Credit Card Fraud Detection Using Machine Learning

Anil Kumar Boggavarapu
Masters in Computer Science
University of Central Missouri
Leesummit, Missouri, USA
axb16470@ucmo.edu

Satya Ishyanth Kadali
Masters in Computer Science
University of Central Missouri
Leesummit, Missouri, USA
sxk55130@ucmo.edu

Pranav Garipelli
Masters in Computer Science
University of Central Missouri
Leesummit, Missouri, USA
pxg14880@ucmo.edu

Abstract— **The amount of credit card fraud has grown significantly over the years. Due to the growing volume of daily digital trades that make credit cards more vulnerable to fraud, they are the most commonly used digital instalment method. Credit card organizations are looking for the best innovations and circumstances to identify and reduce exchange extortion at the credit score card. The objective is to identify specific instances of false extortion.**

Neural groups, genetic algorithms, and k-manner bunching are some of the structures that can be used to identify credit card falsification. Over $4 trillion has been generated globally via persistent misrepresentation within the economy. The solution is to rely on project-wide data storage capabilities and cutting-edge field research techniques that support the use of automatic reasoning (AI) and AI (ML) approaches to deal with live one stride in the front of lawbreakers. ML competencies, misrepresentation and consistence devices can make investments their strength handling extra-complicated extortion issues.

We presented a machine learning-based transaction fraud detection model using feature engineering. The algorithm can learn from its mistakes and thus improve. You can improve stability and performance by processing as much data as possible. These algorithms can be used in the detection of online fraud transactions. A dataset of specific online transactions is used in these. Then, with the help of machine learning algorithms, we can identify unique or unusual data patterns that can be used to detect fraudulent transactions. For the best results, the XGBoost algorithm, which is a cluster of decision trees, will be used. This algorithm has recently taken over the world of machine learning. This ML algorithm outperforms others in terms of accuracy and speed.

Keywords— Fraud detection, Machine learning, Xgboost algorithm, classification, Data pre-processing, Prediction.

I. INTRODUCTION

Credit card fraud is likely the most serious threat to today's business relationships. One of the electronic installment solutions is handled by credit cards. A credit card is a thin rectangular piece of plastic or metal given to a buyer by a bank or financial administration organization to work with instalments to a vendor of labor and products. The advancement of technology has opened up a few new avenues for fabricating demonstrations. Organizations face significant financial, functional, and psychological risks as a result of these acts. Universes are becoming extremely fast in comparison to previous years at this time. Trading is the primary driver of the rapid growth. Card-based installments account for roughly 51% of all exchanges. Despite the benefits of electronic payments, credit card companies are experiencing an increase in card extortion as a result of numerous new innovations. The advancement and development of innovation has opened up a few new avenues for submitting fraudulent demonstrations. Organizations face significant financial, functional, and psychological risks as a result of these acts. In 2018, the estimated monetary loss due to Visa misrepresentation increased to $24.26 million. According to the PR Newswire Association LLC, worldwide extortion losses totaled US $ 27. billion in 2019.

To put a stop to this abuse, necessary preventive measures can be implemented, and the behavior of such fraudulent practices can be studied to minimize it and protect against future occurrences. In other words, this is a highly relevant problem that necessitates the attention of communities such

https://github.com/Ishyanth/Credit_Card_Fraud_Detection

as machine learning and data science, and the solution is automatable. This issue is especially difficult to learn about because it is characterised by various factors such as class imbalance. The number of valid transactions outnumbers the number of fraudulent transactions by a wide margin. Furthermore, transaction patterns' statistical properties frequently change over time.

However, these are not the only challenges associated with implementing a real-world fraud detection system. In practice, automated tools quickly scan a massive stream of payment requests to determine which transactions to authorize. All authorized transactions are analyzed using machine learning algorithms, and any that are suspicious are reported. Professionals investigate these reports and contact cardholders to determine whether or not the transaction was genuine. The automated system receives feedback from the investigators, which is then used to train and update the algorithm, ultimately improving fraud detection performance over time. So, in this project, we attempted to use Machine Learning to create a detection system for such types of fraud.

