IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 9, Issue 4, 2021

BASED ON BACKFLOW LEARNING, A NEW NOISE-ADAPTIVE TWO-LAYER ENSEMBLE MODEL FOR CREDIT SCORING

Dr.Siddiqui Riyazoddin Alimoddin[1], Md.Abdul Rawoof[2]

Professor[1], Asst.Prof[2]

Department of cse

NAWAB SHAH ALAM KHAN COLLEGE OF ENGINEERING & TECHNOLOGY NEW MALAKPET, HYDERABAD-500024

ABSTRACT

It has become more crucial to use machine learning and artificial intelligence algorithms in categorization challenges like credit scoring. Commercial banks and loan lenders have an essential information management task: building an ensemble learning model that has been demonstrated to be more accurate and resilient than individual classifiers. Extreme gradient boosting (EGB), gradient boosting decision tree (GBT), support vector machine (SVM), random forest (RF), and linear discriminant analysis are all combined into an unique noise-adapted two-layer ensemble model for credit scoring based on backflow learning. Base classifiers may be boosted in strength and variety by using a novel backflow learning technique that retrains them on misclassified data points. Two-layer ensemble modelling is used to combine the predictions of all of the basic classifiers to provide a final predictive result. As a result, an isolation forest-based noise adaptation strategy is being suggested to solve the issue of noise data, which has long been considered a key impediment to the accuracy of prediction models. Outlier scores are calculated for each data point in the training set to identify noise data, which are then boosted to generate the noise-adapted training set. A comparison of the proposed model's performance against that of existing benchmark models is carried out using three credit datasets from the UCI machine learning library. Our suggested model outperforms other models in terms of several performance metrics, according to the findings of our experiments.

INDEX TERMSMachine learning and feature engineering are some of the terms used to describe the many aspects of credit scoring.

INTRODUCTION

It is no longer possible to make decisions based just on human expertise in today's rapidly evolving loan lending industry. The risk that consumers will be unable to pay back their debts is known as the probability of default (PD) and has received the greatest study attention among all risk management jobs [1]. Loan lenders have to make wise decisions in order to minimise losses and increase profits. In order to distinguish between "good" and "bad" consumers, the term "repayment" has been coined. Misclassifying clients will result in two separate forms of losses. If a client who is genuinely "good" is labelled "bad," the loan providers will suffer the resulting loss of revenue. Loans made to "bad" customers, on the other hand, will be non-performing and result in enormous losses. A strong PD model is essential for financial institutions to thrive in today's competitive environment [2] and optimise profits. In recent years, credit scoring has become more dependent on machine learning methods [3]. Classification difficulties may be solved using machine learning methods such as support vector machine (SVM), artificial neural network (ANN), and decision tree (DT). LDA and LR, two of the most often used statistical techniques in business despite recent advances in machine learning, remain popular because of their ease of implementation [9]. The power of ensemble classifiers is a promising study area, even though machine learning approaches may be utilised to grasp complicated models. Noise is a concern in real-world data because it reduces prediction accuracy and raises the computational cost of creating classifiers [17]. Researchers have yet to include noise data into their prediction algorithms, as far as we know. It is becoming more critical to identify noise in datasets while doing feature engineering. The goal of this research is to develop a noise detection technology that is both accurate and efficient. The isolation forest (IF) algorithm [18] is offered as a novel noise adaptation strategy for dealing with noise data. A tree-based ensemble noise detection approach with linear time complexity is the IF algorithm. Each tree formed by the IF algorithm is distinct from all other trees. It is possible to use the IF algorithm to analyse big datasets. Other noise-detection methods don't have these specific qualities. A noise-adapted training set is created by calculating the

