Comparative Analysis of Algorithms for Credit Card Fraud Detection

Aashi Maharjan
Department of Computer Science
University of Nevada, Las Vegas
Las Vegas, USA
mahara2@unlv.nevada.edu

Partha Chudal
Department of Computer Science
University of Nevada, Las Vegas
Las Vegas, USA
chudal@unlv.nevada.edu

***Abstract*— With the advancement of technology and E-commerce, the usage of the online transaction has increased dramatically. Credit cards are excessively employed for online purchasing and thus resulting in many serious fraudulent activities. Among them credit card fraud is one of the leading factors for tremendous financial losses which affect millions of people as well as different financial companies. It is important to keep safe and prevent access of our account transaction from intruders. A very powerful detection technique is required to detect such frauds, not after it has happened but before the fraud occurs. The emerging data mining techniques assist in detecting and identifying such fraudulent behaviors. Different approaches to machine learning can be employed to predict suspicious and non-suspicious transactions by implementing numerous classification algorithms. It is important to seek out a better approach to overcome such fraud and learn from past frauds so that we can formulate new methods in the future. In this paper, the main focus is to make the comparative analysis of different learning techniques to analyze credit card fraud detection and using the dataset, it finds out the best algorithm to detect fraud based upon the performance.**

Keywords— suspicious, fraud detection, transaction, Classification algorithms

I. INTRODUCTION

Credit card fraud is regarded as an illegal activity in which anyone tries to use the physical card information without the consent and knowledge of the cardholder [1]. Credit card can be dealt in two ways: one is online fraud that can be detected through mobile phone, internet, web, shopping and other is offline fraud that detects the card which is stolen by using their personal details [3]. Credit card fraud is one of the malicious activities that occur in an online transaction. Generally, credit card fraud refers to the unauthorized access of credit or debit card for making payments. These are simply the fraudulent source of funds used in different transactions. The use and popularity of credit card for online shopping and making various payments has been increased with the rising of E-commerce. These days customers are focusing on the popular payment method with credit card for paying bills and making online shopping in an easy and convenient way. Along with their increasing usage, credit card fraud is increasing day-by-day resulting in a global loss as single fraud can lead to millions of losses [4]. Fraud is at an alarming rate with the emergence of technologies causing huge financial loss. It happens either by stealing someone's credit/debit card physically or when credit card information including card holder's name, card number, expiry date, secured code and so on is stolen directly from physical credit card [2]. Fraud detection predicts whether it is fraudulent or nonfraudulent from the thousands of datasets. Many fraud detections detect the stream of data and learn fraud patterns [5]. Different machine learning models have been developed to detect such activities. There are some reasons due to which machine learning algorithms cannot solve the problem completely.

Some of them are:

- Real information and customers sensitive transaction data is not revealed due to some privacy reasons which becomes insufficient for fraudulent prediction.
- Due to the unavailability of real-world data, it is quite difficult to implement models into the actual detection systems [3].
- The true detection of fraud is complex, and it is hard to find the system that predicts fraudulent transaction instantly and quickly.

The banking fraud can't be completely but at least we can prevent from happening and its occurrence to certain level using machine learning techniques. In this paper, four different classification methods are tested for their dataset in fraud detection: Decision tree, Support Vector Machine, Naïve Bayes and Logistic Regression. Using these algorithms, we are going to make a comparative analysis based upon their performance.

II. BACKGROUND

A. Classification and Machine Learning

Classification is one of the datamining algorithms that classifies items from the large collection of datasets having different features. The main idea of classification is to predict the target class and the popular one is binary classification. Machine learning is the technique to provide computers, the ability to learn without being programmed explicitly. It can be classified as supervised and unsupervised depending upon the datasets provided. Once the model has been

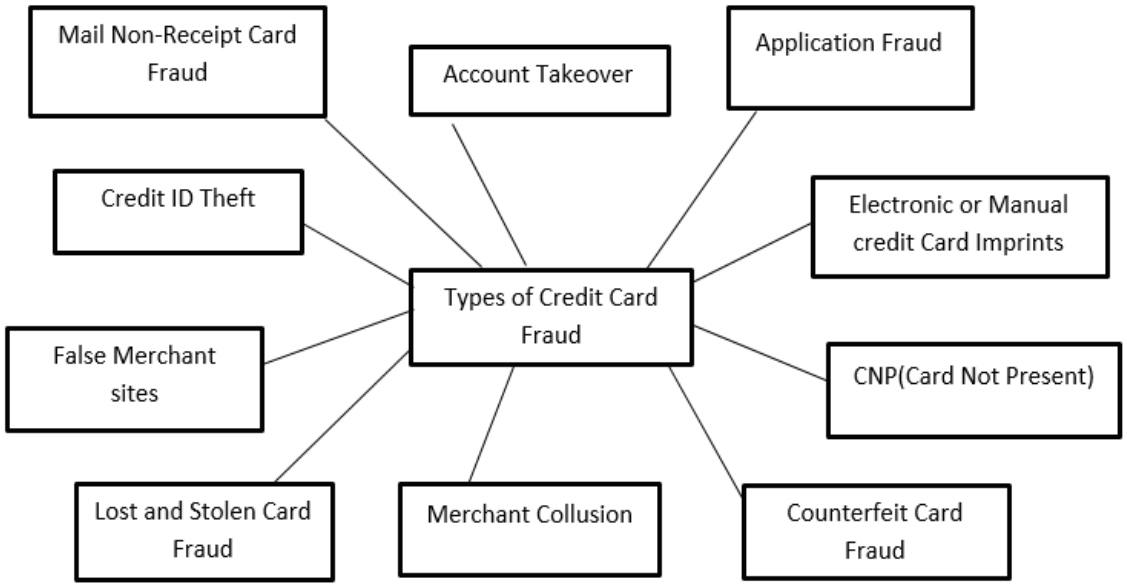

Fig. 1. Types of Credit Card Fraud

trained we will be able to predict the targeted values. We are going to implement supervised machine learning algorithms for classification.

III. Literature Review

It is very important to review history and other related works before understanding different techniques while analyzing credit card fraud detection. It is mandatory to innovate new techniques for day to day increasing frauds. In January 2019, Yashvi Jain et al. presented a paper on comparative analysis for credit card fraud detection in which they implemented different techniques such as Artificial Neural Network, Decision Tree, Fuzzy Logic, K-Nearest Neighbour and some others. They concluded that among all ANN performed best but they have few drawbacks. Optimization techniques will be implemented soon for better results [4]. Vijayshree B. Nipane in [6] implemented a hybrid approach of artificial intelligence to reduce financial losses to a great extent. They proposed a fraud detection system that provides a different level of security, which analyses the spending behavior pattern of the cardholders. He also proposed the architecture for efficient fraud detection and those techniques are useful for finding the data which are not fitted and doesn't belong to the current data pattern by implementing two strong algorithms (Support Vector Machine and Decision Tree). The system provided security to detect patterns in a transaction. Rishi Banerjee in et. al [2] demonstrated the best algorithms to utilize datasets with high imbalances and observed Support vector machine resulted in the best performance rate for credit card fraud detection. The paper also proposed better metrics for determining the false negative rate to measure effectiveness. Masoumeh Zareapoor et.al in [1] focused on the study of nine frauds detection methods to review the methodology of different detection methods based upon credit card. They considered accuracy, speed and cost so further mentioned the weakness and strengths of different techniques.

IV. Classification Techniques

A. Support Vector Machines

A Support Vector Machine is a very popular method for classification and one of the discriminative classifiers, which is defined by separating hyperplane. If quadratic/log-loss is replaced with some other function, then the solution will be sparse. Then the predictions depend upon the subset of training data which is known as support vector and the combination of kernel trick and modified loss function is known as SVM [9]. Generally, we use LIBSVM as the library for SVMs in which we built a

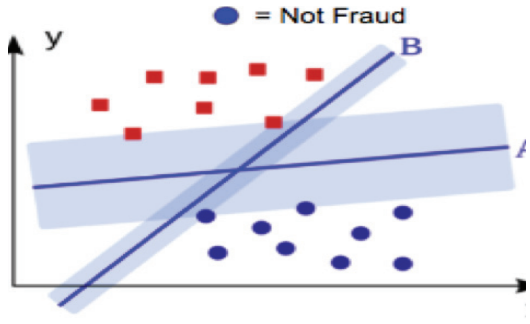

Fig. 2. Example of a SVM[10]

model using training datasets and then based on testing datasets it can predict different information [6]. SVM depends mainly upon decision planes that

separate different classes. This method needs a lot of training datasets so that we can predict better accuracy. Two properties known as kernel representation and margin optimization can be used to learn complex regions. Using RBF kernel, we can find a linear model so that all training instances can be correctly classified into two classes -positive and negative [1]. There can be many lines that will perfectly separate training data, but the main idea is to pick the best one that maximize the margin. We need to confirm each point resides in correct place of boundary. Similarly, in credit card fraud detection it determines whether each test instance lies within learned region, which indicates it is normal otherwise, anomalous. SVMs also support for multidimensional features [8]. This model also provides better time efficiency and higher accuracy as compared to other algorithms.

B. Logistic regression

Logistic regression is regarded the popular approach to classification as shown in fig. 4. This algorithm uses both regression and sigmoid function to perform binary classification depending upon various factors [2]. They are easy to fit data, meaning that the algorithms are simple to implement, and are very fast that takes linear time [9]. The model is used for predicting binomial and multinomial outcomes, which uses a sigmoid function to estimates the values of parameters coefficient. It also measures values of different attributes and decides whether to process such transaction or not [4]. We can compute linear combination of inputs by defining:

(1)

Where indicates the sigmoid function, which is also known as the logistic or logit function.

(2)

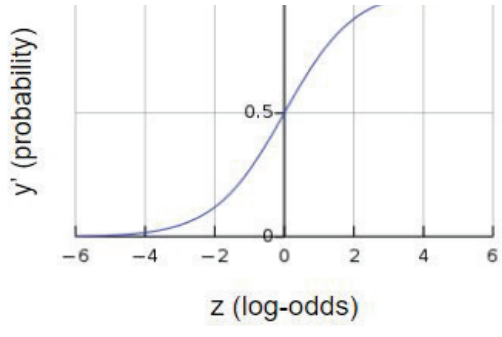

Fig. 3. General Form of Logistic (Sigmoid) Function [11]

LR models are easy to interpret and we can define log odds as:

(3)

Defining log odds, the equation can be written as:

$$z = b + w_1x_1 + w_2x_2 + \ldots. + w_Nx_N \qquad (4)$$

z represents the log-odds of the example and w is the weighted

values. LR models can easily be extended to handle non-linear decision boundaries by using kernels. [9]

C. Naïve Bayes

Naïve Bayes is also one of the powerful algorithms for classification, which can be used for automated detection of events. It falls under a supervised learning method that uses training dataset having targeted classes so that it can predict the outcomes or class of future instances. This model uses conditional probability so that we can calculate the probability of an event using its prior knowledge.

(5)

Naive Bayes classifiers is based on Bayes' theorem, and the term naive assumes that the features in a dataset are mutually independent [data science]. The algorithm is simple, easy and efficient to predict classes in both binary and multiclass classification problems. This model has been used to predict whether the given transaction is fraud or non-fraud in less time than other algorithms.

D. Decision Trees:

A Decision tree is simply a tree, which consists of root node as the top node, branch node that indicates the outcome of a test, leaf node that holds class label and an internal node that indicates a test on the attribute. This model is used for classification and prediction. The classifier breaks down the complex problem into many simpler ones or subproblems by constructing decision trees and then pruning the subtrees. The main idea of the algorithm is to build a decision tree at first and then apply decision rules to determine different classes. The input data must have a class label as fraudulent or non-fraudulent in case of credit card fraud detection. Starting from the root node as a single node with all training datasets, it splits the node into different child nodes either in binary or multiple fashions. Before classification, it reads decision rule one by one from the decision table then finds the perfect match. If no match is found, then it chooses the rule having the highest risk level and thus decides either fraud or not [12]. It is very

reliable and highly efficient as compared to other algorithms [1].

V. Procedures and Experiments

A. Datasets

The dataset has to be collected and properly analyzed before implementing model. The real datasets were prepared during research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (University Libre de Brucellas) on big data mining and fraud detection. The datasets include numerous transactions made by credit cards in September 2013 by European cardholders and present transactions that occurred in two days, where there are 492 frauds out of 284,807 transactions. The dataset is not balanced. Among all transactions, the positive class (frauds) account for 0.172%. It contains only numerical values as input variables and are the result of a PCA transformation. Unfortunately, the original features are not revealed due to privacy issues. The total principal components obtained with PCA are V1, V2, ... V28. But the features 'Time' and 'Amount' have not been transformed.

- Time: The Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.
- Amount: The feature 'Amount' is the transaction Amount and can be used for example-dependent cost-sensitive learning.
- Class: The Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

B. PCA Process:

Principal component analysis (*PCA*) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. When data have a standard deviation too much larger than mean, it could be useful to apply a transformation to reduce it. Furthermore, when there are many variables, we can build a new set of uncorrelated quantities that explain as better as possible the variance. This is can be done, applying a Principal Component Analysis (PCA) [7].

C. Model Training and Data Set:

The total data set has been splited into training and testing data. 80% of the dataset has been used for training and the remaining 20% has been used for testing. In this way, models were trained and tested on a CSV file containing different attributes. For all algorithms, accuracy, precision, the recall was calculated. As logistic regression performs binary classification, it worked well with the given dataset. It has the target variable to the output probability is converted into 0 for negative and 1 for positive which can be checked with targets. Since Naive Bayes takes every factor independently so every numerical factor was tested independently. In the Support vector machine, various data were plotted to find the optimal line before predicting transaction as fraudulent or not. Decision trees applied decision rules to find the best match in both classification and prediction

VI. Comparative Analysis

A comparison table was prepared to compare different machine learning models. We calculated the true positive rate, false positive rate, time and accuracy generated by the systems. We also compared the ROC curve, error rate, and kappa statistics for different classification techniques to measure the performance and make good analysis.

- Accuracy: It represents the fraction of the total number of transactions that have been detected correctly (fraudulent and non-fraudulent).
- True Positive: It is the number of transactions to represent fraudulent transactions were correctly classified as fraudulent.
- False Positive Rate: The false positive rate is represented as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events.

$$FP = \frac{FP}{FP + TN}$$

- Precision: Precision represents precise/accuracy of model. Precision is a good measure to determine when the cost of false positive is high.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Recall: Recall calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). It should be the model metric that we use our best model when there is high cost associated with False Negative.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- ROC curve: It stands for Receiver Operating Characteristic (ROC) curve. In this, for different cut-off points, the true positive rate (Sensitivity)is plotted in function of the false positive rate (100-Specificity). If the ROC curve is closer to the upper left corner, then the overall accuracy of the test will be high. The area under the ROC curve (AUC) is a measure that parameter can distinguish between two groups (fraudulent/non-fraudulent).

TABLE I. COMPARATIVE ANALYSIS OF CREDIT CARD FRAUD DETECTION

SN	Parameters	Support Vector Machine (SVM)	Logistic Regression	Naïve Bayes	Decision Tree
1	Accuracy	99.93%	99.91%	97.67%	99.83%
2	Precision	99.90%	99.99%	99.80%	99.90%
3	TP Rate	99.90%	99.90%	97.70%	99.90%
4	FP Rate	24%	39.60%	15.30%	36.90%
5	Recall	99.99%	99.99%	97.70%	99.90%
6	ROC Area	88.20%	97.60%	96.60%	95.90%
7	Kappa Statistics	83.88%	72.78%	12.89%	74.79%
8	Speed of Detection	41.35 s	21.55 s	3.96 s	52.81 s

- Kappa Statistics: The most commonly used kappa statistics test interrater reliability. The kappa can range from −1 to +1. The rater reliability is very important in Kappa statistics that gives the correct representations of the variables measured while collecting data in study.
- Speed of detection: It represents the amount of time taken by the model for both training and testing the datasets. The models were fast and efficient, and the time taken was measured in seconds.

Three of the detection systems (Support Vector Machine, Logistic Regression, Decision Tree) showed 99% accuracy for the given datasets. Naïve Bayes has a bit low accuracy as compared to others but has high processing speed. All those discussed in this paper have their own weakness and strength. For Decision Tree, it took a long time to build the model as compared to others. Logistic regression is not expensive to train whereas naïve Bayes and decision tree are expensive to train. All the algorithms have good precision, which is around 99%. Kappa Statistics is fine for logistic, SVM and Decision Trees. High detection rate(precision) is offered by all four algorithms. Low False positive rate is given by Naïve Bayes whereas logistic regression gave high false positive rate than others. For Support Vector Machine FP rate is medium. If we plot ROC area for logistic regression, then it will give finest one than others. Since the ROC area for logistic regression is about 97.60%. It is notable that Naïve Bayes has the fastest speed of detection and others have average speed. The time take for building and testing model is fast. Sometimes there occurs a huge gap between models for detection due to the true unavailability of complete data, as they are not revealed due to privacy. Although researchers have been going on at an increasing pace, we lack a strong and yet powerful algorithm that can perform in all situations. We also need efficient parameters to evaluate and measure performance that gives better comparative results among different approaches.

VII. CONCLUSION AND FUTURE SCOPE

The main idea of this paper was to make a good comparison analysis among different algorithms based on various important parameters. Basically, in machine learning, accuracy is regarded as the important measure. The best accuracy i.e. 99.93% is achieved by using SVM model and other models were above 95% which is pretty good. So SVM came out to be the most successful. Today, everyone is seeking better approach and technology that can detect fraudulent transaction on the spot where it is happening so that it can stopped in a minimum cost. So, the major task is to build the system that is accurate, precise, efficient and more over fast detecting system. It is required to find out the perfect solution to fulfill the gaps between algorithms and to overcome the drawbacks by creating the hybrids of different techniques. Sometimes, depending upon the environment and the applications, it might give higher accuracy. In order to increase the accuracy, we can try unsupervised learning in future to achieve better performance.

REFERENCES

[1] Masoumeh Zareapoor, Seeja.K.R, and M.Afshar.Alam, "Analysis of Credit Card Fraud Detection Techniques:based on Certain Design Criteria,"International Journal of Computer Applications(0975-8887), vol. 52-No.3, pp. 35–42,August 2012.

[2] New Jersey's Governor's school of Engineering and Technology July 27,2018.

[3] Gaikwad, Jyoti R., et al. "Credit Card Fraud Detection using Decision Tree Induction Algorithm." International Journal of Innovative Technology and Exploring Engineering (IJITEE)4(2014).

[4] Yashvi Jain, NamrataTiwari, ShripriyaDubey, Sarika Jain, "A Comparative Analysis of Various Credit Card Fraud Detection Techniques,"International Journal of Recent Technology and Engineering(2277-3878), vol. 7-No.5S2, pp. 402–407,January 2019.

[5] Ankur Rohilla, "Comparative Analysis of Various Classification Algorithms in the Case of Fraud Detection,"International Journal of Engineering Research and Technology(IJERT)(2278-0181), vol. 6-No.09, pp. 118–122,September 2017.

[6] Vijayshree B. Nipane, Poonam S. Kalinge, Dipali Vidhate, Kunal War,

Bhagyashree P. Deshpande, "Fraudulent Detection in Credit Card System Using SVM and Decision Tree", IJSDR(2455-2631),vol. 1-No.5, pp. 590–594,Mayt 2016.

[7] Hastie, Trevor, Trevor Hastie, Robert Tibshirani and J. H. Friedman. The Elements of Statistical Learning: Data Mining , Inference, and Prediction. New York: Springer, 2001.

[8] Sunil Bhatia, Rashmi Bajaj, and Santosh Hazari, “Analysis of Credit Card Fraud Detection Techniques, ”International Journal of Science and Research(IJSR) (2319-7064), vol. 5-No.3, pp. 1302–1307, March 2016.

[9] Kevin P. Murphy, snf Francis Bach “Machine learning a probabilistic approach,” 1st ed., MIT Press, 2010.

[10] Support Vector Machines: A simple Explanation, KDnuggets Anaytics Big Data, Data Mining and Data Science[online].

[11] File: Sigmoid -function -2 svg, File:Cholestrol(chemical Structure).svg- Wikimedia Commons.[online].

[12] Dipti D. Patil, V.M. Wadhai, J.A. Gokhale “Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy”. International Journal of Computer Applications. Volume 11-No.2, pp. 23-30, 2010.