II. Motivation

Many businesses are expanding rapidly over the world at the moment. Companies strive to deliver the greatest services to their clients. Companies process massive amounts of data on a daily basis for this reason. This data also includes the clients' personal and financial information. As a result, firms must store data in order to handle it, and data security is critical. If this data is not secured, it may be exploited by other companies or, in the worst-case situation, stolen. In rare circumstances, financial information is taken and utilised for fraudulent transactions, causing harm to the parties involved. Today, online buying has become a widespread daily purchase habit. The perpetrators are engaging in malicious operations such as Trojan and spoofing. When criminals steal cardholder information, the increase in fraud incidents has become a severe issue. Credit card fraud detection is an important subject that has piqued the interest of the machine learning and computational intelligence fields, where a plethora of automation solutions have been presented. Data is available all around the world, and businesses of all sizes are loading information with great volume, variety, speed, and value. This data is derived from a variety of sources, including social media followers, likes, and comments, as well as user purchasing habits. All of this data is utilised to analyse and visualise hidden data patterns.

III. Main contributions & Objectives

Online transactions are fast rising, and credit cards will be used mostly. Loss of physical credit cards or loss of credit card information will result in a substantially higher payment. The hackers are there to commit fraud against people. As a result, there was a requirement to detect fraudulent transactions and safeguard online credit card transactions. To analyse this problem and combat credit card theft, we created a new system called "Online credit card fraud detection and prevention system" that employs machine learning. We tested the system's accuracy by experimenting with several algorithms. This system use the random forest algorithm to determine if a transaction is legitimate or fraudulent.

IV. objective

The goal is to distinguish and precisely recognise erroneous extortion recognition. There are several methods for differentiating credit cards. misrepresentation, such as neural networks, genetic algorithms, and k-means clustering The cost of consistent deception in the economy is more than $4 trillion globally. The solution consists in relying on cutting-edge investigation and performing extensive data stockpiling skills that aid the utilisation of computerised reasoning (AI) and AI (ML) approaches to deal with staying one step ahead of lawbreakers. ML capabilities, misrepresentation, and consistency teams can focus their efforts on more sophisticated extortion concerns.

Our Objective is to Implementing below Algorithm for fraud detection

- **Logistic Regression**
- **SVM (Support Vector Machine)**
- **DT (decision-tree)**
- **XGBoost**

V. Related work

There are significant financial losses as a result of credit card fraud incidents. Criminals utilise Trojan and phishing

technologies to steal credit card information from others. card. As a result, the fraud detection approach is critical.

Because of fraud detection methods, we can detect fraud when a criminal uses a false card to defraud a consumer. Two types of algorithms are employed in this paper to train the behaviour feature of normal and fraudulent transactions.

The purpose of data analytics is to discover hidden patterns and use them to make informed decisions in a range of scenarios. The publicly available datasets on credit card fraud are highly mismatched. We analyse the most essential variables that may contribute to improved credit card fraudulent transaction detection accuracy.

Theft of sensitive credit card information or physical theft of a credit card is considered credit card fraud. For credit card identification, there are several machine learning algorithms available. Multiple algorithms are utilised in this study to identify whether a transaction is fraudulent or not. This study made use of a credit card fraud detection dataset.

Oversampling was accomplished using the "Synthetic Minority Over Sampling Technique (SMOTE)." The dataset was split into two parts: training and testing data. The methods used in the study included Logistic Regression, XGBoost, DT, SVM. The results indicate that each algorithm can detect credit card theft with high accuracy.

VI. Literature Survey

"A Comparison of XGBoost"- Gonzalo, Martinez

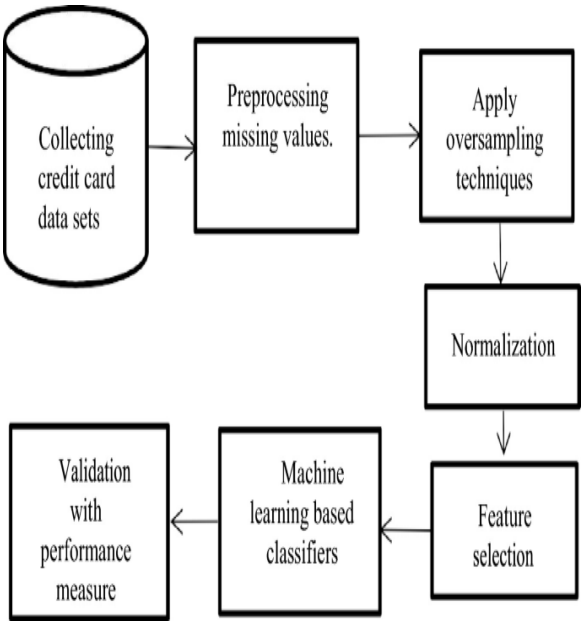
This paper proposes a practical examination of how this novel technique performs in terms of training speed, generalisation performance, and parameter setup. Furthermore, a thorough comparison of XGBoost, random forests, and gradient boosting was carried out using carefully tuned models as well as the default settings. The results of this comparison may indicate that XGBoost is not always the best option, but it does have advantages over other algorithms.

