Michael Loos

Credit Card Fraud Detection

In 2019, credit card issuers lost $28.65 billion due to fraud, despite large efforts to prevent fraudulent accounts and transactions. Being able to identify and reduce fraud can lead to significant reductions in losses for credit card issuers and prevent customers from dealing with the hassle of canceling their credit card and waiting for a new one to be issued. I work in the credit card industry and while I am not directly responsible for identifying fraudulent transactions, they do impact my job. The goal of this case study is to be able to use cardholder and transaction data to train models to predict which transactions are fraudulent and which are not. The models will then be applied on test data to determine the accuracy of the models. Variables in the data include transaction amount, transaction location, merchant type, and several cardholder characteristics such as gender, age, and location.

Data set: <https://www.kaggle.com/kartik2112/fraud-detection>

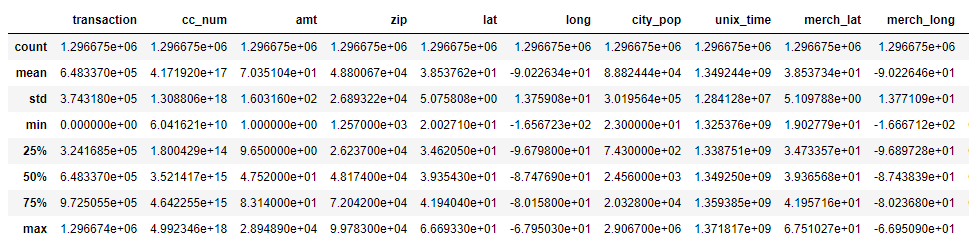
**Questions to answer:**

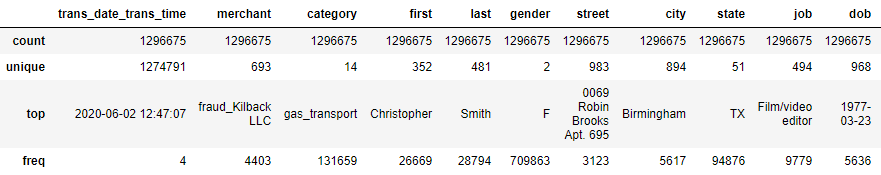
* What do the variables look like?
* Does merchant or merchant category correlate with fraud?
* Are there high risk areas? Does distance from home to the merchant matter?
* Do personal traits like gender, age, or location matter?
* How does spend for each person impact fraud?
* What impact does transaction frequency have?

**1. Data Exploration**

**Variables:**

* Transaction: Transaction number
* trans\_date\_time: Transaction date and time
* CC\_Num: Credit Card Number
* Merchant: Merchant
* Category: Merchant category
* Amt: Transaction amount
* First: Cardholder first name
* Last: Cardholder last name
* Gender: Cardholder gender
* Street: Cardholder street
* City: Cardholder city
* State: Cardholder state
* Zip: Cardholder zip code
* Lat: Cardholder Latitude
* Long: Cardholder Longitude
* City\_pop: Cardholder City Population.
* Job: Cardholder job
* Dob: Cardholder date of birth
* Unix\_time: Unix Time
* Merch\_lat: Merchant Latitude
* Merch\_long: Merchant Longitude
* Is\_fraud: Fraud indicator, 1 if fraud, 0 if not

