International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

Credit Card Fraud Detection Analysis

Dushyant Singh[1], Saubhagya Vardhan[2], Dr. (Mrs.) Neha Agrawal[3]

[1]Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, New Delhi, India
[2]Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, New Delhi, India
[3]Assistant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, New Delhi, India

---

Abstract:- *With the advent of cashless economy the use of credit cards as a method of transaction is on the rise and so is the rise in the number of credit card fraud transactions. To prevent this and instil confidence in people that credit card is a secure method of transaction data analysis of a credit card transaction dataset has been done. The process becomes challenging because of two major reasons-first, behaviour of such transactions changes very frequently and secondly because datasets of such transactions are highly skewed.*

The dataset consists of transaction data from Europe and contains over two hundred and eighty thousand transactions. This dataset is analysed and pre-processed. This analysis is used to design and implement a credit card fraud detection algorithm using Machine Learning techniques. The algorithm designed will detect patterns that fraudulent transactions follow and prevent them. Furthermore, the algorithm will have a high accuracy and low processing time

Key Words: **Credit Card Fraud Detection, Random Forest, Local Outlier, SVM, Machine Learning**

1. INTRODUCTION

Credit card frauds are prevalent in modern society. Detecting such transactions is a daunting task when normal procedures are used, hence development of such credit card fraud detection projects has become momentous, in major fields whether it may be academic or the business community. Data mining and machine learning techniques are notable and popular methods used in solving credit fraud detection problem. Fraud detection in credit card involves identifying of those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as genetic algorithm, artificial neural network frequent item set mining, machine learning algorithms, migrating birds optimization algorithm, comparative analysis of logistic regression, SVM, decision tree, neural networks and random forest. These analyse the spending patterns on every card and to figure out any inconsistencies with respect to the "usual" spending patterns. Since humans tend to exhibit specific behaviourist profiles, every cardholder can be represented by a set of patterns containing information such as the typical purchase category, the time since the last purchase, the amount of money spent, etc. deviation from such patterns is recognised as a potential fraud transaction by the system.This paper evaluates some such fraud detection strategies and identifies what are the difficulties in identifying the frauds.

1.1 Difficulties in Credit Card Fraud Detection

Fraud detection systems are prone to several difficulties and challenges enumerated bellow. An effective fraud detection technique should have the ability to address these difficulties in order to achieve best performance.

Imbalanced data: The credit card fraud detection data has imbalanced nature which means that very small percentages of all credit card transactions are fraudulent. This makes the detection of fraud transactions very difficult and imprecise.

Different misclassification importance: In fraud detection, different misclassification errors have varying importance. Misclassification of a normal transaction as fraud is not as harmful as detecting a fraud transaction as normal. Because in the first case the mistake in classification can be identified in further investigations.

Overlapping data: Many transactions may be considered fraudulent, even though they are normal (false positive) and conversely, a fraudulent transaction may seem to be legitimate (false negative). Hence, obtaining a low rate of false positive and false negative is a key challenge of fraud detection systems.

Lack of adaptability: Classification algorithms are usually faced with the problem of adaptability. The supervised and unsupervised fraud detection systems are inefficient in detecting new patterns of normal and fraud behaviors.

Fraud detection cost: The system should take into account both the value of fraudulent transaction that is detected and the cost of preventing it. For example, no revenue is obtained by stopping a fraudulent transaction of a few dollars.

Lack of standard metrics: there is no standard evaluation criterion for assessing and comparing the results of fraud detection systems to decide which is most efficient.

2. RELATED WORK ON CREDIT CARD FRAUD DETECTION

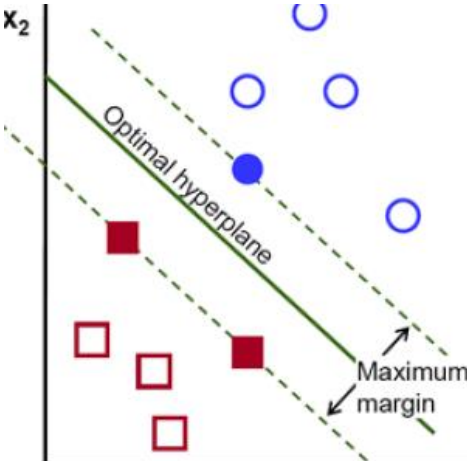
This paper represents a research on European credit card holders, where data normalization is done before doing Cluster Analysis of the dataset. The data is MLP trained and Machine learning algorithms are employed for generating accurate results. Promising results are obtained by using normalized data. This research was based on both supervised and unsupervised learning. Significance of this

paper was to find new methods for fraud detection and to increase the accuracy of results.

