International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

TECHNIQUES FOR DETECTING CREDIT CARD FRAUD

Ashish[1]*, Ms. Upasna[2]*
* *Department of Computer Science & Engineering, Ganga Institute of Technology and Management, Jhajjar Haryana (India)*

ABSTRACT: Credit card fraud is a critical crime. Fraud detection have impacts on many industries such as banks, retail, financial services, health care, etc. According to the Federal Trade Commission reports, data shows that credit card theft has raised by 44.6% from year 2019 to 2020. Newly released Federal Trade Commission data shows that in year 2021 about 389,737 credit card theft report are received. This research work focus to explore possible ways to identify credit card fraudulent activities that have a negative impact on financial institutions. In this paper Machine learning algorithms such as Random Forest, Decision Trees and Xgboost, K-Means, Logistic Regression and Neural Network are implemented for detection of fraudulent transactions. A comparative analysis of these algorithms is performed to determine the best model for predicting credit card frauds. Our result shows that Random Forest Algo give highest Accuracy, Precision and AUC score for credit card fraud detection.

Keywords: Credit Card, Credit Card Frauds, Frauds Detection, Machine Learning Algo, Financial sectors, Machine Learning, Re-sampling Methods.

1. INTRODUCTION

Fraud can be defined as a criminal scheme aimed at making money or personal gain. Fraud prevention and fraud detection programs are two main ways to avoid fraud and loss due to fraudulent activities. Fraud detection systems become operational when fraudsters go beyond fraud prevention programs and start fraudulent activities. Credit card fraud is any form of theft or fraud involving a credit card. The purpose of credit card fraud is to purchase goods without paying, or to steal money from someone else's account. Credit card fraud occurs one of the biggest threats to financial institutions and businesses today. Credit card fraud can be defined as "when an unauthorized person uses a credit card for personal use without the consent or knowledge of the card owner and the issuer of the card has no idea what the card is being used for." The objectives of the study is to obtain fraudulent credit card purchases over non-fraudulent transactions using a machine learning algorithm such as Random Forest, Decision Trees and Xgboost, K-Means, Logistic

Regression and Neural Network to predict fraud effectively and accurately.

2. REVIEW OF LITERATURE

Maniraj & saini et al. (2019) demonstrate the modeling of a data set using ML with Credit Card Frauds Detection. The authors try to detect transactions that are 100% fraudulent as they minimise the unfair fraud classification. The focus was on analyzing and preprocessing datasets and using multiple anomaly detection algorithms such as the Local Factor Isolation Forest algo on the PCA transformed Credit Cards Transaction Data. The results show that the algorithm achieves more than 99.6 accuracy, but its precision is about 28% when using a tenth of a set of data. However, as all data is entered into an algorithm, the accuracy increases to 33%. We expect this increase inaccuracy due to the significant difference between legal and genuine transactions [1].
Sadineni Parveen kumar (2020) in their paper they deals with various machine learning techniques such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Trees, Logistic Regression and Random Forest to identify frauds carried out by credit cards [2].
Mohari & Ankit et al. (2021) said those faulty activities conducted through credit cards could be tackled with Data Science, Machine Learning together with Deep Learning techniques. Another benefit of this is that it helps banks and other financial institutions to detect fraud before it can cause serious damage. On the other hand, the hackers need a small amount of data to carry out their malicious acts; this makes the victims vulnerable to danger. There are various techniques and methods of unsupervised learning. Mohari et al. (2021) described ten of them and compared them in their research. They compared Logistics Regression, Random Forest, AdaBoost, ANN, Genetic Algo, HMM, KNN Classifier, Decision tree, Isolation Forest, and Local Outlier Factor. Out of all the ten methods, their results show that Local

Outlier Factor fraud accuracy is greater than the rest of the algorithms [3].

3. DATA SOURCE

The dataset for this research work is obtained from Kaggle, and it was generated using Sparkov Data Generation, a GitHub tool created by Brandon Harris. The dataset is a simulated credit card transaction that contains legalize and fraudulent transactions. It contains the credit card of 1000 customers making transactions with a pool of 800 merchants.

4. METHODOLOGY

The methodology used in this study to classify the non-fraudulent transactions from the fraudulent transactions. Below Figure 1 shows the steps used in this work.

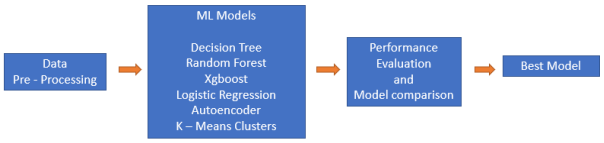

Fig. 1: Classification Methodology

5. DATA PREPROCESSING

Pre-processing data is required prior to using a machine learning algorithm, taking into account different models that produce different specifications in predictions, and data training can affect predictive production. The objectives of data preprocessing are to clean and refine the data into an area that includes more concise discrimination, non-existent values, and more. A resampling method such as under-sampling and over-sampling was performed on the imbalanced original dataset to avoid any kind of bias and over-inclusion in our training model. We have adopted Python data library pandas and machine learning sci-kit to fulfill these pre-processing. The steps are shown in fig.2