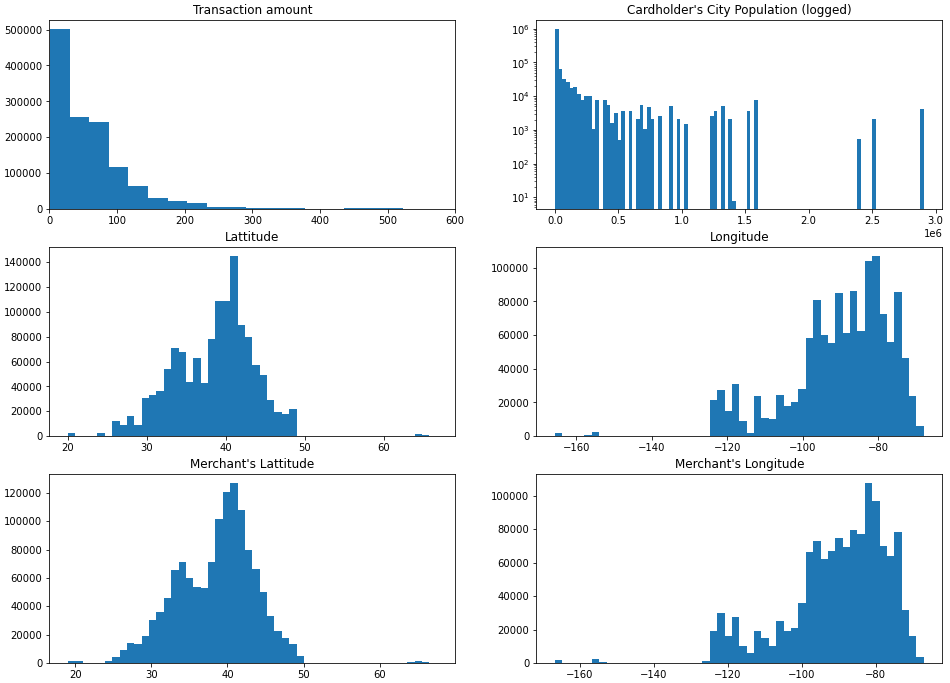


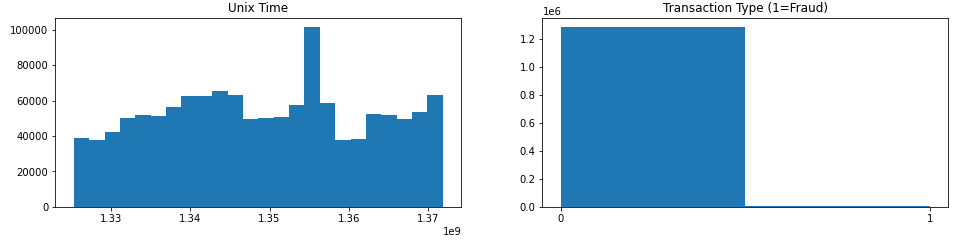


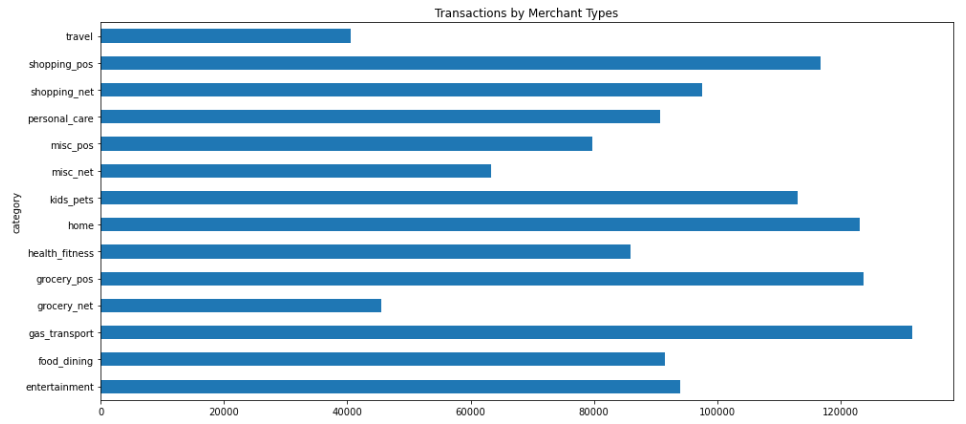
  
Observations:

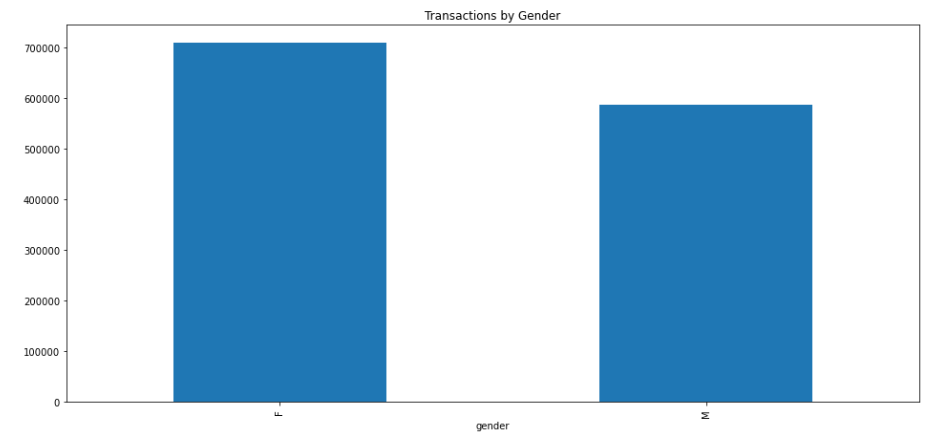
Less than 1% of transactions in the data set are fraud

