|| Volume 5 || Issue 9 || September 2020 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Rahul Powar, Rohan Dawkhar, Pratichi

BE Students, Computer Engineering, Dr D Y Patil School Of Engineering Academy, Pune, India
rahulpower2012@gmail.com, rohan.dawkhar@gmail.com, pratichi3@gmail.com

---------- ***----------

Abstract: - Due to rapid growth in the field of cashless transactions or digital transactions, credit cards are widely used in almost every work and hence there are more chances of fraudulent transactions. These fraudulent transactions can be identified by analyzing various behaviors of credit card customers from previous transaction history datasets. If any deviation is noticed in the behavior from the available patterns, there is the possibility of fraudulent transaction. Machine learning techniques are widely used to detect the frauds. In this paper, we have used KNN technique to detect the frauds .The performance of this techniques is evaluated based on the accuracy, sensitivity, precision and recall.

Keywords: - *Machine learning, Credit Card, KNN technique, T-SNE technique, removing outliers, fraudulent*

---------- ***----------

I INTRODUCTION

These days, payments through the credit cards are common as it is easy and less time taking. But here a big problem comes with it and that is fraudulent transactions and many surveys have shown how credit card frauds have been increased in past few years.

Credit card fraud detection is a very popular but also a difficult problem to solve. Firstly, due to issue of having only a limited amount of data, credit card makes it challenging to match a pattern for dataset. Secondly, there can be many entries in dataset with truncations of fraudsters which also will fit a pattern of legitimate behaviour. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and the results of researches are often hidden and censored, making the results inaccessible and due to this it is challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature is nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes limitation to exchange of ideas and methods in fraud detection, and especially in credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviours always different that is the legit transaction in the past may be a fraud in present or vice versa. With the advancement of machine learning techniques, machine learning has been identified as a successful measure for fraud detection. A large amount of data is transferred during online transaction processes, resulting in a binary result: genuine or fraudulent. Within the sample fraudulent datasets, features are constructed. These are data points namely the age and value of the customer account, as well as the origin of the credit card. There are hundreds of features and each contributes to varying extents towards the fraud probability. We have used KNN algorithm of the machine learning on the dataset to detect the fraudulent transactions. Behavioural pattern of spending money depends on past history of transactions and attributes such as location, daily expenses, transaction time of cardholder can be compared with current transaction details to detection credit card frauds. Deviation from such behaviour helps to detect fraud with more accuracy. With the deviation in behavioural data, we used different data mining techniques to detect the fraud. The model can then be used to identify whether the new transaction is fraud or not.

Problem Definition

Major problem is that online payment does not require physical card. Anyone who knows the details of the card can make fraud transactions. Card holder comes to know only after the fraud transaction is carried out.

Problem Solution

Unfortunately, it contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues original features are not provided and more background information about the data is also not present. Features V1, V2,...,V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Objectives:

The objective of proposed system is to build Credit Card Fraud Detection System Using Machine Learning .This proposed system uses KNN Technique to detect the frauds.

II LITERATURE REVIEW

1. The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada.

Author:- "Kosemani Temitayo Hafiz, Dr.Shaun Aghili, Dr. Pavol Zavarsky."

This research paper focuses on the creation of a scorecard from relevant evaluation criteria, features, and capabilities of predictive analytics vendor solutions currently being used to detect credit card fraud. The scorecard provides a side by side comparison of five credit card predictive analytics vendor solutions adopted in Canada. From the ensuing research findings, a list of credit card fraud PAT vendor solution challenges, risks, and limitations was outlined. All the sub topics should be numbered as shown above. Numbering should be made correctly.

2. BLAST-SSAHA Hybridization for Credit Card Fraud Detection.

Author:- "Amlan Kundu, Suvasini Panigrahi, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar"

This paper propose to use two-stage sequence alignment in which a profile Analyser (PA) first determines the similarity of an incoming sequence of transactions on a given credit card with the genuine cardholder's past spending sequences. The unusual transactions traced by the profile analyser are next passed on to a deviation analyser (DA) for possible alignment with past fraudulent behaviour. The final decision about the nature of a transaction is taken on the basis of the observations by these two analysers. In order to achieve online response time for both PA and DA, we suggest a new approach for combining two sequence alignment algorithms BLAST and SSAHA.

