International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

Credit Card Fraudulent Transaction Detection

Dr. Naveen M, S Aishwarya Rao, Rubina Shrestha, Pooja Tiwari

Department of Information Science and Engineering, R.R Institute of Technology, Bengaluru, India

ABSTRACT*:* Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used. This paper investigates the performance of KNN, Naïve Bayes and Random Forest for credit card fraud detection. Dataset of credit card transactions is collected from Kaggle and it contains a total of 2,84,808 credit card transactions of a European bank data set. To detect such frauds, it is important to check the usage patterns of a user over the past transactions. Comparing the usage pattern and current transaction, we can classify it as either fraud or a legitimate transaction.

KEYWORDS: Fraud detection, Credit card, Random forest. KNN, Naïve Bayes.

I. INTRODUCTION

Credit card fraud is a huge ranging term for theft and fraud committed using or involving at the time of payment by using this card. The purpose may be to purchase goods without paying, or to transfer unauthorized funds from an account. Today, fraud detection systems are introduced to control one-twelfth of one percent of all transactions processed which still translates into billions of dollars in losses. Credit Card Fraud is one of the biggest threats to business establishments today. However, to combat the fraud effectively, it is important to first understand the mechanisms of executing a fraud. Credit card fraudsters employ a large number of ways to commit fraud. In simple terms, Credit Card Fraud is defined as "when an individual uses another individuals' credit card for personal reasons while the owner of the card and the card issuer are not aware of the fact that the card is being used". Card fraud begins either with the theft of the physical card or with the important data associated with the account, including the card account number or other information that necessarily be available to a merchant during a permissible transaction. Card numbers generally the Primary Account Number (PAN) are often reprinted on the card, and a magnetic stripe on the back contains the data in machine-readable format. It contains the following Fields:

- Name of card holder
- Card number
- Expiration date
- Verification/CVV code
- Type of card

There are more methods to commit credit card fraud. In the Traditional approach, to be identified by this paper is Application Fraud, where a person will give the wrong information about himself to get a credit card. There is also the unauthorized use of Lost and Stolen Cards, which makes up a significant area of credit card fraud. There are more enlightened credit card fraudsters, starting with those who produce Fake and Doctored Cards; there are also those who use Skimming to commit fraud. They will get this information held on either the magnetic strip on the back of the credit card, or the data stored on the smart chip is copied from one card to another. Site Cloning and False Merchant Sites on the Internet are getting a popularmethod of fraud for many criminals with a skilled ability for hacking. Such sites are developed to get people to hand over their credit card details without knowing they have been swindled.

II. RELATED WORK

In [4], as per the information from the United States Federal Trade Commission, the theft rate of identity had been holding stable during the mid 2000s, but it was increased by 21 percent in 2008. Even though credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints In 2000, out of 13 billion transactions made annually, approximately 10 million or one out of every 1300 transactions turned out to be

