**Credit Default Analysis**

**Final Report**

**November 21, 2019**

**Credit One**

**S Womack**

**Objective:**

Investigate Credit One rise in customer default rates.

**Methodology:**

Data Visualization:

Look for relationships between the different features available in our data set, including customer default status, through the use of various charts and graphs.

Data Modeling:

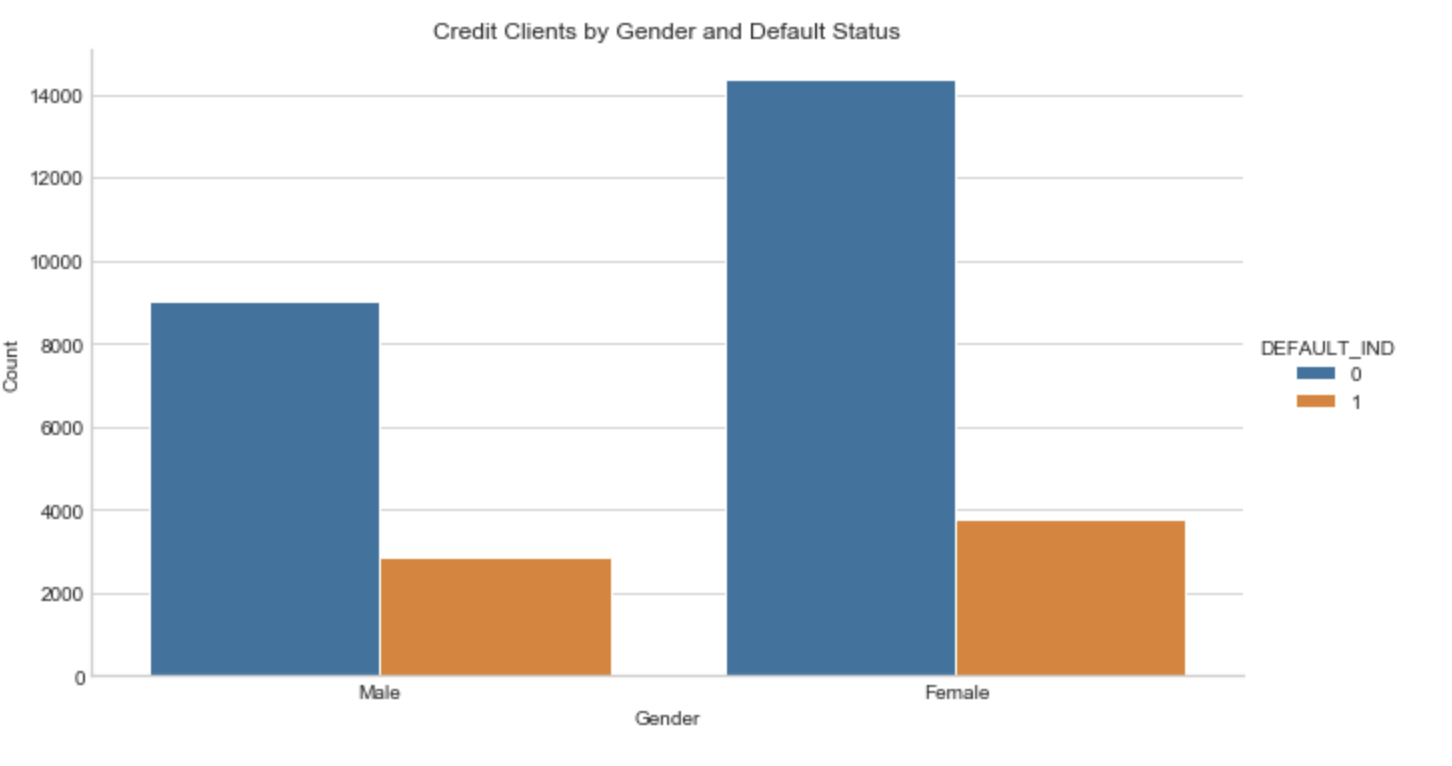
Through the use of machine learning, build classification models to predict if a credit customer will default, and uncover the more important attributes in those predictions.

