Auto loan default prediction based on Stacking model

LingLing Zeng[1a], Jin Sun*[2b], YiMin Zhou[3c]

[1.] School of Economics, Wuhan University of Technology, Hubei Wuhan, China.
[2.] School of Economics, Wuhan University of Technology, Hubei Wuhan, China.
[3.] School of Data Science, Southwestern University of Finance and Economics, Sichuan Chengdu, China.

[a]E-mail:2955926860@qq.com
[b]Corresponding author. E-mail:2944113158@qq.com
[c] E-mail:844277310@qq.com

Abstract. With the rapid development of auto loan and serious credit problems exposed in the industry, auto loan default situation needs to be improved. Based on the customer data provided by a car loan platform, the paper studies the problem of vehicle loan default prediction. Firstly, the data is preprocessed, and then the processed data is used to build logistic regression, random forest, XGBoost and LightGBM models to predict whether the vehicle loan defaults, respectively. Secondly, the above single models are fused in turn and forecasted again. The results show that compared with a single model, the Stacking model has higher accuracy. Comparing the forecasting effects of the fusion models, it was found that the forecasting effect of LR-Stacking model was the best, with the ac-curacy of 83.62%.

Keywords: auto loan default prediction; model fusion; LR-Stacking model

1 Introduction

The small and scattered auto loan business has great development prospects due to its advantages of low entry threshold, low borrowing amount, high liquidity and short term, but due to the lack of risk prevention work, the non-performing loan rate of the industry remains high.

At present, scholars at home and abroad have carried out a lot of research on the application of machine learning in credit. Liu etal.（2016） established models such as decision tree, logistic regression, random forest and BP neural network to predict the loan default of automobile credit data of commercial banks, and found that the logistic regression model performed best. [1] Shu etal.（2017） used binary selection models to predict auto loan default, and the results showed that the Logistic model was more suitable for predicting whether customers would default. [2] In addition, the continuous advancement of technology has spawned new ideas, some scholars have improved the accuracy of predictions by ensemble models.[3~5] Chen Yaofei et al.(2017) compared

Y. Jiao et al. (eds.), *Proceedings of the 3rd International Conference on Internet Finance and Digital Economy (ICIFDE 2023)*, Atlantis Highlights in Economics, Business and Management 1,
https://doi.org/10.2991/978-94-6463-270-5_31

logistic regression with GBDT and other algorithms and found that the XGBoost algorithm greatly improved the prediction accuracy and model training speed.[6].Zhou etal. (2020) used XGBoost algorithm combined with random forest to construct an XGBoost-RF model to evaluate personal credit risk, and the results show that the improved random forest model has more advantages in prediction effect. [7]

Based on the above research, this paper uses the stacking fusion model to study the problem of vehicle loan default prediction, aiming to better use intelligent methods to identify credit risks, improve the audit efficiency and default prediction accuracy of loan users, reduce the credit risk of the platform, and promote the healthy and stable development of the platform.

2 Data sourcing and preprocessing

2.1 Data sources

The experimental data in this paper is derived from the loan records of a credit platform provided in the competition question on Ali Tianchi. The total amount of data exceeded 250,000, including 52 feature fields, from which 150,000 pieces of data known to be in breach of contract were extracted for experimental research.

2.2 Data preprocessing

2.2.1 Construction features.

First, loan risk tends to be positively correlated with return, and loan interest rate is an important factor in vehicle loan default prediction. Therefore, for the accuracy of the forecast results, the annual interest rate of the master and secondary accounts is calculated based on the existing characteristics. In addition, accounts that commit loan fraud may have clustered registrations during the same period, and the number of defaulting accounts will be continuous. Therefore, the neighbor fraud characteristics are constructed by averaging the target variable (loan_default) of the ten accounts before and after each account.

2.2.2 Data binning.

According to the characteristics of the data, different types of binning operations are performed on the feature data.the loan-to-asset ratio (loan_to_asset_ratio) is discretized by the isowidth bins to increase its nonlinear expression ability, the asset cost (asset_cost) characteristics are divided into 10 bins with equal frequencies. In addition, in order to effectively retain the information of outliers and missing values, the eight characteristic data with many outliers and missing values are customized binning, and the outliers and missing values are divided into separate 10 bins for processing.

2.2.3 Data destination encoding.

Object coding, also known as mean coding, is a way to encode features in combination with target values. In order to more directly understand the impact of some categorical variables on vehicle loan default, seven categorical characteristics such as job type (employment_type) and branch (branch_id) that issued the loan were objectively encoded. To prevent overfitting, choose a ten-fold target code that adds regularization to the mean encoding.