**Graphs**



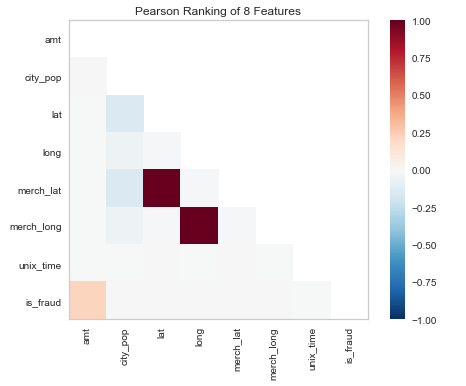






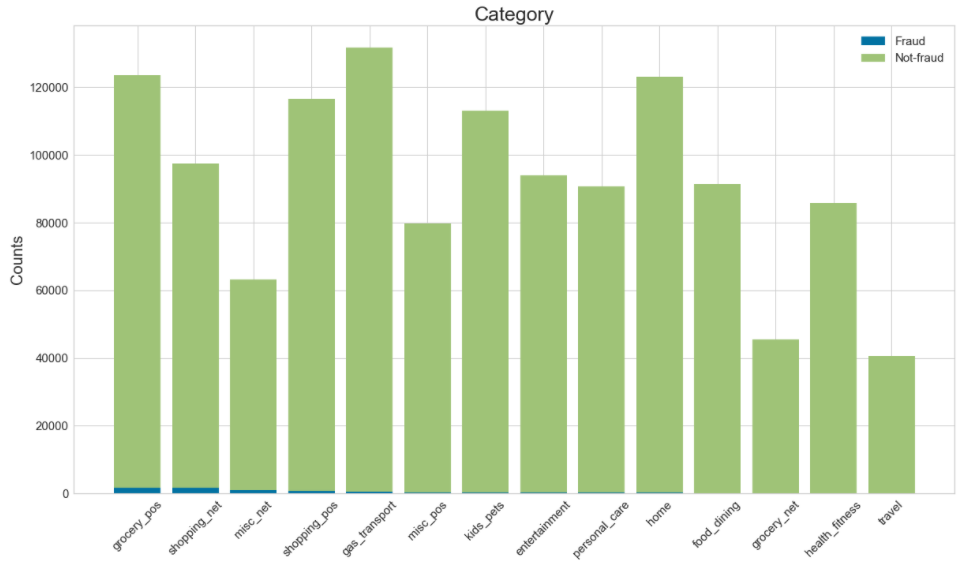
Observations:

a. Transactions skew to relatively low values. Are fraud transactions similar?  
b. Merchant locations appear to be similar to cardholder locations



Observations:

a. The biggest correlations are between location, which makes sense  
b. Population and location have negative correlations (although these variables are more ordinal)  
c. Fraud and transaction amount are slightly correlated (although the fraud variable is more categorical here)





Observations:

Because of relatively low fraud volume, it is difficult to tell but it seems a couple categories may be disproportionate.

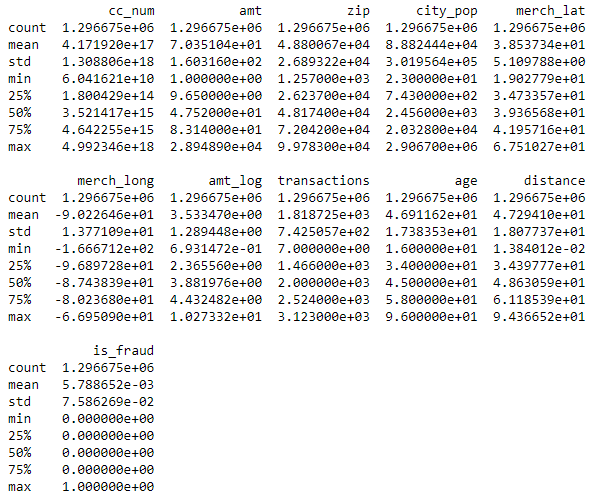
2. **Dimensionality Reduction   
  
Converting Variable Types and Feature Reduction**

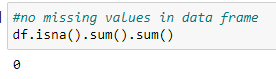
a. Create variable for log of transaction amount  
b. Convert date of birth (dob) to age

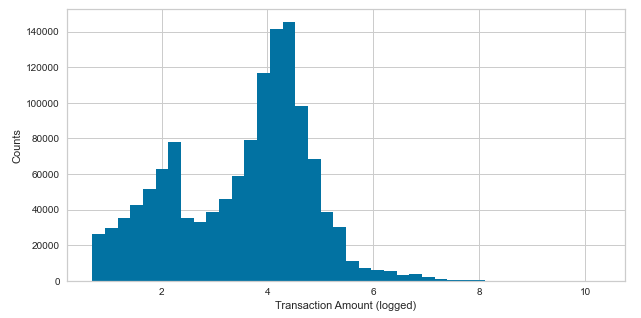
c. Calculate distance beetween cardholder home and merchant location

