Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/85/20240868

Analysis of Credit Default Prediction Based on Logistic Regression, Random Forest and KNN Model

Yutong Wen[1,a,*]

[1]*School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China*
a. 1084366562@stud.tjut.edu.cn
**corresponding author*

Abstract: Addressing the prevalent issue of limited access to loans for individuals with inadequate or non-existent credit histories is imperative to foster financial inclusion and safeguard vulnerable populations from unscrupulous lenders. With the help of unique characteristics, this research seeks to improve financial accessibility for the unbanked population by creating a safe and satisfying financing environment. The paper uses an imbalanced dataset that was gathered from Kaggle to examine how machine learning techniques can be applied to forecast credit risk. specifically, Logistic Regression, Random Forest, KNN, and Synthetic Minority Over-sampling Technique (SMOTE) are utilized in this study. The findings reveal an enhancement in accuracy following the resolution of the imbalanced dataset challenge. By employing these machine learning models, the research seeks to not only bridge the credit gap for the underserved but also mitigate the risks associated with lending to individuals lacking conventional credit histories. A thorough strategy for managing credit risk in the context of financial inclusion is shown by the investigation of different algorithms and the application of methods. These results contribute to the discourse on leveraging machine learning to create more equitable lending practices, ensuring that the unbanked population can access financial resources responsibly.

Keywords: Credit risk, machine learning, Random Forest, Logistic Regression, KNN

1. Introduction

With the currency stability achieved through the Economic Plan "Plano Real" in 1994, financial loans became a lucrative business for banks, as they no longer experienced significant profits from currency devaluation. In order to replace this source of profitability, there was a recognized need to diversify investment alternatives as the inflationary period came to an end. Subsequently, financial institutions made efforts to expand their credit portfolios. However, indiscriminate lending to all applicants was not a viable approach, necessitating the development of methods to evaluate loan candidates [1]. Some years ago, when individuals applied for loans, they had to complete a proposal for evaluation by one or more analysts. These analysts would then provide their opinions on the loan request. While effective, this process was slow, especially when dealing with a large number of requests. Consequently, financial institutions introduced a model for credit concession analysis to expedite the evaluation of loan proposals. Credit scorecard systems are extensively utilized in the banking industry today. By awarding scores, these systems evaluate the credit standing of borrowers who are already enrolled in a program or who are looking to apply for new loans. The many methods used to calculate

DOI: 10.54254/2754-1169/85/20240868

credit scores are created by different lenders. Whatever methods are used, credit scoring always answers the same basic question: how likely is it that a payment would be missed within a given time frame, usually a year [2].

In the conventional lending process, loan institutions typically rely on the "5C" principle (Character, Capacity, Collateral, and Conditions) for a subjective assessment of borrowers' capabilities. This method heavily depends on the personal experience and knowledge of the creditors, drawing most information from customer relationships. Nevertheless, this approach encounters substantial limitations, particularly in the realm of retail credit, where customers exhibit high mobility and urgent demands for business expansion [3].

Business Analytics (BA) is a practical approach that leverages Business Intelligence (BI) techniques to meet the needs of businesses, particularly in predicting behaviors and outcomes. BA utilizes the capabilities of Information Technology (IT), including Data Mining, to address business requirements. The fundamental concept of BA involves integrating the expertise from the IT domain with that of the business domain to achieve effective collaboration.

In the realm of IT, Machine Learning (ML) has been a valuable contributor to solving business prediction challenges. A noteworthy machine learning approach proposed in focuses on predicting customer churn in the telecommunications industry using Genetic Programming (GP) [4]. Churn management holds significant importance in the business domain, directly impacting customer retention strategies. Analyzing customer behaviors is crucial and serves as an early warning mechanism, subsequently prompting business activities related to risk prevention and ensuring business continuity [5].

2. Data and Method

2.1. Data

The data set's origin is an organization that provides finance to underprivileged clients with little or minimal credit history, allowing them to borrow money securely and conveniently both online and off [6]. The data set comprises 307511 observations (each a distinct loan) and 122 features (variables), which encompasses TARGET. The dependent variable that represents the default event is the variable TARGET. If a client does not make the payment, it takes the value of 1, and if the money is made as expected, it takes the value of 0. 67 characteristics have values that are missing. In this work, they have been filled with median.

