JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Credit Card Fraud Detection Using KNN And Naive Bayes Algorithm

Sudiksha Wandre
Department of Computer Science Engineering
Universal college of Engineering,
Vasai,India
sudiksha.wandre01@gmail.com

Shefali Desai
Department of Computer Science Engineering
Universal college of Engineering,
Vasai,India
desaishefali507@gmail.com

Akruti Patel
Department of Computer Science Engineering
Universal college of Engineering,
Vasai,India
akrutipatel06@gmail.com

Asst. Prof. Hezal Lopes
Department of Computer Science Engineering
Universal college of Engineering,
Vasai,India
hezal.lopes@universal.edu.in

***Abstract*— There are totally different patterns in fraud. They constantly alter their behavior, necessitating the employment of unsupervised learning. Fraudsters gain access to trendy technology that enables them to commit fraud through internet transactions. Fraudsters build the idea that consumer behavior and fraud trends evolve over time quickly. As a result, fraud detection systems should be capable of detective work on-line transactions as a result of some fraudsters use on-line mediums to commit fraud once and then switch to different techniques.The aim of the projected system is that unsupervised learning could be a technique of learning that doesn't need direction.Credit cards that acknowledge the importance of accelerating individuals' shopping for power and allowing them to fulfill their daily wants, like vesture and technology with the rising use of credit cards, the frequency of CC (credit card) scams has skyrocketed. The largest credit card frauds area unit outline because the unethical use of credit cards by hackers or credit card users United Nations agency refuses to pay back the number owed. Credit card scams are also discovered by analyzing credit card purchase trends and exploiting historical information. This information analysis will assist banks and different businesses.As the bank doesn't provide information of customers .The aim is to take data sets from local sources and look for anomalies in patterns of fraud activity that have changed over time supporting this data back in time. A competent fraud notification system ought to be ready to detect the following forms of fraud. The fraud group action ought to be exactly recorded, and also the detection ought to be straightforward.**

Keywords—Credit card fraud detection, classification, KNN, Naive Bayes Machine learning algorithm.

I. INTRODUCTION

Today use of Credit Cards even in developing countries has become a typical state of affairs. individuals use it to buy, pay bills and for on-line transactions but with the upward push inside the variety of customers' cards, the instances of fraud in Credit Cards have conjointly been on upward push.Card connected frauds cause globally a loss of billions of greenbacks. Fraud is often any conduct with the purpose to deceive is categorized as deceit. to obtain gain by any manner while not the information of the cardholder and also the establishment bank. card fraud will be tiring in various ways.By dropping or taking cards, via way of means of generating fake or counterfeit cards, via way of means of organic studies of the primary site, via way of means of erasing or enhancing the magnetic strip gift at the card that contains the user's data, by phishing, by skimming or by stealing information from a merchant's face.

II. REVIEW OF LITERATURE

In the paper presented by G.Srinivas [1] proposed a way of victimizing machine learning to observe credit card fraud. Initially, customary models were used at the moment hybrid models came into the image that created use of AdaBoost and majority choice strategies. In public out there knowledge set had been wont to appraise the model potency and another knowledge set used from the establishment and analyzed the fraud. Then the noise was extra to the information sample through that the strength of the algorithms might be measured.

The experiments were conducted on the premise of the theoretical results that show that the bulk of pick strategies bring home the bacons} good accuracy rates so as to observe the fraud within the credit cards. For more analysis of the hybrid models, noise of approximately 100% and half-hour has been added to the sample data[2]. Many pick strategies have achieved an honest score of zero.942 for half-hour additional noise[2]. Thus, it had been over that the pick methodology showed abundant stable performance within the presence of noise.

In the paper presented by Satish B Basapur [3] proposed deep learning topologies for the detection of fraud in on-line cash transactions. This approach springs from the unreal neural network with in-built time and memory elements like long run short term memory and a number of {other|and several other} other parameters.

