Credit Card Fraud Detection Using Machine Learning.

Mr. Bhavsar Akshay
UG Student,
Department of Computer Engineering
SNDCOE & Rc Yeola

Mr. Shaikh Rizwan Nisar
UG Student,
Department of Computer Engineering
SNDCOE & Rc Yeola

Monica Subhash Jadhav
UG Student,
Department of Computer Engineering
SNDCOE & Rc Yeola

Prof. Chandgude A. S
Assistant Professor,
Department of Computer Engineering
SNDCOE & Rc Yeola

ABSTRACT
It is important that companies that produce credit cards are able to detect fraudulent credit card transactions customers need to pay for things they don't need to buy. These issues can be addressed through data science and its importance, along with machines, cannot be emphasized enough. The goal is to show an artificial dataset using machine learning during credit card fraud. The problem of detecting falsification of credit cards involves modeling, ex-credit card transactions, data turned out to be in the position of fraud. This model is then used to determine whether the new operating system is fraudulent or not. Our goal is to find 100% fraudulent transactions here, and at the same time, minimize false rating scams. Fraud detection, credit cards-a typical example of a presentation. In this process, we focus, analyze, and pre-process data sets, as well as placing multiple anomaly detection algorithms as an inconvenient factor Isolation algorithm for forests of ATP-transformed credit card transaction data.

Keywords
Credit card fraud, applications of machine learning, data science, isolation forest algorithm, local outlier factor, automated fraud detection.

1. INTRODUCTION
"Fraud" in the operating room using a credit card is, unauthorized and criminal, using your account, no one but the owner of this account. The necessary proactive measures that can be taken to put an end to this violence and conduct so-called deceptive practices can be studied in order to minimize it, as well as in order to protect ourselves from similar incidents in the future. In other words, fraudulent use of a credit card can be defined as a person who uses someone else's credit card, for personal reasons, and the owner and issuing authorities know that this card is being accessed. Fraud detection involves monitoring the activities of user groups to assess whether it is being used, or to avoid abusive behavior such as fraud, trespassing, and negligence.

This is a very relevant issue that requires attention from the community, education and data science, and the solution to this problem can be automated. This problem is particularly challenging as a learning perspective, which is characterized by a number of factors like: Class imbalance. The total number of valid transactions is greater than the success rate of fraud and forgery. In addition, the operational model or statistical characteristics change over time. It's not just the problems that arise from implementing systems during real-time fraud detection. A real-world example of a mass flow for payment needs to be quickly scanned by automated funds to determine which of the applications to accept. Machine learning algorithms to use, analyze, all transaction confirmations and report suspicious cases. These reports are verified by specialists who contact the cardholder to check whether the transaction is original or fake. You provide feedback for an automated system that uses training and algorithm updates to ultimately improve the performance of the fraud detection system over time.

Fig No 1 Fraud Detection System

Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies. These frauds are classified as:
- Credit Card Frauds: Online and Offline
- Card Theft
 Account Bankruptcy  
 Device Intrusion  
 Application Fraud  
 Counterfeit Card  
 Telecommunication Fraud

Some of the currently used approaches to detection of such fraud are:

- Artificial Neural Network
- Fuzzy Logic
- Genetic Algorithm
- Logistic Regression
- Decision tree
- Support Vector Machines
- Bayesian Networks
- Hidden Markov Model
- K-Nearest Neighbour.

