

Introduction

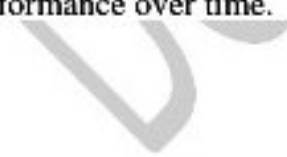
Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used.

Due to rise and acceleration of E-Commerce, there has been a tremendous use of credit cards for online shopping which led to High amount of frauds related to credit cards. In the era of digitalization, the need to identify credit card frauds is necessary. Fraud detection involves monitoring and analyzing the behavior of various users in order to estimate detect or avoid undesirable behavior. In order to identify credit card fraud detection effectively, we need to understand the various technologies, algorithms and types involved in detecting credit card frauds. Algorithm can differentiate transactions which are fraudulent or not. Find fraud, they need to passed dataset and knowledge of fraudulent transaction. They analyze the dataset and classify all transactions.

Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behavior, which consist of fraud, intrusion, and defaulting.

Machine learning algorithms are employed to analyses all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent.

The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.



S P Maniraj [1] In this paper, they describe Random forest algorithm applicable on Find fraud detection. Random forest has two types. They describe in detail and their accuracy 91.96% and 96.77% respectively. This paper summaries second type is better than the first type.

Suman Arora [2] In this paper, many supervised machine learning algorithms apply on 70% training and 30% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree and KNN algorithms compare each other i.e. 94.59%, 95.27%, 94.59%, 93.24%, 90.87%, 90.54% and 94.25% respectively. Summaries of this paper, SVM has the highest ranking with 0.5360 FPR, and stacking classifier has the lowest ranking with 0.0335.

Kosemani Temitayo Hafiz [3] In this paper, they describe flow chart of fraud detection process. i.e. data Acquisition, data pre-processing, Exploratory data analysis and methods or algorithms are in detail. Algorithms are K- nearest neighbor (KNN), random tree and Logistic regression accuracy are 96.91%, 94.32%, 57.73% and 98.24% respectively.

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers.

The basic rough architecture diagram can be represented with the following figure:

When looked at in detail on a larger scale along with real life elements, the full architecture diagram can be represented as follows:

First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets.

Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data.

The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one.

We plot different graphs to check for inconsistencies in the dataset and to visually comprehend it:

This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.

This graph shows the times at which transactions were done within two days. It can be seen that the least number of transactions were made during night time and highest during the days.

This graph represents the amount that was transacted. A majority of transactions are relatively small and only a handful of them come close to the maximum transacted amount.

After checking this dataset, we plot a histogram for every column. This is done to get a graphical representation of the dataset which can be used to verify that there are no missing

Credit Card Fraud Detection

any values in the dataset. This is done to ensure that we don't require any missing value imputation and the machine learning algorithms can process the dataset smoothly.

After this analysis, we plot a heatmap to get a colored representation of the data and to study the correlation between our predicting variables and the class variable. This heatmap is shown below:

The dataset is now formatted and processed. The time and amount column are standardized and the Class column is removed to ensure fairness of evaluation. The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms work together: This data is fit into a model and the following outlier detection modules are applied on it:

- Local Outlier Factor
- Isolation Forest Algorithm

These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection.

This free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provides a lot of simple and efficient tools which can be used for data analysis

and machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries.

We used Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files.

Detailed explanations about the modules with pseudocodes for their algorithms and output graphs are given as follows:

1. Local Outlier Factor

It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbors.

Credit Card Fraud Detection

More precisely, locality is given by k -nearest neighbors, whose distance is used to estimate the local data.

The pseudocode for this algorithm is written as:

On plotting the results of Local Outlier Factor algorithm, we get the following figure:

By comparing the local values of a sample to that of its neighbors, one can identify samples that are substantially lower than their neighbors. These values are quite anomalous and they are considered as outliers.

As the dataset is very large, we used only a fraction of it in our tests to reduce processing times.

The final result with the complete dataset processed is also determined and is given in the results section of this paper.

2. Isolation Forest Algorithm

The Isolation Forest isolates observations by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature.

Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent to the path length root node to terminating node.

The average of this path length gives a measure of normality and the decision function which we use.

The pseudocode for this algorithm can be written as:

On plotting the results of Isolation Forest algorithm, we get the following figure:

Partitioning them randomly produces shorter paths for anomalies. When a forest of random trees mutually produces shorter path lengths for specific samples, they are extremely likely to be anomalies.

Once the anomalies are detected, the system can be used to report them to the concerned authorities. For testing purposes, we are comparing the outputs of these algorithms to determine their accuracy and precision.

Predict & Update

- Streaming Logistic Regression Model with Stochastic Gradient Descent

