Report for the Home Credit Default Risk Competition

Chen Liu
Department of Mathematics
cliudh@connect.ust.hk

1 Introduction of the Dataset

As shown in the competition description, this competition requires us to predict the clients' repayment abilities using various kind of alternative data `https://www.kaggle.com/corwinliu9669/homecredit-default-risk/output`. The aim is to help the corporation to evaluate clients without sufficient credit histories. This competitions offers us a train and a test set, we need to obtain a model from train set and utilize this model to predict the test set. The holder of this competition also provides us some extra data to tackle this problem including applicants' data from other institutions and applicants' previous information in this corporation.

2 Analysis of Data

This section includes some analysis of the data provided in this competition.

2.1 Main Table Data

For the main table data(application_train.csv and application_test.csv), they share the same kinds of features. The first column denotes the id for each client. And the target in the train split stands for the label for each client, which means whether the corporation should give a loan. This target is the one we should get for the test split. For the prediction, we need to give a prediction($pred \in [0, 1]$). The evaluation of our prediction uses the area under the ROC curve. The detailed definition of this metric can be obtained from `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`. The feature columns in the main table includes information to describe the financial condition or other parts of the client such as whether the client owns a car.

2.2 Extra Data

The extra data includes bureau_and_balance, previous_applications,pos_cash,installments_payments and credit_card_balance. For bureau_and_balance, it includes the information of the client's credit from other institutions such as monthly balance. For previous_applications, the historical application information of the clients is given. Th pos_cash,installments_payments and credit_card_balance contains information from other sources.

3 Methods

This section includes the way to tackle this problem, including data processing as well as model fitting.

3.1 Data Processing

Feature Converting The features in this competition varies in the data type. For the numerical features, we utilize them for model fitting directly. For the categorical features, I attempt to convert

them in a simple way. If the feature contains two categories, it will be encoded as binary feature, for features with more categories, one-hot vector strategy is selected.

Data Cleaning Then it comes to data cleaning, I follow the instructions of `https://www.kaggle.com/mayn35/lightgbm-with-simple-features-new-work` to drop 4 features, 'CNT_CHILDREN','DAYS_REGISTRATION','DAYS_ID_PUBLISH','HOUR_APPR_PROCESS_START'. For the extra data, I only use bureau. For bureau, I extract the numerical feature from the original csv and add them to the feature matrix.

3.2 Model Fitting

For the model fitting parts, I choose gradient boosting decision tree [1]. And I use the implementation of LightGBM [2] for its fast-speed and reliable performance.

4 Experiments

This part includes the experiment results, both the one with and without using extra table are shown. For LightGBM, I set the parameter as follows, the number of esitimators is 1000, the learning rate is 0.01, the number of leaves is 50 and the maximum depth is 5. To improve the performance, I split the train set into two parts and get the final prediction with the average of the models trained with the part. We can get the conclusion that adding bureau adds to the prediction performance of the model. The detailed results are shown in Table. 1 Due to the time constraints, here I only test the combination of one extra data, Bureau. The other combinations can be explored in a similar way to get a better understanding of the data. The future works include how to dig more useful information and drop some spurious features.

Data	Public Score	Private Score
Main Table	0.74264	0.74140
Main Table + Bureau	0.75221	0.75402

Table 1: This table illustrates the results of the model trained with different data source.

References

[1] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.