"Detection of Credit Card Fraud in Data Mining Using the XGBoost Classifier" - Rahul Goyal, Amit Kumar Manjhvar, and Vikas Sejwar

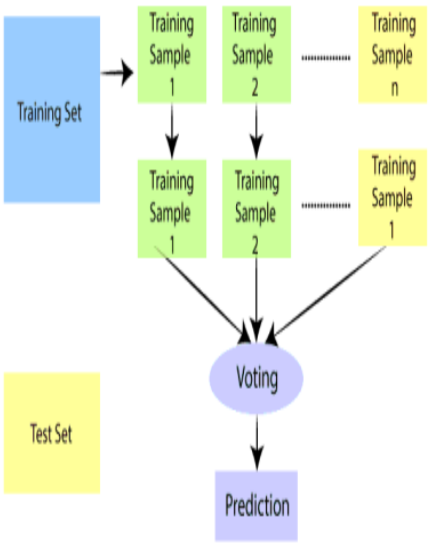
To classify fraud activities, the proposed system in this research paper employs a combination of the SMOTE technique and the Xgboost classification algorithm. Synthetic Minority Oversampling Technique is referred to as SMOTE. This is a technique for increasing the number of cases in a balanced manner in your dataset. SMOTE takes the entire dataset as input, but only the minority cases are increased in percentage. Using XGBoost, they measured and validated their performance for credit card frauds using only publicly available datasets.

VII. Proposed Framework

Architecture



Workflow for multimodal



Analysis

Fraud detection is a set of operations used to prevent money or property from being obtained fraudulently. false representation Fraud can be committed in a variety of ways and in a wide range of industries. To make a decision, the majority of detection systems use a range of fraud detection datasets to generate a connected picture of both legitimate and invalid payment data. This decision must take into account IP address, geolocation, device identity, "BIN" data, global latitude/longitude, historical transaction trends, and transaction information. In practise, this implies that merchants and issuers use analytically based solutions to detect fraud, which leverage internal and external data to apply a set of business rules or analytical algorithms.

Credit Card Fraud Detection Using Machine Learning is a data research process. by a team of data scientists and the creation of a model that will yield the greatest outcomes in terms of detecting and preventing fraudulent transactions This is accomplished by aggregating all relevant information of card users' transactions, such as Date, User Zone, Product Category, Amount, Provider, Client's Behavioural Patterns, and so on. The data is then fed into a subtly trained model that looks for patterns and rules to determine if a transaction is fraudulent or lawful. All major banks, including Chase, employ fraud monitoring and detection systems.

Credit Card Fraud Techniques and Prevention

Associations and banks that use them make excellent proposals. arrangements for security To address these concerns, but at the same time Following the development of the same fraudsters' simple methods a period of time As a result, it is critical for future development. Techniques for recognising and counteracting Location of Fraud because the primary goal of avoidance is to distinguish to distinguish between legitimate and bogus exchanges and to prevent phoney movement. In the event that the framework fails to detect and prevent bogus exercises, extortion detection dominates. Frameworks for managed extortion discovery In light of this, new exchanges are labelled as fraudulent or certified. characteristics of both deceptive and genuine exercises Anomalies exchanges are identified as potentially fake. trades in individual extortion location frameworks.

System configuration

This project may be run on standard hardware. We ran the entire project on an Intel I5 processor with 8 GB RAM and a 2 GB Nvidia Graphic Processor. It also has two cores that run at 1.7 GHz and 2.1 GHz. The first half of the process is the training phase, which takes about 10-15 minutes, and the second part is the testing phase, which just takes a few seconds to generate predictions and calculate accuracy.

