Group 4

Summary:

This project investigates into home credit default risk based on some historical or alternative data of customers. It uses logistic regression and light GBM for modelling. The light GBM model has an overall higher prediction accuracy. It is assumed to be because GBM has a more complex structure. Also, they found that a set of more comprehensive and related data can lead to better prediction.

Strength:

1. Novel method for dimension reduction

In data processing part, their methods for dimension reduction are simple but turns out very effective. While many projects try to fill in all the missing values in the dataset, they decided to drop attributes with more than 70% missing values. Also, they apply standardization on the dataset.

2. Deep insights about a state-of-art model, light GBM

In analysis and conclusion parts, the authors proposed some incisive thoughts about light GBM and how to further improve.

3. Clear structure of the poster

The poster is very structure and easy for audience to follow. It clearly explains the results of data analysis, the reasons for feature selection methods, the choices of models and their mechanisms. ROC graphs clearly show performances of the models.

Weakness:

1. Unclear goal

This project only discusses two models. If the goal is to better predict default risk, maybe some other models or ways of improving accuracy could be considered. Or if the goal is to compare model performance, maybe the title of the project could be adjusted to better states the purpose.

2. Discussion part

The report could be further improved by adding a discussion part to talks about the problems they think they have in the project and discuss possible ways for further improvement.

Evaluation on quality of writing (1-5): 3

The poster uses clear language along with intuitive graphs. The logic in the poster is well organized. One thing I think could be improved is that the authors could consider adding the Feature Importance graph in the poster to better present the result of modelling.

In addition, there are some typos and grammar mistakes. While most of them are minor, there is one significant typo in *3. Feature Selection* part, where *correlation* is mistakenly written as *corr1elation*.

Also, it could be more straight forward if the authors could give some information about the final model in the poster. The poster only shows the result for model comparison yet mentions no details about the final model.

Evaluation on presentation (1-5): 5

The presentation is delivered in a well-structured way. Speakers presented their ideas and work clearly. Their slides clearly show the main ideas of the project. One minor thing I think could be improve is the introduction of light GBM. Instead of using many sentences with small font size to introduce the model, the speakers could consider just put some key words with larger size on the slides, which could help the audience to get the main idea more effectively.

Evaluation on creativity (1-5): 3

I think they presents some creative methods in the project. The model they use, light GBM, is a state-of-art model, just been developed recent years. Also, the decision to standardize dataset is less seen in data analysis projects and novel.

Confidence on your assessment (1-3):3