Credit Card Fraud Detection using Machine Learning and Data Science

Step 1: Prototype Selection

Abstract

For clients to avoid being charged for products they did not buy, credit card issuers must be able to recognise fraudulent credit card transactions. Data Science may be used to solve these issues, and coupled with machine learning, it is of utmost relevance. With the use of credit card fraud detection, this research aims to demonstrate the modelling of a data set using machine learning. Modelling previous credit card transactions using information from those that turned out to be fraudulent is part of the Credit Card Fraud Detection Problem. The validity of a new transaction is then determined using this approach. The goal here is to minimise inaccurate fraud categories while detecting 100% of the fraudulent transactions. A classic example of classification is the detection of credit card fraud. The analysis and pre-processing of data sets, as well as the use of several anomaly detection techniques, such as the Local Outlier Factor and Isolation Forest algorithm, to PCA-transformed Credit Card Transaction data, have been the main points of this approach.

Introduction

Credit card fraud refers to the unauthorised and unwelcome use of a credit card account by someone other than the account owner. The abuse can be stopped with the use of necessary preventative measures, and the behaviour of such fraudulent acts can be researched to lessen it and safeguard against recurrence. In other terms, credit card fraud is the use of another person's credit card for personal gain when neither the cardholder nor the organisation responsible for providing the card are aware that the card is being used.

Monitoring user populations' behaviour is a key component of fraud detection since it helps identify, detect, and prevent undesirable behaviours including fraud, intrusion, and defaulting.

This is a really pertinent issue that has to be addressed by communities like machine learning and data science, where an automated solution is possible.

From the standpoint of learning, this issue is particularly difficult because it is characterised by many characteristics, like class imbalance. There are significantly more legitimate transactions than fraudulent ones. Additionally, the statistical characteristics of the transaction patterns frequently vary over time.

The deployment of a fraud detection system in the real world is not without difficulties, though. In instances from the real world, the enormous volume of payment requests is swiftly reviewed by automated tools to choose which transactions to authorise.

Algorithms for machine learning are used to analyse all permitted transactions and flag any that seem suspect. Professionals look into these reports and get in touch with the cardholders to confirm whether the transaction was legitimate or fraudulent.

The automated system receives feedback from the investigators, which is then utilised to train and update the algorithm to eventually improve the performance of fraud detection over time.

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 suchfraud are:

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

Methodology

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers.

