FRAUD DETECTION IN CREDIT CARDS USING MACHINE LEARNING

Tummalla Ekshita, Garapati Naga Venkata Prudhvi, Vanamadi Anusha, Abdul Kaleem

Gudlavalleru Engineering College

Abstract-
Credit Card Fraud is growing substantially with the evolution of modern technology. Credit card fraud happens frequently and leads to massive financial losses. Online transaction have increased drastically significant no of online transaction are done by online credit cards. Identification of fraud credit card transactions is important to credit card companies for the prevention of being charged for items transaction of items which the customer did not purchase. So this paper aims to illustrate the modeling of a knowledge set using machine learning with Credit Card Fraud Detection.
The proposed model will determine whether a new transaction tends to be fraud or legitimate. So, the objective of the paper is to detect 100% of the fraud transactions while reducing invalid fraud classifications. **Here an extensive review is done on the existing and proposed models for credit card fraud detection and has done a comparative study on these methods. So different classification models are applied to the training data and the model performance is assessed on the basis of evaluation metrics such as accuracy, precision, recall, f1 score, confusion matrix. The aim of the study is to determine the best classifier by training and testing using machine learning methods ie both supervised and unsupervised techniques that delivers better results.**

Keywords— ***Credit card, classification, proposed methods, Supervised techniques***

I. INTRODUCTION

In technical terms, criminal deception which brings personal or financial gain is expressed as fraud. To eliminate these frauds, we can follow two methods i.e. fraud prevention & fraud detection. People or credit card holders has to be cautious and follow preventive steps that involves fraud prevention which prevents the fraud to happen in the first place. Other option to recur the loss of fraudulent transaction is fraud detection.
The illegitimate usage of credit card or its information is considered as credit card fraud. The transactions using credit card can be categorized in two ways i.e. using credit cards in person and in digital transactions. The impostors use this confidential information that includes credit number, expiry date & OTP i.e. verification number to accomplish transaction through internet or telephone. The statistics states that there is an exponential rise of usage in credit cards all over the world and credit cards frauds also increases with the same rate. To diminish this loss occurred by the credit card frauds, an operative system of detecting fraud is necessary to lessen or eradicate these cases. One of the simple ways is to detect fraud is to analyze the spending patterns on every card and to find out any distinction to the "usual" spending patterns. For this, analyzing the existing data purchase of cardholder is the fine way to lessen the rate of successful credit card frauds. Even main concern for banks and customers is to find a best process for detecting fraudulent operations through machine learning techniques. Since nowadays, massive amount of data available for each customer and activities, artificial intelligence can be used to effectively to identify suspicious patterns in transactions. To increase the accuracy and performance of the analyses, many institutions are investing in the improvement of ML algorithms. Various studies have been made on detecting fraudulent transactions of credit card. Finally, some approaches of machine learning like "artificial neural networks, rule- induction techniques, decision trees, logistic regression, and support vector machines etc.. " are useful. These approaches can be used in combined manner or standalone.

II. OBJECTIVES

The aim of the project is to implement machine learning techniques for credit card fraud detection with respect to time and amount of transactions and to build a classification model with accuracy.

III. LITERATURE SURVEY

In previous work and studies shows that many methods have been implemented to detect fraud using supervised, unsupervised algorithms and hybrid ones. Fraud types and patterns are changing everyday. It is important to build a intelligent model to detect frauds in credit cards using machine learning. Many literatures relate to anomaly or fraud detection in this domain are published already and are accessible for public practice. Here are some discussed machine learning models and algorithms and fraud detection models used in earlier studies.

In [1] data mining techniques are discussed are time consuming with huge data. Overlapping is another trouble with credit card transaction data. Imbalanced data distribution is also a major issue to overcome using sampling techniques.

In [2] discuss about data that is Fraud transaction is very less compared to normal transaction data. It may lead to serious bais situations.. Also discuss about difficulties in dealing categorial data. Many machine learning algorithms are not compatible with categorial data. Discuss about the detection cost and adaptability as a challenge. Prevention cost and cost of fraudulent behavior are taken into consideration

In [3] Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction. It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

A extensive survey[4] presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms drew an unexpected accomplishments in some areas, they failed to provide a prominent and consistent result for fraud detection.

