Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230112

Credit default prediction based on blending learning model

Yaoxi Li[1, †], Yuxuan Tian[2, 4, †], Jianan Zhuo[3, †]

[1]School of Arts & Sciences, Bellevue College, Bellevue, WA, 98007, USA
[2]School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei, 066004, China
[3]Dulwich College Suzhou, Suzhou, Jiangsu, 215021, China

[4]202113155@stu.neu.edu.cn
[†]These authors contributed equally.

Abstract. Currently, the preventions of credit default usually will be evaluated by user's credit value before loaning from banks. However, for the loan user, who have no existing record of loaning and the situation of low credit value, it cannot precisely recognize the risk of credit default. After a credit default, the bank not only doesn't get the signed compensation and principal in time, but also the debtor needs to bear the expensive corresponding late fees and credit costs. Therefore, reducing credit defaults can decline more burden of debtors and creditors. In this paper, the authors evaluate multiple machine learning models including algorithms belong to machine learning and deep learning, using blending model to boost the prediction effect and accuracy, while proposing an optimization design to further enhance the stability, accuracy and generalization capacity of proposed algorithm, so as to effectively decrease the credit default rate and the risk of bank loss in practice.

Keywords: credit default, machine learning, deep learning, blending model.

1. Introduction

Compared to other loaning methods, peer-to-peer (P2P) lending allows borrowing and investing at a lower cost and in a more convenient process [1]. With P2P lending platforms being blocked by the state, low-risk lending options in the country are gradually decreasing, so the bank loans become one of the best loaning options once again. Due to the high default rate of bank loans and the inability to identify credit risks, banks generate huge losses. Because of the dominance of credit risk in bank threats [2], banks mostly use credit values to make predictions of credit defaults [3,4] at present. But in fact, it is difficult to obtain high accuracy in predictions based on credit values alone. In order to protect the interests belonging to both banks and investors, a system capable of conducting credit default prediction is greatly important [5]. In recent years, as the rapid growing of data science techniques, artificial intelligence algorithms have been widely used in various fields. Building special models through both machine and deep learning techniques, such as the combination of the plain Bayesian model [6] and deep neural network, provides new ideas and possibilities for predicting whether a credit is in default or not. The paper [7] builds and trains models on credit data from an improved extreme gradient boosting tree i.e. XGBoost after adding adaptive optimization algorithms to provide highly accurate credit default prediction, the paper [8] selects a variety of classical

DOI: 10.54254/2755-2721/8/20230112

algorithms, including random forests, gradient boosting trees, and other optimization boosting algorithms, such as extreme gradient enhancement algorithm and light gradient boosting algorithm using a special fusion framework stacking to build a model for credit default prediction and obtain a high-performance prediction model demonstrating the superior performance of fusion algorithms in credit prediction.

All the aforementioned methods achieve promising performances in credit forecasting, but are based on small datasets for simulation and training. And especially for stacking model [9] has its own internal cross-validation used internally to generate data to make the dataset as balanced as possible while using training data in depth, but applied to realistic situations when faced with large amounts of data, it consumes huge time and arithmetic costs and can lead to overfitting of the model due to the complexity of the algorithm and data. Therefore, choosing the blending model, which is improved from Stacking model, utilizes the variability of different machine learning models for predicting multiple variations of credit default data to polymorphically assign weights of each model, so as to result in a multi-dimensional integrated credit default prediction.

Comprehensive analysis of the above, this paper uses hybrid model based on ensemble learning method as the main idea to build a prediction model for credit prediction, including SVM, Bernoulli NB, Decision Tree [10], Logistic Regression [11], Random Forest [12] and Gradient Boosting [13]. Regarding a total of six machine learning models as level 0 layer—the base learners and deep neural network (DNN) as level 1 layer—the meta-learner to build the blending model for experiments, and use them to throw out improvements.

2. Method

2.1. Dataset

This data used in this paper is South German Credit Data Set collected by Beuth University of Applied Science Berlin, consisting of 20 predictor variables, most of which are the personal information of the debtors and the information the credit contract, such as the debtor's deposits, the debtor's marital status, etc. The data set investigated the credit records from 1973 to 1975, including 700 good and 300 bad credit loans. Table 1 shows part of the data set.

