Group 7

Summary:

This project investigates whether full dataset or a selected subset is better at prediction based on Home Credit Default Risk dataset on Kaggle. To compare the difference, principal component analysis (PCA) is conducted. Then the author manually selected some features among the result from the PCA to eliminate multicollinearity. Then he performed modelling on both full dataset and selected subset, models he used include Logistic regression, Linear discriminant analysis and Random Forest. The result shows that premium subset helps to improve the performance of logistic regression but has no effect on the other models.

Strength:

The author uses many visualizations to help the audience perceive the result. Also, the author gives many incisive thoughts about dimension reduction and how to avoid high correlation between features. He thoroughly analyzed the data and looked into the mechanisms of each model.

Weakness:

The major problem is that models, such as random forest, already perform sorts of feature selection inside its function, such as limiting depth or number of leaves. The author did not mention how he avoid such feature selection when using the whole dataset for modelling. As the author mentioned in the conclusion part, premium subset only works for logistic regression but not the other models. Could it be because the other models already did some similar reduction inside function implementation?

In addition, there are some points that are unclear in the project. For example, the author did not explain why he manually dropped features that have correlation coefficient higher than 0.6 with other features. Does this criterion have some reasons behind? It might happen that even though two features are 0.6 correlated with each other, adding both in the model would increase prediction accuracy, which is the goal of the project.

Evaluation on quality of writing (1-5): 3

The poster is well structured and uses precise language to explain its methodology. The figures in the poster are very intuitive. However, the introduction part in the poster is not concise enough and the language style tends to be subjective. Such introduction is more suitable for a report but not a poster. Instead, I suggest the author focus on describe the major goal of the project in a neutral language.

Evaluation on presentation (1-5): 5

The PowerPoint is well organized, and the author provides many graphs to intuitively show results of data analysis and modelling. The speaker delivered his presentation clearly and thoroughly introduces the logic of his project.

Evaluation on creativity (1-5): 3

In my point of view, the question proposed in the project, which is whether using whole dataset or a selected subset could predict more accurately, has been widely discussed. Methods, such as Ridge and Lasso regularization, have been proven to effectively improve prediction accuracy by eliminating less important features. Maybe the author could investigate further and focus on, for example, the effectiveness of dimension reduction on different models.

Confidence on your assessment (1-3):3