Credit Card Fraud Detection Using Machine Learning

Mrs.A.S.Malini[1], J.M Shajitha Banu[2], M.I Sharmila Fathima[3]
1Assistant Professor, P.S.R.R College of Engineering, Sivakasi
[2,3] UG Student, P.S.R.R College of Engineering, Sivakasi

***Abstract*— Anomaly Detection is a way of recognizing suspicious occurrences of events and data items that may cause difficulties for the authorities. Security difficulties, server breakdowns, bank fraud, building structural weaknesses, clinical abnormalities, and other issues are often related with data anomalies. In today's digital money milieu, credit card fraud has become a big and major concern. These transactions are carried out with such finesse that they resemble legal transactions. As a result, the goal of this research work is to create an autonomous, highly efficient classifier for fraud detection that can detect fraudulent credit card transactions. Many fraud detection strategies and models have been proposed by researchers, including the use of various algorithms to identify fraud trends. In this paper, we look at the Isolation forest, which is a machine learning approach used to train the system using H2O.ai. In the domain of anomaly detection, the Isolation Forest was not widely utilized or researched. The version's overall performance was tested largely using commonly established metrics: accuracy and recall. Kaggle provided the test data for our study.**

***Index Terms:* Anomaly Detection, Isolation Forest, Credit Card Fraud Detection, Classification using Machine Learning.**

INTRODUCTION

With the growth of the Internet and technological advances in wireless communication technologies and network connection in recent years, the use of Internet banking or e-banking in everyday life has increased. However, it was revealed that this fraudulent behavior associated with online purchases, especially when using a credit card, occurs at a quick rate [3]. These illicit operations seek to remove illegitimate cash from an account or buy goods and services without paying their own money, causing significant harm to credit card customers and financial institutions [8]. Credit card fraud has become a big impediment to e-commerce development, having a considerable impact on the economy.

Thus, identifying fraud is critical, and the behaviors of these unlawful activities may be observed in the background to eradicate it and prevent it in the future [12].To prevent such scams, we needed an Automated Fraud Detection System capable of distinguishing between authentic and fraudulent transactions [9].

Fraud detection involves monitoring the activities of users population in order to estimate, perceive or avoid abnormal behaviour, which consist of intrusion, fraud, and defaulting.

To address this issue, machine learning may play an important role in developing detecting systems that can aid in the prevention of Credit Card fraud [10]. Machine learning refers to approaches for extracting valuable information from massive amounts of data in order to help in decision-making and prediction accuracy [7][8].

The Credit Card Fraud Detection system primarily entails differentiating between legitimate and fraudulent transactions [11].

Various difficulties encountered in developing a fraud detection system

- Incorrect data: Less than 0.5 percent of credit card transactions are fraudulent.
- Operational Efficiency: Flagging a transaction takes less than 8 seconds.
- Incorrect Flagging: Avoid bothering legitimate consumers.

Training such a fraud detection system may be accomplished in three ways:

Supervised: The supervisor directs the machine using the well-"labeled" dataset in this sort of learning. It implies they have extensive information about the data items and observations that has already been labeled with the appropriate solution. After

construction, the model is given a fresh dataset to examine the model that classifies the data.

Semi-Supervised: Semi - supervised learning involves training a system with both labeled and unlabeled datasets. It is a hybrid of supervised and unsupervised learning. Furthermore, it is used more often than supervised approaches. Unlabeled data outnumber labeled data in the dataset.

Unsupervised: Unsupervised conclusions are made from datasets that include input data but no labeled