outlier score of each data point and then boosting those data points in the training set. As a result of this research, a new noise adaptive two layer ensemble model for credit rating based on backflow learning has been suggested. For starters, an outlier score is calculated for each data point and the noise data that are subsequently boosted in the training set are identified using the IF-based noise adaption (IFNA) technique. Overfitting is less likely to occur if the training set has more noise than the original dataset, which has less noise than the original dataset. XGBoost (extreme gradient boosting) [19], Gradient Boosting Decision Tree (GBDT) [20], SVM, RF, and LDA are among of the classifiers that will be used in this innovative backflow learning technique. According to the suggested backflow learning technique, the training set is expanded by including data entries with low classification confidence during the base classifier training stage. It is then retrained using the expanded training set to improve base classifier performance. Additionally, the best-performing basic classifiers on the validation set are kept in reserve. Finally, all of the improved base classifiers are combined using two-layer voting and stacking, which not only improves the proposed model's predictive performance but also reduces the unpredictability of the optimised base classifiers. For datasets that contain outliers, the proposed approach to IF-based noise adaptation (IFNA) allows base classifiers to better identify outliers and improve their learning efficiency. The proposed backflow learning approach, on the other hand, achieves an efficient learning strategy for data entries with low classification confidence during the base classifier training stage. As for the rest of the document, it is structured as follows. Several relevant studies are discussed in Section 2, including several by other researchers. Our suggested model is discussed in depth in Section 3. Machine learning algorithms outperform classical statistical methods in terms of prediction power [10] in Section 4 of the paper. For a classification model to be more accurate and resilient, it needs an ensemble method that combines the predictions of many classifiers by majority voting, weighted voting, or other combination logic. Using an ensemble technique, instead than relying only on a base classifier to handle the wide variety of actual datasets, allows us to better compensate for the weaknesses of individual classifiers [11]. Ensemble models such as bootstrap aggregating (bagging), random forest, random subspace, and stacking have consistently outperformed individual classifiers [16] in terms of performance [12, 13]. Section 5 now analyses the outcomes of an ensemble technique based on measurements. We summarise our findings in Section 6, which includes the results of our measurements and a thorough analysis of the findings in Section 5. Section 6 outlines the outcome of our investigation.

TYPES OF NOISE AND NOISE DETECTION METHODS

Data mining experts are usually concerned about the presence of noise in their datasets. Class and attribute noise are the two most common forms of noise in real-world datasets [17]. Attribute noise refers to data that has the erroneous values for the attributes in question. Various investigations have been carried out to determine if class noise or attribute noise is more pervasive. Twala [21] found that class noise has a considerable impact on machine learning models compared to attribute noise. According to Zhu & Wu [17], who came at the same result, it may be useful to clear up the test set's noise. In classification difficulties, class noise is still a concern since it is exceedingly difficult to identify and relabel [22]. Many researchers have studied distance-based attribute noise detection algorithms [23-24]. In terms of computing complexity, these strategies are simple and effective. Isolation forest (IF) was suggested by Liu et al. [18] as a novel solution to anomaly detection, which has linear computing complexity and minimal memory requirements. Furthermore, the study shows that the IF algorithm outperforms many existing noise detection techniques and may be used to examine large datasets with multiple dimensions. Class noise detection using an ensemble approach is a hot topic right now in academia. Reusing noise data in an iterative noise cleaning process was abolished by Sáez et al. [25]. There is a risk of information loss when noise data is deleted, according to Luengo et al. [26]. Since noisy data is labelled incorrectly, their approach may help rectify and regulate filtering process sensitivity. Noise data may be deleted, and the class label can be corrected, but this will result in a loss of information. The approach of processing noisy data points according to the outlier score of each data point was selected to differentiate between noise and non-noise data in the dataset and to reduce the negative impacts of noise on the prediction. To enhance out of sample performance of parametric and non-parametric models for credit risk assessment, local outlier factor (LOF) [27] approaches were suggested, which improved the outcomes of predictive models in bankruptcy prediction. Due to the importance of credit data and the fact that it is often constrained in terms of volume, additional information must be provided to aid decision-making [28]. [27] The temporal

ISSN 2347–3657
Volume 9, Issue 4, 2021

complexity of the LOF technique, which uses density-based noise detection, is (). A data point's local outlier factor is calculated by comparing the densities of the data points immediately around the present one. Noise detection using the IF technique has a linear time complexity, and its time complexity is () [18]. It is a cutting-edge algorithm that addresses the needs of data processing in the context of big data. However, the suggested IF-based method of noise adaptation does not simply erase the noise data found by the IF algorithm, but instead first calculates the outlier score of each data point to find the noisy training data that are then amplified into a noise-adapted training set.

