ORIGINAL RESEARCH ARTICLE | UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Comparison of Different Ensemble Methods in Credit Card Default Prediction

Azhi Abdalmohammed Faraj[1,2], Didam Ahmed Mahmud[1], Bilal Najmaddin Rashid[1]

[1]Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq, [2]Department of Computer Engineering, College of Engineering, Dokuz Eylül Üniversitesi, İzmir, Turkey

ABSTRACT

Credit card defaults pause a business-critical threat in banking systems thus prompt detection of defaulters is a crucial and challenging research problem. Machine learning algorithms must deal with a heavily skewed dataset since the ratio of defaulters to non-defaulters is very small. The purpose of this research is to apply different ensemble methods and compare their performance in detecting the probability of defaults customer's credit card default payments in Taiwan from the UCI Machine learning repository. This is done on both the original skewed dataset and then on balanced dataset several studies have showed the superiority of neural networks as compared to traditional machine learning algorithms, the results of our study show that ensemble methods consistently outperform Neural Networks and other machine learning algorithms in terms of F1 score and area under receiver operating characteristic curve regardless of balancing the dataset or ignoring the imbalance

Index Terms: Ensemble methods, Credit card default prediction, Balanced and imbalanced dataset, Stacking and XGBoosting, Neural networks

1. INTRODUCTION

In the aftermath of the Global Financial Crisis of 2008–2009, many who took mortgages defaulted when they could not pay leading to many credit card issuers routinely encountering a credit debt crisis. Numerous occasions of over-issuing credit cards to unfit candidates have raised concerns. Concurrently a considerable percentage of cardholders regardless of their repayment capabilities heavily relied on credit cards and resulted in heavy credit debts. This has negatively affected banks and consumer confidence.

Access this article online

DOI:10.21928/uhdjst.v5n2y2021.pp20-25	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

The problem of credit card defaulting is binary classification problem applicants will either default or repay their credit debts, however determining the probability of defaulting from the perspective of risk management offers more value than a result of a binary classification [1]. Improving the accuracy of fraudulent activities by only one percent can have a major impact on reducing the loss of financial institutions [2].

The aim of a credit default detection model is to solve the problem of categorizing loan customers into two groups: good customers (those who are expected to pay off their full loans in a already agreed upon time period) and bad customers (those who might default on their payments). Customers who pay their bills on time are more likely to repay their loans on time, which benefits banks. Bad customers, on the other hand, can cost you money. As a result, banks and financial institutions are increasingly focusing on the development of credit scoring models,

Corresponding author's e-mail: Azhi Abdalmohammed Faraj, Department of Information Technology, College of Commerce, University of Sulaimani, Sulaimani, Iraq/Department of Computer Engineering, College of Engineering, Dokuz Eylül Üniversitesi, İzmir, Turkey.
E-mail: azhi.faraj@univsul.edu.iq

Received: 13-03-2021 Accepted: 26-06-2021 Published: 19-07-2021

as even a 1% improvement in the quality of bad credit applicants will result in substantial potential savings for financial institutions. Therefore; organizations and scholars have conducted extensive research on credit score models, which is a significant financial management practice. Several studies have discussed the superiority of ensemble learning, as new machine learning models are proposed. Ensemble learning has been incorporated into the application of credit scoring [3].

Ensemble learning is a machine learning technique in which several machine learning algorithms are trained and combined to generate a final output that is superior to individual algorithm outputs. Ensemble learning strategies are divided into two types: Homogeneous and heterogeneous ensembles. Each base learner form is built in a different way using various machine learning techniques in the heterogeneous ensemble technique. The final forecast and the same dataset are generated by statistically combining each individual base learner prediction. Each base learner is used on different subsets of the entire training dataset in homogeneous ensemble techniques. To satisfy requirements and achieve a good ensemble, two necessary and critical conditions must be met: diversity and accuracy [4].

This research aims to answer three questions, first how well ensemble methods work on credit default predictions? Second how do they compare to NN and other traditional algorithms when used on skewed datasets? Third how does balancing the dataset affect the relative performance gain in Ensemble methods?

The ensemble techniques used in this research are Bagging, Boosting (AdaBoosting and XGBoosting), Voting, and random forests (RF).

