EVALUATING PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUD DETECTION

Rakesh.B.S[1], Tamilarasan.S[2], Sandeep.K.H[3]
[1, 2,3]Dept. of CSE, Brindavan College of Engineering ,Bengaluru, India
[1]rak.appu@gmail.com, [2]stamilarasan74@rediffmail.com, [1]sandeepkh45@gmail.com

Abstract
Credit card exchanges have become regular spot today as is the cheats related with it. One of the most widely recognized usual way of doing things to do extortion is to get the card data illicitly and use it to make online purchases. For credit card companies and merchants, it is in-practical to identify these deceitful exchanges among a huge number of ordinary exchanges. On the off chance that adequate information is gathered and made accessible, machine learning algorithms can be applied to tackle this issue. In this work, well known directed and unaided machine learning algorithms have been applied to recognize credit card fakes in a profoundly imbalanced dataset. It was discovered that solo machine learning algorithms can deal with the skewness and give best characterization results. For frauds, the credit card is an easy and friendly target because without any risk a significant amount of money is obtained within a short period.
Index Terms— Credit Card ,Extortion, Machine learning, Dataset

I. INTRODUCTION

Because of usability and cash obtaining alternative, Credit cards are being utilized as an instalment instrument by both on the web and disconnected purchasers in a major manner [1]. In any case, this accommodation has accompanied it's a lot of difficulties as well. Credit card based exchanges have become a significant powerless objective for criminals, hackers and perpetrators. Online utilization of Visa requires just the card data to be entered and not present the card genuinely. At times, an additional validation factor of sending a One Time-Password (OTP) is thought of. In all others, where this isn't required, explicitly for global exchanges, it very well may be utilized for unapproved buys. Such utilization is called Card-Not-Present as rather than physical card just subtleties of card are required.

With strategies like card taking, shoulder surfing, purchasing Mastercard data and web traffic sniffing getting conceivable, it is anything but difficult to take the card data. Card holder, giving bank just as vendor every one of the three become survivors of a Visa misrepresentation, as it is one of them who needs to shoulder the weight of extortion. By and large, it is the obligation of card holder to recognize the extortion and report fake exchanges to the giving bank. The bank at that point examines the issue and on the off chance that proof of extortion is discovered, at that point the procedure for switching the credit for the exchange is started. This procedure is non-genuine time and has no assurance of effectively settling the issue [2]. Primary stakeholder is the credit card issuing company as with increase in frauds done on its cards the company's reputation suffers a lot. Thus, it is up to the issuer to implement a fraud prevention and detection mechanism. For preventing frauds, companies issue periodic advisories to its customers on do's and don'ts of safe card usage.

In some cases, extra factors of authentication like OTP and security question are employed to deter fraudulent usage. However, fraud cases are inevitable despite these prevention mechanisms

[3]. Thus, when a fraud occurs and is reported the bank must put in resources for post mortem analysis and try to recover and punish the perpetrator. The turnaround time for this detection has been several days which doesn't prove useful to deter the frauds. Fraud Detection Systems (FDS) are automated machine learning based solutions that credit card companies employ to detect the fraudulent transactions even before end users feedback [4]. Goal of such a system is to detect the fraudulent transaction before it is committed to the database and thus prevent the fraud from taking place. An ideal FDS should also minimize the false detections where a genuine transaction is interrupted causing inconvenience to the end-user.

In the rest of this paper, Section II contain the related work, Section III contain the methodology. Section IV contain the Results and discussion, and Section V contains conclusion.

II . RELATED WORK

Data mining for credit card fraud: A comparative study In this research authors evaluated using two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud[1]. The choice of these two techniques, together with logistic regression, for this study is based on their accessibility for practitioners, ease of use, and noted performance advantages in the literature. The third technique included in this study is logistic regression. It is well-understood, easy to use, and remains one of the most commonly used for data-mining in practice. It thus provides a useful baseline for comparing performance of newer methods.

Logistic regression: Qualitative response models are appropriate when dependent variable is categorical. In this study, our dependent variable fraud is binary, and logistic regression is a widely used technique in such problems. Binary choice models have been used in studying fraud[1].

Support vector machines: Support vector machines (SVMs) are statistical learning techniques that have been found to be very successful in a variety of classification tasks. SVMs work in the high- dimensional feature space without incorporating any additional computational complexity[1] .

