Credit card fraud detection using Machine Learning

Abdul Jamsheed V
National Institute of Technology
Tiruchirappalli, TamilNadu, India-620015
205219001@nitt.edu

Dr. U Srinivasulu Reddy
National Institute of Technology
Tiruchirappalli, TamilNadu, India-620015
usreddy@nitt.edu

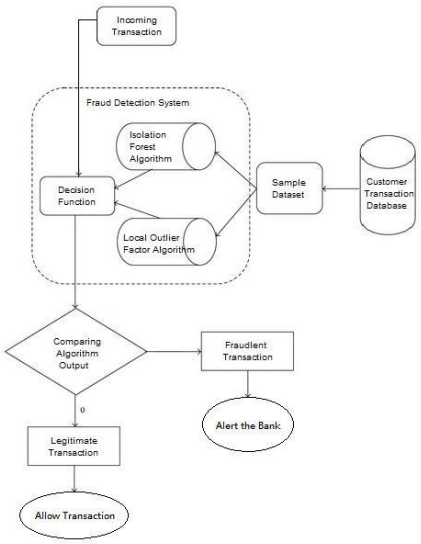
Abstract*-- credit card plays important role in today's economy. People use credit card for all daily transaction and now it becomes unavoidable part of our household, business and global activities. Even though there is huge benefits in using credit cards when used safely and carefully, significant financial damage has been reported by fraudulent activities. Many techniques have been proposed to confront the credit card fraud detection. However each techniques have its own advantages and disadvantages in detecting and reducing these fraudulent transactions. It is vital that credit card companies are detecting these fraudulent activities so that customers are not charged with unnecessary expenses. These problems can be tackled with machine learning techniques to a great extent. This project intends to illustrate the modelling of a data set using different machine learning techniques in order to detect the credit card fraud activities. With this model we will check whether the new transaction is fraudulent or non-fraudulent. Our objective is to minimize the error and get 100% accuracy.

Keywords: credit card fraud detection, machine learning, logistic regression, LDA, KNN, support vector classifier, random forest classifier.

I. INTRODUCTION

In credit transactions, fraud is unauthorized and unwanted usage of an owners account by someone who intended to take advantage of the situation or circumstances. Many necessary steps are taken by banking companies to minimize the fraudulent activities by intruders. In other words credit card fraud can be defined as a person uses some others credit card for transaction without proper permission or in some other case owner is unaware of the transaction. Credit card detection is a very relevant problem which requires attention of the machine learning engineers and data scientists. These communities can automate the fraudulent activities and bring better solution to detecting these types of malicious activities and thereby reducing it. This problem is challenging due to the unbalanced data, and other major factors. Always the valid transaction outnumber the fraudulent transactions. This makes the solution harder.

These are not only the challenges faced in the real world, but just an overview of the major cause of credit card fraud. In real world massive stream of payment requests are coming in every second which is huge challenge to monitor. Machine Learning algorithms helps in solving these huge challenges to an extent.



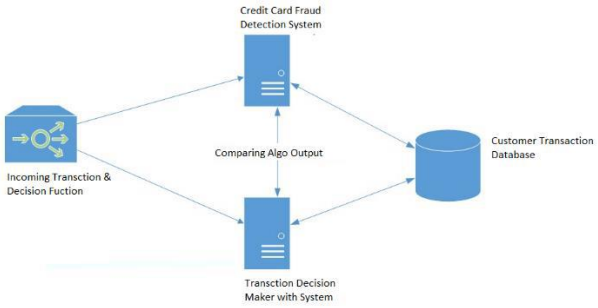
II. LITERATURE REVIEW

Fraud is unlawful and criminal act intended to result in financial loss or personal benefit. Credit card fraud is increasing day by day. So the research in credit card fraud detection is also got a good hype. Numerous researches have already done in the field of credit card fraud. Many research papers are published and are publicly available in many e-platforms. Clifton Phua on his paper of data mining application have mentioned the disadvantages of automatic fraud detection. In a different paper published by Suman, who is a research scholar in GJUS&T at hisar mentioned about the supervised and unsupervised approach for solving credit card fraud detection. Even though this type of solution has been provided by many other researchers, he told how the accuracy can be improved on the existing algorithms. Even though the solution gained a high appreciation it failed to get consistent results for all transaction.

