Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/46/20230356

A Novel Ensemble Machine Learning Model for Credit Risk Prediction

Songtao Chen[1,a,*]

[1]*Financial Mathematics, University of International Business and Economics, Xiaoguan street, Beijing, China*
a. 201957017@uibe.edu.cn
**corresponding author*

Abstract: Due to the increasing uncertainty of credit risk in today's society, assessing the size of credit risk has become an indispensable part of the lender, and whether it can accurately assess the size of credit risk has become extremely important, which is directly related to the benefit loss of the lender. In light of this, numerous machine learning models have been employed to enhance the prediction accuracy and robustness of credit risk assessments. This paper proposes a hybrid model to better improve the prediction accuracy and robustness of the model. Firstly, this paper collected a data set about credit risk from kaggle, which has 11 independent variables and 1 dependent variable. The dependent variable is a binary variable, representing default or not. Then the data set is cleaned and sorted, and the data set is divided into training set and test set. Then seven machine learning models were used to fit and predict the data, and the three models with the best fitting effect were found through the two indexes of Aread Under Curve (AUC) and Accuracy: Random Forest, Gradient Boosting and Categorical Naive Bayes, and then mix these three models to obtain a mixed model. The experimental results show that compared with the seven machine learning models, the hybrid model has improved in AUC and still ranks first in Accuracy. Therefore, the hybrid model can well improve the accuracy of predicting credit risk and the robustness of the model.

Keywords: machine learning, ensemble model, credit risk prediction

1. Introduction

Credit loans constitute an integral component of a banking institution, playing a pivotal role in evaluating the credit of the borrower, which largely determines whether the funds lent by the banking institution can be recovered in full and timely. However, in the current banking system, the credit assessment of borrowers primarily relies on the documentation provided by customers. In practice, numerous banks lack an effective and complete credit risk assessment system, resulting in the failure to recover a large amount of loans and cause serious losses. Therefore, the establishment of a robust credit risk assessment system assumes paramount importance, as it enables the accurate evaluation of borrower credibility, mitigation of default risk, and minimization of losses stemming from defaults.

As early as the 1950s, some transnational banks in Europe and the United States have initiated the utilization of expert analysis method and CART risk analysis method to appraise the loan credit risks [1, 2]. The underlying principle involves a bank’s credit risk assessment expert team assigning scores to loan applicants' repayment capacity, upon which loan approval decisions are based. However, this

DOI: 10.54254/2754-1169/46/20230356

method depends on the personal experience and emotion of the appraisal experts and is easily affected by the subjective factors of the appraisal experts, with randomness and uncertainty.

Machine learning is a branch of artificial intelligence that aims to enable computer systems to automatically learn and improve from data through computer algorithms and make predictions and decisions based on learned patterns and laws, without having to be explicitly programmed. The main characteristics of machine learning are as follows: machine learning algorithms can automatically learn and extract features from data, adapt and improve according to new data and situations, so as to continuously improve performance and accuracy, and can use the learned patterns and rules to predict and classify new data. The advantages of machine learning include processing complex problems, automating decision making, adaptability and generalization capabilities, and processing large-scale data.

For instance, the application of machine learning algorithms was also considered in the credit evaluation. Wiginton applied the logistic regression method to the credit evaluation of loans and got a good result [3]. Additionally, Makowski proposed a loan credit risk prediction model based on decision tree, which further improved the prediction accuracy [4]. In this period, the prediction models are grounded in rigorous mathematical theories. Although they have assisted banks in mitigating loan risks and exhibited a degree of prediction accuracy, their prediction accuracy still has a large room for improvement. Additionally, Blanco et al. built a small loan credit risk assessment model based on shallow neural networks [5].

In order to reduce the instability of data fitting prediction by a single machine learning model, this paper will use 7 machine learning models for fitting prediction based on the machine learning algorithm. The results of performance evaluation determine three models, which are combined to build a hybrid model.

2. Method

2.1. Dataset Description and Preprocessing

This study used the dataset collected from [6]. There is a total of 612 lines, 11dependent and independent variables and one independent variable. The independent variables include the Dependents, Education etc. The subtype variables are Gender, Married etc. The numeric variables are Applicant_Income, Coapplicant_Income, Loan_Amount, Term. The dependent variable is Status, which is a type of variable with two states Y and N, where Y means good reputation and N means bad reputation. For the data processing process, by observing the number of missing values of each independent variable, it can be found that most of the missing values are distributed in typed variables, so this paper directly deletes the whole row of data corresponding to missing values. This is followed by the re-encoding of all typed variables via the LabelEncoder function. Following sorting operations, data consisting of 499 rows were obtained. The dataset is then resampled by using the SMOTE algorithm to balance out the difference in sample size between the different categories in the dataset. Then the MinMaxScaler object is used to normalize the feature matrix X, and the feature values are scaled to a fixed range to eliminate the scale difference between the features. Finally, the dataset is partitioned into training and test sets, with an 8:2 ratio being adopted for this division.