IV. DATABASE PREPARATION

The dataset used in this project was provided by Kaggle which contains transactions made by credit cards in September 2013 by European cardholders. The dataset shows transactions occurred within two days. The dataset is highly unbalanced since it has 492 (0.17%) frauds out of 284,807 normal transactions. All the features or parameters in the dataset are numerical. Due to customers confidentiality, the columns were renamed to V1, V2, V3 ..., V28, and its parameters are applied to a PCA transformation which results to one or more of the smallest principal components, resultant in a lower-dimensional dataset that conserves the most data variance. The only two exceptions were the features Time and Amount, expressed in the seconds passed between each transaction and the transaction amount, respectively. The feature Class is the dependent variable and takes two values: that is

0 for normal transactions
1 for fraudulent transactions.

V. PROPOSED TECHNIQUE

The proposed system overcomes the issue in an efficient way. It aims at analyzing the number of fraud transactions that are present in the dataset. We use different machine learning algorithms such as Naïve Bayes, Decision Tree Random forest algorithm, SVM, to classify the credit card transactions in the dataset.
The algorithm is as follows

Algorithm for processing:

Step 1: Read the dataset.

Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Apply dimensionality reduction methods such as PCA to the dataset for more reliability.

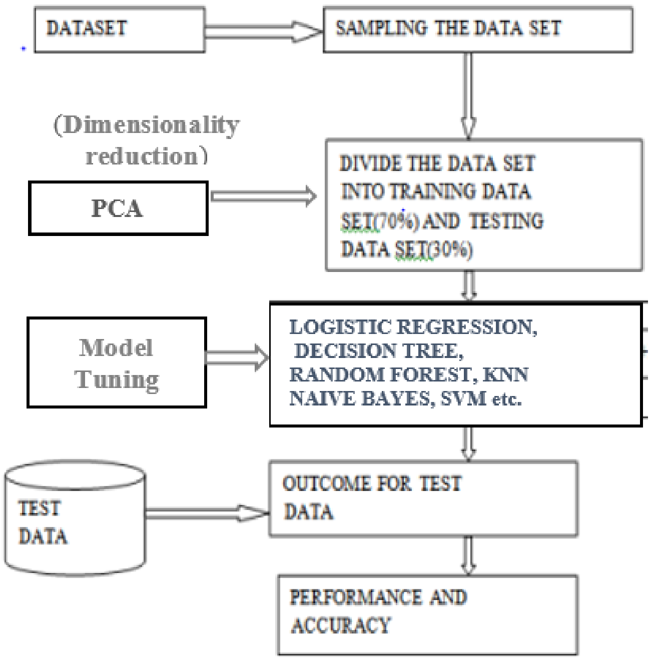
Step 4: Feature selection and Split the dataset into two parts i.e., Train dataset and Test dataset.

Step 5: Implement different machine learning algorithms by fitting training dataset.

Step 6: Accuracy and performance metrics has been evaluated to know the efficiency for different algorithms.

Step 7: Then retrieve the best algorithm based on efficiency for the given dataset.

The system architecture or pipeline is as follows


DATASET
SAMPLING THE DATA SET
(Dimensionality reduction)
PCA
DIVIDE THE DATA SET INTO TRAINING DATA SET(70%) AND TESTING DATA SET(30%)
Model Tuning
LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST, KNN NAIVE BAYES, SVM etc.
TEST DATA
OUTCOME FOR TEST DATA
PERFORMANCE AND ACCURACY

At first, all necessary dependencies are imported, then data preparation is performed. Later, data preprocessing is performed. It's very important task in this project. Data processing involves data cleaning, handling missing values, feature selection, sampling, normalization, dimensionality reduction (PCA), splitting the dataset etc...

DATA PREPROCESSING:

A. *Normalization*

For this a StandardScaler is used to transform the data so that its distribution will have a mean value 0 and a standard deviation of 1. This is a crucial step in that the data is transformed to be easily taken by the ML algorithm.

B. *Feature selection and splitting*

After transforming the Amount and Time features, let's split our dataset into train and test data. The size of the test data is 0.25

C. *Balancing the dataset*

the dataset is extremely unbalanced. Since there is a severe skew in the class distribution (284,315 entries in Class = 0 and 492 in Class = 1), training dataset could be biased and impact the machine learning algorithm to show unsatisfactory outcomes.

D. *Dimensionality reduction with PCA*

Imbalanced dataset classifications pose a challenge for predictive modelling as most of the machine learning techniques used for classification are based on assumption of an equal number of examples for each class in the dataset. This leads to models that have poor predictive performance, especially for the minor class. This issue is solved by Principal Component Analysis (PCA) algorithm

.

IMPLEMENTED CLASSIFICATION ALGORITHM

A. LOGISTIC REGRESSION:

Logistic regression becomes a classification technique with decision threshold on the linear regression output. Here the threshold value is a very important thing in logistic regression and is reliant on the classification problem itself. Generally, it uses sigmoid function for classification.

B. NAIVE BAYES:

Naive Bayes models compute the probability of an example to be of a certain class, based on prior knowledge. It is based on Naïve Bayes Theorem, that assumes that are features are independent of each other.

. P (A/B) = (P (B/A) * P (A)) / P (B)

Where, P (A) – Priority of A P (B) – Priority of B

P (A/B) – Posteriori priority of B

C. DECISION TREE:

Decision tree can be used for both classification and regression problems. Working is same for both same but some formulas are different. It uses the entropy and information gain for the formula.

D. RANDOM FOREST:

The random forest is a supervised learning algorithm in which it randomly creates and combines multiple decision trees into one unit called "forest." The aim is not to depend on a single decision tree. It improves accuracy and solves overfitting problem.

E. KMeans:

K-Means is an unsupervised learning method which is used to solve the clustering problems. It groups the dataset into different clusters. Here K defines the number of clusters.

F. SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine or SVM is a popular Supervised Learning algorithms in which it creates the best line or decision boundary that can separate n-dimensional space into classes so that a new data point can easily put in the correct category in the future.

G. K-NEAREST NEIGHBOR (KNN):

K-NN is a non-parametric algorithm which assumes that the similarity between the new data and existing data and put the new example into the category that is most similar to the available classes.

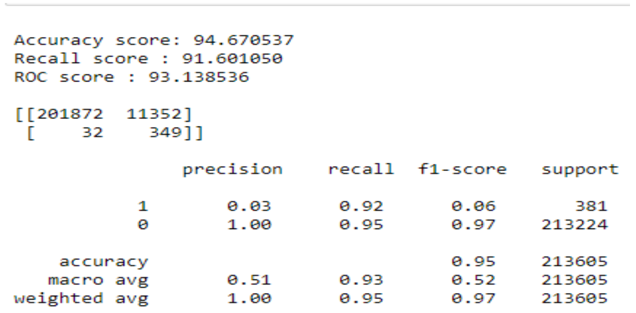
VI. Experimental Results and Discussions

In this project, different machine learning algorithms are defined and implemented with Kaggle dataset for credit card fraud detection. Since the dataset is highly imbalanced, we sample the dataset. Here we split 80% dataset for training and remaining for testing.

The **performance evaluation** of the algorithms is done based on F1 score, precision, recall (sensitivity) and accuracy, confusion matrix. Generally, f1 score, ROC score and confusion matrix is best for these kinds of projects.

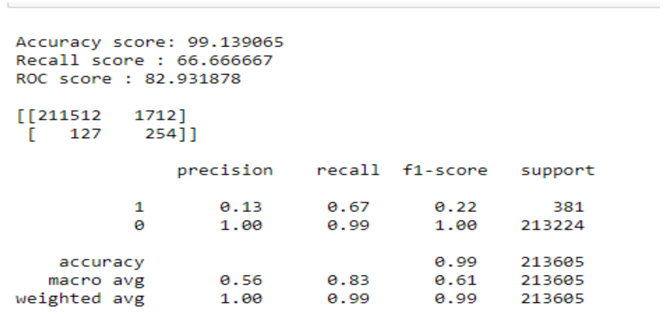
The given below is the results of all algorithms.

Logistic Regression evaluation:



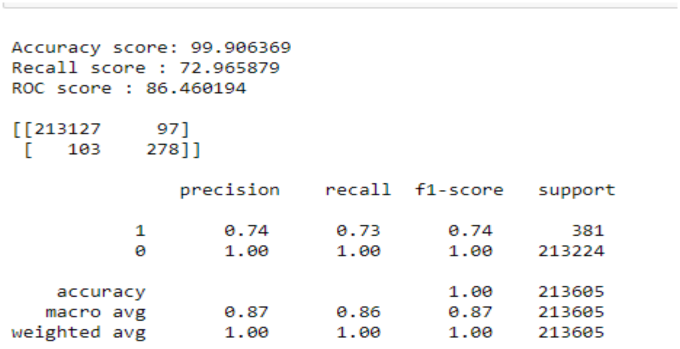
```
Accuracy score: 94.670537
Recall score : 91.601050
ROC score : 93.138536

[[201872  11352]
 [    32    349]]

              precision    recall  f1-score   support

           1       0.03      0.92      0.06       381
           0       1.00      0.95      0.97    213224

    accuracy                           0.95    213605
   macro avg       0.51      0.93      0.52    213605
weighted avg       1.00      0.95      0.97    213605
```

Naïve Bayes Evaluation:



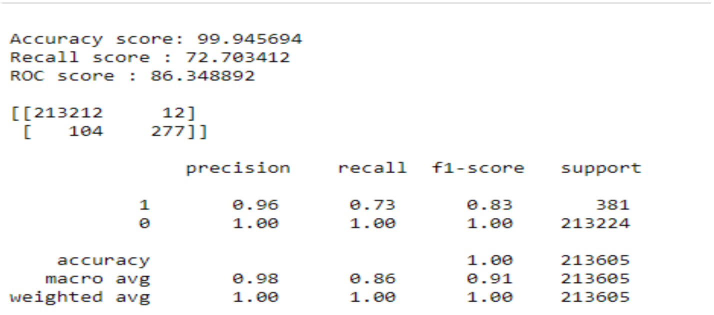
```
Accuracy score: 99.139065
Recall score : 66.666667
ROC score : 82.931878

[[211512   1712]
 [   127    254]]

              precision    recall  f1-score   support

           1       0.13      0.67      0.22       381
           0       1.00      0.99      1.00    213224

    accuracy                           0.99    213605
   macro avg       0.56      0.83      0.61    213605
weighted avg       1.00      0.99      0.99    213605
```

Decision tree Evaluation:



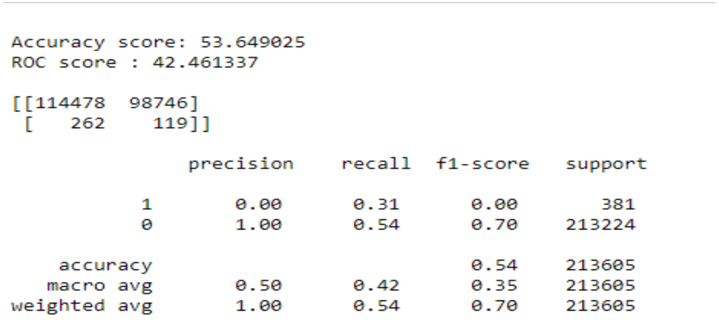
```
Accuracy score: 99.906369
Recall score : 72.965879
ROC score : 86.460194

[[213127     97]
 [   103    278]]

              precision    recall  f1-score   support

           1       0.74      0.73      0.74       381
           0       1.00      1.00      1.00    213224

    accuracy                           1.00    213605
   macro avg       0.87      0.86      0.87    213605
weighted avg       1.00      1.00      1.00    213605
```

Random Forest Evaluation:



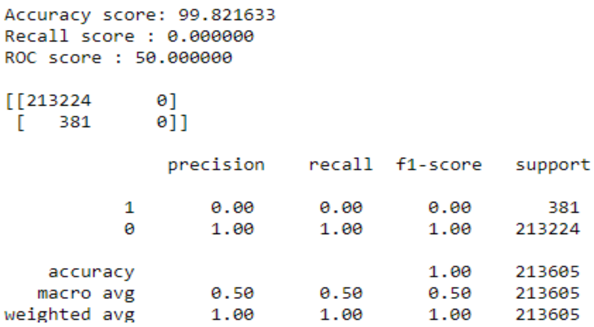
```
Accuracy score: 99.945694
Recall score : 72.703412
ROC score : 86.348892

[[213212     12]
 [   104    277]]

              precision    recall  f1-score   support

           1       0.96      0.73      0.83       381
           0       1.00      1.00      1.00    213224

    accuracy                           1.00    213605
   macro avg       0.98      0.86      0.91    213605
weighted avg       1.00      1.00      1.00    213605
```

