RESEARCH ARTICLE OPEN ACCESS

Credit Risk Analysis Using Machine Learning Techniques

Omer Yavuz Can[1], Ahmet Gurhanli[2]
[1](Department of Computer Engineering, Istanbul Aydin University,
[2](Department of Computer Engineering, Istanbul Aydin University,
Corresponding Author: Omer Yavuz Can

ABSTRACT
It can be easily observed that the general public is putting in more and more loan requests in the banking system recently, which can be regarded as a positive development for the banks, while at the same time presenting a considerable risk. Accurate risk management in the banking and finance sector is related to efficient and optimized use of the current resources, assessment of possible risks and taking timely precautions. It is of utmost importance for the banks to predict the problematic loans in terms of long-term stability. Giving credits to the applicants is one of the fundamental activities of the banks, however; the same activity brings significant risks. As part of their founding purpose, the banks do not avoid taking risks, and they choose to manage them. The banks should perform their risk management in the way to keep the damages resulting from the amount of loans they give to a minimum. Considering the above and in order to speed up the lending procedures in banks while making advantageous decisions, different algorithmic models and classifications, machine learning techniques such as artificial neural networks were started to be used lately, data mining being at the first place. In this study, the accuracy of the applicants' eligibility status for loans was determined by making use of several machine learning techniques. The open-access dataset from the German Credit Data UCI was employed. Based on the 1000 customers in this study's dataset, a 75,60% success rate was achieved in the XGBoost classifier, which has the best success rate among the studies conducted with the XGBoost classifier previously. In addition, the success rate is the highest among the other algorithms used in various studies made.
Keywords – Credit, Credit Risk, Credit Risk Analysis, Machine Learning, Data Mining

---

DATE OF SUBMISSION: 03-01-2020 DATE OF ACCEPTANCE: 18-01-2020

I. INTRODUCTION

The bank loan requests have reached great numbers today and people are having difficulties paying their loans. This causes a considerably problematic course of events for both the banks and the loan applicants. Banks have started to use artificial intelligence and machine learning, which is a sub-branch of artificial intelligence lately, in order to resolve the situation. Basically, Artificial Intelligence is the algorithmic structure that enables a machine to find human-like solutions to a given problem [1]. At times, this was achieved by writing codes such as the programs that carry out mathematical computations, but as the problems became more complicated, different methods were started to be employed. Chess can be given as an example to the said problems. There are a lot of possibilities in chess which cannot be coded or that would take a very long time to code. Program 'Deep Blue,' which beat the World chess champion Garry Kasparov in 1996, being built for this purpose, was developed with machine learning. Machine learning is the data analysis technique wherein a computer program solves given problems itself, through experience, without human interaction [2]. Machine learning algorithms use calculation methods in order to 'learn' the data without relying on a model, or predetermined equation. As the number of present learning examples increase, the performance increases as well. Therefore, the banking and finance institutions are tending towards machine learning.

Logistic Regression

Logistic Regression Analysis was designed with the purpose of optimizing the relation between the variables and rendering them an acceptable model by using the least number of variables for the relation between dependent and independent variables [3].

Linear Discriminant Analysis

There are many possible techniques to be used for data classification. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the two most commonly used techniques for classifying data [4]. The goal of linear discriminant analysis is to assign the correct data classification with the least amount of errors [5].

K-Nearest Neighbor

In this method, classification is made according to the determined threshold value by

calculating the similarities between the data in the learning cluster and the actual behavior of the data, with the average of k data that is considered to be nearest to each other [6].

Decision Tree

Decision Tree takes the variables in the data cluster as a node. Branching is performed by checking the actualization status of the variables in the node. After all the data is classified, branching stops and the leaves of the tree constitutes the classification labels [7].

Naive Bayes

Naive Bayes classifier assumes that the data properties are independent and makes learning considerably easier [8].

Random Forest

Random forest generates a lot of decision trees. It is a simple learning method used for other tasks that produces and processes the class that is the average estimate of the mode of classes or the singular trees [9].

Support Vector Machines

The purpose of the support vector machines is to define the hyperplane which provides that two classes at hand are separated from each other in the best possible way [10].

Extreme Gradient Boosting

XGBoost came to be known by a research project conducted by Tianqi Chen for the first time. It was initially used as a terminal application which can be configured by using a libsvm configuration file. After being used in the winning solution of Higgs Machine Learning Challenge, it started to be recognized in ML contest circles. Shortly after, Python and R packages were generated and now, XGBoost has package applications for Julia, Scala, Java, and other languages.