2. LITERATURE REVIEW

Fraud, which is illegal, or criminal deception aimed at obtaining financial or personal benefits. This is a deliberate act that is contrary to the law, requirement and, basically, products or policies in order to achieve non-financial benefits. Many texts related to the detection of anomalies / fraud in this area have already been published and are available for public use. The results of a broad survey conducted by Clifton Foix and his colleagues showed that the methods used in this field include software for data mining, automated fraud detection, and adversarial testing. In another article by Suman, GIUS&T research scientists focused on HCE, as the development of supervised and unsupervised learning methods for detecting credit card fraud. Although these methods and algorithms have been unexpectedly successful in a particular field, they have not been able to provide reliable and consistent solutions for fraud detection. A similar study in the topic is presented by Wen-Fan Yu and Wang, and they are used to mine waste algorithms in order to accurately predict fraudulent transactions in one of the payment data emulation experiments to determine the amount of a commercial bank. Outlier mining is the field of data mining, which, in principle, allows you to use cash in the Internet sphere as well. The point is that the detected objects are independent of the underlying system, i.e. operations that are wrong. They take attributes, customer behavior, and the main values of this attribute, and calculate the difference between the control value of the object's attribute and the pre-set value. Unconventional methods like a hybrid data mining/complex network algorithm, classification that allows you to detect illegal copies of the actual map dataset in a service that is based on a network, reconstruction algorithm allows you to create representations, even though selection from the reference group has been shown to be effective, usually in an online operations environment. His attempts to move to a whole new level. Attempts to improve mutual information and feedback in the event of a fraudulent transaction. In this case, any fraudulent transaction will be approved and the system will be alerted, and feedback should be provided to deny the current operation process. An artificial genetic algorithm is a great way to shed more light on this domain name, get scanned, on the other hand. It turned out that, correctly, it is necessary to find fraudulent transactions and minimize the number of false positives. Even if it is along with the classification case with the variable misclassification costs.

3. PROPOSED SYSTEM

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers. The basic rough architecture diagram can be represented with the following figure:

![Fig. No 2 System Architecture](image)

When looked at in detail on a larger scale along with real life elements, the full architecture diagram can be represented as follows:

![Fig. No 3 Architecture Diagram](image)

First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets.

Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one. We plot different graphs to check for inconsistencies in the dataset and to visually comprehend it.
Fig No 4 Count of Fraudulent vs Non Fraudulent Transactions
This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.

Fig No 5 Distribution of Time Feature
This graph shows the times at which transactions were done within two days. It can be seen that the least number of transactions were made during night time and highest during the days.

Fig No 6 Distribution of Momentary value Feature.
This graph represents the amount that was transacted. A majority of transactions are relatively small and only a handful of them come close to the maximum transacted amount. After checking this dataset, we plot a histogram for every column. This is done to get a graphical representation of the dataset which can be used to verify that there are no missing any values in the dataset. This is done to ensure that we don’t require any missing value imputation and the machine learning algorithms can process the dataset smoothly.

Fig No 7 Machine Learning Algorithm
After this analysis, we plot a heatmap to get a colored representation of the data and to study the correlation betweenout predicting variables and the class variable. This heatmap isshown below:

Fig No 8 Heatmap of Correlation
The dataset is now formatted and processed. The time and amount column are standardized and the Class column is removed to ensure fairness of evaluation. The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms
work together: This data is fit into a model and the following outlier detection modules are applied on it:

- Local Outlier Factor
- Isolation Forest Algorithm

These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection.

This free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provides a lot of simple and efficient tools which can be used for data analysis and machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries. We've used Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files.

Detailed explanations about the modules with pseudo codes for their algorithms and output graphs are given as follows.

3.1 Local Outlier Factor

It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours. More precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data. The pseudo code for this algorithm is written as:

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]

# Generate some regular novel observations
X = 0.3 * rng.randn(10, 2)
X_test = np.r_[X + 1.5, X - 1.5]

# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(behaviour='new', max_samples=100,
                      random_state=rng, contamination='auto')
clf.fit(X_train)
Y_pred_train = clf.predict(X_train)
Y_pred_test = clf.predict(X_test)
Y_pred_outliers = clf.predict(X_outliers)

# Plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50),
                     np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

On plotting the results of Local Outlier Factor algorithm, we get the following figure:

![Local Outlier Factor](image)

By comparing the local values of a sample to that of its neighbours, one can identify samples that are substantially lower than their neighbours. These values are quite anomalous and they are considered as outliers.

As the dataset is very large, we used only a fraction of it in outtests to reduce processing times.

The final result with the complete dataset processed is also determined and is given in the results section of this paper.

3.2 Isolation Forest Algorithm

The Isolation Forest ‘isolates’ observations by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature.

Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent to the path length root node to terminating node. The average of this path length gives a measure of normality and the decision function which we use.

The pseudocode for this algorithm can be written as:

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[2001]

# Plot the level sets of the decision function
xx, yy = np.meshgrid(np.linspace(-5, 5, 50),
                     np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

On plotting the results of Isolation Forest algorithm, we get the following figure:

![Isolation Forest](image)
Partitioning them randomly produces shorter paths for anomalies. When a forest of random trees mutually produces shorter path lengths for specific samples, they are extremely likely to be anomalies.

