TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

Fraud Detection on Smart Cards Using Machine Learning Algorithms

[1]Dr.B.Rama Devi, [2]Ms.K.Sri Harsha, [3]Ms.Y.Himaja, [4]Mr.B.Nagendra Babu
[1]Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India. ramaburri5@gmail.com
[2,3,4] B.Tech Students, Department of Information Technology
Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India.

Article Info
Volume 83
Page Number: 2495 - 2501
Publication Issue:
May - June 2020

Article History
Article Received*: 11August 2019*
Revised*: 18November 2019*
Accepted*: 23January 2020*
Publication*:10May2020*

Abstract:

One of the toughest problem in financial services is smart card fraud. Every year millions of dollars are going to be lost due to smart card fraud [4]. In recent times online transactions had become one of the most important part of our lives. Due to increase in number of transactions the fraudulent transactions [5] are also increasing rapidly. The main aim of this paper is to find out the finest and accurate model to detect the smart card fraud. Here some of the previously implemented machine learning algorithms [3] are chosen. Among those the top techniques that gives maximum accuracy levels are selected. In order to work on these algorithms the datasets that contains previous smart card transactions [4] are used. Some of the data pre-processing and data normalization techniques are applied on this raw data. To detect and reduce the fraud some of the machine learning algorithms like logistic regression, decision tree, support vector machine, k-nearest neighbour etc., are used. Among these decision tree provides more accuracy rate than the other algorithms and is stated as best for smart card [3] fraud detection.

KEY WORDS:*Smart card, Fraud detection, Machine learning, Classification algorithms, Logistic regression, Support vector machine, Decision tree, K-nearest neighbour.*

I. INTRODUCTION

The smart card is one of the most prominent mode of payment for online as well as offline purchases that are regular purchases. A smart card is a tool which provides the cashless shopping across the world. Since the usage of smart card is increasing day by day the number of fraudtransactions are also increasing rapidly. In order to overcome these frauds we have to use several statistical based models and expert rules. The main objective of this paper is to go through some of the fraud detectingmethods.

The word fraud [6] indicates unwanted or unauthorized access of some one's account by others. When a person uses other's smart card illegally for his personal reasons then it is stated as smart card fraud. In this case the owner of the card is unawareof these fraudulent activities.

Hence we have to find out the number of fraudulent transactions among the total number of transactions. Due to high acceleration of E- Commerce the smart cards usage has been increased rapidly. Hence this vast usage of smart cards is resulting in high amount of fraudulent transactions. As we are living in a digital world, hence there is a need to identify and eradicate these transactions. Hence identifying these fraudulent transactions [1] is mentioned technically as fraud detection.

The financial losses due to these smart card frauds [6] effect all the merchants, bans and individual clients too. These types of frauds also affect the reputation of organizations. Suppose a card holder is a victim of a certain organization then he no longer trusts them and not only that he chooses to continue with their competitive organizations.

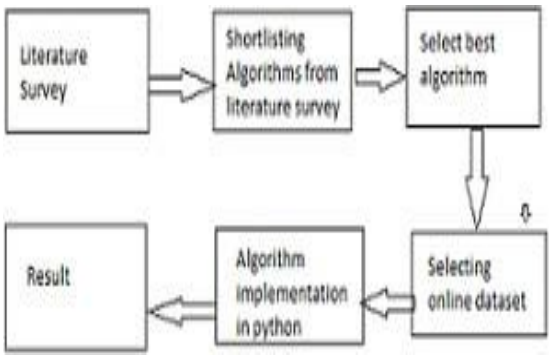
May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

Through fraud detection [1] we have to analyse and identify the behaviour of different smart card users. There by we can stop their illegal behaviour. To do this we have to go through different techniques and mechanisms. These techniques help us to detect the smart card frauds. Hence to do so we have to study different machine learning algorithms [1] which help us in fraud detection these algorithms are used to find out the difference between fraudulent transactions and normal transactions. Machine learning and data science [4] are two fields where these type of problems can be automated. Hence we use the algorithms in these two fields to analyse the transactions and differentiate the authorized ones and suspicious ones.