Table 1. Variables and their corresponding descriptions.

Column name	Content	Variable type	Sample
laufkont	debtor's checking account status	discrete	[1, 1, 2, 4]
laufzeit	monthly credit duration	quantitative	[18, 9, 12, 12]
verw	motivation behind the credit is needed	quantitative	[2, 0, 9, 0]
sparkont	savings of debtor	discrete	[1, 1, 2, 1]
beszeit	debtor's current employment duration	ordinal; discretized quantitative	[2, 3, 4, 3]
alter	age in years	quantitative	[21, 36, 23, 29]
wohn	type of housing the debtor lives in	discrete	[1, 1, 1, 1, 2]
beruf	superiority of debtor's job	Ordinal	[3, 3, 2, 2]
pers	number of persons who financially supported by the debtor	binary, discretized quantitative	[2, 1, 2, 1]
gastarb	the debtor is a foreign worker or not?	binary	[2, 2, 2, 1]

2.2. Data processing

In terms of data pre-processing, the data, which mainly contains quantitative and categorical type of data, were analyzed and processed by adopting linear correlation test, variance inflation factor(VIF) and significance test for credit dataset, and visualized by inter-variable heat map, as shown in Figure 1. Two sets of variables with strong linear relationships can be directly derived from the heat map of linear correlation coefficients, which are laufzeit and hoehe, moral and bishkred. Soon afterwards, in

DOI: 10.54254/2755-2721/8/20230112

order to avoid the impact of multiple cointegration caused by strong linear relationships on the subsequent model training and follow-up fitting process, the variance inflation factors (VIF) are calculated for them, which are 1.64076866 and 1.23681419 respectively. On account for their VIFs are all less than 5, there is no impact on the following process and it's unnecessary to delete the two sets of variables. According to the heat map, it can be easily found that some variables have a strong correlation with credit default, which are status, duration, credit_history, amount, savings, and property, assisting in the follow-up making recommendations and conclusions.

Figure 1. Correlation analysis of variables.

Finally, significance test (Mann-Whitney U test) was conducted, and the data table was shown in Table 2. From the significance test results, it is known that all variables for credit default have correlation, instead of chance and randomness. Therefore, the dataset is not deleted.

Table 2. Pearson product-moment correlation coefficient of variables.

Variable name	Pearson correlation coefficient	Variable name	Pearson correlation coefficient
laufkont	3.178e-246	wohnzeit	3.180e-296
laufzeit	0.0	verm	1.151e-242
moral	2.659e-286	alter	0.0
verw	2.571e-102	weitkred	0.0
hoehe	0.0	wohn	3.471e-290
sparkont	3.401e-153	bishkred	4.379e-147
beszeit	0.0	beruf	0.0
rate	7.108e-295	pers	1.175e-308
famges	0.0	telef	1.239e-156
buerge	5.693e-93	gastarb	0.0

DOI: 10.54254/2755-2721/8/20230112

2.3. Blending model

Blending model is optimized stacking model, which reduces the internal cross-validation of the Stacking model, and reduces the number of operations and lowers the high requirements of computational power. At the same time, it also inherits the basic structure of stacking: using two-level algorithm in series, there are multiple base learners on level 0, and there is only one meta-learner on level 1. It also has the advantage of learning rules trained by machine learning models from multiple dimensions. The base learner on level 0 is trained conventionally - fitting the relationship between the data and the real label, outputting the prediction results, and forming a new feature matrix, and then the meta-learner on level 1 is trained and predicted on the new feature matrix.

The main idea of blending is to separate data into two categories: training and verification set. In the first round of training, use the training set data to train multiple models, and then use the verification set to make predictions. The flow chart in Figure 2 shows the structure of the model. The prediction tag obtained is used for the second round of continuous training. The specific training process is as follows: First, separating data to training, verification and test sets, then input the training set for each base learner at level 0 for training, and then input the verification set to verify the generalization capacity of the model so that it can output the prediction for the verification data (prediction results are probability values). After that, the predictions of all base learners in level 0 for the verification set are spliced into a prediction result matrix of the probability of each column corresponding to the classification category to form a new feature matrix. Finally, the feature matrix and test set label are input into the meta-learning machine (deep neural network DNN) for supervised training and learning, and the super-parameter learning is carried out with the ability of the deep neural network to fit any function in theory, and the final classification model is obtained.

