**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING AND DATA SCIENCE**

*Abstract-*Financial fraud in this environment is the fast-growing issue since the mobile channel can facilitate nearly any type of payments. Since in this environment any malicious activities can be took happens like fraud transaction Such problems can be tackled by data science and its importance, along with Machine Learning, cannot be overstated. As we know there are many technology which day by day comes in market and there are hoaxer who studied the technology and used them in malicious activities like fraud detection.

So our project objective is to detect the fraudulent cases while minimizing the fraud cases by applying 2 major algorithms-LOCAL OUTLIER FACTOR AND ISOLATION FOREST ALGORITHM. This totally works on pre-processing data.

***Keywords****-Credit card fraud, applications of machine learning, data science, isolation forest algorithm, local outlier factor*

1. INTRODUCTION

If we see the fraudulent cases there are more in our country. 'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Earlier fraud cases were more because when someone do transaction then don’t even identify who utilize his/her credit card but now whoever do transaction the respective message is given to respective phone number. Basically in this project we have done on the pre-processed data using machine learning and as well as data scince.

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

This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time.

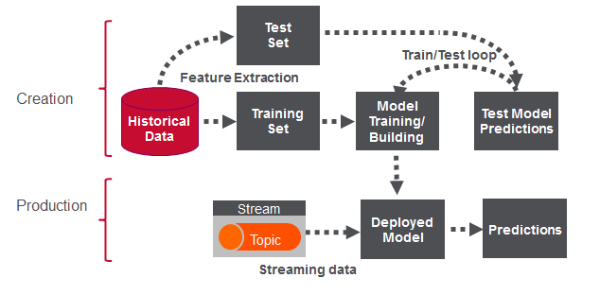
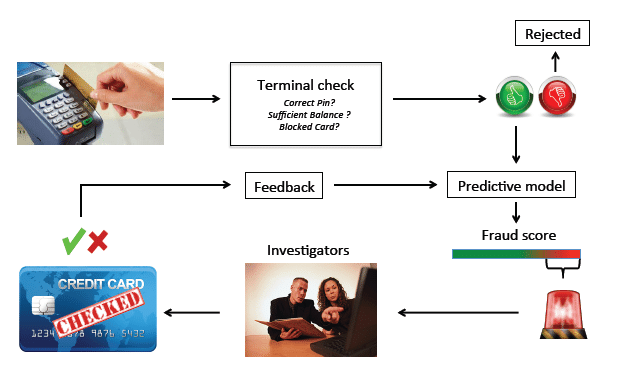
Some of the currently used approach to detec fraud case are-

1. FUZZY LOGIC
2. GANETIC ALGORITHM
3. ARTIFICIAL NEURAL NETWORK
4. LOGISTIC REGRESSION
5. K-NEAREST NEIGHBOUR
6. LITERATURE REVIEW

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit.As fraud detection is the critical part of measures implemented of maintaining an attack tolerant data base system. In this data science uses an effective way to give a fraudulent cases while generating value at the same time. These data sets came across by kaggle in which there were 31 features out of which 28 were anonym zed and labeled v1 to v28.The remaining 3 features were time, amount as well as tell whether the transaction were fraudulent or not. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection.Earlier same detection was done by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank.

