Chan Koon Lam
20748995
Peer review of group 13 project:

Summary of the report:
The goal of the report is to use ML algorithms to predict the probability of the home credit default risk of an individual, based on the loan profile and characteristics of the individual. Group 13 has used exploratory data analysis to examine the data before inputting the data into different models. Furthermore, the group has used feature engineering especially domain knowledge to extract features from the existing data set. For the modelling part, the group has combined a 5-fold cross validation together with Light Gradient Boosting Machine as their solution model. As a result, they have also found that EXT_SOURCE_2 and EXT_SOURCE_3 are the most important features in their model that decides the probability of the home credit default risk of an individual.

Strengths of the report:
- The use of Exploratory Data Analysis is a good practice to help understanding the data in a broader view and discover more insights which will help the group in later model selection and fine tuning. The group has outlined the distribution of the target variable, showing that most of the applicants in the training data does not have risk in repaying their loan, with only 8% applications out of the total has a problem/risk in repaying their debt. Also, they have dropped columns (features) that has more than 60% of missing data, which could help building a more precise and efficient model.
- The group has used Domain knowledge feature to derive new features from the existing data. These new features are likely to help deriving a more precise model as they are more likely to be influential to the output (TARGET). Domain knowledge is often useful in machine learning as they help to identify new features that are likely to be related/important to the output. In this case, the group has used their expertise knowledge on finance as domain knowledge to create four more new features. We can also see that one of the domain knowledge feature - TERM is one of the important features in the feature importance graph in section 5.
- After their first attempt on the model, they have spotted the problem of overfitting and applied Cross-validation as well as adjusted the early stopping parameters in the model to resolve overfitting problem. It is a good practice to review the performance of models and try to find out what is the reason behind poorly performed models.

Weakness of the report:
- In section 2.2.2, the group has replaced outliers of the DAYS_EMPLOYED feature with Nan values and reused it as a feature. I think the group could run a Z-score test to determine outliers for different features as a more general methodology, instead they have only handled the outliers in a few specific features, this could result in a bias when handling/examining the importance of features. On the other hand, the abnormal data in the DAYS_EMPLOYED column contributes to a significant percentage of total entries, it might be better to just drop the feature instead.

Evaluation on Clarity and quality of writing - 4:
- Most part of the report is clearly written, methodologies are well accompanied with examples and figures and the structure of the report is also well organised. In sentence " we try to replace abnormal days employed with NaNand label a new feature state abnormal issue." In section 2.2.2, there is a typo where a space is missing for NaNand.

Evaluation on Technical Quality -4:
- Overall the report is technically sound without obvious flaws in the reasoning. In the feature engineering section, the group could better support their selection of their domain knowledge features with the backing of empirical evidence. For example, the group has included YEARS_EMPLOYED_PERCENT as a new feature by diving the year of employment by age of the applicant, it would be more convincing if they could provide empirical evidence that the effect of working competence of an applicants has to their debt repaying ability.

Overall rating: 4

Confidence on assessment: 3