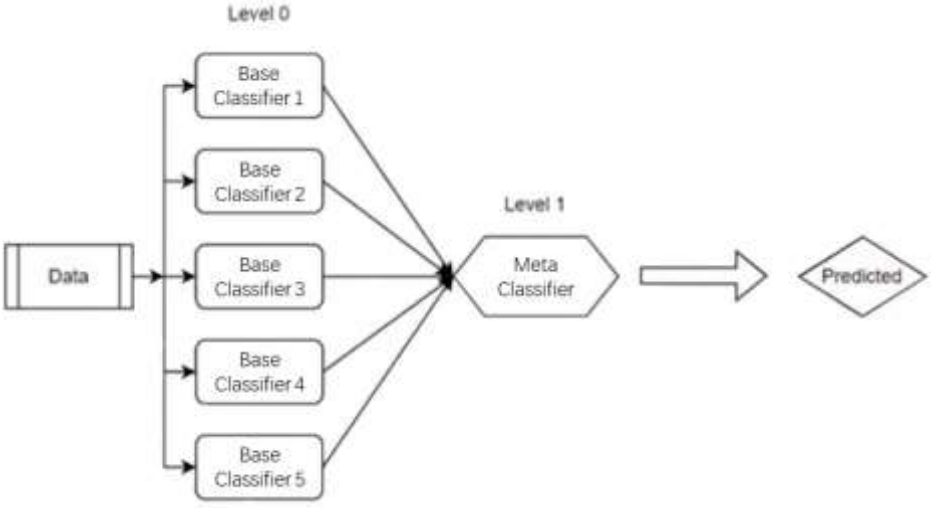

Figure 2. Flow chart of blending model.

2.4. Decision tree

The decision tree model is divided into classification tree and regression tree. The two tree structures are based on the simple logic tree principle - if-then, as shown in Figure 3. The difference is that for classification tree, the last leaf node represents the label to be classified, while each node of the regression tree is the logical attribute to be returned.

DOI: 10.54254/2755-2721/8/20230112

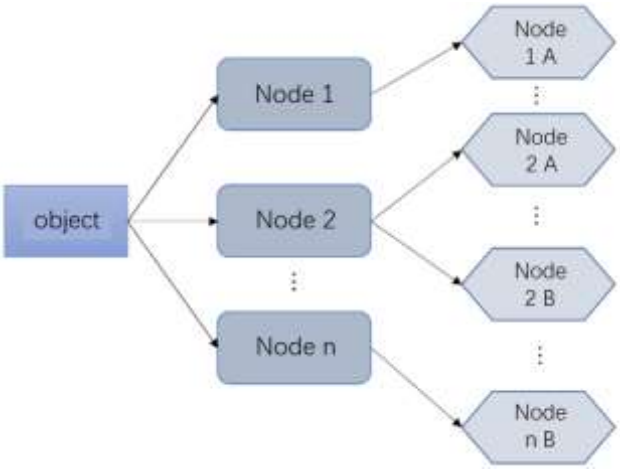

Figure 3. Diagram of a basic decision tree.

The decision tree is constructed by a set of rules about if-then. Each root-to-leaf path is constructed as a special rule with mutually exclusive and complete properties. However, due to the uncertainty of the complexity of the polynomial of the decision tree, it cannot be directly solved. However, because the decision tree is based on the natural structure of the classification tree and the regression tree, it can be explained easily, and the method of verifying the solution is usually used to approximate the solution.

2.5. Logistic regression

It is a statistical analysis method generally used to make two-classification or multi-classification prediction results for data sets. Logical regression model is a machine learning model built by analyzing the relationship between existing independent variables by constructing polynomials, and obtaining dependent variables (probability of occurrence of binary events) to deal with classification problems.

$$g(y) = \frac{1}{1+e^{-1}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\cdots+\theta_n x_n)}} = \frac{1}{1+e^{-\theta^T x}} \tag{1}$$

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \theta^T x \tag{2}$$

The function of the logistic regression model is derivable at any order. It can use many optimization algorithms to help solve the optimal solution. It has good mathematical properties, good interpretability and training speed. For example, the coefficients of independent variables represent the weights of different characteristics. However, the premise of the accuracy of the logistic regression model is that there must exist strong linear relationship among the dependent and independent variables and a high requirement for the sample size.

