NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

THREE-STAGE ENSEMBLE MODEL: REINFORCE PREDICTIVE CAPACITY WITHOUT COMPROMISING INTERPRETABILITY

by

Martinho de Matos Silvestre

Thesis proposal presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with specialization in Risk Analysis and Management.

Advisor: Roberto Henriques, PhD

November 2018

ABSTRACT

Over the last decade, several banks have developed models to quantify credit risk. In addition to the monitoring of the credit portfolio, these models also help deciding the acceptance of new contracts, assess customers profitability and define pricing strategy. The objective of this paper is to improve the approach in credit risk modeling, namely in scoring models to predict default events. To this end, we propose the development of a three-stage ensemble model that combines the results interpretability of the Scorecard with the predictive power of machine learning algorithms. The results show that ROC index improves 0.5%-0.7% and Accuracy 0%-1% considering the Scorecard as baseline.

KEYWORDS

Ensemble Modeling, Probability of Default, Credit Scoring, Scorecard, Logistic Regression, Decision Tree, Artificial Neural Network, Multilayer Perceptron, Random Forest, Machine Learning

CONTENTS

1. INTRODUCTION

Over the last decade, several banks have developed models to quantify credit risk (Basel Committee on Banking Supervision, 1999). The main objective of credit risk modeling is to estimate the expected loss (EL) associated with credit portfolio that is based on the Probability of Default (PD), the Loss Given Default (LGD) and the Exposure At the time of Default (EAD). The portfolio's expected loss is given by the product of these three components (Basel Committee on Banking Supervision, 2004).

This paper focuses only on PD that is typically computed using scoring models built from historical information of several actual/past customers. The compiled data set will include various attributes and whether the customer has defaulted. Specifically, the credit scoring objective is to assign credit applicants to either good customers (non-default) or bad customers (default), which makes it part of the classification problem domain (Anderson, 1978).

Currently, credit scoring models are used by about 97% of banks that approve credit card applications (Brill, 1998). Using scoring models increases revenue by augmenting volume, reducing the cost of credit analysis, enabling faster decisions and monitoring credit risk over time (Brill, 1998). From the previous, credit risk measurement has become increasingly important in the Basel II capital accord (Basel Committee on Banking Supervision, 2003; Gestel et al., 2005).

In the banking industry, credit risk modeling has been based mostly on logistic regression due to the need to conciliate predictive and interpretative power. Recall that regulators require banks to explain credit application decisions, thus transparency is fundamental to these models (Dong, Lai, & Yen, 2010; Hand & Henley, 1997). In this paper, a three-stage ensemble model is proposed to reinforce the predictive capacity of a scorecard (logistic regression) without compromising its transparency and interpretability.

2. LITERATURE SURVEY

In recent years, several attempts have been made to improve the accuracy of Logistic Regression (Lessmann, Baesens, Seow, & Thomas, 2015). Louzada et al. (2016) reviewed 187 credit scoring papers and concluded that the most common goal of researchers is the proposition of new methods in credit scoring (51.3%), mainly by using hybrid approaches (almost 20%), combined methods (almost 15%) and support vector machine along with artificial neural networks (around 13%). The second most popular objective is the comparison of new methods with the traditional techniques, where the most used are Logistic Regression (23%) and neural networks (21%). One of these studies was done by West (2000), who compared five neural network models with traditional techniques. The results show that neural networks may improve the accuracy from 0.5% to 3%. Additionally, logistic regression was found to be an alternative to the neural networks. In turn, Gonçalves and Gouvêa (2007) obtained very similar results using Logistic Regression and neural network models. However, the proposed new methods tend to require complex computing schemes and limit the interpretation of the results, which makes them difficult to implement (Liberati, Camillo, & Saporta, 2017).

Lessmann et al. (2015) state that the accuracy differences between traditional methods and machine learning result from the fully-automatic modeling approach. Consequently, some advanced classifiers do not require human intervention to predict significantly more accurately than simpler alternatives. Abdou and Pointon (2011) carried out a comprehensive review of 214 papers that involve credit scoring applications to conclude that until now there is no overall best statistical technique used in building scoring models that can be applied to all circumstances. This result is aligned with the Supervised Learning No-Free-Lunch (NFL) theorems (Wolpert, 2002).

Marqués et al. (Marqués, García, & Sánchez, 2012) evaluated the performance of seven individual prediction techniques when used as members of five different ensemble methods and concluded that C4.5 decision tree, Multilayer Perceptron and Logistic Regression were the best algorithms for most ensemble methods, whereas the nearest neighbor and the naive Bayes classifiers appear to be the worst. Gestel et al. (2005) suggested the application of a gradual approach in which one starts with a simple Logistic Regression and improves it using Support Vector Machines to combine good model readability with improved performance.

Summing up, we verified that the most common goal of researchers is the proposition of new algorithms and comparison of new methods with the traditional techniques, where Logistic Regression and machine learning algorithms take a central place. Additionally, it seems that there is no overall best statistical technique used in building scoring models that can be applied to all circumstances. Thus, a natural way to improve the state-of-the-art is to consider an ensemble architecture that can combine traditional methods (as Logistic Regression) with complex algorithms (as machine learning algorithms).

