A LightGBM Based Default Prediction Method for American Express

[1st] Zhiren Gan[1,a], [1st] Junyuan Qiu[2,b], [2nd] Fuli Li[3,c], [3rd] Qian Liang[4,d]

[a] gzr19970608@gmail.com, [b] michaelqiu2000@gmail.com, [c] lifuli1202@outlook.com, [d] lchris2021@163.com

[1]Shenzhen University, Shenzhen, China,

[2]South China University of Technology, Guangzhou, China,

[3]Shanghai Lixin University of Accounting And Finance, Shanghai, China

[4]Yunnan Normal University, Yunnan, China

Abstract. With the progress of economy and science and technology, the credit card business has developed rapidly in the financial industry because of its convenient and high profits. However, with the sharp increase in the number of credit card users, the problem of credit card violations has become more prominent. If corresponding measures are not taken in a timely manner to control it, it will cause serious losses to banks and other financial institutions. The task of predicting personal default risk can be seen as a binary classification task. In this study, we utilize data provided by the American Express Company to predict default and mitigate the default risk for consumer finance companies using a model called LightGBM. We discuss related work in the second section, while our methodology and experiments are presented in sections III and IV. In order to assess the performance of our experiments, we conduct experiments using different types of models. We also define new experimental metrics. The results indicate that among these models, LightGBM achieved the highest metric of 0.692, surpassing Xgboost, Lasso, and Catboost by 0.007, 0.032, and 0.008 respectively..

Keywords: credit default, feature engineering, LightGBM

1 Introduction

With the progress of economy and science and technology, the credit card business has developed rapidly in the financial industry because of its convenient and high profits, and has been widely loved by financial institutions and users. However, with the sharp increase in the number of credit card users, the problem of credit card violations has become more prominent. If corresponding measures are not taken in a timely manner to control it, it will cause serious losses to banks and other financial institutions. Therefore, it is very necessary to score the credit of users who hold credit cards and predict in advance that users who are in default are of great significance to the long-term development of banks and other financial institutions.

The problem of default prediction is commonly approached as a binary classification task. Our team addresses this problem by leveraging machine learning algorithms, such as constructing a

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334590

credit evaluation model. In this research, we utilize publicly available data from American Express on Kaggle to investigate default prediction and devise models that can accurately identify customers at risk of default, thus mitigating default risk for consumer finance companies. Our default prediction model is built upon LightGBM. Section II offers a comprehensive review of relevant studies, while sections III and IV provide a thorough explanation of our methodology and experimental procedures.

2 Related Work

In the choice of methods for studying models, there are mainly three types. The first is a credit scoring model based on statistical methods, mainly including discriminant analysis, logistic regression, etc. For example, [1] proposes that Fisher discriminant analysis can be used in credit scoring; [2] introduced discriminant analysis into the study of personal credit scoring models, and the study showed that the multivariate linear discriminant analysis model has good predictive ability and robustness.

The second approach primarily involves utilizing machine learning models, with commonly employed methods including SVMs, random forest, XGBoost, neural networks, and Logstie regression [3]. Additionally, K-means clustering, support vector machine, and random forest models were established to predict borrowers' default risk. The findings revealed that the random forest method yielded superior prediction performance compared to other approaches.

[4] compared XGBoost with Logistic regression and GBDT, and the results showed that XGBoost had better prediction effect and short training time; In the study of credit risk control strategy model of Internet consumer finance, [5] compares the XGBoost model with the Logosidian regression model, Bayesian model and SVM model, and confirms that the XGBoost model has more advantages in mining important factors affecting the overdue loans of credit customers.

The third approach involves a combination of existing models, where the predicted probabilities from one model are used as input variables for another model. For instance, [6] proposed a hybrid credit scoring model that incorporates Logistic regression and neural network, demonstrating that the hybrid model outperformed individual models in terms of prediction accuracy and robustness. Similarly, [7] employed support vector machine, random forest, and XGBoost to construct a credit prediction model. Comparative analysis with Logistic regression revealed that the performance of the three individual algorithms surpassed that of Logistic regression. By employing a weighted fusion approach, their combined performance exhibited enhanced resolution, improved prediction accuracy, and suitability for personal credit evaluation in online credit scenarios. Additionally, [8] introduced the Stacking integrated learning algorithm in the credit evaluation system, resulting in improved effectiveness, thus suggesting its practical applicability.

- Our Contribution

a. Our credit default prediction model employs LightGBM.

b. We present our dataset and conduct a thorough analysis.

c. During the experimental process, we conduct comparative experiments which reveal that our model outperforms other models.

3 Methodology

The LightGBM algorithm [9, 10, 11, 12, 13] proposed by Microsoft in 2017, is an improved gradient lifting algorithm based on the GBDT algorithm, which can be applied to classification, regression and sorting problems. Through the improvement and optimization of four major aspects, the LightGBM algorithm solves the problem of traditional gradient enhancement algorithm in massive data, and reduces the complexity of the model. It not only reduces the memory occupation, but also greatly improves the calculation speed and prediction accuracy of the model.