Loss occurred due to fraud can be prevented with the help of two mechanisms they are fraud prevention and fraud detection [5]. Prevention is a method which stops the fraud happening in the initial stage, whereas detection is a method which is applied when a fraudulent transaction is attempted. Smart card fraud is concerned due to illegal usage of card for purchases. Transactions can be done both digitally and physically. In physical transactions the smart card should be involved during the transactions. On the other hand digital transactions can be done through telephone or internet. During digital transactions the card holder need to mention card number, expiry date that indicates when the card will be expired along with the verification pin generated. These details are given through a website or phone.

Due to rise in e-commerce the smart card usage is also increased parallel. Hence the smart card transactions are also increasing from trillions to billions from one decade to another. The fraudsters choose internet as a tool by hiding their location information. Financial industries are the primary effectors due to this smart card fraud. To overcome the loss occurred because of these smart card frauds the merchants areincreasing the cost of the products similarly they are reducing the incentives and discounts just to overcome the loss occurred due to these frauds.

Various machine learning algorithms [6] are used for fraud detection. Algorithms like neural networks, support vector machine, decision tree, logistic regression and some of the other algorithms are used. Along with these a few deep learning models are also used. In addition to these Ada-boast and J48 are two other techniques which can also be used for fraud detection [3]. These algorithms are applied on the smart card data sets. These data sets contain the information related to the smart card transactions of different users.



The project can be done based on the above architecture.

Initially conduct a literature survey by going through different references where this project is already implemented. From that literature survey shortlist some fine and good algorithms and techniques that tends to suit this project. Among those shortlisted algorithms select the best algorithm for fraud detection. Parallely gather some online datasets which contain information related to previous smart card transactions. Now train and test the algorithms with the help of these online datasets. The implementation is done in python programming language with the help of some other tools. Finally the result is obtained in terms of accuracy. The algorithm which gives maximum accuracy values is stated as the best on for fraud detection.

II. Literature Survey

Gosh and Reilly [1] had proposed a method based on neural networks to detect the fraud detections on smart card. Their system Consist of a three-layer feed forward network. Bayesian network is another technique to detectfrauds.

May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

Bentley [3] suggested an algorithm that depends on genetic programming. Through this algorithm we can apply some logical rules which are capable of differentiating suspicious and non-suspicious transactions. Here we use a scoring process which compresses last 3 months transactions if due is greater than 3 months it is considered as suspicious.

Kokkinaki [1] proposed a system based on decision trees for fraud detection. Generally this tree consists of nodes and edges which are labelled with attribute names and attribute values. Here they used an intensity factor that could satisfy some conditions. This is stated as the ratio of transactions that satisfy some conditions to the total number of legitimate transactions.

An unsupervised smart card detection method is proposed by button and handwhich includes frequency of transactions and abnormal spending behavior. Here the mean amount technique is used in fraud detection. This is stated as one of the most difficult and time consuming fraud detection process. Since there exists millions of transactions everyday hence this is highly skewed data. This data is highly contiguous than these fraudtransactions.

Che-hui lien [4] had proposed a system using data mining technique for fraud detection. I n this paper they compared the performance of six classification techniques. Hence they find out the accuracy of six data mining techniques. Based on their survey artificial neural networks algorithms had performedbest.

A technique based on outlier mining is proposed by Wen-Fang Yu [5] and Na Wang. This is based on internet fields where it detects the transactions that are fraudulent among overall records.

Neda Sultana Halvaiees [4] introduced a technique related to immune systems. This system will differentiate between self and non-self transactions. Hence to distinguish among these there should be B-cells and T- cells.

Abhinav Srivastava [2] implemented a model named Hidden Markov Model for the identification of fraud. In thus method it identifies the spending behavior of the cardholder and there by detects the fraud.

A comparison between six classification techniques is made by I-Chang Yeh [4]Che- hui Lien. This comparison will identifies which classification algorithm has least error rate and then states it as the best classificationmodel.

Dr.M.Kishore Kumar [1], Siva Parvathi Nelluri, Shaik Nagul implemented a detection system with the help of card holder's profile. Based on users profile this model identifies whether the transaction is legitimate or not.

Bram Vanschooen, Sam Maer, Karl Tuyls, Winkel et al proposed a fraud detection system using the fields of Artificial Neural Networks and Bayesian Belief Network. By comparing the results of these two methods they deliver the finest one.

Raghavendra Patidar [3] and Lokesh Sharma introduced a topology based neural network. This model consists of neurons that are hidden in the topology with the help of geneticalgorithms.