According to the potency of those elements in fraud detection, virtually eighty million on-line transactions through credit cards are pre-labeled as dishonorable and legal.They have used a excessive overall performance allotted cloud computing environment.The researchers' research provides an accurate reference to the sensitivity analysis of the projected parameters of fraud detection performance.The researchers additionally projected a framework for the parameter standardization fraud detection with deep learning topologies.This permits the financial organization to decrease the losses by avoiding dishonorable activities.

In the paper presented by R.Vateli [4] proposed a Deep autoencoder that is employed to extract the most effective characteristics of the data from the credit card group action. His can upload greater softmax applications to remedy the class labels problemsAN overcomplete autoencoder is employed to map the info into a high dimensional area and a thin model was utilized in a descriptive manner that provides advantages for the classification of a sort of fraud.

Deep learning is one of the foremost motivated and powerful techniques being utilized for the detection of fraud within the credit card. These kinds of networks have a posh distribution of knowledge that is incredibly troublesome to acknowledge. Deep autoencoders are utilized in some stages to extract the most effective options of the info and for classification functions. Also, higher accuracy and low variance are achieved among these networks.

In the paper presented by John O. Awoyemi [5] proposed associate degree investigation through which the performances of many algorithms were evaluated once they were applied on credit card fraud knowledge that's extremely inclined. The cardholders' 284,807 transactions were used as a supply to come up with the dataset of credit card transactions. On the inclined knowledge, a hybrid approach of under-sampling and oversampling is performed. On raw and preprocessed knowledge, there are 3 totally different techniques applied in Python.

Based on sure parameters like preciseness, sensitivity,accuracy, balanced classification rate and so on, the performances of those techniques are unit evaluated. It's seen through the achieved results that as compared to naïve mathematicians, the performance of k-NN is best.

In the paper presented by X. Zhao [6] study on the usually found crime inside the Credit card applications. There area unit sure problems featured once the prevailing non-data mining approaches area unit applied to avoid fraud. a unique data processing layer of defense is projected for finding these problems. For detective work the frauds inside numerous applications, 2 algorithms named Communal Detection and Spike Detection that generate novel layers.

There is an outsized moving window, higher numbers of attributes and numbers of link sorts obtainable which may be searched by CD and Coyote State algorithms. Thus, results may be generated by the system by intensely consuming a large quantity of your time. Since the attackers don't get time to switch their behaviors with relevance to the algorithms being deployed in real time, there's no true analysis achieved even after a daily update of the algorithms. Therefore, it's unimaginable to properly demonstrate the thought of ability. These problems may be resolved by making sure enhancements within the projected algorithmic rule in future work

In the paper presented by Sameena Naaz [7] investigated many strategies that have been used for detective work and the improper transactions and furnished a comparative examination among them.The fallacious transactions may be detected by utilizing either one amongst these or group action any of those ways. The model will probably be trained in a very additional correct manner by adding new options. Many data processing techniques square measure being employed by bank and credit card corporations for detective work fraud behaviors. The conventional usage pattern of shoppers relying upon their past activities may be known by applying any of those ways.

Therefore, a comparative analysis is formed here by learning completely different fraud detection techniques projected over the years

III. PROBLEM STATEMENT AND PROJECT SCOPE

A.Problem Statement

With the expansion of e-commerce websites, people and financial firms believe on-line services to hold out their transactions that have an exponential increase within the credit card frauds. Credit Card fraudulent transactions cause a loss of a big quantity of cash. To design a good fraud detection system is important so as to scale back the losses incurred by the purchasers and monetary firms. The detection of faulty transactions in credit cards

B.Project Scope

Credit card fraud detection could be highly regarded however conjointly a difficult drawback to unravel because of the difficulty of getting only a restricted quantity of information, credit cards build it challenging to match a pattern for a dataset. There are often many entries within the dataset with truncations of fraudsters which conjointly can work a pattern of legitimate behavior. Also, the problem has several constraints. Because the bank never gives authorization information of customers and the data sets aren't easily accessible to the general public, the analysis results are often buried and controlled, making them inaccessible; it's difficult to benchmark for the models built.

Building a system that improves the strategies formed by the fact that the protection concern imposes a limitation to

exchange of concepts and strategies in fraud detection, and especially in credit card fraud detection.

