MAFS 6010Z Project 1: LightGBM Model on Default risk Data

Team Member: Fu Qiyin, Liu Enjie, Yu Xintong, Zhao Encong
{qfuab, ezhaoab, eliuac,xyubi} @connect.ust.hk
Kaggle ID: math6010z_Fu_Yu_Liu_Zhao

1. Introduction

We try to use a variety of different source data to predict unbanked clients' repayment abilities. We compared several models and selected LightGBM as training model to due with the data with large-scale. After feature processing and first model training, we found some of features more important and selected 391 features to train again, which improved 0.3% of AUC(0.786 overall). To sum up, it shows that personal information, historical behavior, current lending behavior and whole defaukt situation can contribute to default risk prediction.

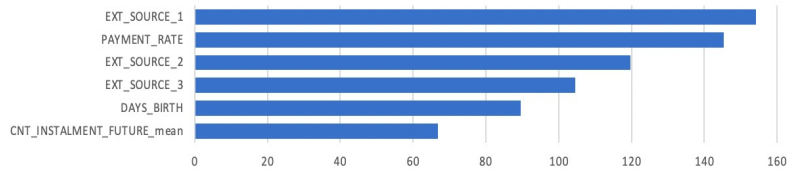
2. Model Selection

LightGBM uses histogram-based algorithms, which bucket continuous feature values into discrete bins. It can reduce cost of calculating the gain for each split, reduce memory usage and handle large-scale data. Because we have large-scale data and a large number of features, LightGBM is a suitable choice.

3. Feature Processing and Model Training

Firstly, we do feature processing. We encode the categorical features, use average minimum and maximize values to aggregate numerical features, standardize some features through rate calculation, and then join the tables together.

Then we first train the model with a higher learning rate and tree depth, then calculate the importance of features and eliminated 211 features with zero contribution, then do training again with a low learning rate and higher tree depth, to improve model accuracy.



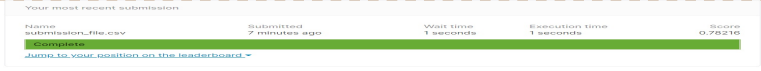
The important level of features

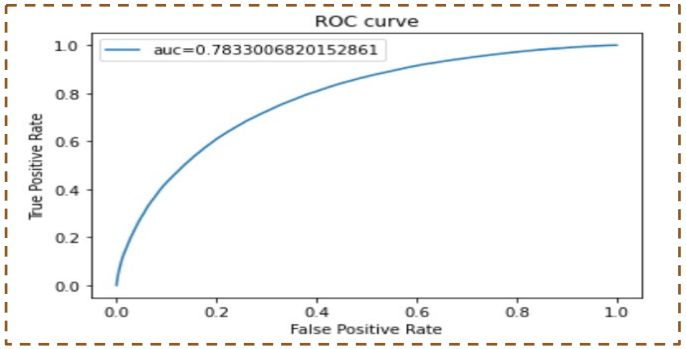
4. Prediction

In the case of credit default risk prediction, a user's credit behavior (deposit record, credit card record, loan record) over a period of time will be helpful. Based on this hypothesis, we selected features to train model and evaluate the model on testing set.

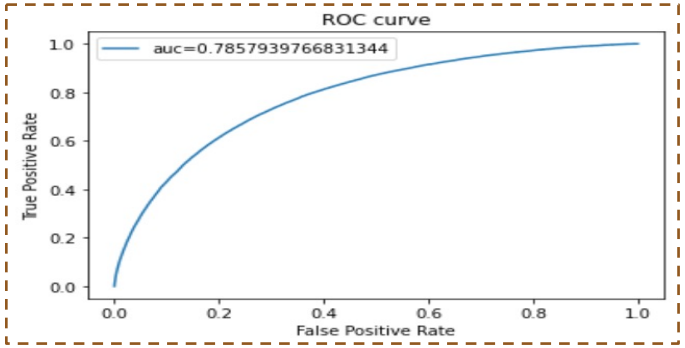
Using 5-fold cross validation, we can find that the first training model can achieve an AUC of 0.783 on testing data. Eliminating 211 features with zero contribution, the second training model can achieve an AUC of 0.786 on testing data, which is slightly larger than the first one.

Uploading our test result to Kaggle, we gain a score of 0.78216.





The ROC Curve of 1st model



The ROC Curve of 2nd model

5. Analysis and conclusion

Based on the contribution to the model, we choose the most 25 important features and , we can conclude that the variables which contribute to the accuracy of the predicted default risk mainly include 4 types: personal information, historical behavior, current lending behavior and whole default situation:

1. Personal information:

Age, car age, gender and working years, which can reflect a person's financial status and paying ability; In addition, the current demand debt and the amount to be repaid by letter of credit show a person's debt situation.

2. Historical behavior:

Default on the last loan (days, amount), previous credit amount, ID card modification, remaining days of previous credit. When a customer has defaulted in the past, we tend to think that this person is more likely to default and does not violate common sense.

3. Current lending feature:

The commodity price of consumer loans can describe the current loan amount.

4. whole default situation:

The maximum overdue amount of credit card and the maximum expected date of credit card, this two types of feature indicate that, among all customers using the company's loan services, how many people have breach of contract.

7. Contribution

Coding

- Fu Qiyin

Data processing and chart

- Yu Xintong, Zhao Encong

Report and Poster

- Liu Enjie, Yu Xintong