3. Result

To assess the strength of the blending model, the authors used Accuracy, Area Under the ROC Curve (AUC), and Model Distinction (KS) to assess the effectiveness of the model. This work also assesses the performances with the Recall rate and F1 Score.

According to Figure 4, the blending model performs superior than the other six machine learning models which are employed by their own base learners. This indicates that the blending model successfully incorporates the rules and functions learned by the machine learning models in multiple dimensions. The blending model also gains knowledge of multidimensional features and achieves a higher accuracy for both the strong and weak models. Furthermore, the prediction for 20 different dimensions of variable information reaches a high standard.

DOI: 10.54254/2755-2721/8/20230112

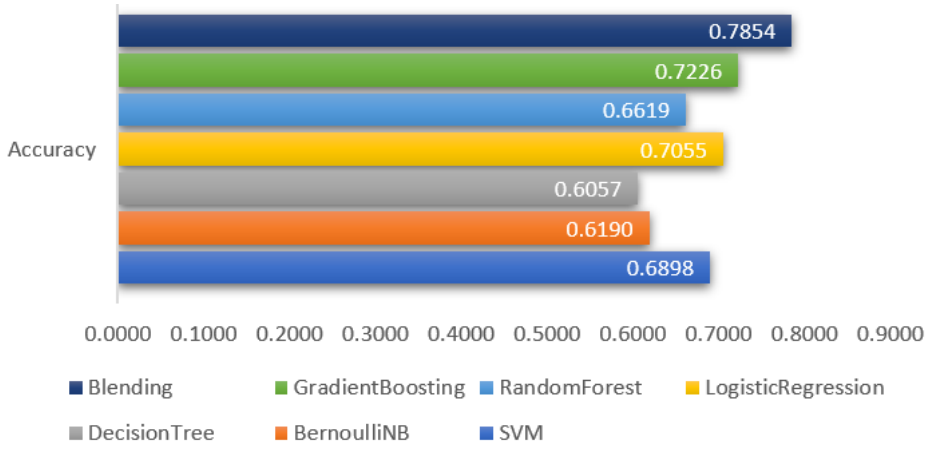

Figure 4. Bar chart for accuracy of each model.

From the AUC/KS index histogram in Figure 5, it can observe that blending ranks the highest in the AUC index and ranks second in the KS index. This indicates that the blending model has strong abilities in sample classification sorting and risk sorting. The blending model also seems to enjoy a certain stability.

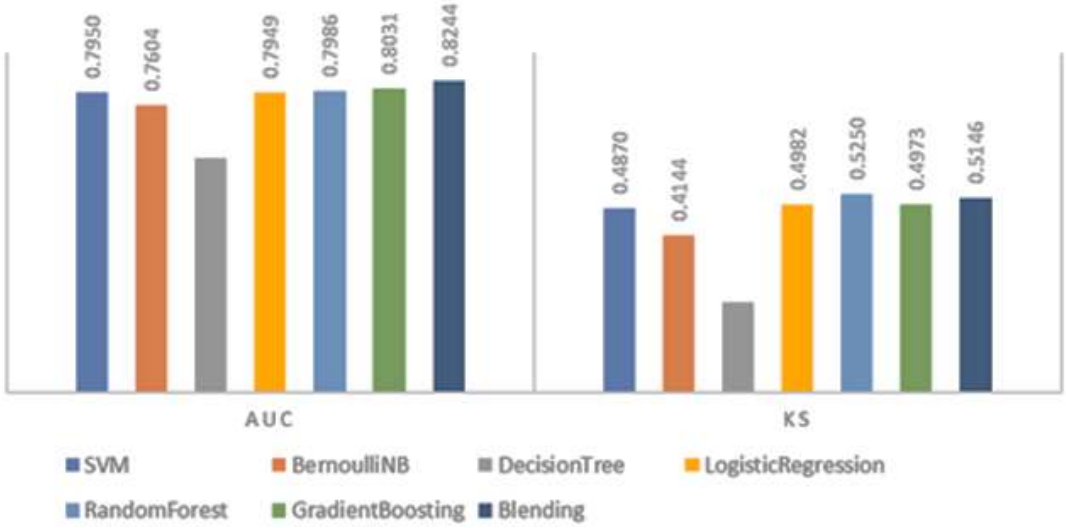

Figure 5. AUC/KS indexes.

