Evaluating Classical and Artificial Intelligence Methods for Credit Risk Analysis

An experimental comparison using a database for B2B clients

Bruno Miguel Querido dos Reis

Department of Engineering and Management, Instituto Superior Técnico

Area: Industrial Engineering and Management

Credit scoring remains one of the most important subjects in financial risk management. Although the methods in this field have grown in sophistication, further improvements are necessary. These advances could translate in major gains for financial institutions and other companies that extend credit by diminishing the potential for losses in this process. This research seeks to compare statistical and artificial intelligence predictors in a credit risk analysis setting, namely the discriminant analysis, logistic regression, artificial neural networks and random forests. In order to perform this comparison, these methods are used to predict the default risk for a sample of companies that engage in trade credit. Pre-processing procedures are established, namely in the form of a proper sampling technique to assure the balance of the sample. Additionally, multicollinearity in the dataset is assessed via an analysis of the variance inflation factors and the presence of multivariate outliers is investigated with an algorithm based on robust Mahalanobis distances. After seeking the most beneficial architectures/settings for each predictor category, the final models are then compared in terms of several relevant key performance indicators. This allows for conclusions to be drawn regarding the performance of statistical and artificial intelligence approaches.

Keywords: *credit scoring, credit risk, artificial intelligence, risk management, discriminant analysis, logistic regression, artificial neural networks, random forest.*

1. Introduction

Companies acquire funds not only from specialized financial intermediaries but also from the respective suppliers (Fabbri & Menichini, 2010). This practice is denominated trade credit and occurs frequently in the B2B market when buyers delay payments to suppliers for merchandise and/or services. If credit is approved for a certain client, there is always the possibility that this client will not honor the agreement to repay the amount in question. On the other hand, if credit is denied, it is possible that a potentially profitable client was handed over to rival companies. Therefore, both of these issues must be taken into consideration when deciding on whether to extend credit to any applicant.

Credit risk, in general, is a topic of the utmost importance in financial risk management, being a major source of concern for financial and banking institutions (Khashman, 2010). In the last decades, quantitative methods to manage credit risk have grown in sophistication. The end-goal is to separate good credit applicants from bad ones. The criterion used in this classification is the ability of the applicants to repay the full amount of the loan. Usually, this is achieved by feeding a predictive model with past customer data, thus finding the relationships between the clients' characteristics and the potential for default (Huang, Liu, & Ren, 2018). There is substantial research material on this topic, as only a small improvement in prediction accuracy may lead to large gains in profitability (Kvamme, Sellereite, Aas, & Sjursen, 2018).

Until recently, to build these credit scoring models, the sole solution was to employ statistical models. The linear discriminant analysis and logistic regression are among the statistical techniques widely used for this purpose (Baesens, Setiono, Mues, & Vanthienen, 2003).

However, the emergence of artificial intelligence (AI) methods provided an opportunity for credit risk professionals. There are numerous studies showing that machine learning tools like artificial neural networks, decision trees and support vector machines, present a chance to improve on the prediction accuracy of statistical models with regards to credit risk (Vellido, Lisboa, & Vaughan, 1999; Huang, Chen, Hsu, Chen, & Wu, 2004; Ong, Huang, & Tzeng, 2005).

Despite significant developments in terms of newer classifiers, the literature on credit risk has not kept pace with the breakthroughs in predictive learning (Lessmann, Baesens, Seow, & Thomas, 2015; Jones, Johnstone, & Wilson, 2015). Indeed, more recent techniques such as random forests and generalized boosting have been explored by a limited number of studies, although some sources report them as superior to previous methods (Jones et al., 2015). It is therefore imperative to further study these new techniques to understand how these compare to older and more established methods of credit scoring with respect to performance and applicability.

This research seeks then to offer a comprehensive view of how statistical and artificial intelligence predictors compare at credit scoring. More specifically, this study focuses on the discriminant analysis, logistic regression, artificial neural network and random forest methods. In order to assess the robustness of these techniques, the predictors are used to determine the default risk for a novel sample composed of companies that engage in trade credit.

2. Theoretical Framework

2.1. Linear Discriminant Analysis

The linear discriminant analysis (LDA) may be defined as a statistical technique utilized to classify an observation into one of several a priori groupings depending on the observation's individual characteristics (Altman, 1968). There are some limitations regarding the validity of this method. It is dependent on stringent assumptions, namely that all variables must present a normal distribution and be mutually independent (Huang et al., 2004; Šušteršič, Mramor, & Zupan 2009).

Considering a certain feature vector with s dimensions, it is important to know what linear function of these values best separates the groups in question. This function corresponds to the expression that follows.

$$f(x) = \lambda_1 x_1 + \cdots + \lambda_s x_s \quad (1)$$

In this formula, λ_i and x_i represent the discriminant coefficient for explanatory variable i and the value for indicator i, respectively. In the LDA, the goal is to find the values for these coefficients that maximize the differences between the groups as measured by a given objective function. The original method proposed by Fisher in 1936 sought to find the coefficients that maximized the ratio of the explained variance to the unexplained variance. This corresponds to the F-ratio, which may be computed with the following expression:

$$F = \frac{\sum_{g=1}^{G} N_g(\bar{y}_g - \bar{y})^2}{\sum_{g=1}^{G}\sum_{p=1}^{N_g}(y_{pg} - \bar{y}_g)^2} \quad (2)$$

This formulation considers a total of G groups in a dataset, with g and y_{pg} being the index for the groups and the observation p of group g, respectively. Additionally, N_g represents the number of cases in each group, while $\bar{y}_g$ is the mean for group g and $\bar{y}$ is the overall sample mean. Analyzing this expression, one can observe that its numerator corresponds to the sums-of-squares between groups and the denominator to the within-groups sums-of-squares (Altman, 1968).

Once the coefficients have been computed to maximize the discriminant power of the function, it is possible to calculate the score for each observation in the sample and assign it to a certain group accordingly.

The LDA technique was first applied to credit scoring by Edward Altman in 1968. This approach is designated by Altman's Z-score and served as the basis for the future applications of discriminant analysis in credit scoring. Altman's method implies assigning each instance to the group it resembles the most. The comparisons are measured by a chi-square value and classifications are made based upon the relative proximity of the instance's score to the various group centroids (Altman,1968).

2.2. Logistic Regression

The logistic regression (LR) is one of the most widespread statistical tools for classification problems in general (Ong et al., 2005). Much as the LDA, it is a technique used in problems with categorical dependent variables displaying linear relationships with the explanatory variables. Despite the similarities, it should be stressed that the logistic regression model does not assume the populations in classification problems to be normally distributed. Unlike the LDA, the logistic regression can deal with various distribution functions (Press & Wilson, 1978; Ong et al., 2005), and is thus, arguably, a better option for credit scoring tasks.

Assuming the case of a binary logistic regression that is used to determine if an event E will happen (e.g. company bankruptcy), then $\pi(x)$ may be defined as the probability of E occurring given the n-dimensional input vector X. As there are only two possible outcomes, 1 - $\pi(x)$ is equal to the probability of the event E not happening. The linear form of the LR model may be obtained by applying the natural algorithm to the odds ratio, which is equivalent to the logit of $\pi(x)$. This leads to the following mathematical formulation:

$$logit\big(\pi(x)\big) = log\frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta X \quad (3)$$

A different formulation of the logistic regression is usually obtained by relating the probability of a given event, E, happening, conditional on the vector X of observed explanatory variables, to the vector X (Press & Wilson, 1978). This corresponds to expression 4, which may be also obtained by manipulating the former formula.

$$\pi(x) = \mathrm{P}(E|x) = \frac{1}{1 + e^{-\alpha-\beta X}} \quad (4)$$

The output of this expression describes a sigmoid curve, taking values between zero and one. After the parameter α and the vector of coefficients β are calculated, it may be used as a predictor. The maximum likelihood method that is commonly used in statistics can be applied to estimate these parameters.

2.3. Artificial Neural Networks

Artificial neural networks (ANNs) started being studied as a possible credit risk predictor in the nineties (Tang et al., 2018) and since then have become a mainstream tool utilized by several financial institutions and other companies.

Neural networks are composed of several artificial neurons, which can be regarded as processing units. These elements are interconnected via synapses that convey values, with each one of these connections having an assigned weight. When a neuron performs a computation, the first step is to do a weighted sum of the inputs, afterward, the result is used in the transfer function that will calculate the neuron's output. Sigmoid, linear and step functions are common transfer functions (Angelini, di Tollo, & Roli, 2008).

All neural networks require the partitioning of the input data into training, validation and testing subsets, which have distinct purposes. The training subset is used in the learning stage of the models, while the validation subset assures that every change in the models' parameters truly reduces the overall error. In the absence of validation, the models may overfit by modeling noise in the training data. Finally, the testing subset provides an independent way to assess the predictive ability of the models.

The first artificial neural network considered in this research is the multilayer perceptron (MLP), which is the most frequently used type of neural network in credit risk assessment (West, 2000), having been tested in various studies. The backpropagation rule is a widely used technique to update the weights of these networks (Zhao et al., 2015; Huang et al., 2018). Backpropagation algorithms are supervised learning tools. These techniques begin by initializing the weights with small random values (West, 2000). Subsequently, the gradient of the error's variation with respect to changes in the weights is computed, and these weights are modified in the direction which reduces the overall error of the network.

The other artificial neural network tested in this research is a radial basis function (RBF) neural network. The first layers of these models just carry the data directly to the ensuing layers. A fundamental aspect of these networks is that the hidden layers are entirely composed of neurons with radial basis transfer functions, such as Gaussian functions (Ayala & Coelho, 2016). The outcome of a radial basis function is dependent on three parameters: the received input vector X, the center of the respective neuron c_j and the spread σ_j. The training that RBF networks undergo allows for the determination of the appropriate number of hidden layers and also the best centers and widths for each hidden neuron (Chen, Wang, Liu, & Wu, 2018). These parameters will be the ones that allow for a minimization of the network's overall error.

The estimation of the centers can be done via a clustering algorithm. The k-means clustering technique, for example, is one of the common and intuitive methods of this type. This algorithm considers a set of initial centers and then iteratively changes the centers to minimize the total within cluster variance (Hastie, Tibshirani, & Friedman, 2008). First, all the input data points are attributed to the closest center, which effectively corresponds to dividing the data into separate subsets. Afterward, each center is recalculated to correspond to the vector of the means for the features of the data points composing the respective subset.

Despite the great promise of ANNs in general, there is a major disadvantage that should be noted. Neural networks work as black boxes, which basically means that it is very difficult to interpret how the results are achieved (Abdou & Pointon, 2011). This may severely restrict the use of such techniques.

2.4. Random Forest

This research also includes the testing of the random forest (RF), which is a much newer artificial intelligence technique. A random forest is a homogenous ensemble predictor. Its predictions are dependent on the individual outputs of various decision trees (DTs). The aggregation of the many outputs obtained into a single outcome may be done by averaging over all the output values when predicting a numerical outcome or by performing a vote when predicting a class (Breiman, 1996). There is evidence that this procedure of model combination can lead to increased accuracy (Paleologo, Elisseeff, & Antonini, 2010; Finlay, 2011; Lessmann et al., 2015).

Assuming it is used for classification purposes, a random forest is analogous to a voting committee. Each decision tree reaches a prediction or classification and then the results of all trees are checked to find what is the output of the majority. It is implied in this logic that the decision trees reach different results and consequently display distinct structures. A fundamental challenge when building a RF is to ensure decision tree diversity. The diversification of decision trees is achieved via two mechanisms, bootstrap aggregating (bagging) and random feature selection.

Bootstrap aggregating is a procedure that allows each tree to use a different sample as input without partitioning the data. These replicate datasets, each consisting of a given number of cases, are drawn at random, but with replacement, from the original dataset (Breiman, 1996).

In contrast, the random feature selection mechanism dictates that each node is assigned a random subset of variables that it may use in the node splitting procedure. This random selection of features at each node decreases the correlation between the decision trees, causing a reduction in the random forest error rate (Bryll, Gutierrez-Osuna, & Quek, 2003; Archer & Kimes, 2008). Random feature selection has been demonstrated to perform better than bagging alone (Dietterich, 2000), namely in problems with several redundant features (Archer & Kimes, 2008). This strategy has also been proven to help prevent the overfitting phenomenon.

However, after the random forest is applied, its results are not easily interpretable, which is inconvenient when it is critical to understand the interactions between the variables of the problem (Breiman, 2001).

2.5. Key Performance Indicators

The models in the following sections are evaluated in terms of several key performance indicators (KPIs). These include the percentage of correctly classified (PCC) instances, which measures the accuracy of the techniques. The sensitivity and specificity are also presented, which measure the imperviousness of the models to type I and type II errors, respectively. Assuming a null hypothesis that the company applying for credit will not default next year, then the sensitivity is equal to the true positive rate and the specificity corresponds to the true false rate.

The area under the curve (AUC) is also computed for all models. The AUC corresponds to the area under the receiver operating characteristic (ROC) curve. Assuming binary outcomes, this curve plots the sensitivity and specificity observed for different thresholds. Finally, the Gini Index is included. This coefficient is a chance standardized alternative to the AUC that measures how well the models separate the existing groups. Greater values for the AUC and Gini Index are desirable, as these are indicative of a higher discriminatory ability. It should be noted that, in cases of conflicting performance ranks, these last two measures are prioritized in this work.

3. Input Data Collection, Analysis and Treatment

3.1. Input Data Collection Process

The data used in the models was obtained from the Orbis financial database. Bureau van Dijk (BvD), a Moody's Analytics Company, is responsible for the capture and treatment of the data present in this database. The access to the database is provided in exchange for a subscription fee, albeit there is a free trial version available online at BvD's website.

The financial information used in this research was extracted for a list of Galp's clients and concerns the fiscal year of 2016. Additionally, the information regarding the clients' financial status in the fiscal year of 2017 was retrieved from the internal data kept by Galp.

3.2. Description of the Input Variables

In order to obtain the most explanatory input variables, several financial and non-financial indicators were extracted from the database or computed from the exported information. This data includes raw financials, equity ratios, growth tendencies, operational ratios, the maturity of the companies, profitability ratios, sectors of activity and structural ratios. The final indicator included, company status in 2017, corresponds to the dependent variable for all the models. In this variable, all companies are assigned to the mutually excluding categories:

- Active: The company remains in operation;
- Insolvent: The company has filed for bankruptcy;
- Undergoing a Special Revitalization Process (SPR): The company has been given a protection against creditors status, preventing an imminent insolvency;
- Non-compliant: The company has failed to pay for the products and/or services provided by Galp.

3.3. Aggregation of company outcomes

The company status variable poses a challenge, as it must be decided whether to aggregate the negative categories under a broader class of bad companies, merge just some of these, or keep all of them separate.

Although there are several possible groupings for the distinct strategies, a preliminary analysis is enough to understand that some seem counterintuitive. The discriminant analysis, as well as the artificial neural networks and other predictive models, offer similar predictions for close inputs, as such, it is detrimental to merge classes that are characterized by very dissimilar inputs. Therefore, one must take this factor into consideration when deciding on the best course of action regarding the aggregation of classes.

Both insolvent and SRP companies display similar very poor financial indicators. Hence, this pair of classes is the most logical choice to undergo merging. Non-compliant companies display better financial indicators in comparison with the other two negative categories, although these indicators remain deteriorated in relation to active companies.

Upon experimenting with the aggregation strategies, it became evident that it is beneficial to keep only two possible outcomes. This is due to the similarity of the inputs obtained for insolvent, SPR and non-compliant classes. Furthermore, the main goal of any creditor is to understand if there is a significant risk of default for any given potential debtor, and it is notorious that the applicants included in these three classes present such a risk. Considering this, it was ultimately decided to pursue a two-outcome aggregation strategy, merging the insolvent, SPR and non-complaint categories in a broader class of bad companies. The active companies remain in a separate class of good companies.

3.4. Sampling Procedure

Although the majority of credit scoring research has not focused on the input samples' characteristics, the size and balance of such datasets have a tremendous potential to affect the performance of the predictive models. This latter characteristic refers to the proportion of the groups in the sample. Ideally, considering a binary outcome scenario, half the instances would belong to one group and the remaining to the other. Some methods are more sensitive than others to changes in the input data's size and structure, but both statistical and AI techniques are affected by these features to varying degrees.

There are two options to manipulate the balance of a sample, under-sampling by reducing the number of instances of the majority group or over-sampling through an increase of the cases in the minority class. In this research, it was decided to under-sample the majority class, which encompasses the cases of good companies. Although over-sampling may produce better results according to Crone & Finlay (2012), this dataset proved extremely unbalanced due to a pronounced deficiency of bad companies, making it difficult to employ this technique. Considering that the minority class is much smaller, over-sampling would cause certain cases in this class to be repeated several times. This repetition may lead the models to overfit, thus degrading the results.

After selecting a subset of instances from the good companies' class, the near perfectly balanced dataset described in Table 1 was obtained.

Table 1 - New distribution of the cases by the categories.

Group	Subgroup	Observations	Percentage of total
Good	Active	1001	50.2%
Bad	Insolvent	701	35.2%
	SPR	265	13.3%
	Non-compliant	27	1.4%

The slightly bigger number of good companies in relation to the total number of bad companies is due to a few detected cases of duplicated corporations in the data. This issue was solved by studying the causes of each repetition and assigning these cases to a sole category.

3.5. Missing and Invalid Data

Another important aspect to be addressed relates to the presence of missing values in the dataset. The usual reasons for missing values in credit scoring problems are that those values were already missing in source data or were out of the theoretical allowed range. The latter motive is quite common in these situations due to typos or transcription errors (Angelini et al., 2008). On the other hand, these lapses may be due to computational errors. After analyzing the dataset, two main types of missing data were detected, NA and NS lapses. The first one corresponds to data that is truly missing, NA being an acronym for not available in the database. On the other hand, NS stands for not significant and is used when indicators expressed as percentages take values near zero. As NS cases do not truly represent missing data, these were replaced by null values in the sample. This approximation allows for the use of such instances.

3.6. Correlation Analysis

The multicollinearity problem refers to the existence of strong correlations between independent variables in a dataset. Many authors have stated before that the logistic model becomes unstable in the eventuality of a strong dependence among predictors, as it seems that no single variable is important when all the others are in the model (e.g. Aguilera, Escabias, & Valderrama, 2006). This weakness is shared with the LDA method.

A common technique used in the detection of multicollinearity involves the computation of the variance inflation factor (VIF). Variance inflation factors over 10 are usually considered to be indicative of multicollinearity. However, certain authors point out that this threshold is very lenient. Indeed, a VIF of 10 for a given independent variable implies that 90% of its variability is explained by the remainder indicators. Another typical threshold is a maximum VIF of 5 (Craney & Surles, 2002). This is a more conservative approach that was deemed adequate, as certain variables displayed VIF values nearing 10 and would not be excluded with the former criterium.

The correlation analysis indicated that there are clear signs of multicollinearity in the original data, with several VIF values exceeding the threshold defined. In order to solve this problem, the variables were removed iteratively until no VIF values were over 5. This removal procedure was performed giving preference to the variables that are more correlated. The final dataset obtained displays no indications of multicollinearity.

3.7. Outlier Analysis

According to Filzmoser (2004), the basis for multivariate outlier detection is the Mahalanobis distance. This metric measures the distance of each instance in the data to a central point in multivariate space. The key feature of this measure is that it considers the correlations between variables, as well as the respective scales (Brereton & Lloyd, 2016). The Mahalanobis distances (MDs) may be computed with following expression:

$$MD = \sqrt{(x_i - \bar{x})S^{-1}(x_i - \bar{x})^T} \qquad (5)$$

This formula considers that x_i is the vector for a given data instance, while $\bar{x}$ is the arithmetic mean of the dataset and S represents the sample covariance matrix.

However, outliers are known to distort the observed mean. A small cluster of outliers may impact the mean in such a way that these are no longer detected as aberrant instances. Additionally, the distortion brought on by the outliers may be so high that normal instances are wrongly labeled as outliers. These occurrences are commonly referred to as masking and swamping, respectively. In order to prevent them, it was decided to examine the presence of outliers by computing MDs with geometric medians (GMs). This indicator is one of the most common robust estimators of centrality in Euclidean spaces (Fletcher, Venkatasubramanian & Joshi, 2008).

In order to compute this parameter, the Weiszfeld algorithm is employed. This is an iterative procedure that with the appropriate initialization values converges to the point that presents the lowest sum of Euclidean distances for all the sample instances.

The computation of the GMs does not tolerate missing values. As such, it is necessary to replace these lapses with usable data. The techniques used for this purpose are called imputation procedures. After analyzing the sample's pattern of missing data and assessing if monotonicity is present, it was decided to proceed with a fully conditional specification imputation method.

This procedure warrants the separation of the sample into two groups, which contain exclusively good and bad companies. Since the whole sample contains two distinct populations with very different characteristics, this splitting is fundamental to assure that the MDs are computed with the GMs of the class (good or bad) to which each instance belongs.

As normality tests proved that various indicators do not follow normal distributions, it was opted to use an alternate exclusion criterion to the comparison of the MDs with a quantile of the chi-squared distribution. There is no guarantee that the MDs follow this specific distribution in the absence of multivariate normality. By building scatter plots with the sample ID numbers and the robust MDs, it is possible, via visual inspection, to detect any potential outliers. These plots are displayed in Figures 1 and 2.

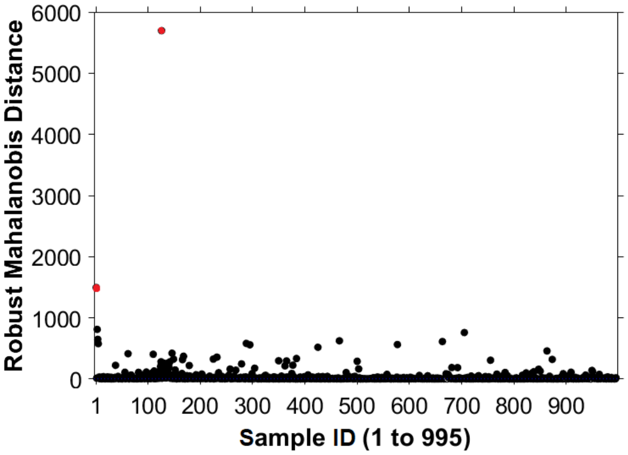

Figure 1 - Scatter plot of the robust Mahalanobis distances for the good companies.

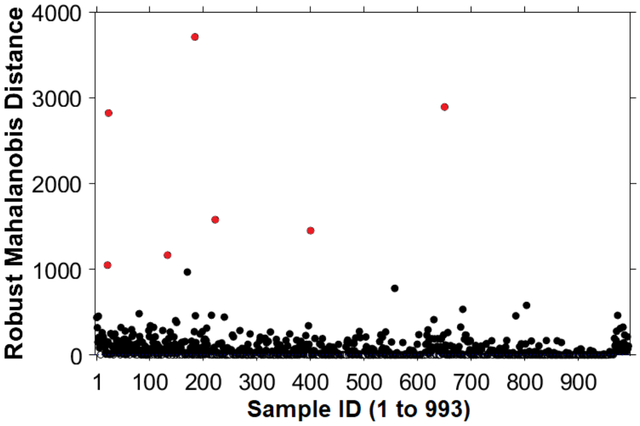

Figure 2 - Scatter plot of the robust Mahalanobis distances for the bad companies.

Some instances in these scatter plots standout for being clearly anomalous. It was decided to label as potential outliers all the cases with robust Mahalanobis distances above 1000. These points are marked in red for easier identification. In order to comprehend to what extent these flagged instances are aberrant, there was an analysis of the indicators presented by these companies. This study reinforced the idea that such corporations display altered values for several indicators.

Considering that the results of the robust MDs analysis were confirmed for good and bad companies by the subsequent findings of extreme values for several indicators in the flagged cases, the decision was taken to label these nine instances as outliers and remove them from the sample. The outlier detection technique implemented in this section was partially based on the work of Semechko (2019). Further details are provided in the reference section.

4. Model Development

4.1. Linear Discriminant Analysis

The discriminant analysis model was applied to the data with IBM SPSS Statistics 25. Considering the capabilities of the software, alternative discriminant analysis models were computed using different combinations of stepwise techniques and entry/removal criteria.

After experimenting with various selection rules, it was found that the best results were obtained by including in the model any explanatory variables with a minimum F value of 3.00 and excluding those with F values inferior to 1.00.

Following the computation of the discriminant coefficients, it was possible to assess the relative importance of the independent variables included in the model. The standardized coefficients are particularly important to assess the discriminating ability of the explanatory variables, as the standardization allows for the comparison of variables expressed in distinct scales. The five variables with the most predictive potential were found to be the shareholder equity ratio, Cash flow / Total assets, return on assets using net income, credit period and the major sector of activity, by descending order of discriminating ability.

The key performance indicators were then computed for the best discriminant analysis model obtained. These are listed in Table 2.

Table 2 - KPIs for the linear discriminant analysis model.

PCC	Sensitivity (%)	Specificity (%)	AUC	Gini Index
80.0	88.9	67.7	0.863	0.726

4.2. Logistic Regression

The logistic regression model was also applied with IBM SPSS Statistics 25. There is no need to use a multinomial logistic regression, as the considered output is dichotomous. Therefore, a binary logistic regression model was implemented.

The first step in the development of this model is choosing the input variable selection procedure. There are a variety of stepwise techniques available in this software, namely forward selection and backward elimination procedures.

After careful experimentation, the best results were obtained using the forward selection stepwise techniques. The maximum number of iterations before model termination was kept at 20, the default setting, as overriding this configuration did not improve the results. In terms of the thresholds used in the stepwise methods, the best results were obtained when the probability for the score statistic must be less than 0.01 for entry and over 0.03 for removal. The option to include a constant in the LR model remained selected.

Furthermore, the user interface allows for the definition of the classification cutoff directly, which was kept at 0.5. Although it is relevant to study the model’s performance under different thresholds, this will be addressed with the computation of the remaining KPIs, namely the AUC. Table 3 presents all of these performance metrics, which are relative to the most robust logistic regression model achieved.

Table 3 - KPIs for the logistic regression model.

PCC	Sensitivity (%)	Specificity (%)	AUC	Gini Index
89.9	93.8	83.5	0.926	0.852

4.3. Multilayer Perceptron

The multilayer perceptron model was applied in the neural networks’ module of IBM SPSS Statistics 25, which offers various options regarding the way the ANNs are structured and the methods through which the results are computed.

First, the partitioning of the data may be set. This involves specifying the fractions of the sample that are allocated to the training, validation and testing datasets. Secondly, the structure of the MLP network may be stipulated in terms of the number of hidden layers, the activation function to be used in these layers and the transfer function of the output layer.

Finally, there are different options for the learning algorithm to be employed in the networks’ development. Considering these possibilities, four different MLP neural networks are proposed, which are detailed in the following table.

Table 4 - Features of the MLP networks tested.

ANN	Number of hidden layers	Number of hidden neurons	Hidden layers’ activation function	Output layer’s activation function	Training algorithm
MLP 1	1	Automatic selection	Sigmoid	Identity function	Scaled conjugate gradient
MLP 2	2	Automatic selection	Sigmoid	Identity function	Scaled conjugate gradient
MLP 3	1	Automatic selection	Hyperbolic tangent	Identity function	Scaled conjugate gradient
MLP 4	2	Automatic selection	Hyperbolic tangent	Identity function	Scaled conjugate gradient

Regarding the partitioning of the data, several combinations were selected in accordance with the best practices in the literature. The first training-testing-validation ratio, 700:300:0, is the most popular partition, being used by numerous authors (e.g. Angelini et al., 2008 and Pacelli & Azzollini, 2010), being also the default setting in SPSS. The second option, 600:150:250, is used by Lai, Yu, Wang, and Zhou (2006). Lastly, the third partitioning, 600:200:200, which varies only slightly in relation to the second alternative, is based on the work of Addo et. al, 2018.

For the comparison between methods to be fair, one must be careful when setting the partitioning strategy in SPSS. The percentage of cases that are attributed to each set may be defined directly in the software's user interface for a given network. However, this introduces the potential for chance to influence the results. As the cases are randomly sampled from the dataset to build the training, testing and validation sets, the results obtained will be strongly influenced by this arbitrary selection. By not guaranteeing the replicability of the partition, the comparison between the different architectures cannot yield meaningful results.

This issue essentially arises because some companies are more difficult to classify than others. Not all instances present overwhelmingly positive or negative indicators. These cases are the ones that contribute the most to the errors committed by the models. If a given partition randomly samples more of these instances than the others, the models using it would tend to display poorer results, although this partitioning strategy is not necessarily inferior to the others. The same reasoning applies to comparisons between different models that use the same partitioning strategy. A given model may perform better solely because it was evaluated with a test set containing a higher percentage of instances that are easier to sort.

In order to overcome this flaw, a strategy was employed that mitigates the potential for the models' results to be influenced by chance. Firstly, three partitioning variables were defined beforehand. These variables contain values that determine the placement of each instance (i.e. whether it is used for training, testing or validation). The variables' values are generated in accordance with the partitioning strategy desired and used for the testing of all the MLP models for that given strategy. This essentially assures that the networks are comparable if the results were obtained for the same partitioning option, as all of these models were developed with similar initial conditions.

However, when comparing networks that used different partitioning strategies, which correspond to different auxiliary partitioning variables, there is still the potential for chance to affect the analysis. Therefore, it was deemed necessary to do multiple runs of the algorithm that generates these variables and then compute the average values for the KPIs.

By averaging out all the performance metrics across the iterations (according to the MLP model and partitioning option considered in each iteration), it was possible to compute the results that are presented in Table 5.

Table 5 - Average values of the KPIs after 5 runs for each combination of MLP model and partition.

Partitioning	ANN	PCC	Sens. (%)	Spec. (%)	AUC	Gini Index
700:300:0	MLP 1	88.10	93.68	79.32	0.940	0.881
	MLP 2	88.14	93.44	79.92	0.942	0.884
	MLP 3	88.32	94.08	79.28	0.949	0.898
	MLP 4	89.12	93.74	81.74	0.951	0.902
600:150:250	MLP 1	88.02	91.86	81.22	0.947	0.894
	MLP 2	88.04	91.98	81.10	0.939	0.878
	MLP 3	89.74	93.28	83.28	0.957	0.914
	MLP 4	90.74	94.54	84.04	0.959	0.917
600:200:200	MLP 1	89.02	93.88	81.00	0.946	0.892
	MLP 2	88.52	93.38	80.38	0.937	0.875
	MLP 3	89.44	94.44	81.08	0.950	0.899
	MLP 4	90.20	94.42	83.14	0.954	0.909

Analyzing the values of the KPIs displayed in this table, which are all relative to the testing set, it can be understood how each MLP network performs for all the partitioning strategies considered. After comparing the models, it was considered that the most robust network is MLP 4 trained with a 600:150:250 training-testing-validation ratio. This artificial neural network displays the best value for the AUC, as well as the greatest Gini index. Finally, a sensitivity analysis is performed that computes importance estimates for each independent variable in the model. These results imply that the most important indicator is the shareholder equity ratio, followed by the Cash flow / Total assets. The variations of the cash flow and equity are considered the third and fourth most relevant variables, respectively.

4.4. Radial Basis Function Neural Network

The RBF neural network model was applied in the neural networks' module of SPSS Statistics 25. In the same way as the MLP models, there is the option to define the percentages that are assigned to the training, validation and testing sets. Additionally, there are two alternatives for the activation function used in the hidden layers, which are ordinary and normalized radial basis functions. The remaining customizable settings are the number of elements in the hidden layers and the overlap among hidden units. The overlapping factor is a multiplier applied to the width of the radial basis functions.

As SPSS offers algorithms that define the optimal number of units in the hidden layers and the best values for the overlapping factors, these features were not set manually. Thus, the software automatically defined the most advantageous architecture regarding these characteristics. Considering that there is no mechanism in place to select the transfer function in the hidden layers that achieves the best results, two alternative RBF networks are studied that differ solely in this aspect.

Table 6 - Features of the RBF networks tested.

Characteristics	RBF 1	RBF 2
Number of elements in the hidden layers	Set automatically	Set automatically
Overlapping factor	Set automatically	Set automatically
Activation function for the hidden layers	Normalized RBF	Ordinary RBF

The partitioning schemes defined in the previous section were also considered for the development of the RBF models. Similarly to what was done before, in order to mitigate the variability in the results that can happen because of the random sampling procedure used to build the various sets in SPSS, two partitioning variables were computed and used iteratively to build the networks and collect the KPIs. By averaging out all the performance metrics across the iterations, it was possible to obtain the results presented in Table 7.

Table 7 - Average values of the KPIs after 5 runs for each combination of RBF model and partition.

Partitioning	ANN	PCC	Sens. (%)	Spec. (%)	AUC	Gini Index
700:300:0	RBF 1	83.54	87.36	77.68	0.892	0.784
	RBF 2	81.64	88.02	72.00	0.884	0.768
600:150:250	RBF 1	81.02	84.54	75.12	0.889	0.778
	RBF 2	81.28	86.76	72.02	0.892	0.785
600:200:200	RBF 1	81.88	84.84	76.94	0.890	0.780
	RBF 2	82.38	85.94	76.22	0.891	0.782

It may be concluded from the results displayed in Table 7 that RBF 2 under the second partitioning option (60% Training, 15% Testing and 25% Validation) outperforms the remaining alternatives.

4.5. Random Forest

The random forest method was applied in MATLAB R2018b. This model can be obtained by using the TreeBagger function available in the software, which builds an ensemble of bootstrapped decision trees for either classification or regression purposes. This function also selects a random subset of predictors to use at each decision split as is described by Breiman (2001) in the original random forest algorithm.

In terms of the settings used, the model is set for classification purposes, as the outcome considered is categorical. The surrogate splits option is activated to handle cases of missing data. If the value for the best split is missing, this technique assesses to what extent alternate splits resemble the best split. Afterward, the most similar split is used, instead of the original optimal division.

Additionally, optional arguments are included in the function to allow for the assessment of the variables' explanatory power and the computation of the predicted class probabilities. The probabilities are especially important, as these are used in the latter plotting of the ROC curve and subsequent computation of the AUC.

The TreeBagger function offers two possibilities for the algorithm that selects the best split at each node, a curvature test (CT) and an interaction-curvature test (ICT). In order to understand which of these algorithms would provide the best results, two distinct random forest models were applied differing in the splitting techniques. The relevant KPIs obtained for both models are displayed in Table 8.

Table 8 - KPIs for the different splitting algorithm options.

Splitting algorithm	PCC	Sens. (%)	Spec. (%)	AUC	Gini Index
CT	96.46	98.59	94.32	0.997	0.994
ICT	96.61	98.49	94.73	0.996	0.992

The random forest using the curvature tests provided the best predictions in terms of AUC, Gini Index and sensitivity. Although the percentage of correctly classified cases is slightly inferior to the one presented by the model trained with the interaction-curvature tests, a higher AUC is prioritized. A critical parameter that must also be defined is the number of decision trees contained in the ensembles. The results displayed so far were obtained with models composed of 50 decision trees, which is a common setting for random forest models. However, it must be analyzed if there are gains to be had by adding more trees or, on the other hand, there is an excess of DTs that does not translate into a reduction of the prediction error and increases the computation time unnecessarily. In order to do this, the out-of-bag prediction error is plotted for a variable number of decision trees in the graph present in Figure 3.

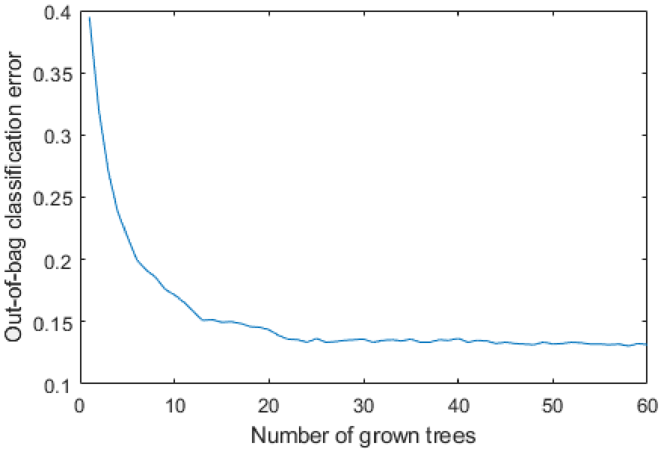

Figure 3 - Out-of-bag prediction error obtained for a variable number of decision trees.

Analyzing Figure 3, one can observe that, when the total number of grown trees is small, there is a rapid decrease of the out-of-bag prediction error with additional DTs in the ensemble. However, these gains in accuracy are progressively smaller, which causes the out-of-bag prediction error to stabilize around an ensemble of 50 trees. The error rate observed for a RF containing 50 decision trees is 0.1319, whereas an ensemble of 60 DTs displays a rate of 0.1314. As this reduction is hardly significant, it was opted to keep the number of decision trees at 50.

Analyzing the estimates of the predictors' importance, it is relevant to point out that the variable with the most explanatory power is the shareholder equity ratio, which displays a remarkable score in comparison with the other indicators. The credit period and the Cash flow / Total assets indicators display the second and third highest importance estimates, respectively. Certain measures, namely the profit per employee and gearing, are also important to the robustness of the model.

5. Benchmarking the models

By compiling the results obtained so far in terms of the relevant KPIs, it is now possible to compare the credit scoring approaches. For each category of predictive methods, the best model in the developmental stage was considered for benchmarking purposes. Table 9 exhibits the values for the performance metrics, as well as a ranking based on the AUC and Gini Index displayed.

Table 9 - KPIs for all the credit scoring models implemented.

Model	PCC	Sens. (%)	Spec. (%)	AUC	Gini Index	Rank
LDA	80.0	88.9	67.7	0.863	0.726	5
LR	89.9	93.8	83.5	0.926	0.852	3
MLP	90.7	94.5	84.0	0.959	0.917	2
RBF	81.3	86.8	72.0	0.892	0.785	4
RF	96.5	98.6	94.3	0.997	0.994	1

Analyzing Table 9, it can be observed that the random forest model is ranked as the best credit scoring model, displaying the highest AUC and Gini Index, while also presenting a remarkable overall accuracy. Over 95% of all instances are assigned correct predictions, with 98.59% of all future good companies being classified as such. In second place, the MLP neural network displayed impressive KPIs, although not up to par with the ones obtained with the random forest. On the other hand, the RBF neural network was the overall worst AI model considered, being even outranked by the logistic regression model.

Regarding the statistical methods, the results fall in line with what was observed in other benchmarking studies. The discriminant analysis proved to be the least predictive model of all the credit scoring methods tested, which may be a result of the violation of this models' assumptions in terms of normality and mutual independence regarding the explanatory variables. The logistic regression is ranked as the third best predictor, behind the MLP neural network and the random forest. This model provides accurate predictions in almost 90% of the cases and demonstrates good sensitivity and specificity, which translate into low rates of type I and type II errors. Despite this, the LR model fell short on the more robust KPIs, namely the AUC and Gini Index, which caused it to be ranked behind some of the AI models.

Considering these results, it can be concluded that the MLP neural network and the random forest outperformed the statistical approaches in the credit scoring experiment. However, the logistic regression proved to be a robust predictor, displaying a high level of accuracy and presenting values for other performance measures that come close to the results of the AI alternatives. This is coherent with the recent rise in popularity of the LR method, which is a solid compromise in terms of prediction performance and ease of implementation. Furthermore, the logistic regression also permits an intuitive interpretation of the model's parameters, overcoming the black-box syndrome of AI predictors.

6. Conclusions and Further Work

6.1. Further Work

Regarding the pre-processing of the input dataset, several measures were taken to assure the quality of the data, which necessarily impacts the performance of the predictive models. However, posterior studies may adopt distinct methodological approaches to address some limitations of the current research. Specifically, the detection of the multivariate outliers could be improved in terms of the rule utilized in the labeling of these instances. As the input data in the sample failed the normality tests, it was not possible to proceed with the typical criterium of labeling as outliers any observations with robust Mahalanobis distances beyond a given quantile of the chi-squared distribution. The detection of the multivariate outliers relied then upon the visual examination of the scatterplots with the robust MDs for each observation in the dataset. Consequently, the labeling process lacks objectivity. Therefore, it would be beneficial to develop a more sophisticated outlier labeling rule that is applicable to multivariate non-normal data.

Further research could also attempt to mitigate the detrimental effects of the missing values in the dataset. Some of the predictor methods applied in this study simply discard such cases, which reduces the size of the sample utilized. In order to deal with this situation in the context of the computation of the robust MDs, a fully conditional specification imputation procedure was put in place. However, the imputed dataset could not be used in the development of some of the models, which limited the applicability of this sample to the pre-processing stage of this project. Thus, additional studies could attempt to employ multiple imputation procedures that are compatible with the implementation of the credit scoring methods.