2.2. Models

Based on the current understanding, Logistic Regression, KNN model, and Random Forest have consistently shown the most effective performance in the realm of classification. Consequently, these three algorithms were specifically chosen for the evaluation of loans in this research. Because of its consistency in error distribution and ease of use, logistic regression is extensively used in a variety of industries. It functions as a binary classification technique, yielding results such as classifying borrowers as superior or inferior. The following is the formula for logistic regression: The probability prediction is denoted by F(x), the constant coefficient is denoted by β0, and the coefficient associated with feature Xi is represented by βi. Maximum likelihood estimate is used to find the coefficients. As such, logistic regression, when applied to a set of features Xi (i=1,...,n), predicts the likelihood that an ensemble will belong to a specific class [7]. Eq. (1) encapsulates the fundamental predictive mechanism of the logistic regression algorithm:

$$Ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \Sigma_i \beta_i X_i \quad (1)$$

DOI: 10.54254/2754-1169/85/20240868

One of the ensemble learning techniques in the tree-based model family is Random Forest. This method is frequently utilized for both classification and regression tasks, and it constructs many decision trees during training. The outcomes of each of these distinct trees are combined to produce the ultimate forecast, which is then classified using a majority vote and regression using the average prediction. Key features include the ensemble of trees, random feature selection at each split to reduce correlations between trees, the use of bootstrap aggregating for improved stability, and a voting mechanism for determining final predictions. Known for its robustness against overfitting, Random Forest is versatile and applicable to various datasets and tasks. Additionally, it provides a measure of feature importance, contributing to its widespread use in practical applications [8]:

$$H(x) = \arg\max_{Y} \Sigma_i I(h_i(x) = Y) \tag{2}$$

Here, Y is the output variable, i is an indicator function, H(x) is the combined classification model, and h_i is a single decision-tree classification model.

KNN (K-nearest neighbors) is a non-parametric, slow learning technique. Since real-world data frequently deviates from recognized theoretical assumptions, being non-parametric means that it draws no assumptions about the foundational data distribution. This is a crucial quality. Such scenarios lend themselves to the usage of non-parametric algorithms such as KNN. Lazy algorithms, particularly KNN, base their decisions entirely on the training dataset [9].

SMOTE (Synthetic Minority Over-sampling Technique), is a widely used technique for resolving class imbalance problems in a variety of fields. The fundamental principle of SMOTE is to apply feature space similarities identified in existing minority cases to generate fresh instances of the minority class. In summary, the SMOTE algorithm does the following for each minority class instance in the imbalanced dataset T: Using the Euclidean distance metric, it first finds the K nearest neighbors for Xi. Next, it chooses one of the K nearest neighbors at random and computes the difference in the feature vectors between Xi and the selected neighbor. Finally, it multiplies the resultant new vector by a stochastic number and adds it to Xi. This process effectively augments the minority class samples, thereby mitigating the challenges posed by class imbalance [10]. Eq. (3) provides the mathematical formulation for creating a new minority sample:

$$X_{new} = X_i + \left(X_i^k - X_i\right)\delta \tag{3}$$

where δ is a random number that falls between 0 and 1, and X_i^k is one of X_i's closest neighbors. As a result, the point on the line segment between X_i and its closest neighbor is the synthesized minority instance X_{new}.

2.3. Evaluation Metrics

Evaluation metrics play a crucial role in predicting credit risk. For credit scoring models, accurate assessment is an indispensable part of ensuring their effectiveness and reliability. This paper uses Confusion Matrix and Accuracy as evaluation metrics. Confusion matrix offers an overview of a classifier's performance in classifying test data. It is typically utilized in these types of contexts where there are two classes, one acting as the positive class and the other as the negative class. Table 1 depicts the matrix's four cells beneath this structure [11].

Table 1: Definition of Confusion Matrix.

