Peer Review For Group 4

Mini-Project 1. MATH 4995

Group Number: 4

Title: Machine Learning Basics Kaggle Contest: Home Credit Default Risk

Summary of the report

The objective of the report is to use light gradient boosting machine as prediction model to predict the problem proposed in Home Credit Default Risk. Multiple table in the given file from Home Credit is used. Feature selection is deployed and filter out features with many missing data or highly correlated with other features. Before using LGBM, logistic regression is also used to analyze the data.

Strengths of the report

As LGBM is used, the Kaggle score is very high, which this effect is consistent with Group 3. The prediction of LGBM would make the prediction very effective. Moreover, more features are used rather them only using the data in training set, which makes more features could be available, and have potential to find better features. The multiple ways of aggregation is also a good way to exhaust all possible features, without understanding them.

Weaknesses of the report

As LGBM is used as the primary methodology, it might be better to state out the thought of the entire report. It is quite possible that the initial thought of prediction is not LGBM, which the progress before having LGBM as the result would also be valuable also. The analysis could be more detail in terms of explaining why a model perform better.

Evaluation on quality of writing: 4

Is the report clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with sty le 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.

In the writing, the information related to methodologies are clearly written, and with the graphical illustrations, it helps readers to understand LGBM. However, the ROC curves is a bit blurry in a way that it is unclear to me what does the result implies, which might need further explanation. Other than that, the flow of the project has been clearly explained. It could been see that the LGBM improves more from merged dataset than logistic regression, which might be one way of exploring further from the report.

Evaluation on presentation: 4

Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.

The presentation did present many things that are stated in the poster or source code, including the aggregation of features, using logistic regression and LGBM. Reasoning on each steps are also provided. The authors have shown to be well aware of the advantages and disadvantages of each of the model they have chosen. And as not everyone knows LGBM, they have also explained it in a understandable way.

Evaluation on creativity: 2

Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations or extensions of existing ideas are also valuable.

LGBM would be one of the common way of giving out desirable results at least in this contest, which makes it one of the popular model in this contest. The way of aggregation would be useful to many of the model with relational data also. In fact, the aggregation to LGBM part should be possible to be generalizable to many different question other than the one proposed by Home Credit. While there is a lack of specificity towards the problem, and that both methods are commonly used, by combining those 2 and giving a desirable result did give insight on the usefulness of aggregation and LGBM.

Confidence on your assessment: 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)

I have read through the poster, source code, and the presentation carefully and that the result is consistent with my understanding and experience.