First of all, the improvement of the leaf growth strategy is different from the leaf growth method used by layer splitting in the GBDT algorithm and the XGBoost algorithm, the LightGBM algorithm uses the deep growth method of splitting by leaf nodes, when splitting, according to the information gain formula Gain Calculate the splitting gain, grow the leaf node with the largest information gain obtained from all leaf nodes at present, and sequentially perform to find the optimal tree structure g(x). Compared with the layered division of leaves in GBDT and XGBoost, splitting by leaf nodes can reduce losses and improve the accuracy of prediction results. At the same time, the division of leaf nodes can also avoid overfitting problems by limiting the minimum value of each leaf node and the depth of the tree.

The second is that LightGBM uses the histogram algorithm, through the discretization of continuous data into k features, so that the information containing k groups constitutes a histogram with a width of k, compared with the XGBoost algorithm when splitting, the original data of the indicator is first pre-sorted, the histogram algorithm divides the original data of the indicator into a series of discrete regions, traverses the discrete data, and looks for the optimal division point, through the simplification of the data, reduces the use of memory, and improves the efficiency of model operation. In the process of histogram traversing attributes, the number of operations is reduced because only k-times information gain needs to be calculated, and the division value we find is not necessarily the most accurate, but a large number of experiments have shown that the impact of discretization on the accuracy of the model is limited.

The third is the unilateral gradient sampling algorithm, which uses the information of the gradient size of the sample as a consideration of the importance of the sample, and believes that the smaller the gradient of the sample, the better the model fit and the smaller the error, and adopts a random sampling strategy for such samples and gives them weight compensation. For samples with large gradients, all are retained to improve the attention to samples that are not well trained, improve the recognition accuracy of the model, and greatly reduce the amount of operation of the model, improving the running speed.

The fourth is the mutex feature bundling algorithm: the algorithm is used to solve the feature sparsity problem of high-dimensional samples, in the feature sparsity space, often many features are mutually exclusive, that is, several features will not be non-zero at the same time (such as data obtained by the one-heat encoding), LightGBM algorithm converts these feature features into graph coloring problem processing, and the mutually exclusive features form a weighted

undirected graph according to the relationship between the feature vectors, according to the principle of least overall feature conflict. Assign the features with a medium size to the resulting node to an existing feature pack, or directly form a new feature pack. In this way, the mutex feature bundling algorithm improves the efficiency of the model by having fewer data features.

4 Experiments

- Experiments data

The objective of this research is to employ a customer's monthly profile in order to forecast the probability of them defaulting on their credit card balance in the future. The target variable is a binary classification determined by evaluating the customer's behavior during an 18-month timeframe following their most recent credit card statement. If the customer fails to make the required payment within 120 days of their statement date, it is considered a default event. The dataset comprises aggregated profile characteristics for each customer at each statement date, which have been anonymized, normalized, and categorized into general groups.

The variables in the dataset are categorized as follows:

B_*: Balance variables

R_*: Risk variables

D_*: Delinquency variables

S_*: Spend variables

P_*: Payment variables

- Feature engineering

We do some feature engineering to get a better feature set. For Delinquency, Spend, Payment, Balance and Risk features, we calculate some statistic features like mean, last value and max features to expand the whole feature set. Besides, feature combination is also used to generate more features with the combination of different kind of features like Payment and Date features. To reduce the memory and speed up the training speed, feature selection is necessary. In this work, the features with a correlation coefficient more than 0.98 would be removed. The feature importance histogram is also used for selecting importance features. In Figure 1, it shows the feature and corresponding importance of top 20 most important features. The figure 1 shows that the features like p_2_last, D_48_last and B_2_last are more important compared with other features like B_32_last and B_32_mean. Therefore, we can choose the features with a descending order from the feature importance histogram. The figure 2 shows the distribution of target versus different feature like S_8, S_9.

