CREDIT ONE FINAL REPORT OUT

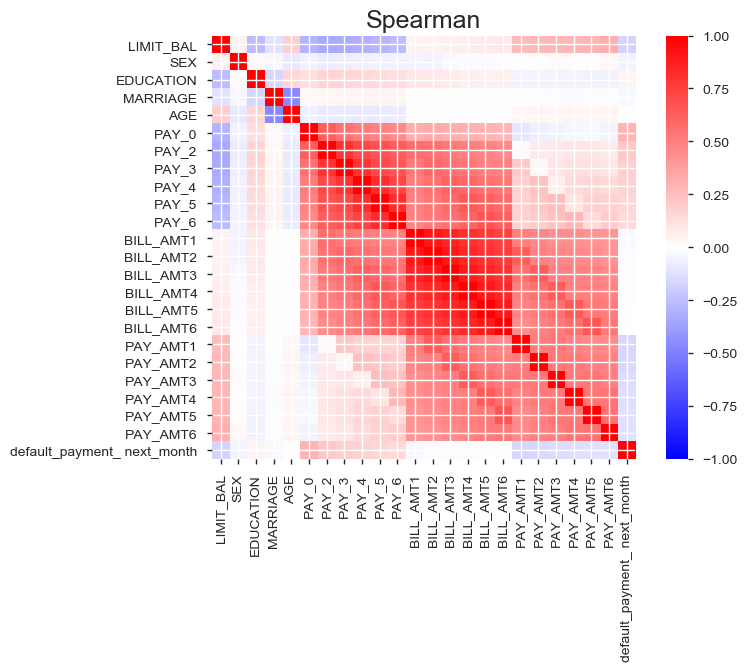
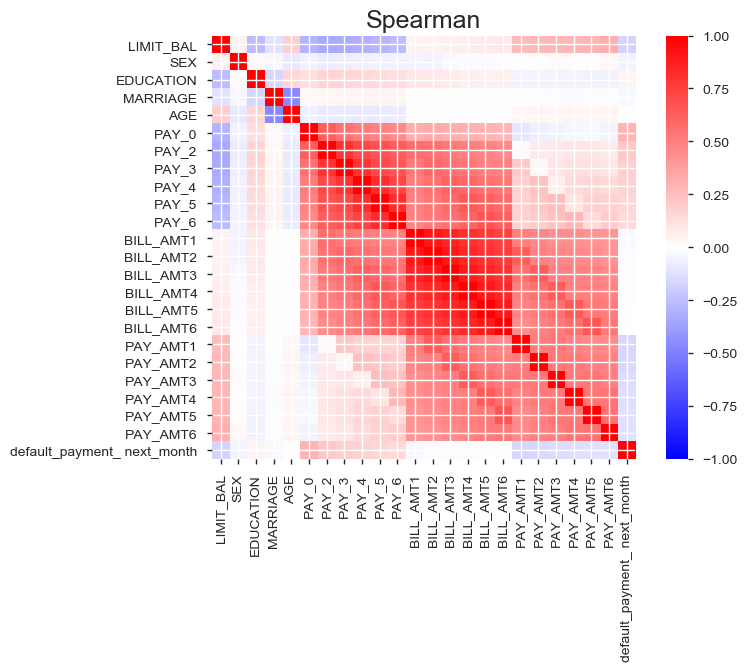
Marko Gorenc

Credit One contracted our team to complete machine learning work with the goal of utilizing existing customer data to identify which customers are more likely to default on their loans and to further utilize that information to predict the lending risk-profile for future customers. The dataset that was provided covered a six-month period completed recently. The features utilized covered 4 main categories, the first was the demographic data of each customer, the next two feature groups were the monthly payment history and the monthly balance history of customers. The final features were of the current limit balance of and a binary feature that recorded whether the customer had defaulted.

The work commenced after the dataset was acquired. Initial data cleanup and transformation was relatively extensive. Two features of column ID #s were removed to prevent the machine learning algorithm from using that insignificant data from making erroneous predications. Further, the headers were simplified and several additional headers within the sheet were removed. Machine learning requires numerical inputs, so where applicable, words were transformed to numbers. The transformation was guided by a “Rosetta Stone” document provided by Credit One, where for the “SEX” feature, male was set to 1 and female was set to 2, etc. Further work completed included dropping duplicate entries and removing null values.

