1. Summary of the report

Group 2 chose data from application.csv to do the default prediction. In terms of variable selection, they at first wanted to choose variables based on the Person correlation between the variables and the target value, however, they found out it worsens the performance of models using this approach. Then they change the approach to look at VIF and eliminate variables that have VIF larger than 10 (according to the code). For models cannot take in NA values they filled NAs by generating normal random numbers for numerical variables and by mode for qualitative variables(according to the code). They evaluated model including LDA, LR, RF, GB, SGD, XGB and LGBM and finally chose LGBM because this model performed the best. Their final public score on Kaggle is 0.748 (AUC value).

2. Strength of the report

The poster formats in a pretty good manner. It's concise and catches the point: they use blocks to present each step with proper amounts of graphs and use bold font to emphasize some words. It is a good point to touch on the reason they think why LGBM performed the best in the conclusion part.

3. Weaknesses of the report

The font in first three graphs is too small and hardly can be seen clearly. Because it's a poster and there's limited space, so it doesn't touch on evaluation result of each model as well as the data processing part, and I had to look at the code, the Jupyter Notebook can be run smoothly and is commented though.

4. Evaluation on Clarity and quality of writing (1-5): 4.4

Is the report clearly written? Yes.

Is there a good use of examples and figures? Yes. Is it well organized? Yes.

Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Yes, like sentences below

However, considering the multicollinearity problem in multiple regression model may largely effect(affect) the variance

For the figures comparing the applicant's age(s), there is no significant different between default and non-default.

However, for this dataset, the method is not the effective one as it actually worsen(s) the performance of the model.

5. Evaluation on Technical Quality (1-5): 3.6

The results are technically sound when looked through the code. I think the data processing and model evaluation is not mentioned enough in the poster. And I don't think the authors clearly assess both the strengths and weaknesses of their approach. Relevant papers are cited but are not compared to the presented work.

6. Overall rating: 4 (good poster)

7. Confidence on your assessment: 3 (I looked at and ran the code they submitted)