3. THEORETICAL FRAMEWORK

3.1. CREDIT SCORING PROBLEM

Credit scoring objective is to assign credit applicants to either good customers (non-default) or bad customers (default) in the format of a classification problem (Anderson, 1978). Specifically, for each customer historical attributes are recorded and whether the contract has defaulted (failed to pay). Thus, the credit scoring model captures the relationship between the historical information and future credit performance. This relationship can be described mathematically as follows:

$$P(y_i = 1|x_{1i}, x_{2i}, \dots, x_{ki}) = f\left(x_{1i}, x_{2i}, \dots, x_{,ki}\right) \quad (1)$$

where $x_1, x_2, \dots x_k$ represent the customer's attributes, y_i denotes the type of customer (for example good or bad), and f is the function, or the credit scoring model, that maps between the customer attributes (inputs) and his creditworthiness (output). In the credit scoring industry, the most popular method to capture this relation is the Logistic Regression (Hand & Henley, 1997). Then a transformation is needed to convert the creditworthiness into a classification (default/non-default). Usually this is done ,using a threshold (c):

$$\hat{y}_i = \begin{cases} 0, & P(y_i = 1|X) < c \\ 1, & P(y_i = 1|X) \geq c \end{cases} \quad (2)$$

In this paper, we aim to improve the approach used in credit scoring models through the development of an Ensemble Model (Three-Stage Ensemble Model) that combines the results interpretability of Logistic Regression with the predictive power of machine learning algorithms.

3.2. BASE LEARNERS

An ensemble model combines several algorithms which are usually called base learners (Zhou, 2012). The base learners used in this paper are addressed in the following subsections.

Scorecard (Logistic Regression)

The scorecard (SC) model consists in a logistic regression on a set of categorical inputs:

$$y_{SC} = \delta_0 + \sum_{i=1}^{k} \sum_{j=1}^{b_{x_i}} \delta_j^{x_i} B_j^{x_i} \quad (3)$$

where δ_0 stands for the independent term, $B_j^{x_i}$ is a binary variable associated to one of the b_{x_i} classes of x_i (the ith input variable) and $\delta_j^{x_i}$ is the coefficient associated to that binary variable.

Prior to scorecard estimation, the numerical inputs must be binned. This process consists in grouping the values that had similar event behavior in the target variable. In the present study, the cutoffs used maximized the Weight of Evidence (WOE), which is a metric for variable Information Value (IV)

(Zeng, 2014). The binning outcome are new categorical input variables, which are then used in a stepwise selection algorithm. Regarding the score points, they increase as the event rate decreases. The estimation parameterization ensures that a score of 200 represents odds of 50 to 1, that is $\frac{P(Non-default)}{P(Default)} = 50$, and an increase of 20 score points corresponds to twice the odds.

Decision Tree

In the paper from Debeljak et al, , a classification decision tree (DT) is used to find the sequence of rules on the input variables that might predict the target (Debeljak & Džeroski, 2011; Pradhan, 2013). This way, it may be considered a Boolean function where the input is a set of hierarchical rules and the output is the final decision:

$$f : \{0,1\}^n \rightarrow \{0,1\} \tag{4}$$

To determine which input is selected to integrate the next rule the variance criterion was used for the numerical variables and the entropy for the categorical. Each splitting rule could only produce a maximum of two subsets, and each subset could not have less than 5 instances. To avoid oversized structures, and possibly overfitting, each tree was limited to 10 branches.

Multilayer Perceptron (Artificial Neural Network)

The Multilayer Perceptron (MLP) is a specific artificial neural network with at least three layers (input, output and hidden layer) were each node connects with every node in the following layer. Due to the simplicity of the network's architecture, MLP is often referred to as "vanilla" neural network (Hastie, Tibshirani, & Friedman, 2009).

Given this is an iterative algorithm, we start by defining how to obtain x_i^j, which is the neuron i from layer j:

$$x_i^j = f^j\left(b_i^j + \sum_{l=1}^{k} w_{li}^j x_l^{j-1}\right) \tag{5}$$

where f^j is the activation function of neurons in the j^{th} layer (in our case we will use always the same activation function for each neuron in a specific layer), $w_{li}^{(j)}$ is the weight associated to input l and neuron i at j layer, b_i^j the bias associated to neuron i at j layer and x_l^{j-1} the input variable l from the previous layer. The formula (5) may be generalized to express all neurons in each layer:

$$\underline{x}^j = f^j\left(\underline{b}^j + \underline{\underline{w}}^j \underline{x}^{j-1}\right) \tag{6}$$

where $\underline{x}^j$ is a vector containing all neurons from layer j, f^j is the activation function of neurons in the j^{th} layer, $\underline{b}^j$ the bias vector from layer j, $\underline{\underline{w}}^j$ the weight matrix from layer j and $\underline{x}^{j-1}$ the input variables vector from the previous layer.

The MLP used in this paper was designed with three hidden layers, each with 3 neurons and the $Tanh$ activation function.

