Peer Review of Project 1 Group 17

- **Summary of the report.**

This report summarized the data consolidation processes of 3 datasets that have been used in the prediction and how they use the LGBM model to predict the probability of default of clients.

- **Describe the strengths of the report.**

The data cleaning and consolidation process of this project are well and carefully designed, functions and steps are exact and well organized. The effect of LGBM model fitting is excellent. The code is tidy and highly readable.

Good understanding of the features. Also, the figure of feature importance clearly demonstrated the effectiveness of domain knowledge features.

- **Describe the weaknesses of the report.**

It is possible to improve the structure of the report. For part 2, 3 and 4, for each dataset, the approaches are similar, so it seems to be no need to repeat the approach for multiple times. Instead, after going through the code, I believe it would be more informative by building the report by different data processing techniques, such as construction of domain knowledge features (regroup of some categorical values for dimension reduction and new features), identification of illogic inputs, aggregation etc.

- **Evaluation on Clarity and quality of writing (1-5): Is the report clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper and provide details of typos.**

4 - Clearly written, good use of figure, good formatting.

Some typos and grammar mistakes:

Part 4 line 5 “set zero”→ “set to zero”.

Part 5 line 8 “R2m”→ “R_2m” to make sure the symbols are consistent in format. However, I suggest using the “Equation” function of MS PowerPoint to type the symbols ($e.g. R_{2m}$, J_m).

- **Evaluation on Technical Quality (1-5): Are the results technically sound? Are there obvious flaws in the reasoning? Are claims well-supported by theoretical analysis or experimental results? Are the experiments well thought out and convincing? Will it be possible for other researchers to replicate these results? Is the evaluation appropriate? Did the authors clearly assess both the strengths and weaknesses of their approach? Are relevant papers cited, discussed, and compared to the presented work?**

4.8 - Sound results, solid reasoning, easy to replicate the results.
Relatively weak evaluation. It would be better to include further fine-tune of the model in the report. For example, apply *gridsearchcv*.

- **Overall rating: (5- My vote as the best-report. 4- A good report. 3- An average one. 2- below average. 1- a poorly written one).**

4.5

- **Confidence on your assessment (1-3) (3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)**

3