Hardware Requirements:

• RAM: 4 GB

• Storage: 500 GB

• CPU: 2 GHz or faster

• Architecture: 32-bit or 64-bit

Software requirements

• Python 3.5 in Google Colab is used for data pre-processing, model training and prediction.

• Operating System: windows 7 and above or Linux based OS or MAC OS.

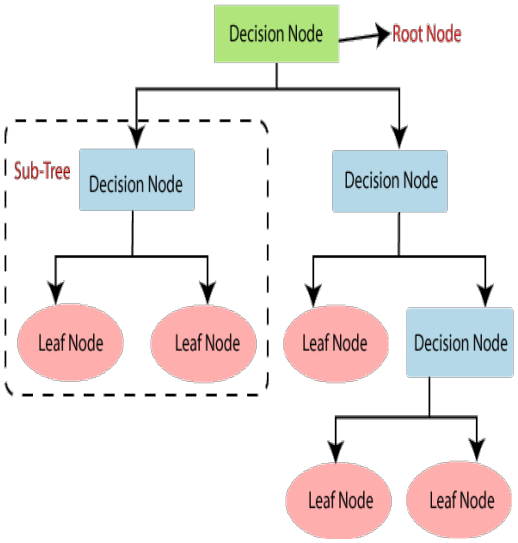
Methodology

In the field of data visualization, patterns, trends, and relationships that might otherwise go unnoticed are highlighted by showing data in a visual context in an effort to better comprehend it.

Excellent graphing libraries in Python are abundant and filled with essential features. A wonderful library for making dynamic or highly customized charts is available in Python. A few well-known plotting libraries are listed below to give you a quick overview:

- **Matplotlib:** low level, provides lots of freedom
- **Pandas Visualization:** easy to use interface, built on Matplotlib
- **Seaborn:** high-level interface, great default styles
- **plotnine:** based on R's ggplot2, uses Grammar of Graphics
- **Plotly:** able to produce interactive plots

Algorithm of Decision Tree Classification



Flow of the Process

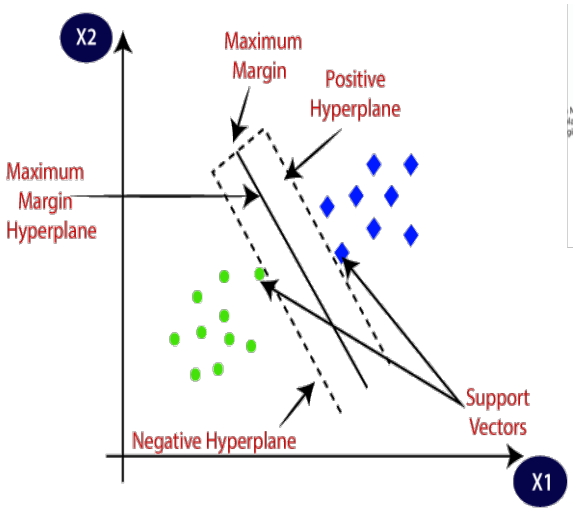
- Step 1: Begin the tree with the root node, which contains the entire dataset, says S.
- Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset (ASM).
- Step 3: Subdivide the S into subsets containing potential values for the best qualities.
- Step 4: Create the decision tree node with the best attribute.
- Step 5: Create new decision trees recursively using the subsets of the dataset obtained in step 3. Continue this process until you reach a point where you can no longer categorise the nodes and refer to the final node as a leaf node.

Support Vector Machine

Support Vector Machine, or SVM, is a well-known Supervised Learning method that is applied to both regression and classification problems. The majority of the time, it is used in Machine Learning to solve classification problems.

The goal of the SVM method is to identify the best decision boundary or line for classifying n-dimensional space so that subsequent data points can be quickly added to the appropriate category. The best possible chosen boundary is a hyperplane.

SVM chooses the extreme vectors and points that help build the hyperplane. Support vectors are used in this technique, which is also known as the Support Vector Machine. Take a look at the image below, which displays two unique categories that are



XG Boost

Another boosting technique introduced as an improvement over Gradient Boosting is XGBoost (Extreme Gradient Boosting). XGBoost reduces the risk of overfitting the dataset, as in Gradient Boosting. It also provides the added benefit of handling missing values on its own. However, the working procedure is the same as with Gradient Boosting.

