Journal of Database Management
Volume 34 • Issue 1

Credit Risk Models for Financial Fraud Detection:

A New Outlier Feature Analysis Method of XGBoost With SMOTE

Huosong Xia, Wuhan Textile University, China
Wuyue An, Wuhan Textile University, China
Zuopeng (Justin) Zhang, University of North Florida, USA*

ABSTRACT

Outlier detection is currently applied in many fields, where existing research focuses on improving imbalanced data or enhancing classification accuracy. In the financial area, financial fraud detection puts higher demands on real-time and interpretability. This paper attempts to develop a credit risk model for financial fraud detection based on an extreme gradient boosting tree (XGBoost). SMOTE is adopted to deal with imbalanced data. AUC is the assessment indicator, and the running time is taken as the reference to compare with other frequently used classification algorithms. The results indicate that the method proposed by this paper performs better than others. At the same time, XGBoost can obtain a ranking of important features that impact the classification results when performing classification tasks, making the evaluation results of the model interpretable. The above shows that the model proposed in the paper is more practical in solving credit risk assessment problems. It has faster response times, reduced costs, and better interpretability.

KEYWORDS

class-imbalance, outlier knowledge, outlier feature, Real-time, XGBoost

1. INTRODUCTION

Fraud is intentional deception to obtain financial gain or cause loss by implicit or explicit tricks (Kou et al., 2019). Fraud violates public laws, in which the swindlers attempt to obtain illegal benefits or produce irreversible losses (Carcillo et al., 2018; Khanuja & Adane, 2018). The damage resulting from fraudulent activities shows that they cost the victims and financial institutions a significant amount of money. According to the statistics from the Internet Crime Complaint Center, there has been a substantial soar in reported fraud activities in the last decade (Hou et al., 2020).

Industries and research institutions have invested heavily to develop effective methods to combat the problem with emerging machine learning, deep learning, big data, and computational intelligence

DOI: 10.4018/JDM.321739 *Corresponding Author

technologies (Cai & Zhang, 2020; Chua & Storey, 2016; Oreski & Oreski, 2014). Their efforts in this perspective have resulted in many approaches that can intelligently differentiate legitimate transactions from fraudulent ones. However, no matter what methods are applied, some common problems still exist and often reduce their performance and efficiency. For instance, one of the most common problems resides in the training data of the past transactions represented by unbalanced distribution, which causes various difficulties of overfitting and results in inferior performances of the implemented classifiers (Altinbas, 2020). These problems occur due to the relatively smaller number of available fraudulent samples than legitimate ones. This type of unbalance prevents the designation of a dependable model of assessment (Khemakhem, Said, & Boujelbene, 2018). Moreover, data heterogeneity and overlap are additional issues that aggravate the problem (Arora & Kaur, 2019). Computational complexity is another challenge for effectively identifying anomalies (Coser, Maer-Matei, & Albu, 2019; Xu et al., 2020; Ye et al., 2018). These problems significantly impact the efficacy of any fraud recognition techniques that may produce a large number of incorrect classifications.

In recent years, most studies on credit risk assessment models for financial institutions have focused on improving imbalanced data or enhancing classification accuracy through multistage modeling and deep learning. Although these methods can somewhat boost accuracy, the following research gaps still exist. First, low time responsiveness dominates as models with higher classification accuracy tend to have higher model complexity. Second, transparency and interpretability are lacking for the existing methods, along with the insufficient analysis of behavior features (Laughlin, Sankaranarayanan, & El-Khatib, 2020). Therefore, to address the research gaps with the motivation of improving high efficiency and interpretability, we study the research questions in this paper as follows:

(1) *How to build an efficient and interpretable fraud detection model based on the characteristics of the financial domain?*
(2) *How to obtain knowledge about the risks associated with credit assessment? And what are the implications for financial institutions?*

To address the above research questions, we employed the XGBoost method suggested by Chen and Guestrin (2016), which has recently attracted attention for its speed and accuracy. In addition, SMOTE is deployed in this research to process imbalanced data. Empirical results demonstrate that the proposed model combining XGBoost and SMOTE achieved the highest AUC score (0.97) for credit risk evaluation. It also shows that the proposed model is better than other classification models discussed in this paper. At the same time, combined with the running time (89 runtime/s), the calculation speed of this model is faster and time-sensitive. Another advantage is that when performing classification tasks, it is possible to obtain a ranking of important features that have a greater impact on the classification results, making the evaluation results of the model interpretable. The above features show that the model proposed in the paper is more practical in solving credit risk assessment problems.

The rest of the paper proceeds as follows. Section 2 reviews prior literature and highlights the contributions of our study. Section 3 summarizes related algorithms. Section 4 introduces the design of the new method for fraud detection. Section 5 outlines the empirical process and findings. Section 6 concludes the paper with theoretical and practical contributions, limitations, and future research directions.

2. LITERATURE REVIEW

There are three main groups of tools or methods to assess credit risk decisions: statistical, machine learning (ML), and deep learning (DL) methods (Hou et al., 2020). Decades ago, statistical methods, such as linear discriminant analysis, logit analysis, and probit analysis, were the dominant methods in this area (Ala'Raj & Abbod, 2016; Feng, Xiao, & Zhong, 2018; Moscatelli et al., 2020). Orgler (1970) applied linear regression into credit risk evaluation to differentiate between good and bad credit

applicants for commercial banks in practical credit-scoring applications. Wiginton (1980) proposed the logistic regression model for bankruptcy prediction. Overall, statistical models provide a satisfactory forecasting device, accommodating accuracy and transparency requirements (Zhou et al., 2019). Their simple, functional form combines the default additive, monotonic predictors into probabilities with good out-of-sample performance (Song et al., 2020; Xia, Liu, & Da., 2018). However, they have a slow capacity to adapt to changes in the state of the economy and a limited ability to model complex non-linear interactions between economic, financial, and credit variables (Kim & Cho, 2019).

