MSBD 5013 Project 1 Peer Review

Group: 3 Reviewer: Yuen Zhikun (20505288)

1. **Summary**

   This group has a comprehensive way to process the data before modelling. They check the missing value very carefully and found that some of the features have more nearly 70% of missing values. Also, they examine the overlapping of different data tables and prioritize the highly overlapping data first. They use some common techniques such as filling mean for the missing value and normalize the numeric data with standard scaler. Later they reduce the features by correlation and encode different types of features. In modelling, they choose LightGBM to fit the data and fine-tune with k-fold cross validation. They achieved 0.75803 AUC on their final submission result.

2. **Strengths**

- Very detailed data visualization, some impressive insights can help dive deeper to the problem, e.g. show the percentage of missing value and loan instance overlapping.
- Impressive approach to select useful features and filter out the table with smaller area of overlapping.
- Good to use K-fold cross validation to reduce the variance and use hyper-parameter searching to fine tune their models.

3. **Weaknesses**

- Some figures are not very helpful to show the message they want, e.g. the two correlation matrices. Too many features in the x and y axis. Not mentioned any useful points from the two figures.
- Not delete the rows with missing features is an appropriate decision but the features with high percentage of missing value are not deleted also. No any explanation of why not delete those features with over 70% missing values.
- Use almost one page to explain the details of gradient boosting tree and LGBM.

4. **Rating:**

- Clarity and quality of writing: 4
  The figures in EDA section can show some points accurately. The confusion matrix of the correlation can use less features or show some of the features are high correlation only.
- Technical Quality: 4
  Good to use Bayesian optimizer to optimize the LGBM.
- Overall rating: 4
- Confidence: 3