III. Methodology

Machine learning algorithms are broadly classified into supervised and unsupervised. Supervised algorithms in turn are divided into regression algorithms and classification algorithms.

Classification algorithms are used for predicting categorical values. When coming to classification algorithms the identification of new observations based on training data isdone.

Here in this project classification algorithms are used for predicting smartcard frauds. In classification algorithms a program gets trained based on the given data set with the help of some machine learning techniques and then gives new observations in terms of classes orgroups.

The classification algorithms are of two forms:

1. Binaryclassifiers
2. Multi-classclassifiers

In binary classifiers the problem has only two possible outcomes that is o or 1, yes or no, male or female, spam or not spam etc.

May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

A problem belongs to multi-class classifier when it has more than two outcomes.
Ex: Classifying types of music, Classifying types of food etc.
In turn classification algorithms are mainly of two different models:

1. Linear Models
   - Logistic Regression
   - Support Vector Machine
2. Non-Linear Models
   - K-Nearest Neighbors
   - Naïve Bayes
   - Decision Tree Classification
   - Random Forest Classification

In classification algorithms there will be a confusion matrix which consists of evaluation data of a particular model. The confusion matrix can also be called as error matrix. The matrix is represented in the form of a table. The table generally stores the results summary in short. When coming to prediction the matrix table store the number of correct and incorrectpredictions.

	Predicted Positve	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

$$\text{Accuracy} = \frac{TP + TN}{\text{TotalPopulation}}$$

IV. Implementation of Algorithms:

Here in this project a set of four main classification algorithms are used for fraud detection. They are stated as follows:

Logistic Regression:
The logistic regression algorithm is similar to linear regression in some aspects in classification algorithms. The difference between these two algorithms is that linear regression is used to forecast values whereas logistic regression is used for classification tasks.

The logistic regression consists of three patterns binomial, multinomial and ordinal. Binomial – The result is in a binary formati.e. it consists of only two possible outcomes either 0 or 1. 0 for true and 1 for false.

Multinomial – In multinomial there exists more than two possible outcomes that are in an unordered format.

Ordinal – In ordinal the result is ordered based on a specific category like good, poor, bad, very good, nice.

In this algorithm mainly binomial form is used to identify and predict the values.

V. Support Vector Machine:

Support vector machine is used for both classification and regression problems in machine learning. However it is mainly used for classification problems.

The main aim of svm is to provide best boundary which separates an n-dimensional area into different classes. So that new values can be placed in related classes in future.

Here the best boundary is called hyper plane and there are some cases on which the classes are separated and those are called support vectors.

In this algorithm we train the model with some basic prototype features and then give some strange features for identification to test the model thereby training and testing can be done parallely.

VI. K-Nearest Neighbor:

This is one of the simplest machine learning algorithms in classification models. In knn the algorithm checks the similarity between the new data value among the already existing classes. If any of the class appears to be most similar to the new data value then if places the new value in that class.

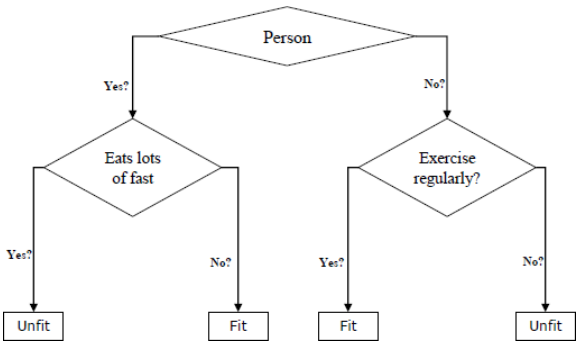
The other names for knn are non-parametric algorithm or lazy learner algorithm.

This algorithm is pretty easy to learn and apply on the data sets among all the other machine learning algorithms.

May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

VII. Decision Trees:

Decision Trees are one of the most powerful algorithms among the classification algorithms in machine learning. The most important parts of a tree are decision nodes that splits the data into various sub parts similarly leaves where we get the result of that particular node.



To construct a decision tree initially a terminal node has to be created which is the parent node of the tree. The terminal nodemay be either minimum or maximum. Through the terminal node the leaf nodes are created which gives the result. Any of the key column value can be taken as a terminal node.

