Critical Analysis Task

Unit Name: COMP2200 Data Science with COMP6200 Co-badged

Student Name: Umut DEMIRHAN

Student Number: 46739106

Date: 29/05/2022

In this task, we were given a dataset, that contains the homeloan records, and a Jupyter notebook file which includes some analysis of the data, machine learning models and evaluations and visualization of the results. However, it is required to find some issues with the processes in the notebook, explain the problems and provide solutions. I have managed to find many issues; however, I will only discuss three of them in this report as I think they are the major issues.

The first issue that I realised is about cleaning the data set. When examining the data, there was one null value detected. It was in the 'not.fully.paid' column, and it was replaced with the mean of the feature which should not be because this feature is Boolean; therefore, it only takes 0 and 1 whether it is yes or no. This negatively affects the accuracy of the data and therefore the model. We have just one null value; Removing it would not affect the analysis and the predictive model. It would be an appropriate resolution.

The second issue is dropping the features with a negative correlation with the label. It is clearly wrong because, positive or negative, if the correlation is close to 1 or -1, it can be a significant predictor. A negative correlation means that if one increases, the other decreases or vice versa. Therefore, instead of dropping negative correlations, we should drop the ones that have no correlation such as 'revol.bal', 'log.annual.inc', 'pub.rec', 'delinq.2yrs' and 'installment' as their correlations are close to zero with the 'credit.policy'.

The third major issue is sorting the data and splitting them into training and test sets in order. It is quite a mistake because we use training sets to train the model and testing sets to test the model whether it is accurately predicting the target variable. Therefore, we should split the data randomly so that the prediction can be unbiased.

In conclusion, there are many issues in this notebook, and these are not coding issues, but logical mistakes that affect the accuracy of the model and analysis. With this task, we realized that coding accurately does not mean that the analysis and the model are also correct. We need to be aware that our interpretation may be wrong, so it is necessary to have a good grasp of the relevant subject. Otherwise, we may cause losing time and money for the company we work with by causing wrong decisions. Lastly, in this case, I did not include any screenshots of the codes or illustrations of the answer because I already explained why they were an issue, how they affect the outcome and how we can resolve those issues. Therefore, I think it would not be appropriate to repeat the answers with screenshots.