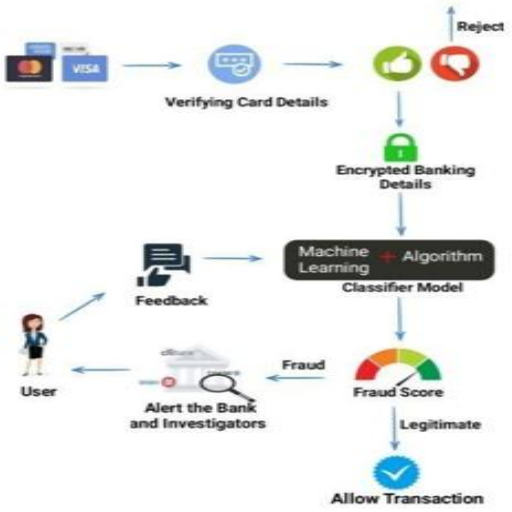

Fig - 1: Classifying Fradulent Transaction

This is the most widely employed strategy of the three. Unsupervised learning is a self- methodology in which the model expects that exceptions are less common in a dataset. It enables you to conduct more sophisticated processing jobs and is more unpredictable than other approaches. H2O framework will be used in this project.

The AI framework employs the Isolation forest, an unsupervised learning approach. Other algorithms detect abnormalities by profiling common data points, while the Isolation forest is an ensemble technique. It generates a tree-like structure that aids in decision- making. These irregularities may be detected at the tree's root and subsequently analyzed [13].

RELATED WORK

The authors of article [1] present an anomaly detection approach based on an artificial neural network and decision tree. This approach is divided into two stages. First, a decision tree is utilized to generate a fresh dataset, which is then sent into a Multilayer neural network to categorize the data. This two-level system has a low false detection rate.

The authors [2] conduct a thorough examination of several machine learning techniques including ANN. They discovered that Artificial Neural Networks provide more exact results than K- Nearest Neighbor (KNN), Logistic Regression, Support Vector Machine (SVM), and Decision Tree[10]. Another article [3] claims that the Random Forest approach, together with Logistic Regression and SVM, gives the most trustworthy effects.

The decision tree strategy outperforms the SVM approach in answering the issue, according to study [4]. Furthermore, when the size of the datasets increases, the accuracy of the SVM-based system outperforms the accuracy of the decision tree-based system. However, the quantity of fraud detected by SVM models is significantly smaller than the total amount of fraud detected by decision tree techniques.

The method utilized in study [5] proposes a unique technique to detecting fraudulent transactions by using various anomaly detection algorithms. Outlier detection and the KNN methods were used in paper [6] to maximize the results in fraud detection situations. The major goal was to increase the rate of fraud detection and eliminate false alarms.

METHODOLOGY

The approach provided in this research, termed Isolation Forest, employs one of the most recent machine-learning algorithms to detect anomalous activity.

A.Forest of Isolation

Isolation forest is an unsupervised ensemble that is built on the notion of "separate-away" anomalies in isolation[13]. There is no point-based distance computation or profiling of regular instances. Instead, the Isolation forest constructs an ensemble of decision trees, with the goal of isolating anomalies via partitions. Here, the decision-making ensemble A tree is formed for a given data collection, and the route length for each data point is determined, and the data points with the smallest average path length are deemed anomalies.

B.H2O.ai

H2O is compatible with the most popular supervised and unsupervised machine learning methods. It is a

completely open-source machine learning platform with linear scalability, ultra- high speed, in-memory, and predictive analytics. It incorporates gradient boosted machines, generalized linear models, deep learning, and the ability to install and tune machine learning models without requiring expertise.

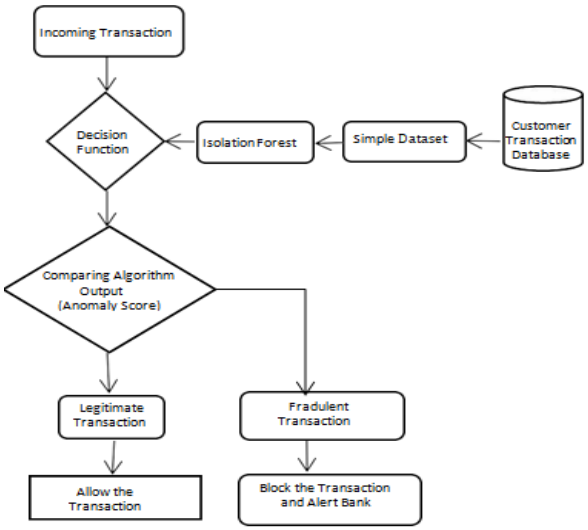

Fig-2: Flow diagram of classification model

A. Dataset Description

The dataset utilized in our study is the only publicly available data collection suitable for developing a fraud detection system.The dataset contains around 500 fraudulent transactions and 284300 reported valid transactions, resulting in a significantly skewed dataset. The dataset comprises variables in numerical form produced by Principal component analysis (PCA) transformation (V1, V2, up to V28), which provide information on different features of credit card transactions. The only features in the dataset that are not modified using PCA are 'Time' and 'Amount.' In addition, the 'Class".