B. ENSEMBLE MODELS

We've already discussed the fact that an ensemble learning model is often better than a base classifier [29-31]. Multiple base classifiers may be combined into an ensemble learner using different fusion strategies in ensemble learning models, and this fusion technique can have a considerable impact on the ensemble learner's performance [13]. As a result, a thorough investigation of fusion approach is necessary. It is common practise to combine the predictions of basic classifiers via voting. Majority voting, in particular, is a popular method because of its simplicity [32]. Votes are cast by all the base classifiers, and the class label with the most votes is utilised as the final result. Breiman was the first to develop the RF, which entails creating numerous decision tree models from randomly chosen subsets and then converging on the conclusion by majority vote [13]. When Brown and Mues [32] evaluated numerous credit scoring systems, they found that RF outperformed them all, explaining why a majority vote is so effective. Another helpful voting method is weighted voting. In weighted voting, each vote is given a varying amount of weight to help determine the relative importance of each vote [33]. With adaptive boosting (AdaBoost), the training set is changed repeatedly while a series of decision tree models are constructed consecutively [33]. Weighted voting is used to calculate the weight of each decision tree based on how accurate it is. Boosting algorithms such as GBDT and XGBoost serve as good examples. Two new credit scoring models based on boosting were suggested by Finlay [34] and Wang & Ma. A credit prediction model may be improved by utilising unanimous voting, as Liang et al. [11] shown in their trials. An further method known as stacking includes training a higher-level classifier (meta-classifier) based on the results of numerous lower-level classifiers [15]. Stacking Stacked meta-classifiers generally outperform basic classifiers in terms of performance Stacking has been the subject of much research. Heterogeneous ensemble models that incorporate bagging and stacking to improve prediction performance have been suggested by Xia et al. [35]. It was found that the bagging and stacking method is superior for credit analysis, according to Wang and colleagues [36]. An improved meta-learning model for bankruptcy prediction was suggested by Tsai and Hsu [37]. In addition, our prior experiments showed that a stacking ensemble outperformed alternative models [38]. The homogeneous ensemble model and the heterogeneous ensemble model are discussed as a consequence of the choice of basic classifiers. For example, RF and AdaBoost both employ DT as the basic classifier for their homogeneous ensemble models. Ensemble models on the other hand, incorporate a variety of categorization techniques. It wasn't until recently that the heterogeneous ensemble model became popular. Compared to normal homogeneous bagging, Coelho and Nascimento [39] illustrated the advantages of heterogeneous bagging. A heterogeneous ensemble classifier, according to Lessman et al. [40], has the benefit of allowing several classifiers to provide their own perspectives on the same data, which may then be used to complement one another. [41] Nascimento et al [41] found that the variety of base classifiers is critical to achieving acceptable accuracy-generalization performance on ensemble classifiers. Ensemble models still have room for improvement, despite the considerable contributions made by these academics. To begin, noise in credit scoring has received little attention from academics. To deal with noise data, an IF-based noise adaptation strategy is provided. In order to create a heterogeneous ensemble model, very few researchers have investigated utilising diverse training sets. In this work, not only the two-layer ensemble modelling strategy, but also the backflow learning approach, which may increase the training efficiency of classifiers, is studied. Backflow learning is presented here to construct a training set for each basis classifier chosen in order to increase the diversity between base classifiers. After that, the ultimate result may be obtained using the two-layer ensemble model, which employs voting and stacking processes on top of each other. Due to the suggested backflow learning strategy increasing the

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 4, 2021

variety amongst the base classifiers, the ensemble model's power may be further boosted with an additional layer of the ensemble model.