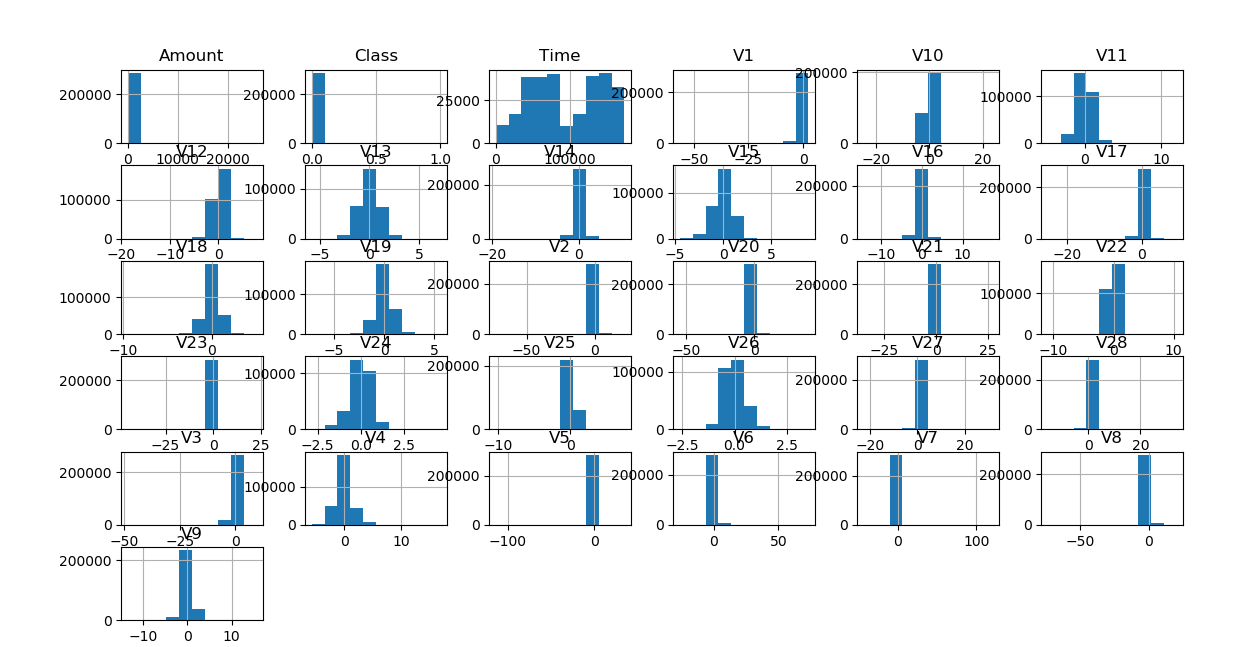
1. METHODOLOGY

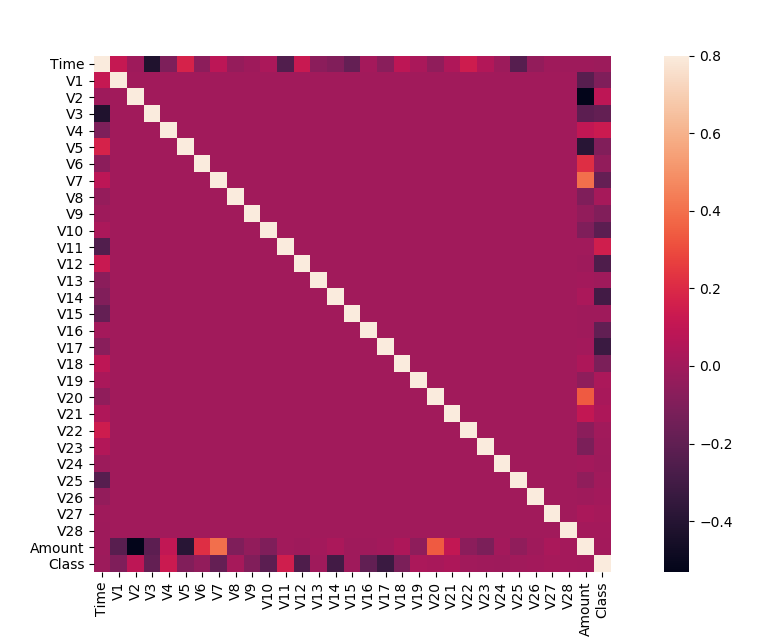
The main aim of this project is to to use the various alagorithm to detect fraud detection done by credit card.The small diagram helps to clear fraud detection-



First of all we fetch the data from kaggle which contains all the fraudulent cases as well as well valid cases.This file was in in the extension .csv file and in this file there were 31 columns out of which v1 to v28 were there to protect sensitive data and there were another columns i.e. time, amount and class in which time shows the gap between first transaction and the following one. Amount shows how much transaction is done and class 0 represents valid transaction and class 1 represents fraud transaction.

