MAFS6010Z: Project 1 Review (For group 19)

Summary of the report

The report summarized how the group performed the prediction by modelling the default risk using logistic regression.

The group checked the dataset provided, including any missing and extreme value, as well as the distribution of default and nondefault records for each feature. They also constructed new features which perform high correlation with default rates.

Since fitting the logistic regression using the whole dataset can only achieve a low prediction score of 0.55-0.66, the group tried to select only part of the features to be included in the model. They performed the features selection by checking the distributions of default and nondefault records, judging by common sense as well as constructing some new features. The improved logistic regression model achieved a prediction score of around 0.71.

The group then performed analysis on the advantages and limitations of the fitted model, and suggested future improvement measures.

Strength of the report

1. The group explained how they perform the feature engineering in details. They not just only checked for the missing and extreme value, but also constructed new features by combining the existing features, which can help reducing the number of features in the model but keeping features with significant correlation with the default rate.
2. The group successfully improved the prediction score by performing feature selection. The prediction score of 0.71 is quite decent given that logistic regression has some major limitations such as assuming linearity between the features and the response.
3. The report is clearly written. It is organized which mentioned the feature engineering first, followed by the introduction of the model, production result and the analytic part. It also includes tables and figures to help illustrating the content.
4. The python coding is clearly presented with comments explaining the purpose in each part.

Weakness of the report

1. For the data preparation part, the group only made use of the application_train dataset and neglected the other dataset provided which could provide valuable information for the prediction. Besides, replacing the null value by the sample mean may not perform well for all feature since it depends on the distribution of each feature variable, and the report does not provide the rationale of such approach. Also, the report does not mention how the categorical features are transformed to fit in the model.
2. The feature selection was based on the distribution of feature variables combined with common sense. Methods such as subset selection together with cross validation can be considered as a supplement to perform the feature selection. Besides, no regularization method is applied (e.g. Lasso or Ridge) to reduce the variance of the fitted model.
3. For the result analysis part, the report does not provide more details on the major features included in the model fitting and the resulted regression coefficients. It also does not present the feature importance. This information can be used for model analysis and help readers understanding the fitted model better.

Evaluation on Clarity and quality of writing: 4

As mentioned in the strength, the report is clearly written and is organized. It is easy to follow.

Evaluation on Technical Quality: 3

The result is technically sound and have no obvious flaws in the reasoning. However, more work can be done in the data preparation, feature selection and result analysis to further improve the fitted model as mentioned in the weaknesses.

Overall rating: 4

Confidence on the assessment: 3