3.3. COMMONLY USED ENSEMBLE LEARNING TECHNIQUES

The idea behind ensemble algorithms is to combine multiple base learners to improve the final model predictive power. By this,it is possible to achieve better performance than by considering solely each of the base learners (Opitz & Maclin, 1999; Polikar, 2006; Rokach, 2010). In the following subsections, some technics commonly used are going to be addressed.

Bootstrap aggregating (bagging)

In this method, a base learner is used in several random samples of the training set. The prediction is obtained by averaging the base learner's outcome in different random samples. The random forest algorithm is widespread use of bagging (Breiman, 1996a).

Boosting

Boosting is an iterative method that increases the weight of misclassified observations using the previous step results. The objective is to decrease the bias error, yet it may lead to overfitting to the training data. A very well-known application of this method is Adaboost (Breiman, 1996b; Schapire, 1990).

Stacking

This method consists of using a base learner to combine the output from different learners. Firstly, all base learners are trained using the training sample. Secondly, each outcome is used as input in a combiner algorithm. The most used combiner is a logistic regression. Comparing with the previous methods, Stacking has the advantage of being able to reduce both error and variance (Wolpert, 1992).

3.4. THREE-STAGE ENSEMBLE ALGORITHM

The proposed algorithm is a Three-Stage Ensemble Model (3SEM) which reinforces the predictive power of a Scorecard without compromising its transparency and interpretability. The concept is based on the idea of achieving a better performance, using several algorithms combined to outperform each one applied individually (Rokach, 2010). Firstly, it is used a Scorecard (SC) model to estimate the probability of default. Secondly, the SC Residual is used as target variable by another base learner. Thirdly, the SC estimate (first step) and SC Residual (second step) are combined using logistic regression. Thus, 3SEM might be considered a variation of the Stacking method.

The objective of the 3SEM is to let SC capture the linear effect, while the base learner algorithm covers the remaining variability. The proposed architecture for the Ensemble Model is presented in Figure 1:

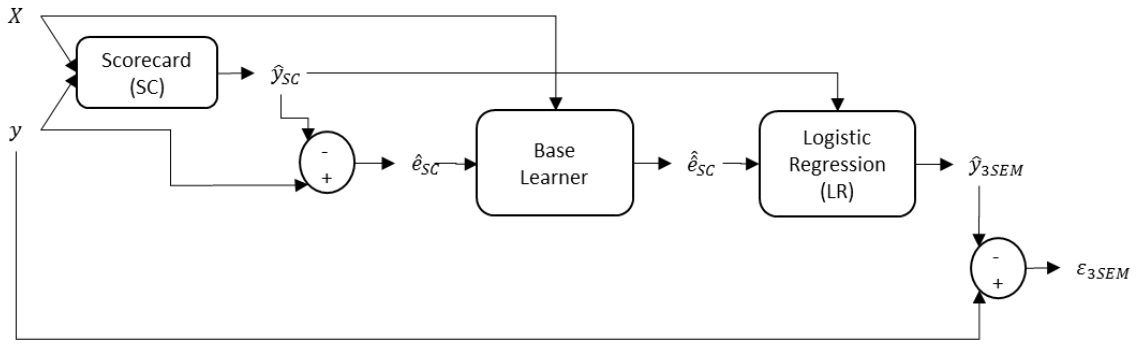

Figure 1: Proposed architecture for the three-stage ensemble model (3SEM)

In Figure 1, X is the set of inputs, y the target variable, while $\hat{y}$ and $\hat{\hat{e}}$ are the target and residual estimates, respectively. The box operator stands for a specific algorithm (for example SC or LR) and the circle is a sum operator with a sign for each of the variables. The components of Figure 1 are better described in Table 1.

Component	Description
y	Target variable
X	Input variables
$\hat{y}_{SC}$	Scorecard estimate
$\hat{e}_{SC}$	Scorecard residual
$\hat{\hat{e}}_{SC}$	Scorecard residual estimate
ε_{3SEM}	Three-Stage Ensemble Model error
$\hat{y}_{3SEM}$	Three-Stage Ensemble Model estimate

Table 1: Three-stage ensemble model components description

This ensemble architecture may also be defined through the mathematical formula(7):

$$y_{3SEM} = P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \hat{y}_{SC} + \beta_2 \hat{e}_{SC} + \varepsilon_{3SEM})}} \quad (7)$$

To estimate (7) the following steps should be done:

1. Estimate a Scorecard ($\hat{y}_{SC}$) using the available data (inputs) and the target (y);
2. Compute the Scorecard residual ($\hat{e}_{SC}$):

$$\hat{e}_{SC} = y - \hat{y}_{SC} \tag{8}$$

3. Estimate the base learner using $\hat{e}_{SC}$ as target and the inputs used on step 1. This estimate is noted $\hat{\hat{e}}_{SC}$;
4. Estimate a Logistic Regression using the target (y), Scorecard estimate ($\hat{y}_{SC}$) and Scorecard residual estimate ($\hat{\hat{e}}_{SC}$) as inputs.

