Proceedings of the Second International Conference on Smart Electronics and Communication (ICOSEC).
IEEE Xplore Part Number: CFP21V90-ART; ISBN: 978-1-6654-3368-6

An Approach for Detecting Frauds in E-Commerce Transactions using Machine Learning Techniques

Abhirami K
Department of computer science and Engineering
School of Engineering and Technology
Christ(Deemed to be University),Bangalore, India
Email: abhirami[dot]k@btech.christuniversity.in

Alok Kumar Pani
Department of computer science and Engineering
School of Engineering and Technology
Christ(Deemed to be University),Bangalore, India
Email: alok[dot]kumar@christuniversity.in

Manohar M
Department of computer science and Engineering
School of Engineering and Technology
Christ(Deemed to be University),Bangalore, India
Email: manohar[dot]m@christuniversity.in

Pankaj Kumar
Department of Computer Science and Engineering
Motihari college of Engineering, Motihari, Bihar
Email : Pankaj[dot]bih@gov.in

2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) | 978-1-6654-3368-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOSEC51865.2021.9591720

Abstract*— This paper is primarily focused on E-commerce fraud detection using machine learning techniques. There are many different ways to detect E-commerce fraud using machine learning approach. In this work, comparison study is conducted between various available machine learning algorithms to detect the online frauds. During the comparative study, focus is underlined on comparison of all the algorithms to identify the fraud transactions. When compared to other algorithms, such as support vector machine, Decision Tree, K-nearest neighbour and Random Forest, it has been observed that Logistic regression gives better result among all machine learning algorithms.

Keywords—Credit card; Frauds; Fraud detection ;Logistic regression; K-Nearest Neighbor; Support Vector Machine(SVM); Random Forest(RF); Decision Tree(DT); Supervised machine learning algorithm.

I. INTRODUCTION

Typical organization looses 5% of its average revenue to frauds each year. RTI report has shown that 2480 frauds involving a huge sum of rupees 32,000 crores from 18 public sector banks in India. As such the frauds pose a serious concern in all the companies. Fraud is a lie that is used to illegally limit the rights of other people, entities, or money. In 2017 to 2018 total of 911 credit card frauds amounting to 65,26 crores according to RBI report which has been illegally transferred from different banks. Due to cashless transaction every people use ATM card and credit card for transaction, so fraud can also be increased. Billions of dollars are lost every year by fraudulent activities. The design of fraud detection algorithms is an important factor in lowering losses, and more and more algorithms are relying on powerful machine learning approaches to help fraud detectors.

the time. Credit cards are used to purchase goods and services using virtual and physical cards, with virtual cards being used for online transactions and physical cards being used for offline transactions. When making a physical card-based purchase, the cardholder must physically present his card to the seller. If the cardholder does not discover the loss of funds on the card, the credit card firm may suffer a financial loss. Fraudsters simply require a few pieces of information to carry out a fraudulent transaction using the online payment method, such as the security code, credit card number, expiration date, and so on.

There are various forms of scams, such as online frauds, in which a person's account is hacked and all of the money is transferred illegally, and offline frauds, in which a person's account is hacked and all of the money is transferred unlawfully. The next form of fraud is credit card fraud, which is the most prevalent type, in which the owner's credit card is stolen or the credit card number is used to make bogus transactions, and the threat and theft of inventory are the other types of fraud. Fraud detection is a problem that applies to a variety of businesses, including banking and finance, insurance, government agencies, and low enforcement, among others. Millions of transactions can be searched using advanced data mining algorithms to find patterns and detect fraudulent transactions. The method of detecting fraudulent rent transactions is known as fraud detection. Customers are not charged for things they did not purchase using a credit card fraud detection approach.

There can be a rule based as well as a data science approach for fraud detection system. In rule based approach, Algorithm written by fraud analyst which are based on a strict rule or rules, changes for detecting a new fraud are all done manually. In this strategy, as the number of customers and data grows, so does the amount of human labour which is

required. As a result, we may conclude that the rule-based approach is time consuming and costly. Some disadvantages of this approach can be that it cannot recognize the hidden pattern, cannot predict fraud by going beyond the rules, cannot respond to new situations, not trained on or explicitly programmed. The data science approach, on the other hand, uses it to take advantage of the vast quantity of data collected from online transactions and model it in a way that helps us to detect fraud in future transactions. We use a variety of data science approaches for this, including machine learning and deep neural networks (DNNs), which are apparent options.

We employed machine learning approach in this paper, which is the scientific study of algorithms and static models that computer systems employ to do a certain task effectively without utilising explicit instructions, instead relying on patterns and interfaces. In order to produce a prediction and a judgement, a machine learning algorithm creates a mathematical model based on sample data, known as training data. Email filtering, face recognition, and other applications use machine learning algorithms.

II. LITERATURE SURVEY

Fraud is defined as an intentional deception carried out for monetary or other benefit. With the increased use of credit cards and internet transactions, fraud is on the rise. This is something that should be taken seriously and can be mitigated by building a system to detect fraudulent and legitimate transactions. This type of technology is based on user-specific card usage patterns [1]. It is difficult to detect and prevent frauds during online transactions. To detect frauds, artificial intelligence or machine learning can be utilised. The Random Forest machine learning algorithm outperforms the decision tree and XGBOOST approaches when accuracy is compared [2]. However, in our situation, the decision tree method is more precise and accurate than the Random Forest approach and provides the best results. Data analytics' goal is to find hidden patterns and use them to make better decisions in a variety of situations [3].

The support vector machine, Nave Bayes, Logistic Regression, and K-Nearest Neighbor algorithms were used in the data set measurement method, variable selection, and detection methods, all appear to have done a good job of detecting online payment fraud, The data set measurement method, variable selection, and detection algorithms used, particularly support vector machine, Nave Bayes, Logistic Regression, and K-Nearest Neighbor, all have a significant impact on credit card payment fraud detection efficiency [4]. Unsupervised approaches handle dataset skewness better than supervised algorithms, outperforming them on all measures, both in absolute terms and in comparison, to other techniques [5]. [6] Authors discussed about various ways of detection and verification methods to identify the fraudulent transactions. [7] Author recommends two phases for the detection of online frauds. In first phase, result will be evaluated for its genuineness. In the second level, author used rule-based techniques for detection of online frauds. [8] proposed tools for identification of the frauds. The author mainly concentrates on reducing the human interaction for the fraud detection process for the online applications. [9] Author analyzed the relationship between individuals' characteristics and the use of online payment services in daily life in the Middle East, North Africa, Afghanistan and Pakistan. [10,11,12,13,14,15,16] discussed the similar aspects of fraud detection in their research work.

III. MACHINE LEARNING:

Machine learning is a vast term that refers to a wide range of algorithms and approaches for classification, regression, clustering, and anomaly detection. It is an artificial intelligence application that allows a system or computer to learn and decide for itself; it will learn, improve, and adapt without having to be programmed to execute any tasks [1]. The three main categories of machine learning are supervised learning, unsupervised learning, and reinforcement learning. The classification task, also known as supervised learning, is used to predict the value of a response variable or the label of a collection of pre-determined categories. As a result, the algorithm learns to predict the unknown sample using data from samples with known response variables and labels in supervised learning. So we're dealing with the categorization problem in terms of fraud detection.

Supervised learning indicates the presence of a supervisor as a teacher. A mathematical model of a set of data is created by an algorithm, which comprises both input and output. It's a sort of machine learning in which a machine is taught or trained using well-labelled data, meaning that some of the data has already been tagged with the correct answers. The machine is then given a new set of examples to analyse the training data and produce a proper result from labelled data using a supervised learning method.

It is used in unsupervised learning to find clusters, outliers, and anomalies in datasets. We don't trust the prediction labels in the fraud dataset to be 100% accurate, thus there will be some incorrect labels. However, we can expect fraudulent transactions to differ sufficiently from the vast majority of regular transactions. As a result, they will be flagged as anomalies or outliers by the supervised learning system. In supervised learning we can use the dimensionality detection which is also called as Principle Compound Analysis (PCA) .PCA is a strong method that may be used to find hidden patterns in data. When the data has a lot of dimensions, it's best to use PCA to reduce the number of features needed for machine learning while keeping the most relevant patterns. It can also be used for K-means clustering algorithms which is used to find the patterns in data using number of clusters. So, in general, Unsupervised learning is the process of teaching an algorithm to function without supervision on data that is

neither classed nor labelled. Without any prior data training, the machine's objective is to group information that is not categorised according to similarities, patterns, or differences. Unlike supervised learning, there is no teacher present, which implies the computer will not be trained, therefore machines are limited to find the structures which are hidden in non-labelled data by ourselves.

Reinforcement learning is concerned with taking appropriate action in order to maximise rewards in a given scenario. It determines the best possible behaviour or path in a given situation using a mixture of software and machines. There is no answer key, and the reinforcement agent decides how to accomplish the task, unlike supervised learning.

IV. DATASET

The Kaggle dataset was utilised to create this system. Kaggle is a web-based data science project that allows people to publish and find datasets, explore and create models in a web-based data science project, and solve challenges.

The dataset derived is an imbalanced dataset and is already a processed one. The dataset contains Time, principle components that are obtained through principle component analysis (PCA) which are from V1 till V28 ,Amount and Class.

The time between the current transaction and the first transaction is indicated by Time. The transaction amount is indicated by Amount, and the target variable is Class, where 1 denotes a fraudulent transaction and 0 represents a legitimate transaction. The values in each row in the dataset gives the information about the count, Mean,Std,P0,P25,P50,P75 and max as shown in figure 1.0.

	Class
count	284807.000000
mean	0.001727
std	0.041527
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

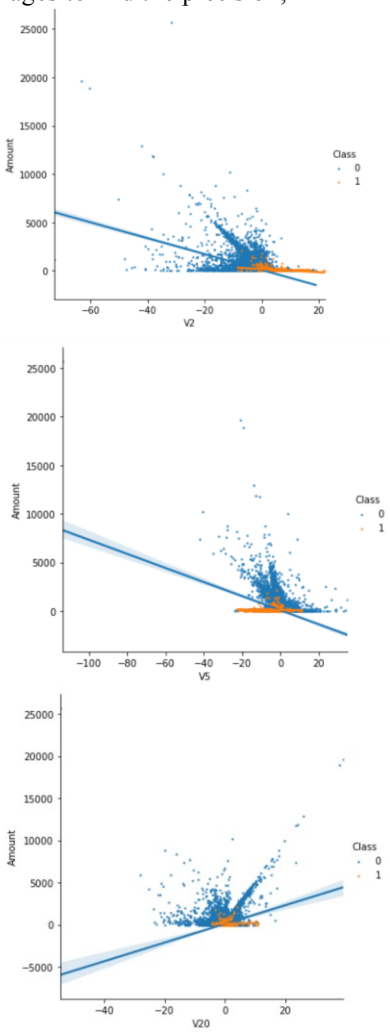
[8 rows x 31 columns]

Figure 1.0

Since the mean is 0.001727 which is close to 0, we can say that we have more legitimate transactions. The fraud in this particular dataset is 0.172% of all other transactions. In this dataset there are 284807 data and out of which 492 are frauds. This dataset was obtained from a previously processed dataset, and it was discovered that there is no association between the principal components (V1 to V28), and the Time and Amount components have not been altered and are presented as is. Some of the attributes have a weak relationship with Time (inverse correlation with V3) and Amount (Direct correlation with V7 and V20, Inverse correlation with V2 and V5) as shown in figure1.1

V. IMPLEMENTATION

The system is implemented in Anaconda jupyter notebook using python. After downloading the process dataset from Kaggle, few dependencies have been installed, then segregated the fraudulent transaction from the normal one to get the number of frauds and legitimate cases in the dataset. Imported necessary packages to find the precision,



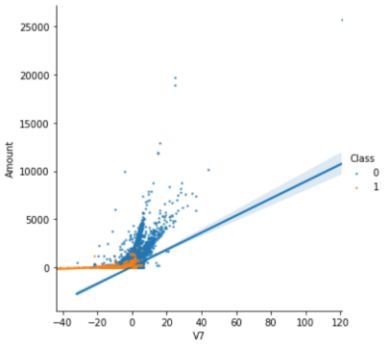


Figure 1.1.Correlation graph

accuracy, confusion matrix, f1 score, recall, decision tree, logistic regression ,support vector machine, random forest and K-Nearest Neighbour. Also plotted graphs to understand the relationship between the variables, and for better understanding of the relationship correlation is also found. Correlation is the tendency of simultaneous variation between two variables.

We can predict a continuous value using regression algorithms and several additional variables. it has 2 variables namely dependent and independent variable. we use regression for prediction. Here we used train and test approach for evaluation since the out of sample accuracy should be improved. The percentage of correct predictions a model generates on data it hasn't been trained on with the actual values of the testing set is known as out of sample accuracy.

In this system, we used 5 models and compared them to get the highest performing model which is the perfect model for prediction. The evaluation metrics used for the comparison are Recall (True positive/(True positive + False negative),Precision(True positive/(True positive + False positive) and Accuracy score ,where recall is the true positive rate, Precision is a measure of accuracy when a class label has been predicted, and Accuracy compares the actual values in the test set to the model's projected value. The true positive, false positive, and false negative values come from the confusion matrix for each model. The next step before evaluation is the classification, categorization is a supervised learning strategy for categorising or classifying unknown items into a discrete set of classes. It tries to figure out the relationship between a set of feature variables and the variables of interest that are of interest. The classification techniques used in this paper for implementation are pointed below:

1.DECISON TREE ALGORITHM

A decision tree is created by dividing the training set into different nodes, each of which contains all or most of the data. By examining the attributes one by one, a decision tree may be built. Choose an attribute from the dataset and compute the attribute's importance in data splitting.

2.Split the data according to the best attribute's value.

Then proceed to each branch and repeat the process for the remaining attributes. We can use this tree to forecast the class of unknown situations after we've built it.

Recursive partitioning is used to classify data in decision trees.

2. LOGISTIC REGRESSION

It's a statistical and machine-learning methodology for classifying records in a dataset based on input field values. We employ one or more independent variables to predict a dependent variable's result. It's similar to linear regression, except instead of predicting a numeric target field, it seeks to predict a categorical or discrete one. In logistic regression, we predict a binary value. If the dependent variable is categorical, it should be dummy-coded or indicator-coded. It can be used to classify items into binary and multiclass categories.

3.K-NEAREST NEIGHBOUR

Choosing the first nearest neighbour and assigning similar value has no assurity that the judgement is correct. As a result, to learn how to label other points, we employ K nearest values or a group of labelled points. Cases are classified using this technique depending on how similar they are to other cases. This classification technique work as follows

1.Picka value for K

2.Calculate the distance from the new case, can use Euclidean distance equation to calculate.

3. Find the K observations in the training data that are the closest to the unknown data point's measurement.

4.Using the most popular response value from the K closest neighbours, predict the unknown data points response.

4.SUPPORT VECTOR MACHINE

It is a classifier that help to understand the pattern in the dataset. After the model is obtained, it can be used to predict the cell with higher accuracy. It classifies based on a separator and is a supervised learning technique. It works by first mapping data to a high-dimensional feature space to categorise data points that aren't otherwise linearly separable, and then predicting a separator for the data.

5.RANDOM FOREST

It calculates the number of decision tree classifiers to be used on distinct sub-samples of the dataset and uses averaging to increase projected accuracy while also preventing over fitting. The implementation is easily shown with a flowchart in figure 1.2.

Since the dataset used for this implementation is already processed and further processing is not done while implementing and comparing the algorithms, the results obtained was not conclusive in nature due to very less difference in the values obtained. To solve this issue, dataset can be pre-processed by dropping the irrelevant column that is 'Time ' from the dataset ,found there is no NULL or NaN values in the dataset and in the dataset the amount column is having higher range when compared to other columns , in-order to reduce the range standard scaling is used. After pre-processing, defined X as feature set and Y as target class using train-test-split method. Before applying the classification algorithm, to balance the dataset under-sampling is done. For under-sampling, we check the no. of samples of both classes and selecting the

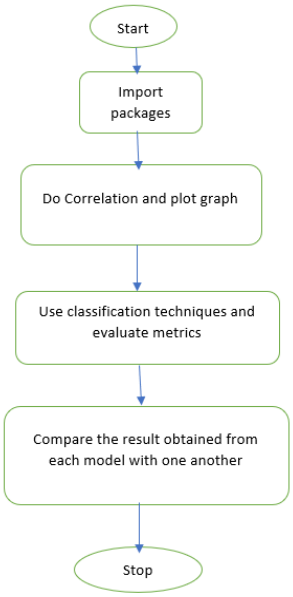


Figure 1.2. Flow chart

smaller number and taking random samples from other class sample to create new dataset. The new dataset has an equal number of samples for both target classes that is class 0 and class 1 and then the algorithms are applied and did the comparison.

VI. RESULT

After implementing the models without pre-processing and sampling, the result obtained was not conclusive in nature(fig1.3)

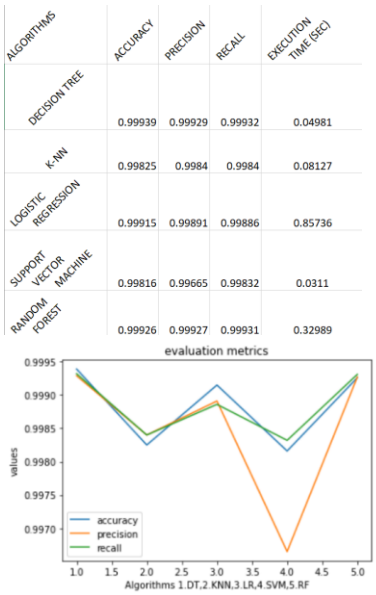


Fig 1.3.without under sampling

To improve the systemand remove the unbalanced dataset, we required pre-processing and sampling technique to obtain a better result(fig1.4)

Fig 1.4 with under sampling technique

VII.CONCLUSION

From the above result, we can conclude that logistic regression classification algorithm is giving maximum performance and decision tree algorithm is giving minimum performance.

REFERENCES

[1] Khatri, S., Arora, A., & Agrawal, A. P. (2020, January). Supervised machine learning algorithms for credit card fraud detection: a comparison. In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 680-683). IEEE.

[2] Jain, V., Agrawal, M., & Kumar, A. (2020, June). Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 86-88). IEEE.

[3] Dhankhad, S., Mohammed, E., & Far, B. (2018, July). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In 2018 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 122-125). IEEE.

[4] Adepoju, O., Wosowei, J., & Jaiman, H. (2019, October). Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques. In 2019 Global Conference for Advancement in Technology (GCAT) (pp. 1-6). IEEE.

[5] Mittal, S., & Tyagi, S. (2019, January). Performance evaluation of machine learning algorithms for credit card fraud detection. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 320-324). IEEE.

[6] Nidhika Chauhan, Prikshit Tekta, (2020), Fraud detection and verification system for online transactions: a brief overview, nternational Journal of Electronic Banking, 2020 Vol.2 No.4, pp.267 – 274.

[7] G Vishnu Manohar ,Biplab Bhattacharjee, Maheshwar Pratap(2021), Preventing misuse of discount promotions in e-commerce websites: an application of rule-based systems, International Journal of Services Operations and Informatics, 2021 Vol.11 No.1, pp.54 – 74.

[8] Dr Padmalatha N A(2020),E-Commerce Frauds and the role of fraud Detection Tools in managing the risks associated with the frauds, International journal of advanced science and Technology, 2021, Vol. 29(4s),pp.38-46.

[9] Fadi Shihadeh(2021), Online payment services and individuals' behaviour: newevidence from the MENAP, International journal of electronic banking, Vol.2 issue 4, pp. 275-282.

[10] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019, January). Real-time credit card fraud detection using machine learning. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 488-493). IEEE.

[11] Raj, S. B. E., & Portia, A. A. (2011, March). Analysis on credit card fraud detection methods. In 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET) (pp. 152-156). IEEE.

[12] Kumar, P., & Iqbal, F. (2019, April). Credit Card Fraud Identification Using Machine Learning Approaches. In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) (pp. 1-4). IEEE.

[13] Popat, R. R., & Chaudhary, J. (2018, May). A survey on credit card fraud detection using machine learning. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1120-1125). IEEE.

[14] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-9). IEEE.

[15] Choudhury, T., Dangi, G., Singh, T. P., Chauhan, A., & Aggarwal, A. (2018, August). An Efficient Way to Detect Credit Card Fraud Using Machine Learning Methodologies. In 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 591-597). IEEE.

[16] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.