Random forests: A random forest model is an ensemble of classification (or regression) trees. Ensembles perform well when individual members are dissimilar, and random forests obtain variation among individual trees using two sources for randomness: first, each tree is built on separate bootstrapped samples of the training data; secondly, only a randomly selected subset of data attributes is considered at each node in building the individual trees[1]

Data mining application in credit card fraud detection system.This study presented an application of artificial neural networks with built-in learning capabilities, which can be used to determine fraudulent and legitimate models from the huge transaction data. A technique of self-organizing artificial neural networks and transaction rules were used to develop a decision aid known as Credit Card Fraud Watch (CCFW), which could run at the background of existing banking software to detect breaches of transaction policy, which cannot be easily detected using other methods [2].

The credit card fraud detection system developed used four clusters of low, high, risky and high risk. Once the transaction is legitimate, it was processed but if any transaction falls into any of these clusters; it was labeled as suspicious/fraudulent. The alert goes off and the reason is given. The fraudulent transaction will not be processed but will be committed to the database. The approach involves the following step. The steps are select an appropriate algorithm; implement the algorithm in software; test the algorithm with known data set; evaluate and refine the algorithm as it is being tested with other known data sets; and show the results.

In this study, the fraud detection system watch consists of two units namely, the withdrawal and deposit unit. Each of the two units is in turn made up of the following subunits: the database interface, the neural network classification, and the visualization. The database interface subunit was tested to ensure that the necessary transaction data was imported and used. The study resulted in a model, which was used to detect abrupt changes in established patterns and recognize typical usage patterns of fraud. The CCF detection system was designed to run at the background of existing banking software and

attempt to discover illegitimate transactions entering on real-time basis. This proved to be very effective and efficient method of discovering fraudulent transaction [2].

Credit card fraud detection model that's handle imbalanced dataset and facilitate knowing of customers' patterns by splitting data into legal (confirmed True transactions) and fraud (Confirmed Fraudster behaviours) patterns to eliminate the problem of imbalanced dataset. Credit card transactions trained using Baum-Welch algorithm in by modeling sequence of operation using Hidden Markov Model (HMM) and dividing transactions into three groups high, medium and low according to transaction amount so that spending profile of cardholder created more easier. Hybrid algorithm proposed in for credit card fraud detection based on combination of Naïve Bayes algorithm with Hidden Markov model and offering OTP (One Time Password) for newly transactions for more security about newly behaviors [3].

An alluring FDS is the one that can recognize a wide range of Mastercard cheats. It chips away at the guideline of learning client explicit card utilization conduct and fraudsters spending designs as opposed to concentrating on extortion vector [9]. On the off chance that drawn out card utilization information of numerous clients and false exchanges happening inside that period are accessible, FDS creation turns into a paired arrangement issue.

The two classes of enthusiasm here are Normal and Fraud exchanges. Existing methodologies of Supervised and Unsupervised AI can be in this way applied to these datasets. In any case, there are a couple of difficulties that come in the method of good characterization results from these calculations. A portion of these difficulties Class Skewness, changes in fraudster conduct to abstain from getting captured, occasional changes in clients conduct, area measurements, absence of truth names, real-time arrangement prerequisites [10].

III. METHODOLOGY

A. *Existing System*

In the above survey it is observed that different types of fraud transactions have been discussed. Also, different types of feature extraction strategies have been addressed like the incremental and the unbalanced nature of the fraud detection problem. Data mining also offers a plethora of techniques to find patterns in data, distinguishing normal from suspicious transactions [5].

Here in our approach some of the traditional methods for classification with unbalanced datasets using sampling techniques shall be applied to balance the dataset. The dataset shall be trained, tested, parameterized and compared using three distinct supervised algorithms such as Logistic Regression, Random Forests and Support Vector Machines and their performance analysis shall be carried out to identify the right algorithm for the dataset [6]

A. *Proposed System*

The objective of our study is to predict the fraudulent transactions with a credit card's details. In the process of predicting the success we have implemented six different machine learning algorithms. We have considered some of the parameters to check the efficiency of a machine learning algorithm. The project has 6 main modules:

a) Gathering data: The Phishing database which is open source includes thousands of instances with 29 parameters. The data source is Kaggle. It has been considered the most comprehensive dataset on Phishing Website Dataset in the world.

b) Preparing the data: Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data. Data preparation is often a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality. We employ pandas' functions to perform data cleaning and filtering out a balanced positive and negative class instances [7].

c) Choosing a model: We have chosen eight different machine learning algorithms here. They are :

(i)Random Forest

(ii)Support Vector Classification

(iii)Artificial Neural Network (ELM Method)
(iv)KNN
(v) Logistic Regression
(vi)Decision Tree
(vii)Gradient Boosting
(viii)XGBoost.

