MAFS6010Z: Project1 Review for Group 2

Liu Enjie eliuac@connect.ust.hk (20745838)

- **Summary of the report:**

This report provides a detailed and clear introduction to the logic and process of data processing, variable selection and model selection. (1)Variable selection: The report first takes [days birth] variable as an example to introduce the idea of single variable selection. They then used Pearson correlation coefficients for multivariate selection, but since it actually made the final model perform worse, all variables were used in the final report. In addition, considering multicollinearity, variables with high VIF has been eliminated. (2) Data processing: The large number of null values in data is treated as a form of variable information. (3) Model selection: After trying 7 classifier models, LGBM was finally selected according to its max AUC value. As a result, this group gain a kaggle test of 0.748.

- **The strengths of the report:**

1. Careful selection of variables: Considering both single variable, multivariable selection and multicollinearity problems, this report filters all variables in detail.

2. Clear report writing: Each step is clearly explained in concise writing, and all logic and methods can be easily understood by the reader through examples and diagrams.

- **The weaknesses of the report:**

1. Lack of interpretation of final results: For the interpretation of the final results, this

report mainly stays on the performance of variables and models. Considering that the credit default risk is a problem with practical significance, readers would prefer to see the explanation close to the practical significance.

- **Score**

Evaluation on Clarity and quality of writing: Score: <u>5</u>

Evaluation on Technical Quality Score: <u>4</u>

Overall rating: Score: <u>4.5</u>

Confidence on your assessment: Score: <u>3</u>