In the big data era, ML technology has gradually replaced traditional statistical methods for credit risk assessment (Munkhdalai et al., 2019), for instance, logistic regression (Rushin et al., 2017), decision tree (Chang, Chang, & Wu, 2018) and support vector machine (SVM) (Ahmad et al., 2018), to predict default probability of a borrower. ML models can handle complex relationships across different variables and large datasets with correlated predictors (Coser, Maer-Matei, & Albu, 2019; Yu et al., 2018). Yao, Crook, and Andreeva (2017) proposed a novel two-stage model based on the least square support vector machine, and the results showed that this model yields better performance than that of the other statistical models. Kim and Cho (2019) designed a method that combines label propagation transductive support vector machine (TSVM) with Dempster-Shafer theory for social lending default prediction. In addition, using social lending platform data, a semi-supervised approach based on SVM has been designed by Li et al. (2017) for reject inference in credit scoring. Serrano-Cinca and Begoa Gutiérrez-Nieto (2016) employed a decision tree (DT) to determine the non-linear relationship between the explanatory variables and the target variable in the credit risk assessment of peer-to-peer (P2P) lending. Moscatelli et al. (2020) compared statistical models in credit risk modeling with ML models, namely Random Forest (RF) and Gradient Boosted (GB) Trees. Their results suggest that the joint use of statistical and ML models by lenders or credit analysts may be beneficial for the accurate assessment of potential borrowers. ML models are relatively non-transparent and may be used to benchmark the probability of defaults obtained using more transparent statistical models. However, single classification models cannot gratify the requirement of accurate prediction models with good generalization ability when handling tremendous amounts of instances and patterns in the big data era. At the same time, due to the unique properties of various data sets, no single classification algorithm applies to all credit data sets (Gicić & Subasi, 2019; Niu et al., 2020). Hence, ensemble methods, such as Random forest (Speiser et al., 2019), AdaBoost (Song et al., 2020), Gradient boosting decision tree (GBDT) (Tomczak & Ziba, 2015), XGBoost (Chang, Chang, & Wu, 2018) algorithms, have been utilized for credit scoring. The shortcomings of each classifier can be compensated by others by aggregating the advantages of multiple classifiers suitable for various local regions to guarantee the final prediction accuracy (Xu et al., 2020; Yu et al., 2018). Marqués, García, and Sánchez (2012) evaluated the prediction performances of five different ensemble methods to suggest the appropriate classifiers for each ensemble approach in the context of credit scoring. Another ensemble method has been proposed by Feng, Xiao, and Zhong (2018), whose classifiers are selected on the basis of their classification performance for credit scoring. Moreover, Xia, Liu, and Da (2018) proposed an ensemble credit model that integrates the bagging algorithm with the stacking method for credit score prediction. Niu et al. (2020) studied the credit risk factors in P2P network lending and constructed a data mining model in risk assessment. Their experiment results show that the best performance is the random forest model. Li et al. (2020) developed a heterogeneous ensemble learning algorithm by combining different classifiers with a linear weight ensemble for the default prediction of social lending.

DL has received significant attention in recent years due to its effectiveness in learning complex and non-linear relationships from data (Fantinato et al., 2021; Fior et al., 2017; Li et al., 2020; Zeng et al., 2018). In particular, Fiore et al. (2017) trained a generative adversarial network (GAN) to output mimicked minority class examples, which were then merged with training data into an augmented training set to improve the effectiveness of a classifier. Experiments show that a classifier trained on the expanded set outperforms the same classifier trained on the original data; however, the computational cost of this algorithm is higher. Bastani, Asgari, and Namavari's (2019) test results indicated that the wide learning

(WL) and DL model in combination with the resampling technique achieve the highest performance on the loan status prediction compared to the benchmark algorithms, including WL, DL, RF, GB, and SVM. DL defines a class of models that can perform complex forecasting tasks when the relationship between predictors and outcomes is unclear or unknown (Li et al., 2019; Yu et al., 2018). Models based on DL typically neglect the inference about the importance of economic determinants for default risk, but, in turn, highly accurate forecasts can be achieved with DL, as established in several studies (e.g., Fonseca, Wanke, & Correa, 2020; Kogeda & Vumane, 2017). This feature of DL models is particularly relevant to credit risk applications, but it comes at the cost of lower transparency relative to statistical models (Gicić & Subasi, 2019; Oh, Hong, & Baek, 2019). Moreover, DL models do not estimate the parameters that relate predictors to the outcome variable (the models are non-parametric). Such a black box feature can make their rationale and forecasts challenging to explain (Jadwal, Jain, & Agarwal, 2020; Guo et al., 2021).

Credit risk problems are often associated with the fact that the class label of bad credits is much smaller than that of good ones. This significantly reduces the effectiveness of binary classifiers, undesirably biasing the results toward the prevailing class while the minority class may also be of interest (Hou et al., 2020). There have been two leading solutions offered for imbalanced data classification problems. The strategies can be divided into two types, data-level solutions and algorithm-level methods. Random oversampling (ROS), random under-sampling (RUS), and the SMOTE algorithm have been widely used to rebalance imbalanced datasets for data-level solutions. For instance, Song and Peng (2019) evaluated several imbalanced classifiers for credit risk prediction using a multi-criteria decision-making (MCDM)-based method and proved that the SMOTEBoost-based model was more effective for imbalanced data classification than other methods. Hou et al. (2020) proposed a novel DES technology called META-DESKNN-MI for the imbalanced classification problem and applied it to credit risk assessment. The SMOTE method was initially used to balance the training data set, and then the idea of DES-MI was introduced to assign higher weights to the minority instances to further improve the performance of the ensemble classifier, with which the frameworks of META-DES and DES-KNN are combined. The results demonstrated the better performance of the ensemble classifier compared with that of a single classifier. Shen et al. (2019) proposed a novel ensemble classification model based on neural networks and classifier optimization techniques for imbalanced credit risk evaluation. They employed the adaptive boosting (AdaBoost) approach to construct the ensemble model by combining back propagation (BP) classifiers, the particle swarm optimization (PSO) technique, and SMOTE. Empirical results showed that the proposed hybrid classification model is superior to the individual classification models. In addition to data-level solutions, algorithm-level methods can help tackle data imbalance problems to improve the learning process of a classification algorithm. For this method type, a minority sample can be assigned with additional weight or higher misclassification costs to increase its influence. For instance, Shen et al. (2020) developed an improved synthetic minority oversampling technique (SMOTE) method combined with the long-short-term-memory (LSTM) network and the AdaBoost algorithm to train and learn the processed credit data. The experimental test results indicated that the proposed DL ensemble model was generally more competitive when addressing imbalanced credit risk evaluation problems than other models. Song et al. (2020) constructed a distance-to-model with adaptive clustering-based multi-view ensemble (DM-ACME) learning method for predicting default risk in P2P lending. Their experimental results demonstrated the superiority of the proposed method in loan default prediction.

However, some constraints in the real world should be considered when establishing a credit risk assessment model because (1) the number of investigators who check the high-risk transactions are limited, and (2) the two types of misclassifications have different costs. Adequate fraud detection allows investigators to take timely actions to potentially prevent financial losses (Xia et al., 2020). However, investigators can only check a few alerts per day in practice since the investigation process can be long and tedious (Yu et al., 2018). An advanced model should provide financial institutions with more sophisticated and accurate tools to provide faster response times, reduced costs, and better interpretability (Hou et al., 2020; Papouskova & Hajek, 2019). The following aspects should also be taken into account when establishing a risk assessment model with superior performance: first,

strengthen the analysis of user behavior; second, judge the correlation between user behavior and credit risk; and third, build a user credit risk assessment system to provide a real-time search function for the platform. A powerful yet effective solution is a promising improvement in this area (Xu et al., 2020).

In summary, prior studies on credit risk assessment models for financial institutions focus on improving imbalanced data or classification accuracy through multistage modeling and DL. Although these methods can somewhat boost accuracy, the following aspects need to be further addressed: low time responsiveness, transparency, interpretability, and analysis of behavior features (Laughlin, Sankaranarayanan, & El-Khatib, 2020).