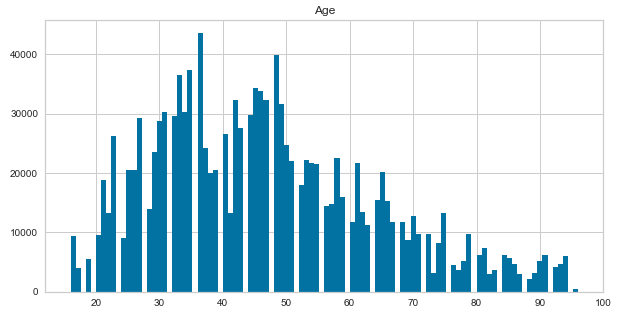
d. Calculate total transactions per card  
e. Drop variables: transaction, merchant, first, last, street, job, dob, unix\_time, merchant

f. Check for missing data







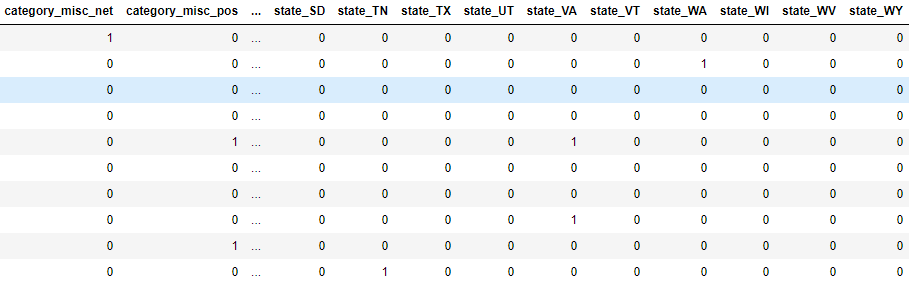
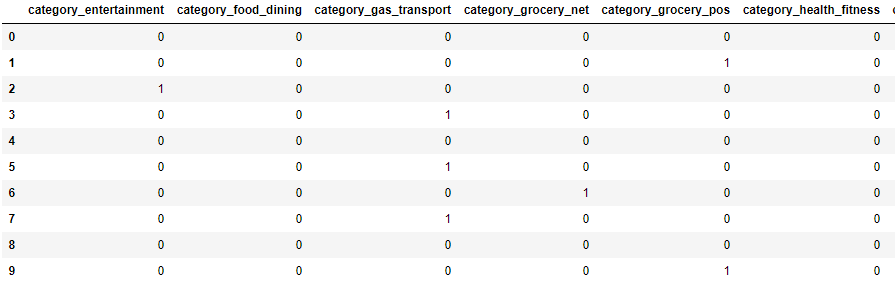


Observations:

a. log of transaction amount (log\_amt) shows a much more readable graph with less skew than the original graph in part 1  
b. Age is much easier to interpret than date of birth. Data appears to have positive skew.  
c. There is no missing data that needs to be accounted for or filled in

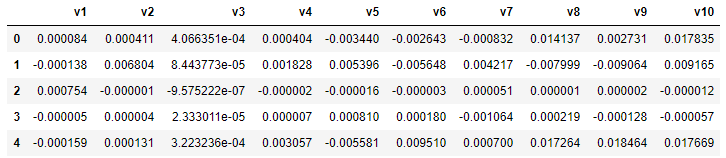
**One Hot Encoding**

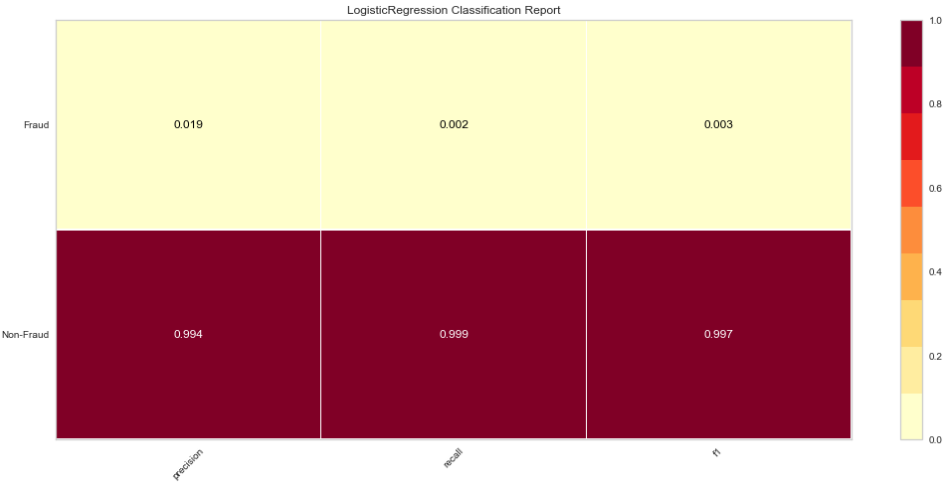
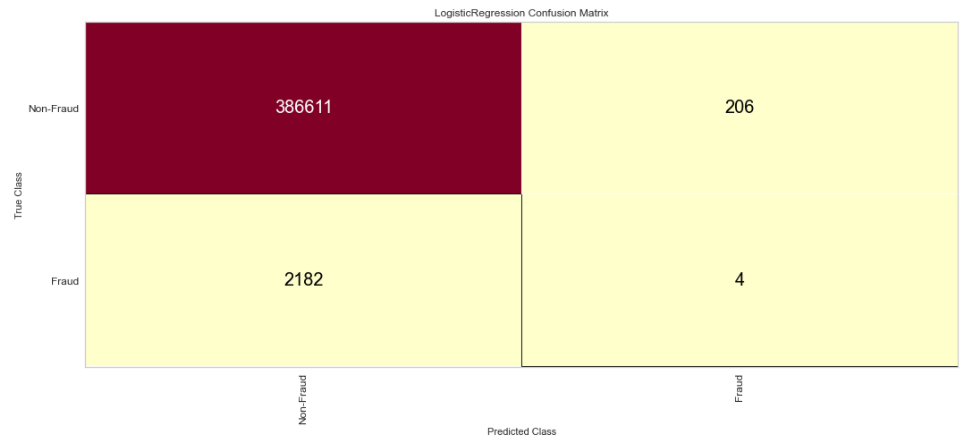
Convert categorical data into numbers; variables: merchant category, gender, city, state