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

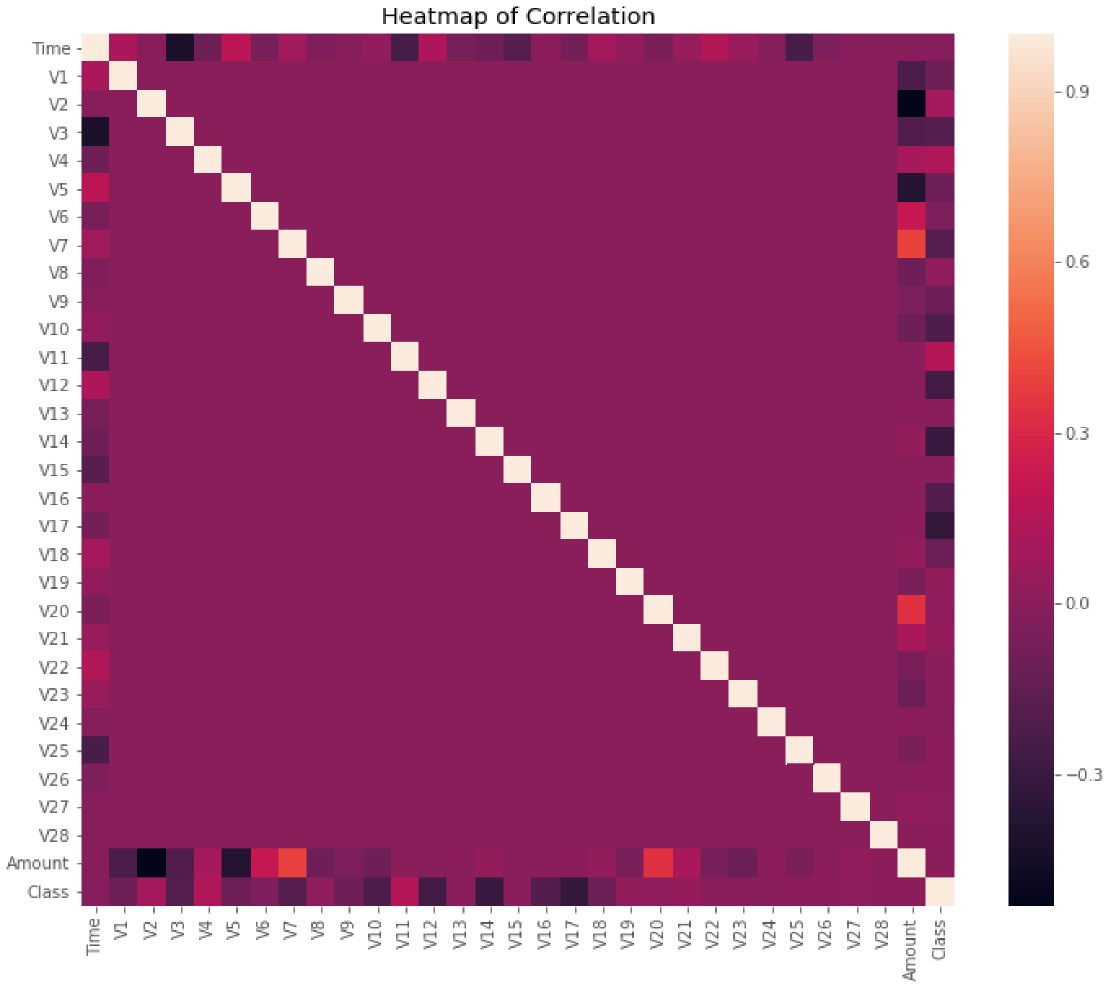
Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data.

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.

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.

After this analysis, we plot a heatmap to get a colored representation of the data and to study the correlation between out predicting variables and the class variable. This heatmap is shown below:



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.

Implementation

This concept is challenging to put into practice since banks are needed, yet they are reluctant to cooperate because to market competitiveness, legal considerations, and the need to secure customer data.

In order to acquire information, we searched for some reference publications that used comparable techniques. In one of these reference papers, it is written: "This technique was applied to a comprehensive application data set provided by a German bank in 2006. Only a summary of the findings is offered below due to banking confidentiality concerns. After using this method, the level 1 list includes a small number of cases, but these cases have a high likelihood of being fraudsters. Due to their high-risk profiles, all of the people identified on this list had their cards closed. For the other list, the requirement is more complicated. The level 2 list is still sufficiently restricted to be examined case by case. Officers in charge of credit and collections thought that half of the cases on this list might include suspicious fraudulent activity. The task is heavy for the final and largest list. Only one-third of them appear to be suspicious. One option is to add a new element to the query in order to increase time efficiency and reduce overhead costs. This element may be the password, email address, or the first five digits of the phone numbers.