VIII. Process of Machine Learning:

Machine Learning is one of the technologies that gives the system some abilities to learn and work itself instead of explicitly programmed.
There was a process about how a project or problem can be solved in machine learning. It includes:

- CollectingData
- DataPreparation
- DataPre-processing
- Analyzing thedata
- TrainModel
- TestModel
- Deploy

Before starting this process there is a need to understand the problem domain first, because a good understanding of the problem helps us to attain better results.

IX. Collecting Data:

The initial step of this process is collecting the data. So choose some of the data sources and collect data from them. The data sources can be mobile devices, internet, databases etc. If more amount of data is collected the accuracy for predicting the problem is also more.

DataPreparation:

In data preparation the collected data is to be prepared into a format that suits the machine learning trainingprocess.

It includes two steps:

Exploring data:

Here it describes the format, the quality and the nature of the data that is collected.

Data Pre-processing:

Pre=processing includes analyzing the data and identifying the mistakes in it.

Data Wrangling:

Wrangling of data is nothing but cleaning of raw data that is collected during collection process and convertingit into useable format. This is one of the main step in this whole process.
Generally the data collected may have various problems like:

- Missingvalues
- Invaliddata
- Duplicatedata
- Noise

So filtering techniques are used to avoid these problems.

Analyzing the data:

In this step the aim is to introduce a machine learning model, after choosing a model build that particular model and then review the result. The model may be nay of the classification, regression and association types.

May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

Train Model:

In this phase the data set has to be trained with the help of a model. Training the data with models is just to get the better outcome. So, for training a variety of patterns, datasets and features areused.

Test Model:

Once a machine learning model is trained on the provided dataset the next step is to test that model. In testing the accuracy values for predicting the model are notified.

Deploy:

In deployment phase we deploy our project into the real-world system. If an accurate prediction value is obtained on a particular model then that model is stated as the best one for predicting a problem.

X. Results:

The datasets that are collected from different sources are initially trained with different selected models and secondly it is tested on the same models. Hence we train 70% of the data and remaining 30% is tested with the selected models among the entire data. As the smart card contains fields like transaction time, place, amount and also some confidential details that are not given in the datasets and are mentioned with some variables or null data. The results are given in the form of accuracy values and model with maximum accuracy value is stated as best for fraud identification.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V21	V22	V23	
0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	...	-0.018307	0.277838	-0.110474	0
1	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	...	-0.225775	-0.638672	0.101288	-0
2	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	...	0.247998	0.771679	0.909412	-0
3	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	...	-0.108300	0.005274	-0.190321	-1
4	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	...	-0.009431	0.798278	-0.137458	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
284802	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	1.914428	4.356170	...	0.213454	0.111864	1.014480	-0
284803	-0.732789	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	0.584800	-0.975926	...	0.214205	0.924384	0.012463	-1
284804	1.919565	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	0.432454	-0.484782	...	0.232045	0.578229	-0.037501	0
284805	-0.240440	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	0.392087	-0.399126	...	0.265245	0.800049	-0.163298	0
284806	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	0.486180	-0.915427	...	0.261057	0.643078	0.376777	0

284807 rows x 30 columns

Fig.1 Processed Data

In the below figure 1 indicates suspicious transactions among the overall transactions in the dataset and 0 indicates the authorized transactions.

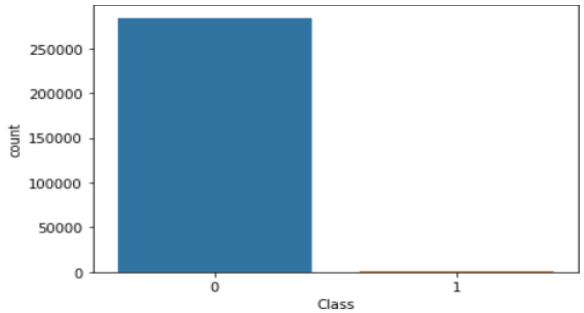


Fig.2 Transactions

Here two fields which are not confidential with respect to user are selected to apply any of the selected models. In this project two fields' amount and time are selected and the results are as below.