3. Research on Credit Card Fraud Detection Model Based on Distance Sum.

Author:- "Wen-Fang YU, Na Wang".

Along with increasing credit cards and growing trade volume in China, credit card fraud rises sharply. How to enhance the detection and prevention of credit card fraud becomes the focus of risk control of banks. It proposes a credit card fraud detection model using outlier detection based on distance sum according to the infrequency and unconventionality of fraud in credit card transaction data, applying outlier mining into credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud. All the sub topics should be numbered as shown above. Numbering should be made correctly.

4. Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models.

Author:- "Navanshu Khare and Saad Yunus Sait"

This paper investigates and checks the performance of Decision tree, Random Forest, SVM and logistic regression on highly skewed credit card fraud data. Dataset of credit card transactions is sourced from European cardholders containing 284,786 transactions. These techniques are applied on the raw and pre-processed data.

5. Credit Card Fraud Detection Using Bayesian and Neural Networks

Author:-"Sam Maes , karl tuyls , Bram vanschoenwinkel and Bernard Manderick"

This paper discuss automated credit card fraud detection by means of machine learning. We apply two techniques suited for reasoning under uncertainity : artificial neural networks and Bayesian belief networks to the problem and show their significant results on real world financial data.

III PROPOSED SYSTEM

Machine learning and its algorithms:

Machine learning is a collection of methods that can automatically identify patterns in data, and then use those patterns to predict future outcomes, or to perform other types of decision making below certain conditions. Machine learning introduces various algorithms, those enable machines to understand the current situations and on the basis of that machines can take appropriate decisions. Machine learning works independently and takes decision at its own. The main two types of machine learning are, supervised learning and unsupervised learning.

Supervised Learning: In supervised learning, the input and its corresponding output is already known. This is called supervised learning because it learns from training data set and creates model from it and when this model applies on new data set it gives predicted results. Decision Tree, naive Bayes etc. are the examples of supervised learning.

Unsupervised Learning: Unsupervised learning is where we have only input data and no corresponding output variable. The main job of unsupervised learning is to build up class labels automatically.

The relationship between the data can be found using unsupervised learning algorithms to discover whether the data can characterize to form a group. This group is known as clusters. Unsupervised learning can be also described as cluster analyses. K Means Clustering, KNN etc. are the examples of unsupervised learning.

Selected online dataset:

In this project, we have used a Kaggle provided dataset of simulated mobile based payment transactions. We analyze this data by categorizing it with respect to different types of transactions it contains. We also perform PCA - Principal Component Analysis - to visualize the variability of data in two dimensional spaces. The datasets contain transactions made by credit cards in September 2013 by European cardholders. These dataset present transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation.

Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise."

Selected algorithm for implementing:

On the Literature review, many algorithms are applied on Fraud detection. Here we have used K – nearest neighbour which is having better accuracy than other algorithms for fraud detection.

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 the credit card fraud detection is k- nearest neighbour algorithm that is a supervised learning algorithm where the result of new instance query is classified based on majority of K-Nearest Neighbour category.

The performance of KNN algorithm is influenced by three main factors:

- The distance metric used to locate the nearest neighbours.
- The distance rule used to derive a classification from k-nearest neighbour.
- The number of neighbours used to classify the new sample.

Among the various credit card fraud detection methods of supervised statistical pattern recognition, the K Nearest Neighbour rule achieves consistently high performance, without a priori assumptions about the distributions from which the training examples are drawn. K- Nearest neighbour based credit card fraud detection techniques require a distance or similar the measure defined between two data instances.