This paper investigates the usefulness of applying different learning approaches

2.1 SVM Model (Support Vector Machine)

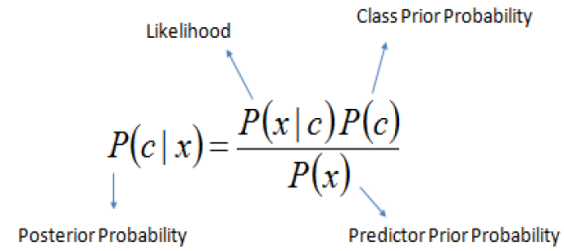
SVM is a popular machine learning algorithm used for regression and classification. It is a supervised learning algorithm that involves analyzing data used for classification and regression. SVM modeling involves two steps, first to train a data set in order to obtain a model & then, to use this model to predict information for testing the data set. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane where the SVM model represents the training data points as points in space and then maps the points such that the points of different classes are divided by a gap that is as wide as possible. Mapping for new data points is done in to the same space and then predicted on which side of the gap they fall



In SVM algorithm, plotting is done as each data item is taken as a point in n-dimensional space where n is number of features, with the value of each feature being the value of a particular coordinate. Then, classification is finally done by locating the hyper-plane that separates the two classes very well.

2.2 Random Forest

Random Forest is an algorithm which is used for classification and regression. It is summarily a collection of decision tree classifiers. Random forest is advantageous over decision tree as it corrects the habit of overfitting to their training set as done in decision tree. In order to train each individual tree a decision tree is built by sampling a random subset of the training set. Each node is then split as per a feature randomly selected from a subset of the full feature set. Since each tree is trained independently, data instances' training is extremely fast in random forest even for large data sets with many features. The Random Forest algorithm has been found to be resistant to overfitting and to provide a good estimate of the generalization error.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Random forest can be used to rank the importance of variables in a regression or classification problem in a natural way.

2.3 Local outlier factor

Local Outlier factor (**LOF**) is an algorithm used for anomaly detection. It is used for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbors.

LOF shares some concepts with DBSCAN and OPTICS such as the concepts of "core distance" and "reachability distance", which are used for local density estimation.

Basic Idea

The local outlier factor is based on the concept of local density, where locality is defined by 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.

The local density is estimated by taking into account the typical distance by 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.

3. Experimental results

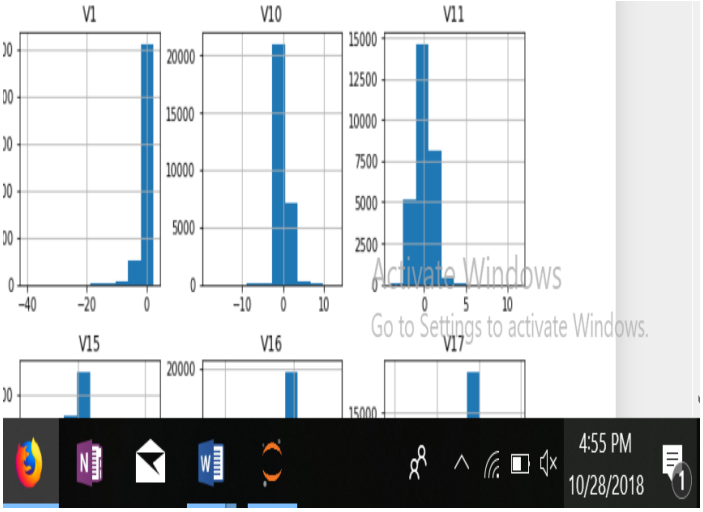
3.1 Observations

1. The data set is highly skewed, consisting of 492 frauds in a total of 284,807 observations. This resulted in only 0.172% fraud cases. This skewed set is justified by the low number of fraudulent transactions.

2. The dataset consists of numerical values from the 28 'Principal Component Analysis (PCA)' transformed features, namely V1 to V28. Furthermore, there is no metadata about the original features provided, so pre-analysis or feature study could not be done.

