International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-8, Issue-3, May 2020
https://doi.org/10.21276/ijircst.2020.8.3.5
www.ijircst.org

Anomaly Detection in Credit Card Transactions using Machine Learning

Meenu, Swati Gupta, Sanjay Patel, Surender Kumar, Goldi Chauhan

ABSTRACT- Anomaly Detection is a method of identifying the suspicious occurrence of events and data items that could create problems for the concerned authorities. Data anomalies are usually associated with issues such as security issues, server crashes, bank fraud, building structural flaws, clinical defects, and many more. Credit card fraud has now become a massive and significant problem in today's climate of digital money. These transactions carried out with such elegance as to be similar to the legitimate one. So, this research paper aims to develop an automatic, highly efficient classifier for fraud detection that can identify fraudulent transactions on credit cards. Researchers have suggested many fraud detection methods and models, the use of different algorithms to identify fraud patterns. In this study, we review the Isolation forest, which is a machine learning technique to train the system with the help of H2O.ai. The Isolation Forest was not so much used and explored in the area of anomaly detection. The overall performance of the version evaluated primarily based on widely-accepted metrics: precision and recall. The test data used in our research come from Kaggle.

KEYWORDS- Anomaly Detection, Isolation Forest, Credit Card Fraud Detection, Classification using Machine Learning.

I. INTRODUCTION

With the growth of the Internet and technological advancements in wireless communication technologies and Network Connectivity in recent years, the use of Internet banking or e-banking in day to day life has increased. But due to this fraudulent activity associated with online purchases, mainly when using a credit card, it observed that they occur at a rapid pace [3]. These illegal activities aim to withdraw illegitimate funds from an account or purchase goods and services without paying their own money, which causes severe damage to the credit card holders and banking organizations [8]. Credit card fraud has become a significant obstacle to e-commerce growth, which has a drastic effect on the economy.

Thus, identification of fraud is crucial and vital, and the actions of these illegal activities may observe in the background to eliminate it and avoid it against repeated incidents [12].

To stop such frauds, we required an Automated Fraud detection system that will be capable of classifying fraudulent transactions from genuine ones [9]. To solve this problem machine learning can play a significant role in building such type of detection system that can help to prevent such Credit Card fraud [10]. Machine learning consists of methods to derive useful information from vast volumes of data to aid in the decision-making process and predictive accuracy [7][8]. The Credit Card Fraud Detection system mainly involves distinguishing fraudulent transactions from the authentic ones [11].

Various challenges while creating fraud detection system

- Unbalance data: Less than 0.5 % of credit card transactions are a fraud.
- Operational Efficiency: Less than 8 sec to flag a transaction.
- Incorrect Flagging: Avoid harassing real customers.

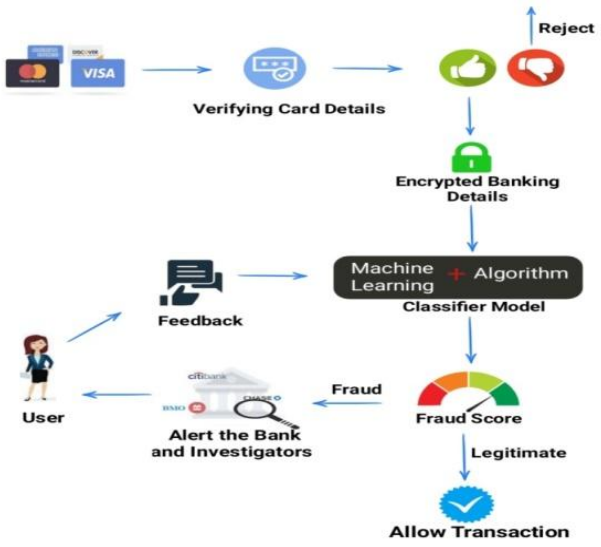

Fig 1: Classifying fraudulent transactions

Manuscript received May 6, 2020

Meenu, Assistant Professor, Department of Computer Science, Amity University, Gurugram, Haryana, India (email:mvijarania@ggn.amity.edu).