In process of KNN, we classify any incoming transaction by calculating of nearest point to new incoming transaction. Then if the nearest neighbour be fraudulent, then the transaction indicates as a fraud. The value of K is used as, a small and odd to break the ties (typically 1, 3 or 5). Larger K values can help to reduce the effect of noisy data set. In this algorithm, distance between two data instances can be calculated in different ways.

For continuous attributes, Euclidean distance is a good choice. For categorical attributes, a simple matching coefficient is often used. For multivariate data, distance is usually calculated for each attribute and then combined. The performance of KNN algorithm can be improved by optimizing the distance metric. This technique required legitimate as well as fraudulent samples of data for training. It is fast technique along with high false alert.

Flow of Project:

We have done Exploratory Data Analysis on full data then we have removed outliers using "Local Outlier Factor", then finally we have used KNN technique to predict to train the data and to predict whether the transaction is Fraud or not. We have also applied T-SNE to visualize the Fraud and genuine transactions in 2-D.

IV SYSTEM ARCHITECTURE

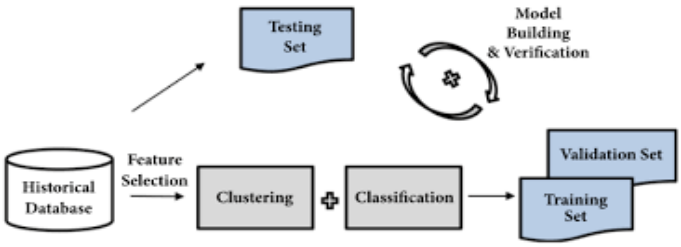

Figure 1 : Credit card fraud detection using machine learning

First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and Class. Class 0 represents a valid transaction and 1 represents a fraudulent one. It contains only numerical (continuous) input variables which are as a result of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. Behavioural characteristic of the card is shown by a variable of each profile usage representing the spending habits of the customers along with days of the month, hours of the day, geographical locations, or type of the merchant where the transaction takes place. Afterwards these variables are used to create a model which distinguish fraudulent activities. The details and background information of the features cannot be presented due to confidentiality issues. The time feature stores the seconds that has elapsed between each transaction along with first transaction in the dataset. The 'amount' feature is the transaction amount. Feature 'class' is the target class for the binary classification.

Four basic metrics are used in evaluating the experiments, namely True positive (TPR), True Negative (TNR), False Positive (FPR) and False Negative (FNR) rates metric respectively.

$$TPR = \frac{TP}{P}$$

$$TNR = \frac{TN}{N}$$

$$FPR = \frac{FP}{N}$$

$$FNR = \frac{FN}{P}$$

where FN , FP ,TP,TN, and are the number of false negative false positive ,true positive and true negative test cases classified while total number of positive and negative class cases under test are represented by P and N. Cases classified rightly as negate are termed with true negative and cases classified as positive which are actually positive are termed with True positive .Cases classified as positive but are negative cases are termed as false positive and cases classified as negative but are truly positive are termed as false negative. The performance of Classifiers is evaluated based on accuracy, precision, specificity and sensitivity.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity (Recall) gives the accuracy on positive (fraud) cases classification. Specificity gives the accuracy on

negative (legitimate) cases classification. Precision gives the accuracy in cases classified as fraud (positive).

V RESULTS:

1.

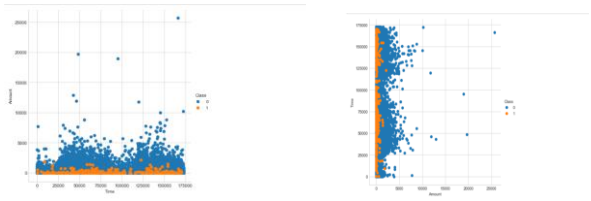

Figure: 2-D Scatter Plot

Observation: From above two plots it is clear that there are frauds only on the transactions which have amount less than 2500.

