Kahlil Cole

Credit Card Fraud

For my machine learning project, I decided to do a credit card fraud analysis. The data set that I used is from Kaggle. The data was pulled from credit card transactions made in September 2013 by European card holders. The variables are V1-V28, Time, Amount, and Class. According to the data set description V1- V28 had confidential names so they got changed. Time is the seconds between the transactions in the data set starting with the first transaction. Amount is the transaction amount. Class has a value of 1 for fraudulent transaction and 0 for normal transaction. I know next time to pick a data set that does not have confidential variables in it so that I can work with the data more.

This project was done in python primarily using the pandas, seaborn and scikit-learn packages. Through my analysis I found that average real transaction was $88.29, and the average fraud transaction was $122.12. After that I set up a random forest classifier to predict the accuracy of the model picking authentic transactions. In this data set the amount of real to fraud transactions is highly unbalanced. The data has 492 frauds out of 284,807 transactions. The frauds account for 0.172% of all transactions. That means what ever model you run on the data is going to be extremely accurate. For the RFC I split the data into a 70% training and 30% test split to evaluate the data. When using the RFC I was looking for the accuracy score, precision, and recall score. The Accuracy score measures how many labels the model got right out of the total number of predictions. The Precision score is the number of correctly-identified members of a class divided by all the times the model predicted that class. The Recall is the number of members of a class that the classifier identified correctly divided by the total number of members in that class.

For the results the accuracy of the model was 0.999 which is extremely accurate at predicting real transactions. However, since the data is highly unbalanced in a real world setting you would not want to use the model. The precision score was 0.948 and the recall had the most realistic score or 0.808. Looking back on the project I would go with a different data set so I could use some different packages and have more variables to play with.