XGBoost was built in order to avoid the previous gradient boosting limitations and use the resources in a correct and efficient manner [11]. It can be used for controlled learning tasks such as Regression, Classification and Sorting. XGBoost is one of the applications of the Gradient Boosting concept, however, the element that renders XGBoost unique is that it uses "a more systematic model formalization in order to control overfitting,'' according to the creator of the algorithm Tianqi Chen [12].

XGBoost library is quite popular among R users. It produces a better estimation performance compared to the other algorithms and completes the tasks more quickly. According to the creators of the XGBoost package (Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang), it makes parallel calculations automatically on a single machine that can be 10 times faster than the current gradient boosting packages [13].

In the XGBoost algorithm, the decision trees are created in an order. The weights have an important role in XGBoost. The weights are assigned to all independent variables and fed to the decision tree that estimates the results afterwards. The weight of the variables determined incorrectly by the decision tree are increased and these variables are then fed to the second decision tree. These individual classifiers are then added together to yield a stronger and more precise model.

Gradient Boosting

Gradient Boosting, is a gradient boosting algorithm. Gradient boosting is one of the machine learning methods used in order to solve classification and regression problems. This method is one that is generated by bringing more than one weak estimation models together [14].

ADA Boosting

Ada-boost or Adaptive Boosting increases the accuracy by joining more than one classifiers. AdaBoost classifier form a stronger classifier by joining more than one weak classifier and therefore obtain a stronger classifier with higher accuracy [15].

II. PURPOSE

This study makes use of the German Credit Data UCI dataset which contains 1000 data. Assessing whether the applicants are suitable for a loan by using machine learning techniques on the dataset is the goal of this study. 300 applicants pose a risk among the data of 1000 people in the dataset. The dataset contains various private attributions such as the sex, age, and profession of the people, the amount in their savings account and whether they own a house. In this study, the people which are eligible for a loan are shown as 1 and the others as 0. Totally 10 attributes were used in order to asses eligibility. Totally 10 machine learning methods were used for these attributions. The results of the applied methods were cross-checked and making the credit eligibility estimation based on the algorithm that yields the best result was aimed.

III. SCOPE

This study can be used by the related individuals in the banking and finance sector in order to make fast decisions and take the necessary precautions by using the personal attributions of the loan applicants. Considering that there is a considerable number of people in need for a loan; the study can provide a swift and continuous

operational flow for the right people. When the machine learning methods employed in this study are integrated with the customer attributions, faster results would be obtained. The results of this research can be used in order to help the assessment process of credit applications rather than making a final determination of customer eligibility for loans.

IV. METHOD

In this study, German Credit Data UCI dataset which contains various attributions of 1000 people belonging to a community was used, IDLE was employed as the compiler in Pyhton language, and 300 people were determined to be posing a risk to receive a loan. The study was conducted on a laptop which has 8 GB RAM, 2 GB Video Card and 2nd generation i5 Processor. The program was run totally 80 times in order to ensure stability. Numpy, Pandas, Matplotlib, Seaborn libraries were used for the graphics used in the study and Scikit-Learn library for the algorithms. The normalized dataset was processed one by one with the machine learning algorithms. 3/4 (75%) of the 1000 data was used for training and the remainder 1/4 (25%) of the data was used for testing. A maximum of 73,6% success rate was achieved by means of the random forest algorithm as a result of the previous studies conducted with machine learning methods in the Python language regarding credit risk analysis. In this study, increasing the accuracy rate by using different algorithms in order to increase the above percentage was aimed.

V. RESULTS

Totally 10 machine learning algorithms were used on the dataset. Considering these algorithms, the accuracy rate was determined to be 74,80% when the dataset was processed through Logistic Regression method. Linear Discriminant analysis method achieved 73,60% success rate. The second lowest success rate was achieved with the nearest neighbor method by 67,20%. As a result of the tests conducted with another algorithm, Decision Tree, the number of branching was determined as 5, and in the case that the randomness is none, a result of 71,20% was obtained. In the study conducted with Naive Bayes method, the lowest score of the study, 64,80% was achieved. Random forest algorithm achieved 73,20% success rate. The support vector machines achieved 71,60% success rate. The highest accuracy rate of the study was achieved with the XGBoost Model, which is 75,60%. Gradient Boosting classifier was successful with %71,60%. Lastly, ADA Boosting classifier had a 70% success rate. The parameters and methods of these 10 algorithms were studied and as can be seen in Figure 1, the highest result was achieved with XGBoost Classifier.