2.

```
In [21]: from sklearn.metrics import recall_score
         KNN_best = KNeighborsClassifier(n_neighbors = best_k, algorithm = 'kd_tree')
         KNN_best.fit(X1, Y1)
         prediction = KNN_best.predict(XTest)
         recallTest = recall_score(YTest, prediction)
         print("Recall Score of the knn classifier for best k values of "+str(best_k)+" is: "+str(recallTest))
         cm = confusion_matrix(YTest, prediction)
         print(cm)
         tn, fp, fn, tp = cm.ravel()
         (tn, fp, fn, tp)
         Recall Score of the knn classifier for best k values of 1 is: 0.8333333333333334
         [[3978   10]
          [   2   10]]
Out[21]: (3978, 10, 2, 10)

In [22]: YTest.value_counts()
Out[22]: 0    3988
         1      12
         Name: Class, dtype: int64
```

Figure: Recall score calculated

3. Confusion Matrix: From the above snapshot,

[[3978 10]

[2 10]]

We can see that value of “True Negative” is 3978 which means that out of 3988 points which belongs to class ‘0’,3978 points are predicted as ‘0’.

Furthermore, from the same confusion matrix we can see that the value of “True Positive” is 10 which means that out of 12 points which belongs to class ‘1’,10 points are detected as ‘1’.

VI CONCLUSIONS

There are total 4000 points in our dataset, out of which 3988 points belongs to class label ‘0’ and 12 belongs to class label ‘1’. This means that our model has performed well despite having very imbalanced dataset.

REFERENCES

[1] STUDY OF DETECTION OF VARIOUS TYPES OF CANCERS BY USING DEEP LEARNING : A SURVEY VINOD B BHARAT DR. NAVNEET MALIK 2019/8/31 International Journal of Advanced Trends in Computer Science and Engineering

[2] Erkin, Erkin et. al., "Privacy-preserving distributed clustering" in EURASIP Journal on Information Security, licensee Springer, 2013.

[3] Ge-Er Teng, Chang-Zheng He, Jin Xiao, Xiao-Yi Jiang, "Customer credit scoring based on HMM/GMDH hybrid model" in, London: SpringerVerlag, 2012.

[4] Ashphak Khan, Tejpal Singh, Amit Sinhal, "Implement Credit Card Fraudulent Detection System Using Observation Probabilistic in Hidden Markov Model", NUiCONE-2012, December. 2012.

[5] Divya Lyer, Arti Mohanpurkar, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE, 2011.

[6] V. Bhusari, S. Patil, "Study of Hidden Markov Model in Credit Card Fraudulent Detection", International Journal of Computer Applications, vol. 20, no. 5, pp. 0975-8887, April. 2011.

[7] Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", IEEE Transactions On Dependable and Secure Computing, vol. 5, no. 1, January-March. 2018

[8] B.Pushpalatha and C.W. Joseph,"Credit Card Treachery Detection Based on the Transaction by Using Data mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, No. 2, pp. 1785-1793, 2017.

[9] Deepika .N and Roopa .H"Analyzing the CC (credit card) Treachery Detection using Data Mining Techniques", IJESE, Vol. 7, No. 6, pp. 12851-12854, 2017.

[10] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare,"Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", 2nd International conference on New IT trends 2017, IEEE, PP 978-988.

[11] B.Pushpalatha, C.Willson Joseph," Credit Card Fraud Detection Based on the Transaction by Using Data mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017, PP 1785-1794.

[12]. You Dai, Jin Yan, Xiaoxin Tang, Han Zhao and Minyi Guo," Online Credit Card Fraud Detection: A

Hybrid Framework with Big Data Technologies", 2016 IEEE TrustCom/BigDataSE/ISPA, PP 1644-1653.

[13]. T Nuno Carneiroa, Gonc¸alo Figueiraa,∗, Miguel Costab," A data mining based system for credit-card fraud detection in e-tail", Journal of DSS, Sep 2016, PP 1-11

Guide : Prof Amol Jadav

Name: Rahul Powar

Email : rahulpower2012@gmail.com

Name: Rohan Dawkhar

Email **:** rohan.dawkhar@gmail.com

Name: Pratichi

Email **:** pratichi3@gmail.com

****