Credit Card Fraud Detection using Isolation Forest

Spoorthy Reddy Jarugu[1]

[1] VIT, India

Introduction

Credit Cards are the most commonly used mode of payment nowadays. The reason is it has multiple features, which make it easy for users to make payments on the spot.
Isolation Forest is an anomaly detection algorithm well-suited for credit card fraud detection. Isolation Forest is an unsupervised algorithm that does not require labelled data to train.
Credit card fraud can be defined as any unauthorized use of a credit card, such as using a stolen credit card or making unauthorized purchases with a valid credit card. This dataset is taken from Kaggle.

The dataset contains transactions made by credit cards in September 2013 by European cardholders.

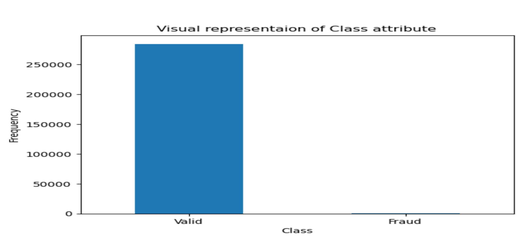
Unfortunately, due to confidentiality issues, original features are not given. The features given are the result of the PCA transformation.

There are a variety of techniques that can be used to detect credit card fraud. One common technique is to use machine learning models to identify patterns in fraudulent transactions.

Data visualization and Algorithm used

Data Visualization

- Number of classes with respect to frequency that are Valid transaction and Fraud transaction



- From the above diagram, we can see that Valid transactions are way more significant than fraud transactions

Algorithm used

Two algorithms are used for comparison

Isolation Forest: is a powerful tool for anomaly detection. It is fast, efficient, and robust to outliers and noise. It can detect anomalies in various applications, including fraud detection, intrusion detection, medical diagnosis, network monitoring, and financial market analysis.

Local Outlier Factor(LOR): Local Outlier Factor (LOF) is an unsupervised anomaly detection algorithm that identifies outliers based on their local density. LOF is calculated by comparing the local density of a data point to the local densities of its neighbours.

Comparing the accuracy of both algorithms

- Both the algorithms gave a reasonable accuracy rate. However, Isolation Forest (IF) is effective in fetching outliers for large datasets, whereas Local Outlier Factor (LOR) algorithms are computationally expensive for large datasets.
- IF is generally more interpretable than LOF.
- LOF has more hyperparameters to tune than IF

Algorithm	Accuracy
Isolation Forest	0.997156
Local Outlier Factor	0.996524

Classification Report

- Classification report is the performance of an Isolation Forest model on a given dataset.

Classification Report :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	28432
1	0.26	0.27	0.26	49
accuracy			1.00	28481
macro avg	0.63	0.63	0.63	28481
weighted avg	1.00	1.00	1.00	28481

- The model has perfect accuracy on the Valid Transaction (i.e., class 0), with a precision of 1.0 and a recall of 1.0. However, the model performs poorly on the Fraud transaction (class 1), with a precision of 0.26 and a recall of 0.27.

Analysis and Evaluation

Analysis

- From the above visual representation of the class feature, we can understand that Valid transactions are more in number than Fraud transactions. When understanding the percentage, it seems 0.17% are fraud transactions in the entire data.
- As it is very low, this data is highly imbalanced data. We need to balance the data using Undersampling

Evaluation

- Undersampling the dataset is done to get accurate results for IF and LOF algorithms.
- We can see that the IF algorithm outperformed the LOF algorithm and obtained an accuracy of 97.5%. Whereas LOF obtained 87.8%.
- Outliers obtained are 350
- Overall, the classification report shows that the model performed very well on class 0 data and moderately well on class 1 data. The model has a high accuracy, precision, and recall for class 0 data. The model has a moderate accuracy, precision, and recall for class 1 data.

Algorithm	Accuracy
Isolation Forest	0.9756
Local Outlier Factor	0.878647

Accuracy Report

Isolation Forest: 350
Classification Report :

	precision	recall	f1-score	support
0	0.99	0.99	0.99	15000
1	0.64	0.65	0.65	492
accuracy			0.98	15492
macro avg	0.82	0.82	0.82	15492
weighted avg	0.98	0.98	0.98	15492

Classification Report

Conclusion

- Flip 01 task I have chosen is to find "Credit Card Fraud Detection using Isolation Forest"
- The Isolation Forest algorithm has a number of advantages over other anomaly detection algorithms :
  - It is robust to outliers and noise.
  - It can be used to detect anomalies in both high-dim… dimensional data.
  - It is easy to implement and interpr…
- Using Isolation Forest for detecting … good accuracy compared to other algorithms.

Acknowledgement

- Flip 01