3. ALGORITHM PRINCIPLE

3.1 SMOTE

This paper uses the synthetic minority oversampling technique (SMOTE) to balance the number of cases from each type to reduce the effects of the class imbalance problem on modeling (Chawla et al., 2002). It generates artificial data online among minority instances, which is easy and effective in enhancing the accuracy of defaults (Niu, Cai, & Cai, 2020). The algorithm flow is as follows.

1. Set the sampling magnification *N* according to the imbalance degree.
2. For each positive class sample *x* of the data set, calculate *k* nearest neighbor samples and randomly select *N* among the *k* samples, denoted as generating *N* new samples according to Equation (1).

Add the *N* samples added to each positive class to the original data set to form a new sample data set as follows:

$$\text{NewMinority} = x + rand(y_i - x)(i = 1, 2, \cdots N), \tag{1}$$

where *rand* is a random number within (0, 1) and New Minority represents the newly synthesized sample.

3.2 XGBoost Feature Extraction Process

The full name of XGBoost is extreme Gradient Boosting. It is a C++ implementation of the Gradient Boosting Machine algorithm by Chen and Guestrin (2016), which has attracted a lot of attention in international big data competitions such as Kaggle and Data Castle with its speed and precision compared to Boosted Tree. XGBoost is an engineering implementation that follows the Boosted Tree idea but at the same time considers both system optimization and machine learning principles to maximize scalability, convenience, and accuracy (Zhou et al., 2019; Xia et al., 2017). The most prominent feature of XGBoost is that it can automatically use the multi-threading of the CPU for parallelism, and the algorithm is improved to enhance the accuracy (Chang, Chang, & Wu, 2018). For the traditional GBDT algorithm, only the first-order derivative information is used. When training the n_{th} tree, the residual of the former *n-1* tree is needed, and it is challenging to implement the distribution. XGBoost performs a second-order Taylor expansion on the loss function and adds the regular term to the optimal solution in addition to the loss function to balance the loss function and the complexity of the model to avoid over-fitting (Xia et al., 2017).

Given the data set $D = \{(x_i, y_i)\}(\left|D\right| = n, x_i \hat{I} R^m, y_i \hat{I} R)$, the integrated model of trees is expressed as

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F,$$
$$F = \{f(x) = w_{q(x)}\}(q : R^m \rightarrow T, w \in R^T), \tag{2}$$

where k is the number of models in the integrated model, F is a set space of regression trees, xi represents the eigenvectors of the first i data points, q represents the index of the leaf, T represents the number of leaves on a tree, and every $f_k(\cdot)$ corresponds to an independent tree structure $q(\cdot)$ and the leaf w.

The objective function consists of two parts, that is,

$$O(\theta) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where the first part $l(\hat{y}_i, y_i)$ is the training error between the predicted value $\hat{y}_i$ and real value y_i of the target, and the second part $\Omega(f_k)$ is the sum of the complexity of each tree as

$$\Omega(f) = \lambda T + \frac{1}{2}\lambda \left\| w \right\|^2, \tag{4}$$

where λ and γ are coefficients. The minimum value of the objective function is its optimal value.

In Equation (3), the objective function of the integrated decision tree model is trained by boosting method, i.e., each time the model is retained, a new function is added to the model as follows:

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0, \\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i), \\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_{21}(x_i), \\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i),
\end{aligned} \tag{5}$$

where $\hat{y}_i^{(t)}$ is the predicted value of sample *i* in the round *t*, and $\hat{y}_i^{(t)}$ retains the predicted value of the model $\hat{y}_i^{(t-1)}$ in the round *t-1* and then adds a new function $f_t(x_i)$. The choice of adding a new function in each round is to minimize the objective function. Thus, the objective function turns to

$$O^{(t)} = L^{(t)} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C. \tag{6}$$

The objective function can be optimized through $f_t(x)$. When the error function $l(\hat{y}_i, y_i)$ is squared error, the objective function can be written as

$$L^{(t)} = \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_i(x_i) + f_i(x_i)^2] + \Omega(f_t) + C. \tag{7}$$

For other error functions in addition to squared errors, Taylor expansion is used to approximately define the objective function as

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), \tag{8}$$

$$\tilde{L}^{(t)} \approx \sum_{i=1}^{n}[l(y_i,\hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) + C. \tag{9}$$

After the constant term is removed, a relatively uniform objective function is obtained as

$$\tilde{L}^{(t)} \approx \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t). \tag{10}$$

We define $I_j = \left\{ i \middle| q(x_i) = j \right\}$ as the jth leaf node, that is,

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \lambda T + \frac{1}{2}\lambda\sum_{J=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \lambda T.\end{aligned} \tag{11}$$

Then the above formula is derived, and the result of the derivation is equal to

$$w_j' = -\frac{\sum_{i\bar{I}I_j} g_i}{\sum_{i\bar{I}I_j} h_i + \lambda}. \tag{12}$$

Substituting the w_j optimal solution w_j' into the objective function, we obtain

$$L^{(t)}_{(q)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \lambda T. \tag{13}$$

Compared to the use of partitioning criteria to minimize the mean square error in GBDT, it is impossible to enumerate all tree structures under normal circumstances. XGBoost uses a greedy algorithm to add partitions to existing leaf nodes each time. The hypothesis sum is a collection of left and right nodes after segmentation. Information gain is as follows:

$$Gain = \frac{1}{2}[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}] - \lambda. \tag{14}$$

The information gain uses a certain value after segmentation to reduce a specific value before segmentation. To limit the growth of the tree, a threshold is added. When the gain is greater than the threshold, the node is divided, but the coefficient of the leaf node is in the regular term. Therefore, it is equivalent to doing pre-reduction when optimizing the objective function.

The algorithm for finding the optimal segmentation and segmentation points using the greedy algorithm is as follows:

Algorithm : Exact Greedy Algorithm for Split Finding

Input : I, instance set of current node

$gain \neg 0$

$G \neg \sum_{i \in I} g_i, H \neg \sum_{i \in I} h_i$

for $k = 1$ *to* m *do*

$G_L \neg 0, H_L \neg 0,$

for j *in* $sorted(I, by \ x_{jk})$ *do*

$G_L \neg G_L + g_j, H_L \neg H_L + h_j$

$G_R \neg G + G_L, H_R \neg H + H_L$

$score \neg \max(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda})$

end

end

Output : Split and default directions with max gain

The algorithm introduces the penalty term of a new leaf in the enumeration process and then calculates the complexity of the tree. When the gain brought by the introduced segmentation is less than a threshold, the segmentation is cut off to avoid overfitting. According to the decision tree generated, the weight of each node can be obtained, and then according to the weighted ranking, important features can be extracted.

4. DESIGN OF THE FUSION METHOD

4.1 Design Concept

The design concept of the new method proposed by this paper is to combine the features of the latest feature extraction algorithms and data samples.

First, for the class imbalance problem, we choose SMOTE for balance processing. This method is simple in practice, keeping the important information in the samples, enhancing the training effect of the model, and overcoming the overfitting problem in the oversampling (Shen et al., 2019).