1.1. Related Work

Advances in technology and the availability of big data have helped researcher improve results on Machine Learning in credit scoring, default prediction, and risk evaluation. Since the purpose of credit management is to improve the business performance and decrease the associated risk, rules must be established to make credit decisions. Hence, clustering algorithm is widely used in the credit default detection systems in the early stage. For instance, William and Huang combined the K-means clustering method with the supervision method for insurance risk identification [5].

Researchers in Saia *et al.* [6] performed credit scoring to detect defaults using the Wavelet transform combined with three metrics three different datasets were used in their experimentation the authors compared their results with RF and improved on RF; however, state of the art results is achieved using neural networks and to get a better perspective neural networks approach needed to be included. The work in Saia and Carta [7] transformed the canonical time domain representation to the frequency domain, by comparing differences of magnitudes after Fourier Transform conversion of time-series data. The authors in Ceronmani Sharmila *et al.* [8] applied an outlier-based score for each transaction, together with an isolation forest classifier to improve default detection. Authors of Zhang *et al.* [9] used data preprocessing and a RF optimized through a grid search step, the feature selection step while preparing the data helped to improve the accuracy of RF.

In Zhu *et al.* [10], deep learning was utilized for the 1st time by applying convolutional neural networks (CNN) approach through the transformation of features to gray scale images, their R-CNN model improved on the area under curve (AUC) of RF and logistic regression (LR) by around 10%. A thorough analysis of different neural networks, such as Multilayer Perceptron and CNNs for credit defaulting can be found in Neagoe *et al.* [11].

Ensemble learning techniques have previously been applied in different credit-related topics for example [12] used RF and majority voting to classify transactions by European cardholders in September 2013 [13], used majoring voting by combining support vector machine (SVMs) and LR, to validate a feature selection approach, called group penalty function the research mainly focuses on robustness of the models. Wang *et al.* [14] used bagging and boosting for credit scoring, Ghodselahi [15] used a hybrid SVM ensemble for binary classification of credit default predictions. The work in Zhang *et al.* [16] ensembles five classifiers (LR, SVMs, neural network, gradient boosting decision tree, and 6 RF) using a genetic algorithm and fuzzy assignment. In Feng *et al.* [17], a set of classifiers are joined in an ensemble according to their soft probabilities. In Tripathi *et al.* [18], an ensemble is used with a feature selection step based on feature clustering, and the final result is a weighted voting approach.

1.2. Overviews of Ensemble Learning

The ensemble methods seek to enhance model predictability by integrating several models to create one stable model. By training several models to train a meta-estimator, ensemble learning aims to enhance predictive efficiency. Base estimators or base learners are considered the component models of an ensemble. The strategies of the ensemble exploit the influence of "the wisdom of crowds," which is focused on the idea that a community's collective judgment is more powerful than any person in the group. Ensemble techniques are widely used in various fields of application, including economic and business analytics, medicine and health insurance, information security, education, industrial production, predictive analytics, entertainment, and many more. Many machine-learning algorithms deal with a tradeoff of fit versus uncertainty (also known as bias-variance), which affects their ability to generalize potential knowledge accurately. To solve this tradeoff, ensemble approaches use multiple models. Two essential components are required for an effective ensemble: (1) Ensemble diversity and (2) model aggregation for the final predictions [19], [20].

1.3. Bagging

Bagging is primarily used in classification and regression, the short form for bootstrap aggregation. By utilizing decision trees, it improves the precision of models, and to a large degree decreases uncertainty. The reduction of variance increases accuracy, hence eliminating overfitting, which is a challenge to many predictive models [19]. Using bootstrapped replicas of the training data, diversity in bagging is acquired: different training data subsets are randomly drawn from the entire training data with replacement. To train a different base learner of the same type, each training data subset is used. The combination strategy of the base learners for bagging is the majority vote. Simple as it is, when combined with the basic learner generation strategies, this strategy can decrease variance. Bagging is particularly attractive when the data available is limited in size. Relatively large portions of the samples (75–100%) are drawn into each subset to ensure that there are sufficient training samples in each subset. This causes a significant overlap of individual training subsets, with many of the same instances appearing in most subsets, and some instances appearing in a given subset multiple times. A relatively unstable base learner is used to ensure diversity under this scenario, so that sufficiently different decision limits can be obtained for small disturbances in different training datasets [21].

