**IBM Data Science Capstone - Credit Card Fraud**

1. **Introduction:**

Everyday hundreds of people are victims of fraudulent credit card transactions across the united states. Over 47% of the worlds credit card fraud happens in the united states alone.

Credit card fraud is one of the biggest issues that our banking systems and government entities have to deal that not only effect them but us a hard working people. I personally have been effected by credit card fraud; Therefore, I understand the impact and headache it has on a person. Not only is it extremely frustrating knowing someone has access to your money but can also ruin your financial situation at anytime.

The ability to be able to defect fraudulent activity on someone credit card within minutes can literally save the card holder hundreds to even thousands of dollars.

1. **Purpose & Audience:**

The purpose of this project is to show how having machine learning algorithms in banking institutions can help protect banking clients from financial theft.

The indented audience is banking institutions which do not currently have machine learning algorithms set up to help detect fraud on the behave of clients.

1. **Data:**

**3.1**

The data being used in this project consist of credit card transactions in September 2013 by European cardholders. The data set is two days worth of transactions for a total of 284,807. Out of the data set, there are approximately 492 fraudulent transactions in the data. The data will be used to create a machine learning model to accurately identify fraud on credit cards.

**3.2**

The data gathered from here: <https://www.kaggle.com/mlg-ulb/creditcardfraud>

1. **Methodology:**

**4.1**

To start, the data set was already very clean so didn’t need to do any data wrangling.

The feature ‘Amount’ is the transaction and amount which was used for the dependant cost. The feature ‘Class’ was the independent variable and took the values of either 1 for fraud or 0 for valid transaction.

**4.2**

Next, visualized all the features of the data set on to graphs to see the distribution of transactions. Due to the length of graphs, please reference the notebook for the visualizations.

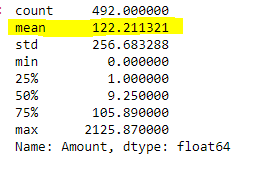
**4.3**

I then determined to separated the valid transaction cases vs the fraud transactions to find out that only 0.17% resulted in fraud transactions. Then did a comparison operation to to see the average cost of fraud transactions to valid transactions and found that the average fraud transaction was a higher amount compared to valid transactions from the data set.

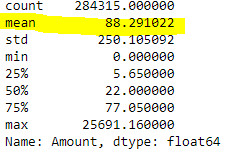
Transaction distribution in data set:



Avg Fraud Transaction:



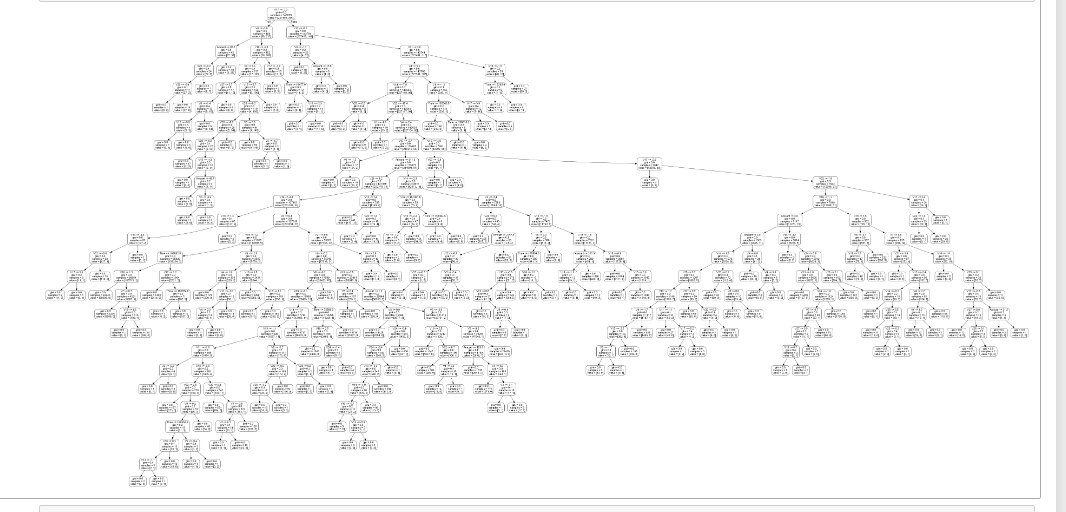
Avg Valid Transaction:



**4.4**

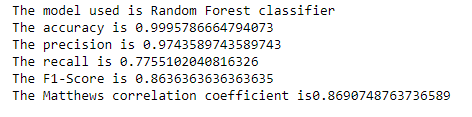
In order to train a machine learning model next, I had to split the data in to Feature variables and the target variable. Please reference the note book.

The machine learning model I used is called a Random Forest Model from scikit-learn python library and is a classification model.



**4.5**

Evaluation then was taken place to have a result of the following:



1. **Results and Conclusion:**

The results of the classification model show why when we make transactions we are often asked to verify our identify due to a low recall value of 77% and that we actually determine fraud activity with a high amount of accuracy but as more and more data becomes available we will be able to more accurately detect fraud. Naturally its safe to have a 2nd way of verification on transactions to bridge the gap between fraud and valid credit card transactions.