2.2. Machine Learning Models

2.2.1. Model Introduction

This paper applies 7 machine learning models, K Neighbors, Categorical Naive Bayes, Gaussian Naive Bayes, Decision Tree, Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines (SVM). More details about these algorithms can be found as follows: 1) K-Neighbors:

DOI: 10.54254/2754-1169/46/20230356

Classifies K neighbors based on an instance. 2) Categorical Naive Bayes: A naive Bayes classification algorithm dealing with discrete features. 3) Decision Tree: A classification algorithm based on the tree structure is used to classify features step by step. 4) Logistic Regression: Linear model algorithm for binary classification problems. 5) Random Forest: An ensemble learning method consisting of multiple decision trees that are classified by voting or averaging. 6) Gradient Boosting: ensemble learning method that corrects by iteratively training a series of weak classifiers to get the final classification result. 7): Support Vector Machines (SVM): binary classification algorithm, by finding the optimal hyperplane classification, has good generalization ability. The above models are selected by hyperparameters, and the optimal hyperparameters shown in Table 1 are obtained for model fitting and prediction.

Table 1: The hyperparameters setting used in this study.

	Model	Hyperparameter
1	K Neighbors	n_neighbors = 12
2	Categorical Naive Bayes	alpha = 0.1 fit_prior = True
3	Decision Tree	max_depth = 10 max_features = 'sqrt' min_samples_leaf = 1 min_samples_split = 10
4	Logistic Regression	C = 10.0 Penalty = 'l2' solver = 'liblinear'
5	Random Forest	max_depth = None max_features = 'sqrt' min_samples_leaf = 4 min_samples_split = 2 n_estimators = 100
6	Gradient Boosting	learning_rate = 0.1 max_depth = 5 max_features = 'sqrt' min_samples_leaf = 3 min_samples_split = 2 n_estimators = 200
7	Support Vector Machines（SVM）	C = 10 gamma = 'auto' kernel = 'rbf' probability = True

2.2.2. Ensemble Model

To further improve the model predictive performance, these three models called gradient Boosting, Random Forest and Categorical Naïve Bayes are combined to form an integrated or hybrid model as shown in Figure 1 due to their satisfactory accuracy. This amalgamated approach enhances the modeling capabilities and interpretability by incorporating logistic regression, which establishes a relationship between the predicted probabilities generated by the three models and the corresponding true values. By leveraging this hybrid framework, a more robust and comprehensive modeling approach is achieved, allowing for improved predictive accuracy and interpretability.

DOI: 10.54254/2754-1169/46/20230356

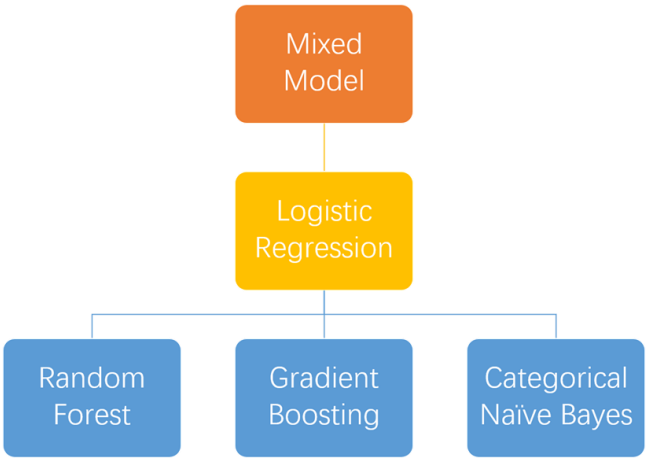

Figure 1: The schematic diagram of the ensemble model (Photo/Picture credit: Original).

3. Results and Discussion

Based on an analysis of the AUC value and accuracy, it is found that Gradient Boosting, Random Forest and Categorical Naive Bayes have the best performance compared with other models shown in Table 2. Therefore, these three models are chosen to be integrated into a mixed model. The hybrid model enhances the modeling and prediction power between the predicted probabilities of the three models and the true values.

Table 2: The performance of various models based on the accuracy and AUC metrics.