2.2.4 Feature post-processing.

This paper performs post-processing of features. First, delete the data column with more than 60% redundancy; Second, delete unnecessary data, such as: whether to fill in the mobile phone number, whether to fill in the ID card, whether to issue a driver's license, whether to fill in the passport and other information. The final contains 46 features.

3 Model building

3.1 Stacking ensemble learning algorithms

Stacking is a heterogeneous ensemble algorithm that fuses different types of learners. The algorithm can be divided into two layers. The first layer is the base model training layer, which contains multiple prediction models; the second layer is the meta-model training layer, which usually chooses a relatively simple prediction model. For the base model training layer, by selecting multiple prediction models to perform K-fold cross-validation on the training set of the original data, the predicted data will be spliced to form a new training set; similarly, the test of multiple prediction models on the original data The set is predicted, and then the predicted results are averaged and spliced to form a new test set. For the meta-model training layer, the training set and test set obtained by using the base model training layer are used as the input of the meta-model, and then a prediction model is selected to train it, and the prediction result is output.

3.2 Stacking model building

Use the Borderline SMOTE method to balance the data. Then, the data processed is divided into training set and test set according to the ratio of 7:3, which are used for model training and prediction respectively, and then the training set data is used to construct four single models of vehicle loan default prediction, such as logistic regression, random forest, XGBoost, and LightGBM, and Bayesian optimization algorithms are used to find the optimal parameter combination. Table 1 lists the parameters of each model after parameter tuning:

Table 1. Parameter setting

Model	Parameter
Logistic regression	C：0.7；penalty:L1；solver:liblinear
Random forest	max_depth=11；n_estimators:120
XGBoost	max_depth=9； subsample=0.9； min_child_weight=10； col_sample_bytree=0.85；lambda:10；eta:0.02
LightGBM	learning_rate:0.01；min_data_in_leaf:150；feature_fraction:0.8；bagging_fraction:0.7

In order to find the best combination of model fusion, this paper uses logistic regression, random forest, LightGBM, Xgboost as meta-learners, and the other three models as base learners to construct stacking fusion models.[8~10] Taking the LR-Stacking model as an example, the specific process is shown in Figure1:

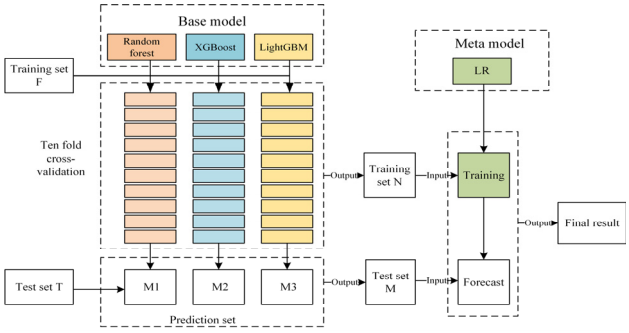

Fig. 1. Framework diagram of the LR-Stacking model

Figure 1 shows the training process of the LR-Stacking model, which uses random forest, XGBoost, LightGBM model as the base model, and logistic regression as the metamodel. First, for the base model, it goes through the following process: taking a random forest as an example, the training set F and the test set T are imported separately, and since K=10, the training set F will be divided into 10 equal parts. Each of the 10 equal parts is used as the validation set in turn, and the remaining 9 copies are used as the training set to train the random forest model, that is, the ten-fold cross-validation. After the training is completed, the prediction set N1 is obtained sequentially. Prediction of the test set T using the trained random forest model then yields the prediction set M1. Next, the XGBoost model and LightGBM model will be trained in the same steps to obtain N2, N3, M2, and M3. Then, the prediction set N1, N2 and N3 are arranged horizontally as the metamodel training set N to train the metamodel logistic regression. M1, M2 and M3 are arranged horizontally to obtain the metamodel test set M, and the test set M is imported into the trained logistic regression to obtain the final prediction result.

