Peer Review – Group 10

• Summary of the report

The Group 10 finished *Home Credit Default Risk Prediction* project. They analyzed data before feature engineering and used some selection strategies during feature engineering. Then they built decision tree models, like Random Forest, XGBoost and LightGBM and compare their performance on this task. Among them, the best model achieved 0.77701/0.77984 AUC on Kaggle. Finally, they also did feature importance analysis and good conclusion.

• The strengths of the report

1. They analyzed the label distribution firstly and have some strategies for selecting features, like dropping features with low correlation coefficient and removing features with 60% missing value.
2. Then, they handled with numerical and categorical attributes respectively, like using statistical value and one-hot encoding.
3. Besides, they used L1-regularizaiton to prevent overfitting and used k-fold cross validation for model selection.
4. Finally, they also used SHAP technique to show important features in LightGBM and put forward that LGD (Loss Given Default) is better than ROC as a metric in this task.

• The weaknesses of the report

1. It may be not clear that what csv data they used for this task.
2. They didn't explain why there are some different performances between these tree models. In other word, if they assessed the strength and weakness of their models, the work would be more completed.
3. Refer to the categorical features procession in LightGBM, maybe it's not a good choice that using one-hot encoder to encode all categorical features.

• Evaluation on Clarity and quality of writing (1-5): 4

Firstly, the report is written clearly and they used good figures to analyze feature correlation before feature engineering. Besides, the report is well organized and completed. It has the introduction of task, the details about data analysis, feature engineering, model selection and training strategies, the illustrations of feature importance and good discussion. Moreover, I find few typos. But the only small drawback is that they did not show what and how many csv data files they sued in this task at the beginning.

- **Evaluation on Technical Quality** (1-5): **5**

Overall, their results are reasonable in terms of the technology they used, because our group had tried the same experiment, so the results are replicated and acceptable. The highlight is that they used a technique to show important features and also put forward that ROC may be not a good metric in this task which proved they had a further thought.

- **Overall rating** (1-5): **4**

A good report.

- **Confidence on your assessment** (1-3): **3**

I have carefully read the paper and checked the results.