Second, the selection and optimization design of the feature extraction model is based on the use of XGBoost, which can process sparse data through distributed and parallel computing. XGBoost method is a powerful emerging tool to mitigate the existing problems in machine learning and data mining. XGBoost can generate better classification accuracy, avoid the overfitting problem, handle missing values, and calculate ten times faster than off-the-shelf packages (Parsa et al., 2019). Compared with deep neural networks, XGBoost can handle tabular data well and has some features not available in deep neural networks, such as the interpretability of the model, the invariance of the input data, and easier adjustment. Also, a prominent feature of XGBoost is that it runs very fast (Li, Abdel-Aty, & Yuan, 2019). The integration of the XGBoost and SMOTE algorithm successfully solves three key problems in the financial fraud domain, namely class-imbalance, model interpretability, and real-time.

Third, combined with the characteristics of business data, it solves the problem of acquiring interpretable outliers' knowledge to provide an effective method for outlier knowledge management (Xia et al., 2017).

4.2 SMOTE and XGBoost Fusion Method Description

The technical roadmap of this paper is shown in Figure 1, which provides an operational process from the logical and basic framework.

For data resampling in the first stage, random undersampling and oversampling are two simple resampling methods. However, these two resampling methods have class underfitting and overfitting problems, respectively. SMOTE is one of the most well-known oversampling techniques that create new

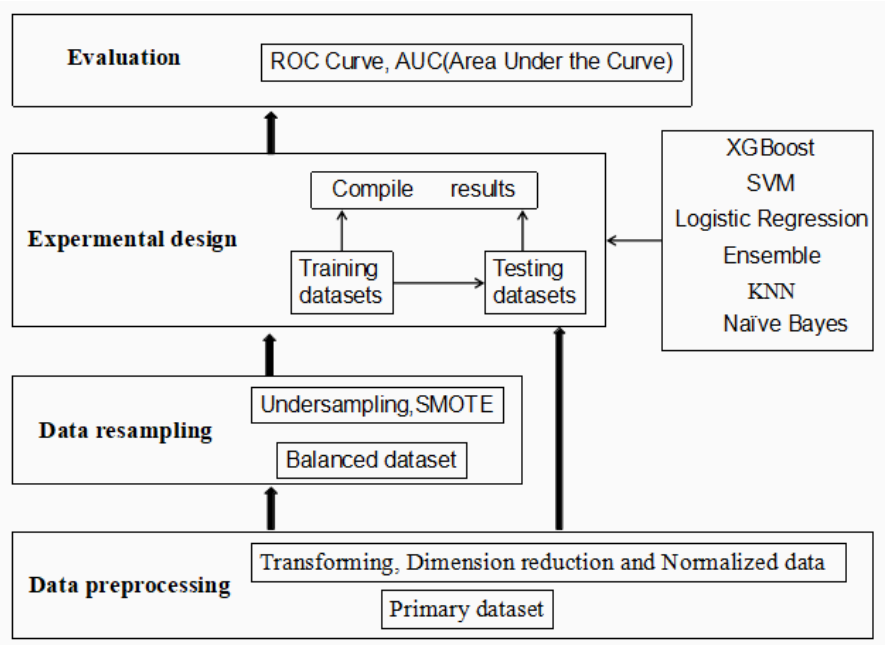
minority samples by randomly interpolating the instances on the line segments of neighboring minority instances (Hou et al., 2020; Shen et al., 2019). Since the introduction of SMOTE, more than 100 variants of the technique have been developed. Kovács (2019) compared 85 SMOTE variants using 104 imbalanced data sets. The results indicate that the classification performance largely depends on the classifier used; after the performance was averaged on 104 data sets, the advantages of the SMOTE variants were not as obvious as expected. In our work, we opted for RUS and SMOTE to address the problem of unbalanced datasets.

The second stage is experimental design; a powerful yet effective solution is a promising improvement in this area. XGBoost yields better classification accuracy and processes calculations ten times faster than the off-the-shelf packages. It can effectively avoid over-fitness and handle missing values. For the sample with incomplete eigenvalues, XGBoost can automatically figure out the splitting direction with learning. Extending Chang (2018), this paper proposes a new method that combines XGBoost and SMOTE; at the same time of classification, important features can be sorted, which makes the model more transparent and the experimental results more explanatory. A series of experiments were conducted on a credit dataset that had been artificially modified using two resampling methods and six prediction models, including SVM, K-Nearest Neighbor (KNN), Naïve Bayes, Multi-Layer Perception (MLP), Logistic Regression, and Ensemble. The third stage evaluates the experimental results; for the classification of unbalanced data, it is appropriate to select the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) value and run time as a reference to verify the superiority of the proposed model.

5. EXPERIMENTAL PROCESS

To validate the feasibility of our proposed credit risk evaluation model, we conduct empirical research based on the credit data sourced from the loan book of a financial institution. This paper tests the performance of the proposed method by comparing it with other algorithms.

Figure 1.
Fusion method combining SMOTE and XGBoost



5.1 Data Description

This paper gathers the historical data of the loan books of a financial institution in Taiwan from the year 2010 to 2017 based on the actual practice of its corporate borrowers. We had a total of 284807 samples, of which 84,102 (1.8%) defaulted, and the remaining 200,705 (98,2%) were non-default cases. Borrowing corporate clients usually provide financial statements when applying for credit facilities. Initially, each sample was described by 424 variables.

The growth in the data set dimensions increases the number of samples required for algorithm learning exponentially, and learning from large data sets requires more memory and processing power (Chang, Chang, & Wu, 2018). In addition, as the dimensionality increases, the sparsity of the data will become higher and higher (Coser, Maer-Matei, & Albu, 2019). PCA can synthesize potentially correlated high-dimensional variables into linearly independent low-dimensional variables, called principal components. The new low-dimensional data set will retain the original data variables as much as possible. The principle is to select the eigenvectors corresponding to k larger eigenvalues to form an Eigen matrix, and the final number of dimension reduction k is determined according to the contribution rate of the eigenvectors; that is, the proportion of the first k eigenvalues is greater than 90% (Chen & Guestrin, 2016). We deleted a series of variables with multiple duplicates, missing values, and high correlation through PCA dimension reduction processing. These 424 variables were reduced to 30, including financial variables and non-financial variables variable.

5.2 Data Processing

The model was implemented in the Python environment using the Scikit-Learn machine learning package. The classifier is very sensitive to the difference of the order of magnitude between data, and the resulting classification error is relatively large (Wang, Ning, & Kong, 2019). To minimize its impact and make the comparison experiment more convincing, we used the preprocessing module in the SKlearn library to standardize the data. Specifically, we cleaned up the abnormal values in the dataset and corrected the erroneous data, so the data became more suitable for the model with respect to smoothing noise, data protocol, etc. Finally, we added the ID attribute and variable names for each variable to normalize them and selected 80% of the data for training and the remaining 20% for testing. The visible sample distribution was extremely unbalanced; we cannot achieve the desired results when this data is directly used to train the model. We then used SMOTE and under-sampling to balance the data. Table 1 shows the sample distribution after under-sampling and SMOTE.

