# MSBD5013 Midterm Project

## Peer Review for Group 9

### 1. Summary of the report

Group 9 makes attempt to build a baseline Logistic Regression classifier and LightGBM classifier to predict the default risk by the probability for each loan example via utilization of multiple datasets. They statistic the TARGET distribution and missing value attributes distribution for reducing several feature columns. They join multiple datasets and groupby with SK_ID_CURR for feature engineering. With 5-fold cross validation for the model, they indicates that more datasets involved such as Main+Previous+Cash will achieve the best AUC performance score for prediction.

### 2. Strengths of the report

(1) The report main content is logistical.

(2) The workflow of this report can be easy to follow-up. It makes an introduction about the aim and the main datasets for usage.

(3) There is no style and grammar problem here.

(4) In modeling part, it is great to provide a Logistic Regression baseline for reference.

(5) They engage enough features in this project for training through Colab platform.

### 3. Weakness of the report

(1) The x-axis for distribution of missing value graph is relatively ambiguous. It will be better if this graph shows the detailed attribute names of columns that have missing values with a percentage over 90%.

(2) In feature engineering part, more reasons need to be provided for exacting numerical / categorical features and for feature selection.

(3) Since the final specific features can be selected through LightGBM model, more detailed reasons for supporting model and feature selection can be further involved more clearly to increase the model interpretability.

### 4. Rating of the report

(1) Evaluation on Clarity and quality of writing (1-5):

3.5

This is a well-organized report and there are also no problems with style, grammar, or typos basically. However, more detailed reasons and support ideas can be involved for more clearly written. More examples and figures can be provided in feature engineering for exploration.

(2) Evaluation on Technical Quality (1-5):

4

The AUC result is good overall in prediction. The model can also be easy to replicate. In technique, besides simply join the several dataset sheets for training, more features can be further extracted for better training and clearer interpretation.

(3) Overall rating (1-5):

3.5

On average, it is a good report.

(4) Confidence on your assessment (1-3):

3

I have carefully read the paper and checked the results.