IV. DESIGN DETAIL

In design details, studying the System Architecture and System Modules in detail. To study the flow and process of the entire project in order to develop the project in an orderly and systematic manner.

A.System Architecture

This system accepts a true time client credit card dealing information, it is more important to search out fraud rate of credit card. information collection: collect input dataset supported transaction details.In figure 1, we show the detailed system architecture

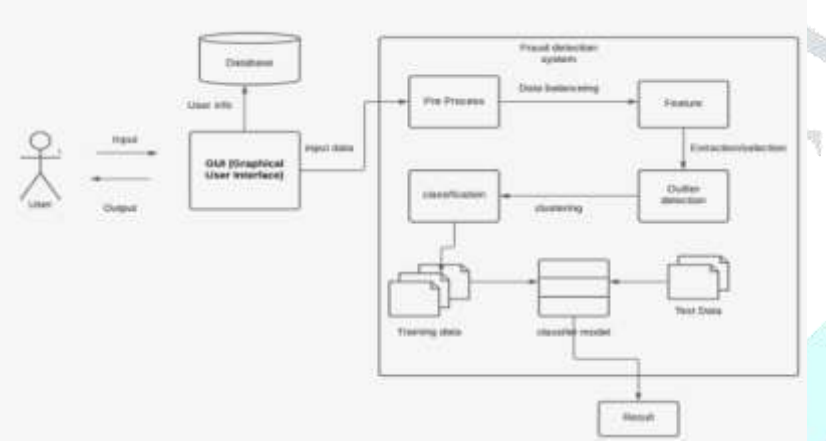

Figure 1. System Architecture

i.Data balancing : Once grouping an oversized set of knowledge bases it's necessary to grasp and separate the balanced knowledge and unbalanced data in 2 forms of category.Class-0 indicates non-fraud and Class-1 indicates fraud.

ii.Feature extraction and selection : Class-1 means that there are 492 samples in total for fraud transactions during this project v1, v2 …v28 options.

iii.Outlier detection: It measures the space between every similar knowledge to the agglomeration technique. The values that don't follow the trained knowledge square measure thought-about outliers.

iv. Classification : The dataset is unbalanced, classifying is very important for classification because it can offer the labels for classification.

After standardizing the whole dataset, it split the dataset into a coaching set likewise as a take a look at a set with a split quantitative relation of zer [2] This means that eighty percent [2] of our knowledge will be ascribed to train data, whereas two hundred percent will be credited to looking at knowledge.

B.System Design

a.Functional requirements

The model ought to be ready to offer correct and trustworthy predictions.The application should show graphical visualization of the expected results and Data data don't seem to be provided within the dataset due to confidentiality problems.

b.Non functional requirements

Non Functional needs can describe how a system should behave and what constraints are decided to impose on its usability.It typically specifies the system's quality attributes or characteristics :

i. Any verification system can be dealt with due to the system's availability.
ii. The accuracy of the system should be as high as the potential for higher prediction.
iii.The system's maintainability should keep track of all of the records that have been created.
iv. The system's usability should meet the needs of a wide range of banking industry users.

C.Training & Testing :

Split data into the training data and testing data take a look at information.Currently the two completely different datasets .Train information are used for coaching our model and also the information that is unseen are used for testing.The coaching set is the set of transactions that are used for coaching the prediction model.

Examine the set of transactions that are used to evaluate the prediction model's performance. The coaching team set out to find a ready-to-predict prediction model,whether or not a deal is real or dishonorable. that may be trusted for this task on the Python sklearn library, which provides easy-to-use functions to coach prediction models.

V. METHODOLOGY

We analyze the dataset and classify the transactions as fraud or legit. During this paper we have a tendency to use 2 completely different algorithms for our planned model on the Kaggle dataset for sleuthing frauds within the credit card system using python.Which are discussed briefly and their performance compared below.

Algorithms being deployed in real time, there is no true Comparison square measure created for these algorithms to see which algorithms offer higher results and might be custom-made by credit card for characteristic frauds.

A.KNN

The algorithmic rule of K-nearest neighbors (KNN) is a type of supervised metric capacity unit algorithmic rule. The K-nearest neighbors (KNN) algorithm to predict the values of new purpose |information based on 'feature similarity.'that means the new knowledge point will be appointed a price supported however closely it matches the points within the coaching set. We will perceive its operating with the assistance of following steps − therefore throughout the primary step of KNN, we have a tendency to load the coaching additionally as we take a look at knowledge. Next, we need to settle on the worth of K i.e. the closest knowledge points.

KNN are often any number. For each purpose within the take a look at knowledge do the following:

Calculate the gap between taking a look at knowledge and each row of coaching knowledge.Using one of the following techniques: euclidean, manhattan or teleaction.The foremost ordinarily used technique to calculate distance is euclidean. Now, support the gap worth, sort them in ascending order. Next, it'll select the highest K rows from the sorted array. Now, it'll assign a category to take a look at the point supported most frequent category of those rows and so End.

B.NAIVE BAYES

c

C. Dataset

In this paper we've taken a local source dataset that is from Kaggle[2]. The dataset is in csv format (creditcard.csv), it contains credit card transactions that were created by customers throughout Sep 2013 in Europe containing 284,807 transactions. By observation the behavior of the transactions credit card transactions square measure characterized into 2 categories: fraudulent and non-fraudulent. Original options and a lot of background data don't seem to be provided within the dataset due to confidentiality problems.

Only numeric input variables are provided at the results of Principal part Analysis (PCA) Transformation. options V1, V2 ... V28 are the principal parts obtained 'Time' and 'Amount' are the only parameters that haven't been updated with PCA.Feature 'Time' contains the seconds passed on between every group action and also the 1st group action within the dataset. The feature 'Amount' is the group action quantity. Feature 'Class' is that the response variable and it takes price one just in case of fraud and zero otherwise.

D. Tools

The list of tools accustomed explore credit card fraud detection analysis is as:
This projected model is enforced in Python. Numpy and Pandas.Seaborn is used for statistical data visualization and for algorithms we used Sklearn.

E.Details of Modules

a. Data Preprocessing
b. Scoring Rule
c. Classification of Alerts
d. Ranking of Alert

a.Data Preprocessing

In this module selected information is formatted, clean and sampled.The data preprocessing steps include the following:

i.Formatting : The data which has been selected may not be in a suitable format. The data may be in a file format and we may like it in relational databases or vice versa.

ii.Cleaning : The process of removing or replacing missing data. The dataset may contain records which may be incomplete or it may have null values. Such records need to be removed.

iii.Sampling : The range of frauds within the dataset is a smaller amount than the dealings, class distribution is unbalanced in credit card transactions.Therefore sampling methodology is employed to resolve this issue.

b.Scoring Rule

Scoring rule share of fraud in transactions is called a score. This module assigns score by matching recent dealings patterns with the past dealings pattern of the cardholder. If the score is larger than the dealings is considered as suspicious and any continuing is stopped.Otherwise it's captive to subsequent modules.

c.Classification

Classification of Alert Here a machine learning model can be used that may train and update the information supported feedback and delayed samples. Classifiers are going to be trained individually using feedback and delayed samples and their chances will be collective to spot alerts. Transactions that will have a high likelihood.will be alerted. thus solely a limited range of alerted transactions square measure reported to investigators.

d.Ranking of Alert

This module ranks every alert supported correctness of security queries. These security queries are going to be created on each occasion whenever the dealings are known to be suspicious.The alerts square measure hierarchic victimization likelihood. If it's found that associate degree alert has a bigger likelihood than alternative alerts then it's added to a queue and also the location of the fraudster is half-track. This feature makes the system user friendly and helps to file complaints against fraud.

F.Approach

Step 1: browse the dataset.
Step 2: random sampling is finished at the records set to create it balanced.
Step 3: Divide the dataset into 2 additives i.e., Train the dataset and take a look at the dataset.
Step 4: Feature choice is implemented for the proposed models.
Step 5: Accuracy and overall performance metrics are calculated to understand the efficiency for one of a kind algorithms.
Step 6: Then retrieve the best technique supported potency for the given dataset.