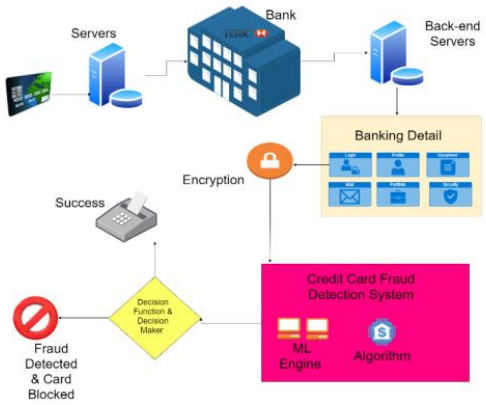
There have been many efforts to improve the accuracy of credit card fraud detection with different machine learning algorithm, every solution lacks the consistency and maintaining their accuracy rate for all transactions. These misclassification has forced scientists to continuously do researches in this field.

III. METHODOLOGY

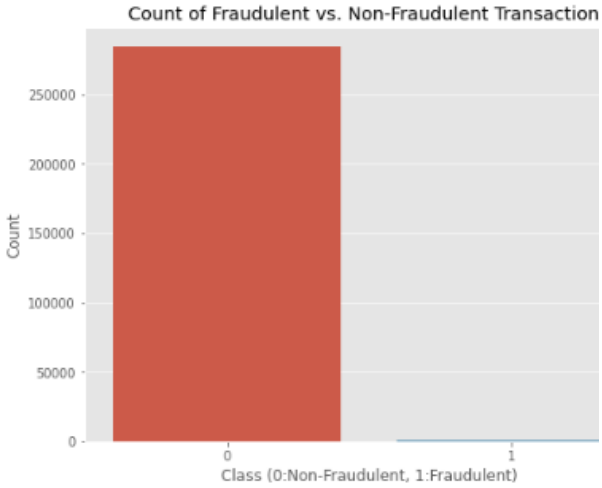
We use an approach in which latest machine learning algorithms are used to detect the anomalous activities in the banking transactions. Basic block diagram can be represented as follows



If we look in to the detail of the above block diagram with real life elements, we will get a full architecture diagram, which is represented as follows:

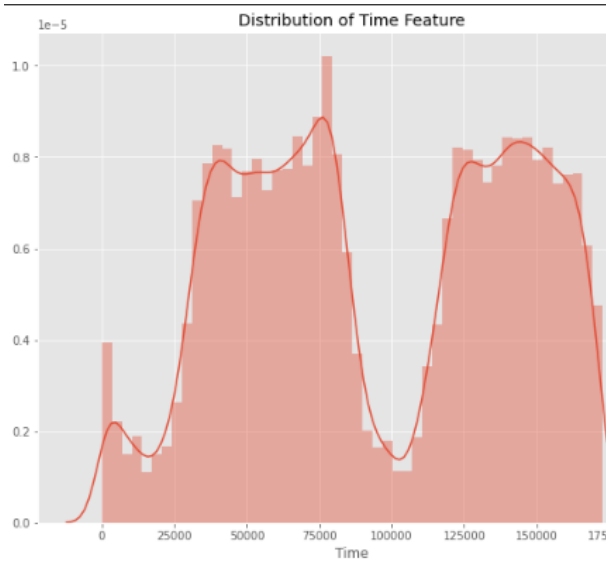


The data set for credit card fraud detection is found from Kaggle website which is a data analysis website. The original data set is sourced from ULB Machine learning Group. This data set consist of 2,84,807 transactions with 0.172% fraud cases. Data set has total of 31 columns out of which 28 are named from v1-v28 to protect sensitive data. Other columns are class time and amount. Class represents the valid transaction. If it is 1 means fraudulent and 0 means non-fraudulent. Time represents the time gap between one transaction and the other. Amount is the money transacted for that transaction. We plot different graphs to visualise the inconsistencies in the data set and rectify them.