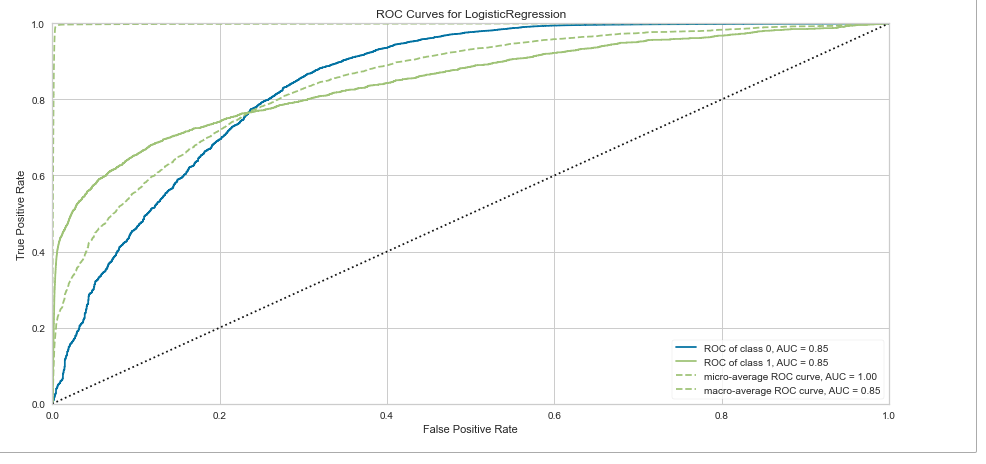
Observations:

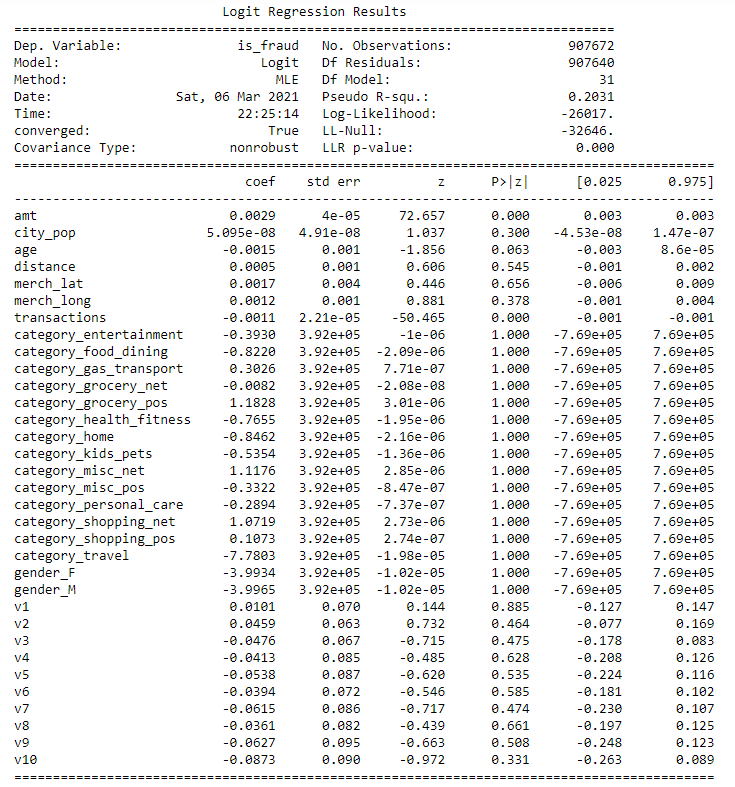
a. While this may prove useful, the data set just became a lot larger. Using state may also be a bit redundant since I already have zip code which could be rolled up to a state level.