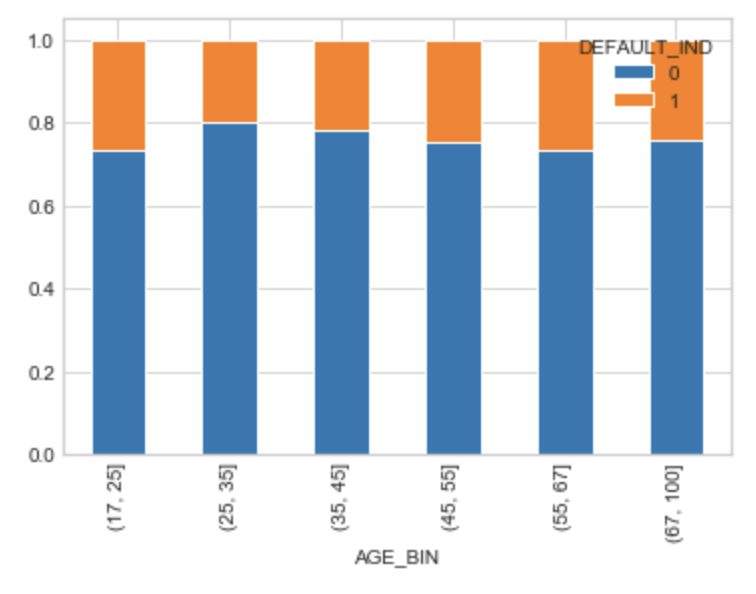
**Findings:**

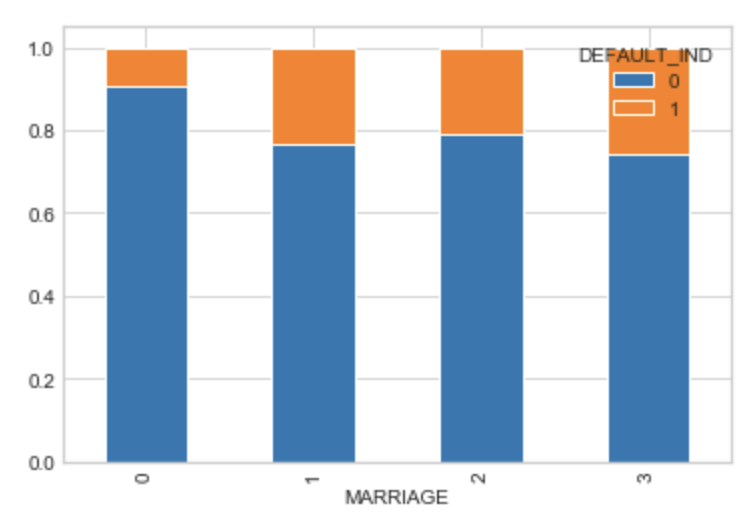
Demographics:

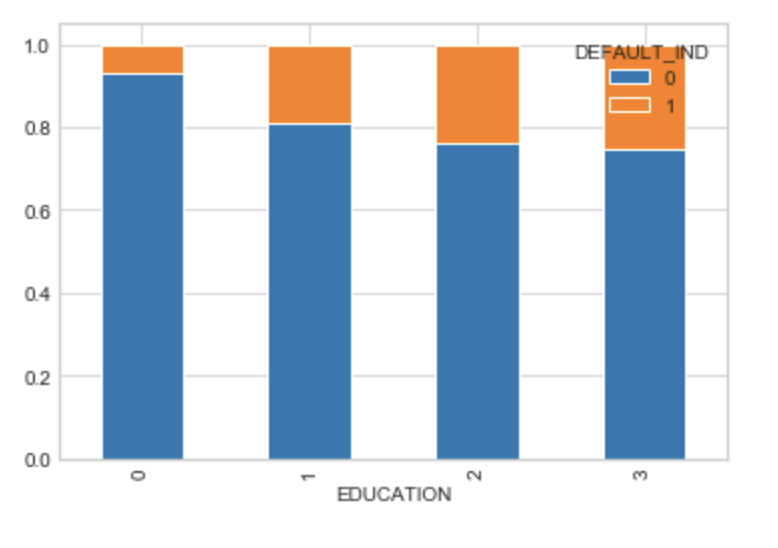
Demographics analyzed included: Gender, Age, Education Level, and Marital Status.

There were no major default status predictors in the demographics data. As you can see in the following charts, default was pretty evenly split across the various demographics available to us.