4 Discussion and analysis

The implementation process of RF-Stacking model, XGBoost-Stacking model, and LightGBM-Stacking model is similar to the LR-Stacking model, and will not be repeated in this article. Finally, the overall model training results are evaluated, and the evaluation results are as follows:

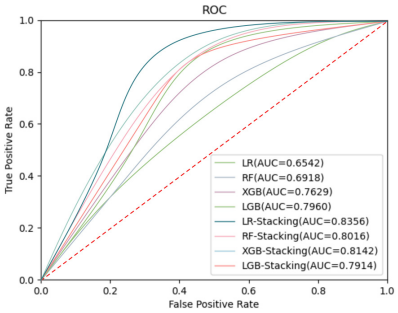

Fig. 2. ROC performance comparison of the eight models

Table 2. Evaluation metrics comparison of the eight models

Model	Accuracy（%）	Precision（%）	Recall（%）	F-score	AUC
Logistic regression	65.39	67.54	59.26	0.6312	0.6542
Random forest	69.08	68.65	70.24	0.6943	0.6918
XGBoost	76.05	72.66	83.52	0.7771	0.7629
LightGBM	79.30	76.48	84.62	0.8035	0.7960
LR-Stacking	83.62	84.22	82.74	0.8347	0.8356
RF-Stacking	80.14	81.33	78.24	0.7976	0.8016
XGBoost-Stacking	81.52	79.83	84.36	0.8203	0.8142
LightGBM-Stacking	79.23	77.95	81.52	0.7969	0.7914

Table 2 compares the evaluation indicators of the prediction results of the single model and the fusion model. Overall, the prediction performance of the fusion model is generally better than that of the single model. In addition, by comparing the evaluation indicators of the fusion model, it is found that the prediction performance of the fusion model is LR-Stacking model, XGBoost-Stacking model, RF-Stacking model and LightGBM-Stacking model from high to low. Among them, the prediction accuracy of the LR-Stacking model is 83.62%, which is the highest among the four models.

Figure 2 shows the ROC curve comparison between the single model and the fusion model, and it can be seen that the LR-Stacking model has the best efficiency, and the AUC value is 0.8356. Among the single models, the LightGBM model has the best performance, with an AUC value of 0.7960. It can be seen that by fusing a single model, the predictive performance of the vehicle loan default prediction model can be effectively improved.

5 Conclusion

Based on the real loan records of a credit platform provided in the competition question on Ali Tianchi, this paper studies the prediction problem of vehicle loan default. Firstly, according to the characteristics of the auto loan industry, the data are analyzed and preprocessed, and the logistic regression, random forest, XGBoost and LightGBM models are constructed to predict whether the vehicle loan defaults. Secondly, in order to further improve the prediction accuracy of the model, the above single model is used as the metamodel in turn, and the remaining three models are used as the base model to construct the Stacking fusion model respectively, and predict whether the car loan default is again predicted, and the comparison of the model evaluation indicators shows that the Stacking model has higher accuracy and stronger robustness, which the LR-Stacking model has the best prediction effect, and the accuracy is as high as 83.62%.

Refrences

1. Liu K. Application of Stochastic Forest and logistic regression model to default prediction[J].China Computer & Communication, 2016, 367(21):111-112. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8
2. Shu Y, Yang Q. Research on Auto Loan Default Prediction Based on Large Sample Data Model[J].Management Review, 2017, 29(09):59-71. DOI:10.14120/j.cnki.cn11-5057/f.2017.09.006.
3. Zhang Liying, Yang Ruojin..The Application Research of Households' Loan Default Prediction Model Based on Machine Learning[J].Financial Regulation Research, 2022, No. 126(06):46-59. DOI:10.13490/j.cnki.frr.2022.06.002.
4. Fabio S, Nicola L. Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities[J]. European Journal of Operational Research, 2023, 305(3):1390-1406. https://doi.org/10.1016/j.ejor.2022.06.035.
5. Meixuan L, Chun Y, Wei L. The network loan risk prediction model based on Convolutional neural network and Stacking fusion model[J]. Applied Soft Computing, 2021, 113(2)171-179. https://doi.org/10.1016/j.asoc.2021.107961.
6. Chen Yaofei, Chen Yijie, Liming.Credit score prediction model based on XGBoost[C]//China Statistical Education Society. Selected winning papers from the 2017 (5th) National University Student Statistical Modeling Competition. 2017:16. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG
7. Zhou Y S, etc. Study on the evaluation of personal credit risk based on the improved random forest model[J]. Credit Reference, 2020, 38(01):28-32. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C4
8. Much A M, Tiara L N, et al. New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning[J]. Intelligent Systems with Applications. 2023, 18(2):123-131. https://doi.org/10.1016/j.iswa.2023.200204.
9. LIU Xiao,ZHOU Rongxi,LI Yuru. Default prediction of credit bonds in China based on Stacking algorithm integration[J].Operations Research and Management,2023,32(03):163-170. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44
10. LI Ruolin, GONG Yishan. Research on loan default prediction model based on model fusion[J].Changjiang Information and Communication,2023,36(03):65-67. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8