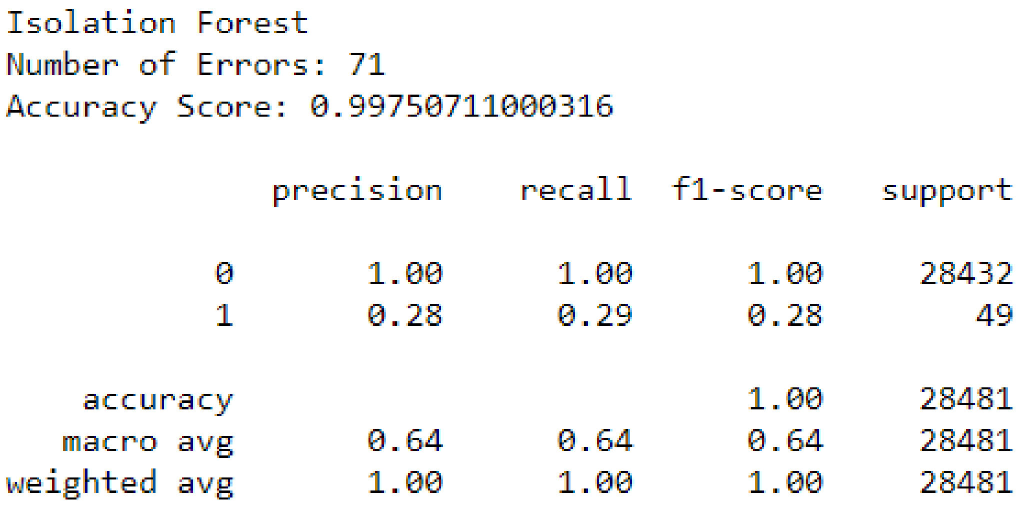
Results

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:

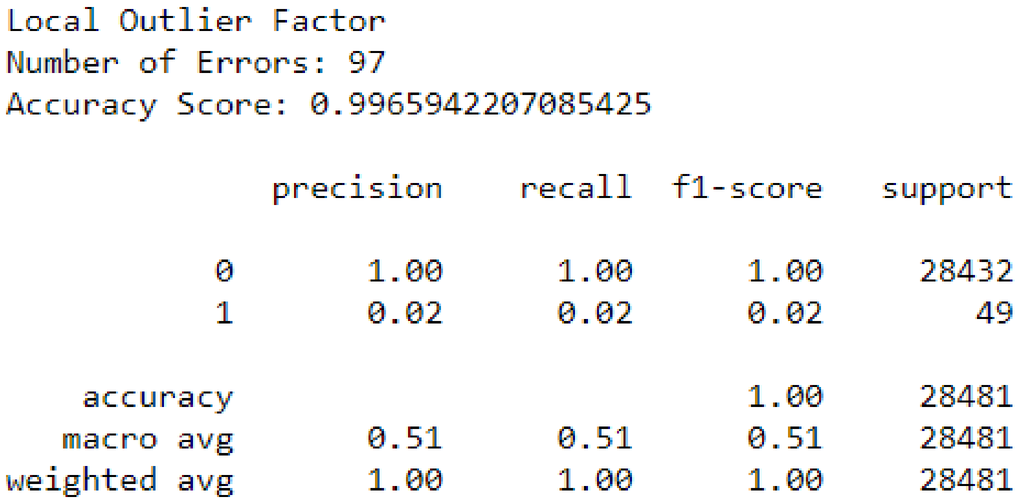


```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

              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
```