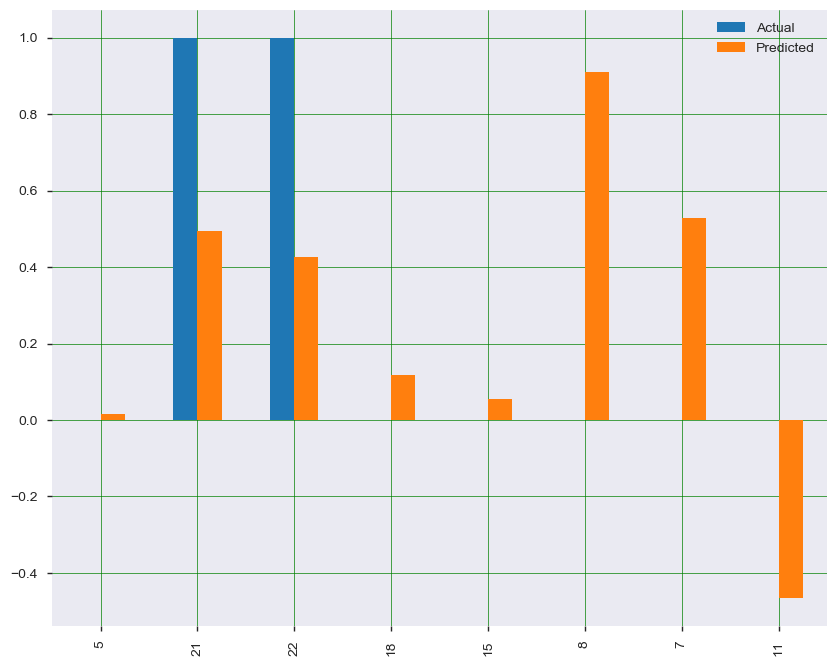
Exploratory data analysis work was completed to visually investigate relationships to focus on during the machine learning process. Because the focus of the work was on who was defaulting, the EDA work focused on demographic data. Presumably, the payment and balance information recorded by Credit One correlates to default, but to understand who was defaulting and to predict who may default in the future, the balance and payment information was largely disregarded. Ultimately a few conclusions on who was more likely to default were drawn from the EDA process: (1) men, and/or (2) single, and/or (3) have no more than a high school education and/or (4) older (45-79) and younger (<30). We present this work with caution due to the weak relationships observed with these demographic groups. These basic observations are likely pointing to the complexity of identifying high risk individuals and the difficulty in identifying trends in 1D or 2D plots.

Heat map of positive and anti-correlation shows that default payment next month strongly correlates with payment info but has only weak correlations with demographic data.



Despite the weak relationships observed during EDA, the team moved forward with machine learning. We utilized regression because of its strength in determining relationships between features and its ability to forecast once relationships are established. The key finding is that when we use demographic data to predict who is going to default, we struggle to come up with a strong, predictive model. The results of the three different regression algorithms used never exceeded values that indicate a relationship was successfully established that we could use to predict who, in fact, may be more likely to default. Subsequent model tuning runs failed to bolster our initial conclusions on the modeling.

The key takeaways from the work completed are that we cannot successfully predict, based on the data provided, if there is a given demographic group that is more likely to default on their loans. To bolster the modeling work, the team may need more data. Credit One provided data from a short period of time over which an increase in defaults occurred. Default increase likely biased the dataset. Additionally, understanding the long-term payment and balance trends may help to flag customers whose behavior changes over the longer periods of time, so an increase in historical data would aid our work. Finally, our team recommends collecting additional demographic information to determine whether the very weak correlations established through EDA hold.



Plot of model (default = dependent variable) results where blue bars plot actual data used in the machine learning process and orange bars show the predicted value after machine learning. Large deviations in the bar length demonstrate the poor overall predication capabilities of our model.