Rebuttal of Project 1 – Group8

CEN Xinxin, MA Xiaoran and WU Xiang
{xcenab, xmabi, xwucb}@connect.ust.hk

1. Congjian Chen:

The project applies several models including both traditional methods and deep methods, and I think the quantitative improvement of model could be mentioned in the poster.

Reply: Thanks for your comments. We think it is a good suggestion. We tried different models and the results showed that LightGBM model got the best results, so we only focused on this model and showed its performance due to the limited space of a poster.

2. Li Yilin:

I don't think this report has any obvious weakness. But if I have to find one weakness then I would suggest the team to expand more on the hyperparameter tuning of the LightBGM model.

Reply: Thanks for your comments. Of course, we tried different hyperparameter settings to tune the model to get good performance. But we think the hyperparameter tuning results will not be the key content to be shown in the poster, so we do not show them. The final hyperparameter we used can be found in the code. We agree it will be better to show this part in the poster.

3. longaihung

➢ *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.*

Reply:

Thanks a lot for your suggestions. But for one suggestion about *Pearson Correlation Analysis*, we just want to clearly explain something rather than rebut. Firstly, we used *Pearson Correlation Analysis* to compute the correlation between features and labels rather than features and features. Then, we could intuitively consider that the higher the positive correlation coefficients are, the more important the features are. Although the drawback is that we only consider the importance of linear correlation between features and labels. Therefore, we could rank the contribution of features according to the level of positive correlation coefficient to select top-k features. However, the k is a hyperparameter, the improvement could be that explaining why we choose 100, 200, 300 and 400 dimensions.

In addition, we agree that it will be better to explain why both L1 and L2 regularization are both used. The reasons are that for the L1 regularization, we can tune this parameter to let some useless feature weights to be zero, and for the L2 regularization, we can tune this parameter to let the feature weights to be even so that avoiding the situation that only some feature weights control the model. Both L1 and L2 regularization can control the over-fitting problem of model.

4. Tsechunlok

The group did not visualize the AUC scores of the tested models. No hyperparameters tuning on LightGBM is mentioned and experimented.

Reply: The same as 2.

5. Yang Yuxin

1. The reasons for table selection and feature extraction may be vague. I have seen the code and haven't seen the explicit reason for choosing tables and features. For example, report said add some additional features in numerical features but without explicit reason. Maybe losing the analysis of the correlation between different features. In the end, I still couldn't figure out how many features were selected as variables.

2. Report said that they have compared with top100,200,300 features and compared different models but it seems there was not such process in the source code (some of the code have not been run). In the source code it seems there was only one method lightgbm came with the result.

Some suggestions:

1. *You can prune the model by optimizing the parameters.*
2. *You can add more visualization and EDA of the data.*

Rebuttal:

Thanks for the comments. As we have mentioned in the *Datasets Aggregation and Data Exploration* part of our report, the reason that we aggregated three additional dataset is that we aimed to depict each applicant with richer information. We used previous_application.csv, credit_card_balance.csv, and POS_CASH_balance.csv because we thought previous application history, credit card balance history and POS/Cash loans history are closely related to clients' financial situation, thus would reflect clients' repayment abilities. These contents have been explicitly mentioned in our report, please double check if you have any questions.

As for EDA, we noticed that your group have done a lot of work on EDA and chose only to investigate the main dataset. In our project, we don't think drawing fantastic chart should be the main work of this project, instead, we should aim to get higher AUC score by aggregating more useful information. Thus, we focused on introducing more meaningful features: both features generated from supplementary datasets and some manual features constructed by ourselves which has certain meanings. Since we have over 300 features, it was impractical and meaningless to compute pairwise correlations like you did, instead we chose to investigate feature importance and that part was included in the report. As you can see from our result, our method got quite great result, verifying the strategy to aggregate more features was effective. We have had basic understandings of the dataset, otherwise we wouldn’t be able to investigate different datasets and came up with an efficient model.

As for our source codes, we indeed did not display the running results of all comparisons in our notebook, because we decided to focus on LightGBM which has the best performance due to the size limitation of poster, so the final notebook mainly shows the results of LightGBM. But that doesn't mean we didn't try or implement what we mentioned in our poster. Because you could obviously see the corresponding processes we saved in our codes and you also can run them to check the all results. However, we agree that we should better display the performance comparison results of all models and make our techniques more quantitative and persuasive, and we will get it next time.

We very appreciate your suggestions but also hope you could understand our rebuttals. We would like to argue the low score you gave us on the technical quality, because we really did a lot on this part and our method was effective according to the final result. Having different workflow and strategy does not mean that our approach is unreasonable and lack of consideration.