B .Anomaly Detection

Anomaly Detection is a way of recognizing suspicious occurrences of events and data items that may cause difficulties for the authorities. Security difficulties, server breakdowns, bank fraud, building structural weaknesses, clinical abnormalities, and other issues are often related with data anomalies.

It consists of two levels of training and testing: Building an Isolation forest during the training stage, then passing each data point through each tree to compute the average number of edges necessary to reach an exterior node during the testing stage.

We begin by randomly picking a characteristic to create many decision trees. Then, in an unexpected manner, we determine a split value from the maximum and lowest values of that randomly chosen property. Each ending node of the tree should ideally include one observation from the data collection, which isolates the sample. We assume that if one discovery in our data set is similar to another, more random splits would be required to exactly isolate the finding, as opposed to isolating an outlier.

```
import h2o
from sklearn.metrics import roc_curve, precision_recall_curve, auc
import pandas as pd
h2o.init(strict_version_check = False)
file = h2o.import_file("creditcard.csv")
seed = 12345
print("Enter the Number of decision trees want to create:")
ntrees=int(input())
isoforest = h2o.estimators.H2OIsolationForestEstimator(
    ntrees=ntrees, seed=seed)
col_names = ['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6',
       'V6' , 'V7', 'V8', 'V9', 'V10', 'V11', 'V12',
       'V13', 'V14','V15','V16','V17','V18','V19','V20',
       'V21','V22','V23','V24','V25','V26','V27','V28','Amount']
isoforest.train(x=col_names, training_frame=file)
predictions = isoforest.predict(file)
```

Fig-3:The initial code for importing the dataset and starting the H2O instance and giving initial predictions.

We compute the route length for each observation since we generated many decision trees that add together to form an isolation forest. The amount of splitting required to differentiate the observation is equal to the length of the route from the root node to the leaf node. This route length is then averaged throughout a forest of decision trees, which acts as a scale for the anomaly and is used to calculate the final anomaly score. The shorter the journey, the more probable it is to be abnormal.

PREDICT	MEAN_LENGTH
0.0559194	6.778
0.0420655	6.833
0.175063	6.305
0.07733	6.693
0.0546599	6.783
0.0357683	6.858
0.0458438	6.818
0.186146	6.261
0.0649874	6.742
0.0420655	6.833

Fig - 4: Initial normalized predicted length and the mean length for the multiple decision trees created.

The H2O frame encompassing the results of the predictions: we forecast presenting a normalized incongruity score.

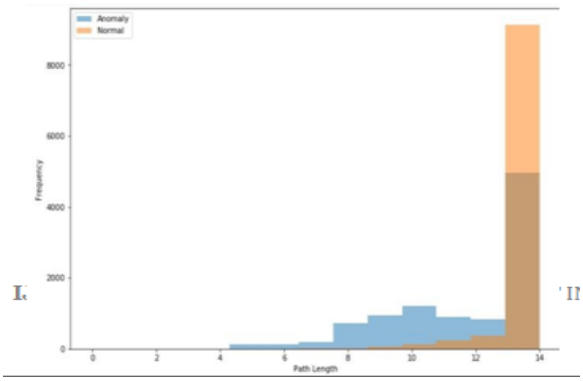

Fig-5:Path length of 15000 transactions of the training dataset.

We are working unsupervised manner! We need a threshold. If we had an estimation of the raw number of outliers in our dataset, we can find the score's equivalent quantile value and use it for our predictions as a threshold.

PROBS	PREDICT QUANTITIES	MEAN_LENGTHQUANTITIES
0.95	0.164736	6.982

Fig-6: Probability of predicting Qualities and Length Qualities for the dataset.

The analog generated quantile price score can be perceived and used it as a limit value for the forecasts made by our generated H2O frame. We use the edge to categorize the anomalous segment in the dataset.

PREDICT	MEAN_LENGTH	PREDICTED_CLASS	CLASS
0.0559194	6.778	0	0
0.0420655	6.833	0	0
0.175063	6.305	1	0
0.07733	6.693	0	0
0.0546599	6.783	0	0
0.0357683	6.858	0	0
0.0458438	6.818	0	0
0.186146	6.261	1	0
0.0649874	6.742	0	0
0.0420655	6.833	0	0

Fig-7: Results of the predicted class with the actual quality of the dataset.

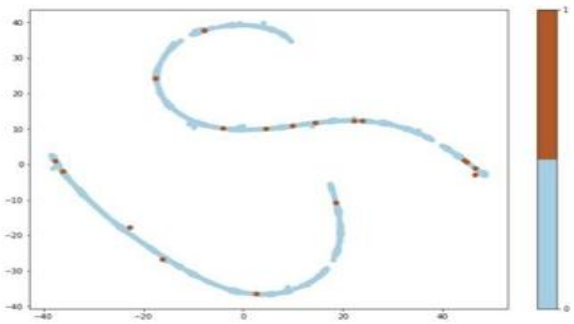

Fig-8: 2D view of the Predicted class by fraud detection model (Genuine: 1000, Fraud: 20).*1. Evaluation*

Because the isolation forest is an unsupervised approach, we need classification metrics that are not reliant on the prediction threshold and provide an accurate score. Area under the Precision-Recall Curve (AUCPR) and Area under the Receiver Operating Characteristic Curve (AUCRC) are two such measurements (AUC).

AUC is a metric that measures how well a binary classifier differentiates between true and false positives. The ideal AUC score is 1, while the wild approximation is 0.5. AUCPR is the precision-recall trade-off of a binary classification using multiple thresholds of the continuous prediction ranking. The maximum AUCPR score is 1; the baseline score is positive class relative count. For an excessively imbalanced dataset, AUCPR is favored over AUC because it is particularly sensitive to true positives, false negatives, and false positives while not caring about True negative.

On average, the binary compound isolation forest implementation outperforms the scikit-learn implementation. The capacity to re-scale too many nodes and function flawlessly with Apache Spark is a big benefit of the binary compound. This allows you to process very big datasets, which is useful in the context of transactional data.

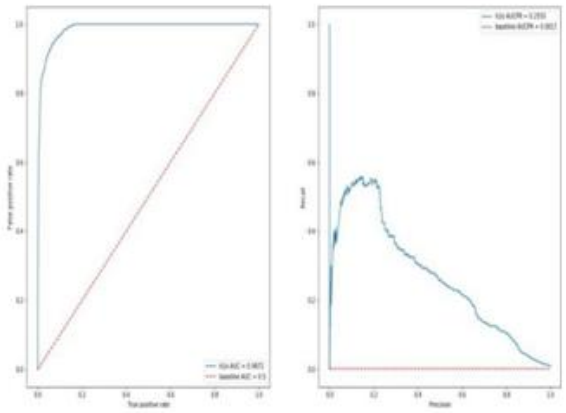

Fig-9: Shows the baseline AUCPR and h20 AUCPR for true positive vs. real negative and precision vs. recall for the dataset.

The prognostic accuracy for this suggested classification model employing the isolation forest to identify fraud in credit card transactions was found to be 98.72 percent by AUCPR, which is very

beneficial, and the fraud detection error was minimized.

CONCLUSION AND FUTURE SCOPE

In this project, we describe a strategy that has an astonishing capacity to detect anomalies from simple inliers by generating numerous decision trees for each data point. We utilize the Area Under Precision-Recall curve (AUC) to evaluate our methods since it produces better results than the Area Under ROC curve. Finally, we show that the efficiency of our method in a fraud detection model is 98.72 percent, indicating that it is substantially superior than existing fraud detection strategies.