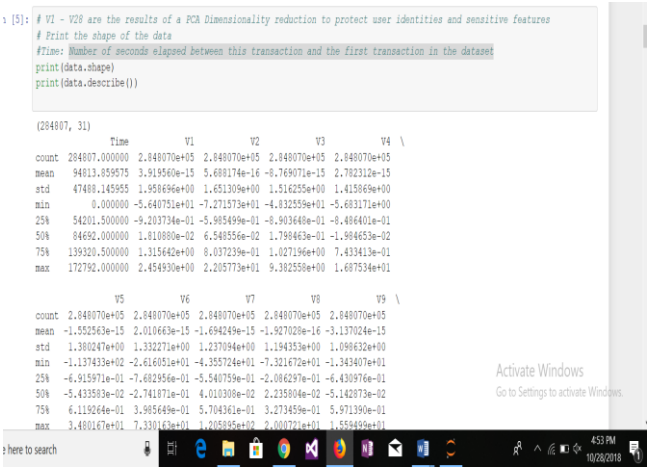
3. The 'Time' and 'Amount' features are not transformed data.

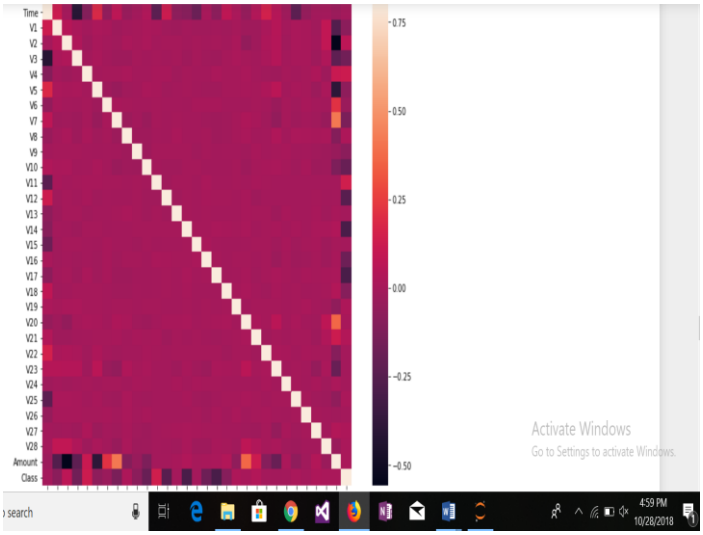
4. There is no missing value in the dataset.



3.2 Inferences drawn

1. Owing to such imbalance in data, an algorithm that does not do any feature analysis and predicts all the transactions as non-frauds will also achieve an accuracy of 99.828%. Therefore, accuracy is not a correct measure of efficiency in our case. We need some other standard of correctness while classifying transactions as fraud or non-fraud.
2. The 'Time' feature does not indicate the actual time of the transaction and is more of a list of the data in chronological order. So we assume that the 'Time' feature has little or no significance in classifying a fraud transaction. Therefore, we eliminate this column from further analysis.
3. Number of seconds elapsed between this transaction and the first transaction in the dataset
4. In Amount, mean is closer to 0, this means we have large no of valid transactions as compared to fraud transactions
5. Mean of fraud transaction is higher than valid transactions, it means when high value transactions are done, it could possibly be a fraud!!
6. The algorithm is harder to train since very few valid cases are present in the dataset
7. Most values (in middle) are closer to 0, depicting no great correlation between v1 to v28 parameters
8. No co-relation between prediction, most fraud transaction done in large amounts
9. No strong co-relation between class, amount or class, time
10. We expected strong co-relation between amount and class since fraudsters would want to gain max from single transaction
11. If that were the case we could have recommended limiting the max amount on transactions, but it is not useful now
12. One reason for this could be to avoid detection and for them to do multiple transaction from single credit card
13. This is odd since the algorithm can predict which transaction was fraudulent and if transaction is done 3-4 times so by limiting transaction amount they would only be reducing their profits
14. The reason behind this could be that police investigation is more likely to bear fruits in case when large amounts of transactions are done
15. Which means they are afraid of being caught, and hence are biased towards safer frauds!!





4. CONCLUSION

The study focuses on analysis of credit card transactions' dataset and identifies patterns that fraudulent credit card transactions follow to help design and implement a credit card fraud detection algorithm.

The data analysis is done by creating histograms of the variables related to the dataset and by creating co-relation matrix of the variables. The data analysis further helps us identify the appropriate Machine Learning techniques applicable to implement the algorithm.

The algorithm is implemented by using Local Outlier Factor Machine Learning technique the results of which shows that the precision of this technique is very high but the accuracy in detecting credit card fraud is low. The algorithm can be used to detect fraudulent transactions which follow the pattern.

REFERENCES

[1] Samaneh Sorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective"

[2] https://en.wikipedia.org/wiki/Local_outlier_factor

[3] https://www.3pillarglobal.com/insights/credit-card-fraud-detection