III. THE PROPOSED MODEL

Figure 1 depicts the three key steps of the proposed approach, which include IF-based noise adaptation, backflow learning of basic classifiers, and two-layer ensemble modelling. Data normalisation and dummy coding are used in the initial phase of data processing. Data normalisation and dummy coding may be used to turn a continuous input variable into a set of dichotomous characteristics. There is a limit to the number of explanatory characteristics that may be maintained for correlations greater than 0.97. After data preparation, the training and test datasets are separated. It's possible to divide training data into a "training" and "validation" set. For the first time in this study, the IFNA method is used to identify and analyse noise data, using the IF algorithm in its place. Noise-adapted training sets are created by first calculating the outlie score of each data point in order to identify the noise data. The more noise data in the noise-adapted training set compared to the raw dataset means that the model is better able to adapt to the noise data and the

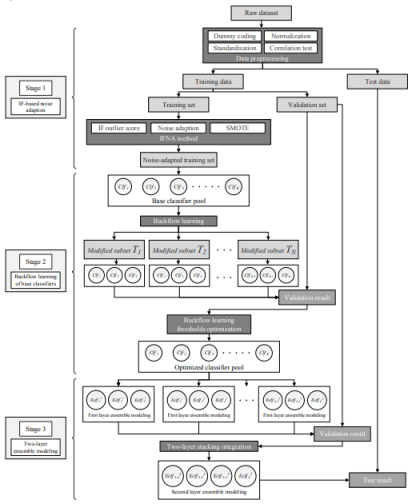

FIGURE 1. Framework of the proposed model

The basic classifiers in the classifier pool are trained in the second step. It represents the first base classifier, and so on. Base classifiers are trained and optimised using a novel backflow learning technique that is based on the noise-adapted training set from the previous step. The optimal ensemble effects are attained in the following stage. According to the suggested backflow learning technique, the training set is expanded by including data entries with low classification confidence during the base classifier training stage. It is then retrained using the expanded training set to improve base classifier performance. Experiments with larger training sets included data items with lower classification confidence, which were utilised to retrain classifiers in order to improve their learning efficiency and decrease model overfitting. After optimising the classifier pool in the second step, a two-layer ensemble technique is used to make predictions on the test data. signifies the first ensemble classifier of the first layer, the second ensemble classifier of the first layer, and so on. In subsections 3.1, 3.2, and 3.3, the specifics of the IFNA method, backflow learning technique, and two-layer ensemble approach are explained. Even said, the IF's outliers may in fact aid in replicating the true distribution of the target function, since they are undersampled samples. As a result, the training set should include more examples of this kind of data. Because the credit dataset has a small sample size, increasing the sample size will increase the suggested model's performance. Figure 2 depicts the IF-based noise adaptation mechanism. By computing the outlier score for each data point, the IF method is applied to the training set of size N

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 4, 2021

to identify the data points that are "likely to be separated." An outlier score, i.e. noise data, is boosted in the training set so that a noise-adapted training set may be created An anomaly may be defined as an outlier in the data space because data points in sparsely distributed areas are statistically less likely to occur than data in dense locations. For the isolation, multiple isolation trees are generated using IF, which builds an IF on M isolation trees, all of which are binary trees with a height H of at least 2. In order to develop an IF, a random subset of the training set of size N is produced. A random hyperplane then divides the data space. The data space is continuously subdivided until the maximum height H is attained or there is only one data point left in each subspace. Isolation trees with a lower average route length L for outliers than for inliers are expected to be found in a subspace sooner than those with a longer average path length L. After that, a subset of size k with the highest outlier score is increased in the training set to produce the noise-adapted training set of size N+k, which is then subjected to the noise increase operation. However, the noise-adapted training set that is created by noise adaptation is immune to noise, and so the model constructed is more robust.