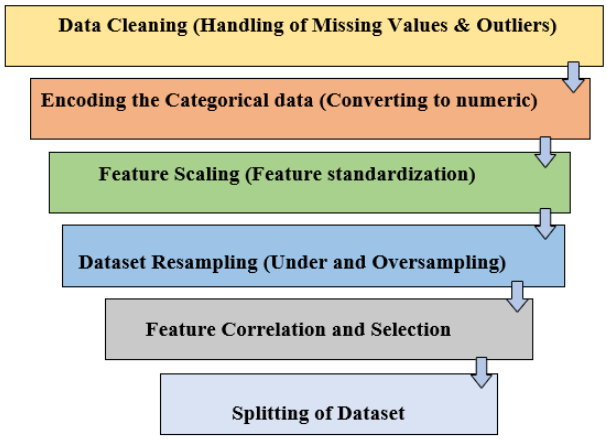

Fig. 2: The Data Preprocessing steps

The credit card dataset was imported using the python import command, and a data purification process was performed. During data cleaning we do two tasks; 1. Remove empty values and missing values, and 2. Outliers handling. The boxplot method is adopted to determine the presence of outsiders in all independent components. outliers were removed using the Inter Quantile Range (IQR) method which is one of the most popular methods of outsourcing as it is very strong for outsiders. After clearing the dataset, we convert any categorical values into numerical values as most machine learning algorithms do better with numerical input. we have used the One-Hot Encoder to convert the categorical variables into numerical values. We did a feature measurement using the Robust Scaler process, also known as robust standardization. Scaling can be achieved by calculating the 50 percent intermediate, 25th, and 75th percentiles. The data used in this study were not balanced; which is why we have done resampling methods such as Undersampling and Oversampling. For feature selection we use the lasso technique, which is a tool that helps minimize the cost function.

6. MACHINE LEARNING MODELS

6.1 DECISION TREE

Decision Tree is a Supervised learning method that can be used for both classification and Regression problems, but mostly it is preferable for solving Classification problems. The most widely algorithm used in Machine Learning applications is called the decision tree model.

Decision trees work very quickly and efficiently, especially when used for mining and analyzing large amounts of data. The decisions or the test are performed on the basis of features of the given dataset [4]. It follows a key root question and branches in which the details are used to form certain components that eventually reach the climax or leaves of the tree. It contains two nodes, decision node, and Leaf nodes [5]. An example can be seen below.

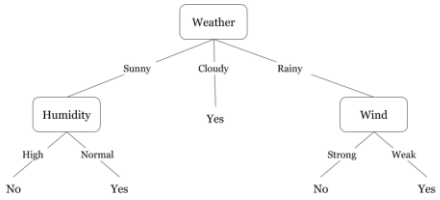

Fig.3: Decision Tress

6.2 RANDOM FOREST

The most widely used machine learning algorithm for Random Forest. It is an accepted method of solving both classification and regression problems. It is a pool with a large number of different decision-making trees called " forest ". Each different tree makes a class prediction. Any class with the most votes is considered to be predictable. Therefore, this process takes a bagging approach to creating a group of decision trees that will build the forest. The strength of this process is that feature selection is unnecessary, and it runs the model quickly and estimates errors intelligently. The downside of this process is that it is sensitive to data with various values and attributes with additional value and can easily flag you as fraud. The 'forest' that this algorithm builds is known as decision tree ensemble, which is often trained with a method called bagging, an application of Bootstrap strategy to a high variance algorithm used in machine learning [6].

6.3 AUTOENCODERS

An Artificial Neural Network (ANN) is an interconnected network of processing nodes, for example, "neurons," which collectively play out an (ordinarily nonlinear) change of contributions to specific ideal outputs. This process uses a set of connected neurons, and neurons contribute to decision-making [7].

6.4 XGBOOST CLASSIFIER

XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems [8].

6.5 K MEANS CLUSTERING

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters [9].

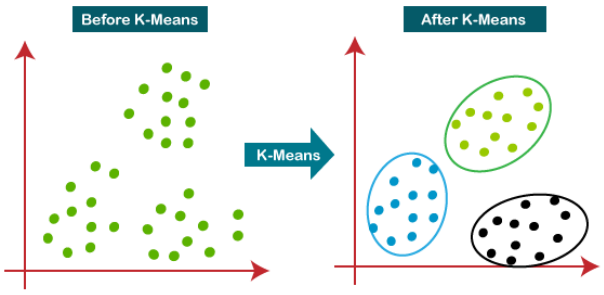

Fig.4: Working of k-means clustering algo

6.6. LOGISTIC CLASSIFICATION

This is a simple method used to solve classification and regression problems. It can be used in identifying emails, spam detection, among others. It creates opportunities for either binomial or multinomial output. Accept sigmoid function in interpreting data

and the relationship between dependent and independent variables. It can also be used in current research work to classify work as fraud or not. It works very well, although it can overfit high-dimensional datasets. It offers better accuracy and does not make assumptions about the scattering of classes in feature space as some other methods do. The cons is that it uses the assumption of linearity between the dependent and independent variables. Classification and regression functions are the two kinds of supervised learning, yet the yield factors of the two are unique.

7. RESULTS

In order to evaluate the performance of our model, we adopted the use of a metric called AUC score and other metrics to measure the performance of our model.

7.1 ACCURACY

Accuracy is defined as the ratio of the number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

7.2 RECALL

Recall is the ratio of true positive to the number of all samples, which should have been recognized as a positive value.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

7.3 PRECISION

Precision is find out by dividing the correct positive number results by the number of positive results that the classifier predicted.

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ \text{Positives}}$$

7.4 F1-SCORE

F1 Score also known as F score or F-measure is the consonant mean of the recall and precision. Its value ranges from 0 to 1, where 0 is considered the worst, and 1 is considered the best. It can be evaluated as.

$$F1 = 2 \times \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

7.5 CONFUSION MATRIX

The Confusion Matrix gives us a complete description of the performance of the model in terms of matrix output. It assess well, particularly when working with a binary classification where we have samples that belong to two classes: TRUE or False, YES or NO.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig.5: Confusion Matrix

- True Positive (TP) is a case where the actual value was positive and the predicted value is also positive.
- False Positive (FP) is a case where the actual value was negative but the predicted value is positive.
- True Negative (TN) is a case where the actual value was negative and the predicted value is also negative.
- False Negative (FN) is a case where the actual value was positive but the predicted value is negative.

7.6 ROC AUC Score

ROC AUC Score: ROC (Receiver Operating Characteristics) AUC (Area Under Curve) is a broadly used metric for model evaluation. AUC is defined as the degree of measurement for separability, which reports on how the model can differentiate between classes. Classification problems should valuate performance with different thresholds been set. A finer model can predict 0 classes as 0 and 1 as 1, while this can be confirmed if the AUC score is high. ROC is the curve probability [10]. This ROC curve map the TPR y-axis against the FPR x-axis.

$$\text{TPR (True Positive Rate) / Recall /Sensitivity} = \frac{TP}{}$$

$$TP+FN$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$FPR = 1 - Specificity = \frac{FP}{TN+FP}$$

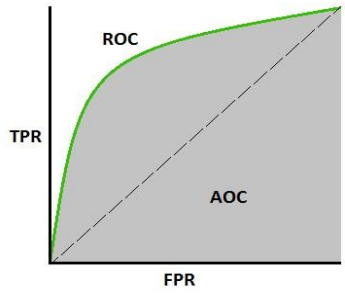

Fig.6: AUC-ROC Curve

8. RESULTS AND COMPARISION

The result of accuracy, precision and AUC score obtained by implementation of preprocessing methods and ML algorithms such as Random Forest, Decision Trees and Xgboost, K-Means, Logistic Regression and Neural Network is shown in below table.

Methods	Accuracy	Precision	AUC score
Random Forest	1.0000	1.0000	1.0000
Decision Tree	0.9522	0.9881	0.9982
Xgboost	0.9853	0.9822	0.9996
Logistic Regression	0.8179	0.7562	0.8761
Autoencoder	0.9821	0.9011	0.9889
K-means cluster	0.5001	0.7312	0.5005

Table 1: Accuracy, Precision and AUC score of ML algorithm

After comparing all results of how ML algorithm has performed based on the metric, Random Forest has the best AUC score, accuracy, and precision, as shown in table. Therefore, we have selected the Random Forest algorithm as the best model for the prediction of credit card frauds.

9. CONCLUSION

Fraud detection is a complex process, In this paper, we applied machine learning techniques to predict whether a credit card transaction is fraudulent or not. Of the various machine learning algorithms implemented in this paper we concluded that Random Forest would be the perfect fit for our model.

REFERENCES

[1] S P, Maniraj & Saini, Aditya & Ahmed, Shadab & Sarkar, Swarna. (2019). Credit Card Fraud Detection using Machine Learning and Data Science. International Journal of Engineering Research and. 08. 10.17577/IJERTV8IS090031.

[2] Sadineni, Praveen Kumar. (2020). Detection of Fraudulent Transactions in Credit Card using Machine Learning Algorithms. 659-660. 10.1109/ISMAC49090.2020.9243545.Engineering Research and. 08. 10.17577/IJERTV8IS090031

[3] Mohari, Ankit & Dowerah, Joyeeta & Das, Kashyavee & Koucher, Faiyaz & Bora, Dibya & Bora. (2021). A COMPARATIVE STUDY ON CLASSIFICATION ALGORITHMS FOR CREDIT CARD FRAUD DETECTION.

[4] Javatpoint contributors. Decision tree classification_algorithm https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[5] Wikipedia contributors (2022). Decision tree learning. In *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Decision_tree_learning

[6] Jason Brownlee. (2021). Bagging and Random Forest for imbalanced Classification. https://machinelearningmastery.com/bagging-and-random-forestfor-imbalanced-classification/

[7] Arden Dertat. (2017). Applied Deep Learning –autoencoder. https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders1c083af4d798

[8] Geeksforgeekscontributors.XGBOOST: https://www.geeksforgeeks.org/xgboost/

[9] K-MeansClustering https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[10] Sarang Narkhede. (2018). Understanding AUC – ROCCurve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5