3.5. PERFORMANCE METRICS

Following Hamdy & Hussein (2016) performance assessment approach, we will rely on Confusion Matrix and ROC index (area under the ROC curve) to compare the predictive quality of the 3SEM and the base learners. These two methodologies are presented in the following subsections.

Confusion Matrix

The Confusion Matrix is a very widespread concept that allows a more detailed analysis of the right and wrong predictions. As depicted in Table 2, there are two possible predictive classes and two actual classes. The combination of these classes originates four possible outcomes: True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) (Powers, 2011).

		Prediction	
		Default	**Non-Default**
Actual	**Default**	True Positive (TP)	False Negative (FN)
	Non-Default	False Positive (FP)	True Negative (TN)

Table 2: Confusion Matrix

These classifications have the following meaning:

- True Positive: includes the observations predicted as default and are actually default;
- False Positive: includes the observations predicted as default but are actually non-default (error type I);
- True Negative: includes the observations predicted as non-default and are actually non-default;
- False Negative: includes the observations predicted as non-default but are actually default (error type II).

To ease up the matrix interpretation the following measures may be computed (Powers, 2011):

Metric	Description	Formula
Accuracy	determines how often the classifier is correct	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	determines how often the classifier is correct predicting Defaults	$\frac{TP}{TP+FN}$
Specificity	determines how often the classifier is correct predicting Non-Defaults	$\frac{TN}{TN+FP}$
Positive Predictive Value	is the proportion of true positive prediction in all defaults	$\frac{TP}{TP+FP}$
Negative Predictive Value	is the proportion of true negative prediction in all non-defaults	$\frac{TN}{TN+FN}$

Table 3: Performance metrics based on th Confusion Matrix

Among the previous metrics, accuracy is easy to understand and takes a central place in the literature (Louzada et al., 2016). However, this metric must be used carefully, especially on unbalanced datasets. For example, in a data set with 1% event rate, a unary prediction of non-event would have an accuracy of 99%, higher than most stochastic models. Clearly, this metric is not robust for comparisons between algorithms applied on datasets with different event rate. However, we may use it to compare models on the same data set.

ROC index

Another measure for assessing predictive power is the Area Under Curve (AUC) Receiver Operating Characteristic (ROC). The curve is created by plotting the true positive rate (Sensitivity) against the false positive rate (1- Specificity) at various cutoff points. The true positive rate is the probability of identifying a default, while the false positive rate is the probability of a false alarm. The AUC=0.5 (random predictor) is used as a baseline to see whether the model is useful or not (Provost & Fawcett, 2013).

The ROC index has the advantage of not requiring the cutoff definition, as the confusion matrix demands. Besides, it is also suitable for unbalanced datasets (Hamdy & Hussein, 2016). However, the use of the ROC Curve as unique misclassification criterion has decreased significantly in the articles over the years. More recently the use of metrics based on the confusion matrix is most common (Louzada et al., 2016).

4. EMPIRICAL STUDY

4.1. STUDY DESIGN

The empirical study consists of estimating the Three-Stage Ensemble Model (3SEM) and compare the results with the base learner included in the algorithm. In this sense, three base learners are used: Logistic Regression (LR), Decision Tree (DT) and Multi-Layer Perceptron (MLP) neural network.

According to Louzada et al. (2016), almost 45% of the reviewed papers in their survey consider either Australian Credit Approval Data Set (AU) or German Credit Data Set (DE). To ensure that our results are replicable and comparable, we use datasets from the University of California at Irvine (UCI) Machine Learning Repository.

These datasets were split into a training set (80%) and testing set (20%) using stratified sampling on the target variable to ensure its representativeness. This procedure is widely used in previous studies and it is meant to improve the assessment metrics quality (Antunes, Ribeiro, & Pereira, 2017; du Jardin, 2016). Furthermore, the 10-fold cross-validation method was used to minimize the influence of variability in the training set (Olson, Delen, & Meng, 2012; Wang, Ma, & Yang, 2014). According to Kohavi (1995) this is the best method for model selection.

4.2. DATASETS

In our study two widely used data sets were employed, the Australian Credit Approval Data Set (AU) and the German Credit Data Set (DE). Both datasets can be found at the UCI Repository of Machine Learning Databases (Lichman, 2013).

The AU has 690 instances, being 307 of good applicants, a binary target and 14 input variables, where 8 are numerical. The DE consists of 1000 records, where 30% are bad applicants, and 20 input variables are available to describe the applicant socio-economical and behavioral attributes. Unlike the previous data set, most variables are categorical (13).

Regarding the cost matrix, AU does not have one while for DE it is recommended to have a five-fold impact of failing in predicting a default against mislabeling a non-default. Nevertheless, the use of any cost matrix is outside the scope of this paper, which means that both failing to predict a default and a non-default have the same cost.

The basic details of these data sets are shown in Table 4.