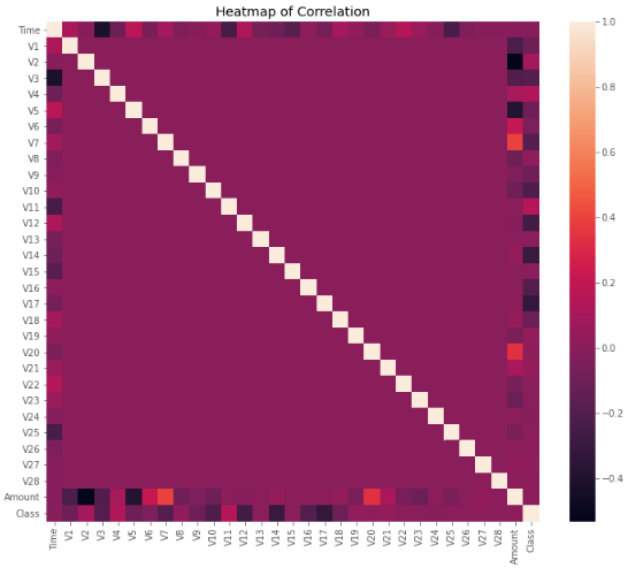


Above graph shows the difference between fraudulent and non-fraudulent transaction. From the figure, it is clear that number of fraudulent transactions are much lower than the legitimate ones.

Following graph shows the time of all the transaction in the data set. Whole transaction has been done within 2 days. Also it can be noted that during night time number of transactions were very low and during day time the transactions are pretty high.

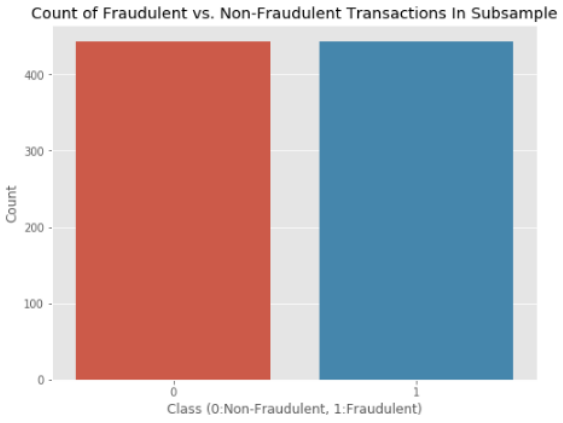


After this analysis, we plot a heat map to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable.
This heat map is shown below:



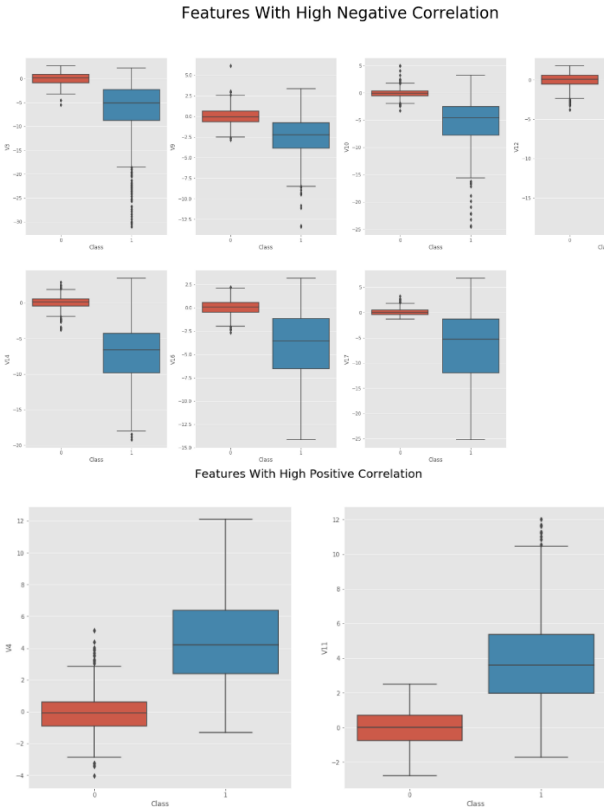
The data set is now formatted. All the columns are standardized so that no disparities will occur in the data. Class column is removed so that difference between fraudulent and non-fraudulent transactions are normalized with equal number of both transactions. These data are now processed by a set of Machine Learning algorithms.

The existing training data set more likely to select non fraudulent class since 99% of data are non-fraudulent. This will make 99% accuracy if all the data are classified as non-fraudulent. So using the original dataset would not be appropriate to classify the actions. We don't want a 99% accurate system that doesn't classify the system as fraudulent and non-fraudulent, but we want to detect the transactions as such and label the accurately.



A. Outlier detection and removal.

There will be many outliers in our data and detecting the outliers is a complex topic. We have to consider the trade-off between the number of transactions and the amount of information available in these transactions. With extreme outliers in the data set, it will affect the result by skewing toward one end and reducing the prediction efficiency. Here we consider the features with a correlation of 0.5 or higher as class variable for outlier removal.



