**Group 7. Home Credit Default Risk**

1. Summary of the report

Logistic Regression, Fandom Forest and GBDT models were utilized in this project to predict whether the clients will default or not. The unbalance issue of the data was detected and was solved by oversampling and down-sampling methods. Besides, Polynomial features were constructed and K-fold cross-validation was used to select hyperparameters. Finally, they got the best AUC score 0.75320 in the private leaderboard.

1. The strengths of the report

The report was clearly written, and in addition to the main dataset, they also merged bureau dataset into the main dataset to generate more features as complementary information. They also checked the distribution of the target column, missing values, and column types, and encoded categorical variables in a proper way. They also used oversampling and down-sampling to solve the unbalanced issue. As for model training and hyperparameter tuning, they compared three different models, Logistic Regression, Random Forest and GBDT, and used K-fold cross-validation to select hyperparameters. They also did some feature extension by introducing Polynomial features into the feature vector. The whole workflow was clear and rational.

1. Weakness of the report

I noticed that in this project, only the main data and bureau data was used to build up the feature vector. And the 6 main features selected was the features with the highest correlation with target, then more polynomial features were further generated using these 6 main features. But I think those features which have lower correlation with the target should also be considered, because they may still serve as supplementary information to further improve the prediction ability. Besides, the idea of generating polynomial features was quite great, but this process significantly enlarged the dimension of the feature vector, while many of these features might not be that important. Thus, I think manually construct some new features which has explainable meanings would be a better choice. The last suggestion is that maybe you can aggregate other datasets into the main data in order to further improve the prediction ability of the model.

1. Evaluation on Clarity and quality of writing: 4

The report is quite good overall. The main work flow was expressed quite clearly. But it would be better with more detailed descriptions, especially in the first half of the report. For example, why did you choose the bureau instead of other dataset as complement to the main dataset? What is the intuition to include many polynomial features? Did these new features really help the prediction? It would help a lot if these kinds of discussions were provided in the poster.

1. Evaluation on Technical Quality: 3

I think only using the main dataset and the bureau dataset is not sufficient. It was acceptable if you provided clear reasons of selecting certain datasets instead of the others. Besides, selecting 6 features which has the highest correlation with the target and then generate more polynomial features seem quite arbitrary. Maybe you can try to provide some evidence to show that this process really improved the performance of the model, then it would be more convincing. A feature importance would be a great choice. In my opinion, generating some manual features would be a better choice than constructing so many polynomial features, because perhaps most of these polynomial features would not help the model to improve.

1. Overall rating: 4- A good report.
2. Confidence on your assessment: 3- I have carefully read the paper and checked the results