Hector Jose Gonzalez Dominguez

Course 5 Task 2

November 12, 2019

Lessons Learned Report

A successful data analysis always starts with the preparation and initial exploration of the data. For the machine learning algorithm to work properly and optimally, we must “feed” clean data to our model. And the more we understand the data we have and its limitations, the higher our chances of achieving our project outcome.

One of the main lessons I learned from this experience is to preprocess your data before doing any modeling. For the *Default of Credit Card Clients* data set, this meant checking for missing values and discretizing some of the variables. Other examples that are also applicable to this task include data cleaning and transformation. The goal is to provide our predictive models with features that better represent the main problem we are trying to solve, resulting in higher model accuracy on unseen data.

Another important lesson I learned is to understand the data we are working with as much as possible. For this, we can use descriptive statistics as well as visualizations. Python and Pandas make this task easier, by allowing the user to quickly understand the makeup of the data and visualize its distribution. The benefits of exploratory data analysis include identifying outliers, comparing features, and addressing collinearity and covariance.

Per this report’s findings, the data science team should investigate the relationship between the variables related to bill statements. These variables are highly correlated with each other – it may cause overfitting in the modeling step. Another recommendation is to explore the difference (if any) in default rates among age groups, gender, and education level.