Once the anomalies are detected, the system can be used to report them to the concerned authorities. For testing purposes, we are comparing the outputs of these algorithms to determine their accuracy and precision.

4. IMPLEMENTATION

This concept is difficult to implement in real life, so it requires the cooperation of banks that are willing to share the details of their competition in the market, as well as for legal reasons, and to protect the data of their users. So we were looking for a kind of reference in the paper, behind, followed by similar methods and results. As already noted, there is a help sheet:

"These technologies are applied in order to fill out the application, information about it was given by a German bank in 2006. For the bank, for confidentiality reasons, below is only a summary of the results. For the application of this technology, the level is 1, and the list includes a small number of cases, but the probability of being an impostor. All the people listed on this list were slightly connected to avoid any risks, due to their high risk profile. The situation is complicated, and second on the list. Level 2 is still limited accordingly so that it can be tested on a case-by-case basis. Credit lines, as well as a collection of officials found that half of the cases on this list should be considered as suspected of fraudulent activity. Discover the latest on the list, and the greatest work experience of the same weight. Not a third of the suspects.

To maximize time, efficiency, and overhead, one of the options for entering a question is a new item, this item can be in the first five digits, phone number, email address, and password, for example, new and new questions, one could apply a level 2 list and a level 3-b list."

5. RESULTS & DISCUSSION

The code prints out the number of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end and both the results are printed. These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 means it was determined as a fraud transaction. This result matched against the class values to check for false positives. Results when 10% of the dataset is used:

<table>
<thead>
<tr>
<th>Isolation Forest</th>
<th>Number of Errors: 73</th>
<th>Accuracy Score: 0.297590711000316</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>recall</td>
<td>f1-score</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>accuracy</td>
<td>macro avg</td>
<td>0.64</td>
</tr>
<tr>
<td>weighted avg</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Local Outlier Factor</th>
<th>Number of Errors: 97</th>
<th>Accuracy Score: 0.9965942207685425</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>recall</td>
<td>f1-score</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>accuracy</td>
<td>macro avg</td>
<td>0.51</td>
</tr>
<tr>
<td>weighted avg</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Results with the complete dataset is used:

<table>
<thead>
<tr>
<th>Isolation Forest</th>
<th>Number of Errors: 659</th>
<th>Accuracy Score: 0.9976061523768727</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>recall</td>
<td>f1-score</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>accuracy</td>
<td>macro avg</td>
<td>0.66</td>
</tr>
<tr>
<td>weighted avg</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Local Outlier Factor</th>
<th>Number of Errors: 935</th>
<th>Accuracy Score: 0.9967176507129008</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>recall</td>
<td>f1-score</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>accuracy</td>
<td>macro avg</td>
<td>0.52</td>
</tr>
<tr>
<td>weighted avg</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
CONCLUSION

Credit card fraud is undoubtedly an act of criminal dishonesty. This article explains the most common methods of fraud, as well as methods of detecting them, and presents the most recent findings in this area. This paper also explains in detail how it can be that the education applied in order to obtain a better result detected falsifications, along with the algorithm, pseudocode and explanation of its implementation, as well as experimental results. While the algorithm allows you to get more than 99.6% accuracy, its accuracy remains up to 28%, about a tenth of the data that needs to be taken into account. But when using the entire data set that the algorithm outputs, the accuracy increases by 33%. This high percentage is exactly to be expected, as there is a big difference between a very reliable and very real operation. Because the entire data set consists of just two days of operation notes, and then only a small portion of the data available on this project was developed for use on a commercial scale. Being based on machine learning algorithms, the program will not only increase its efficiency over time, as more information is added to it.

ACKNOWLEDGMENTS

A very firstly we gladly thanks to my project guide Prof A.S. Chandgude, for his valuable guidance for implementation of proposed system. We will forever remain a thankful for their excellent as well as polite guidance for preparation of this report. Also we would sincerely like to thank to HOD A.S. Chandgude and other staff for their helpful coordination and support in project work.

REFERENCES


[3] “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence

