XXI SEMEAD
Seminários em Administração

novembro de 2018
ISSN 2177-3866

CREDIT DEFAULT PREDICTION MODEL VIA EXTREME GRADIENT BOOSTING WITH EMPIRICAL DATA FROM TAIWAN

VINICIUS MARRA
UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU)

Agradecimento à orgão de fomento:
This study was supported by Federal University of Uberlândia

CREDIT DEFAULT PREDICTION MODEL VIA EXTREME GRADIENT BOOSTING WITH EMPIRICAL DATA FROM TAIWAN

1. INTRODUCTION

Credit Scoring is concerned with assessing financial risks and supporting managerial decision making in the money lending business. So a credit score is an estimate of the probability that a borrower will show some undesirable behavior in the paying back the credit to the lender. This is a classic credit-scoring problem. (DIONNE, 2013)

As a result, there is a ongoing demand to automate credit-approval decision process in an attempt to improve the efficiency of risk management methods. This commitment for a bank as a lending strategy is emphasized by the regulatory framework of Basel II, which gave them a range of increasingly sophisticated options for calculating capital requirements. (BIS - BANK FOR INTERNATIONAL SETTLEMENTS, 2006). This ambition to avoid loss is simply a feature of prudent banking and helps the maintenance of first pillar of the accord, which deals with monitoring regulatory capital necessity.

The credit crisis of 2008 and its impacts focused a harsh spotlight onto credit management within the financial industry. The aftermath of the crisis showed that methods and systems employed should be reevaluated in a pursuit to improve the situation in sovereign credit risk management and to minimize possible losses with another economic turmoil. (ACHARYA; DRECHSLER & SCHNABL, 2014). Thus, it is beneficial to promote model researches that are capable of foreseeing risky clients that are not exclusively tied to the economic and financial circumstances surrounding a specific crisis.

The automation and improvement of these processes and models are especially decisive because of an ever-increasing amount and variety of data being generated regarding clients. Thus, systems that are able to predict defaults and distress are imperative so that both parties (lender and borrower) can take either preventive or corrective actions. (WANG, WANG & LAI, 2005; LAI *et al*. 2006b).

Therefore, to uncover new insights and expand analytical capacity, models that use Machine Learning offer considerable advantages over those that rely on human judgement or traditional statistical models. (CHEN & LIU; 2004). One of the main benefits of these methods is their ability to run across large volumes of data to predict an outcome and their relative lack of limitations. (HUANG, CHEN & WANG; 2007).

In a broad study, Jones, Johnstone, and Wilson (2015) compared the performance of models ranging from traditional classifiers (logit/probit and linear discriminant analysis) to machine learning classifiers, such as neural networks, support vector machines, and recent learning techniques such as generalized boosting, AdaBoost, and random forests. In their paper, they demonstrated that the latter outperformed all other methods.

Besides its efficiency in data analysis, machine learning has aspects that should be further researched to foment its adoption in credit lending companies. Wang and Ma (2011) point out that models should combine accuracy and usability. Khashman (2010) says that models should focus on designing, training and implementing systems with more outputs, which could indicate the reason why a credit application had been rejected, without having to understand its statistical calculations in the background.

Chen *et al*. (2011) suggest that improving the interpretability of ensembles is another important yet largely understudied research direction. Bae (2012) recommends exploring and building models on different datasets. This is a special issue since banks may be reluctant to disclose their costumers' information.

Guo *et al.* (2016) goes in the other direction insisting that models have been relying purely on numeric and financial variables, therefore, it was recommended to experiment with non-financial variables, such as: corporate governance-related factors (e.g., management ability, reputation, type of ownership, future plans, etc.), macroeconomic conditions on the corporate and consumer performance and even social data.

Kim and Kang (2010), Finlay (2011), Brown and Mues (2012), Tsai, Hsu and Yen (2014) and Kim, Kang and Kim (2015) recommended development of models taking in consideration boosting and bagging methods, which are based on a constructive strategy of formation.

This present study tested Extreme Gradient Boosting (XGBoost), a state-of-the-art machine learning method, which is used for supervised learning problems (Chen and Guestrin, 2016), which the term Gradient Boosting was proposed by Friedman (2001). XGBoost is an enhancement and based on his original model. We have chosen this model because of demonstrated efficiency, accuracy and practicability of its algorithm (Chen and Guestrin, 2016). Besides that, its capacity to do parallel computation on a commonplace machine causes it to be alluring. It also has additional features for doing cross validation and displaying important variables.

In our study, we confirmed its performance and accuracy when compared to benchmark models (Logistic and Random Forest). XGboost had and accuracy rate of 82.29% against 82.01% from Random Forest and 81.93% from Logistic Regression.

These findings should contribute to the literature on credit risk prediction in few ways. The application of an open-source method based on XGBoost represents an advance in the use of new techniques to predict consumer default probability. This becomes an important fact because the depiction of the algorithms are publicly available so that banks and future uses can avoid the "black box" concept that complex models have. Finally, we would also like to encourage the alliance of finance and computer science in pursuit of a structured, accurate decision-making and understanding process.

The remainder of the paper is organized as follows: Section 2 presents a review of the literature on machine learning. Section 3 describes the data. Section 4 discusses the method. Section 5 presents the results of the analysis. Section 6 offers the main conclusions and managerial implications of the paper.

2. LITERATURE REVIEW

Basel II Accord requires financial institutions to disclose Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD) in such a manner that accuracy and speed to analyze and predict data are paramount. Thus, an ever-increasing demand for smart and agile processes have exposed an opportunity for machine learning research in credit risk management.

Basel II also requires companies to disclose risk management practices, which demands more reliable and accurate models to classify and quantify risk. (BIS - BANK FOR INTERNATIONAL SETTLEMENTS, 2006). For this reason, the adoption of Machine Learning algorithms have gained a lot of attention within the financial industry. Machine Learning relates to the study of pattern recognition and computational learning theory in artificial intelligence.

This field explores the study and construction of algorithms that can learn from and make predictions on previously collected data (SAMUEL, 1959). In credit risk management, one could argue that each dataset is peculiar and unique in each circumstance, so the data relationships can be quite complex, non-normal, non-linear and reflect structural changes

such as demographic or market trends. Therefore, the construction and improvement of such models able to understand this dynamic is a continuous process. (GALINDO & TAMAYO, 2000).

Different studies have compared performance from different methods. Alfaro *et al.* (2008) confirmed that the AdaBoost outperforms Neural Networks and AdaBoost test error was 8.898% against an error of 12.712% for neural network. Heo and Yang (2014) compared several machine-learning algorithm success ratios and tested it against Altman's famous Z-score: AdaBoost (78.5%), ANN: (77.1%) SVM (73.3%), DT (73.1%) and Altman Z-score (51.3%).

On Table 1, we can see that various authors endeavored with different machine learning algorithm. These researchers have highlighted the capability of the models, but also pointed out their disadvantages, such as their obscure nature, greater computational burden, proneness to overfitting and empirical nature of construction.

Table 1 - Machine Learning Methods applied to credit risk prediction

Application in a credit risk context	Machine Learning Algorithm
Tsai and Wu, 2008; Chauhan, Ravi and Chandra, 2009; Kim and Kang, 2010; Du Jardin, 2010; Chuang and Huang, 2011; Marcano-Cedeño et al., 2011; Jeong, Min and Kim, 2012; Blanco et al., 2013; Lee and Wu Sung, 2013; López and Sanz, 2015; Zhao et al., 2015; Yu, Yang and Tang, 2016	Artificial Neural Networks
Sun and Li, 2012; Wang and Ma, 2012; Hens and Tiwari, 2012; Hsieh et al., 2012; Harris, 2015; Danenas and Garsva, 2015; Sun et al., 2017	Support Vector Machine
Li, Sun and Wu, 2010; Cho, Hong and Ha, 2010; Zhang et al., 2010; Gepp, Kumar and Bhattacharya, 2010; Wang et al. 2012; Kim and Upneja, 2014	Decision Trees
Sun, Jia and Li, 2011; Wang and Ma,2011; Wang, Ma and Yang, 2014; Kim and Upneja, 2014; Heo and Yang, 2014; Kim, Kang and Kim, 2015; Sun et al., 2017	Boosting Algorithms

Even though almost all methods can be used to assess credit risk, recently – due to the increasing complexity and size of datasets – researchers have been combining different classifiers and technics, which integrate two or more classification methods. These approaches have been showing higher precision in predictability than individual methods. Combination of classifiers have flourished in credit risk assessment. Some examples are

Neural Discriminant Technique (Lee *et al.,* 2002), neuro-fuzzy (PIRAMUTHU, 1999; MALHOTRA & MALHOTRA, 2002) and fuzzy SVM (WANG *et al*., 2005).

In real world credit dataset, we have a scenario where the number of observations associated to one class is rather lower than those belonging to the other class. For this reason, we have an imbalance issue, which Brown and Mues (2012) examined in their study investigating several kinds of credit scoring algorithms and results demonstrated that Random Forest – a decision tree based algorithm - and gradient boosting performed relatively well in imbalanced datasets.

3. THE MODEL

Our goal was to test a classification model to determine the PD of credit card clients, also monitor which variables should be observed to anticipate the event. Wang and Ma (2011) applied RS-Boosting and had better results mainly on reducing type II error, but from their work we detected that interpretability of ensembles is another important yet largely understudied research direction.

We compared the predictive ability of XGboost, which showed evidence of its accuracy in more than half of the winning solutions in machine learning challenges hosted at Kaggle (He, 2016).

In finance, the application of this model is relatively new. He, Zhang and Zhang (2018) compared XGboost performance against other models for credit scoring and obtained the best ranking results four times out of the six datasets, which indicates that it has excellent performance. Carmona, Climent and Momparler (2018) tested the model to predict failure in the U.S. banking sector and concluded that XGBoost has greater predictive power than both Logistic Regression and Random Forest methods. Xia *et al.* (2018) point out the superiority of the model as a meta-classifier. Xia *et al.* (2017) highlighted comparisons with different baseline models and showed the superiority of the XGBoost-based model in terms of predictive performance.

Contrasting bagging algorithm that fits the base models in parallel, boosting approach is to build models in sequential fashion, once a single regression tree is too weak to be used in practice. Therefore, the tree ensemble model sums the prediction of multiple trees. (FRIEDMAN, 2001). XGBoost uses K additive base learners $f_k(x)$ to approximate the final model $F_K(x)$ to minimize the loss function provided. The Gradient descent method calculates the partial derivative with respective to zero and tries to optimize the loss function by tuning different values of coefficients to minimize the error. This loss function measures how well the model fits the current data and the process of boosting continues until the loss function reduction becomes limited. (CHEN & GUESTRIN, 2016)

The final goal of a learning problem is to determine a roadmap, where $\boldsymbol{y}$ is the expected prediction and $\boldsymbol{x}$ are the characteristics vectors. (YUFEI *et al*., 2017). In order to build the map, the proposed model requires multiple parameters to be set. Controlling the proper combination of parameters is fundamental to optimize and improve the model. (Chen and Benesty, 2016). General parameters are discussed in the following topics:

3.1 Parameters

- *Number of rounds or maximum number of iterations*: the optimal number of rounds or trees required in XGBoost model;

- *Maximum depth or size of a tree:* is the number of splits in each tree. It is used to control overfitting because higher depth allows the model to learn relationships that are highly specific to a particular sample;
- *Learning rate*: first introduced by Friedman (2002), is generally a small positive number (ranging from 0 to 1) that determines how quickly the algorithm adapts or the contribution of each tree to the growing model. A low value means that the model is more robust to overfitting.
- *Gamma:* minimum loss reduction required to make the next split on a leaf node of the tree. The larger its value, the more conservative the algorithm will be.
- *Column and observation sample:* a subsample ratio of variables and observations when constructing each tree. The column and observation sample denotes the fraction of variables and observations, which should be randomly sampled for each tree. Their value ranges from 0 to 1 and prevents overfitting and speeds up computations of the algorithm.
- *The minimum child weight:* indicates the minimum sum of instance weight required in a child node. If the tree separation step results in a leaf node whose sum of precedent weight is less than the value assigned to this parameter, then the building process will stop further partitioning.
- *Regularization or penalty term on weights:* The regularization term controls the complexity of the model to help avoid overfitting.

4. EXPERIMENTAL SET-UP

4.1 Credit dataset

In this experiment, a real-world credit dataset is utilized to verify the performances of our model. The dataset for this project consists of publicly available information from credit card clients from Taiwan and it is available at UCI Machine Learning Repository. Yeh and Lien (2009) explored this dataset through six major classification techniques and concluded that ANN outperformed all other technics. The data set consists of 30.000 observations and 24 attributes containing gender, education profile, marital status, age, history of statement balance, payment status and binary status of default (1 or 0). The features used in this paper are similar to traditional credit scoring datasets, which include demographic variables, solvency, and creditworthiness of the borrower.

Table 2 - Characteristics of credit scoring data set.

Inputs	Data set size	Training set size	Test set size	Goods/bads
24	30.000	21.000	9.000	70/30

4.2 Benchmark models

Our main purpose is not only to contrast different machine learning methods, but also to shed some light on a recent model and motivate the union between finance researches and computer scientists. We compared XGboost with a conventional method (logistic regression) and a modern machine learning approach (the random forest algorithm). Besides its performance, XGboost as another important aspect which refers to identifying and displaying most important variables to the model.

Logistic regression is one of the most accepted and used technics on theoretical ground, given that two discrete classes (either good or bad) have been defined beforehand. (KIM, 2011; LI & SUN, 2011; Li *et al.*, 2011). Given a training set of N data points $D = \{(xi, yi)\}_{i=1}^{N}$, with input data $x_i \in R^N$ and corresponding binary class labels $y_i \in \{0,1\}$, the logistic regression approach to classification (*LOG*) and tries to estimate the probability P(y = 1|x) of good and bad clients as follows:

$$P(y = 0|\, x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (1)$$

Random Forest is an ensemble tree-based method that uses bagging to achieve diversified subsets of the entire training set to build individual trees. The algorithm is a classifier consisting of a selection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1,...\}$ where the $\{\Theta_k\}$ are independent distributed random vectors and each tree casts a unit vote for the most popular class at input **x**. (BREIMAN, 2001). This technique require tuning for two parameters, the number of trees and the number of attributes used to grow each tree.

5. RESULTS

This dataset provides a large and representative sample of Taiwanese credit card holders, which is a gain over other publicly available datasets drawn from financial institutions due to its sample size. On table 3 we have a simple descriptive statistic for all variables. For the response variable we have binary feature – default payment (Yes = 1, No = 0), followed by 23 explanatory variables.

Table 3 - Descriptive Statistics of Dataset

Descriptive Statistics					
Row	**Observations**	**Min.**	**Max**	**Mean**	**Std.dev**
DEFAULT	30000	0	1		
SEX	30000	1	2		
LIMIT_BAL	30000	10000	1000000	167484,32	129747,662
EDUCATION	30000	0	6	1,85	0,790
MARRIAGE	30000	0	3	1,55	0,522
AGE	30000	21	79	35,49	9,218
PAY_0	30000	-2	8	-0,02	1,124
PAY_2	30000	-2	8	-0,13	1,197
PAY_3	30000	-2	8	-0,17	1,197
PAY_4	30000	-2	8	-0,22	1,169
PAY_5	30000	-2	8	-0,27	1,133
PAY_6	30000	-2	8	-0,29	1,150
BILL_AMT1	30000	-165580	964511	51223,33	73635,861
BILL_AMT2	30000	-69777	983931	49179,08	71173,769
BILL_AMT3	30000	-157264	1664089	47013,15	69349,387
BILL_AMT4	30000	-170000	891586	43262,95	64332,856
BILL_AMT5	30000	-81334	927171	40311,40	60797,156
BILL_AMT6	30000	-339603	961664	38871,76	59554,108
PAY_AMT1	30000	0	873552	5663,58	16563,280
PAY_AMT2	30000	0	1684259	5921,16	23040,870
PAY_AMT3	30000	0	896040	5225,68	17606,961
PAY_AMT4	30000	0	621000	4826,08	15666,160
PAY_AMT5	30000	0	426529	4799,39	15278,306
PAY_AMT6	30000	0	528666	5215,50	17777,466

wwe tuned the XGBoost model parameters to fit the best model and identify structure in the data. Even though dataset partition and best parameter tuning deserve another discussion, this specific topic is beyond our interest in this study.

Four popular evaluation metrics – accuracy, Type I error rate, Type II error rate, area under the receiver operating characteristic curve (AUC) – were employed to assess the performance of models. Our main evaluation metric is AUC (Area under the curve), which is an alternative discrimination capability measure based on the receiver operating characteristic (ROC) curve. The ROC curve plots *true positive rate values* (TPR) against the *false positive rate values* (FPR) at various threshold settings. The true-positive rate is also known as sensitivity and false-positive rate is known as probability of false alarm and can be calculated as (1 − specificity). In other words, the AUC score measures how well the model discriminate between the two classes.

We have tune the parameters according to Carmona, Climent and Momparler (2018). Controlling parameters can avoid overfitting and can ensure its generalization. We trained

the model with 1000 iterations or rounds, a maximum tree depth of 5, learning rate of 0.1, a gamma of 0, a subsample ratio of variables of 0.8, a minimum child weight of 1 and lastly, a regularization value of 0.

Table 4 – XGboost Parameters

Parameters	Values
Number of Iterations	1000
Maximum depth	5
Learning rate	0,1
Gamma	0
Observation sample	0,8
Min. Child Weight	1
Regularization	0

After we evaluated the model, we tested it and on table 4 we can see the results from the performance metrics. We have assessed the model's performance on a dataset different from the one used to estimate it. Thus, we randomly divided the observations in 70% for training and 30% for testing the model. On the larger dataset, we trained and fit the XGBoost, while the second was used to test it. We conducted the dataset partition for both Logistic Regression and Random Forest in the same manner, 70% for training and 30% for testing the models.

In regards to evaluation metrics, we have displayed four different measures for all three models. On table 5 we have a confusion matrix with four basic elements: true positives (TP) indicating that the prediction of good credit and consistent with its real value; false negatives (FN) means that the prediction result of the sample is classified as bad credit but its real label indicates good credit. Likewise, false positives (FP) are those bad credit samples classified as good credit and those within bad credit samples correctly predicted as bad credit are labeled as true negatives (TN).

Table 5 - Confusion Matrix

		Predicted Values	
		Positive	Negative
Real Values	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

ACC is the measure of correct prediction of classifier compared to the overall data points. It is the ratio of the units of correctly predicted and total number of predictions made by the classifiers. We calculated ACC as follows:

$$ACC = \frac{TP+FN}{TP+FP+FN+TN} \quad \textbf{(2)}$$

The Under the ROC Curve (AUC) measures the ability of a binary machine learning model to predict a higher score for positive examples as compared to negative examples. Normally the threshold for two class is 0.5. The closer to limit value of 1, the better the algorithm classifies into the classes.

The Kolmogorov–Smirnov test (KS) was another metric used to measure the goodness-of-fit of the model, which is used for testing normality, where TPR stands for true positive rate and FRP is false positive rate.

$$KS = \max_{t}(|TPR(t) - FRP(t)|) \quad \textbf{(3)}$$

Lastly, type I and II error rates were used as indicators to further explore the label prediction capability of models over good and bad loans, respectively. In our study, Type I error rate denotes the proportion of misclassified good loans, and Type II error rate refer to the proportion of misclassified values. We computed their values as described in Eqs. **(4)** and **(5)**.

$$Error\ I = \frac{FP}{FP+TN} \quad \textbf{(4)}$$

$$Error\ II = \frac{FN}{TP+FN} \quad \textbf{(5)}$$

For the Taiwanese dataset, the XGBoost model achieves the best ACC (0,8229) with 0,0539 type I error and 0,8914 type II error. Random Forest performs relatively close to our proposed model with ACC score of 0,8201 and 0,0556 for Type I error and 0,8914 for Type II error.

Table 6 - Results Taiwanese dataset

Model	*ACC*	*AUC*	*KS*	*Type I Error*	*Type II Error*
Logistic	0,8193	0,7258	0,3840	0,0509	0,6286
Random Forest	0,8201	0,7652	0,4152	0,0556	0,8928
XGboost	0,8229	0,7725	0,4220	0,0539	0,8914