	Positive	Negative
Positive	TP (True positive)	FN (False negative)
Negative	FP (False negative)	TN (True positive)

DOI: 10.54254/2754-1169/85/20240868

Accuracy is characterized as the proportion of correctly predicted outcomes to the overall amount of forecasts manufactured [12], which can be calculated as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

3. Results and Discussion

3.1. Features Engineering

Within the dataset, all these data types have the following numbers: 65 for float64, 41 for int64, and 16 for objects. In this study, categorical variables with two or fewer categories are encoded using Label Encoding; categorical variables with more than two categories are encoded using One-Hot Encoding. This paper employed the Spearman rank correlation method to assess the correlation between variables. The results obtained through this paper provide valuable insights into the interrelationships among variables, aiding in a more profound understanding of data features and their associations. Through Spearman correlation analysis in the research, this paper able to unveil patterns of relationships between variables and the target variable. This has offered robust support for subsequent data interpretation and analysis. These are the correlations between features and target as shown in Table 2. This paper chooses DAYS_BIRTH, EXT_SOURCE_3, EXT_SOURCE_2, EXT_SOURCE_1, as features. The three EXT_SOURCE elements that exhibit negative correlations with the target demonstrates that a client's likelihood of repaying the loan increases as the value of EXT_SOURCE increases. Additionally, DAYS_BIRTH and EXT_SOURCE_1 show a positive correlation, suggesting that the client's age may be a contributing factor to this score (Fig. 1).

Table 2: Correlations between features and Target.

Most positive correlations		Most negative correlations	
DAYS_BIRTH	0.0783	EXT_SOURCE_3	-0.1663
DAYS_EMPLOYED	0.0749	EXT_SOURCE_2	-0.1473
REGION_RATING_CLIENT_W_CITY	0.0608	EXT_SOURCE_1	-0.1511

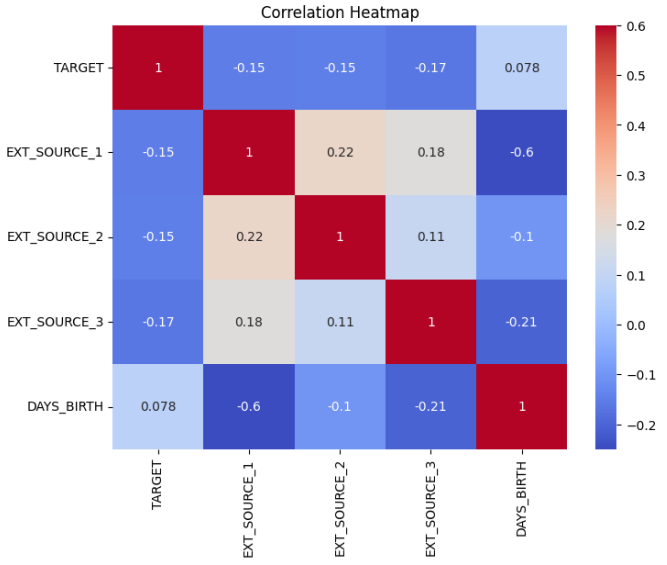

Figure 1: Correlation Heatmap between features and Target (Picture credit: Original).

In this visualization, the color red represents loans that were not repaid, while blue corresponds to paid loans. Various relationships within the data are discernible. However, it is crucial to address the

DOI: 10.54254/2754-1169/85/20240868

significant imbalance in the number of instances labeled as 1 (not repaid) and 0 (paid) as shown in Fig. 2. However, it is crucial to address the significant imbalance in the number of instances labeled as 1 (not repaid) and 0 (paid). This major distinction could have a negative impact on the model's ability to forecast future events. To mitigate this issue, this paper uses the SMOTE method to artificially increase the number of instances, particularly those labeled as 1, and thus create a more balanced dataset for improved model training and prediction accuracy as depicted in Fig. 3. Before using SMOTE model, the number of 0 is 282686 and the number of 1 is 24825. After using SMOTE model, the number of 0 increases to 452264, and the number of 1 increases to 246008.

Figure 2: Scatter Plot of features and Target (Picture credit: Original).

