Credit Card Fraud Detection

*Using Artificial Intelligence and Clouding Platforms*

Write your Name here

ABSTRACT

*In these times of where we have information and computational power, both scaling up like never before, it is pertinent to find solution of the former in the latter.* As we move towards digitalization, we observe that processes that were once done manually are no more bearable to do without taking help of modern technological measures.  *Particularly in the cases of anomaly detection where computation, in the form of machine learning and deep learning predictions has proven its mettle in finding those odd ones out through the recognized patterns from the loads of data. In our case where we are out for finding suspicious one out of all the activities can be seen as anomaly detection. It is no less exciting to see how humungous datasets with fleeting pace of information that can not be conquered through ordinary measures, are tackled with cloud computing and artificial intelligence.*

# Introduction

Data Science is much about learning from the patterns and blueprints. The trails left by an activity have the potential to help an investigator understand the intent of the action. Hence, it cannot be denied that a fraudulent activity does not leave something different at all than an execution that had a fair intention to take place. In this report, we shall try to bring out those hidden patterns that could be decisive in separating those activities. With a large and identified records of transactions, a study of the dataset of such transactions can help us bring out what trends could be there that can be used to predict a potentially fraudulent activity. For predictions, we have numerous algorithms that can be used and even tailored to fit on what results we already have, learn the patterns at depth, and make a prediction of an upcoming transaction with a very practical, and at times, astounding accuracy. This can help prevent the disaster for all the parties involved.

Problem Statement:

It is highly important that stake holders and companies are much more geared up against the menace of credit card frauds more than ever. According to reports, the frauds has been doubling between the years of 2019 and 2020. 2021 saw no less than 1.7 million people being impacted by the fraudulent activities in the U.S. According to an article published at [www.precisesecurity.com](http://www.precisesecurity.com), Americans reported 271,823 cases of credit card fraud in 2019. This is an increase of 72.4 percent from 2018, when there were 157,715 cases of credit card fraud reported. Losses from fraud involving cards used for payment worldwide reached $27.85 billion in 2018. They are projected to rise to $35.67 billion in five years and $40.63 billion in 10 years, forecasted by www.prnewswire.com.

# Data

The dataset used for this report is generated by Brandon Harris using Sparkov Data Generation. It consists of over 1.8 million transactions taken the years of 2019 and 2020. The transactions are done from nearly 1000 different customers with around 700 merchants. The transactions involved sums up to more than 129 million. The dataset comes with around 23 different attributes of users. The attributes include information about the customer, the purchase made, merchant, and the transaction of money. For customer, information such as its first and last name, customer’s unique number, date of birth, occupation, and gender and address is provided. For merchant, its name, address such as geographical coordinates, zip, city, state is given as well. The dataset includes information about the transaction such as the time and place of transaction, its unique identification number, and the amount involved. For product, we have its name, and the category it belongs to. Finally, we have a target variable called “is\_fraud” that contains ground truth label about whether the transaction was fraud or not. In this dataset of 1.8 million, we have 1,842,743 transactions as fair and a relative tinier number of 9,651 number of fraud transactions.

## Data Preparation

The dataset was remarkably clean in terms of null/void, missing, corrupt, values. The outliers had to be respected as the purpose is of the exercise is to perform fraud detection, the problem that has been framed as an anomaly detection. The columns containing date/time attributes were fixed back by assigning the corresponding format. There was also some cleaning performed such as converting the strings (words) into numeric values wherever possible and dismantling the values into units to let model capture the patterns better.

## Initial Analysis

During the analysis, some patterns emerged even before going any further. It was found that the fraudulent transactions are conducted between a range of some 200 USDs to 900 US dollars. The mean value of the fraud amount is a little less than 400 US dollars. Looking at the full picture, the actual range of transactions we have ranges from a few pennies to about 35000 US dollars. It was obvious that the fraudsters were meticulous with selecting the amount they are going to slap their victims with. Apart from the amount they chose to target with, the fraudsters were also found quite picky with the categories of goods they looted the customers with. Most of the products were online and point of sale shopping, groceries, and miscellaneous items. The victims of these credit card frauds were mostly Teachers, Lawyers, Community workers, horticultural consultants, medical staff, writers, and very surprisingly, charted accountants. Demographically, most prone to get fraud states were New York, Pennsylvania, Texas, and California. At city level, New York, Dallas, Beaver Falls, Glendale, Tulsa, and Huntsville were among the most dangerous cities. Gender wasn’t very decisive. Another pattern showed up in time series where frauds seemed to happen more in the late hours. Whereas if we examine it through out the year, frauds were much likely to happen in the months of February and March.

## Pre-Processing Data for Machine Learning

Finally, we had most of the things we could do with manual inspection, and we had to move on to the data preparation for modelling. As the principle of machine learning, said as “Garbage in, Garbage out”, the features had to be chosen carefully rather than going with all or a random selection. Once we have arrived at the stage of using algorithms, the assumptions of the algorithms must be taken care of. For an instance, Linear models do not go with the features that happen to have multi collinearity. Some algorithms have an assumption that the features are independent of each other. Therefore, these aspects were must to be taken care of.

