A LightGBM based Default Prediction Method for American Express

Zhiren Gan[1],
Shenzhen University,
Shenzhen, China,
gzr19970608@gmail.com,

Junyuan Qiu[1],
South China University of Technology,
Guangzhou, China,
michaelqiu2000@gmail.com,

Fuli Li[2],
Shanghai Lixin University of Accounting And Finance,
Shanghai, China,
lifuli1202@outlook.com

Qian Liang*,
Yunnan Normal University,
Yunnan, China
lchris2021@163.com,

***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 personal credit risk prediction problem can be regarded as a binary classification problem. In this paper, we use data published by American Express in Kaggle to conduct a study of default prediction, and reduces the default risk of consumer finance companies through the effective identification of defaulting customers through models. We utilize LightGBM for default prediction. Related work is described in section II, and we introduce our methodology and experiment in section III and IV. To evaluate our experiment's performance, we do compared competitions. We define the new experimental metrics. The result shows that LightGBM owns the highest metric 0.692 among these models, which is 0.007, 0.032, 0.008 higher than Xgboost, Lasso and Catboost respectively.**

***Keywords*-credit default, feature engineering, LightGBM**

I. 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 personal credit risk prediction problem can be regarded as a binary classification problem, and it is a common method to solve the problem by using machine learning algorithms to build a credit evaluation model. This paper uses the match data published by American Express in Kaggle to conduct a study of default prediction, and reduces the default risk of consumer finance companies through the effective identification of defaulting customers through models. We utilize LightGBM for default prediction. Related work is described in section II, and we introduce our methodology and experiment in section III and IV.

II. 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 is mainly a machine learning model, and commonly used methods include SVMs, random forest XGBoost, neural networks, [3] established Logstie regression, K-means clustering, support vector machine, and random forest models to predict borrowers' default risk, respectively, and the results showed that the prediction effect of random forests was better than other methods.

[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 is a combinatorial approach based on existing models, where the predicted probabilities of one model are used as input variables for another model. For example, [6] proposed a hybrid credit scoring model based on Logistic regression and neural network, and the study showed that the prediction accuracy and robustness of the hybrid model were better than those of a single model. [7] used support vector machine, random forest and XGBoost to establish a credit prediction model, and compared with Logistic regression, the experimental results show that the performance of the three single algorithms is better than Logistic regression, and they are voted weighted fusion, and their performance is better than that of the single--model, which has better resolution, higher prediction accuracy, and is more suitable for Internet credit personal credit evaluation; [8] introduced the Stacking integrated learning algorithm in the credit evaluation system, and the effect has been improved, which is worth putting into use in practice.

- ● Our Contribution
- ✓ We utilize LightGBM to predict the credits' Default.
- ✓ We introduce our dataset and do some analysis.
- ✓ In the experiments process, we do the comparing experiments and the result shows that our model performed better than the other models.

III. 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.

IV. Experiments

- ● Experiments data

The objective of this paper is to predict the probability that a customer does not pay back their credit card balance amount in the future based on their monthly customer profile. The target binary variable is calculated by observing 18 months performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days after their latest statement date it is considered a default event.

The dataset contains aggregated profile features for each customer at each statement date. Features are anonymized and normalized, and fall into the following general categories:

D_* = Delinquency variables
S_* = Spend variables
P_* = Payment variables
B_* = Balance variables
R_* = Risk 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.

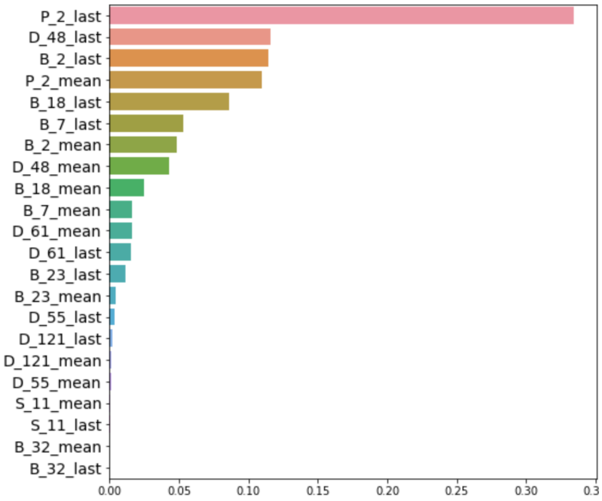

Figure 1: feature importance figure

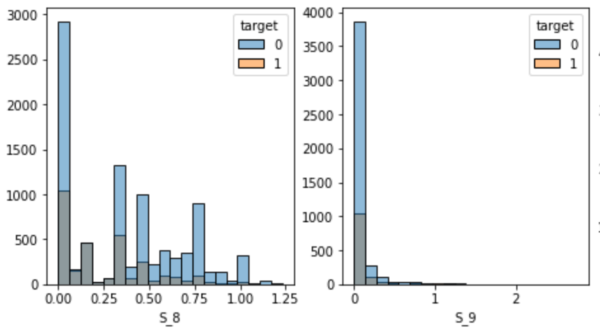

Figure 2: feature distribution

- 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.

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 M is the mean of two measures of rank ordering: Normalized Gini Coefficient G and default rate D captured at 4%.

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

The default rate captured at 4% is the percentage of the positive labels (defaults) captured within the highest-ranked 4% of the predictions, and represents a Sensitivity/Recall statistic. For both of the sub-metrics G and D, the negative labels are given a weight of 20 to adjust for down sampling.

- Experiment result

To evaluate our experiment's performance, we do compared competitions. The higher metric is, the better our model will be. The experiment result is shown in table 2. Our LightGBM owns the highest metric 0.692 among these models, which is 0.007, 0.032, 0.008 higher than Xgboost, Lasso, Catboost respectively.

Table 2: Experiment result

Models	Metric
Xgboost	0.794
Lasso	0.769
Catboost	0.793
Lightgbm	0.801

V. Conclusion

In our paper, we do feature engineering and using LightGBM as our model to predict credit default. We introduce related work in section II and the model in section III. In section IV, our experiment is stated. We introduce some part of feature engineering and give the concrete parameters for lightgbm. In the experiment part, our model LightGBM owns the highest metric 0.692 among these models, which is 0.007, 0.032, 0.008 higher than Xgboost, Lasso, Catboost respectively.

VI. Acknowledgement

Thanks to Zhiren Gan and Junyuan Qiu for providing some guidance of the topic in this paper. We would also like to thank Fuli Li and Meng Luo for their suggestions and analysis of the model training and model evaluation. In the future, we will spend more time in data analysis to build a better feature set. Besides, the deep learning method will be also investigated to build a better model.

Reference

[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 statis-tics, 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.