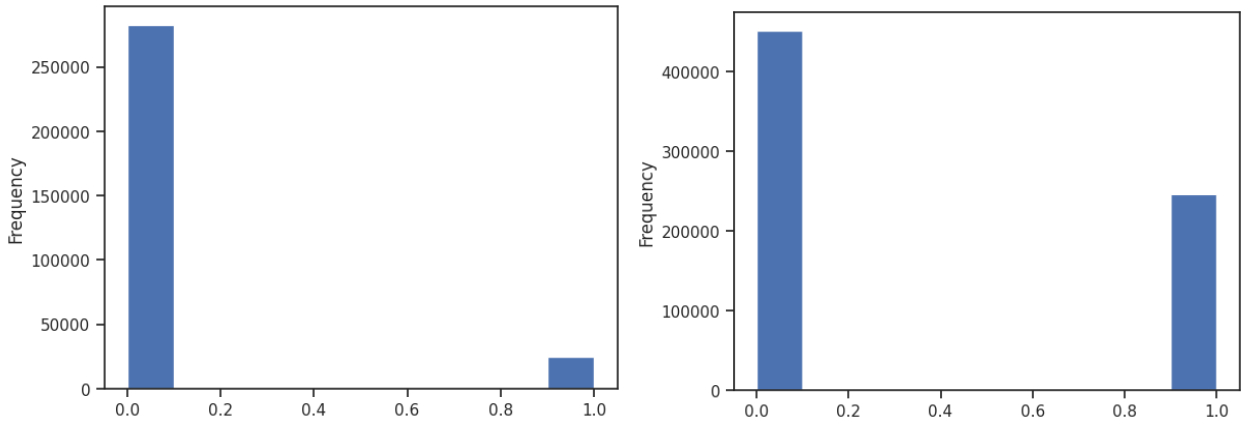

Figure 3: The number of target (Picture credit: Original).

3.2. Model Performance

In this paper, the dataset was separately input into logistic regression, random forest, and KNN model for analysis. In the scatter plot illustrated in Fig. 4, a distinct separation between two classes can be

DOI: 10.54254/2754-1169/85/20240868

observed, indicating that logistic regression is effective in classifying the data points. It reveals a clear decision boundary, demonstrating the capacity of the model to discriminate between the classes. The accuracy of Logistic Regression improves from 67% to 71%.

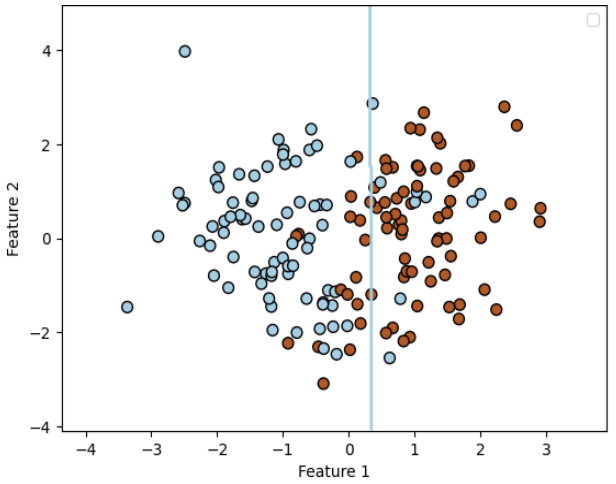

Figure 4: Logistic Regression Output (Picture credit: Original).

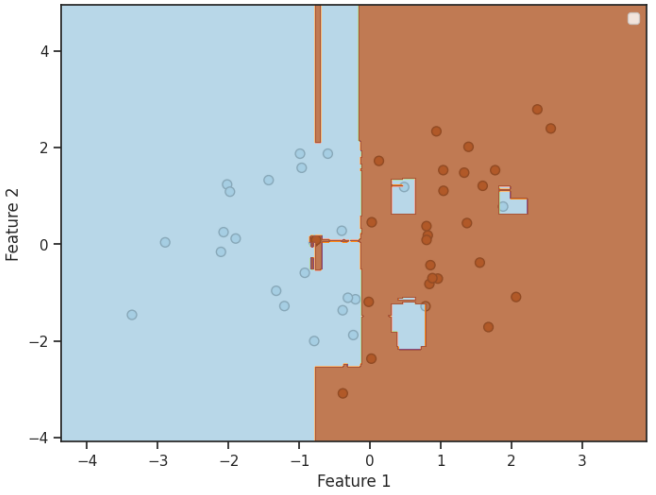

Figure 5: The results of Random Forest (Picture credit: Original).

Seen from Fig. 5, it can be observed a noticeable separation between two classes, signifying the effectiveness of the random forest model in classifying the data points. The plot highlights a well-defined decision boundary, showcasing the model's discriminative prowess across classes. This implies that the random forest is adept at handling the classification task at hand, as evidenced by the clear grouping of points associated with different classes. The distinct separation and discernible pattern evident in the scatter plot visually confirm the robust classification performance of the random forest model. The accuracy of Random Forest improves from 68% to 89%.