Table 3. Classification results of blending algorithm.

	precision	recall	f1-score	support
0	0.9092	0.8235	0.8642	221
1	0.6616	0.6835	0.6723	79

The list of model evaluation reports based on the blending model is shown in Table 3. The blending model has a higher recall for confirmed credit defaults and a lower recall for confirmed non-credit defaults, which is caused by the small sample size in dataset. This results in the insufficient learning of some models and lower prediction performance, while the F1 score is in the same situation. However, the minimum value is higher than 67%, which indicates that the quality of the blending model is relatively high, but also limited by the factor of imbalance condition existing in the dataset itself.

DOI: 10.54254/2755-2721/8/20230112

4. Discussion

The machine learning models selected as the base learners for blending Model are most weak models for this dataset, but it appears a higher performance in blending without weighting. In general, when the prediction performance of the base learners in level 0 is relatively similar, it is appropriate to use uniform blending, i.e., setting the same or no weights corresponding to each algorithmic model. While when the prediction performance of the base learners is relatively different, researchers can use linear blending by adding a different weight for each different machine learning model through a weighted average method to make better prediction results. In addition to weighting, voting methods or other algorithms can be tried to participate in the computation of fusion. Meanwhile, SSA (Sparrow Search Algorithm), POS (Particle Swarm Algorithm), and GA (Genetic Algorithm), for example, can be added to automatically adjust the parameters and assist in the computation of machine learning models that are partially used as base learners to improve the stability and prediction ability of the models for data anomalies.

5. Conclusion

After six machine learning models are fused into a new hybrid model through the blending algorithm, the accuracy and F1 Score are significantly improved, while the prediction for 20 different dimensions of indicators reaches a high level. Therefore, the blending model can use the collected multidimensional big data with algorithms to ensure the security of loans, evaluate the credit of customers who need loans, and help banks identify new groups of customers who may experience credit defaults, further avoiding major financial losses and reducing the risk of lending defaults. Therefore, the authors believe it is advisable to focus on scoring and predicting the data information of credit users when they take out a credit loan in terms of their account status, the duration of loan length, loan history, savings, gender, and property (customers' most valuable possessions).

References

[1] Ding, J., Huang, J., Li, Y., & Meng, M. (2019). Is there an effective reputation mechanism in peer-to-peer lending? Evidence from China. Finance Research Letters, 30, 208-215.

[2] Caruso, G., Gattone, S. A., Fortuna, F., & Di Battista, T. (2021). Cluster Analysis for mixed data: An application to credit risk evaluation. Socio-Economic Planning Sciences, 73, 100850.

[3] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.

[4] Soui, M., Gasmi, I., Smiti, S., & Ghédira, K. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. Expert systems with applications, 126, 144-157.

[5] Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 425, 76-91.

[6] Catal, C., Sevim, U., & Diri, B. (2011). Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm. Expert Systems with Applications, 38(3), 2347-2353.

[7] Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble learning or deep learning? Application to default risk analysis. Journal of Risk and Financial Management, 11(1), 12.

[8] Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. Expert Systems with Applications, 161, 113567.

[9] Vlamis, P. (2007). Default risk of the UK real estate companies: is there a macro-economy effect?. The Journal of Economic Asymmetries, 4(2), 99-117.

DOI: 10.54254/2755-2721/8/20230112

[10] Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39, 261-283.
[11] Gai, K., Zhu, X., Li, H., Liu, K., & Wang, Z. (2017). Learning piece-wise linear models from large scale data for ad click prediction. arXiv preprint arXiv:1704.05194.
[12] Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9(9), 1-9.
[13] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.