**Feature Reduction using Truncated Singular Value Decomposition**Using TSVD, city and state reduced from 945 features to 10

**3. Modeling**

**Logistic Regression**

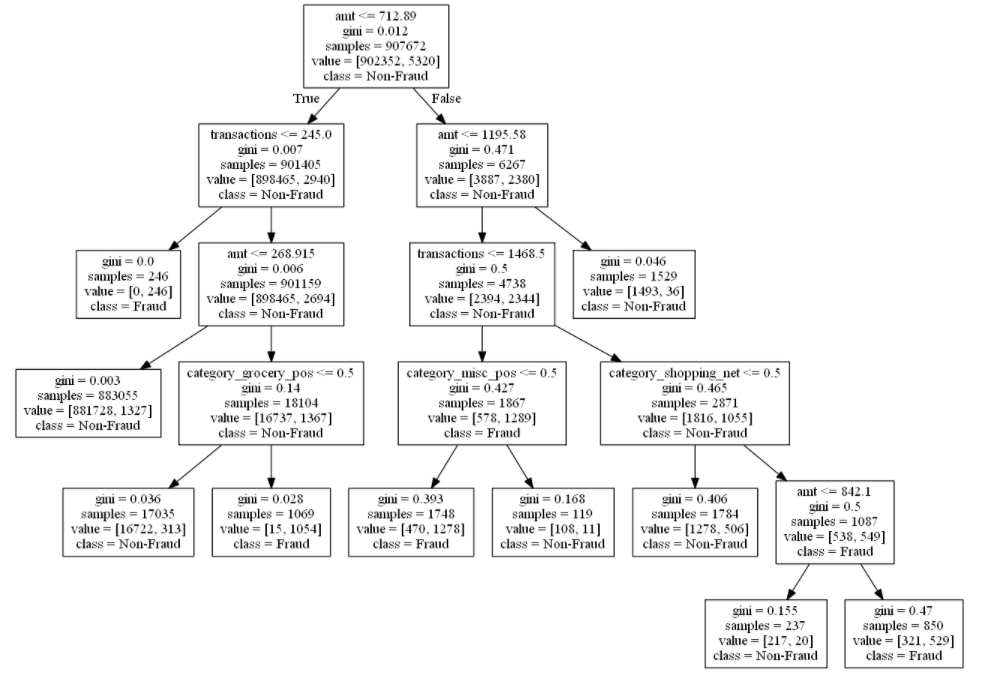


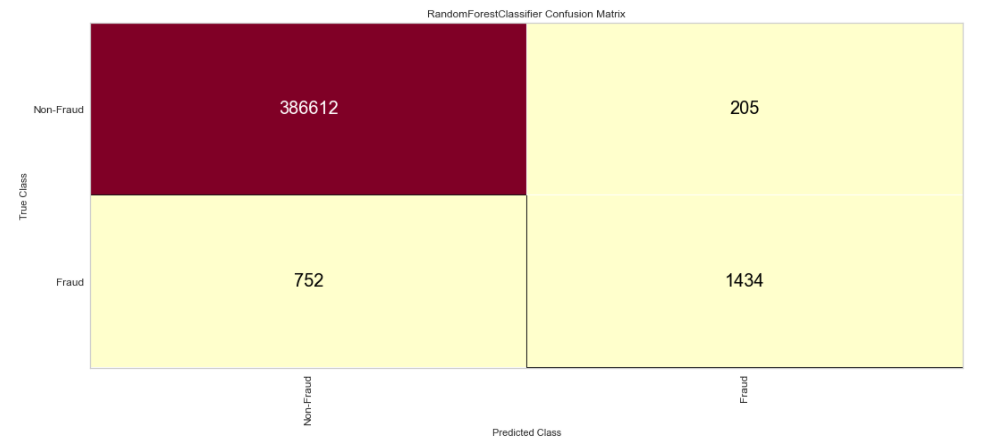
  
Observations:

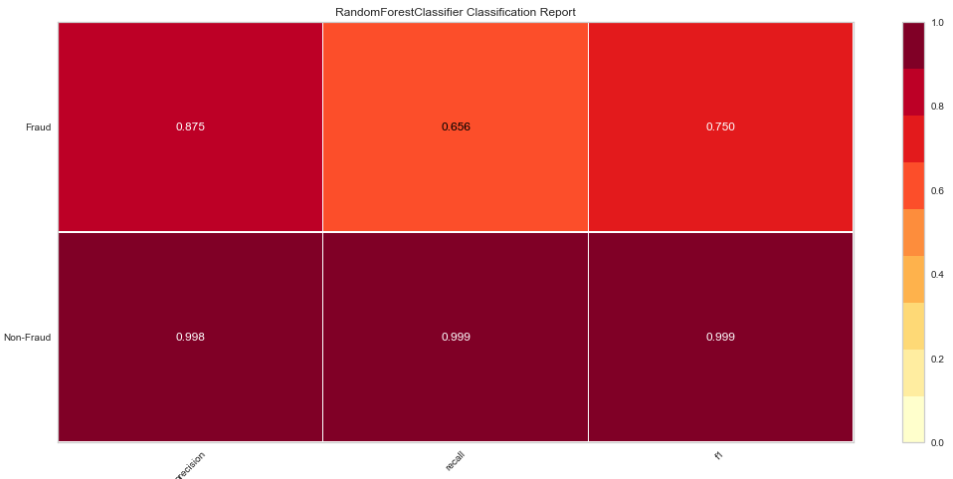
a. Obviously this model did not perform well. If you look at the confusion matrix, none of the Fraud transactions were predicted as Fraud and Several Non-Fraud transactions were inaccurately predicted as Fraud.