Presented in Fig. 6, a distinct separation between two classes is evident, highlighting the effectiveness of KNN model in classifying the data points. The plot emphasizes a well-defined decision boundary, showcasing the model's discriminative capabilities across classes. This suggests that the KNN model excels in handling the current classification task, as indicated by the clear grouping of points associated with different classes. The evident separation and discernible pattern in

DOI: 10.54254/2754-1169/85/20240868

the scatter plot visually affirm the robust classification performance of the KNN model. The accuracy of KNN model is 84%.

After inputting the dataset into the model, a satisfactory accuracy was obtained, and a confusion matrix was generated for a more thorough evaluation of the model's functionality. The accuracy is recorded as shown in Fig. 7, indicating the model's success rate in the overall classification task. The confusion matrix further supplies detailed information regarding the classification outcomes of the model listed in Table 3.

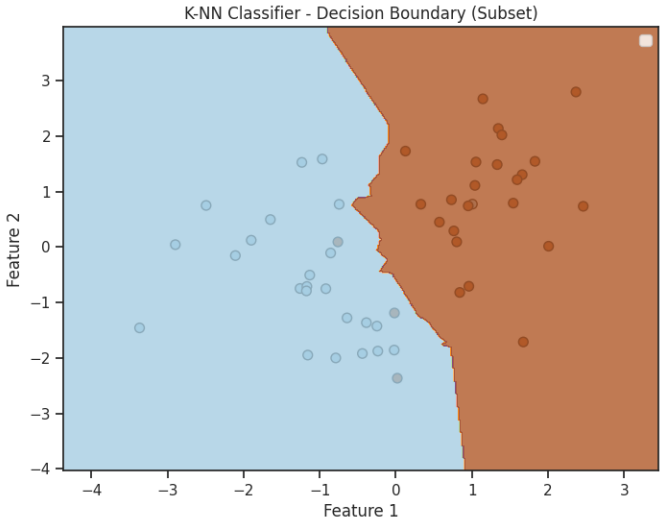

Figure 6: KNN model's outcomes (Picture credit: Original).

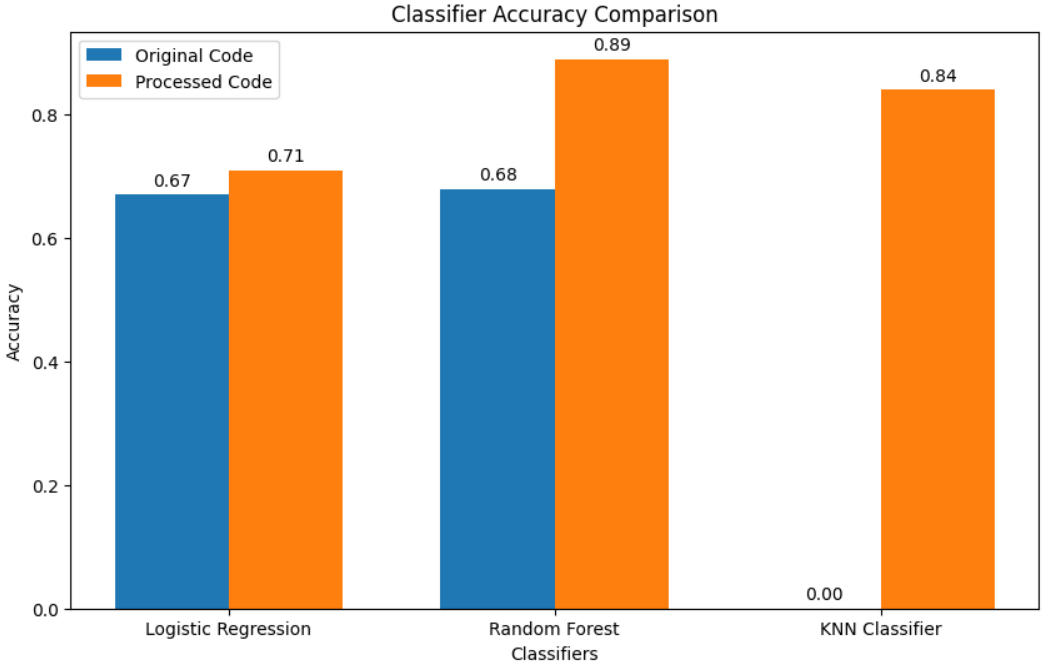

Figure 7: The accuracy of models (Picture credit: Original).

