies may be defined according to one or more variables. Besides the fact that there is the possibility that several models may be needed to calculate the

credit limit (depending on the class of each company), each model associated with a class may contain different variables.

Since one of the major constraints in the study is the presence of many missing values, additional effort should be made to ensure as many cases as possible (without missing values). The more complete the database, the batter the quality of the study.

In a further study it would be advantageous to create several classes of companies, and define a model for each one, considering the same method, which would involve a lot of work and effort. There are two different approaches that can be adopted to define the classes of enterprises: in an arbitrary way, by analyzing the data, without a theoretical foundation or by using AI technology to detect patterns in data. It should be noted that this work only applies to the development of models by statistical methods, since AI techniques should be able to detect the presence of classes without any type of treatment.

Regarding the pre-processing of the data, which is a crucial element in the development of this type of studies, although several measures have been taken to ensure the quality of the database, in further studies in can be adopted other methodology to address some limitations experienced in the current research. Since one of the major constraints in the study is the presence of many missing values, additional effort should be made to ensure as many cases as possible.

7. References

A. Marill, K. (2004, January). Multiple Linear Regression. *Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression*, pp. 94-102.

Abu Bakar, N. M., & Mohd Tahir, I. (2009). Applying Multiple Linear Regression and Neural Network to Predict Bank Performance. *Internationa Business Research*, 176-183.

Bazzi, M., & Chamlal, H. (2015). Rating models and its Applications: Setting Credit Limits. *Journal of Applied Finance & Banking, Vol. 5, no. 5*, 201-216.

Benoit, K. (2011, March 17). Linear Regression Models with Logarithmic Transformations.

Dotai, J. (2015, December 28). *Everything you need to know about Artificial Neural Networks*. Retrieved from Medium: https://medium.com/technology-invention-and-more/everything-you-need-to-know-about-artificial-neural-networks-57fac18245a1

Franzese, M., & Iuliano, A. (2019). *Correlation Analysis*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/medicine-and-dentistry/correlation-analysis

IBM. (2019). *Training (Multilayer Perceptron)*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/neural_network/idh_idd_mlp_training.html

Kang, N. (2017, June 27). *Multi-Layer Neural Networks with Sigmoid Function- Deep Learning*. Retrieved from Towars Data Science: https://towardsdatascience.com/multi-layer-neural-networks-with-sigmoid-function-deep-learning-for-rookies-2-bf464f09eb7f

Kenton, W. (2019, June 4). *Credit & Debt*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/credit.asp

Lou, K.-R., & Wang, W.-C. (2017). Optimal trade credit and order quantity when trade credit impacts on both demand rate and default risk. *Journal of the Operational Research Society*.

Marr, B. (2018, September 24). *What are Artificial Neural Networks- A simple explanation for absolutely anyone*. Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2018/09/24/what-are-artificial-neural-networks-a-simple-explanation-for-absolutely-anyone/#1836f4e11245

Méric, S. (2018, October 12). *AI delivers a paradigm shift for credit management*. Retrieved from The Global Treasurer: https://www.theglobaltreasurer.com/2018/10/12/ai-delivers-a-paradigm-shift-for-credit-management/

Militký, J. (2011). *Fundamentals of soft models in textiles*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/chemical-engineering/radial-basis-function-networks

Nicholson, C. (2017). *A beginner's guide to Multilayer perceptrons (MLP)*. Retrieved from Pathmind: https://pathmind.com/wiki/multilayer-perceptron

Osborne, J. (2002). Pratical Assessment, Research, and Evaluation: Vol.8, Article 6. *Notes on the use of data transformations*. Retrieved from https://scholarworks.umass.edu/pare/vol8/iss1/6

Pope, P., & Webster, J. (1972). The Use of an F-statistic in Stepwise Regression Porcedures. *Technometrics, Vol.14, No2*, 327-340.

Renaud, O., & Victoria-Feser, M.-P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 1852-1862.

Rogers, K., & Boyd Enders, F. (2013, December). *Collinearity*. Retrieved from Britannica: https://www.britannica.com/topic/collinearity-statistics

Sadeghkhani, I., Ketabi, A., & Feuillet, R. (2012, May 1). Radial Basis Function Neural Network Application to Power System Restoration Studies. *Computacional Intelligence and Neuroscience*, pp. 1-10. Retrieved from Hindawi.

Sharma, A. (2017, March 30). *Understanding Activation Function in Neural Networks*. Retrieved from Medium: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0

Singh Chauhan, N. (2019, October 3). *Introduction to Artificial Neural Networks (ANN)*. Retrieved from Towards Data Science: https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9

Witten, I. H., & Pal, C. J. (2017). *Extending instance-based and linear models*. Retrieved from Science DIrect: https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron