Credit Card Fraud Detection using Machine Learning Algorithms

Kanika Singhal, Ayush Agrawal, Ayush Gupta and Harshit Gupta

July 13, 2021

Credit Card Fraud Detection using Machine Learning Algorithms

(**Kanika Singhal** , Assistant Professior in Department of **Information Technology, Galgotias College of Engineering and Technology**)

Authors Name : Ayush Agrawal
Department of **Information Technology**
Galgotias College of Engineering and Technology
Greater Noida , India

Email id : ayushsamrat1998@gmail.com

Authors Name : Ayush Gupta
Department of **Information Technology**
Galgotias College of Engineering and Technology
Greater Noida , India

Email id : ayushgupta6598@gmail.com

Authors Name : Harshit Gupta
Department of **Information Technology**
Galgotias College of Engineering and Technology
Greater Noida , India

Email id: harshitgupta584@gmail.com

***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.**

I. INTRODUCTION

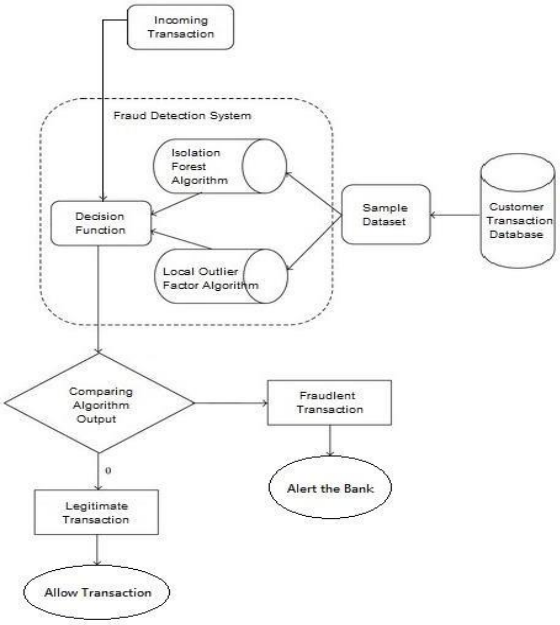
"Fraud" in operating with a credit card is the unauthorized and criminal use of your account by anyone other than the owner of that account. The necessary proactive measures that can be taken to put an end to this violence and conduct so-called deceptive practices can be studied in order to minimize 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 purposes, and the owner, and the issuing authorities, do not suspect 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 behaviour such as fraud, trespassing, and negligence.

This is a very urgent problem that requires the attention of the community, education and science, and information about its solution 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 operation pattern will change their statistical characteristics 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 should be used to analyse all operations, notify and report suspicious cases. These reports are tracked by our specialists who will apply the cardholder to check whether the transaction is original or fake. You are giving feedback to an automated system that uses training and algorithm updates, so that, we can ultimately improve the performance of cheating detection over a certain period of time

Fraud detection techniques are constantly being prepared to protect criminals adapting to lie strategies. Our system is classified as:

- Credit card fraud from online and offline
- Short thefts
- Consider Bankruptcy
- Copyright Block
- The App Is a Scam
- Fake Cards
- Telecommunications fraud

Here are some of the methods currently used to identify such practices:

- Artificial Neural Networks.
- Fuzzy logic
- Genetic algorithm
- Logistic regression
- Decision tree
- Machine Prop Direction
- Bayesian networks
- Hidden Markov model
- K-Nearest Neighbour

II. 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, reqlament 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, GJUS&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 discovered abandoned mining Distance and amount 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. All detected objects are independent of the underlying system, i.e. transactions that are not valid. They take attributes, customer behaviour, 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 from a completely new perspective. 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.

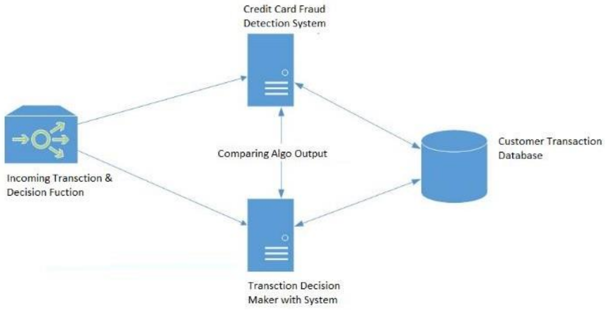
An artificial genetic algorithm is a great way to shed more light on this domain name, get scammed, on the other hand.

It turned out that, correctly, it is necessary to find fraudulent transactions and minimize the number of false positives. Even if it was related to a classification problem, the variable costs the correct classification.

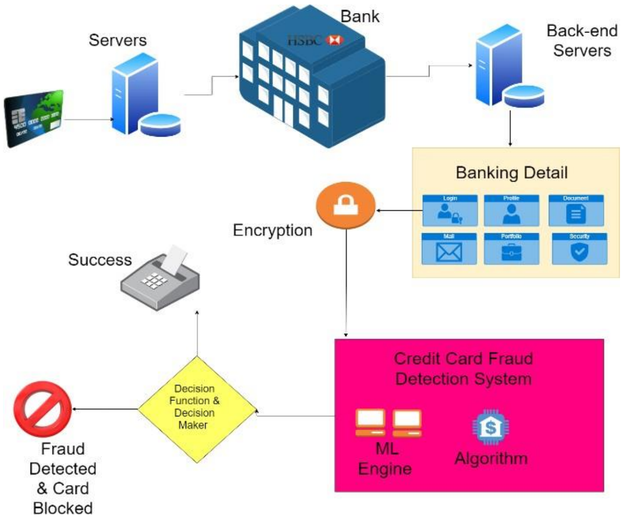
III. METHODOLOGY

The approach to this work is based on the use of the latest machine learning algorithms to detect abnormal activity, known as outliers.

Important for a rough architecture, the diagram can be represented by the following figure:


Credit Card Fraud Detection System
Incoming Transction & Decision Fuction
Comparing Algo Output
Customer Transaction Database
Transction Decision Maker with System

When a detailed overview of a large scale, along with, real-life, fully architecture, the diagram can be presented as follows:

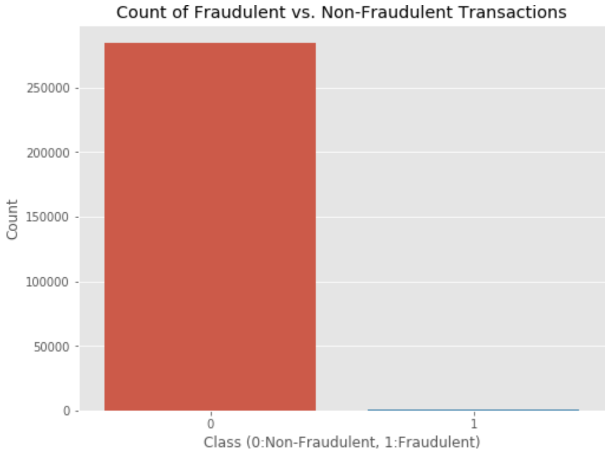


First of all, we get our Kaggle dataset, analyse this data to a website that provides these datasets.

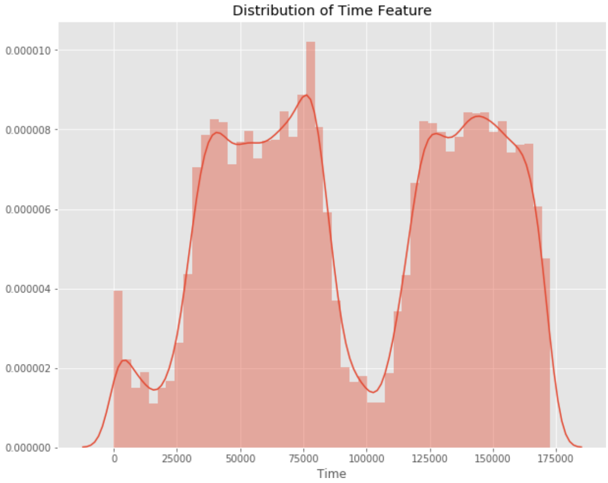
This dataset has 31 columns of 28 named v1-v28 to protect sensitive information.

The remaining columns are presented in Time, Amount, and Class. Time-displays the time between the first operation and after. Quantity is the amount of money in circulation. Score 0 for a valid military operation to start, and 1 is false.

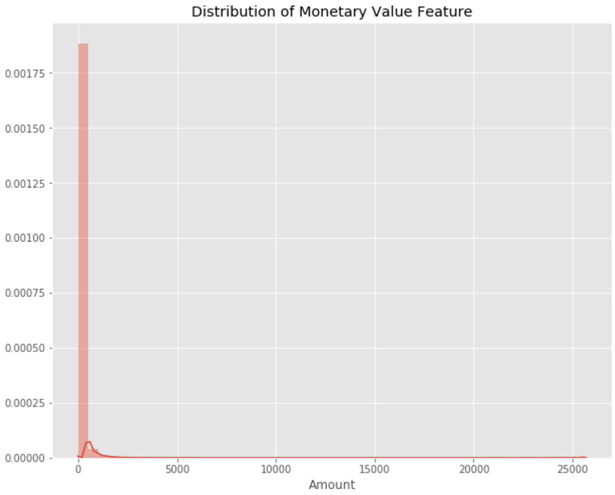
We will draw a graph to check the inconsistency of devices, data, and the visual meaning of this:



This chart shows that the number of fake transactions is a very low number of legitimate ones.

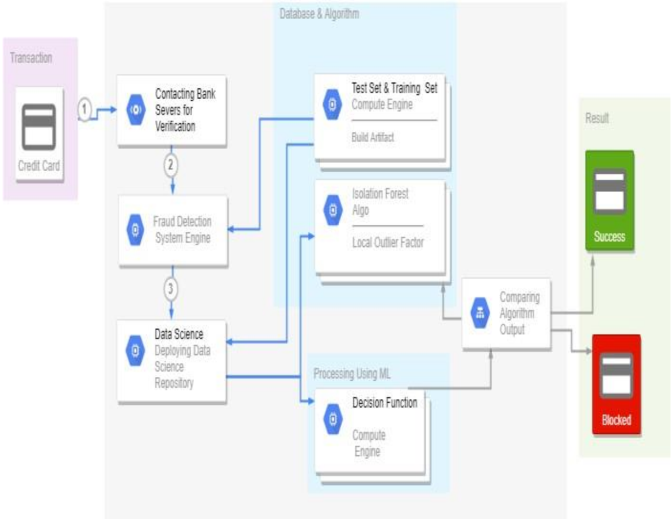


This chart shows when trades were made in a two-day event. It can be seen that the minimum number of transactions was made during the night, and this is the greatest day.

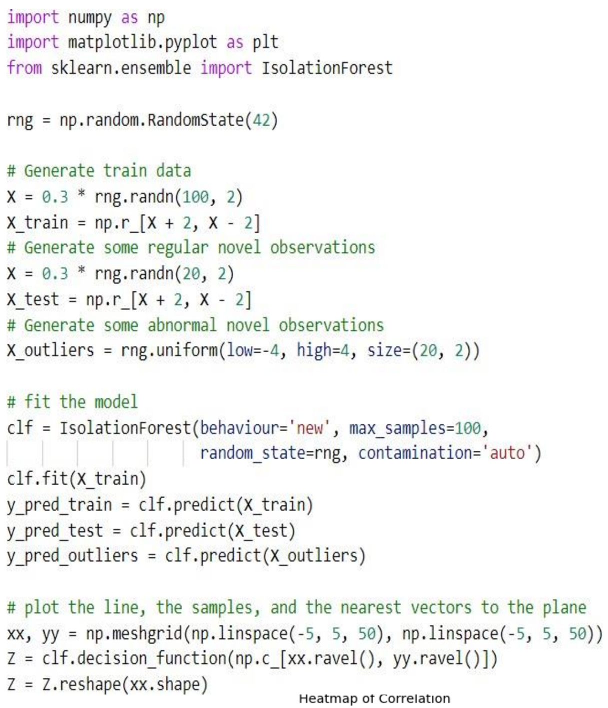


This graph represents the volume of products or services. Most transactions are relatively small, and only a few of them approach the maximum amount in circulation.

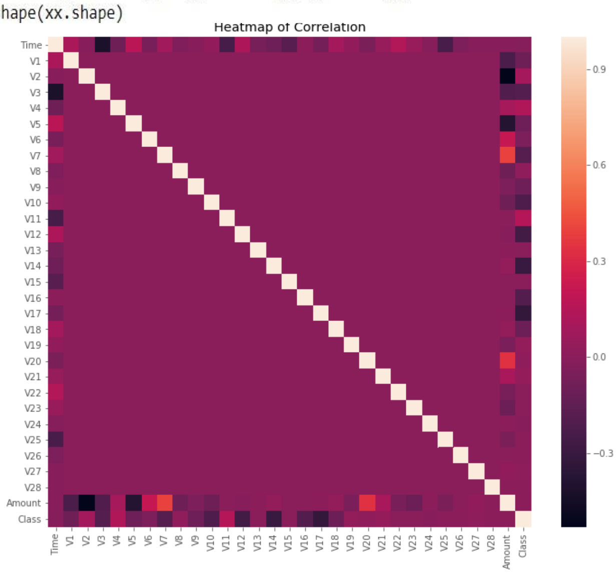
Then check if this is a data set, and plot a histogram of each column. This is in order to get a graphical representation of the data, sets that can be used to make sure that the data set has no missing values. To ensure this, we have no missing imputation values and machine learning algorithms are able to process the information efficiently.



For this analysis, we build heatmaps in order to get colourful presentations of information, to test the interaction between predictive variables and a class variable. This is later shown below:



```
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(20, 2)
X_test = np.r_[X + 2, X - 2]
# 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)
```



Now the data is formatted and edited. The time, as well as the number of columns, column, standard, and class columns must be removed to ensure the validity of the estimate. The data is processed using the complete module algorithms. Below is a module diagram explaining how these algorithms work together with metrics according to the model and the following outlier detection modules will be applied:

- Local Adverse Factors
- Isolation Forest algorithm

These algorithms are part of sklearn. The chorus module of the sklearn package is part of an ensemble on the stages of methods and events for classification, regression, and outlier detection.

It is a free and open source Python library built using the NumPy, SciPy, and matplotlib modules, giving you a very simple and powerful tool that you can use for data analysis and machine learning. There are many classification, clustering and regression algorithms and the system is designed to work with numerical and scientific libraries.

We use the Jupyter Notebook platform to make a Python program to demonstrate the method by which it is proposed. This application can also be implemented in the cloud using the Google Collab platform, which supports all python notebook files.

Detailed explanation of pseudocode modules their algorithms and relaxation programs all rights reserved below:

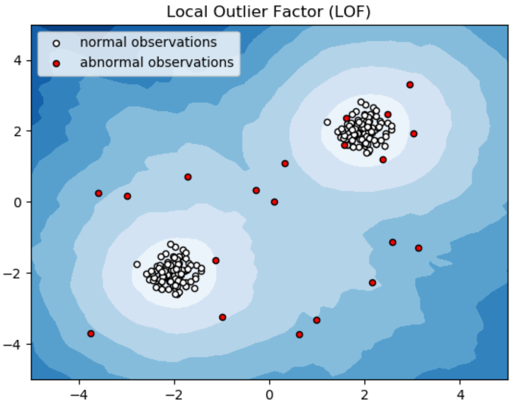
A. *Local Outlier Factor*

This is an unsupervised outlier detection algorithm. Local unfavourable factor " refers to the abnormal data of each sample. It measures local deviations of a data sample relative to neighbouring ones.

More precisely, the point provided by the k-nearest neighbours, the distances and their estimates are used for the local

The pseudocode for this algorithm is written as:

Graph of the results of local negative factor, algorithm, we get the following picture:



By comparing a local host with a sample of its neighbours, you can make an example that is significantly lower than that of their neighbours. They are quite ammonoid and they are considered the basis.

Because the data set is very large, we use only a small part of the test in order to reduce the required time.

Final processing results complete information is defined and specified in the work results section.

B. *Isolation Forest Algorithm*

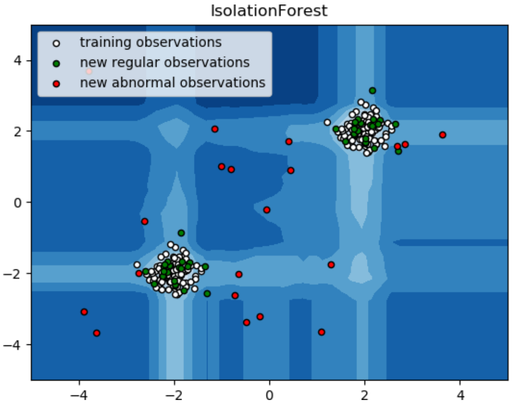
Isolation for isolation and observation, by randomly selecting an object, and then randomly selecting the split values between the maximum and minimum values of this function.

Recursive partitioning can be represented by a tree of the number of elements to split the sample, as well as the wavelength from the root node to the last node.

The average value for this route length is a measure of normality and the resolution of the function we use.

The pseudocode for this algorithm can be written as:

Graph of the results of the forest isolation algorithm, we get the following picture:



Splitting them randomly to bring shorter paths for anomalies. When there are random trees in the forest that are mutually shorter than the path length for specific samples, they are highly likely to be anomalies.

If detected, the system's data can be used for submission to the relevant authorities. For testing purposes, we compare the performance of these algorithms to evaluate accuracy and accuracy

IV. 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.

Therefore, we look in some references in the paper, for, following this, similar methods and results obtained. As noted in one of the reference sheets:

"This technology was applied in order to complete the application of the submitted information, the German bank in 2006. For the bank, for confidentiality reasons, below is only a summary of the results. Applying this technology to bid 1 to the list in a small number of cases, but the probability of being an impostor.

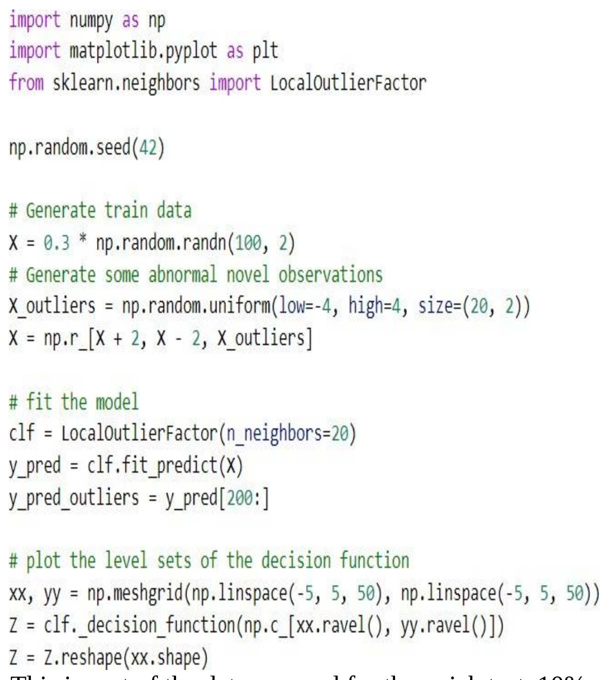
All people listed on this list have been briefly contacted to avoid any risks due to their high risk profile. The situation is complicated, and second on the list. Level 2 is still limited appropriately for case-by-case verification.

The Credit Line and meeting of officials determined that half of the cases on this list should be considered 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, a level 2 list and a level 3 list can be applied.

V. RESULTS

This code outputs the number of false positives that are detected and compared to the actual value. This is used to calculate estimates, accurate and correct algorithms.



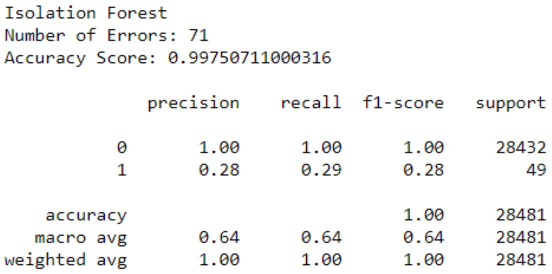
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

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))
X = np.r_[X + 2, X - 2, X_outliers]

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]

# 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)
```

This is part of the data we used for the quick test, 10% of the data set. A complete collection of information was also used at the end of the two print results.

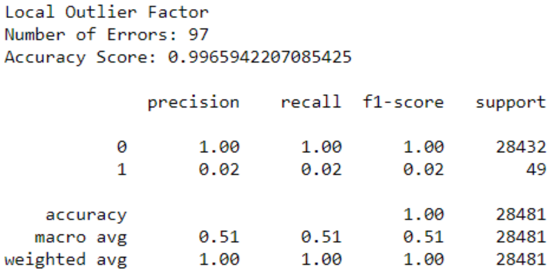
These results are combined with the classification report issued by the algorithm to output levels of 0, meaning that the operation is determined to be real, and 1 means that a fraudulent operation was established.

This map output can be used with values in the class to control for false-positive results.

Results when 10% of the dataset is used:

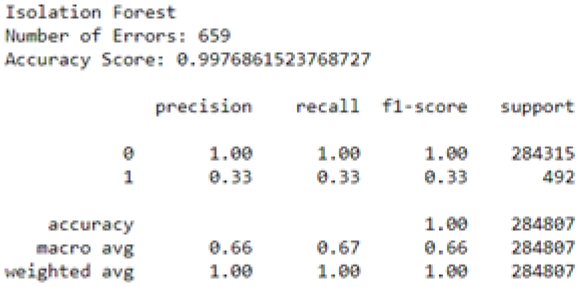


```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.28      0.29      0.28        49

    accuracy                           1.00     28481
   macro avg       0.64      0.64      0.64     28481
weighted avg       1.00      1.00      1.00     28481
```

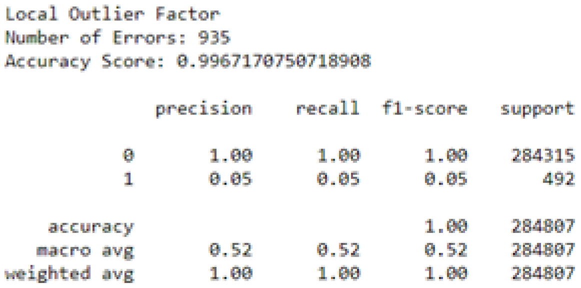


```
Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

Results with the complete dataset is used:



```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.33      0.33      0.33       492

    accuracy                           1.00    284807
   macro avg       0.66      0.67      0.66    284807
weighted avg       1.00      1.00      1.00    284807
```



```
Local Outlier Factor
Number of Errors: 935
Accuracy Score: 0.9967170750718908

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.05      0.05      0.05       492

    accuracy                           1.00    284807
   macro avg       0.52      0.52      0.52    284807
weighted avg       1.00      1.00      1.00    284807
```

VI. CONCLUSION

Credit card fraud is undoubtedly an act of criminal dishonesty. This article has identified the most common methods of fraud, along with methods for detecting them, and presents the most recent results in this area. This article explains in detail how it can be that education is applied in order to obtain a better result of fraud detection, along with the algorithm, pseudocode and an 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 the entire data set is given to the algorithm, 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 only two days of operation notes, and then only a small part of the information that is available, and this project is intended, therefore, for its use on a commercial scale. Based on machine learning algorithms, the program will not only improve its efficiency over time, as it will add more information.

VII. FUTURE ENHANCEMENTS

Until we have managed to achieve the goal with 100% accurate fraud detection, we will leave at the time of creating the system, in a state where enough time and a lot of information is available to get closer to the goal. As with any project, there is a lot of room for improvement.

The very nature of this project makes possible several algorithms that can be integrated together as a module, their results combined to improve the accuracy of the final result.

This model can be further improved with the addition of more and more optimization algorithms. But the output of these algorithms should be formatted in the same way as the others. If these conditions are met, the module is easy to integrate, just as you can do in the code. This allows you to ensure the degree of modularity and versatility of your project.

For more, there is room to improve set information. As shown earlier, the accuracy of the algorithm increases as the size of the data set increases. Therefore, by taking in more information, you will definitely see a model that is more accurate for fraud detection and reduces the number of false positives. But this requires support or official support from the banks themselves.

REFERENCES

[1] “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

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road,
Clayton, Victoria 3800, Australia

[3] “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4] “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

[5] “Credit Card Fraud Detection through Parenclitic Network Analysis- By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018,
Article ID 5764370, 9 pages

[6] “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

[7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by 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.