1.4. Boosting and RF

Boosting is a form of machine-learning as well. Whereas bagging and RF use autonomous learning, sequential learning is used for boosting. In boosting method, by integrating multiple instances into a more reliable estimation, the simple concept is to improve the precision of a poor classification method [22].

RF is a decision tree-based ensemble learning algorithm. It is simple to implement and can be used for both regression and classification tasks. The bootstrap method is used by RF to collect samples from the original results. Every tree assigns a classification, and the forest selects the classification that receives the most votes among all trees. The degree of randomness is determined by the parameter m, which is the number of decision trees. The borrower is presumed to have d attributes in the RF [23].

Random fore effects of the classification produced from multiple datasets of training are organized and combined to improve the accuracy of the prediction. However, bagging uses all input variables to build each decision tree, RF uses subsets to create each decision tree that are random samplings of variables. This means that forest randomness is best adapted for high-dimensional data processing than bagging [24].

1.5. Stacking

Stacking, another tactic of the ensemble, is also known as stacked generalization. This approach works by allowing many other related learning algorithm predictions to be put together by a training algorithm. Regression, density calculations, distance learning, and classifications have been widely applied by stacking. It may also be used during bagging to calculate the error rate involved [25].

2. MATERIALS AND METHODOLOGY

2.1. The Dataset

The dataset contains information about 30,000 consumers for each consumer 23 attributes marked X1 to X23 [Table 1] are stored. The dependent variable represents whether a customer has defaulted (1) or repaid (0). All the client's data are recorded in September 2005 in Taiwan. As with all types of risk assessment datasets, the ratio of positive to negative samples causes a major imbalance in the dataset, in this dataset, only 22% of the clients have defaulted. There are no missing values in the dataset however there are 35 duplicated rows in the dataset, these have been removed.

- X1: Amount of the given credit (NT dollar): It includes both the individual consumer credit and his/her family (supplementary) credit
- X2: Gender (1 = male; 2 = female)

TABLE 1: Results of the imbalanced dataset

Ensemble methods	Accuracy	Precision	Recall	F1
Neural network	82.01	66.85	36.73	47.41
Bagging	79.43	55.45	34.62	42.62
Ada boost	81.83	68.06	33.33	44.75
XGBoosting	82.11	68.16	35.6	46.77
Voting ensemble	81.88	68.32	33.41	44.87
Stacking	81.86	65.73	37.26	47.56

- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
- X4: Marital status (1 = married; 2 = single; 3 = others)
- X5: Age (year)
- X6–X11: History of past payment. We tracked the past monthly payment records (from April to September 2005); as follows: X6 = the repayment status in September 2005 X7 = the repayment status in August 2005 X11 = the repayment status in April 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for 1 month; 2 = payment delay for 2 months; 8 = payment delay for 8 months; 9 = payment delay for 9 months and above
- X12–X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September 2005; X13 = amount of bill statement in August 2005 X17 = amount of bill statement in April, 2005
- X18–X23: Amount of previous payment (NT dollar). X18 = amount paid in September 2005; X19 = amount paid in August 2005. X23 = amount paid in April 2005.

2.2. Evaluation Metrics

The dataset used in this research is imbalanced if this is not handled then accuracy will not provide a meaningful result because even if the model only predicts the output to be 0 it will still get 78% accuracy regardless of the dependent features. It can be presumed that those responsible for issuing these credit cards believed that every cardholder will not default otherwise it would not have been issued in the first place, thus we can conclude that the human level accuracy for this dataset is approximately 78%, This is an example of when a machine learning performs better than humans. It should be noted that misclassifying a positive example as negative will have higher cost and damage than predicting a negative class to be positive.

This means that the model with better performance on the positive cases should be preferred. Some of the common metrics for classification include accuracy, precision, recall, receiver operating characteristic (ROC), and AUC [6]. All these common metrics will be presented for each model.

In the context of credit card default recall means out of all defaulters how many did the model get correct while precision measures the correctness of the model based on its predictions. F1 score is the harmonic mean of recall and precision. In this research, all the common metrics will be presented however for the assessment of ensemble methods we will focus on the F1 score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2.3. Methodology

We used the steps shown in Fig. 1 for each of the ensemble methods mentioned in section 1. In addition, LR and decision trees were also used however because their results were outperformed by neural networks, we opted to not include them in the results section and decided to use NN as a benchmark for performance comparison. To measure the effects of imbalance on the data all algorithms have also been tested after the down sampling of the datasets their results have included in subsequent sections.

