MSBD 5013 Statistical Prediction

Peer Review

Review Group: 9
Reviewer: LO, Ngai Hung

1. Summary of the report

This group worked on Home Credit Default Risk data set. Dataset analysis, how to perform feature engineering, model used and experiment results are discussed in this report.

All csv files with 308k training samples and 46k testing samples were used. To handle missing data, all features with missing value more than 90% samples were deleted. After that, all remaining missing value were replaced with 0. Numerical features were aggerated by count, mean, max, min and sum.
One hot encoding was applied on categorical features. Moreover, count, mean and sum on encoded categorical features were created on the data other than Main dataset.

The baseline model used was Logistic regression and LightGBM was used for further investigating the problem. In order to prevent over-fitting max-depth of LightGBM was adopted. 5-fold cross validation and grid search tuning was applied for model selection during model training.

Due to limited hardware resources, instead of using all feature, different combinations of features were used on model training. As a baseline model, Logistic regression achieved 0.68453 Kaggle score. The best Kaggle score was 0.76614, which is obtained by using LightGBM and features from Main, Previous and Cash dataset.

2. Describe the strengths of the report

- Used Figures to visualize the data distribution.
- Discussed the key information and concept in a simple and clear way.
- Mentioned the reason and thought behind most of the strategic used.
- Clearly explained the model selection process

3. Describe the weaknesses of the report

- To handle missing value, this group tried to delete attributes which have more than 30 % missing values, but the performance dropped. It would be better to discuss more why 30 % is chosen as the threshold. Is there any analyst was done on using different threshold?
- Can mention more setting on the model parameters.
  The report mentioned by using max-depth of LightGBM , it can prevent model to be over-fitting. However, according to the code, default value of max-depth was used, which means no limitation on depth. It would be a good idea to clarify this in the report.
- According to the code, reg_alpha and reg_lambda it applied when training LightGBM. It would be great to briefly mention regularization was adopted and explain why both L1 and L2 regularization was used.
- The term “Test (Kaggle Score)” was shown in experiment results. However, Kaggle score is generally described as public/private score. It may cause the reader a bit confuses whether it means private score or not.
-

4. Rating:

Evaluation on Clarity and quality of writing: 4
Evaluation on Technical Quality: 4
Overall rating: 4
Confidence on your assessment: 3