5.3 XGBoost output feature importance ranking

Figure 2 shows the results of ranking the importance of features. The analysis results show that Growth in Sales, Current ratio, Employee loyalty, and Gross margin all substantially impact enterprise credit evaluation. The impact of return on equity, normAmout, sales & marketing, and capitalization is also significant. The analysis results have specific practical significance. To sum up, financial institutions focus on the financial situation of enterprises when conducting credit assessments. Only when enterprises become bigger and stronger can they obtain financial services more efficiently. What is interesting is that the proportion of employee loyalty also has a higher percentage. It is a

Table 1.
Positive and negative sample distribution

	train set	test set	Pos-sample	Neg-sample	Total sample
Raw data set	199364	85443	84102	200705	284807
Under-sampling	688	296	492	492	984
SMOTE	353807	88450	214803	227454	442257

literal reflection of enterprise credit, strength, potential, and corporate culture as a soft indicator. Therefore, enterprises should pay attention to the cultivation of employee loyalty at ordinary times. This also provides a new perspective for financial institutions to conduct a credit risk assessment of enterprises, namely, evaluation of the credit risk of the enterprise from employees' income, credit, and cultural background., AUC is used to verify the reliability of the feature analysis.

5.4 Comparative Analysis of Experimental Results

In this part, to verify the effectiveness of the proposed model, we compared it with various benchmark models in the original dataset, SMOTE, and under-sampling datasets, respectively. To simulate a real credit assessment scene, we divided the data set into two parts, the training and test set. By convention, we implemented 80%/20% of the training/test partition, the former for model training and the latter for performance evaluation. For enhancing the robustness of the experiment, 80 cycles of experiments were performed, and the evaluation criteria were averaged.

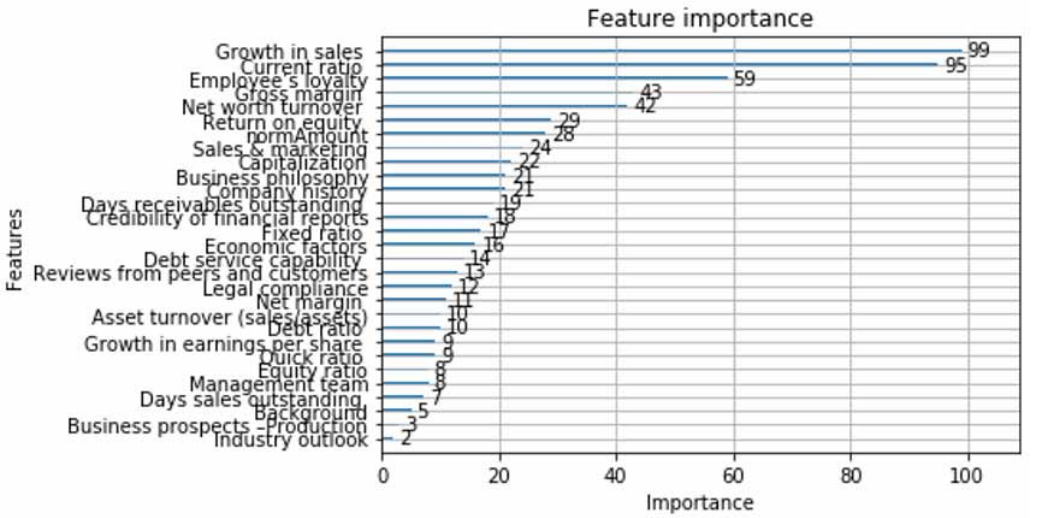
The AUC and Runtime were used as performance measures. These performance measures are suitable for measuring the classification performance of imbalanced data (Fiore et al., 2017). AUC is the area of the ROC curve where the x-axis represents the false positive rate and the y-axis represents the true positive rate. The definitions of false positive and true positive rates are provided in equations (15) and (16), respectively, where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, and TN is the number of true negatives. The running time is also an important index to be considered in the credit risk assessment models (Fiore et al., 2017; Xia et al., 2020; Yu et al., 2018).

$$\text{True positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{16}$$

Through comparison experiments, the results show that XGBoost had the shortest calculation time and the highest AUC value compared with other algorithms on the SMOTE data set. Figures 3-5

Figure 2.
Ranking of feature importance



show the ROC curves of the various algorithms on the three data sets. It can be seen that XGBoost had the largest ROC curve area and was higher than the ensemble algorithm on the original data set and SMOTE data set. For the under-sampling data set, the ROC curve area of logistic regression was the largest, consistent with the previous research conclusions (Im et al., 2012), indicating that the model trained by the logistic regression algorithm for small-scale data sets is the best.

Table 2 shows the AUC and Runtime of various classification algorithms on the SMOTE data set. We can see that the proposed credit risk assessment model had the highest AUC value (0.97), better than KNN (0.91), SVM (0.93), Naïve Bayes (0.947), MLP (0.958), Logistic Regression (0.954), and Ensemble algorithm (0.96). XGBoost also had the fastest Runtime (56.161), better than KNN (8806.192), SVM (85.129), Naïve Bayes (8836.729), MLP (8839.503), Logistic Regression (8854.343), and Ensemble algorithm (17608.932). Through comparative analysis, XGBoost has more advantages in large-scale high-dimensional data.

Table 3 indicates that Logistic Regression had the highest AUC value (0.963), better than Ensemble (0.958) and XGBoos (0.95). Logistic regression had the best performance. This is consistent with previous research conclusions. After under-sampling, the number of samples became smaller; Logistic Regression performed better on small samples.

Table 4 demonstrates that XGBoost had the highest AUC value (0.942), better than KNN (0.538), SVM (0.72), Naïve Bayes (0.938), MLP (0.92), Logistic Regression (0.906) and Ensemble algorithm (0.93). XGBoost also had the fastest Runtime (89.798), better than KNN (3054.946), SVM (103.606), Naïve Bayes (3081.592), MLP (3088.817), Logistic Regression (3117.940), and Ensemble algorithm (5639.206). But XGBoost's AUC value was lower than the AUC value on the data set processed by the SMOTE; it indicates that the advantage of XGBoost on large-scale data sets was more prominent.

Table 5 shows the difference between the AUC of the model proposed in this paper and other algorithms, which indicates that the class imbalance affected the training effect. It is necessary to solve the imbalanced problem. In summary, the model proposed in this paper is more suitable for credit risk assessment, and the analysis of outlier features is also effective.

Figure 3.
ROC of the classification algorithm after SMOTE

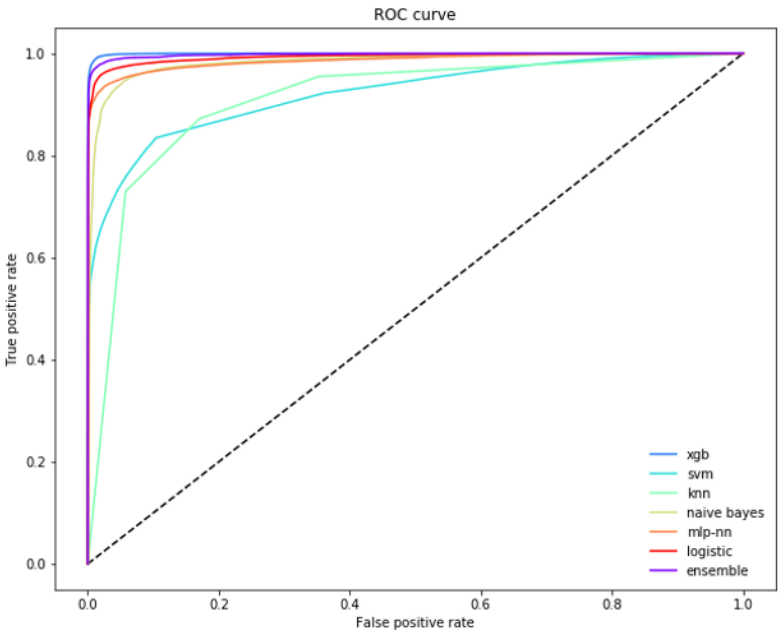


Figure 4.
ROC of the classification algorithm on the original sample

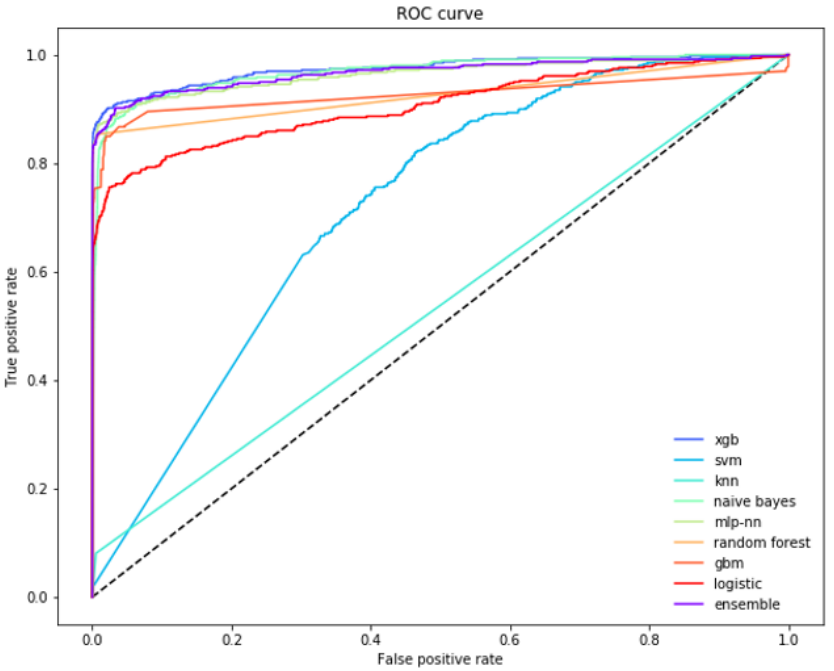


Figure 5.
ROC of the classificationalgorithm after under-sampling

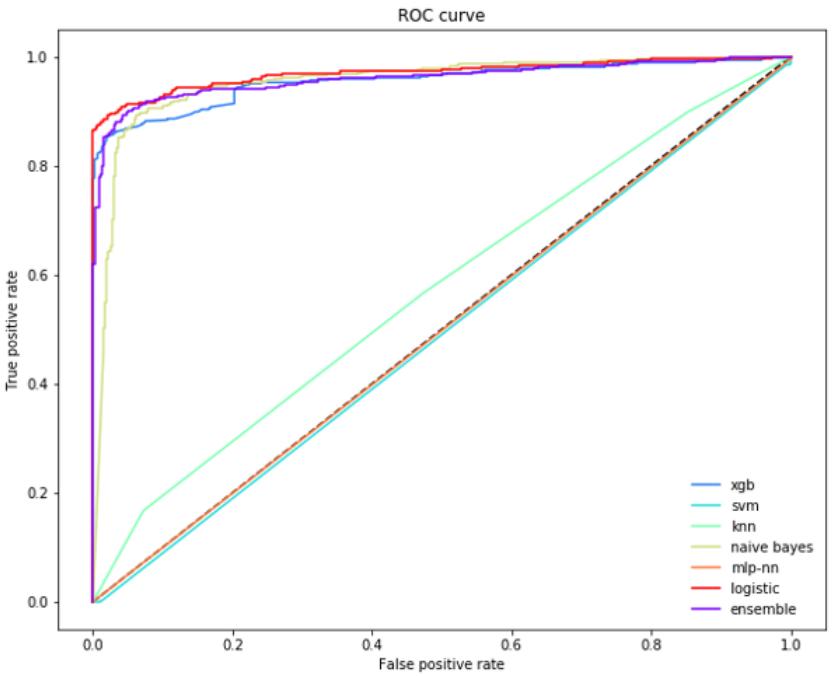


Table 2.
AUC and runtimes for classification algorithms after SMOTE

	AUC	Runtime
XGBoost	0.97	56.161
SVM	0.93	85.129
KNN	0.91	8806.192
Naïve Bayes	0.947	8836.729
MLP	0.958	8839.503
Logistic Regression	0.954	8854.343
Ensemble	0.96	17608.932

Table3.
AUC and runtimes for classification algorithms after undersampling

	AUC	Runtime
XGBoost	0.95	0.062
SVM	0.49	0.052
KNN	0.571	0.032
Naïve Bayes	0.94	0.023
MLP	0.498	0.096
Logistic Regression	0.963	0.020
Ensemble	0.958	0.102

Table 4.
AUC and Runtimes for classification algorithms on the original data set

	AUC	Runtime
XGBoost	0.942	89.798
SVM	0.72	103.606
KNN	0.538	3054.946
Naïve Bayes	0.938	3081.592
MLP	0.92	3088.817
Logistic Regression	0.906	3117.940
Ensemble	0.93	5639.206

Table 5.
The difference in AUC between under-sampling and the original dataset

	Original data set	Under-sampling
XGBoost	0.028	0.2
SVM	0.21	0.44
KNN	0.372	0.339
Naïve Bayes	0.009	0.007
MLP	0.038	0.46
Logistic Regression	0.048	-0.09
Ensemble	0.03	0.02

6. CONCLUSION

This paper proposes a new method for credit risk assessment combining SMOTE and XGBoost. To verify the effectiveness of the method, it was compared with SVM, KNN, Naïve Bayes, MLP, Logistic Regression, and Ensemble with AUC and Runtime as the evaluation index of the model. The results show that XGBoost had the shortest time and the highest AUC value. At the same time, XGBoost can obtain a ranking of important features that impact the classification results when performing classification tasks, making the evaluation results of the model interpretable. The model proposed in the paper is more practical in solving credit risk assessment problems; it has faster response times, reduced costs, and better interpretability. The main contributions of this paper are as follows.

6.1 Theoretical Contribution

The primary contribution of this study is the development of an improved credit risk assessment model. While the model completes the classification task, the importance ranking of features that affect the classification results can be obtained, and new knowledge can be discovered through the analysis of important features. Our study provides a powerful yet effective solution to simplify the complex nature of model development, which is a promising improvement and can be readily applied by financial institutions in practice. The credit risk evaluation method proposed by this paper shows exceptional differentiating accuracy evaluated with AUC values, compared to the other common classifiers such as logistic regression, SVM, and MLP. At the same time, it can meet the requirements of the financial domain on interpretability and real-time.

Another contribution of this study is to compare the performance of resampling models when applied together with statistical and machine learning methods to predict the probability of default settings using a real-world credit dataset. Specifically, the results demonstrated performance gains when introducing resampling strategies into artificial intelligence methods, which significantly benefits the borrowers' solvency prediction accuracy. This can provide decision-makers with efficient and effective tools to more accurately assess credit risk.

6.2 Practical Contribution

First, the implementation of sampling strategies may help financial institutions reduce erroneous classification costs compared to the unbalanced data. These strategies improved the performance of prediction models concerning unbalanced data independent of the classifier used. They can also identify insolvent clients better than unbalanced data. After a detailed analysis, we concluded that the ideal distribution is certainly not of origin for a classifier.

Second, by analyzing the ranking of essential features derived from XGBoost, helpful knowledge of financial fraud risk can guide financial institutions to reduce the risks correlated with enterprise defaults and improve loan efficiency. This can serve as a reference to the credit department of financial institutions to reduce the risks associated with corporate defaults and improve loan business efficiency. For example, financial institutions should focus on the asset strength and credit of enterprises. It is interesting that employees' loyalty is an indicator that financial institutions should focus on in the credit assessment, so enterprises should cultivate soft power at ordinary times. This also provides a new perspective for financial institutions to conduct a credit risk assessment of enterprises, i.e., evaluating the credit risk of the enterprise from employees' income, credit, and cultural background.

Third, the model proposed by this paper is relatively simple and has strong operability, which makes some contributions to financial institutions and banks. Although some existing models can achieve higher classification accuracy (such as DL and multistage modeling), their model structure is complex and challenging to implement. While we have presented our framework in the context of credit risk assessment, it is quite general and can be readily extended to other application domains characterized by significant class imbalance rates.

6.3 Limitations and Future Research Direction

The limits of the research are as follows. This model is not very efficient in detecting new patterns of normal and fraudulent behaviors. For the class imbalance, scholars have proposed a large number of new algorithms. Considering the ease of using the classification model, we used a relatively simple SMOTE and only compared it with the under-sampling. We did not use the newly proposed algorithm. In future research, we will try to use a new algorithm to solve the class imbalance problem. Besides, for fairness, we did not optimize the parameters; all models used default parameters. In this paper, after balancing the data set, the ratio of positive and negative samples is 1:1. In future research, different sampling rates can be tried to improve the classification performance of the model.

REFERENCES

Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. *IEEE Access : Practical Innovations, Open Solutions*, *6*, 33789–33795. doi:10.1109/ACCESS.2018.2841987

Ala'Raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89–105. doi:10.1016/j.knosys.2016.04.013

Altinbas, H. (2020). Metaheuristic Algorithms and Modern Credit Classification Methods: A Systematic Review. *Istanbul Business Research*, *49*(1), 146–175.

Arora, N., & Kaur, P. D. (2019). A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*(2), 96–101.

Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. *Expert Systems with Applications*, *134*(6), 209–224. doi:10.1016/j.eswa.2019.05.042

Cai, S., & Zhang, J. (2020). Exploration of credit risk of p2p platform based on data mining technology. *Journal of Computational and Applied Mathematics*, *372*, 372. doi:10.1016/j.cam.2020.112718

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). Scarff: A scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, *41*(4), 182–194. doi:10.1016/j.inffus.2017.09.005

Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of extreme Gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*(3), 112–116. doi:10.1016/j.asoc.2018.09.029

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(1), 321–357. doi:10.1613/jair.953

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review*, *45*(1), 1–23. doi:10.1007/s10462-015-9434-x

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. doi:10.1145/2939672.2939785

Cho, P., Chang, W., & Song, J. W. (2019). Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 16925–16939. doi:10.1109/ACCESS.2019.2896474

Chua, C. E. H., & Storey, V. C. (2016). Dealing with dangerous data: Part-whole validation for low incident, high risk data. *Journal of Database Management*, *27*(1), 29–57. doi:10.4018/JDM.2016010102

Coser, A., Maer-Matei, M. M., & Albu, C. (2019). Predictive models for loan default risk assessment. *Economic computation and economic cybernetics studies and research/Academy of Economic Studies, 53*(2), 149-165.

Fantinato, M., Peres, S. M., Kafeza, E., Chiu, D. K., & Hung, P. C. (2021). A Review on the Integration of Deep Learning and Service-Oriented Architecture. [JDM]. *Journal of Database Management*, *32*(3), 95–119. doi:10.4018/JDM.2021070105

Feng, X. D., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing*, *65*, 139–151. doi:10.1016/j.asoc.2018.01.021

Fiore, U., Santis, A. D., Perla, F., Zanetti, P., & Palmieri, F. (2017). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 448–455.

Fonseca, D. P., Wanke, P. F., & Correa, H. L. (2020). A two-stage fuzzy neural approach for credit risk assessment in a Brazilian credit card company. *Applied Soft Computing*, *92*(5), 55–61. doi:10.1016/j.asoc.2020.106329

Gicić, A., & Subasi, A. (2019). Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, *36*(2), e12363. doi:10.1111/exsy.12363

Guo, Y., Jiang, S., Qiao, H., Chen, F., & Li, Y. (2021). A new integrated similarity measure for enhancing instance-based credit assessment in P2P lending. *Expert Systems with Applications*, *175*, 114798. doi:10.1016/j.eswa.2021.114798

Hou, W. H., Wang, X. K., Zhang, H. Y., Wang, J. Q., & Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems*, *208*, 106462. doi:10.1016/j.knosys.2020.106462

Im, J. K., Apley, D. W., Qi, C., & Shan, X. (2012). A time-dependent proportional hazards survival model for credit risk analysis. *The Journal of the Operational Research Society*, *63*(3), 306–321. doi:10.1057/jors.2011.34

Jadwal, P. K., Jain, S., & Agarwal, B. (2020). Spectral Clustering and Cost-Sensitive Deep Neural Network-Based Undersampling Approach for P2P Lending Data. [IJITWE]. *International Journal of Information Technology and Web Engineering*, *15*(4), 37–52. doi:10.4018/IJITWE.2020100103

Khanuja, H. K., & Adane, D. (2018). Monitor and Detect Suspicious Transactions With Database Forensic Analysis. *Journal of Database Management*, *29*(4), 28–50. doi:10.4018/JDM.2018100102

Khemakhem, S., Said, F. B., & Boujelbene, Y. (2018). Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines. *Journal of Modelling in Management*, *13*(4), 932–951. doi:10.1108/JM2-01-2017-0002