3. RESULTS AND DISCUSSION

The results of the ensemble learning were recorded in two separate trials, first with the original imbalanced dataset and second after the imbalance aspect were eliminated.

When the ratio of positive samples to negative sample is approximately 82% accuracy cannot be used as a reliable measure and as shown in Table 1 all models retrieve an accuracy of around 80% which is equivalent to predicting the performance

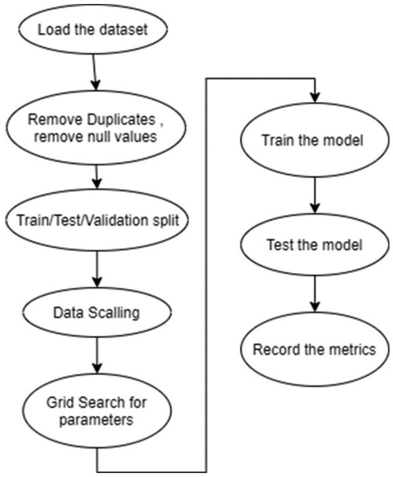
Load the dataset
Remove Duplicates , remove null values
Train/Test/Validation split
Data Scalling
Grid Search for parameters
Train the model
Test the model
Record the metrics

Fig. 1. Proposed method.

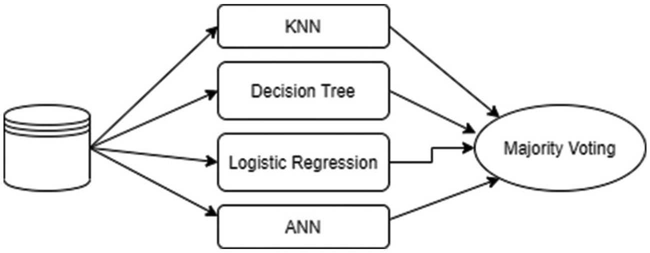

Fig. 2. Voting ensemble used in this research.

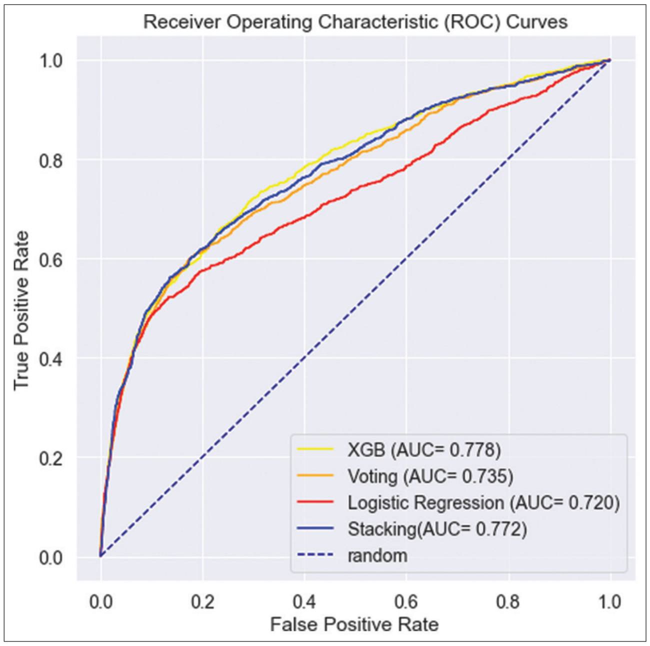

Fig. 3. Receiver operating characteristic curve for imbalanced dataset.

of ensemble methods as compared to regular prediction methods, a variety of other methods were tested such KNN, LR, decision trees, and neural networks. Neural networks performed the best as also confirmed in Cheng Yeh and Lien [1], Hand and Henley [2]. Therefore, for comparison purposes, the results of the artificial neural network are also presented with the ensemble methods for both cases. Fig. 2 show the structure of the voting ensemble used in this study, additionally, for the stacking ensemble, the same algorithms were used in the first level and later LR was applied as the final estimator. In both cases, the data were scaled using a min-max scaler.

