C2T2 EDA Lessons Learned Report

DongMei Li, May 28, 2021

By performing the EDA, two main lessons were learned that the Data Science team can use for future projects like this.

1. The EDA prior to data preparation is no trivial. It is critical to prepare the data for the machine learning (ML) model building, as indicated by the BADIR Framework. It also determines to what extent we can go from what find in the EDA analysis to the underlying "why" (consistent with one lesson learned from a similar problem that was addressed last year).

Regression results from the selected features, which will be used as the parsimonious ML model building, show an R squared of 0.31, meaning 31% of the credit variation can be explained by the selected features. Even the regression results with all the variables render an R squared of 0.37%. Either way, the unexplained variance constitute a large portion.

Further, based on the correlation results, the few demographic features are not strongly correlated with "credit".

If possible, Data Science team can review relevant literature to see what other demographics and non-demographic features can be used to predict the credit limit and then check with Guido to see if they store these data. The goal is to prepare a better dataset for ML model building. At the very least, we can help Credit One collect better credit data in the future.

2. It is highly important to perform both data visualizations and statistical models. Graphing shows both demographics and other features are related to the target (credit limit). However, correlations provide more information and nuances.

- For instance, payment history turns out to generally have pretty large correlation coefficients and the weights of payment history, the bill amount, and the amount of previous payments vary from month to month. Therefore, the month that has the largest correlation coefficient was selected for those three sets of features to build the ML model.
- According to the graphs, "SEX" seems a good feature, but its correlation with "credit" is pretty low, as such it was not selected for the ML model building.
- "default" is also associated with "credit" but its correlation is not as large as I assumed.
- Another assumption is that college degree might matter a lot. However, it turns out that "graduate school" has stronger correlation with "credit", as such selected for the ML model building.

Informed by correlation and regression, the attributes in the data that are found to be statistically significant to predict "credit" are: 'default', 'AGE', 'MARRIAGE', 'EDUCATION_graduate school', 'PAY_2', 'BILL_AMT5', 'PAY_AMT6'.