For checking the outlier removal we usually take IQR(inter- quartile range) into consideration. All transaction outside the 1.5 times of IQR are neglected in many case. But if we remove all the transaction outside 1.5 IQR it will lead to reduce the training data size. Instead we take 2.5 time IQR for our training data.

B. Dimensionality reduction

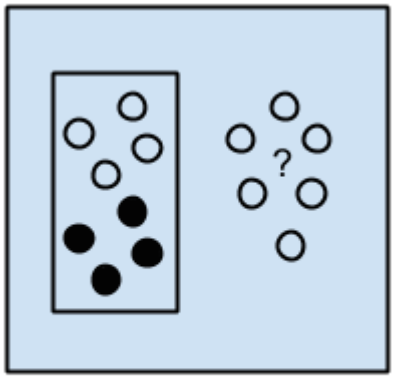
Visualizing the data will make people understand the data very easily. But generating a data of 20 dimensions in a single plane is a difficult part. Here comes the role of dimensionality reduction algorithms. We use t-SNE dimensionality reduction technique to project our higher dimensional data into a single plane. The scatter plot diagram is shown below.

IV. IMPLEMENTATION

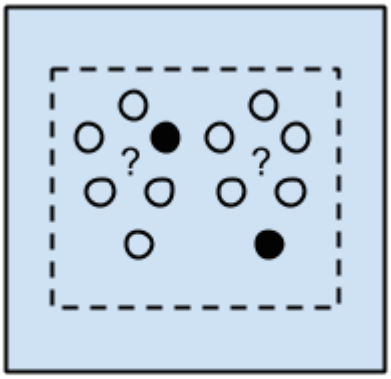
Implementation of credit card detection is difficult in real life because we need prior approval of banking organizations. They won't do this due to the competition and security issues. So we have taken Kaggle data set and done a theoretical approach to solve this problem. Following algorithmic methods are used for the implementation of credit card fraud detection in this paper.

1. Supervised Learning

A model is prepared from labelled data and make predictions. Model is re trained when the predictions are wrong.

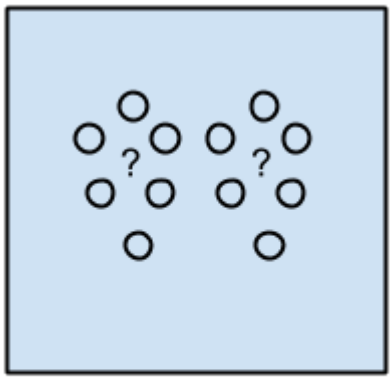


Supervised Learning Algorithms



Semi-supervised Learning Algorithms

2. Unsupervised Learning
A model is prepared from unlabelled data. The systems is trained to learn from inherent structure of the input.

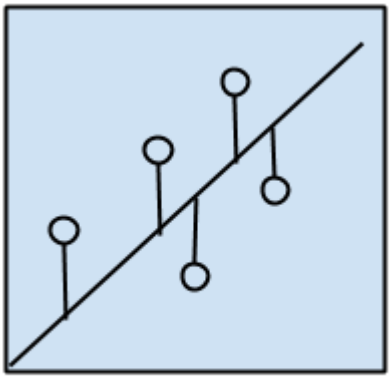


Unsupervised Learning Algorithms

3. Semi-Supervised Learning
Input data is a mixture of labelled and unlabelled data. Or in other way it is learning in between the supervised and un supervised learning.

Regression Algorithms

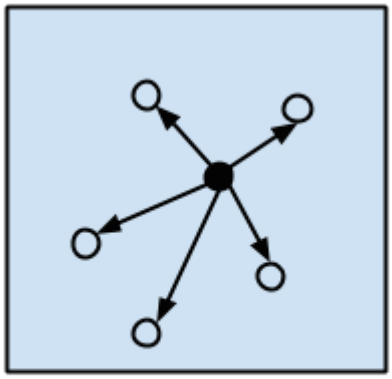
Model in which it predicts the output values based on the input features fed into the system.