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

fraudulent.Also, 0.05% (5 out of every 10,000) of all monthly active accounts was fraudulent. Today, fraud detection systems are introduced to control one-twelfth of one percent of all transactions processed which still translates into billions of dollars in losses. In [1][7] authors aim to solve problem of imbalance in data along with Concept drift by using MCC, SVM and SMOTE.These classifiers were applied on three different groups later rating scores are generated for every type of classifier. These dynamic changes in parameters lead the system to adapt to new cardholder's transaction behaviors timely followed by a feedback mechanism to solve the problem of concept drift. It was observed that the Matthews Correlation Coefficient was the better parameter to deal with imbalance dataset. MCC was not the only solution. By applying the SMOTE, they tried balancing the dataset, where it was found that the classifiers were performing better than before. The other way of handling imbalance dataset was to use one-class classifiers like one-class SVM.The papers [2][6] represent a research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Logistic regression has an accuracy of 97.7% while SVM shows accuracy of 97.5% and Decision tree shows accuracy of 95.5% but the best results are obtained by Random forest with a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6% in problem of credit card fraud detection with dataset provided by ULB machine learning. In the study of [3][8], four classifier algorithms Logistic Regression, K-Nearest Neighbor, Naive Bayes and Decision Tree are developed. The hybrid sampling approach is used to convert imbalanced dataset into balanced dataset so as to get more accurate results. Later, the classifier algorithms are applied on the balanced dataset.The performance of these machine learning algorithms are studied based on performance metrics such as accuracy, sensitivity and specificity Two neural network algorithms (Multi-Layer Perceptron and Chebyshev Functional Link Artificial Neural Network) performance is compared based on Mean Squared Error (MSE). Results from the experiment shows that performance of KNN is better than other machine learning algorithms. It was observed that MLP outperforms as compared to Chebyshev Functional Link Artificial Neural Networks, even when CFLANN gives lower MSE because of the time taken by CFLANN to converge to a solution. Finding a fraud transaction is a timely action and hence it is the accuracy as well as the time required balance which is found in MLP better than CFLANN. In [5], Local Outlier Factor (LOF) rule is used which associate unattended anomaly detection technique that computes the native density deviation of given information with relevance of its neighbors. It considers as outliers the samples that have a considerably lower density than their neighbors. This instance shows a way to use LOF for outlier detection that is that the default use case of this figure in Scikit-learn.

III. PROPOSED SYSTEM

The system is developed into two parts namely the data preprocessing and the evaluation.

I. Data Preprocessing

The primary data collected from the online sources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. Data Preprocessing is a technique that is used to convertthe raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

The various preprocessing steps carried out are:

A. ***Data Cleaning*** - Filling in missing values is an important task in the data cleaning process.Tuples with meaningless value were removed from the files as they do not contribute to producing important data as well as they do not bias the data. Additional changes are removing unnecessary columns, separating the date time column into two.

B. ***Data Integration***- Before the data were subjected to further change the two data sources were integrated together since fraudulent and genuine record files were in two separate files.

C. ***Data Transformation*** -Here, all the categorical data were consolidated into an understandable numerical format. The transactional dataset contains several data types with several ranges. Therefore, data transformation comprises of data normalization. Data normalization scales the attribute data to fall in a small numeric range.

D. ***Data Reduction*** - The principal component analysis which is well-known that PCA is a popular transform method. Applying this method resolves the feature selection issue from the perspective of numerical analysis. PCA performed feature selection successfully by finding the suitable number of principal components.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

The imbalanced data must be converted to a balanced form for better accuracy. Hence, SMOTE (Synthetic Minority Over-sampling Technique) is used. It is an over-sampling method which creates synthetic samples of the minority class. Hence making the minority class equal to the majority class. SMOTE does this by selecting similar records and altering that record one column at a time by a random amount within the difference to the neighboring records.The various steps in SMOTE are:

1. Take sample `from the dataset
2. Consider the k nearest neighbour
3. Take vector between the neighbours and the current data point
4. Multiply the vector with a random number X in the range 0 to 1
5. Add the result to the current data and the synthetic data is formed.

II. Evaluation

The evaluation step includes splitting, training and testing the model. The various algorithms carried out are:

A. *K Nearest Neighbor*

KNN classifier is a supervised machine learning technique where result of new instance query is classified based on a majority of KNN category. This algorithm uses distance rule to derive a classification from K-Nearest neighbours. In CC fraud detection, we classify any incoming transaction by calculating distance of the nearest point to new incoming transaction. Then, if the nearest neighbour is fraudulent, then the transaction indicates as fraud.

Pseudocode for KNN Model

1. The k-nearest neighbor algorithm is imported from the scikit-learn package.
2. Create feature and target variables.
3. Split data into training and test data.
4. Generate a k-NN model using neighbors' value.
5. Train or fit the data into the model.
6. Predict the future.