Swati Gupta, Assistant Professor, Department of Computer Science, Amity University, Gurugram, Haryana, India

Sanjay Patel, Department of Computer Science, Amity University, Gurugram, Haryana, India.

Surender Kumar, Department of Computer Science, Amity University, Gurugram, Haryana, India.

Goldi Chauhan, Department of Computer Science, Amity University, Gurugram, Haryana, India.

To train such a fraud detection system can carry out in three ways:

• Supervised: In this type of learning, the supervisor instructs the machine using the well-"labeled" dataset. It means they have detailed information about the data items and observations, already tagged with the correct solution. After building, the model provided with a new dataset to analyze the model that classifies the data.

• Semi-Supervised: In semi-supervised learning, to train a machine done by using both labeled and unlabeled datasets. It is a combination of both supervised and unsupervised learning. Moreover, it employed more commonly than supervised techniques. The dataset contains more unlabeled data than labeled one.

• Unsupervised: In unsupervised inferences from the datasets that consist of input data without labeled responses. Among all three strategies, this is the most commonly used method. Unsupervised learning is a self-methodology in which the model assumes exceptions occur less frequently in a dataset. It allows you to perform more complex processing tasks and can be more unpredictable compared with other methods.

In this project, we will use framework H20.ai framework implement the Isolation forest, which is an unsupervised learning method. Other algorithms identify anomalies with the help of profiling usual data points, but the Isolation forest is an ensemble method. It creates a tree-like structure that helps to make decisions. These anomalies can detect near the root of the tree and then can further analyze [13].

II. RELATED WORK

In paper [1], the authors propose an anomaly detection methodology using an artificial neural network and decision tree. This methodology consists of two-level, firstly a decision tree is used to provide a new dataset, which is passed into the Multilayer neural network to classify the data. This two-level system results in a meager false detection rate. A detailed study of various machine learning algorithms and ANN is done by the authors [2]. In which they found Artificial Neural Network gives more precise results compared to K-Nearest Neighbour (KNN), Logistic regression, Support Vector Machine (SVM) and Decision tree[10]. Another paper [3] suggests that the Random Forest methodology provides the most reliable effects, accompanied by Logistic Regression and SVM.

In paper [4], the outcome is that the decision tree approach performs better than the SVM approach in solving the problem. Moreover, as the size of the datasets grows, the degree of accuracy of SVM-based system surpasses the accuracy of decision tree-based system. However, the amount of fraud observed by SVM models is still far less than the sum of frauds identified by decision tree approaches. The system used in paper [5] presents a novel approach to detect fraudulent transactions uses multiple anomaly detection algorithms to detect fraud. Paper [6] applied Outlier detection and the KNN methodology to optimize the result in fraud detection problems. The primary aim was to improve the fraud detection rate and reduce false alarm.

III. METHODOLOGY

The method presented in this paper uses one of the latest machine-learning algorithms to identify anomaly behavior called Isolation Forest.

A. Isolation Forest

Isolation forest is an unsupervised ensemble, and it based on the concept of isolation “separate-away” anomalies[13]. No point-based distance calculation and no profiling of regular instances are done. Instead, the Isolation forest builds an ensemble of decision trees; the principle behind this technique is to isolate anomalies through partitions. Here ensemble of decision Trees is created for a given data set, and path length is calculated for each data, and the data points which have the shortest average path length are considered as anomalies.

B. H2O.ai

H2O supports the most widely used supervised and unsupervised machine learning algorithms. It is a fully open-source, ultra-high performance, in-memory, and predictive analytics machine learning platform with linear scalability. It includes gradient boosted machines, generalized linear models, deep learning, and allowing without needing expertise in deploying or tuning machine learning models.

