Summary of the report:

This group used application {train / test}.csv and bureau.csv for training and prediction. The group has done some data preprocessing and feature engineering to encode categorical variables and increase data diversity respectively. Oversampling and undersampling methods have been applied to the training data. Oversampling yielded the best results. Three models including Logistic Regression, Gradient Boosting Tree (GBDT) and Random Forest (RF) have been experimented and both GBDT and RF achieved the highest AUC.

Describe the strengths of the report:

The group has done some feature engineering to increase the variance of dataset. The group identified that the data is extremely imbalanced and have done oversampling and undersampling to tackle this problem.

Describe the weaknesses of the report:

Normalization may not be required for tree based algorithms such as GBDT and Random Forest. The effectiveness of the polynomial features from feature engineering have not been investigated. Some data from the dataset are not used, which may provide additional information that could potentially increase the AUC of the models.

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

4

Evaluation on Technical Quality (1-5):

4

Overall rating (1-5):

3.5

Confidence on your assessment (1-3):

3