1、 Summary of the report.

The fifth group of classmates completed the topic "home credit default risk" very well. Firstly, they describe and summarize the data as a whole, calculate the correlation between every independent variable and target variable, select the variables which have the highest correlation with the target, and they deleted some variables with more than 40% missing value, for non-numerical columns, they used the one-hot encoder method to convert them into numerical. After marking NAs separately, they used cross validation to calculate the accuracy of the model. They used Naïve Bayes, Randomforest, AdaBoost, XGBoost, LightGBM and other models with the increasing CV score, and they finally used LightGBM as the final model.

The latter part is the further improvement of the model results. The first is feature engineering, which processes some "outsiders", adds some previously deleted variables, generates some new variables according to the existing variables, uses PCA for dimension reduction, and finally obtains the validation score of 0.78587

2、 Describe the strengths of the report.

(1) They have a relatively complete analysis of the data and extract the data columns most needed for prediction.
(2) The missing value and outlier are handled well.
(3) Many models are compared comprehensively, and the model with the best prediction effect is used reasonably.
(4) Some new variables are generated creatively, and convincing results are achieved.

3、 Describe the weaknesses of the report.

(1) Feature Engineering and feature selection can be considered at the beginning of the report, that can make the report more logical, rather than simply as a means to improve the effect of the model.
(2) As they mentioned, most related or most important features are listed, without reasonable explanation on how these features affect credit default.

4、 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.

Evaluation on Clarity and quality of writing (4)

The writing of this report is very clear, but there are still some typos such as “Figure 9: most importance features in the new training dataset”, where importance should be replaced by important, there are many figures but not enough examples, the report is well organized and few problems with style and grammar, I think the report should unify figures style.

5、 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?

Evaluation on Technical Quality (3)

the result is technically sound and there are no obvious flaws, and the result is convincing and easily be replicate. But the authors did not assess the strengths, and the analysis of weakness is not clearly enough. Relevant papers are cited and discussed.

6、 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- A good report

7、 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, 1My assessment can be wrong)

2- I just browse the paper without checking the details