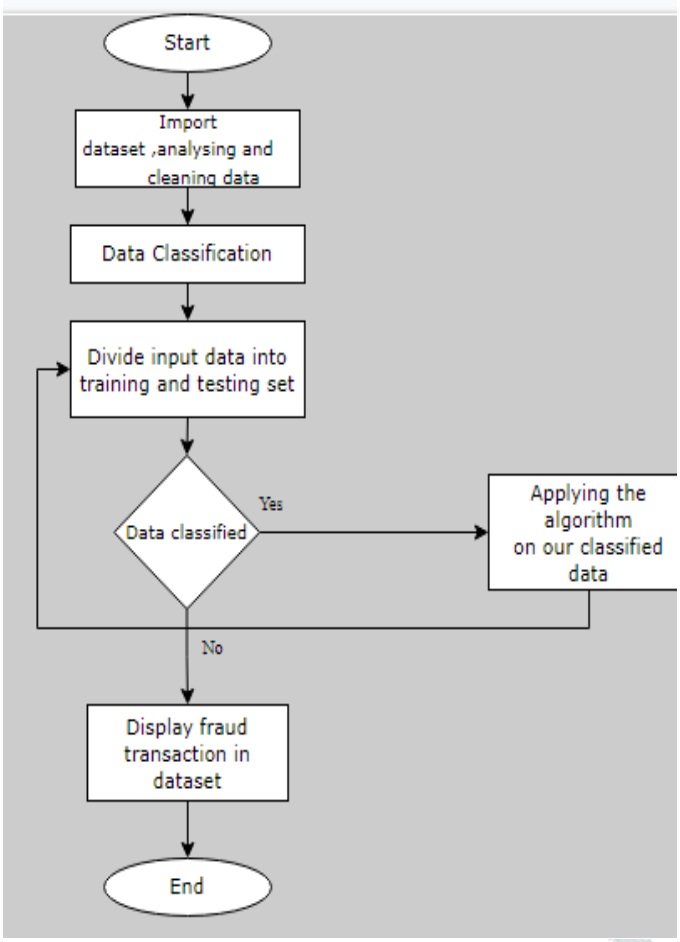

Figure 2. Flowchart of Proposed algorithm

VI.EXPERIMENTAL SETUP

A. Evaluation Metrics

Many type duties use honest evaluation metrics like Accuracy to test overall performance among models, because of accuracy is easy stay to enforce and generalizes to pretty sincerely binary labels.To classify the transactions as fraud or non-fraud we would like any other requirements of correctness that vicinity unit as:

i. Precision
ii. Recall
iii.Support.

i.Precision: It's the quantitative relation of well anticipated Positive observations to the anticipated wonderful observations.

Precision: TP/TP*FP

ii.Recall: It's a quantitative relation of well anticipated wonderful observations to all observations withinside the real class affirmative.

Recall: TP/TP+FP

iii.Support: Is the quantity of occurrences of each class in accurate goal values.

VII .RESULTS AND DISCUSSION

The exclusive purpose of this system was to detect the fraud from the given details. Credit Card Fraud Detection being a webapp gives the user a comfort to predict its term from any device, the user just needs an active internet connection and give required informative details. We have kept the interface simple yet appealing at the same time.
This chapter includes the snapshots of the actual outputs that were seen by the user and this chapter also contains the results of the proposed system.
The system is designated as such only an administrator has the access to the system to fill in details and make changes, as it is mentioned above information about customers banks does not provide or give authority for local purposes , the goal of our project is to create a system for a single trusted organization that offers cards to customers.

```
[ ] from sklearn.naive_bayes import GaussianNB

[ ] drop_list = []
    X_train, X_test, y_train, y_test = split_data(df, drop_list)
    y_pred, y_pred_prob = get_predictions(GaussianNB(), X_train, y_train, X_test)
    print_scores(y_test,y_pred,y_pred_prob)

    Index(['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11',
           'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21',
           'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Class', 'Time_Hr',
           'scaled_Amount'],
          dtype='object')
    train-set size:  227845
    test-set size:  56962
    fraud cases in test-set:  98
    train-set confusion matrix:
     [[222480   4971]
     [    69    325]]
    test-set confusion matrix:
     [[55535  1329]
     [   15    83]]
    recall score:  0.84693877551
    precision score:  0.0587818696884
    f1 score:  0.109933774834
    accuracy score:  0.976405322847
    ROC AUC: 0.963247971529636
```

Figure 3. Result for Naive Bayes Algorithm

```
[ ] from sklearn.neighbors import KNeighborsClassifier
    clf = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
    clf.fit(x_train, y_train)

    KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                         metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                         weights='uniform')

[ ] y_pred = clf.predict(x_test)
    print("Training Accuracy: ",clf.score(x_train, y_train))
    print("Testing Accuracy: ", clf.score(x_test, y_test))
    cm = confusion_matrix(y_test, y_pred)
    print(cm)
    print(classification_report(y_test,y_pred))

    Training Accuracy:  0.9996161138558128
    Testing Accuracy:  0.9995084407741356
    [[71065     8]
     [   27   102]]
                  precision    recall  f1-score   support

               0       1.00      1.00      1.00     71073
               1       0.93      0.79      0.85       129

        accuracy                           1.00     71202
       macro avg       0.96      0.90      0.93     71202
    weighted avg       1.00      1.00      1.00     71202
```

Figure 4. Result for KNN Algorithm

As shown in Figures 3 and 4, it displays the outcome of detection, i.e. whether or not the information provided is fraudulent, the accuracy of the input provided by the user.

VIII. CONCLUSION AND FUTURE SCOPE

Chances of credit score card frauds are growing hugely with the growth in utilization of credit score playing cards for transactions.This paper examines the detection of credit card fraud using a variety of Machine Learning methods on an online dataset.The projected machine is carried out in PYTHON. Analyzing the dataset gave the best accuracy charge of KNN so NAIVE bayes.

The scope of this project is to detect frauds when the data is collected in large amounts as it is not possible for now to generate reports for individual transactions. The future of this project shall be to identify the fraud in transactions on a personal level at high speed so that the user gets notified about the consequences at that time. In short, this project is applicable for mid size organization as they will have numerous amount of data

XI.ACKNOWLEDGMENT

We take this opportunity to express our deep sense of gratitude to our project guide Mrs.Hezal Lopes for her continuous guidance and encouragement throughout the duration of our mini project work. It is because of her experience and wonderful knowledge; we can fulfill the requirement of completing the mini project within the stipulated time. We would also like to thank **Dr. Jitendra Saturwar**, Head of computer engineering department and **Mrs. Vishakha Shelke, Mr. John Kenny,** Project Coordinators for their encouragement,whole-hearted cooperation and support.

We would also like to thank our Principal, **Dr. J. B. Patil** and the management of Universal College of Engineering, Vasai, Mumbai for providing us all the facilities and the work friendly environment. We acknowledge with thanks, the assistance provided by departmental staff, library and lab attendants

X. References.

[1] Munira Ansari, Hashim Malik,Siddhesh Jadhav, Zaiyyan Khan J. “Credit Card Fraud Detection” IJERT NREST - 2021 Conference Proceedings

[2] Kaggle Machine Learning Group ULB —Credit Card Fraud Detection, 03 May 2021

[3] Vinaya D S , Satish B Basapur, Vanishree Abhay, Neetha Natesh 4“Deep Learning Detecting Fraud in Credit Card Transactions.” 2018 Systems and Information Engineering Design Symposium (SIEDS), 2019

[4] Randhawa, Kuldeep, et al. “Credit Card Fraud Detection Using AdaBoost and Majority Voting.” IEEE Access, vol. 6

[5] Random Forest for Credit Card Fraud Detection.” 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2019.

[6] O. Harrison, “Machine Learning Basics with the K-Nearest Neighbors Algorithm”, 1. [Accessed: 18-Sep-2019].

[7] M. Pichler, V. Boreux, A. Klein, M. Schleuning, and F. Hartig, “Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks", Methods in Ecology and Evolution, pp. 1–13, 2019