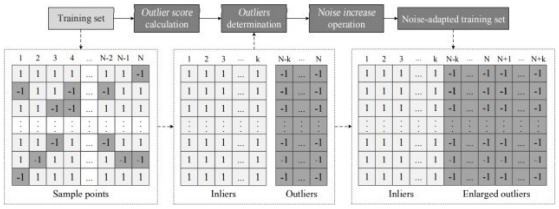

FIGURE 2. IF-based noise adaption process

For the forecast, there is a problem with data imbalance. The misclassification of the classifier may occur due to the imbalance of the majority and minority samples in the dataset, resulting in a decrease in model performance. By creating minority sample points at random, Chawla et al. [42] came up with the synthetic minority over-sampling technique (SMOTE) method to address the data imbalance issue. Using the SMOTE technique, it is possible to increase the number of minority class samples, thereby increasing the model's resilience. The BACKFLOW APPROACH TO LEARNING Using a backflow learning technique, the training process of each base classifier is shown. The classifier will output the probability of a positive or negative class label after training, representing the degree of confidence in the prediction. As an example, if p=0.95, the classifier has a high degree of confidence in predicting the class label as positive. Similarly, if the likelihood of the class label being negative is p=0.05, the classifier is quite confident in this conclusion. To put it another way, the classifier loses its discriminative capability since it is unsure about its prediction if the probability is less than roughly 0.5. This information is shown in the form of a graph in Figure 3.

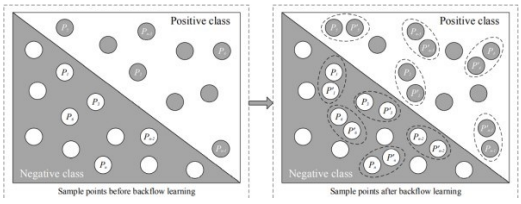

FIGURE 3. The change of sample points in backflow learning

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9, Issue 4, 2021

Backflow learning may be considerably affected by the values of!$& and!$'. For the present research, a grid search is used to find the best possible combination of!$& and!$'. The combination with the best performance is considered to be the optimum threshold. Based on fresh training data, all base classifiers are retrained so that the optimal base classifiers may be obtained

C. TWO-LAYER ENSEMBLE APPROACH

Ensemble modelling is used in the last step to incorporate basic classifiers that were improved in the previous stage. Because voting and stacking classifiers are regarded to have strong generalisation power and performance scores, the optimised basic classifiers may be mixed. A two-layer technique, where the second layer ensemble models are generated on top of the first layer, increases generalisation power even further. Combining numerous strong classifiers often results in the model being destroyed by a poor classifier with performance that is much lower than the model average. Second layer ensemble modelling alleviates this difficulty since its input classifiers are voting or stacking classifiers with minimal correlation.

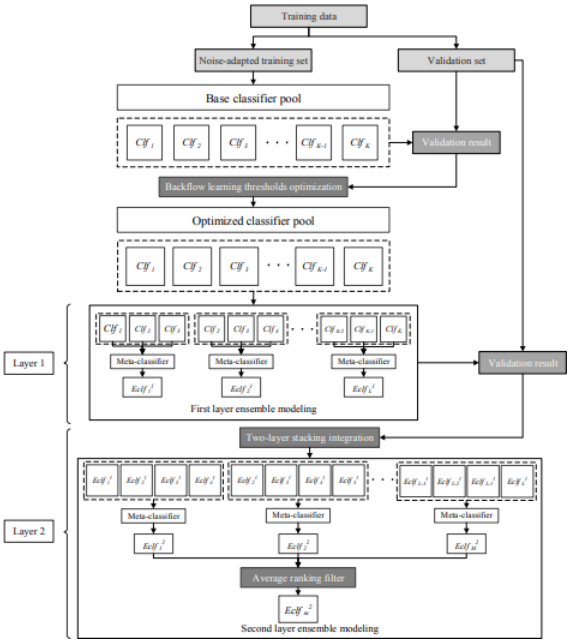

FIGURE 4. Schematic diagram of the two-layer ensemble mode

IV. EXPERIMENT A. DATASETS

Three credit datasets from the UCI machine learning collection are examined in this experiment [44], including the Australian, Japanese, and Polish credit datasets. Data scientists use and study these datasets a great deal. Table 1 provides a breakdown of the datasets. A total of 690 samples are included in the Australian credit dataset, with 307 being positive and 383 being negative. It has a dimension of 14, which includes six continuous qualities and eight category ones. For future calculations, the category attribute's labels have been changed. For instance, the labels of a categorical attribute with just three classes are changed to "1", "2", and "3".

