Hunter Messer

C2T2

Lessons Learned

One of the biggest improvements I have made up to this point is the data cleaning aspect of the analysis. This is the first data set where there were more than the traditional issues (blank cells, duplicates, etc.). After evaluating the data, I noticed that not only were the headers in the second row of the dataset, but there was also a duplicate header mixed within the data. This required writing the SQL file as a CSV file to my local drive, moving the header up, and indexing and deleting the duplicate header row. This allowed me to re-write the clean dataset to my local drive and re-read it within the Jupyter Notebook so that the data type of each feature could be interpreted. These steps were necessary to be able to explore the data and perform various statistical processes in the EDA.

With that, I also utilized Pandas Profiling Report to speed up the EDA. The Profiling Report tool quickly analyses the data set and provides descriptive statistics on the variables, along with a correlation matrix of the variables.

Based on the population of the data set, there is a 30% default rate, which seems to be significantly high. There is a positive correlation between default and the first bill month, as well as a negative correlation between not defaulting and the first bill month. In other words, there is a tendency for loan customers to default when there is a payment delay on the first month, and conversely, a tendency for customers to not default when they pay in full the first month. Furthermore, there is a positive relationship between the credit given and defaulted loans as well credit given and late payments.

There is no correlation between education and defaulted loans. Oddly enough, the majority of customers who defaulted either had a graduate degree (31%) or a university degree (50%). There is also no relationship between marital status and defaulted loans. However, defaulted loans tend to lean towards a younger population, as 33% of defaulted loans were within 20-29 years old and 34% within 30-39 years old.

Based on the above insights, I would recommend being more stringent on the amount of loan credit being given to customers. It seems that these customers are overextending their financial means and taking out loans that they simply can’t afford. If wanting to limit risk, I would also recommend extending loans to an older population as they tend to default less than younger individuals. This is most likely because they have established more security in regards to finances, and can handle unforeseen circumstances that cause so many to default on their loans. Lastly, I would recommend implementing some form of monitoring system that tracks customers who miss their first payment. As stated before, there is a tendency for those who miss their first payment to eventually default on their loan.