Regression Algorithms

Instance-based Algorithms

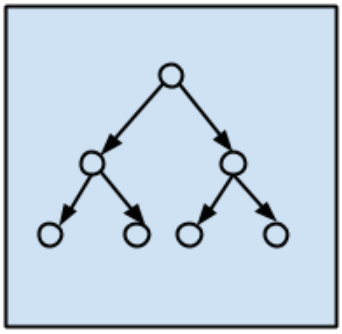
Instance based algorithms are also known as memory based learning that instead of performing explicit generalizations, compares new problem with instance. Such type of algorithms are K Nearest Neighbour (KNN) algorithm and Support Vector Machine (SVM)



Instance-based Algorithms

Decision Tree Algorithms

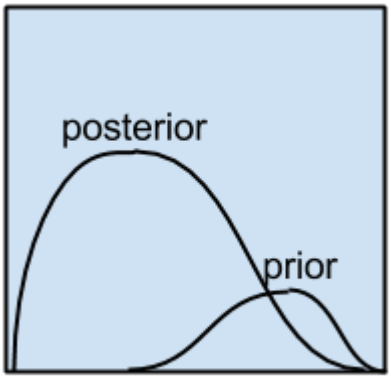
Decision tree models construct a tree like model and make decisions based on the levels of the tree. Classification and regression are examples of this type of model.



Decision Tree Algorithms

Bayesian Algorithms

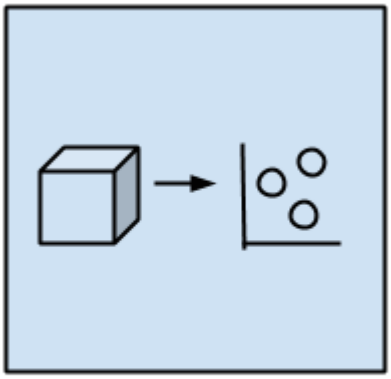
It's a probabilistic classifiers which is based on Bayes theorem. Naïve Bayes algorithm is an example of this type of algorithm.



Bayesian Algorithms

Dimensionality Reduction Algorithms

Dimensionality reductions techniques are generally used to plot a high dimensional data into a low dimensional plane.These algorithms are used to visualize the data or simplify the data. PCA and LDA are few of them.



Dimensional Reduction Algorithms

To test the performance of our algorithms, we first performed an 80/20 train-test split, splitting our balanced data set into two pieces. We will perform analysis on following models and then compare the results.

- Logistic Regression
- Naïve Bayes
- K Nearest Neighbour (KNN)
- Decision Trees
- Support Vector Classifier
- Random Forest Classifier
- XGBoost Classifier

No	Algorithm	Result of performance			
		Accuracy	Precision	Recall	F-score
1	Random Forest	0.9840	0.9997	0.985	0.994
2	XGBoost	0.9759	0.9998	0.975	0.985
3	Logistic Regression	0.9709	0.9997	0.976	0.984
4	K Nearest Neighbour	0.9640	0.9948	0.977	0.985
5	Naïve Bayes	0.9594	0.9956	0.962	0.976
6	Decision Tree	0.9223	0.9965	0.956	0.975
7	Support Vector Machine	0.9562	0.9976	0.953	0.975

V. RESULTS

Results of all machine learning model can be analysed for the above diagram.

From the detailed analysis of all algorithm we can see that a few algorithms quite significantly outperformed others. They are Random Forest and XGBoost. But the clear winner is Random Forest. Random Forest not only considers the highest accuracy, but also takes into account the business value of the project. So taking Random Forest over XGBoost is a reasonable approach.

VI. CONCLUSION

Credit card fraud detection is a criminal offence without a doubt. This project focus on detecting and minimizing the credit card fraud activities with different machine learning algorithms. This paper also explained in detail how such machine learning models are trained to get a better result with implementation and experimental results. The best algorithm reached over 98% accuracy and 99% precision. Recall and F-score also are quite high. This is possible because of the effective cleaning of data. Pre-processing the data is a crucial step in the analysis of machine learning algorithm. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

VII. FUTURE WORK

We didn’t end up with our goal of 100% accuracy, means that there is future work to do to make it 100% accurate. Since data science field is growing exponentially there will be an improvement in these algorithms to make it 100% accurate. We have analysed the dataset with many algorithms to get a better result. Future scope of this project is to combine these algorithms and get a 100% accuracy of the system. More room for improvement can also be found in the dataset. These process will make the system more accurate and reduce false positives.

REFERENCES

[1] “Credit Card Fraud Detection Based on Transaction Behaviour –by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] Clifton Phua1, Vincent Lee1, Kate Smith1 & Ross Gayler2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3] “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008.

[4] “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence

[5] “Credit Card Fraud Detection through Parenclitic Network Analysis- By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6] “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE Transactions On Neural Networks And Learning Systems, Vol. 29, No. 8, August 2018

[7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016