The main constraint of the fraud detection system is the lack of a balanced dataset for training purposes, as well as the scarcity of the dataset. The study result will be more efficient and qualitative if financial institutions make accessible the crucial data set of different fraudulent actions.

REFERENCES

[1] R. M. Jamail Esmaily, "Intrusion detection system based on Multilayer perceptron neural networks and decision tree," in International Conference on Information and Knowledge Technology, 2015.

[2] Jain, Y. & Tiwari, N. & Dubey, S. & Jain, Sarika. (2019). "A comparative analysis of various credit card fraud detection techniques" in International Journal of Recent Technology and Engineering. 7. 402-407.

[3] S. J. K. T. J. C. W. Siddhartha Bhattacharya, "Data Mining for credit card fraud: A comparative study," Elsevier, vol. 50, no. 3, pp. 602613, 2011.

[4] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines," Int. Multi-conference Eng. Comput. Sci., vol. I, pp. 442–447, 2011.

[5] S P Maniraj , Aditya Saini , Shadab Ahmed , Swarna Deep Sarkar, 2019, Credit Card Fraud Detection using Machine Learning and Data Science, International Journal of Engineering Research & Technology (IJERT) Volume 08, Issue 09 (September 2019),

[6] N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 255-258.

[7] Ishu Trivedi, Monika, Mrigya, Mridushi, "Credit Card Fraud Detection", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016.

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak "Plastic Card Fraud Detection using Peer Group Analysis" Springer, Issue 2008.

[9] Ekrem Duman, M. Hamdi Ozcelik "Detecting credit card fraud by genetic algorithm and scatter search". Elsevier, Expert Systems with Applications, (2011). 38; (13057– 13063).

[10] S. Benson Edwin Raj, A. Annie Portia "Analysis on Credit Card Fraud Detection Methods", IEEE-International Conference on Computer, Communication and Electrical Technology, (2011), pg.152-156.

[11] Khyati Chaudhary, Yadav, Mallick, "Review of fraud detection techniques: credit card", International Journal of Computer Applications (0975- 8887), Volume 45-No 1, May 2012.

[12] Quah, J. T. S., and Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4).

[13] Wen-Fang YU and Na Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum", International Joint Conference on Artificial Intelligence 2009.

[14] F. Z. El Hlouli, J. Riffi, M. A. Mahraz, A. El Yahyaouy, and H. Tairi, ''Credit card fraud detection based on multilayer perceptron and extreme learning machine architectures,'' in Proc. Int. Conf. Intell. Syst. Comput.Vis. (ISCV), Jun. 2020, pp. 1–5

[15] S. Khatri, A. Arora, and A. P. Agrawal, ''Supervised machine learning algorithms for credit card fraud detection: A comparison,'' in Proc.10th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence), Jan. 2020,pp. 680–683.

[16] Pandas. Accessed: Sep. 27, 2021. [Online].

Available: https://pandas.pydata.org/

[17] T. T. Wong and P. Y. Yeh, ‘‘Reliable accuracy estimates from k-fold cross validation,’’ IEEE Trans. Knowl. Data Eng., vol. 32, no. 8, pp. 1586–1594, Apr. 2019.

[18] “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[19] “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[20] Y. Sahin, S. Bulkan, E. Duman, "A cost-sensitive decision tree approach for fraud detection", Expert Systems with Applications, vol. 40, no. 15, pp. 5916-5923, 2013.

[21] Iwasokun GB, Omomule TG, Akinyede RO. Encryption and tokenization-based system for credit card information security. Int J Cyber Sec Digital Forensics. 2018;7(3):283–93.

[22] Maniraj SP, Saini A, Ahmed S, Sarkar D. Credit card fraud detection using machine learning and data science. Int J Eng Res 2019; 8(09).

[23] Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Proc Comput Sci. 2019;165:631–41.

[24] Thennakoon, Anuruddha, et al. Real-time credit card fraud detection using machine learning. In: 2019 9th international conference on cloud computing, data science & engineering (Confluence). IEEE; 2019.

[25] Campus K. Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math. 2018;118(20):825–38.