Guido Rossum

Data Science Team

The following report summarizes the lessons learned while performing the Exploratory Data Analysis on the data set of Credit One customers who have increasingly defaulted on loans they secured from various partners. Credit One needs to provide more appropriate credit limits for customers, in order to reduce default rates, resulting in customer retention and better revenues.

Whenever working in Jupyter Notebook, clarity of data, comments, and coding should be prioritized. For clarity, the “PAY\_0” through “PAY\_6” columns were modified to “SEPT\_STATUS” through “APRIL\_STATUS”, and a table was created and imported to define the “PAY\_STATUS” values. When performing a regression problem, it is ideal to limit the number of relevant features. So firstly, the ‘customer ID’ column was removed and ‘default payment next month’ was changed to ‘default’ for brevity. Next, all nominal columns must be converted to numerical for regression. The “SEX” and “default” columns were confirmed to have only 2 unique entries each, so Label Encoder was used to create binary values. “EDUCATION”, having 4 unique entries, required One Hot Encoding. The next step required converting the “PAY\_STATUS” values of -2 (no credit consumption) and -1 (paid in full) to 0 (revolving credit) values. This change ensured that customers who were paying on-time were not penalized within the regression algorithm.

The various plots using matplotlib and seaborn revealed that most customers have credit limits below $500k and are skewed towards the lower end ($10k-$200k) with a maximum credit limit of $1 million. Most customers are single or married, as opposed to divorced/other, 60% of customers are female and the majority of customers are Millennials (ages 26-40) followed by Gen X (ages 41-55).

The correlation matrix revealed that Pay Status, Sex, Marriage, and Education (sans graduate school) exhibit low correlation with Credit Limit and could be removed as features. The next steps are to choose and evaluate an appropriate model within an acceptable degree of accuracy.

The main lessons learned from this data analysis exercise are:

1. Don’t make assumptions about the data. Let the data tell you what is relevant.
2. Determine a data analysis process to follow and exercise discipline to execute the plan.
3. Use a table of contents and clear and concise comments within the Jupyter Notebook for organization and ease of future reference.