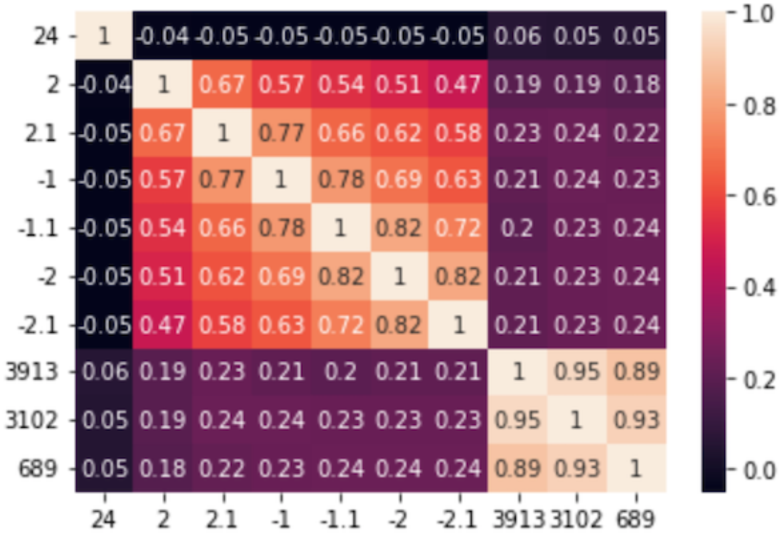
Jack Dando
Course 2 Task 3
University of Texas

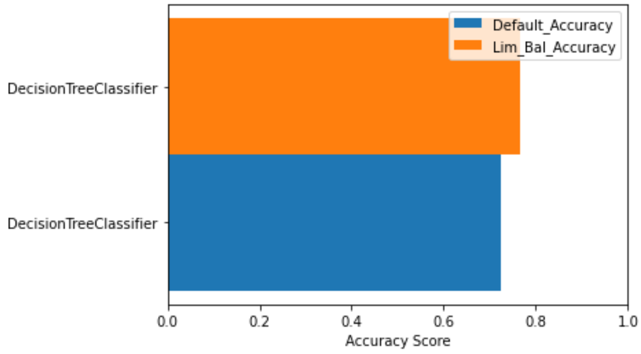
I was asked to analyze some data for Credit One in hopes of limiting the recent increase of defaulted loans from various companies using Credit One as the scoring service. I think that the best way to ensure customers can and will pay their loans is to be very cautious when issuing them credit to begin with. It might be in Credit One's best interest to do further research into finding predictors that will provide them with more accurate results than the ones provided in this data set. Based on the data provided to us, I do not believe that it is possible to approve customers with very high certainty. I think that it could be possible to get a fairly decent certainty at some point, but the data provided to us is simply not going to get us there. The two features we used as dependent variables provided with some valuable insights, however there are other potential features not provided to us that could possibly be used to develop a more useful model for determining whether or not we issue someone credit. Some potential features to add to the dataset could be the length of persons credit history instead of age, an average of the 6 payment/billing history features so they are easier to analyze. The best thing I can think of would be to develop some kind of formula that finds an average percentage of credit usage over the course of a year based on the billing and payment history, which would be over 100% in cases where the customer wasn't very good about making their payments on time. Another way to further tweak and increase the accuracy of our model would be to increase the number of bins for the dependent variable.

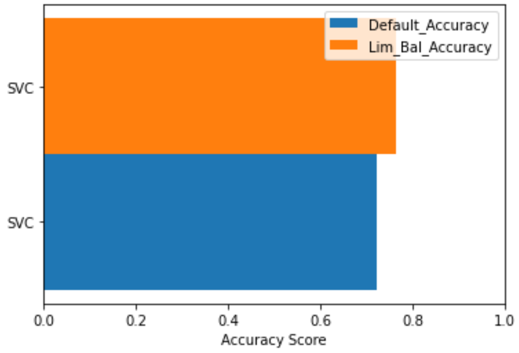
In data pre-processing, finding out if there's a high correlation between the predictor variables are important to avoid the multicollinearity. Multicollinearity is observed when two or more independent predictors are highly correlated to each other in a multiple regression model. Multicollinearity is a problem in regression model since it undermines the statistical significance of an independent predictor.



In the plot, we can see strong correlations with limit balance between several attributes, such as between PAY_4 and PAY_5. Highly correlated variables (>0.7) should be removed as predictors in the regression models.

We attempted classification machine learning with limit balance as the dependent and classification machine learning with default as the dependent. For classification with limit balance, we divided the limit balance into 2 classes. Then we used SVC and a decision tree classifier on the limit balance classes. I got 76% accuracy on the test dataset with SVC as well as a decision tree classifier. For classification with default, I got an accuracy of 72% for both decision tree and SVC. The plots are shown below.


Default_Accuracy
Lim_Bal_Accuracy
DecisionTreeClassifier
DecisionTreeClassifier
Accuracy Score


Default_Accuracy
Lim_Bal_Accuracy
SVC
SVC
Accuracy Score