B. *Naïve Bayes Classifier*

This algorithm is based on Bayes Theorem with an assumption of independence among features. It assumes that the presence of particular feature in a class is unrelated to the presence of any other feature. This algorithm is easy to build and is useful for large datasets. Bayes Theorem provides formula for calculating posterior probability of features in a class.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)} \quad \ldots \text{eq. (1)}$$

Where,

P(c|x) = Posterior Probability of class

P(c) = Prior Probability of class (or Class Probability)

P(x|c) = Likelihood probability of features of given class (Conditional Feature)

P(x) = Prior Probability of feature

Working of Naïve Bayes

1. Convert the data set into a frequency table
2. Create Likelihood table by finding the probabilities.The main objective function in Naïve Bayes classifier is to maximize theposterior probability given the training data for each class

$$Y_i^{pred} = \underset{c_i}{\operatorname{argmax}} \left(p(c_i) * \pi p(x_j|c_i) \right) \quad \ldots \text{eq. (2)}$$

whereπ is for the range ($1 \leq j \leq m$)

3. Use Naïve Bayesian equationto calculate the posteriorprobability for each class. The class with the highest posterior probability is the outcome of prediction.

C. ***Random Forest***

Random Forest is a supervised classification algorithm. This involves building several decision trees and combining them with outputs to improve generalization of the model. This method of combining trees is also known as Ensemble method. In this algorithm, a subset of training dataset is sampled randomly so that to train individual tree, then decision tree is built. Each node of the tree then splits on a feature selected from a random subset of full feature set.

Pseudocode for creating Random Forest

1. Randomly select "K" features from total "m" features where k << m.
2. Among the "K" features, calculate the node "d" using the best split point.
3. Split the node into daughter nodes.
4. Repeat thesteps until "l" number of nodes has been reached.
5. Build forest by repeating steps for "n" number times to create "n" number of trees.

Pseudocode for Random Forest Prediction

1. Take the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target
3. Consider the high voted predicted target as the final prediction from the random forest algorithm

The important steps in choosing the best algorithms are:

Step 1:Import the dataset.

Step 2: Convert the data into data frame format.

Step 3: Do random oversampling using SMOTE package.

Step 4: Decide the amount of data for training data and testing data.

Step 5: Give 80% of data for training and rest for testing.

Step 6: Assign train dataset into the model.

Step 7: Choose the algorithm among 3 different algorithms and create the model.

Step 8: Make predictions for test dataset using each algorithm.

Step 9: Calculate accuracy for each algorithm.

Step 10: Apply confusion matrix for each variable.

Step 11: Compare the algorithms for all variables and find out the best algorithm.

IV. RESULTS

In this study, three classifier algorithms K-Nearest Neighbor, Naive Bayes and Random Forest are developed. To evaluate the models, the dataset is divided in 80:20 ratio where 80% transactions are considered for training the models and 20% remaining for testing the accuracy of the models. The hybrid sampling approach is used to convert imbalanced dataset into balanced dataset so as to get more accurate results. Later, the classifier algorithms are applied on the balanced dataset. The performance measures are derived from confusion matrix. It is a 2*2 matrix that contains 4 outcomes produced by binary classifiers.

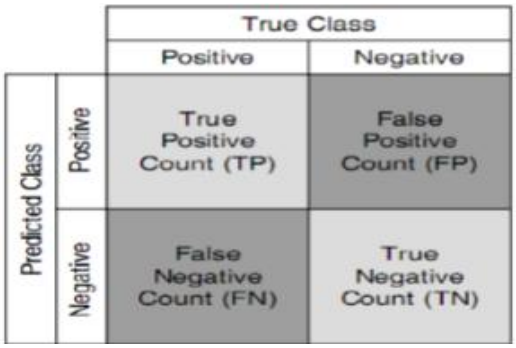

Fig.1: Confusion Matrix

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

The output of the metrics depended on the results obtained by:

