Group 8

• Summary

This project completes the classification task of loan default risk prediction with a

mature data analysis process using several datasets from Home Credit. The best

model (LightGBM) reaches a public score of 0.77786 and private score of 0.77412.

• Strength

The project clearly explains data pre-processing methods, including the aggregation

methods for several datasets and detailed methodology for dealing with missing

values & outliers. Feature engineering is divided respectively for numerical and

categorial variables, which is effectively used in real world. Different types of models

(GLM and tree models) are used to complete the classification task from various

aspects, which indicates the full workload of the group.

• Weakness

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.

• Evaluation on Clarity and Quality of Writing: 5

The poster is well arranged with clear figures of both feature importance and ROC

curve. It’s easy for me to get the points by the highlighted headlines. The reasons

why not all datasets are utilized are clearly and reasonably explained in the poster.

Also, the Analysis and Conclusion section provides the insights of the results, which is

of great value for business decisions.

• Evaluation on Technical Quality: 5

In the feature engineering section, Pearson correlation analysis and PCA both

contribute to feature importance analysis. The model construction section explains

the strength of different models. Great prediction results (public/private score =

0.77786/0.77412) are a matter of course.

• Overall Rating: 5

• Confidence on Your Assessment: 2