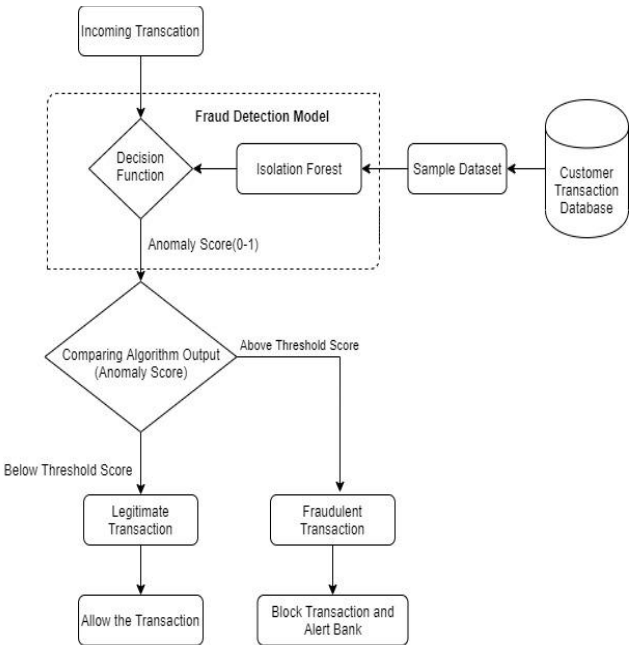

Fig 2: Flow diagram of classification model

1. Dataset Description

Dataset used in our research is the only accessible data collection that is appropriate for building a fraud detection system. The dataset includes approximately 500 fraudulent transactions, and 284300 reported legitimate transactions, which makes it highly imbalanced. Here the dataset contains variables in numerical form that obtained after applying Principal component analysis (PCA) transformation (V1, V2, up to V28), which contains information on various properties of credit card transactions. 'Time' and 'Amount' are the only features in the dataset that is not transformed with PCA. And the 'Class'

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-8, Issue-3, May 2020
https://doi.org/10.21276/ijircst.2020.8.3.5
www.ijircst.org

function only contains only binary value 0 or 1, 1 in case of fraud transaction and 0 in a genuine transaction.

2. Anomaly Detection

It has two stages training and testing: Training stages involve building Isolation forest, and testing stages involve passing each data point through each tree to calculate the average number of edges required to reach an external node.

We begin by building several decision trees by selecting an attribute randomly. Then we choose a split value from the maximum and minimum values of that randomly selected attribute in an unpredictable way. Ideally, each terminating node of the tree contains one observation from the data set, which isolates the sample. We presume that if one finding in our data set is identical to another, it would require further random splits to isolate the finding precisely, 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.

As we created multiple decision trees, which sum as an isolation forest, for each observation, we calculate the path length. The amount of splitting needed to distinguish the observation is equivalent to the path length from the root node to the leaf node. Then this path length is averaged over a forest of a decision tree, which serves as a scale for the anomaly and further use for determining the final anomaly score. Less the path length, the more likely it is to be anomalous.

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 containing the results of the predictions: we forecast presenting a normalized incongruity score.

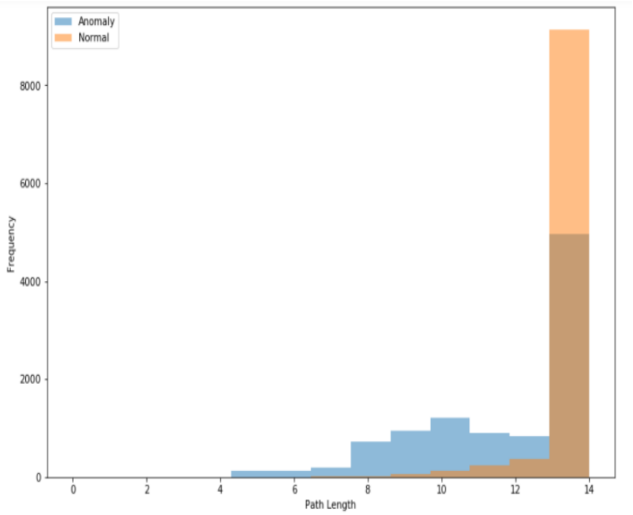

Fig 5: Path length of 15000 transactions of the training dataset.

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	predictQuantiles	mean_lengthQuantiles
0.95	0.164736	6.982

Fig 6: Probability of predicting Quantiles and mean length Quantiles for the dataset.

The corresponding generated quantile price score can be observed and used it as a limit value for the predictions made by our generated h20 frame. We use the edge to classify the abnormal 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.