Characteristics	AU	DE
Number of Instances	690	1000
Number of good applicants	307	700
Number of bad applicants	383	300
Number of categorical attributes	8	13
Number of numerical attributes	6	7

Table 4: Basic details of the data sets

4.1. RESULTS AND DISCUSSION

In this section, we compare each base learner with the homologous 3SEM. To illustrate the discriminative power of the model, the default rate distribution through predicted status (probability of default) is presented in Figure 2. The test data set was ascending sorted by predicted status, thus the default rate is expected to be monotonically increasing. This is usually a requirement in a probability of default model.

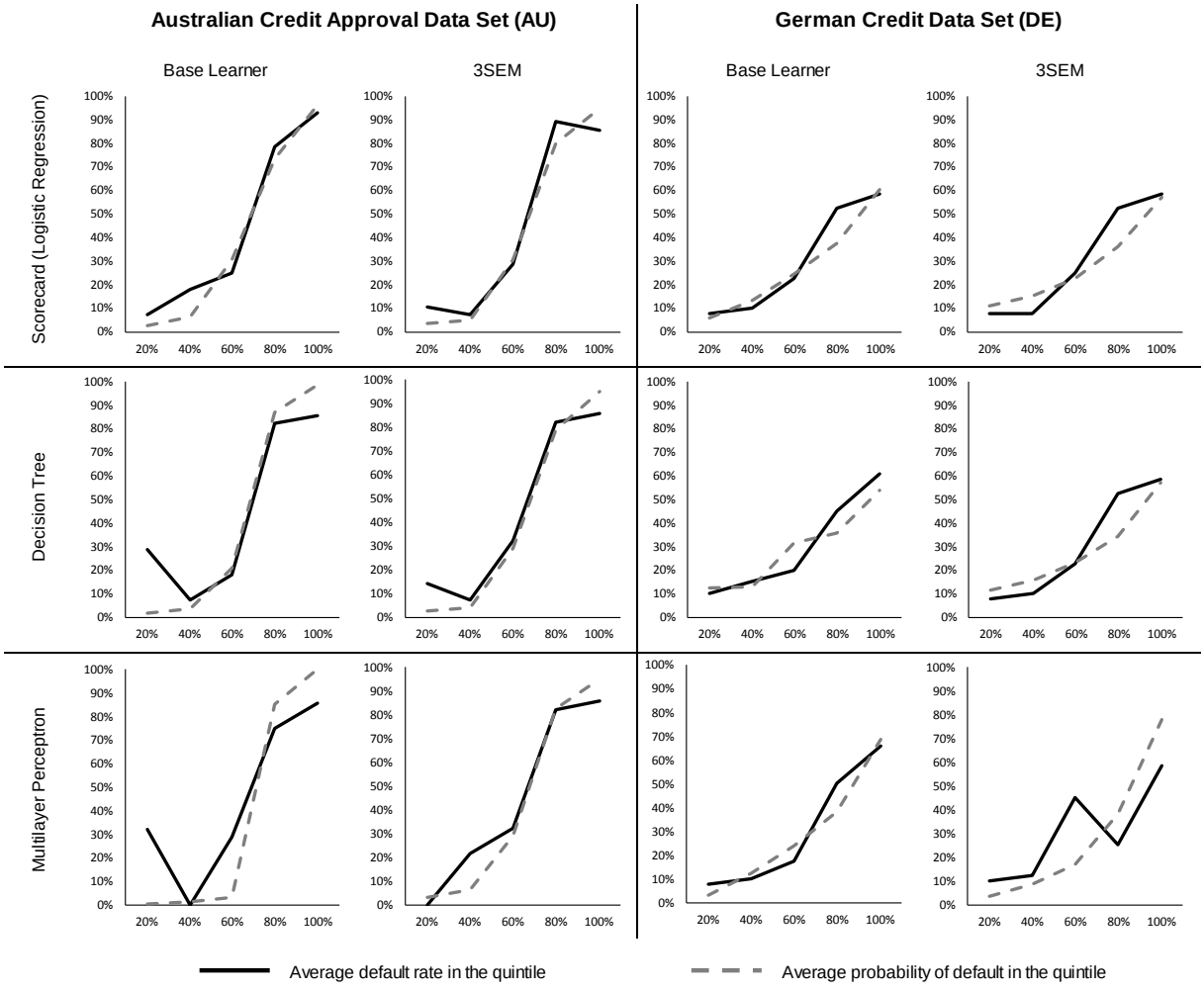

Figure 2: Default rate and predicted status (probability of default) distribution through quintiles for the 10th fold

In Figure 2, the quintiles are presented in the x-axis while y-axis show the default rate and predicted status (probability of default) distribution through quintiles for the 10th fold.Analyzing these plots, we identify that for AU only SC and 3SEM MLP have a monotonic default distribution, while on DE only 3SEM LR and 3SEM MLP violate this behavior.

However, classification accuracy along with interpretability are the most important criteria in choosing the credit classification approach (Zhu, Li, Wu, Wang, & Liang, 2013). Thus, the performance of 3SEM is compared with three base learners.

	Australian Credit Approval Data Set			German Credit Data Set		
	Base Learner	3SEM	Δ	Base Learner	3SEM	Δ
Logistic Regression	**88.80**	**89.40**	0.60	80.30	**80.70**	0.40
Decision Tree	82.60	85.50	2.90	75.60	80.30	4.70
MLP	77.90	89.10	11.20	**81.40**	72.60	-8.80

Table 5: Area Under the ROC Curve (ROC Index) for AU and DE

The results in Table 5 show that the 3SEM improves the ROC index in all base learners except for MLP in DE. Additionally, Logistic Regression seems to be the winner algorithm for both AU and DE. In the case of LR, the 3SEM improves the ROC index by 0.7% (+0.6pp) in AU and 0.5% (+0.4pp) in DE.

	Australian Credit Approval Data Set			German Credit Data Set		
	Base Learner	3SEM	Δ	Base Learner	3SEM	Δ
Logistic Regression	**85.00**	**85.00**	0.00	**74.53**	**75.27**	0.75
Decision Tree	81.43	80.71	-0.71	72.79	74.68	1.89
MLP	80.00	84.29	4.29	73.83	73.73	-0.10

Table 6: Accuracy (%) for AU and DE

Regarding Accuracy, the results in Table 6 reinforce the previous findings. Also, we conclude that the algorithm fitting is not consistent between data sets. Namely, MLP seems to be the least adjusted model for AU, while for DE is the best base learner (without considering 3SEM). However, Logistic Regression reinforces its place as winner algorithm and that 3SEM improves the accuracy of this estimator for DE by 1% (+0.75pp).

The results obtained are now compared with 30 papers reviewed by Louzada et al. (2016). Table 7 summarizes the accuracy on AU and DE in these papers.

Paper	AU	DE	Paper	AU	DE
Baesens et al. (2003)	90.40	74.60	Nieddu et al. (2011)	87.30	79.20
Hsieh (2005)	98.00	98.50	Marcano-Cedeno et al. (2011)	92.75	84.67
Somol et al. (2005)	92.60	83.80	Ping and Yongheng (2011)	87.52	76.60
Lan et al. (2006)	86.96	74.40	Yu and Li (2011)	85.65	72.60
Hoffmann et al. (2007)	85.80	73.40	Chang and Yeh (2012)	85.36	77.10
Huang et al. (2007)	87.00	78.10	Wang et al. (2012)	88.17	78.52
Tsai and Wu (2008)	97.32	78.97	Hens and Tiwari (2012)	85.98	75.08
Tsai (2008)	90.20	79.11	Vukovic et al. (2012)	88.55	77.40
Tsai (2009)	81.93	74.28	Marques et al. (2012)	86.81	76.60
Luo et al. (2009)	86.52	84.80	Ling et al. (2012)	87.85	79.55
Lahsasna et al. (2010)	88.60	75.00	Sadatrasoul et al. (2015)	84.83	73.51
Chen and Li (2010)	86.52	76.70	Zhang et al. (2014)	88.84	73.20
Zhang et al. (2010)	91.97	81.64	Liang et al. (2015)	86.09	74.16
Liu et al. (2010)	86.84	75.75	Tsai et al. (2014)	87.23	76.48
Wang et al. (2011)	86.57	76.30	Zhu et al. (2013)	86.78	76.62

Table 7: Accuracy (%) results for AU and DE according to Louzada et al. (2016)

Comparing the results obtained with some other studies we identify that the 3SEM's accuracy is not very high. The accuracy obtained in the literature ranges from 81.93% - 98% in AU and 72.6% - 98.5% while 3SEM got 85% and 75.27%, respectively. So, despite the proposed ensemble architecture potential, there is still room for future developments to improve its predictive power.

4.2. FURTHER WORK

This study has some limitations that give space for further research. On the one hand, the results should be verified using other data sets. Recalling the Supervised Learning No-Free-Lunch (NFL) theorems (Wolpert, 2002), there is no overall best statistical technique used in building models, thus the best technique always depends on the data set specificities. We expect that regardless of the data used, the performance of the 3SEM will always be at least as good as the best algorithm that integrates it (LR, DT and MLP). However, it is still to be determined the propensity for overfitting.

On the other hand, in the context of machine learning, there is a multiplicity of classification algorithms. The selected algorithms, accompanied by the set of choices such as the activation function or the neuron structure (MLP), are only illustrative. We stress that there is no hard evidence that the algorithms used are the best fit. Thus, in future work, there is room for further study using other techniques.

Finally, a generalization of the ensemble architecture should be developed, turning the algorithm into an n-stage ensemble model. In this approach, the researcher would apply more powerful methods from layer to layer. Once residuals are used as the target in the next layer, the largest fit is obtained in the first layers. Thus, the simplest algorithms produce the majority of the prediction, preserving most of the interpretability.

5. CONCLUSION

Credit scoring models attempt to measure the risk of a customer failling to pay back a loan based on his characteristics. In the banking industry, the most popular model is the scorecard which conciliates predictive and interpretative powers. Notice that banks are required to explain the credit application decisions, thus transparency is fundamental to these models. In this paper, we propose a new ensemble framework for the credit-scoring model to reinforce the predictive capacity of a scorecard without compromising its transparency and interpretability.

Our three-stage ensemble model consists on a Stacking of the Scorecard estimate and the Scorecard residual estimate, obtained through a base learner. Thus, the Scorecard estimate accounts for the majority of 3SEM predictive power, while the base learner aims to help to correct the prediction failures. This ensemble framework may be considered as estimation by layers, where modeling is done using more powerful methods from layer to layer. The advantage of this approach lies in the use of residuals as the target in the next layer. As the largest fit is obtained in the first layers, the majority of the model components is produced by the simplest algorithms, preserving the interpretability of most of the prediction.

Results indicate that the default rate distribution produced by the Scorecard is monotonic, which is usually a requirement in the probability of default models, yet there is no evidence that 3SEM keeps this behavior. Furthermore, the ROC index improves by 0.7% (+0.6pp) in AU and 0.5% (+0.4pp) in DE. However, Accuracy only improves in DE by 1% (+0.75pp).

Several other paths are still open. Firstly, other algorithms and parameterizations may be tested to check if the second stage contribution may be improved as there is no hard evidence that the algorithms used are the best fit. Secondly, a generalization of the ensemble architecture should be developed, turning the algorithm into an n-stage ensemble model. Finally, the results should be obtained also for other data sets, to ensure that they are not a lucky guess.

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. https://doi.org/10.1002/isaf.325

Anderson, T. W. (1978). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Antunes, F., Ribeiro, B., & Pereira, F. (2017). Probabilistic modeling and visualization for bankruptcy prediction. *Applied Soft Computing*, *60*, 831–843. https://doi.org/10.1016/j.asoc.2017.06.043

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Basel Committee on Banking Supervision. (1999). Credit Risk Modelling: Current Practices and Applications. *Bank for International Settlements*.

Basel Committee on Banking Supervision. (2003). The New Basel Capital Accord. *Bank of International Settlements*.

Basel Committee on Banking Supervision. (2004). International Convergence of Capital Measurement and Capital Standards. *Bank for International Settlements*.

Breiman, L. (1996a). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (1996b). *Bias, variance and arcing classifier*. Berkeley.

Brill, J. (1998). The Importance of Credit Scoring Models in Improving Cash Flow and Collections. *Business Credit*, *100*(1).

Chang, S.-Y., & Yeh, T.-Y. (2012). An artificial immune classifier for credit scoring analysis. *Applied Soft Computing*, *12*(2), 611–618. https://doi.org/10.1016/J.ASOC.2011.11.002

Chen, F.-L., & Li, F.-C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, *37*(7), 4902–4909. https://doi.org/10.1016/J.ESWA.2009.12.025

Debeljak, M., & Džeroski, S. (2011). Decision Trees in Ecological Modelling. In *Modelling Complex Ecological Dynamics* (pp. 197–209). Springer. https://doi.org/10.1007/978-3-642-05029-9_14

Dong, G., Lai, K. K., & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, *1*(1), 2463–2468. https://doi.org/10.1016/J.PROCS.2010.04.278

du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research*, *254*(1), 236–252. https://doi.org/10.1016/J.EJOR.2016.03.008

Gestel, T. Van, Baesens, B., Dijcke, P. Van, Suykens, J. A. K., Garcia, J., & Alderweireld, T. (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. *Journal of Credit Risk*, *1*(4).

Gonçalves, E., & Gouvêa, M. (2007). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. *POMS 18th Annual Conference*.

Hamdy, A., & Hussein, W. B. (2016). Credit Risk Assessment Model Based Using Principal component Analysis And Artificial Neural Network. *MATEC Web of Conferences*, *76*, 02039. https://doi.org/10.1051/matecconf/20167602039

Hand, J., & Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Computer Journal of the Royal Statistical Society Series a Statistics in Society*, *160*(3), 523–541.

Hastie, T., Tibshirani, R., & Friedman, J. H. (Jerome H. . (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). Springer-Verlag New York. https://doi.org/10.1007/978-0-387-84858-7

Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications*, *39*(8), 6774–6781. https://doi.org/10.1016/J.ESWA.2011.12.057

Hoffmann, F., Baesens, B., Mues, C., Van Gestel, T., & Vanthienen, J. (2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European Journal of Operational Research*, *177*(1), 540–555. https://doi.org/10.1016/J.EJOR.2005.09.044

Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications*, *28*(4), 655–665. https://doi.org/10.1016/J.ESWA.2004.12.022

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. https://doi.org/10.1016/J.ESWA.2006.07.007

Kohavi, R., & Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1137--1143.

Lahsasna, A., Ainon, R. N., & Wah, T. Y. (2010). Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier. *Maejo International Journal of Science and Technology*, *4*(1), 136–158.

