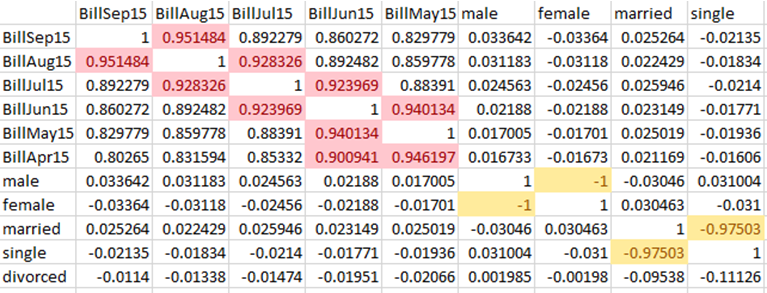
**Credit Default Risk Task 3: Build and Evaluate Models**

Credit One has seen an increase in the number of customers who have defaulted on loans. This has become an issue as Credit One is beginning to lose business. It is our primary goal to determine which customers are less likely to default through the data given which included current customer demographics and characteristics. Furthermore, our goal is to identify and predict “customer default” using specific features from historical payment, default records and customer demographics. We began this project by defining the business objectives and performing Exploratory Data Analysis.

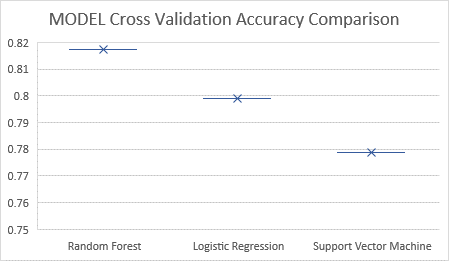
In EDA, we viewed descriptive analytics such as min, max, mean, and median of each attribute. Additionally, we discretized age group into 6 bins which were (21 - 30), (31-40), (41-50), (51-60), (61-70), (71-79). Moreover, we also reviewed interesting insights into the data itself and characteristics of customers that defaulted and did not default. Expanding on this process, we prepared and cleaned the data through transformation tasks via python. For example, we adjusted the values in the categorical features (Education, Marital Status). In Education, we noticed that values 5 and 6 were extra categories that were not necessary, so we consolidated. We did this with other attributes such as Marital Status.

Because we have categorical attributes, we needed to dummy each category through a process known as One Hot Encoder. Categorical data is mostly a constraint of the efficient implementation of machine learning algorithms. Consequently, we need to convert categorical data to numerical (through one hot encoding). The attributes we did this to were Education, Gender, and Marital Status. One Hot Encoder adds more features for each label. Moreover, we needed to drop Education, Gender, and Martial Status attributes. Next step in this process, we normalized the data. Since each column have different ranges in data, we need to normalize to put each column in equal scale.

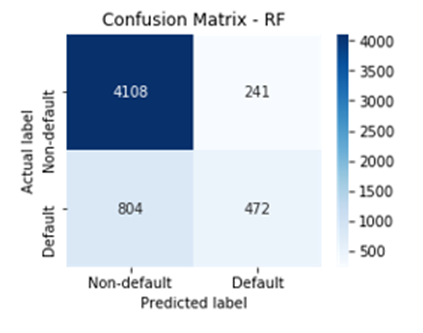
We found no high correlations between all columns and Default Payment (absolute value of .80). However, through correlation matrix, we found a few features that were highly correlated. Below is screenshot of highly correlated Independent variables:

This is also known as multicollinearity. If two independent variables contain essentially the same information to a large extent, one gains little by using both in the model. In our data set, we found 5 pairs of independent variables that were highly correlated with each other. To avoid multicollinearity, we dropped only one of the paired independent variables. Since we had 5 paired independent variables that were highly correlated, we dropped 5 features which were ‘BillMay15’, ‘BillAug15’, ‘BillJul15’, ‘female’, and ‘single’.

We built three Classification models, tuned each model and selected the best model based on their Cross-Validation Score. We then used the best model with the highest CV score for prediction. Prediction accuracy compares the accuracy of predictions to the actual values in the validation (Test) dataset. The 3 models we used were Support Vector Machine, Random Forest, and Logistic Regression. Below were the results of relative to accuracy score:



Reviewing the previous scores, we can clearly see that Random Forest perfomed the best out of all the models. The model accuracy scores were the following: Random Forest – 81.74%, Logistic Regression – 79.90%, and Support Vectore Machine – 77.87%. To further understand the importance of this accurace we can view the below confusion matrix



Random did an a relatively great job at correctly predicting Non- Default customers. In conclusion, we cannot control customer spending habits, nor can we change customers’ spending or payment habits. However, we can apply predictive analytics using the customer demographics and historical payment & default data to predict a customer’s likelihood to make payments or default their credit with reasonable accuracy.

The process we used in this task can be applied to existing and new customer base to determine if a person should be extended credit. Based on our findings, we can reduce loan limits to existing customers that are at risk of default. We can decide credit worthiness of new customers using the same features outlined in the EDA dataset.