Review1_group16

1. Summary of the report

The report first give a brief introduction of the procedures and the key findings of important features, then they have done the data cleansing, feature engineering and also, dealt with the missing value and imbalanced data. For the model training and testing, they choose random forest as the main model and choose two methods including SMOTE and Under Sampling to deal with the imbalanced data.

2. Describe the strengths of the report.

It is a completed report and the procedure of data cleansing and feature engineering part are both clear and great. What impress me most is the method of SMOTE to tackle with imbalanced data. Though it is not technically hard, the idea behind paying attention to the imbalanced data is worth praised.

3. Describe the weaknesses of the report.

The report is not clearly enough since I found the model -- random forest they are using from their codes instead of directly known from reports. Besides, for the imbalanced data, it is better to check the AUC instead of accuracy before dealing with the imbalanced data. For the feature selection, they just choose top 10 correlation features, while I think it may be better if they compare the result since there may be some features that are not linearly correlated with the target. What's more, I didn't see the result of their kaggle test score and there is no clear group member contribution.

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.

Point: 3.5. I think the overall logic is clear while there miss some basic points, for example, it would be better to state the use of randomforest model, and your team do the feature selection according to randome forest etc. There is no reference. Found one typo formatting -- in part two, third points, there is no space between ' stillhave'.

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?

Point: 3.5. The kaggle competition scores based on AUC, while I didn’t see any related test in the modeling. Whats’ more, I think it would be better if you deal with the imbalanced data at the very beginning data processing part, instead of at the very end, since the choice of modeling part is based on the judgement of data.

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).

Point: 3.5.

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, 1- My assessment can be wrong)

Point: 3.