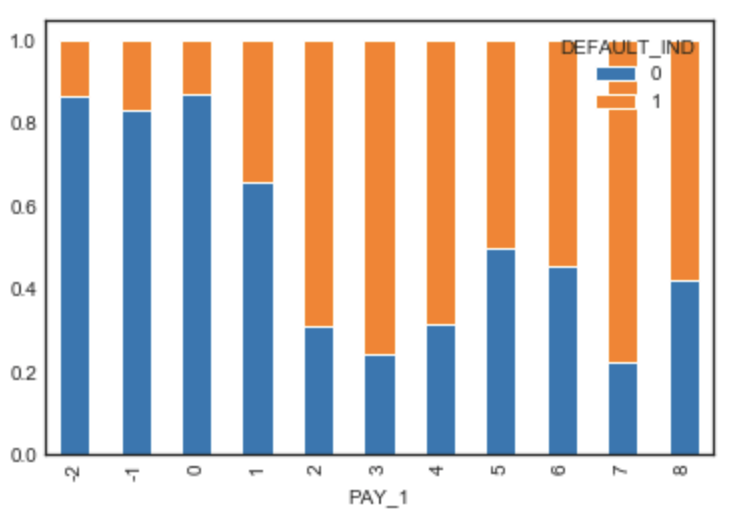




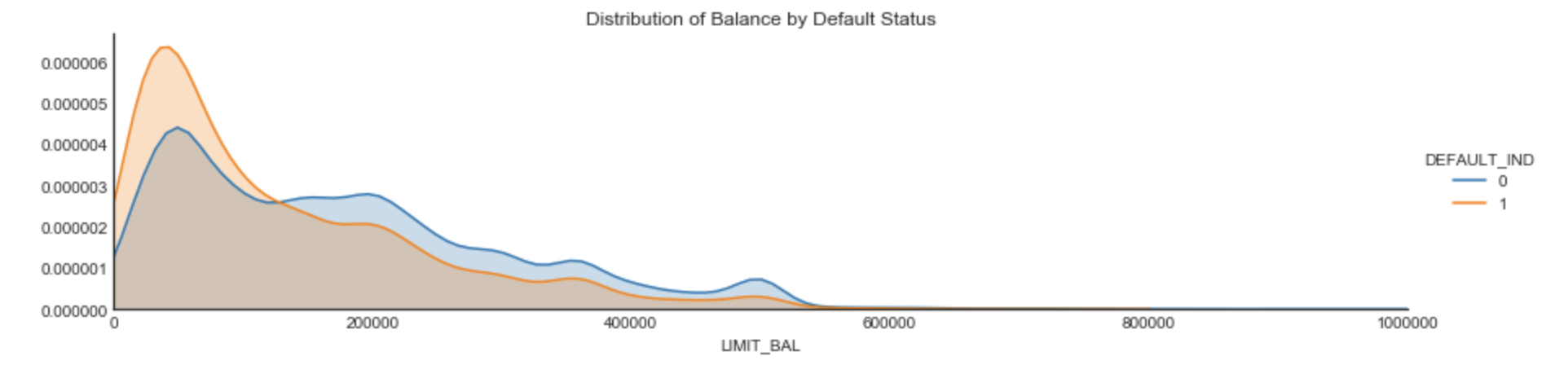
Credit Transaction History:

Unlike demographics, there were some predictors found in the credit transaction attributes. Particularly credit limit and payment status, both of which are obvious factors given some domain knowledge.

The longer the monthly bill has gone unpaid, the more likely to have a default status:

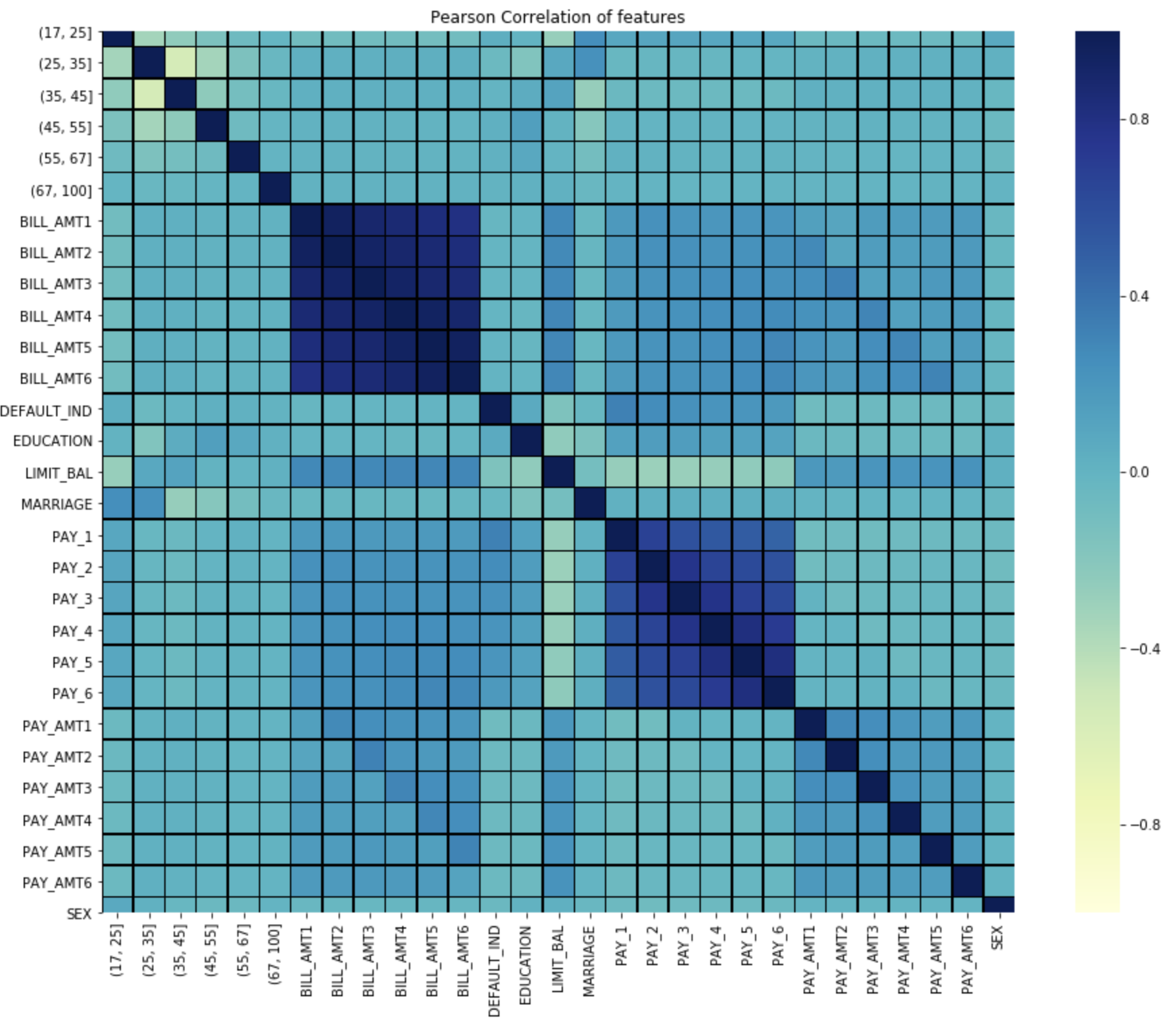


The lower the credit limit, the more likely for the account to be in default status:



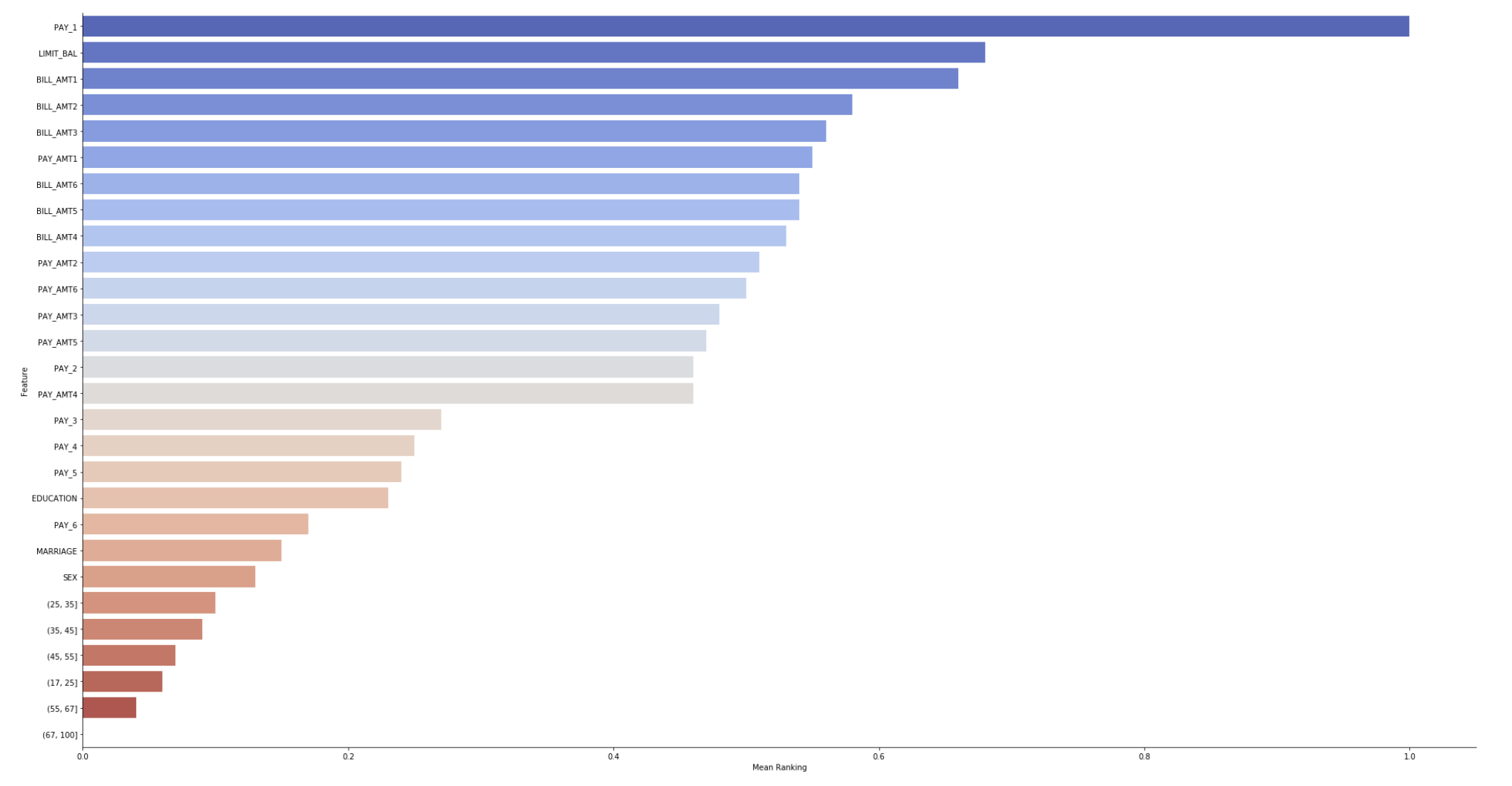
Correlation:

The correlation between features can be seen in the following heat map. There are no features correlated directly to default status (DEFAULT\_IND) but the billing amount features are all correlated to each other, as are the payment type features.



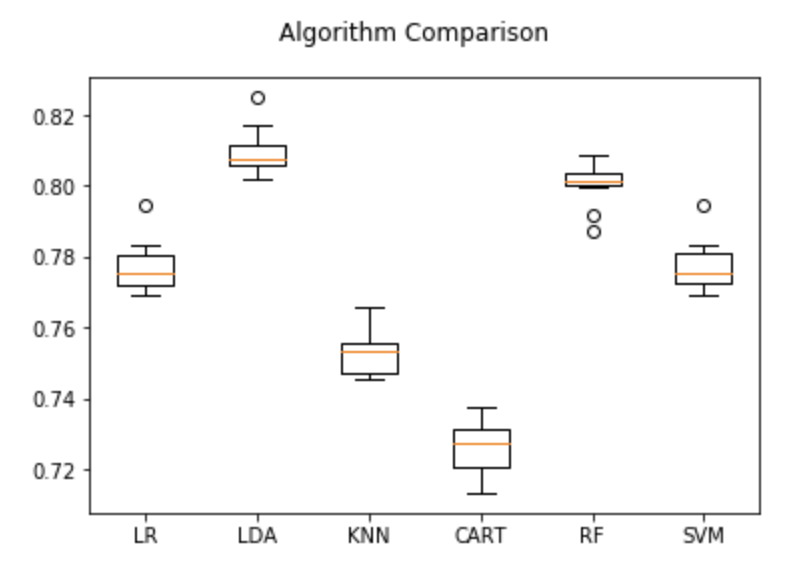
Most Important features:

The following bar chart highlights which features are most important to predicting default status, and which are the least. This confirms what we saw in our initial exploratory analysis: demographic features are not predictors whole payment status (PAY\_1) and credit limit (LIMIT\_BAL) are the strongest:



