Group 7

Summary

Overall the team did a good job in modeling the default risk based on the given data.

Strength

1. The team did a decent amount of data preprocessing to fully understand the data.

2. The team used SMOTE (desample+oversample) to deal with the skewness of the data.

3. The team used 2 ensemble methods including GBDT and random forest

Weaknesses

1. Naming inconsistency. For example, the second to last csv file in preprocessing.ipynb was named

"train.csv" but the first file read in main.ipynb was named "train\_scale.csv". Without consistent naming

I wasn't sure what data were used for the final models.

2. Code comments. The code would be easier to read with more comments.

3. Report writing. Due to the word limit of a poster, there were not enough space for details for all the

work they have done.

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

3.5

Due to the word limit of a poster, there were not enough space for details for all the work they have done.

For example, I do see that logistic regression was used as a baseline comparison, but there was no

discussion on the performance. Moreover, the report did not mention which model was selected and why.

Evaluation on Technical Quality (1-5):

4.5

Missing values.

I do see that missing values were calculated for the 122 features in app\_train. However I do not fully

understand how was the missing values treated. I do see that the median imputer should work on the

missing numerical value, but what about categorical features?

There were 6 features that were generated from the bureau, are there any missing values when you

join it with the app\_train?

Overall rating (1-5):

4

Confidence on your assessment (1-3)

3