Kim, A., & Cho, S. B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, *81*, 193–199. doi:10.1016/j.engappai.2019.02.014

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. doi:10.1016/j.eswa.2014.08.025

Kogeda, O. P., & Vumane, N. N. (2017). A model augmenting credit risk management in the banking industry. [IJTD]. *International Journal of Technology Diffusion*, *8*(4), 47–65. doi:10.4018/IJTD.2017100104

Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera-Viedma, E. (2019). Machine learning methods for systemic risk analysis in financial sectors. *Technological and Economic Development of Economy*, *25*(5), 716–742. doi:10.3846/tede.2019.8740

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*(7), 445–463. doi:10.1016/j.asoc.2019.105662

Laughlin, B., Sankaranarayanan, K., & El-Khatib, K. (2020). A service architecture using machine learning to contextualize anomaly detection. [JDM]. *Journal of Database Management*, *31*(1), 64–84. doi:10.4018/JDM.2020010104

Li, P., Abdel-Aty, M., & Yuan, J. (2019). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident; Analysis and Prevention*, *135*(8), 667–672. PMID:31783334

Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020). Heterogeneous Ensemble learning with feature engineering for default prediction in peer-to-peer lending in china. *World Wide Web (Bussum)*, *23*(1), 23–45. doi:10.1007/s11280-019-00676-y

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, *74*, 105–114. doi:10.1016/j.eswa.2017.01.011

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

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, *161*, 113567. doi:10.1016/j.eswa.2020.113567

Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*, *11*(3), 699. doi:10.3390/su11030699

Niu, A., Cai, B., & Cai, S. (2020). Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm. *Complexity*, *2020*, 2020. doi:10.1155/2020/8563030

Niu, K., Zhang, Z., Liu, Y., & Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information Sciences*, *536*, 120–134. doi:10.1016/j.ins.2020.05.040

Oh, J. H., Hong, J. Y., & Baek, J. G. (2019). Oversampling method using outlier detectable generative adversarial network. *Expert Systems with Applications*, *133*, 1–8. doi:10.1016/j.eswa.2019.05.006

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications*, *41*(4), 2052–2064. doi:10.1016/j.eswa.2013.09.004

Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of Money, Credit and Banking*, *2*(4), 435–445. doi:10.2307/1991095

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, *118*, 33–45. doi:10.1016/j.dss.2019.01.002

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2019). Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident; Analysis and Prevention*, 136. PMID:31864931

Rushin, G., Stancil, C., Sun, M., Adams, S., & Beling, P. (2017). Horse race analysis in credit card fraud-deep learning, logistic regression, and Gradient boosted tree. *Systems and Information Engineering Design Symposium* (SIEDS), (pp. 117-121). IEEE.

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending. *Decision Support Systems*, *89*(C), 113–122. doi:10.1016/j.dss.2016.06.014

Shen, F., Zhao, X., Kou, G., & Alsaadi, F. E. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, *98*, 106852. doi:10.1016/j.asoc.2020.106852

Shen, F., Zhao, X., Li, Z., Li, K., & Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A*, *526*, 121073. doi:10.1016/j.physa.2019.121073

Song, Y., & Peng, Y. (2019). A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 84897–84906. doi:10.1109/ACCESS.2019.2924923

Song, Y., Wang, Y., Ye, X., Wang, D., & Wang, Y. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in p2p lending. *Information Sciences*, *525*, 182–204. doi:10.1016/j.ins.2020.03.027

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. doi:10.1016/j.eswa.2019.05.028 PMID:32968335

Tomczak, J. M., & Zięba, M. (2015). Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications*, *42*(4), 1789–1796. doi:10.1016/j.eswa.2014.10.016

Wang, B., Ning, L., & Kong, Y. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301–315. doi:10.1016/j.eswa.2019.02.033

Wang, Y., & Zhang, Y. (2020). Credit Risk Assessment for Small and Microsized Enterprises Using Kernel Feature Selection-Based Multiple Criteria Linear Optimization Classifier: Evidence from China. *Complexity*, *2020*, 2020. doi:10.1155/2020/2394948

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, *15*(3), 757–770. doi:10.2307/2330408

Xia, H., Long, J., Li, F., & He, T. (2017). Outlier Detection and Outlier Knowledge Based on High Frequency Keywords: A Study from the Literature Analysis Perspective. *Journal of Intelligence*, *32*(5).

Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. doi:10.1016/j.eswa.2017.02.017

Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, *24*, 30–49. doi:10.1016/j.elerap.2017.06.004

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications*, *159*, 159. doi:10.1016/j.eswa.2020.113615

Xia, Y. F., Liu, C. Z., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, *93*, 182–199. doi:10.1016/j.eswa.2017.10.022

Xu, D. Y., Chen, J., Zhang, X. Y., & Hu, J. G. (2020). A novel ensemble credit scoring model based on extreme learning machine and generalized fuzzy soft sets. *Mathematical Problems in Engineering*, *2020*, 2020. doi:10.1155/2020/7504764

Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, *263*(2), 679–689. doi:10.1016/j.ejor.2017.05.017

Ye, X., Dong, L. A., & Ma, D. (2018). Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score. *Electronic Commerce Research and Applications*, *32*, 23–36. doi:10.1016/j.elerap.2018.10.004

Yu, L., Zhou, R., Tang, L., & Chen, R. (2018). A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing*, *69*, 192–202. doi:10.1016/j.asoc.2018.04.049

Zeng, W., Xu, H., Li, H., & Li, X. (2018). Research on Methodology of Correlation Analysis of Sci-Tech Literature Based on Deep Learning Technology in the Big Data. [JDM]. *Journal of Database Management*, *29*(3), 67–88. doi:10.4018/JDM.2018070104

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A*, *534*, 122370. doi:10.1016/j.physa.2019.122370

Huosong Xia is a professor in the school of management at Wuhan Textile University. He graduated from Huazhong University of Science and Technology in China and was a visiting scholar at Eller College of Management of the University of Arizona, USA from 2006 to 2007. His main research interests are knowledge management, data mining, e-Commerce, and logistics information system. His publications have appeared in over 100 referred papers in journals, book chapters, and conferences. He has obtained research funding from 4 projects with National Social Science Foundation of China and National Science Foundation of China.

Wuyue An is a master candidate in the school of management, Wuhan Textile University. In 2018, She got her bachelor's degree from the school of software, Zhengzhou University. Her main research interests are knowledge management, data mining, e-Commerce, and Logistics Information System.

Justin Zuopeng Zhang is a faculty member in the Coggin College of Business at University of North Florida. He was previously an Associate Professor of Management, Information Systems, and Analytics at State University of New York at Plattsburgh. He received his Ph.D. in Business Administration with a concentration on Management Science and Information Systems from Pennsylvania State University, University Park. His research interests include economics of information systems, knowledge management, electronic business, business process management, information security, and social networking. He is the editor-in-chief of the Journal of Global Information Management, an ABET program evaluator, and an IEEE senior member.