DOI: 10.54254/2754-1169/85/20240868

Table 3: The confusion matrix of models.

	TP	FP	TN	FN
Logistic Regression	94534	34624	4581	5916
Random Forest	87186	10933	38372	3264
KNN model	76263	7176	42029	14187

4. Limitations and Prospects

This paper uses Logistic Regression, Random Forest, and KNN model for credit risk research has certain limitations. Firstly, these models rely on historical data for training and may lack flexibility in rapidly changing market and economic conditions. Secondly, model performance is influenced by feature selection and data quality, and inadequate choices or poor data quality can result in decreased predictive accuracy. Additionally, for extreme cases like emerging financial products or market uncertainties, traditional historical data models may not provide sufficient information. Looking ahead, improving model performance can be achieved by integrating more data sources and employing advanced feature engineering methods. Incorporating emerging technologies such as deep learning can help capture complex credit risk patterns and enhance model generalization.

5. Conclusion

The incorporation of machine learning tackles to forecast credit risk is examined in the one. The accuracy of Random Forest and Logistic Regression both significantly improved when the unbalanced data problem was resolved, which was an impressive improvement. Upon comparison, it was observed that the Random Forest and KNN models are more suitable for this dataset. However, the paper acknowledges a limitation in the feature selection process, which may impact the models' accuracy. To enhance model performance, the suggestion is made to incorporate more complex features. The primary contribution of this paper lies in predicting credit risk through machine learning techniques, achieving a substantial boost in accuracy by addressing the imbalanced data challenge. Furthermore, the comparative analysis of different models offers valuable insights for selecting appropriate models. Nevertheless, the paper notes limitations in feature selection and provides recommendations for future research directions, offering valuable guidance in forecasting credit risk.

References

[1] *Gouvêa, M.A. and Gonçalves, E.B. (2007) Credit risk analysis applying logistic regression, neural networks and genetic algorithms models. In POMS 18th annual conference, 17.*

[2] *Dong,G., Lai, K.K. and Yen, J. (2010) Credit scorecard based on logistic regression with random coefficients. Procedia Computer Science, 1(1), 2463-2468.*

[3] *Li, Y. (2019) Credit risk prediction based on machine learning methods. In 2019 14th International Conference on Computer Science & Education (ICCSE) pp. 1011-1013.*

[4] *Faris, H., Al-Shboul, B. and Ghatasheh, N. (2014) A genetic programming based framework for churn prediction in telecommunication industry. In Computational Collective Intelligence. Technologies and Applications: 6th International Conference, ICCCI 2014, Seoul, Korea, September 24-26, 2014. Proceedings 6 pp. 353-362.*

[5] *Ghatasheh, N. (2014) Business analytics using random forest trees for credit risk prediction: a comparison study. International Journal of Advanced Science and Technology, 72, 19-30.*

[6] *Koehrsen, W. (2018). Start Here: A Gentle Introduction. Retrieved from https://www.kaggle.com/code/willkoehrsen/start-here-a-gentle-introduction/notebook*

[7] *Namvar, A., Siami, M., Rabhi, F. and Naderpour, M. (2018) Credit risk prediction in an imbalanced social lending environment. arXiv preprint arXiv:1805.00801*

[8] *Tang, L., Cai, F. and Ouyang, Y. (2019) Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China. Technological Forecasting and Social Change, 144, 563-572.*

[9] *Jiang, H., Ching, W.K., Yiu, K.F.C. and Qiu, Y. (2018) Stationary Mahalanobis kernel SVM for credit risk evaluation. Applied Soft Computing, 71, 407-417.*

DOI: 10.54254/2754-1169/85/20240868

[10] *Abedin, M.Z., Guotai, C., Hajek, P. and Zhang, T. (2023) Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. Complex & Intelligent Systems, 9(4), 3559-3579.*
[11] *Sammut, C. and Webb, G.I. (2011) Encyclopedia of machine learning. Springer Science & Business Media.*
[12] *Deng, X., Liu, Q., Deng, Y. and Mahadevan, S. (2016) An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences, 340, 250-261.*