(i) Random forest: Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

(ii) Support Vector Classification: "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

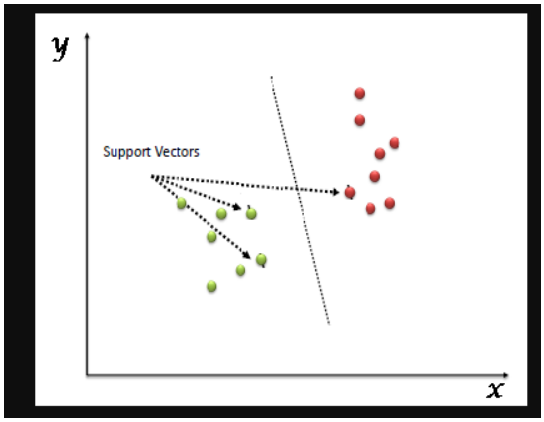

Fig 3a. Support Vector Coordinates

Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

(iii) Artificial Neural Network (ELM Method): Artificial neural networks are one of the main tools used in machine learning. As the "neural" part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize.

(iv) KNN: K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion
detection.

(v) Logistic Regression: It's a classification algorithm that is used where the response variable is *categorical*. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome. E.g. when we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail. This type of a problem is referred to as Binomial Logistic Regression, where the response variable has two values 0 and 1 or pass and fail or true and false. Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.

(vi) Decision Tree: Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

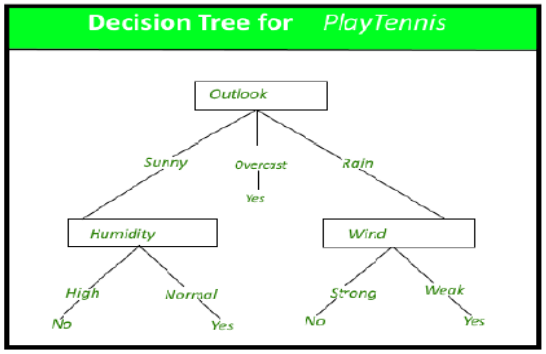

Fig 3b: Decision tree Example

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

(vii) Gradient Boosting: Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. The gradient boosting algorithm (gbm) can be most easily explained by first introducing the AdaBoost Algorithm. The
AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore Tree 1 + Tree 2[8].

(viii) XGBoost: s a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform
all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

d) Training: Training your model is the bulk of machine learning. The objective is to use your training data and incrementally improve the predictions of the model. Each cycle of updating the weights and biases is one training step. In supervised machine learning, the model is built using labelled sample data, while unsupervised machine learning tries to draw inferences from non-labelled data.

e) Evaluation of the model: After training the model comes evaluating the model. This entails testing the machine learning against an unused control dataset to see how it performs. This might be representative of how the model works in the real world, but this does not have to be the case. The larger the number of variables in the real world, the bigger to training and test data should be.

f) Prediction: Once you have gone through the process of collecting data, preparing the data, selecting the model, and training and evaluating the model, it is time to answer questions using predictions. These can be all kinds of predictions, ranging from image recognition to semantics to predictive analytics [9].

IV. RESULTS AND DISCUSSION

Results of test sets are listed below; the results show that Gradient Boosting and Logistic Regression have a higher training accuracy and KNN gives least training accuracy

Fig 4a: Experimental Accuracy values over different Machine learning algorithms

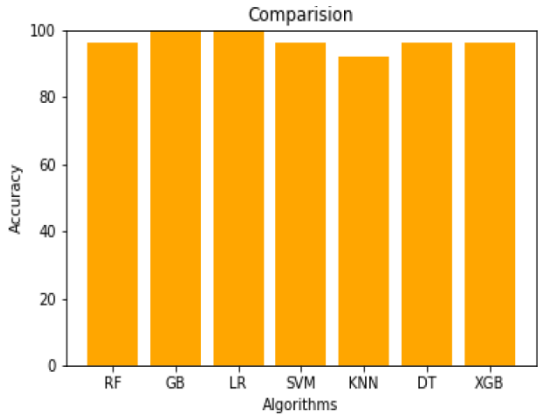

Fig 4 b: Accuracy graph Comparison of different Machine Algorithms

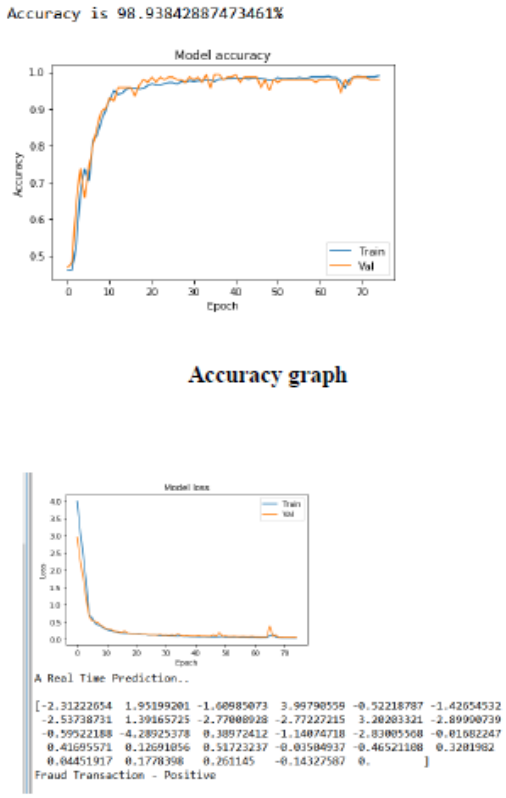

Fig 4 c: Graphical Representation of Accuracy and Model loss graph

From the Fig 4a and 4b the practical implementation of different machine learning algorithms can be seen and through which we got accuracy ranging from 96% to 100%.

Here we can depict that the KNN has the least accuracy (96%) among all whereas through Gradient Boosting and logistic Regression we were able to achieve the highest accuracy (100%) among all the algorithms. Hence we can conclude that Gradient Boosting and Logistic Regression algorithms are best approaches in order to evaluate performance of any credit cards in real time implementation, also from the Figure 4c. the graphical representation of accuracy can be seen through experiments which is of 98.9% ,with model loss graph we can find that percentage of loss was around 2%.

V. CONCLUSION

In this paper we are introducing Credit card fraud detection which is a peculiar classification problem due to very high imbalance in instances of normal and fraudulent transactions as examples. A number of popular algorithms in supervised, ensemble and unsupervised categories were evaluated on different metrics. It is concluded that unsupervised algorithms handle the dataset skewness in better ways and hence perform well over all metrics absolutely and relatively to other techniques. There were few NaN values in the result table where the classifier couldn't detect even a single true positive or true negative value. After training our models on the features, it is estimated that Gradient Boosting and Logistic Regression offer 100%(percentage)of accuracy in on predicting fraudulent transactions of Credit Card transactions. The findings of the presented paper can be used in the coming times to improve predictions for the same.

VI. REFERENCES

[1].Siddhartha Bhattacharyya , Sanjeev Jha, Kurian Tharakunnel, J. Christopher Westland Data mining for credit card fraud: A comparative study,2011

[2].Francisca Nonyelum Ogwueleka Data mining application in credit card fraud detection system,2011 , pp 100-115.

[3].Mohamed Hegazy, Ahmed Madian, Mohamed Ragaie, Enhanced Fraud Miner: Credit Card Fraud Detection using Clustering Data Mining Techniques,2016

[4].Andrea Dal Pozzolo, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, Gianluca Bontempi,2014 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2014, pp. 20-25.

[5] L. Zheng, G. Liu, C. Yan and C. Jiang, "Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity," in

IEEE Transactions on Computational Social Systems, vol. 5, no. 3, pp. 796- 806, Sept. 2018.
[6]Vaishali. Article: Fraud Detection in Credit Card by Clustering Approach. International Journal of Computer Applications 98(3):29-32, July 2014.
[7] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, ”Credit card fraud detection using machine learning techniques: A comparative analysis,” 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017, pp. 1-9.
[8] L. Zheng et al., ”A new credit card fraud detecting method based on behavior certificate,” 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, 2018, pp. 1-6.
[9] SurajPatil*, VarshaNemade, PiyushKumarSoni, Predictive Modelling for Credit Card Fraud Detection Using Data Analytics, International Conference on omputational Intelligence and Data Science (ICCIDS 2018)
[10] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, ”Deep /learning detecting fraud in credit card