## Summary of the report

The group divided the whole project into various part, including a brief introduction, data cleaning process, feature engineering process, comparison of the results of different models, conclusion and future work.

## Strengths of the Report

This group has tried to make use of data sets other than the main data set (i.e., application_train.csv and application_test.csv) and illustrated that the data set has imbalanced target variable. The report gives some examples with the features containing outliers (e.g., "DAYS_EMPLOYED") and graphs to visualize the difference of the distribution before and after data cleaning.

## Weaknesses of the Report

Perhaps it is a good idea to state the reason of using principal components analysis at the beginning of the report, e.g., reduce the time of model training.

## Clarity and Quality of Writing (5)

The report clearly shows all the procedures of various process, including data preparation, model development and model evaluation, in a well-organized structure. Therefore, readers can easily follow the report.

However, perhaps the report would be better if brief explanation of the techniques being used could be given. For instance, the report can elaborate more about "principal component analysis" and the rationale behind. Also, it would be a good idea to compare the models training time with and without "principal component analysis".

## Evaluation on Technical Quality (4)

Principal components analysis: although this technique can greatly reduce the computational resources required for the model training process, it cannot boost the performance of the different models

Feature engineering: this group has created features from the original data set, e.g., 'CREDIT_INCOME_PERCENT' which is constructed by dividing 'AMT_CREDIT' by 'AMT_INCOME_TOTAL'

Directly replace all missing values with median may not be the most appropriate way as there are some features having more than 50% missing values. Such way may dilute the relationship of that particular feature to the target variable.

Various models: this group has built different models for results comparison (i.e., logistic regression, random forest, LightGBM), which the models without "principal component analysis" always outperform the models with PCA respectively.

**<u>Overall Rating (4)</u>**

**<u>Confidence on the Assessment (3)</u>**

After reading all the code and researching, I understand all the techniques being used by this group and the rationale behind.