Data Description

Dataset Link:

https://www.kaggle.com/datasets/kartik2112/fraud-detection

Motive for a Dataset!

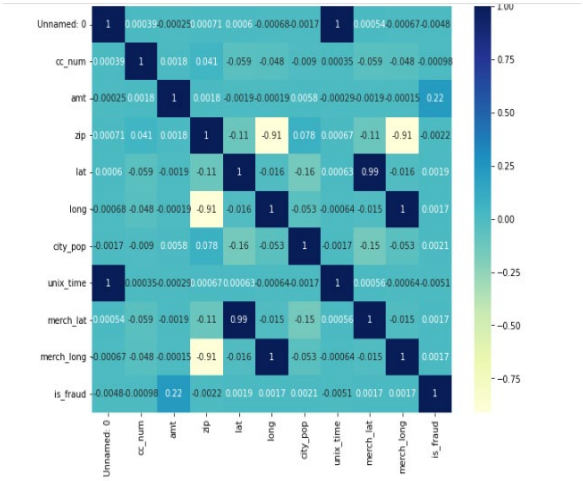
This data collection has been published in order to gain insights into credit card defaulters based on the relevant attributes! inside in the application data set, we have attributes such as income total, amt application, amt credit, and around 122 columns. the fascinating part is that if you want to see patterns and variances, we can also leverage the previous application data set to gain more insights.

Inspiration

We accepted this data set as our homework and tried our hardest to complete the EDA to the best of our abilities!

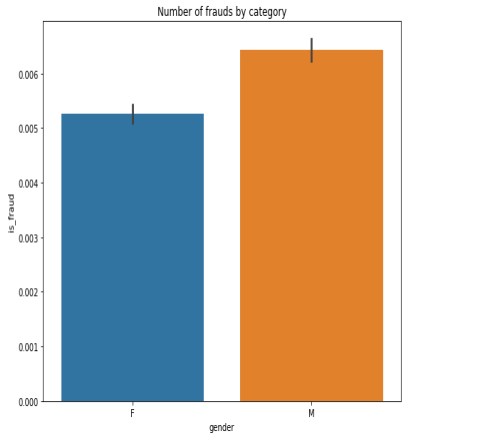
VIII. RESULTS

Correlation matrix



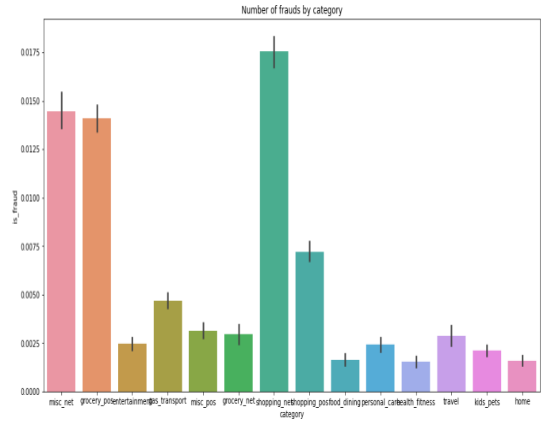
Box Plot

```
[50]: <AxesSubplot:title={'center':'Number of frauds by category'}, xlabel='gender', ylabel='is_fraud'>
```



Bar Graph

```
[51]: <AxesSubplot:title={'center':'Number of frauds by category'}, xlabel='category', ylabel='is_fraud'>
```