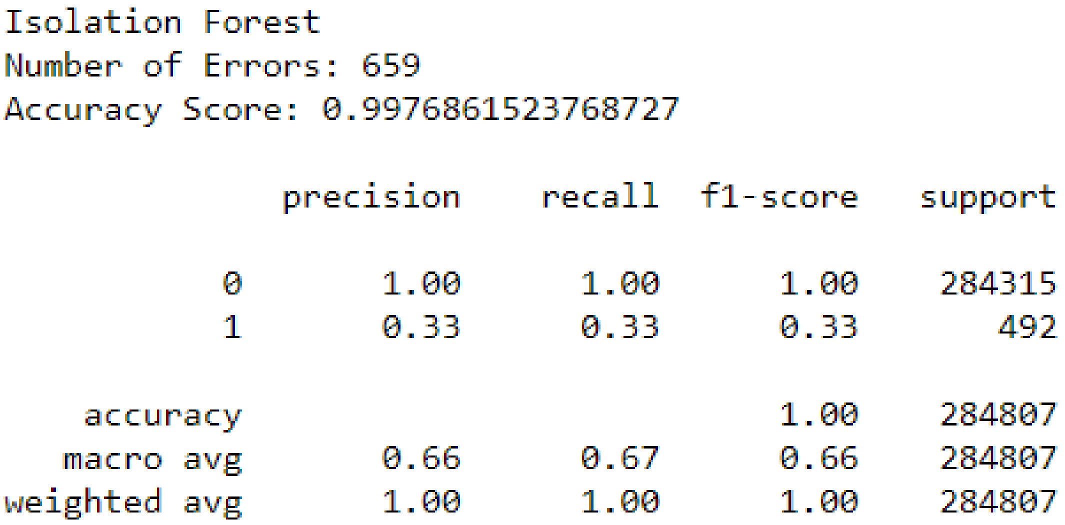
```
Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

              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:

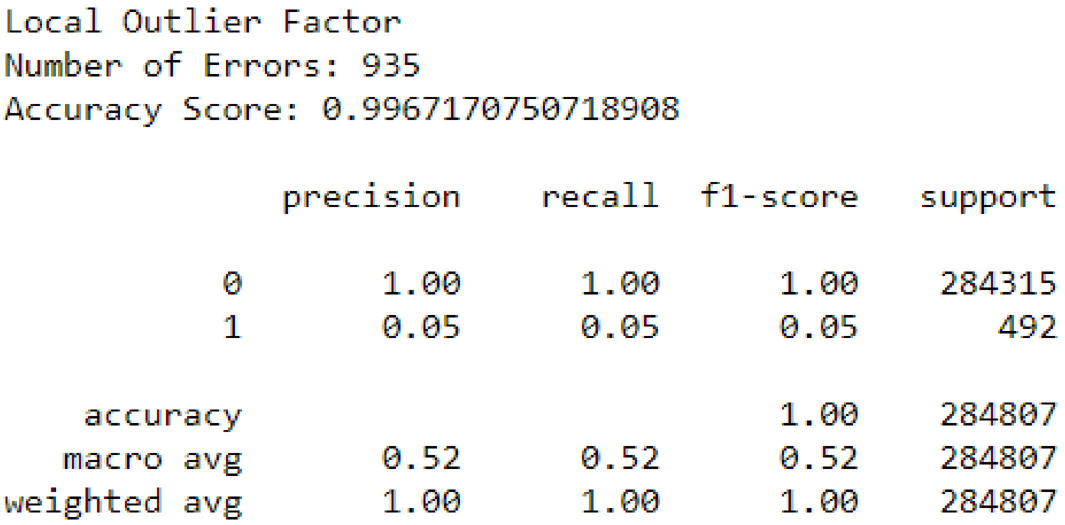


