International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 08 | Aug - 2021 ISSN: 2582-3930

A Brief Analysis on Credit Card Fraud Detection

Devansh Akhilesh Shukla, Vishal Rajendra Patil, Nikhil Tiwari and Ajeet Jain

Mumbai University , India

***Abstract*— Growing E-commerce and many other pay first services have increased the risk of frauds. To counter this deceit various service providers are trying to improvise on their service. Researchers started using different machine learning methods to detect and analyse frauds in online transactions.Different Machine learning techniques like K-Nearest Neighbors,Logistic Regression,Random Forest Tree and some boosting techniques like XG Boost and a brief analysis is done based on their accuracy. The main aim of the paper is to design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyse the past transaction details of the customers and extract the behavioural patterns.**

Keywords—Fraud Detection,Credit Card,K-Nearest Neighbors,Logistic Regression,Random Forest Tree,XG Boost.

1. Introduction

Service providers are facing new challenges with most of the services going online promoters fear loss of trust from their customers This is one of the major issues of this contemporary world where intruders make use of the slightest loophole in the system to carry out fraud transactions.Credit card fraud can be done in many ways. There are several types of credit card fraud. Some of them are Merchant collision, site cloning and credit card generator or some man in the middle activity. Government and parties involved are putting substantial efforts but still frauds are on rise as the number of transactions every year are outnumbering previous year. Due to digitization, most people are now preferring online shopping which requires payment transactions through credit card, debit card, or Net banking rather than going for the regular mode. And as we know, in online payments the only requirement is about sensitive information like passwords, CVV numbers, OTP, etc.There is no requirement for any physical card. But in any case, if this sensitive information is compromised then it will lead to huge losses for both the service providers as well as for the customer. Therefore, in such a scenario, a credit card fraud detection technique is required to tackle the challenge faced by the cardholder. This makes tech research and development wings to invest heavily on machine learning algorithms.

If we look at the below chart we can see that the fraud has been increasing gradually.

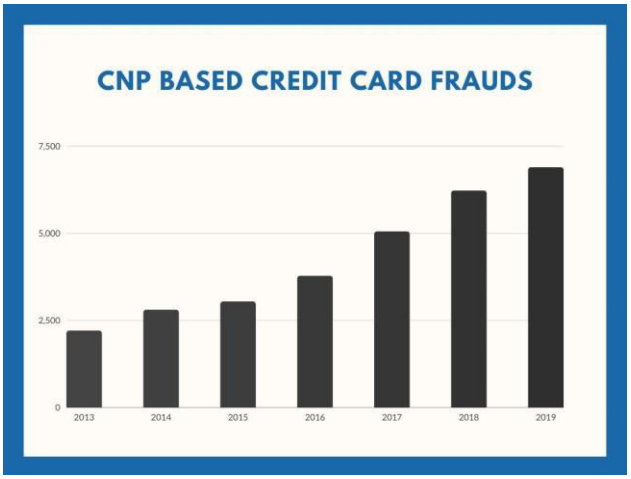

Fig1.CNP bases credit card frauds

2. Literature Survey

2.1 A Cost-Sensitive Decision Tree Approach for Fraud Detection

As information technology is developing the fraud is also increasing as a result financial loss due to fraud is also very large. A cost sensitive decision tree approach has been used for fraud detection. A cost called misclassification cost is used which is taken as varying as well as priorities of the fraud also differs according to individual records. So common performance metrics such as accuracy, True Positive Rate (TPR) or even area Under Curve cannot be used to evaluate the performance of the models because they accept each fraud as having the same priority regardless of the amount of that fraudulent transaction or the available usable limit of the card used in the transaction at that time. For avoiding this a new performance metric which prioritizes each fraudulent

transaction in a meaningful way and it also checks the performance of the model in minimizing the total financial loss. The measure used is Saved Loss Rate (SLR) which is the saved percentage of the potential financial loss that is the sum of the available usable limits of the cards from which fraudulent transactions are committed. Different methods are used for cost sensitivity. They mainly include the machine learning approach, decision tree approach. In the machine learning approach two techniques called oversampling and undersampling are performed, in which the latter obtained a good result. In decision tree approach, decision tree algorithms are used in which misclassification cost is considered in pruning step. A cost matrix is used to find the varying misclassification cost. After finding the misclassification cost the one with minimum value is used. By finding the misclassification cost not only the node value is obtained but also it predicts whether the transaction is fraudulent or not. This study using misclassification cost has made a significant improvement in fraud detection

2.2 Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective

In this study mainly two approaches namely misuses (supervised) and anomaly detection (unsupervised) techniques are being used. After this a classification is also used for checking the capability to process categorical and numerical data. In the first approach the data is classified as fraud based on previous data. With the help of this dataset classification models are also created, which can predict whether the data is fraud or not. The different classification models used are decision tree, neural network, rule induction etc. This has obtained a successful result and this approach is also called a misuse approach. While the second approach is based on account behavior. A transaction is said to be fraudulent if it possesses the features opposite to the user's normal behavior. The behavior of the user's model is extracted and accordingly classified as fraudulent or not. This technique of finding fraud is also called anomaly detection.

2.3 Credit Card Fraud Detection Using Hidden Markov Model

As the E-commerce technology is increasing day by day the use of credit cards has also increased. As a result of this, credit cards are also increasing. In all fraud detection systems, fraud will be detected only after the fraud has taken place. In this study a sequence of operations are modelled using Hidden Markov Model (HMM) and this can be used for the detection of fraud. It is trained with the normal behavior of the card holder. If the incoming transaction is not accepted by the trained HMM with high probability it is considered as fraudulent otherwise not. A hidden Markov Model represents a finite number of states with sufficiently high probability. The transition between the states are handled by these probability values. A possible outcome will be generated based on the probability distribution. This outcome will be visible to the external users that is the states are hidden to the users hence the name. It is a perfect solution for predicting fraud transactions in addition it also provides an extreme decrease in the number of false positive transactions recognized by fraud detection systems. For prediction purposes three values are being used namely low, medium, high.

2.4 Detecting Credit Card Fraud by Modified Fisher Discriminant Analysis

This employee is a linear discriminant called fisher Discriminant. The Linear discriminant is a supervised learning algorithm in which the input region is divided into boundaries called decision boundaries or decision surfaces. The discriminant algorithm is modified to update the weights. This method tries to find the best dimensional hyperplane by which the within class variance is minimized to reduce the overlap and between class variance is maximized. It is a kind of supervised learning algorithm in which the input region is divided into decision regions where the boundaries are called decision boundaries.These boundaries are linear functions of input vector x. There is actually a comparison between fisher discriminant and modified fisher discriminant. FDA captures more positive cases whereas modified FDA gets more profit by classifying the most profitable transactions. In total FDA can give more fraud transactions whereas modified FDA relies on maximizing total profit.

3. Proposed Work

A kaggle dataset is used to do the brief analysis of credit card fraud and a suitable machine learning technique is used which has the highest accuracy. For every machine learning model a F1 -score and accuracy is calculated and a brief analysis is done. USing the machine learning techniques we can make out that the model requires bagging or boosting to improve the accuracy.

3.1 System Architecture

The system architecture is given in Figure 1.

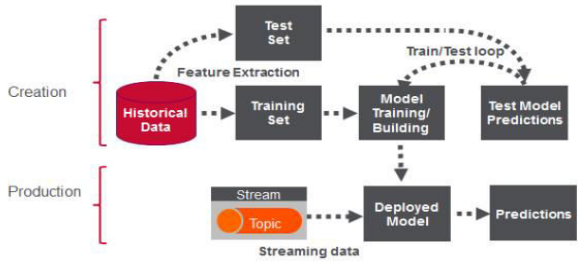

Fig. 2 Proposed system architecture

3.1.1. Training Set: The training set is the material through which the computer learns how to process information. Machine learning uses algorithms – it mimics the abilities of the human brain to take in diverse inputs and weigh them, in order to produce activations in the brain, in the individual neurons.

3.1.2. Test Set: A test set in machine learning is a secondary (or tertiary) data set that is used to test a machine learning program after it has been trained on an initial training data set. ... A test set is also known as a test data set or test data.

3.1.3. Model Training: Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.These are some of the machine learning models which are used:

A)Decision Tree Algorithm-Decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter / differentiator in input variables.

B)K-Nearest Neighbour Algorithm-The concept of K-nearest neighbour analysis has been used in several anomaly detection techniques. One of the best classifier algorithms that have been used in credit card fraud detection is k- nearest neighbour algorithm, which is a supervised learning algorithm where the result of a new instance query is classified based on the majority of K-Nearest Neighbour category.

C)Support Vector Machines-Support vector machine is a method used in pattern recognition and classification. It is a classifier to predict or classify patterns into two categories; fraudulent or non fraudulent. It is well suited for binary classifications.SVM is correlated to and has the basics of non-parametric applied statistics, neural networks and machine learning. SVM is described in the following passage. It was developed from the theory of Structural Risk Minimization. In a binary classification problem, the decision function of SVM is shown in Eq.(1), f x sgn(x.w) b where, x is the input vector which contains weight and b is a constant. Eq.(1) is used to find the decision boundary between two classes. The parameter values of w and b have to be learned by the SVM on the training Phase and b are derived by maximizing the margin of separation between the two classes. The criterion used by SVM is based on the margin maximization between the two classes.

D. Random Forest Algorithm: Random Forest is also called Random Decision Forest (RFA) which is used for Classification, Regression and other tasks that are performed by constructing multiple decision trees. This Random Forest Algorithm is based on supervised learning and the major advantage of this algorithm is that it can be used for both Classification and Regression. Random Forest Algorithm gives you better accuracy when compared with all other existing systems and this is the most commonly used algorithm. In this paper the use of Random forest algorithm in credit card fraud detection can give you accuracy of about 90 to 95%.

E. XG Boost: XGBoost has an inherent ability to handle missing values. When XGBoost encounters nodes at lost value, it tries to split left & right hands & learn all ways to the highest loss. This is when the test runs on the data. XGBoost, namely Extreme Gradient Boosting (XGBoost)algo is supervised learning algo based on synthesis. It includes (written) an objective function consisting of a loss function (d) & regularization term (β):

$$\Omega(\theta) = \underbrace{\sum_{i=1}^{n} d(y_i, \hat{y}_i)}_{Loss} + \underbrace{\sum_{k=1}^{K} \beta(f_k)}_{regularization} ,$$

Fig 3.XGBoost equation

Where yi is predictive value, n is a training set for no. of instances, K is no. of trees generated & fk is synthesis from a tree. The term Regularization is defined as:

$$\beta(f_t) = \gamma T + \frac{1}{2}\left[\alpha \sum_{j=1}^{T} | c_j | + \lambda \sum_{j=1}^{T} c_j^2\right],$$

Fig4.Regularization equation

3.1.4 Deployed Model: Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is one of the last stages in the machine learning life cycle and can be one of the most cumbersome..

4. Results and Evaluation

4.1 Analysis of Data:.After exploring the dataset and doing feature extraction the count of frauds and the percentage was calculated it is given as follows:

Total number of cases	Non-fraud cases	Fraud cases	% Fraud Cases
284807	284315	492	0.1738

Figure 5.Analysis of fraud cases

4.2 Evaluation:
After creating different models for the dataset an analysis has been done by which we have come to a conclusion that KNN(K-Nearest Neighbours) algorithm has the best accuracy. Some bagging and boosting techniques like XG Boost have been used but that didn't have a great impact as compared to KNN.

ML techniques	F1 Score	Accuracy Score
Decision Tree	0.8105	0.9993
KNN	0.7865	0.9993
SVM	0.5	0.9987
Random Forest	0.7657	0.9992
XGBoost	0.8449	0.9995

Fig 6. Accuracy and F1-score for different techniques

The F1-score for KNN(K-Nearest Neighbours) is 0.7865 which is quite good. The best F1-score was provided by XGBoost.

ACKNOWLEDGMENT
It is our privilege to express our sincerest regards to our supervisor Ajeet Jain for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work.
.

REFERENCES

1. Rahul Powar, Rohan Dawkhar, Pratichi "CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING" September 2020
2. M. Suresh Kumar, V Soundarya "Credit Card Fraud Detection Using Random Forest Algorithm" IEEE Xplore Feb-2019
3. Rahul Goyal, Amit Kumar Manjhvar, Vikas Sejwar "Credit Card Fraud Detection in Data Mining using XGBoost Classifier" may-2020
4. V. Dheepa and R. Dhanapal Research and Development Centre, Bharathiar University, India "BEHAVIOR BASED CREDIT CARD FRAUD DETECTION USING SUPPORT VECTOR MACHINES" July-2012
5. Ms. Amruta D. Pawar1, Prof. Prakash N. Kalavadekar2,Ms. Swapnali " A Survey on Outlier Detection Techniques for Credit Card Fraud Detection" Apr-2014
6. Vladimir ZASLAVSKY and Anna STRIZHAK INFORMATION & SECURITY. An International Journal, Vol.18: 2006 "CREDIT CARD FRAUD DETECTION USING SELF-ORGANIZING MAPS"
7. Jiang, Changjun et al. "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism." IEEE Internet of Things Journal 5 (2018): 3637-3647.
8. Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).
9. Randhawa, Kuldeep, et al. "Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, vol. 6, 2018, pp. 14277–14284., doi:10.1109/access.2018.2806420.
10.https://www.ftc.gov/news-events/press-releases/2019/02/imposter-scams-top-complaints-made-ftc-2018.
11.https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.