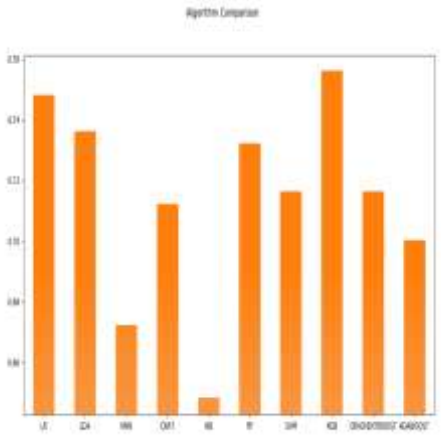

Figure 1: Algorithm Comparison

In order to measure the accuracy of the algorithms used in the study, a complexity matrix was created for each algorithm. The results of the complexity matrix of the algorithms were given in Table 1. TruePositive section in Table 1 shows the number of data having the positive value (risk status 1) that is correctly estimated by the model. TrueNegative section shows the number of data having the negative value (risk status 0) that is correctly estimated by the used model. FalsePositive section shows the number of data having the positive value incorrectly estimated by the used model. FalseNegative section shows the number of data having the negative value that is incorrectly estimated by the used model.

Table 1: Complexity Matrix Results of the Algorithms

ALGORITHMS	COMPLEXITY MATRIX RESULTS			
	TRUE POSITIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE NEGATIVE
Logistic Regression	159	19	44	28
Linear Discriminant	156	22	44	28
K-Nearest Neighbor	150	28	54	18
Decision Tree	146	32	41	31
Naive Bayes	124	54	34	38
Random Forest	160	18	49	23

Support vector machine	172	6	65	7
XGBoosting	166	12	49	23
GRABoosting	161	17	55	17
ADABoosting	158	20	55	17

VI. CONCLUSION

Predetermining whether the loan applicants pose risks has a great importance for banking and finance sectors. This study has shown that a higher success rate can be achieved than the studies conducted previously on the subject of assessing the credit eligibility status of the applicants with 10 attributions, wherein 300 people among 1000 were found to be posing risks. The XGBoost algorithm has yielded the most accurate and highest rate in order to train the dataset.

While the highest success rate was shown to be 73,60% in the previous studies conducted with random forest algorithm, in this study, XGBoost algorithm was found to have the highest success rate among the previous studies conducted in order to assess the credit eligibility of the customers, by 75,60%.

In addition, the accuracy rates of the algorithms used in the study as a resource were increased. However, the XGBoost classifier was emphasized since it had the highest success rate. The highest accuracy rate was achieved with the algorithm which was processed together with the attributions integrated into training, and credit eligibility status was estimated.

REFERENCES

[1]. Pirim, A. G. H. (2006). Yapay zeka. Journal of Yaşar University, 1(1), 81-93.
[2]. Hameed, A. A., Karlik, B., & Salman, M. S. (2016). Back-propagation algorithm with variable adaptive momentum. Knowledge-Based Systems, 114, 79-87.
[3]. Verhulst, Pierre-François. "Notice sur la loi que la population suit dans son accroissement." Corresp. Math. Phys. 10 (1838): 113-126.
[4]. Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." Annals of eugenics 7.2 (1936): 179-188.
[5]. Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27.
[6]. Edwards, Ward, and Detloff von Winterfeldt. "Decision analysis and behavioral research." Cambridge University Press 604 (1986): 6-8.
[7]. McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." AAAI-98 workshop on learning for text categorization. Vol. 752. No. 1. 1998.
[8]. Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.
[9]. Vapnik, Vladimir N., and Aleksei Yakovlevich Chervonenkis. "The uniform convergence of frequencies of the appearance of events to their probabilities." Doklady Akademii Nauk. Vol. 181. No. 4. Russian Academy of Sciences, 1968.
[10]. Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." R package version 0.4-2 (2015): 1-4.
[11]. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.
[12]. Margineantu, Dragos D., and Thomas G. Dietterich. "Pruning adaptive boosting." ICML. Vol. 97. 1997.

Omer Yavuz Can "Credit Risk Analysis Using Machine Learning Techniques" *International Journal of Engineering Research and Applications (IJERA*), vol.10(01), 2020, pp 56-59.