3.1. The Imbalanced Dataset

Not default for everyone and consequently is the same as human-level error. A better metric would be the F1 score which is the harmonic mean of recall and Precision Fig. 3. Stacking produced the best result which is 47.56 marginally better than the 47.41 of neural networks. In terms of area under the ROC curve Stacking and XGBoosting produced the best results.

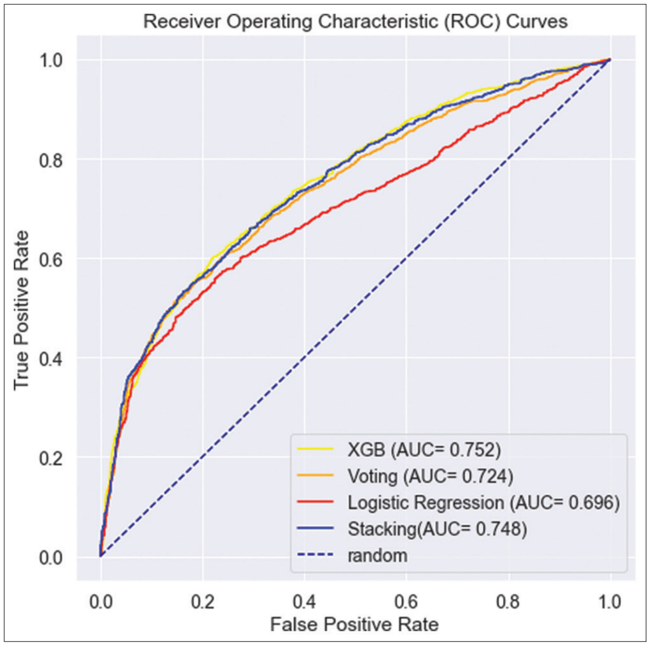

Fig. 4. Receiver operating characteristic curve for the balanced dataset.

TABLE 2: Results of the balanced dataset

Ensemble methods	Accuracy	Precision	Recall	F1
Neural network	68.53	71.9	59.86	65.33
Bagging	64.84	65.79	60.49	63.02
Ada boost	68.49	71.47	60.56	65.57
XGBoosting	68.76	71.42	61.58	66.13
Voting ensemble	67.75	72.57	56.1	63.28
Stacking	68.22	72.58	57.59	64.22

3.2. The Balanced Dataset

For balancing the dataset down sampling was used since there are 6630 positive samples, the same number of negative samples was kept, and the rest was discarded. The samples were randomly shuffled before feeding them to ANN and Ensemble methods. Since the dataset is balanced now accuracy can also be taken into account as shown in Table 2 we can see that XGBoosting is slightly outperforming all the others in all the metrics. XGBoosting is also the fastest in terms of time consumption. Fig. 4 shows the the ROC curves for the balanced dataset, in Which XGBoosting produced the best result.

4. CONCLUSION

The Credit default prediction using ML algorithms has a crucial role in many financial situations including personal loans, insurance policies, etc. However, establishing a model that improves the previous rule-based predictions is weakened by the data imbalance problem in datasets, where the number of unreliable cases is quite smaller than the number of reliable cases.

In this paper, we examine different ensemble methods for credit card default prediction in an imbalanced dataset and compare the results with neural networks. Most research in the literature have either focused on the balanced dataset or a skewed one however we have included both in scenarios to provide a better perspective of the performances of each used algorithm. We tested the results first without altering the imbalance aspect of the dataset in which we used AUC as a metric and ignored accuracy and later by down sampling the majority class. Our experiments show that XGBoosting performs better in both cases as compared to other ensemble methods and also better than neural networks.

REFERENCES

[1] I. Cheng Yeh and C. H. Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients". *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473-2480, 2009.

[2] D. J. Hand and W. E. Henley. "Statistical classification methods in consumer credit scoring: A review". *Journal of the Royal Statistical Society*, vol. 160, no. 3, pp. 523-541, 1997.

[3] Y. Li and W. Chen. "A comparative performance assessment of ensemble learning for credit scoring". *Mathematics*, vol. 8, no. 10, p, 1756, 2020.