Kmeans evaluation:



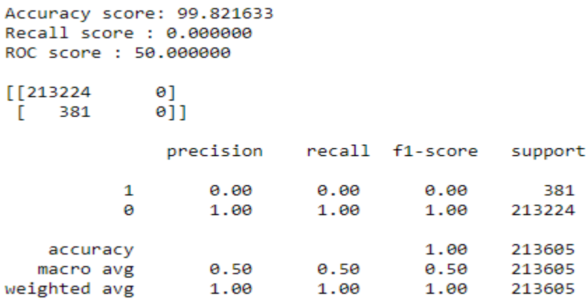
```
Accuracy score: 53.649025
ROC score : 42.461337

[[114478  98746]
 [   262    119]]

              precision    recall  f1-score   support

           1       0.00      0.31      0.00       381
           0       1.00      0.54      0.70    213224

    accuracy                           0.54    213605
   macro avg       0.50      0.42      0.35    213605
weighted avg       1.00      0.54      0.70    213605
```

Support Vector Machine (SVM) evaluation:



```
Accuracy score: 99.821633
Recall score : 0.000000
ROC score : 50.000000

[[213224      0]
 [   381      0]]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       381
           0       1.00      1.00      1.00    213224

    accuracy                           1.00    213605
   macro avg       0.50      0.50      0.50    213605
weighted avg       1.00      1.00      1.00    213605
```

K Nearest Neighbor (KNN) evaluation:



```
Accuracy score: 99.821633
Recall score : 0.000000
ROC score : 50.000000

[[213224      0]
 [   381      0]]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       381
           0       1.00      1.00      1.00    213224

    accuracy                           1.00    213605
   macro avg       0.50      0.50      0.50    213605
weighted avg       1.00      1.00      1.00    213605
```

The above metrics are helpful in performance evaluation. We can find the best algorithm from the given results.

VII. CONCLUSION

In this paper, we discussed applications of machine learning like Naïve Bayes, Logistic regression, Random Forest, SVM, KNN etc. ... and it proves that accurate in deducting fraud transactions and minimizing the number of false signals. The objective of the study was to find best method for this problem statement. Precision score, recall score, f1-score, confusion matrix and accuracy score are used to evaluate the performance of the models.

By comparing all them, we found that random forest performs better than all the algorithms like logistic regression, naïve bayes, svm, knn methods.etc. Kmeans method performs less than all the algorithms which shows that supervised algorithms outperform unsupervised algorithms. Second best method with better accuracy is SVM. KNN is also performs better with high accuracy but it is very costly in terms of computation while prediction.

.

In conclusion, this project suggests that If these ML algorithms are applied into banking credit card fraud detection systems, then the probability of fraud transactions can be predicted soon after credit card transactions. And a series of anti-fraudulent transaction strategies can be adopted and used to prevent banks from great financial losses.

VIII. REFERENCE AND ACKNOWLEDGE:

[1] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, “Random forest for credit card fraud detection”, IEEE 15th (ICNSC),2018.

[2] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, “A Tool for Effective Detection of Fraud in Credit Card System”, published in International Journal of Communication Network Security ISSN:

[3] International Journal of Scientific & Engineering Research, ISSN 2229-5518 IJSER © 2012 http://www.ijser.org Fraud Detection of Credit Card Payment System by Genetic Algorithm

[4] INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 10, OCTOBER 2020 ISSN 2277-8616 216 IJSTR©2020 www.ijstr.org Credit Card Fraud Detection Using Supervised Learning Approach

[5] K. Chaudhary, B. Mallick, "Credit Card Fraud: The study of its impact and detection techniques", International Journal of Computer Science and Network (IJCSN)

[6] M. J. Islam, Q. M. J. Wu, M. Ahmadi "Investigating the Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers", IEEE International Conference on Convergence Information Technology.

[7]. Dinesh L. Talekar, “Credit Card Fraud Detection System-A “ Survey, IJMER 2014.

[8]. SamanehSorournejad, Zahra Zojaji, ,” A Survey of credit card fraud detection techniques: Data and techniques oriented perspective”

[9] Lakshmi S V S S, “Machine learning for credit card fraud detection,” International Journal Of Applied Engineering Research 2018.