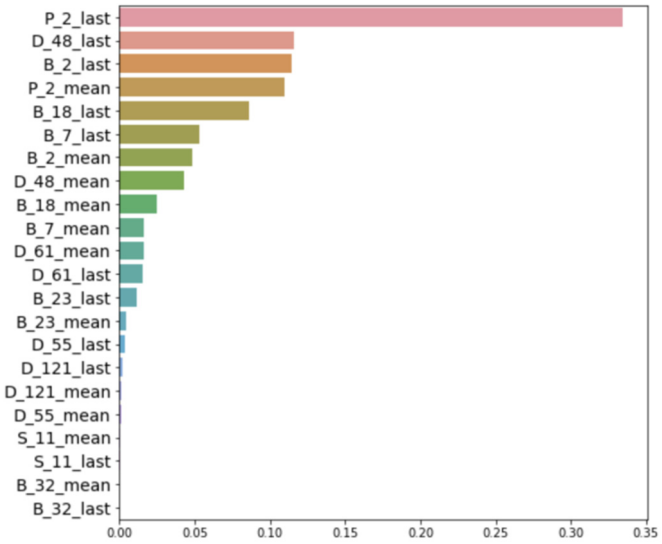

Figure 1: feature importance figure

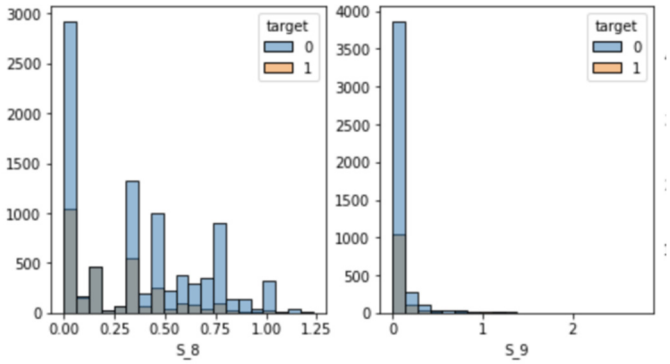

Figure 2: feature distribution

- Training parameters

The LightGBM's parameters are got according to empirical methods and grid search. For example, we will choose the training parameters show in the following table 1.

Table 1: Training parameters

n_estimators	1200
learning_rate	0.03
reg_lambda	50
min_child_samples	2400
num_leaves	95
colsample_bytree	0.19
max_bins	511

- Evaluation metrics

The evaluation metric for this study is denoted by M, which is the average of two rank ordering measures: the Normalized Gini Coefficient G and default rate D captured at 4%.

$$M=0.5\cdot(G+D)$$

The default rate recorded at 4% corresponds to the proportion of positive labels (defaults) identified within the top 4% of the predictions, serving as a measure of Sensitivity/Recall. In order to account for down sampling, a weight of 20 is assigned to the negative labels for both sub-metrics G and D.

- Experiment result

The default rate captured at the 4% threshold indicates the proportion of positive labels (defaults) identified among the highest-ranked 4% of predictions, which serves as a measure of Sensitivity/Recall. In both sub-metrics G and D a weight of 20 is assigned to the negative labels to appropriately adjust for down sampling. The table 2 show the result for each model.

Table 2: Experiment result

Models	Metric
Xgboost	0.794
Lasso	0.769
Catboost	0.793
Lightgbm	0.801

5 Conclusion

Our paper focuses on feature engineering and utilizes the LightGBM model for credit default prediction. Section II provides an overview of related work, while section III details the model employed in our study. In section IV, we present our experimental setup, including specific aspects of feature engineering and the parameter settings for LightGBM. The experimental results demonstrate that our LightGBM model outperforms other models, achieving the highest metric of 0.692. This metric surpasses Xgboost, Lasso, and Catboost by 0.007, 0.032, and 0.008 respectively.

Acknowledgments. We would like to express our gratitude to the individuals who provided guidance and support for this paper. Zhiren Gan and Junyuan Qiu share the same contribution.

References

[1] Napolitano A. Alleviating class imbalance using data sampling: Examning the effects on classification algorithms [D], Department of Computer Science and Engineering, Florida

[2] Van Hulse J, Khoshgoftaar T M, Napolitano A. Experimental perspectives on learning from imbalance data(C]. in Proceedings of the 24th International Conference on Machine; Learning, Corvallis, OR, USA, 2007:935-942

[3] ChawlaNv, BowyerKw, HallL0, et al. SMOTE: synthetic minoriyover-sampling technique [J]. Journal of artificial intelligence research, 2002, 16:321-357.
[4] Han H, Wang W Y, Mao B H. Borderline- -SMOTE: a new over- sampling method in imbalanced data sets learning[C]//International conference on intelligent computing. Springer,Berlin, Heidelberg, 2005: 878- 887.
[5] Rayhan F, Ahmed S, Mahbub A, et al. CatBoost: cluster-based under-sampling with boosting for imbalanced classification[C], 2017 2ndInternational Conference on Computational
[6] Friedman J. H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001, 1189- 1232
[7] Quinlan J R. Induction of Decision Trees[M]. Kluwer Academic Publishers, 1986: 81-106.
[8] Bre iman, L, Random forests [J]. Machine Learning, 2001, 45(1) :5-32,
[9] Ke G, Meng Q, Finley T, etal. Lightgbm :A highly efficient gradient boosting decision tree[C]/1 Advances in Neural Information Processing Sys tems.2017:3146-3154.
[10] Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset[J]. International Journal of Computer and Information Engineering, 2019, 13(1): 6-10.
[11] Zhang J, Mucs D, Norinder U, et al. LightGBM: An effective and scalable algorithm for prediction of chemical toxicity–application to the Tox21 and mutagenicity data sets[J]. Journal of chemical information and modeling, 2019, 59(10): 4150-4158.
[12] Yan J, Xu Y, Cheng Q, et al. LightGBM: Accelerated genomically designed crop breeding through ensemble learning[J]. Genome biology, 2021, 22(1): 1-24.
[13] Tang M, Zhao Q, Ding S X, et al. An improved lightGBM algorithm for online fault detection of wind turbine gearboxes[J]. Energies, 2020, 13(4): 807.