Note: ACC=accuracy, AUC=Area Under the Curve, KS = Kolmogorov-Smirnov

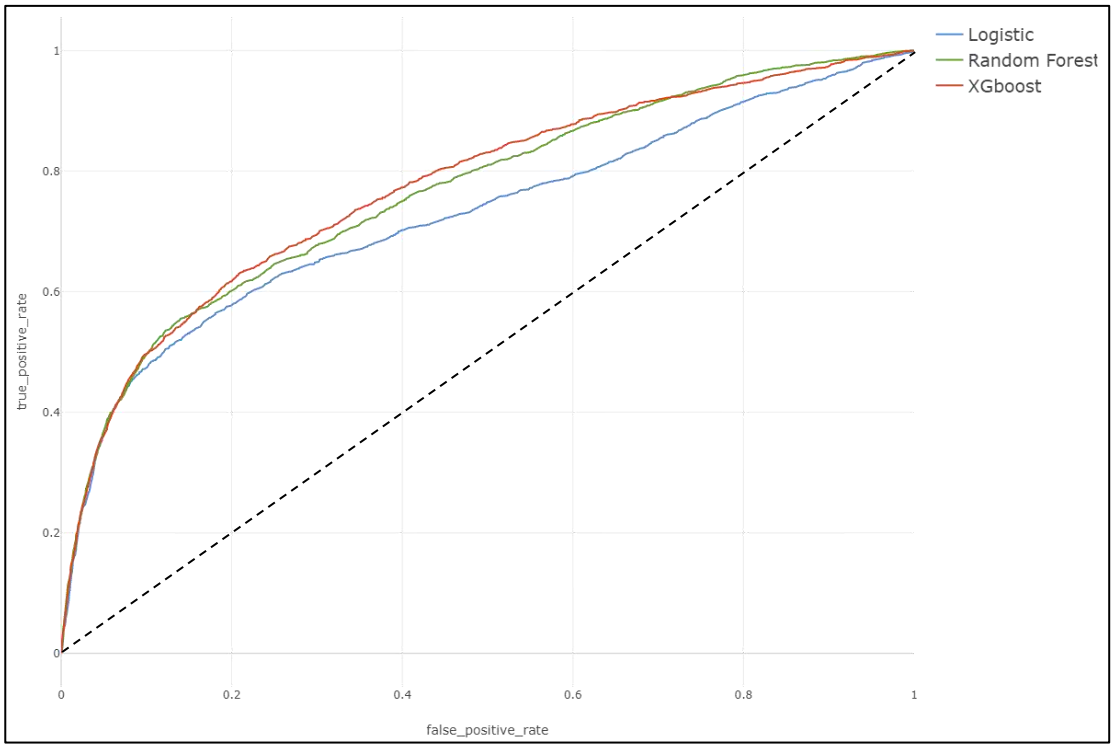

Fig. 1. ROC curves for each model.

5.1 Variable importance

Another important aspect of XGboost is that it displays the most important variables with a high relative influence on the response variable. Abdou (2009) suggested future researches to investigate the behavior of customers who had defaulted and determine in particular what variables may affect early default. This becomes an important matter because credit-lending companies could track and monitor specific features in relation to the timing of the loan period.

Chen and Li (2010) also stated that feature selection approach may uncover essential features and how these features affect the credit models. Fewer features mean that credit department can concentrate on collecting relevant and essential variables.

On Fig. 2 we can observe that variables “MARRIAGE” and “SEX” provide little discriminatory values, moreover companies should not solely consider in their credit policies factors over which they have little or no control, but further researches should be conducted on aspect related to these variables. Models should not treat equally a single mother as it treats men, as they tend to have higher earning power than women. By doing so, credit models would avoid penalizing vulnerable clients.

In regards to variables that contributes the most to the model, we can infer that a client that delays his first payment, “PAY_0”, will more likely default on remaining payments. In addition to that, the amount of spending also contributes to the probability of default, “BILL_AMT1”. “EDUCATION” has and increasing importance for XGboost.

When we contrast variable importance with a traditional econometric model, we note that “MARRIAGE”, “SEX” and “EDUCATION” offers no special information to Logistic Regression. “PAY_0” remains as most important variable however, “BILL_AMT1” does not contribute to the model the same way it does to XGboost.