1. True positive (TP)- number of fraud transactions predicted as fraud
2. True negative (TN)- number of legal transactions predicted as legal
3. False positive (FP)- number of legal transactions predicted as fraud
4. False negative (FN)- number of fraud transactions predicted as legal

The performance of the machine learning algorithms is evaluated based on True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR) and False Negative Rate (FNR) which are given as follows:

TPR = TP/P … eq. (3)

TNR = TN/N …eq. (4)

FPR = FP/N …eq. (5)

FNR = FN/P …eq. (6)

The performance of the machine learning algorithms are studied based on performance metrics such as accuracy, sensitivity and specificity which are given as follows:

Accuracy = (TP + TN)/ (TP + FP + TN + FN) …eq.(7)

Sensitivity= (TP)/ (TP + FN) …eq.(8)

Specificity= (TN)/ (FP + TN) …eq.(9)

Precision = TP/ (TP + FP) …eq.(10)

Recall= TP / (TP + FN) …eq. (11)

F-Measure = 2TP / (2TP + FP + FN) …eq. (12)

Sensitivity gives the accuracy on positive (fraud) cases classification. Specificity gives the accuracy on negative (legal) cases classification. Precision gives the accuracy in cases classified as fraud (positive)

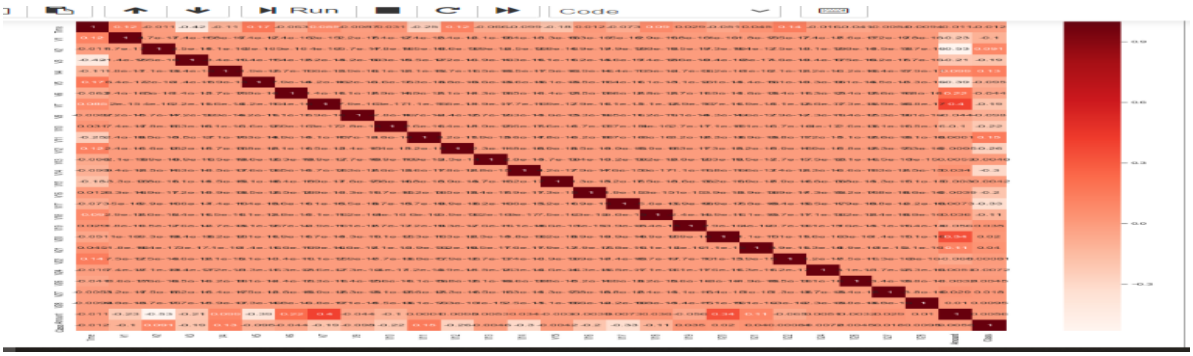

Fig.2: Confusion Matrix Graph

Metric	KNN Classifier (%)	Naïve Bayesian (%)	Random Forest (%)
True positive rate	97.38	79.30	99.99
False positive rate	6.000	0.559	0.014
True negative rate	94.39	99.06	99.98
False negative rate	2.439	34.511	0.005
Precision	94.19	94.29	99.98
Recall	97.38	79.30	99.99
F-measure	95.76	88.18	99.99
Sensitivity	97.38	79.30	99.99
Specificity	94.39	99.06	99.98
Accuracy	95.84	86.71	99.99