	Model	Accuracy	AUC
1	K Neighbors	0.7445	0.7955
2	Categorical Naive Bayes	0.8175	0.8619
3	Decision Tree	0.7737	0.7713
4	Logistic Regression	0.7591	0.8288
5	Random Forest	0.8248	0.9015
6	Gradient Boosting	0.8321	0.8940
7	Support Vector Machines（SVM）	0.7518	0.8241
8	Ensemble Model	0.8321	0.9082

It can be seen from the table that the top three in Accuracy and AUC are Gradient Boosting, Random Forest and Categorical Naive Bayes. The successful application of these models can be attributed to their common advantages, namely 1) 1) Robustness: These models perform well when dealing with noise and outliers and have a certain robustness. They are able to learn from data and adapt to different patterns and noise. 2) Dealing with nonlinear relationships: These models are able to capture and model nonlinear relationships between input features. They can build more complex models by combining multiple decision rules or basic models to better fit the data. 3) Processing high-dimensional features: These models can efficiently process data sets with a large number of features. They are able to automatically select important features and are not susceptible to dimensional disasters. 4) Do not need too much data preprocessing: compared to some other models (such as neural networks), these models have relatively low requirements for data preprocessing. They are often able to handle missing values and incomplete data without excessive data cleaning and preprocessing steps.

DOI: 10.54254/2754-1169/46/20230356

In this paper, these three models are synthesized into a hybrid model, whose Accuracy and AUC are the first among all models. The reasons for its excellent performance are as follows: 1) The advantages of combining multiple models: Hybrid models combine different types of models or algorithms, using the advantages of each model to make up for the shortcomings of other models. By combining the predictions of multiple models, the overall performance and accuracy can be improved. 2) Dealing with complex data distributions: Hybrid models can use multiple sub-models, each of which is modeled for a different data distribution to better adapt to complex data situations. 3) Model generalization ability improvement: hybrid models can better deal with overfitting and underfitting problems of models by integrating different models. The combination of different models can improve the generalization ability of the model, making it perform better on new and unseen data. 4) Strengthen decision-making ability: Hybrid models usually involve the combination of models and integrated learning techniques, and more accurate and reliable results can be obtained by synthesizing the decisions of multiple models. This strengthening of decision-making ability can improve the model's performance in complex tasks. In the future, neural network models can be also considered in the ensemble model for possible improvement of the prediction performance due to their excellent capabilities in other tasks [7-10].

4. Conclusion

This study focuses on the development of seven distinct machine learning models for the purpose of fitting credit risk data. Through rigorous evaluation, three models exhibiting the highest levels of Accuracy and AUC are identified as potential candidates for constructing a mixed model. This amalgamation of models holds the potential to enhance both the robustness of the model and the accuracy of credit risk predictions. Empirical findings indicate noteworthy improvements in both accuracy and AUC metrics. Future research endeavors aim to refine the mixed model by incorporating weights assigned to each selected model or introducing additional models, with the objective of further enhancing the model's robustness, as well as evaluating the potential for improved fitting and predictive capabilities.

References

[1] *Ramos, L., Novo, J., Rouco, J., et al. (2018) Multi-expert analysis and validation of objective vascular tortuosity measurements. Procedia Computer Science, 126:482-489*

[2] *Wu, W., Zhang, Y., Li, Z., et al. (2019) Remote sensing land cover classification based on hierarchical classification of iterative CART algorithm. Remote Sensing technology and application, v.34;No.165(01):70-80.*

[3] *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):757-770.*

[4] *Makowski, P. (1985) Credit Scoring Branches Out. JCredit World. 1985(75):30-37.*

[5] *Blancoa, M. R., Lara, J. et al. (2013) Credit scoring models for the microfinance industry using neural networks: evidence from Peru. Expert Systems with Applications, 40(1): 356-364.*

[6] *Kaggle (2023) https://www.kaggle.com/datasets/mirzahasnine/loan-data-set?datasetId=2991957&select=loan_train.csv.*

[7] *Kaastra, I., Boyd, M. (1996) Designing a neural network for forecasting financial and economic time series. Neurocomputing, 10(3): 215-236.*

[8] *Yu, Q., Chang, C. S., Yan, J. L., et al. (2019) Semantic segmentation of intracranial hemorrhages in head CT scans, 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 112-115.*

[9] *Al-Shayea, Q. K. (2011) Artificial neural networks in medical diagnosis. International Journal of Computer Science Issues, 8(2): 150-154.*

[10] *Patel, J. L., Goyal, R. K. (2007) Applications of artificial neural networks in medical science. Current clinical pharmacology, 2(3): 217-226.*