ISSN 2347–3657
Volume 9, Issue 4, 2021

TABLE 1. Description of three credit datasets

Datasets	Sample size	No. of positive samples	No. of negative samples	Dimension of input space
Australian	690	307	383	15
Japanese	690	383	307	16
Polish	7027	271	6756	65

VI. CONCLUSION

Data mining is now focusing on credit scoring as a potential research area. Effective noise detection and learning methods for low classification confidence may increase the model's performance. As a result, a new noise-adapted two-layer ensemble model for credit rating based on backflow learning has been suggested. Prior to developing an adaptable training set, an entirely new IFNA technique to dealing with noise data was presented, in which the outlier score of each data point was first generated in order to identify the noise data. To train the basic classifiers, a novel backflow learning technique was devised and used. To arrive at a final prediction, we used a two-layer stacking and voting ensemble. The suggested model's performance was assessed using five performance indicators: accuracy, precision, 7 2.89:, AUC, and Brier score. According to the findings, the suggested model outperformed existing industry-standard credit scoring algorithms. Many flaws have to be solved in the suggested paradigm, despite its promising results. There was just one test for the IFNA technique, for example. In the future, other noise-detection approaches, such as replicating neural networks, may be investigated and tested. In addition, grid search was employed to find the best parameters for backflow learning, which was computationally and time-consuming. The number of feasible combinations of base classifiers rose quadratically as the number of accessible base classifiers was raised in the two-layer ensemble technique. We want to investigate the use of dynamic clustering and game theory in credit scoring in the future. A credit scoring model based on a two-layer ensemble has less interpretability than a single-layer ensemble, which is not acceptable in a real-world context. Future research will address these kinds of concerns.

SUPPORTING INFORMATION

Three sets of raw data have been published to Fig share (https://doi.org/10.11084/m9.figshare.8068262): one for the Australian dataset (Australian.csv), another for the Japanese credit dataset (JapanData.csv), and a third for the Polish dataset (Polish.csv). From the University of California, Irvine's machine learning library, we get this raw data.

REFERENCES

[1] Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183 (3), 1447-1465.

[2] Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis, 50 (4), 1113-1130.

[3] Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2012). Machine learning in financial crisis prediction: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42 (4), 421-436.

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

[5] West, D., Dellana, S., & Qian, J. X. (2004). Neural network ensemble strategies for financial decision applications. Computers & Operations Research, 32 (10), 2543-2559.

[6] Li, X., Ying, W. Y., Tuo, J. Y., Li, B., & Liu, W. H. (2004). Applications of classification trees to consumer credit scoring methods in commercial banks. In Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Hague, Netherlands, pp. 4112–4117, October 10– 13, 2004.

[7] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annals of Human Genetics, 7 (2), 179-188.

ISSN 2347–3657
Volume 9, Issue 4, 2021

[8] Hand, D. J. (2002). Superscorecards. IMA Journal of Management Mathematics, 13 (4), 273-281. [9] Alaraj, M., &Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. Knowledge-Based Systems, 104, 89-105.

[10] Davis, R. H., Edelman, D. B., &Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. IMA Journal of Management Mathematics, 4 (1), 43-51.

[11] Liang, D. R., Tsai, C. F., & Wu, H. T. (2015). The effect of feature selection on financial distress prediction. Knowledge-Based Systems, 73, 289-297.

[12] Breiman, L. (1996). Bagging predictors. Machine Learning, 24 (2), 123-140.

[13] Breiman, L. (2001). Random Forests. Machine Learning, 45 (1), 5-32.

[14] 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.

[15] Wolpert, D. (1992). Stacked generalization. Neural Networks, 5 (2), 241-259.

[16] Nanni, L., &Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 36 (2), 3028-3033.

[17] Zhu, X. Q., & Wu, X. D. (2004). Class noise vs. attribute noise: A quantitative study. Artificial Intelligence Review, 22 (3), 177-210.

[18] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, pp. 413-422, December 15-19, 2008.

[19] Chen, T. Q., &Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, pp. 785-794, August 13-17, 2016.

[20] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29 (5), 1189-1232