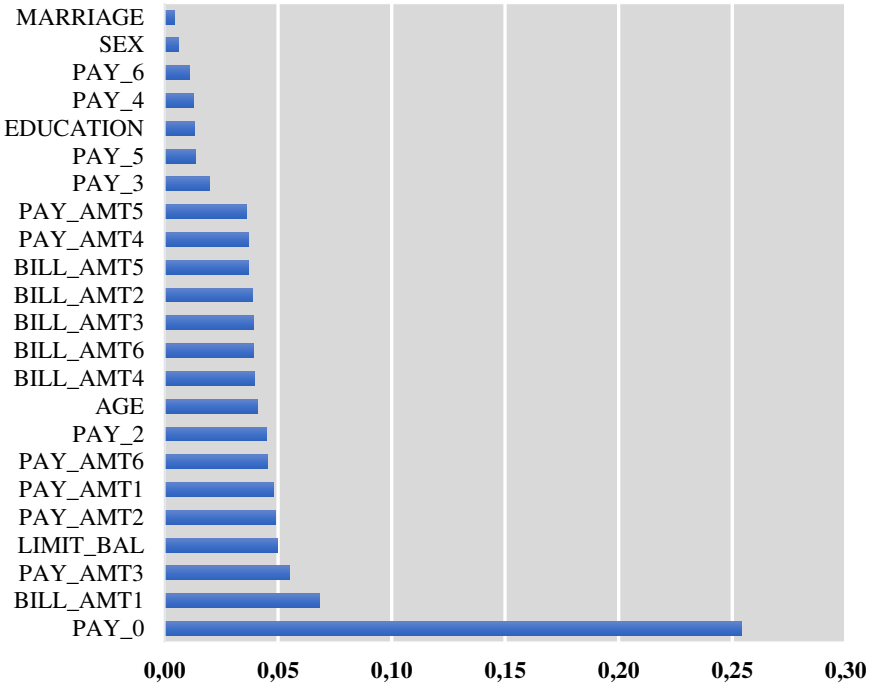

Fig. 2. Importance of variables (XGboost).

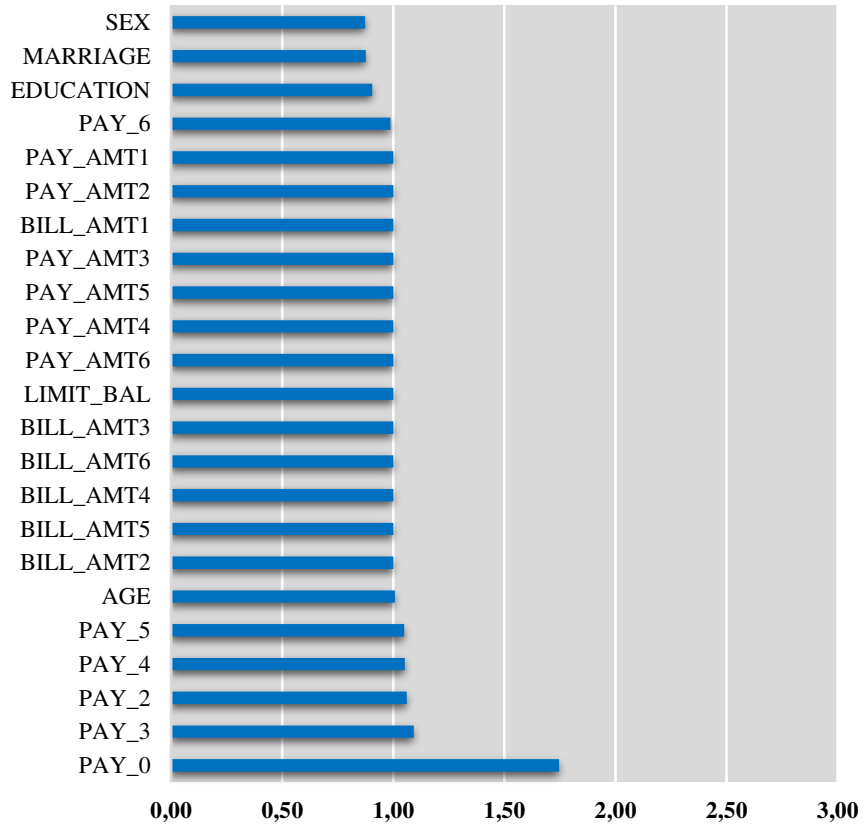

Fig. 3. Importance of variables (Logistic Regression).

6. CONCLUSIONS

Our main goal in this study was to predict probability of default payments on credit card clients. To this end, we tested the efficiency of a new machine learning model called XGboost. This method is an evolution of other boosting methods such as AdaBoost and boosted classification trees, and it has been applied in recent bank failure studies and credit scoring. As a secondary goal, this study focused on predictive power rather than exploring the construction of the model and we also wanted to highlight the benefits of machine learning algorithms applied in finance researches.

Our study showed that XGBoost has greater predictive power than both Logistic Regression and Random Forest methods, considering the parameters used. In addition to that, XGboost has an important feature. This aspect corroborates with Chen and Li (2010), Wang and Ma (2011), Tsai, Hsu and Yen (2014); Zhao et al. (2015) where they suggested that models and credit scoring results should focus on giving explanations on the reasons for rejection, which are important to both applicants and financial institutions. Since there is no ground truth answer to the most representative features (e.g. input variables), the proposed method demonstrated that delay on the first payment classifies a client with high probability of future default and the amount of spending also contributes to that.

The higher predictive power of the model tested in this study should encourage researches to join forces with computer scientists to add a dynamic to econometric models commonly used in the study of finances. Moreover, in an attempt to extend the current limits of performance and interpretability, XGboost tracks variables that can add an extra predictive weight to the model and operational agility.