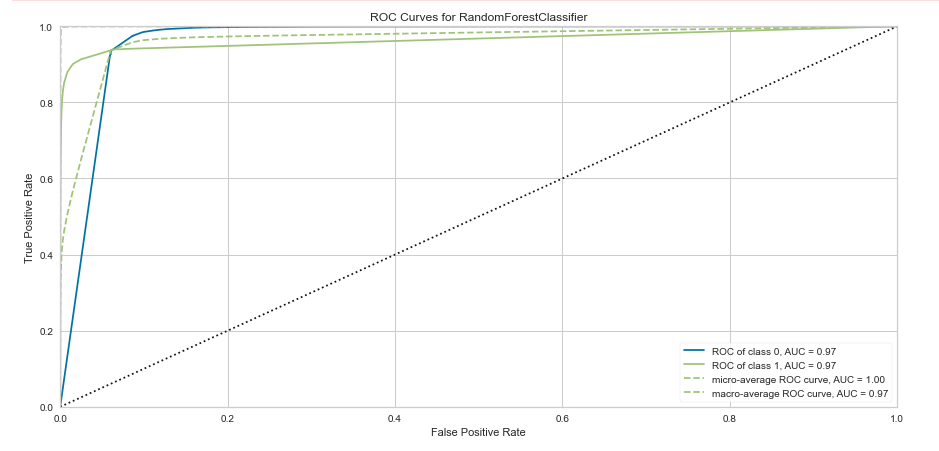
b. I would have assumed the opposite to have happened due to over-fitting of the model.

c. Transaction Amount and number of transactions are the only significant variables at a 0.05 significance level.

**Decision Tree**





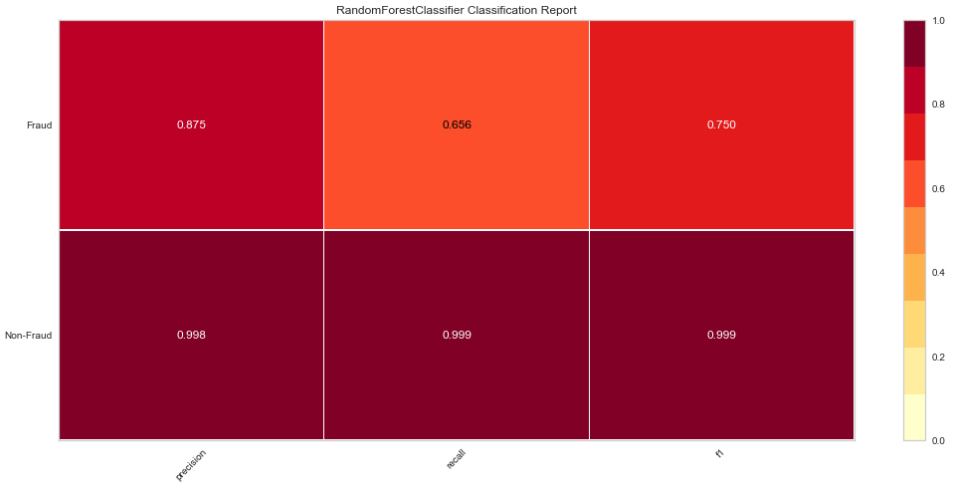
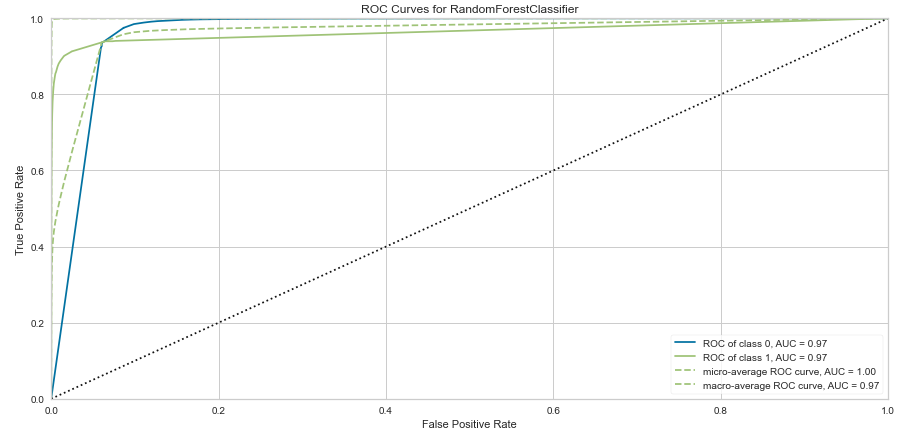
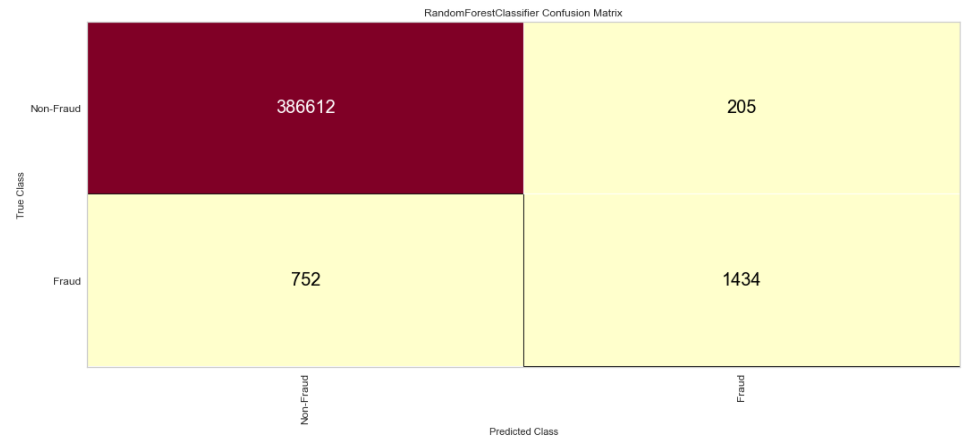
  