Lan, Y., Janssens, D., Chen, G., & Wets, G. (2006). Improving associative classification by incorporating novel interestingness measures. *Expert Systems with Applications*, *31*(1), 184–192. https://doi.org/10.1016/J.ESWA.2005.09.015

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Liang, D., Tsai, C.-F., & Wu, H.-T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*, *73*(1), 289–297. https://doi.org/10.1016/j.knosys.2014.10.010

Liberati, C., Camillo, F., & Saporta, G. (2017). Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. *Advances in Data Analysis and Classification*, *11*(1), 121–138. https://doi.org/10.1007/s11634-015-0213-y

Lichman, M. (2013). UCI machine learning repository. *University of California, School of Information and Computer Science*.

Ling, Y., Cao, Q., & Zhang, H. (2012). Credit Scoring Using Multi-Kernel Support Vector Machine And Chaos Particle Swarm Optimization. *International Journal of Computational Intelligence and*

Applications, *11*(03), 1250019. https://doi.org/10.1142/S1469026812500198

Liu, X., Fu, H., & Lin, W. (2010). A Modified Support Vector Machine model for Credit Scoring. *International Journal of Computational Intelligence Systems*, *3*(6), 797. https://doi.org/10.2991/ijcis.2010.3.6.10

Louzada, F., Ara, A., & Fernandes, G. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

Luo, S.-T., Cheng, B.-W., & Hsieh, C.-H. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. *Expert Systems with Applications*, *36*(4), 7562–7566. https://doi.org/10.1016/J.ESWA.2008.09.028

Marcano-Cedeño, A., Marin-De-La-Barcena, A., Jimenez-Trillo, J., Piñuela, J. A., & Andina, D. (2011). Artificial Metaplasticity Neural Network Applied To Credit Scoring. *International Journal of Neural Systems*, *21*(04), 311–317. https://doi.org/10.1142/S0129065711002857

Marqués, A. I. I., García, V., & Sánchez, J. S. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, *39*(11), 10244–10250. https://doi.org/10.1016/j.eswa.2012.02.092

Nieddu, L., Manfredi, G., D'Acunto, S., & la Regina, K. (2011). An optimal subclass detection method for credit scoring. *International Journal of Economics and Management Engineering*, *75*, 349–354.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464–473. https://doi.org/10.1016/J.DSS.2011.10.007

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Ping, Y., & Yongheng, L. (2011). Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications*, *38*(9), 11300–11304. https://doi.org/10.1016/J.ESWA.2011.02.179

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences*, *51*, 350–365. https://doi.org/10.1016/J.CAGEO.2012.08.023

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. (M. Loukides & M. Blanchette, Eds.) (1st ed.). Sebastopol: O'Reilly Media, Inc.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Sadatrasoul, S., Gholamian, M., & Shahanaghi, K. (2015). Combination of feature selection and optimized fuzzy apriori rules: The case of credit scoring. *International Arab Journal of*

Information Technology, *12*(2), 138–145.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Somol, P., Baesens, B., Pudil, P., & Vanthienen, J. (2005). Filter- versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems*, *20*(10), 985–999. https://doi.org/10.1002/int.20103

Tsai, C.-F. (2008). Financial decision support using neural networks and support vector machines. *Expert Systems*, *25*(4), 380–393. https://doi.org/10.1111/j.1468-0394.2008.00449.x

Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, *22*(2), 120–127. https://doi.org/10.1016/J.KNOSYS.2008.08.002

Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, *24*, 977–984. https://doi.org/10.1016/J.ASOC.2014.08.047

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649. https://doi.org/10.1016/J.ESWA.2007.05.019

Vukovic, S., Delibasic, B., Uzelac, A., & Suknovic, M. (2012). A case-based reasoning model that uses preference theory functions for credit scoring. *Expert Systems with Applications*, *39*(9), 8389–8395. https://doi.org/10.1016/J.ESWA.2012.01.181

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230. https://doi.org/10.1016/J.ESWA.2010.06.048

Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61–68. https://doi.org/10.1016/J.KNOSYS.2011.06.020

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361. https://doi.org/10.1016/J.ESWA.2013.09.033

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*(11–12), 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wolpert, D. H. (2002). The Supervised Learning No-Free-Lunch Theorems. In *Soft Computing and Industry* (pp. 25–42). London: Springer London. https://doi.org/10.1007/978-1-4471-0123-9_3

Yu, J.-L., & Li, H. (2011). On Performance of Feature Normalization in Classification with Distance-Based Case-Based Reasoning. *Recent Patents on Computer Science*, *4*(3), 203–210. https://doi.org/10.2174/2213275911104030203

Zeng, G. (2014). A Necessary Condition for a Good Binning Algorithm in Credit Scoring. *Applied Mathematical Sciences*, *8*(65), 3229–3242. https://doi.org/10.12988/ams.2014.44300

Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, *37*(12), 7838–7843.

https://doi.org/10.1016/J.ESWA.2010.04.054

Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. *European Journal of Operational Research*, *237*(1), 335–348. https://doi.org/10.1016/J.EJOR.2014.01.044

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC Press.

Zhu, X., Li, J., Wu, D., Wang, H., & Liang, C. (2013). Balancing accuracy, complexity and interpretability in consumer credit decision making: A C-TOPSIS classification approach. *Knowledge-Based Systems*, *52*, 258–267. https://doi.org/10.1016/J.KNOSYS.2013.08.004