From the results, we believe that credit-lending corporations can develop a cautionary system that would warn clients on behaviors that could affect their credit score. Adding to that, managers can avoid financial default by taking early appropriate action rather than waiting for the event to happen.

In conclusion, this study offers some interesting prospects for future researchers to enhance the model by:

- Expanding the dataset with different variables could improve model's robustness;
- Test more advanced base learners;
- Evaluate the optimal dataset split for training and testing the model;
- Experiment with different parameters in different datasets, such as corporate variables or emerging markets datasets;
- Include qualitative variables (e.g., social and behavioral information)

7. REFERENCES

Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. Expert systems with applications, 36(9), 11402-11417.

Acharya, V., Drechsler, I., & Schnabl, P. (2014). A pyrrhic victory? Bank bailouts and sovereign credit risk. The Journal of Finance, 69(6), 2689-2739.

Bae, J. K. (2012). Predicting financial distress of the South Korean manufacturing industries. Expert Systems with Applications, 39(10), 9159-9165.

Basel Committee on Banking Supervision. (2006). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive

Version, Bank for International Settlements. Available at: https://www.bis.org/publ/bcbs118.pdf. (accessed August 8, 2017).

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

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

Carmona, P., Climent, F., & Momparler, A. (2018). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics and Finance.

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.

Chen, H. L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S. J., & Liu, D. Y. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. Knowledge-Based Systems, 24(8), 1348-1359.

Chen, S. Y., & Liu, X. (2004). The contribution of data mining to information science. Journal of Information Science, 30(6), 550-558.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

Chen, T., & Benesty, M. (2016). XGBoost: eXtreme gradient boosting. R package version 0.4-3. https://cran.r-project.org/web/packages/XGBoost/vignettes/XGBoost.pdf.

Dionne, G. (2013). Risk management: History, definition and critique. Risk Management and Insurance Review, 16(2), 147-166.

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38(4), 367–378.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Computational Economics, 15(1-2), 107-143.

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From footprint to evidence: An exploratory study of mining social data for credit scoring. ACM Transactions on the Web (TWEB), 10(4), 22.

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. Expert Systems with Applications, 98, 105-117.

He, T. (2016). An Introduction to XGBoost R package. Available at:http://dmlc.ml/rstats/2016/03/10/XGBoost.html (accessed July 2016).

Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. Applied soft computing, 24, 494-499.

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.

Islam, S. R., Eberle, W., & Ghafoor, S. K. (2018). Credit Default Mining Using Combined Machine Learning and Heuristic Approach. arXiv preprint arXiv:1807.01176.

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.

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 Finance, 56, 72–85.

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

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. Expert systems with applications, 37(4), 3373-3379.

Kim, M. J., Kang, D. K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, 42(3), 1074-1082.

Kim, S. Y. (2011). Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. The Service Industries Journal, 31(3), 441-468.

Lai, K. K., Yu, L., Wang, S., & Zhou, L. (2006). Credit risk analysis using a reliability-based neural network ensemble model. In International Conference on Artificial Neural Networks (pp. 682-690). Springer, Berlin, Heidelberg.

Lee, T. S., Chiu, C. C., Lu C. J. & Chen, I. F. (2002). Credit Scoring Using the Hybrid Neural Discriminant Technique. Expert System with Applications, Vol. 23, No. 3, pp. 245-254.

Li, H., & Sun, J. (2011). Empirical research of hybridizing principal component analysis with multivariate discriminant analysis and logistic regression for business failure prediction. Expert Systems with Applications, 38(5), 6244-6253.

Li, H., Lee, Y. C., Zhou, Y. C., & Sun, J. (2011). The random subspace binary logit (RSBL) model for bankruptcy prediction. Knowledge-Based Systems, 24(8), 1380-1388.

Liu, H. W., & Chen, W. D. (2004). Neural Network Model for Credit Analysis of Tally Clients. Industrial Engineering Journal-Guangzou. 7(2), 25-28.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro- fuzzy systems. European Journal of Operational Research, 136, 190–211.

Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. European Journal of Operational Research, 112, 310–321.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research & development, 3(3), 210-229.

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.

Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. Expert Systems with Applications, 38(11), 13871-13878.

Wang, Y. Q., Wang, S. Y., & Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk. IEEE Transactions on Fuzzy Systems, 13, 820–831.

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Systems with Applications, 93, 182-199.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, 225-241.

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

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