S o we show some bar graphs which tells valid as well as fraud cases-

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



The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms work together:

1. outlier fraction factor algorithm
2. isolation forest algorithm

These algorithms are a part of sklearn. This free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provide a lot of simple and efficient tools which can be used for data analysis and machine learning. This project is done on python using jupyter notebook.

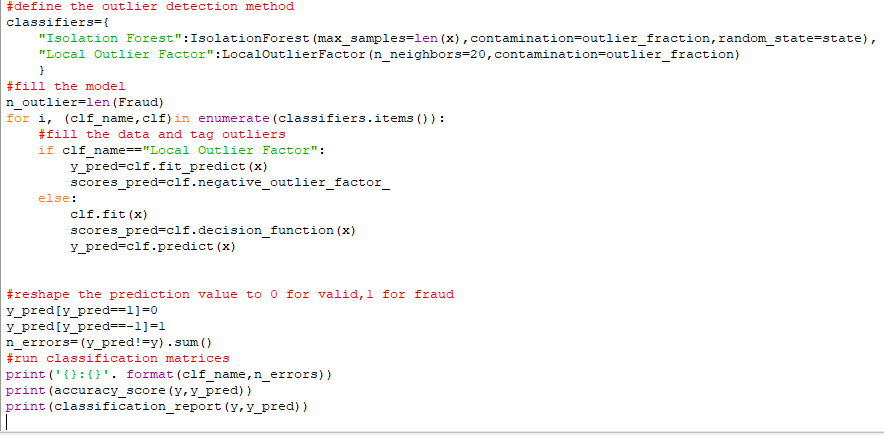
1. The local outlier factor is based on a concept of a local density, where locality is given by {\displaystyle k} nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be [outliers](https://en.wikipedia.org/wiki/Outlier).

The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters.

By comparing the local values of a sample to that of its neighbours, one can identify samples that are substantially lower than their neighbours. These values are quite amanous and they are considered as outliers. As the dataset is very large, we used only a fraction of it in out 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.

1. **Isolation forest** is an  unsupervised learning algorithm for  anamoly detection that works on the principle of isolating anomalies, instead of the most common techniques of profiling normal points. In statistics an anomaly is an observation or event that deviates so much from other events to arouse suspicion it was generated by a different mean. Anomalies in a big dataset may follow very complicated patterns, which are difficult to detect “by eye” in the great majority of cases. This is the reason why the field of anomaly detection is well suited for the application of machine learning techniques.

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.

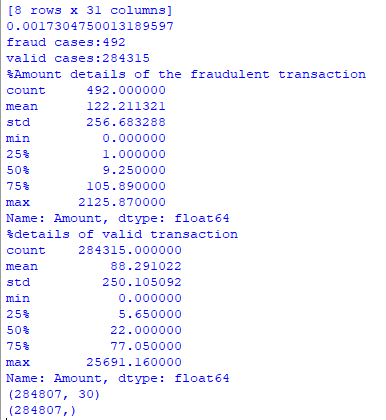
1. IMPLEMENTATION

The idea is difficult to implement in real life because we need a cooperating from banks which are not willing to share information due to thir market competition, and also due to legal reasons and protection of data of their users. Therefore, we looked up some reference papers which followed similar approaches and gathered results. As stated in one of these reference papers: “This technique was applied to a full application data set supplied by a German bank in 2006. For banking confidentiality reasons, only a summary of the results obtained is presented below. After applying this technique, the level 1 list encompasses a few cases but with a high probability of being fraudsters.

Credit and collection officers considered that half of the cases in this list could be considered as suspicious fraudulent behaviour. For the last list and the largest, the work is equitably heavy. Less than a third of them are suspicious.

1. RESULTS

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives. Results when 10% of the dataset is used:



1. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.

1. FUTURE ENHANCEMENT

While we couldn’t reach out goal of 100% accuracy in fraud detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives. However, this requires official support from the banks themselves.