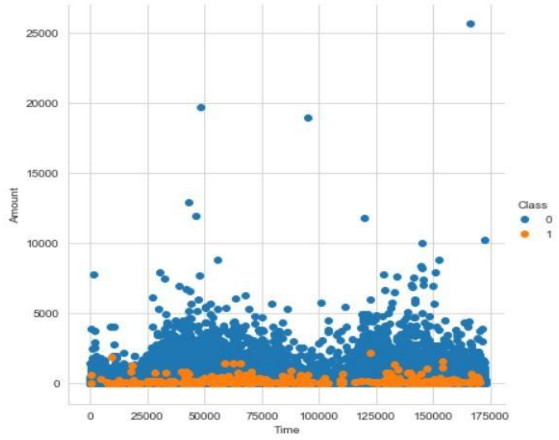


Fig.3.0 Result-1

May – June 2020
ISSN: 0193-4120 Page No. 2495 - 2501

In the above graph the blue indicates non- fraudulent transactions and orange indicates fraudulent transactions.

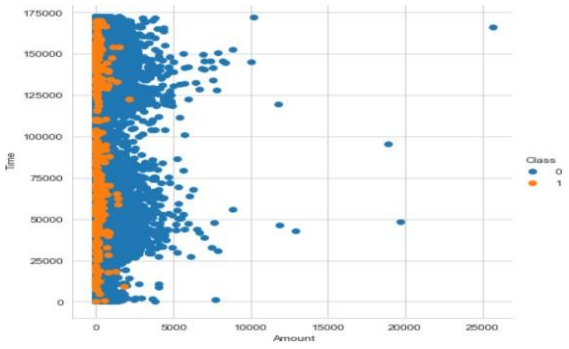


Fig.3.1 Result-2

If the amount and time are taken in an inverse manner then the results are as above.

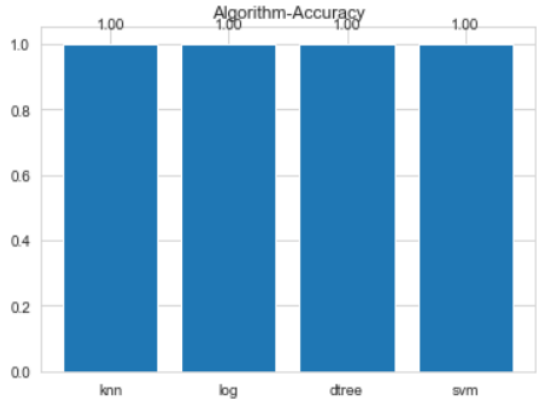


Fig.4 Accuracy

Among the four selected algorithms like k nearest neighbor, logistic regression, decision tree and support vector machine the maximum accuracy is for decisiontree.

Name	Logistic regression	K-nearest neighbor	Support vector machine	Decision tree
Accuracy	98.0%	98.6%	98.1%	100.0%
Method	Classification method	Machine learning method	Supervised learning	Classification method
Training:Testing	70:30	70:30	70:30	70:30
Inbuilt packages	Logistic regression	Kneighbors classifier	Svm classifier	Decision tree classifier

Implementation of algorithms

XI. Conclusion

In this paper a brief description about the process of machine learning and a discussion about some classification techniques are done. Afterthe implementation decision tree reached the maximum accuracy level of 100% and is best suited for the smart card fraud detection.

References:

1. Credit card Fraud Detection based on Machine Learning Algorithms International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 44, March2019
2. Credit card fraud detection using hidden markov model (hmm) akinmuleya benjamin olaseni matric no:135022022
3. Credit Card Fraud Detection using Machine Learning and Data Science International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 8 Issue 09, September-2019.
4. Machine Learning Approaches For Credit Card Fraud Detection International Journal of Engineering & Technology. website:www.sciencepubco.com/php.ijet
5. Credit Card Fraud Detection&Prevention A Survey IJIRST –International Journal for Innovative Research in Science & Technology| Volume 4 | Issue 1 | June2017 ISSN (online): 2349-6010.
6. Rama Devi Burri, Ram Burri, Ramesh Reddy Bojja, Srinivasarao Buraga "Insurance claim Analysis using Machine learning Algorithms, "International Journal of innovative technologyandExploring Engineering (IJITEE), Volume-8, Issue-6S4, April-2019, ISSN: 22278-3075.
7. Rama Devi Burri, Y.Venkata Raghava Rao, V.B.V.N. Prasad "Machine learning methods for software Defect Prediction a Revit." International journalofinnovativetechnology andExploring Engineering (IJITEE), Volume-8, Issue-8, June-2019, ISSN:2278-3075.