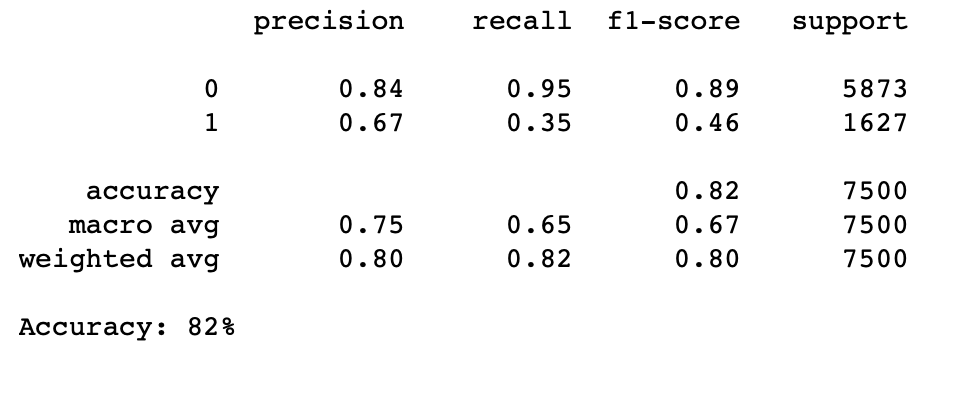
Predictive Models:

Various algorithms were tested in training classification models to predict customer default. The two best performing algorithms were Random Forest and Linear Discriminant Analysis:

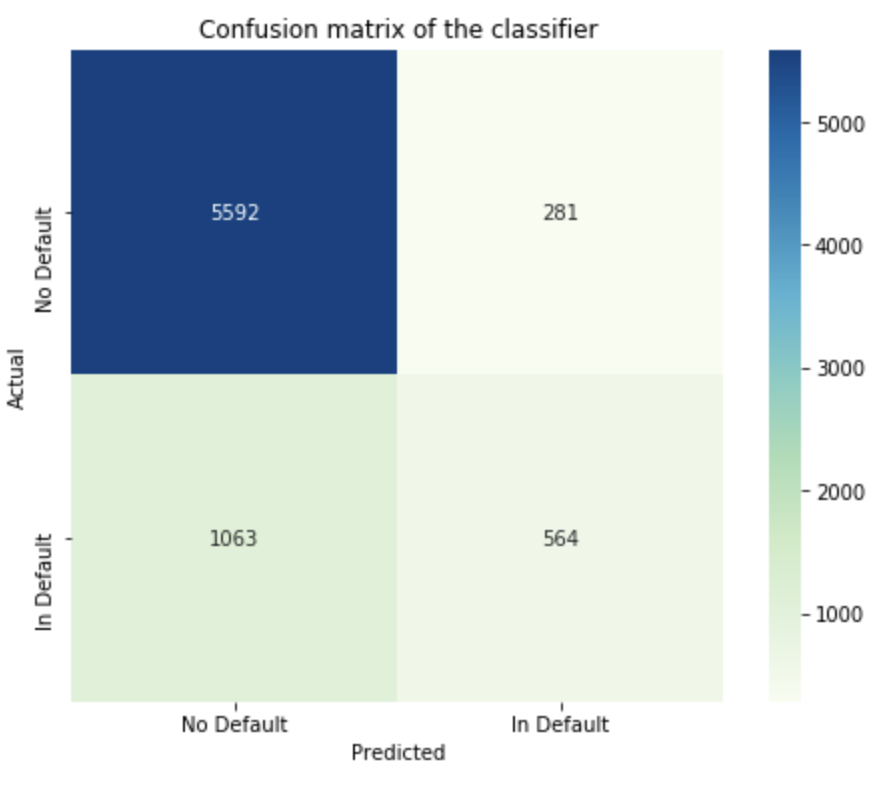


Through parameter tuning, Random Forest performed slightly better and was picked as the model to complete this analysis. Linear Discriminant Analysis had a faster run time, so if this analysis were to be performed on a much larger set of data it would likely be the preferred model.

Final **accuracy was 82%** with the following additional parameters:



A confusion matrix can be used to assess the performance of the final model on ground truth:



**Conclusions:**

While there are no big surprises in the results found, we have confirmed what was seen in initial analysis. Demographic are not a big driver in customer default. From the data available, only financial transactions data can be used as predictors and specifically credit limit and payment type/status.

Given this, there are likely other patterns in financial data that can help determine customer default. As the client has suspected about a loss in business customers leading to higher defaults, this could be true and confirmed with additional data.

As more attributes become available, they can be added into the model to improve and refine predictions.