Before even getting to know how well each feature could contribute, we had to translate the strings into numbers that machine could understand. This led to multiple options. However, for the data types we had in our string columns (and it was nominal), only one hot encoding seemed to be the right choice for the translation. Doing One hot encoding can increase the dimensionality exponentially. Therefore, I began to chose features based on how expensive their expansion could be. We still had numeric features, which were easy to be assessed with Pearson’s correlation. The features that could be correlated to each other were the amount of transaction, the population of the city where the transaction took place, age and gender of the customer, and date, day, month, year, hour, and day of the week of transaction. It was found that the features had very minor correlation with the target, except for the amount of transaction that happened to share around 0.21 value with the target variable. Other variables shown little to no correlation at all.

The features that I chose for One hot encoding were category of the bought product and the State. The reason to pick State was rather about a compromise. Most of the categorical features had a large number of categories. For example, there are around 500 occupations titles and around 700 merchants. Creating dummy variables (One Hot Encoding) and training the model for these could be a very expensive process. Therefore with 14 categories and some 50 states, we had 64 additional columns. So altogether, we have a dataset with 72 features, including the target variable.

## Using Artificial Intelligence, Machine Learning for Prediction:

One problem that we would be facing is that our dataset is extremely unbalanced. This is how classification like detections such as credit card frauds or even cancer is how supposed to be. Simple machine learning techniques work much better at balanced datasets. One of the major traps to avoid doing unbalanced classification is not to be fooled by accuracy metrics. For example, if we have 1 fraud case out of 100 cases and the model we train learns and predicts every sample as a fair transaction, we will still be having a 99% accuracy! Hence, a novice observer will quick fall into this trap. There ware many ways to deal with this including over sampling, under sampling, and choosing the right performance metric for evaluation. To begin, we will perform under-sampling by choosing the same number of fair transactions as we had fraud transactions in the dataset. So, for the purpose of training the model, we have 9651 number of fair and similar number of fraud transactions, making our dataset synthetically balanced. This also allowed us to run cross validation cycles for validation purposes across the dataset. The error metric chosen is recall, rather than overall accuracy. Recall is the measure of the ratio between the number of positive samples correctly classified as positive to the total number of positive samples. Recall tells about model’s ability to bring out identify the samples correctly we are interested in. We will also use precision, that tells how much positive class samples there in all the classified ones are.

## Models and Performances

To being with algorithms and model training, I started with Logistic Regression. To feed in my data in the model, the data had to be scaled. Therefore, feature scaling was performed with Standard Scaler. Next in the pipeline is the algorithm of logistic regression itself. As we have much smaller dataset for now, we can utilize with KFold Cross Validation techniques to overcome any discrepancies that usually happen while splitting the small datasets. The recall score was not impressive at all, lurking around somewhere near 75%, it meant that out only 75% of the fraudulent transactions were identified. The sensitivity of the problem statement demands that the recall should not be any less than 95%.

The desirable results could be envisioned as we want at least 95% of the fraudulent transactions to be picked up. Not only this but we do not mind fair transactions to be picked up as fraudulent ones. In fact, I would go as far as saying that we might welcome those as it could be safely assumed that they might followed similar patterns to the actual fraudulent ones.

Our next algorithm pick was Random Forest Classifier. It is a tree algorithm, meaning that it does not cater the distances among the data points. Hence, there are so such assumptions that the data is scaled down. With the same preprocessing configurations except for scaling features, random forest classifier was attempted on the balanced dataset. The results have been tremendous, with a recall of around 96%. This was exactly we were looking for.

Another algorithm that has been topping at most of the machine learning forums is XG Boost Classifier. Similar to the random forest algorithm, it is decision tree-based ensemble machine learning algorithm that uses gradient boosting to make predictions. Again, using the same pre-processing configuration, XG Boost classifier was modelled. The results were again exceptional, with a recall of somewhere around 95%. It was a bit lesser than the results of Random Forests. However, the difference between both performances is somewhat negligible.

Having done this all, our next step will be to create a web app with flask platform. The flask application ensures that model and predictions can be accessed with any browser, like any other web application.

Lastly, the web application can be deployed on any cloud platform like Amazon Web Services or Google Cloud Platform. I have chosen google cloud platform to deploy this web app.

## Critical Analysis

There could be many ways to make this even better in many aspects. Firstly, the data we had wasn’t very clear. More clarity with data can result in much better feature selection, hence improved model building. Secondly, the dataset did not have the quality features. Weak correlations were exposed during analysis. If there is a universal consensus on using Artificial Intelligence technology for credit card fraud detection, such information that could be useful may be gathered and made central to an authority. There would be legal constraints. But, in the world that is being rapidly changing with technological advances, we must be better prepared with our security than the ones who are up to violate it.

References:

1. <https://www.self.inc/info/credit-card-fraud-statistics/>
2. <https://www.fool.com/the-ascent/research/identity-theft-credit-card-fraud-statistics/#:~:text=Nearly%2084%2C000%20Americans%20reported%20new,account%20fraud%20increased%20by%205%25>.
3. <https://moneytransfers.com/news/content/credit-card-fraud-statistics>
4. <https://www.cardrates.com/advice/credit-card-fraud-statistics/>
5. <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>
6. <https://www.chase.com/personal/credit-cards/education/basics/credit-card-safety#:~:text=Credit%20card%20safety%20tips%3A&text=Don't%20share%20account%20information,Report%20fraud%20ASAP>
7. <https://www.sbp.org.pk/cpd/pdf/learning/A%20Guide%20to%20Use%20Credit%20Cards.pdf>