The review of the project of Group 13.

Summary:

The project of Group 13 can be roughly divided by 4 parts: analyzing the data, dealing with missing values and outliers, using numbers to label categorical variables, and selecting features. The parts worth mentioning are as follows.

First, they used bar chart to show the distribution of the data and found that 8% of the applicants failed to repay the loan.

Then they used a column of bool variables to label the features containing missing values and calculate the percentage of missing values in each column. The columns containing missing values more than 60% are dropped because of difficulty of calculating the mean and median. After that, the missing values of remaining columns are replaced by median.

After labelling the categorical variables, they began to train the model. They adopted cross-validation and lightGBM to drop redundant features and select desirable ones, and to avoid overfitting, they set an early stopping parameter to stop the training when the performance was not improving. Then they use the selected features to train the model.

Strength:

They try to use their financial knowledge to introduce several new features to the model, which makes their model more accurate and specific for financial data. Also, they use lightGBM to train their data and then save plenty of efforts.

Weakness:

They only consider the features of linear forms and ignore the possibility of their quadratic or cubic forms such as a^2, a^3, which may weaken the model' s fitting ability. Perhaps this is why the predicting ability of their model is not good enough.

Writing:

The clarity and quality of writing (1-5): 4

The report is clearly written, and the size of fonts are appropriate. The figures are good used and well organized. But the figures are not clear enough, for example, the figure under section 2.1 Distribution of target variable.

Technical Quality:

Evaluation on Technical Quality (1-5): 4

There is no obvious flaws in the reasoning. And their claims are mainly well-supported. But as I mentioned in the "Weakness", the model' s fitting ability need imoroving. The authors didn' t assess their work about the the strengths and weaknesses. Relevant reference papers are cited in the end of the report, but they didn' t compare their work to the presented work.

Overall rating: 4

Confidence on my assessment (1-3): 2