MAFS 6010Z AI in Fintech - Project 1 Rebuttal

Group Members: WONG Hoi Ming (20641276), WONG Sik Tsun (20038819)

For reviewer 1,

- no specific questions or recommendations have been made and thus we skip this part.

For reviewer 2,

- He/she suggests we have more discussions on features constructions and try different data processing techniques. The key data processing tools we used include i) aggregating multiple records with minimum, maximum, sum and mean, and ii) combining variables with ratios, averages, and differences. We agree with his suggestion and believe other methods such as using medians or treating data with time weights could be attempted. As for the reasoning of features construction, we briefly touched on that on the report/codes and will elaborate more here. For occupation type, we believe different jobs will have different sensitivity with respect to economic cycles and in bad times some are easier to lose jobs than others. Also, we reckon that various occupations should have similar sources of risks, like housing and realtor or banking and insurance. We therefore apply re-grouping using this rationale. For example, we group "School", "Kindergarten" and "University" to “Education”, a group with less variability in incomes and higher credibility. We also apply regrouping of goods category by their nature. This is to reduce the dimension of datasets in one-hot encoding.
- Besides, the review also suggests we could do more evaluation and fine-tuning of the model with functions like *gridsearchcv.* We agree and believe this is common point that different groups of the class could work on and improve. We would consider using different regularization parameters, learning rates and maximum tree depths.

For reviewer 3,

- The review suggests that we could have more discussion on the reasoning as well as the strength and weaknesses of our work. We believe one of the strengths would be the model selection, since gradient boosting decision tree can fit well on large scale and high dimensional data, and it can also handle well on categorical data and data with missing value. Besides that, we believe the inclusion of the domain knowledge features is one of the key factors in improving the prediction accuracy, which is reflected in the feature importance plot. The weaknesses would be that we did not try to adjust the model parameters, such as the tree depths and learning rates, and we did not include all the datasets provided, which may contain valuable information for the prediction. Based on the weaknesses we identified, we have already included some of the possible improvement points in the report.
- Besides, the review also suggests that we could include citation/comparison and to improve the formatting of the report. We agree that we could perform comparison by fitting on more different models such as logistic regression and with different model parameters, we can then apply cross validation technique to compute the prediction score with the training data and select the model with the best performance. We also agreed that we can put the citation and improve the formatting by highlighting the key messages and creating bulleted or numbered list. Our

knowledge on the gradient boosting methodology and the python coding is mainly gained from the previous courses (e.g. MAFS6010S & MAFS6010W) and some online research.