[4] M. Akour, I. Alsmadi and I. Alazzam. "Software fault proneness prediction: A comparative study between bagging, boosting, and stacking ensemble and base learner methods". *International Journal of Data Analysis Techniques and Strategies*, Vol. 9, No. 1, pp. 1-16, 2017.

[5] G. Williams and Z. Huang. "Mining the knowledge mine: The hot spots methodology for mining large real world databases". In: *Proceedings of the 10th Australian Joint Conference on Artificial Intelligence, Perth, Australia*, 1997.

[6] R. Saia, S. Carta and G. Fenu. "A wavelet-based data analysis to credit scoring". In: *ICDSP 2018: Proceedings of the 2nd International Conference on Digital Signal Processing, ACM, 2018*, pp. 176-180, 2018.

[7] R. Saia and S. Carta. "A fourier spectral pattern analysis to design credit scoring models". In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning, ACM*, p. 18, 2017.

[8] V. Ceronmani Sharmila, K. K. R., S. R., S. D. and H. R. "Credit card fraud detection using anomaly techniques". In: 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), pp. 1-6, 2019.

[9] X. Zhang, Y. Yang and Z. Zhou. "A novel credit scoring model based on optimized random forest". In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference* (*CCWC*), pp. 60-65, 2018.

[10] B. Zhu, W. Yang, H. Wang and Y. Yuan. "A hybrid deep learning model for consumer credit scoring". In: *2018 International Conference on Artificial Intelligence and Big Data* (*ICAIBD*), pp. 205-208, 2018.

[11] V. Neagoe, A. Ciotec and G. Cucu. "Deep convolutional neural networks versus multilayer perceptron for financial prediction". In: *2018 International Conference on Communications* (*COMM*), pp. 201-206, 2018.

[12] I. Sohony, R. Pratap and U. Nambiar. "Ensemble learning for credit card fraud detection". In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2018.

[13] J. Lpez and S. Maldonado. "Profit-based credit scoring based on robust optimization and feature selection". *Information Sciences*, vol. 500, pp. 190-202, 2019.

[14] G. Wang, J. Hao, J. Ma and H. Jiang. "A comparative assessment of ensemble learning for credit scoring". *Expert Systems with Applications*, vol. 38, no. 1, pp. 223-230, 2011.

[15] A. Ghodselahi. "A hybrid support vector machine ensemble model for credit scoring". *International Journal of Computer Applications*, vol. 17, no. 5, pp. 1-5, 2011.

[16] H. Zhang, H. He and W. Zhang. "Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring". *Neurocomputing*, vol. 316, pp. 210-221, 2018.

[17] X. Feng, Z. Xiao, B. Zhong, J. Qiu and Y. Dong. "Dynamic ensemble classification for credit scoring using soft probability". *Applied Soft Computing*, vol. 65, pp. 139-151, 2018.

[18] D. Tripathi, D. R. Edla, V. Kuppili, A. Bablani and R. Dharavath. "Credit scoring model based on weighted voting and cluster based feature selection". *Procedia Computer Science*, vol. 132, pp. 22-31, 2018.

[19] P. Bühlmann. "Bagging, boosting and ensemble methods". In: J. Gentle, W. Härdle and Y. Mori, (eds.), *Handbook of Computational Statistics*. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg, 2012.

[20] G. Kunapuli. "*Ensemble Methods for Machine Learning*". MEAP Publication, Shelter Island, New Work, 2020.

[21] S. Hamori, M. Kawai, T. Kume, Y. Murakami and C. Watanabe. "Ensemble learning or deep learning? Application to default risk analysis". *Journal of Risk and Financial Management*, vol. 11, p. 12, 2018.

[22] R. E. Schapire and Y. Freund. "*Boosting: Foundations and algorithms*". *Kybernetes*, vol. 42, no. 1, pp. 164-166, 2013.

[23] B. Niu, J. Ren and X. Li. "Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending". Information, vol. 10, p. 397, 2019.

[24] A. Mayr, H. Binder, O. Gefeller and M. Schmid. "The evolution of boosting algorithms. From machine learning to statistical modeling". *Methods of Information in Medicine*, vol. 53, no. 6, pp. 419-427, 2014.

[25] R. Sikora and O. H. Al-laymoun. "A modified stacking ensemble machine learning algorithm using genetic algorithms". *Journal of International Technology and Information Management*, vol. 23, p. 1, 2014.