6.2. Conclusions

This research allowed for the comparison of statistical and AI predictors, adding significantly to the academic literature by designing a credit scoring experiment using a novel dataset with financial and other relevant data for a selection of Portuguese companies. Credit scoring methods were successfully implemented based on this information and used to distinguish between good and bad applicants in the timespan of a year.

As the statistical predictors are particularly susceptible to multicollinearity in the data and to the presence of outlier instances, there was a thorough pre-processing of the dataset prior to the implementation of the models. This procedure included a correlation analysis to remove certain indicators that displayed high VIF values, which corresponded necessarily to the ones presenting the highest dependencies upon the remaining independent variables. Regarding the outlier issue, there was a detection technique in place based on robust Mahalanobis distances that allowed for the identification of certain aberrant instances in the multivariate space. Additionally, a proper sampling technique was defined in order to build a balanced dataset, as the base data was extremely unbalanced.

After experimenting with different settings and architectures, it was possible to select the most robust models for each category of predictors. This allowed for the comparison of the KPIs computed for statistical and AI alternatives. The benchmarking study completed found that the artificial intelligence methods outperformed the more conventional statistical approaches. The random forest model demonstrated the most potential, followed by the MLP neural network. The RBF neural network and the logistic regression were the fourth and third most robust models respectively, whereas the discriminant analysis was the worst performing model overall.

Regarding the statistical approaches, the results are coherent with the findings of previously published benchmarking research articles. The discriminant analysis is dependent on strict assumptions in terms of normality and mutual independence regarding the input variables, which was a contributing factor to its disuse among credit risk professionals and may explain the poor performance obtained in this experiment. The logistic regression proved to be a robust predictor, displaying a high level of accuracy and presenting values for other performance measures that come close to the results of the AI alternatives. This is consistent with the recent rise in popularity of the LR method, which demonstrated to be a solid compromise in terms of prediction performance and ease of implementation.

The random forest models, along with the MLP artificial neural networks, display tremendous potential in the credit scoring field. In contrast with the statistical techniques, these methods can model hidden non-linear relationships between the explanatory variables and the dependent variable, being also more robust to multicollinearity and the presence of outliers. Besides these advantages, these methods do not make assumptions regarding the probability distributions of the input data. These factors may have contributed to the observed superiority of the AI approaches. The major drawback of these alternatives continues to be the black-box syndrome, which makes the interpretation of the results almost impossible. This may restrict the use of such models in certain settings due to regulatory requirements.

7. References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2–3), 59–88.

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks*, 6(2):38.

Aguilera, A., Escabias, M., & Valderrama, M. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8), 1905-1924.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4), 589-609.

Angelini, E., di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *Quarterly Review of Economics and Finance*, 48(4), 733–755.

Archer, K., & Kimes, R. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4), 2249-2260.

Ayala, H., & Coelho, L. (2016). Cascaded evolutionary algorithm for nonlinear system identification based on correlation functions and radial basis functions neural networks. *Mechanical Systems and Signal Processing*, 68–69, 378–393.

Baesens, B., Setiono, R., Mues, C., Vanthienen, J. (2003). Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Management Science*, 49(3), 312-329.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Brereton, R., & Lloyd (2016). Re-evaluating the role of the Mahalanobis distance measure. *Journal of Chemometrics*, 30(4), 134-143.

Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6), 1291-1302.

Chen, X. & Wang, D. & Liu, Z. & Wu, Y. (2018). A Fast Direct Position Determination for Multiple Sources Based on Radial Basis Function Neural Network. *10th International Conference on Communication Software and Networks (ICCSN)*, 381-385.

Craney, T., & Surles, J. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, 14(3), 391-403.

Crone, S., & Finlay, F. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1), 224-238.

Dietterich, T. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2), 139-157.

Fabbri, D., & Menichini, A. (2010). Trade credit, collateral liquidation and borrowing constraints. *Journal of Financial Economics*, 96(3), 413-432.

Filzmoser, P. (2004). A multivariate outlier detection method. *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, 1, 18-22.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2), 368-378.

Fletcher, P., Venkatasubramanian, S., & Joshi, S. (2008). *2008 IEEE Conference on Computer Vision and Pattern Recognition.*

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York, USA: Springer

Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. *Cognitive Systems Research*, 52, 317–324.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, 37(4), 543–558.

Jones, S., Johnstone, D., & Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking and Finance*, 56, 72–85.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9), 6233–6239.

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102, 207–217.

Lai, K., Yu, L., Wang, S., & Zhou, L. (2006). Credit risk analysis using a reliability-based neural network ensemble model. *Artificial Neural Networks – ICANN 2006*, 682–690.

Lessmann, S., Baesens, B., Seow, H., & Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136.

Ong, C. S., Huang, J. J., & Tzeng, G. H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, 29(1), 41-47.

Pacelli, V., & Azzollini, M. (2011). An Artificial Neural Network Approach for Credit Risk Management. *Journal of Intelligent Learning Systems and Applications*, 03(02), 103–112.

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, 201(2), 490-499.

Press, S., & Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, 73(364), 699-705.

Semechko, A. (2019, September 24). *Detect outliers in multivariate datasets*. Retrieved from https://tinyurl.com/y53e5lhk

Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. Expert Systems with Applications, 36(3), 4736-4744

Tang, Y., Ji, J., Gao, S., Dai, H., Yu, Y., & Todo, Y. (2018). A Pruning Neural Network Model in Credit Classification Analysis. *Computational Intelligence and Neuroscience*, 2018(4), 1-22.

Vellido, A., Lisboa, P. J. G. & Vaughan, J. (1999). Neural networks in business: A survey of applications (1992-1998). *Expert Systems with Applications*, 17(1), 51-70.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, 27(11–12), 1131–1152.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perception neural networks for credit scoring. *Expert Systems with Applications*, 42(7), 3508-3516.