MSBD 5013 Statistical Prediction

Peer Review

Review Group: 8
Reviewer: LO, Ngai Hung

1. Summary of the report

This group worked on Home Credit Default Risk data set. The method used on feature exploration, data cleaning, model selection, model tuning, and performance analysis are discussed in this report.

Four datasets were selected based on whether it closely related to clients' financial situation. Maximum, minimum, mean and standard deviation were used to aggregated client data and merged to main dataset by using SK_ID_CURR as the key. To handle missing values, the group tried to remove features with more than 30% missing value, but the performance of this method was not good. Finally, instead of removing the features, they used model that can handle NaN value or replace missing values with the mean.

Percentage features were constructed on numerical features. Binary encoding, ranking encoding and one-hot encoding were applied to categorial feature according to the characteristic of the features. Pearson correlation and PCA were used to perform features dimensionality reduction. However, the performance of model also dropped compared to using all features.

Four models, including logistic regression, SMV, Random Forest and LightGBM were used. LightGBM was chosen as final model and obtained private and public scores as 0.77412 and 0.77786 by using 5-fold validation. In this model, annuity credit and external sources are the most important features.

2. Describe the strengths of the report

- Discussed the key information and concept in a simple and clear way.
- Easy to follow the concept flow of the data analysis process.
- Mentioned the reason and thought behind most of the strategic used.
- Provided examples to help on method explanation. For example, how to construct percentage example on numerical features.

- Used Figures to visualize part of the findings or results.

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 numerical result. For example, the result of models other than LightGBM and the amount of performance dropped after removing features by PCA or Pearson correlation.
- Can discuss more on the model training part. 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.
- Pearson Correlation Analysis would generate positive and negative correlation coefficients. Both strong positive and negative correlated features would be consider as more useful in general. In this group, the top 100, 200, 300 and 400 best features are selected to experiment respectively. It would better to explain how the coefficients be ranked before selecting top n-values.

4. Rating:

Evaluation on Clarity and quality of writing: 5
Evaluation on Technical Quality: 4
Overall rating: 4
Confidence on your assessment: 3