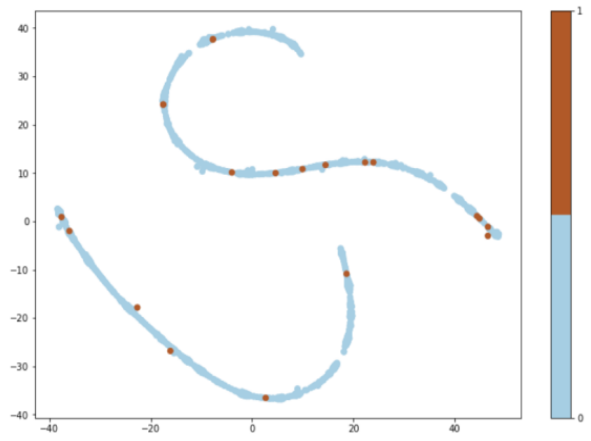

Fig 8: 2D view of the Predicted class by fraud detection model (Genuine: 1000, Fraud: 20).

3. Evaluation

Since the isolation forest is an unsupervised technique, we need classification metrics that should not be dependent on the prediction threshold and give an exact value of scoring. For this, two such metrics is Area under the Precision-Recall Curve (AUCPR) and Area under the Receiver Operating Characteristic Curve (AUC).

AUC is a statistic measuring how often a type of binary classification distinguishes between real and false positives. The optimal AUC score is 1; wild approximation is the baseline value of 0.5. AUCPR is a binary classification's precision-recall trade-off utilizing various thresholds of the continuous prediction ranking. The highest score for AUCPR is 1; the baseline score is positive class relative count. AUCPR is preferred over AUC for an extremely unbalanced dataset because the AUCPR is very sensitive to true positives, false negatives, and false positives while not worrying about True negative.

We can see that the binary compound isolation forest implementation on the average scores equal to the scikit-learn implementation. The significant advantage of the binary compound is that the ability to rescale too many nodes and work seamlessly with Apache Spark. This enables you to method extraordinarily large datasets, which could be crucial within the transactional information setting.

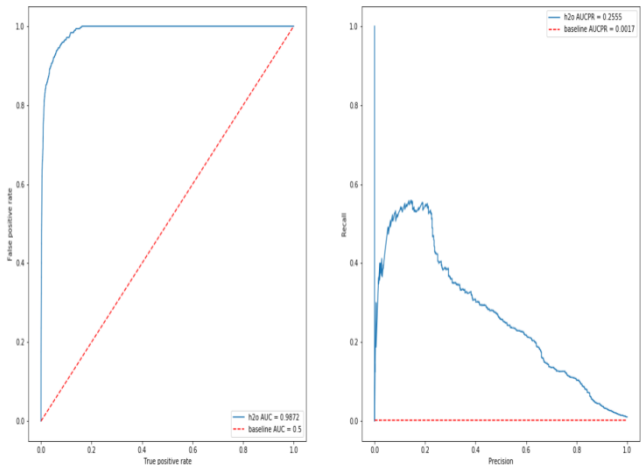

Fig 9: Shows the baseline AUCPR and h20 AUCPR for true positive vs. real negative and precision vs. recall for the dataset.

The Predictive accuracy for this proposed classification model using the Isolation forest to detect fraud in credit card transactions observed to be 98.72 % by AUCPR, which is significantly useful, and the fraud detection error reduced.

IV. CONCLUSION AND FUTURE SCOPE

We present a technique in this project that exhibits an amazing ability to distinguish anomalies and mere inliers by creating several decision trees for every data point. For better evaluation of our methodology, we use Area Under Precision-Recall curve (AUC), which shows better results than the Area Under ROC curve. Lastly, we demonstrate the efficiency of our approach in a fraud detection model observed to be 98.72%, which indicates a significantly better approach than other fraud detection techniques.

The only limitation to the fraud detection system is the unavailability of the balanced dataset for training purposes and the shortage of the dataset. If the financial institutions make available the critical data set of various fraudulent activities, the research outcome will be more efficient and qualitative.

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. Multiconference 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), 1721-1732.

[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.

