**Group 6. Home Credit Default Risk**

1. Summary of the report

Multiple machine learning methods were used to predict whether the clients will default or not in this project, including Random Forest, Neural Network and LightGBM. Feature selection and feature engineering techniques were used, resulting in an improvement in AUC score. It is impressive that this group used all of the provided data in this project, and applied different data pre-processing methods according to the properties of the models. The model with the best AUC score (0.79236 in the private leaderboard) was the LightGBM model.

1. The strengths of the report

The report was clearly written, including nearly all the key points and technical details of their project. They did very careful work in data preparation, including merging all 8 datasets, handling missing values, handling outliers, aggregating numerical features and categorical features. It was also impressive that they built many meaningful manual features based on their understandings of the original features. Also, after they aggregated all the datasets together, they also did investigation on the feature importance in order to select features. In terms of model training, they compared different models including Random Forest, Neural Network and LightGBM. They used cross-validation to tune the hyper-parameters, and used early-stopping technique to avoid overfitting. Of course, the AUC score on the Kaggle private leaderboard shows how great this work is. Their meticulous handling of the data and models earned them this excellent score.

1. Weakness of the report

I should say it is not weakness of the report, I’m just providing some of my point of view for your reference. I noticed that you used Min-Max Normalization, I wonder have you tried some other normalization methods and compare the results? Because I think Min-Max Normalization is sensitive to extreme values / outliers, in some occasions it might make the distribution of feature not ideal, even though you have handled some extreme values.

I also want to ask have you ever considered using some dimension reduction methods, such as PCA, to reduce the dimensionality of your features? Because I have seen your efforts to reduce the memory occupied by the large sparse matrix by precision reduction, maybe PCA is also a method worth trying.

1. Evaluation on Clarity and quality of writing: 5

The report is clearly written and well organized, and there are clear figures and descriptions in the report. The Report is generally good, with only a minor typo in the first page: I think you mean the dataset is extremely *unbalanced*, but what you wrote was extremely *balance*. But this minor typo does not harm the great quality of the whole report.

1. Evaluation on Technical Quality: 5

The results are technically good. And there is no obvious flaw in the reasoning. There are detailed descriptions related to theoretical analysis and experimental results, which are well thought out and convincing. Besides, they provided detailed codes of data preparation, model training, and evaluation. Thus, it is possible for other researchers to replicate these results. They also provided Kaggle results and some references related to this project.

1. Overall rating: 5- My vote as the best-report
2. Confidence on your assessment: 3- I have carefully read the paper and checked the results