  
Observations:

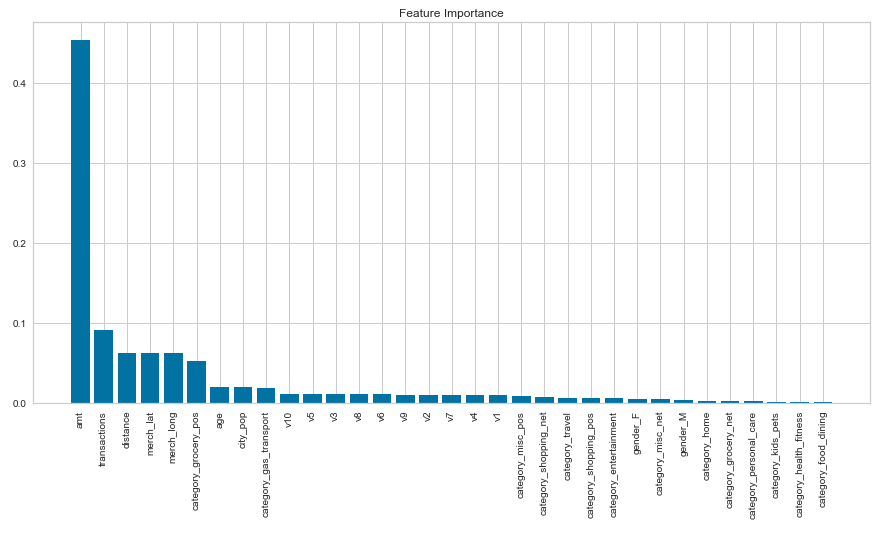
a. While still not great, the decision tree does much better at predicting Fraud, although it inaccurately predicts Fraud on more Non-Fraud transactions than the logistic classifier.

b. With enough depth and leaves, we can probably have our decision tree be perfect.

c. Amount and number of transactions are biggest splits within the decision tree. Man fraud cases are non-grocery transactions over $268.91 and online shopping transactions over $842.10.

**Random Forrest**





Observations:  
a. The random forest performed better than the logistic regression and decision tree models in every metric.

b. As with the logistic regression and decision tree, transaction amount and the total transactions on a credit card are the most important features, while distance from the cardholder’s home, merchant location, and non-grocery spending contribute slightly.

**Conclusion**

The random forest proved to be the best model and has a high accuracy due to the large number of correctly predicted non-fraud transactions, but still produced a high number of false positives. 14% of predicted fraud transactions were false positives. While we would rather over predict fraud than under predict in order to eliminate fraud losses, too many false positives would cause inconveniences to customers as they would either have to confirm their transactions or unlock their card if a stop was put on it. Only 64.5% of of fraud transactions were correctly identified as fraud. This would still leave a company exposed to a lot of potential fraud. If we assumed just $1 billion in transactions and 1% of those transactions were fraudulent, the fraud losses would be just over $3.55 million. While our models were far from perfect, it seems the amount of a transaction and the frequency transactions are made are the best predictors of fraud. Location, distance, and merchant type also contribute, but at a smaller level. These would be the features to build around in further model development to prevent fraudulent credit card transactions.