Data Columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15012 entries, 123118 to 1295733
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             15012 non-null  int64
 1   trans_date_trans_time  15012 non-null  object
 2   cc_num                 15012 non-null  int64
 3   merchant               15012 non-null  object
 4   category               15012 non-null  object
 5   amt                    15012 non-null  float64
 6   first                  15012 non-null  object
 7   last                   15012 non-null  object
 8   gender                 15012 non-null  object
 9   street                 15012 non-null  object
 10  city                   15012 non-null  object
 11  state                  15012 non-null  object
 12  zip                    15012 non-null  int64
 13  lat                    15012 non-null  float64
 14  long                   15012 non-null  float64
 15  city_pop               15012 non-null  int64
 16  job                    15012 non-null  object
 17  dob                    15012 non-null  object
 18  trans_num              15012 non-null  object
 19  unix_time              15012 non-null  int64
 20  merch_lat              15012 non-null  float64
 21  merch_long             15012 non-null  float64
 22  is_fraud               15012 non-null  int64
dtypes: float64(5), int64(6), object(12)
memory usage: 2.7+ MB
```

SVM Classification Report

Classification report

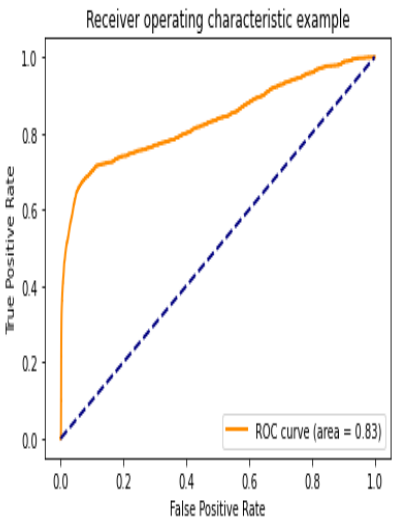
	precision	recall	f1-score	support
0	1.00	0.91	0.95	553574
1	0.03	0.69	0.05	2145
accuracy			0.91	555719
macro avg	0.51	0.80	0.50	555719
weighted avg	0.99	0.91	0.95	555719

DT Classification Report

Classification report

	precision	recall	f1-score	support
0	1.00	0.93	0.96	553574
1	0.01	0.25	0.03	2145
accuracy			0.93	555719
macro avg	0.51	0.59	0.50	555719
weighted avg	0.99	0.93	0.96	555719

Logistic Regression



Classification report

	precision	recall	f1-score	support
0	1.00	0.94	0.97	553574
1	0.01	0.21	0.03	2145
accuracy			0.94	555719
macro avg	0.51	0.58	0.50	555719
weighted avg	0.99	0.94	0.97	555719

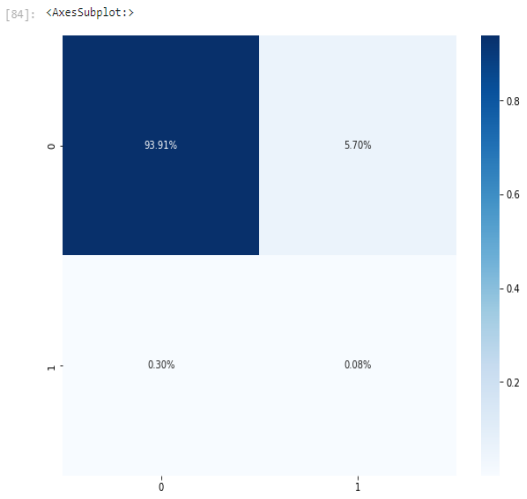
LR Report

Classification report

	precision	recall	f1-score	support
0	1.00	0.91	0.95	553574
1	0.03	0.69	0.05	2145
accuracy			0.91	555719
macro avg	0.51	0.80	0.50	555719
weighted avg	0.99	0.91	0.95	555719

XGBoost

Confusion Matrix



Classification Report

REFERENCES

[1] S. Bachmayer, "Artificial Immune Systems," pp. 119-131 in Artificial Immune Systems, vol. 5132, 2008.M. Krivko, "A Hybrid Model for Plastic Card Fraud, "Expert Systems with Applications, vol. 37, no. 8, pp. 6070-6076, August 2010.

[2] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study,"

[3] Decision Support Systems, vol. 50, no. 3, Feb. 2011, pp. 602-613.

[4] "Plastic card fraud detection via peer group analysis," Advances in Data Analysis and Classification, vol. 2, no. 1, pp. 45-62, Mar. 2008.

[5] O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, "Credit Card Fraud Detection Using Machine Learning As A Tool".

[6] Global Facts (2019). Topic: Startups worldwide. [online] Available at: https://www.statista.com/topics/4733/startupsworldwide/ [Accessed 10 Jan. 2020].

[7] Legal Dictionary (2019). Fraud - Definition, Meaning, Types, Examples of fraudulent activity. [online] Available at: https://legaldictionary.net/fraud/ [Accessed 15 Jan. 2020].

[8] A. Mishra, C. Ghorpade, "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE.

[9] M. Krivko, "A hybrid model for plastic card fraud detection systems," Expert Systems with Applications, vol. 37, no. 8, pp. 6070–6076, Aug. 2010.