```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

              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
```

Feasibility

This project can be developed and deployed within a few months as SaaS for banks to use.

Viability

This project is viable to survive in the long-term future but improvements are necessary as new advancements emerge.

Monetization

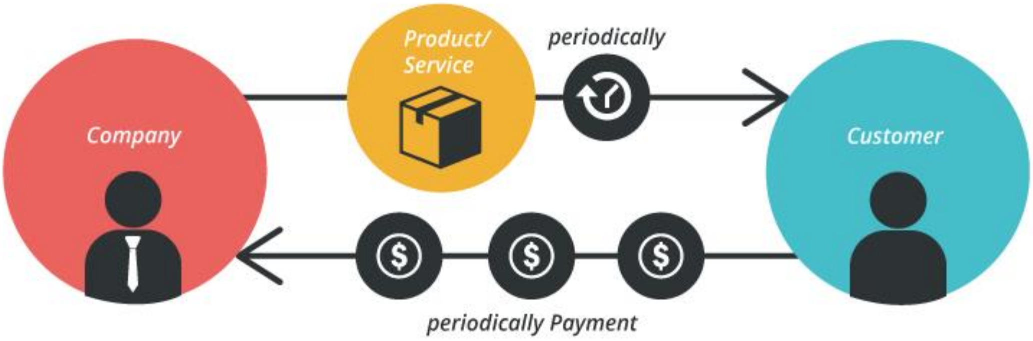
This service is directly monetizable as it can be used by banks.

Step 2: Prototype Development

GitHub Link:
https://github.com/rohansaxena2002/CreditCardFraudDetection

Step 3: Business Modelling

It is advantageous to employ a subscription-based model for this service, wherein some services will initially be offered for free to encourage client retention and boost our customer base. In order to continue using the service for their business, they will later be charged a membership fee. Customers that use the subscription business model pay a set sum of money at predetermined intervals to have access to the company's goods or services. How to turn casual users into paying customers is the main issue.



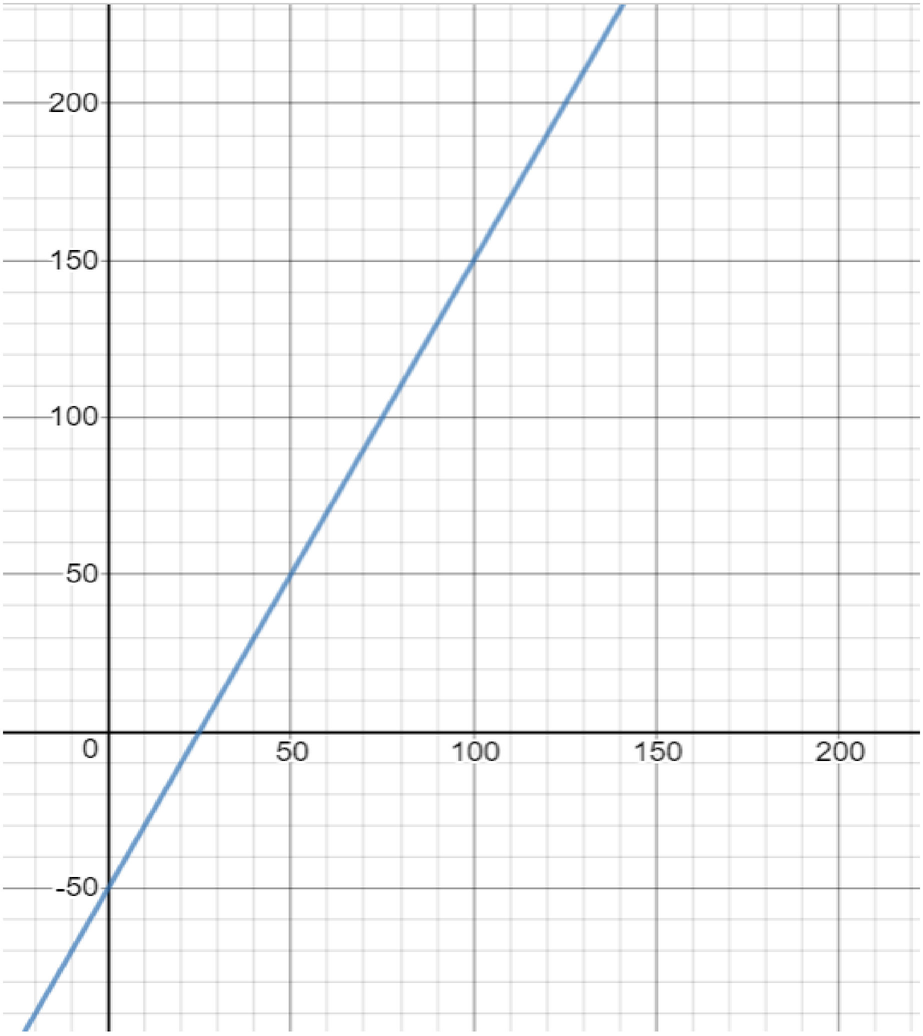
Step 3: Financial Modelling

This is a B2B service. Banks can use this product for fraud detection.

Let us consider the cost per customer to bank is Rs 100 per month. So, m =100.

Let total monthly customers = x and total monthly maintenance = c.

Thus, net profit y = mx – c.



Conclusion

Market basket analysis is being used by an increasing number of companies to acquire beneficial insights about associations and hidden relationships. However, for small businesses, this extension is a fantastic opportunity to boost sales and help them develop and grow their business.