Table:1:Performance Measure for Various Algorithms

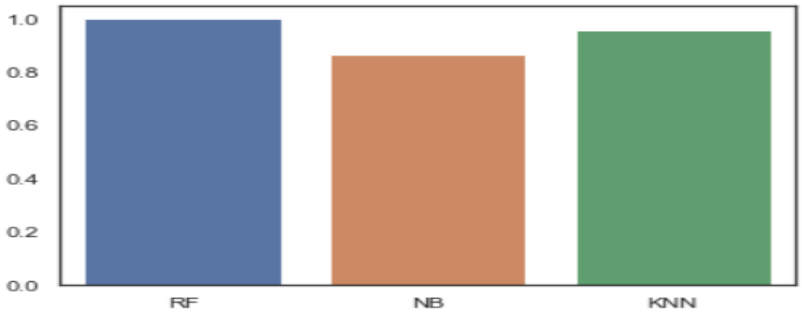

Fig.3: Accuracy Graph

V. CONCLUSION AND FUTURE WORK

Credit card fraud is escalating significantly with the advancement of modernized technology and became an easy target for frauds. Credit card fraud has highly imbalanced publicly available datasets. we apply supervised machine learning algorithms to detect credit card fraudulent transactions using a real-world dataset.Out of 3 supervised machine learning algorithms used for detecting credit card fraud transaction, from Table.1 we can see that Random forest algorithm gives 99.99% accuracy whereas Naïve Bayes gives 86.71% and KNN gives 95.84% accuracy. From Fig.3, Random Forest algorithm is found to be most reliable and accurate for fraud detection.Furthermore, we employ these algorithms to implement a classifier using machine learning methods. We identify the most important variables that may lead to higher accuracy in credit card fraudulent transaction detection.Since datasets are imbalanced, data preprocessing for balanced data is necessary for training the model. With credit card fraud cases increasing day by day, it should be considered serious issue and enhance the "Credit card fraudulent transaction detection" solution in real bases for victory over the frauds.

- Further enhancement can be done by making this system secure with the use of certificates for both merchant and customer.
- As technology changes new checks can be added to understand the pattern of fraudulent transactions and to alert the respective card holders and bankers when fraud activity is identified.
- The dataset available on day to day processing may become outdated, it is necessary to have updated data for effective fraud behavior identification.
- To this extent, the incremental approach is necessary in making the system to learn from past as well as present data and capable of handling the both.
- Fraudster uses different new techniques that are instantaneously growing along with new technology makes it difficult for detection. Also, the nature of access pattern may vary from one geographical location to another

(such as urban and rural areas) that may result in a false positive detection. In such a case a future enhancement may be based on new multiple models with varying access pattern needs attention to improve the effectiveness.

- Privacy preserving techniques applied in distributed environment resolves the security related issues preventing private data access.

REFERENCES

[1] Geetha S, International Conference On Recent Trends In Advanced Computing 2019, Icrtac 2019 Credit Card Fraud detection Using Machine Learning algorithms Vaishnavinathdornadulaa*,

[2] Navanshu Khare and 2Saad YunusSait, International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 825-838 Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models.

[3] Deepti DigheSnehaPatil Shrikant Kokate, Detection of Credit Card Fraud Transactions using Machine Learning Algorithms and Neural Networks: A Comparative Study 978-1-5386-5257-2/18/$31.00 ©2018 IEEE.

[4] Lakshmi S Selvani, Deepthi Kavila , Machine Learning For Credit Card Fraud Detection System, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 24 (2018) pp. 16819-16824 © Research India Publications.

[5] Kumar, Pawan and Fahad Iqbal. “Credit Card Fraud Identification Using Machine Learning Approaches.” 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) (2019): 1-4.

[6] Rimpal R. PopatJayeshChaudhary, A Survey on Credit Card Fraud Detection using Machine Learning , Proceedings of the 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018) IEEE Conference Record: # 42666; IEEE Xplore ISBN:978-1-5386-3570-4

[7] Majapuh, LjiljanaBrkic , Detecting Credit Card Fraud Using Selected ML Algorithms MIPRO 2019, May 20-24, 2019, Opatija Croatia

[8] Anuruddha Thennakoon, CheeBhagyan2, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi, Real-time Credit Card Fraud Detection Using Machine Learning 978-1-5386-5933-IEEE

[9] OngShu Yee, Saravanan Sagadevan and NurulHashimahAhamedHassainMalim, Credit Card Fraud Detection Using Machine Learning As Data Mining Technique, School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia.

[10] www.google.com

[11] www.kaggle.com

[12] www.tutorialspoint.com

[13] www.datacamp.com

[14] www.machinelearningmastery.com