Credit Card Fraud Detection

Team: Data Cynics
(Hrithik Diwakar, Vinyas Vittal, Varshit Singh Yadav, Karthik M)

ABSTRACT: Credit card frauds are easy and friendly targets. E-commerce and many other online sites have increased the online payment modes, increasing the risk for online frauds. At the current state of the world, financial organizations expand the availability of financial facilities by employing of innovative services such as credit cards, Automated Teller Machines (ATM), internet and mobile banking services. Besides, along with the rapid advances of e-commerce, the use of credit card has become a convenience and necessary part of financial life. Due to increase in fraud rates, researchers started using different machine learning methods to detect and analyze frauds in online transactions.

INTRODUCTION AND BACKGROUND

Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used.
This is a very relevant problem which demands the attention of communities such as machine learning and data science where the solution to this problem can be automated and find a satisfying solution.
The aim is to detect the different frauds in credit card using different machine learning techniques and then build the best model that produces accurate result.

PREVIOUS WORK

Title: Credit Card Fraud Detection using Machine Learning Algorithms
Author: Vaishnavi Nath Dornadulaa S Geethaa
Source: Sciencedirect.com
Objective: The main aim of the paper was to design and develop a novel fraud detection method for Streaming Transaction Data. Techniques used is sliding window strategy, Synthetic Minority Over-Sampling Technique and Matthews Correlation Coefficient to handle imbalance data. Followed by a feedback mechanism to solve the problem of concept drift.
Main Claims:
The dataset is highly unbalanced.
It is observed that the Matthews Correlation Coefficient was the better parameter to deal with imbalance dataset. MCC was not the only solution. By applying the SMOTE, it is possible to balance the dataset, where it is found that the classifiers were performing better than before.
The other way of handling imbalance dataset is to use one-class classifiers like one-class SVM.
Finally observed that Logistic regression, decision tree and random forest are the algorithms that gave better results.
Take away from the paper:
I got to know how to handle imbalance datasets. And also, accuracy and precision are never good parameters for evaluating a model. But accuracy and precision are always considered as the base parameter to evaluate any model.

PROPOSED SOLUTION

The target attribute in the dataset is ‘Class’, which takes the value 1 in case of fraud and 0 otherwise. As a part of the pre-processing and explanatory analysis of data the following observation are made:
1)Number of columns 21
2)Sample size 284807
Number of quantative variable – 28
(V1, V2………..,V28) are the principal component obtained with PCA.
The feature amount is the transaction amount that can be used for example dependent cost- sensitive learning.
Number of qualitative variable-1 (class) value 0 represent non-fraud and 1 represent fraud transaction.
Missing values were checked using an in-built python function .0 missing values are found.
A Box plot is made to compare the fraud and normal transaction the transaction from normal cases are way higher than that for transaction.
Histograms are plotted to find the distribution of attributes like amount and number of transactions We can notice that in fraud transaction the amount is very less.
Correlation matrix is also created to get the idea of how future correlate with each other and can help us predict those features that are our most relevant.

Using Python, the following observations were made on the dataset as a part of the initial Exploratory Data
Analytics and Visualization part.

Number of Columns: 31
Number of Samples: 284807
Number of Quantitative variables: 28
Number of Qualitative variables: 3

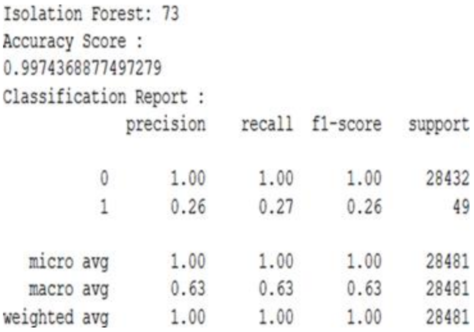
Model Prediction

Now it is time to start building the model. The following are the algorithms we use to detect anomaly in credit card transaction.

1.Isolation Forest Algorithm

Isolation forest is one of the machine learning algorithm widely used for credit card anomaly detection. The outliers get isolated during the formation of decision trees internally.
It is an unsupervised algorithm. It does its job of isolating the outliers by selecting a feature from the set of features randomly and then a split value is selected randomly between the maximum and minimum values of the feature. This random partitioning of features will produce shorter paths in trees for the anomalous data points, thus distinguishing them from the rest of the data.

How Isolation Forests Works?



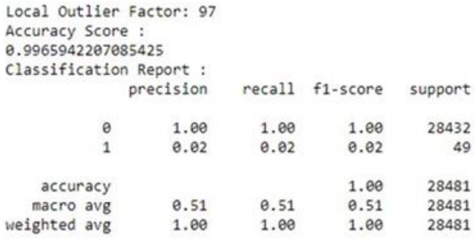
```
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.26      0.27      0.26        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.63      0.63      0.63     28481
weighted avg       1.00      1.00      1.00     28481
```

Experimental Results:

Isolation Forest detected 73 errors. Isolation Forest has a 99.74% accuracy.
The algorithm first randomly creates decision trees that are isolated. The score is then calculated based on the depth of the leaf nodes to isolate.

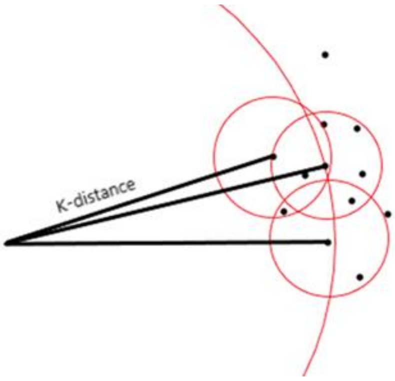
Using Isolation Forest algorithm, we can save lot of memory space and we can also detect the anomalies faster.

2.Local Outlier Factor
The Local outlier factor algorithm is outlier detection technique. It computes the deviation in local density of the given datapoint with respect to its neighbors. It considers the outlier samples that have a significantly lower density than their neighbors.



```
Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :
              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
```

Local density is determined by estimating distances between data points that are neighbors (k-nearest neighbors). Local density can be calculated for each point. By comparing these we can check which data points have similar densities and which have a lesser density than its neighbors. The datapoints with lower densities are the outliers. Firstly, k-distances are distances between points that are calculated for each point to determine their k-nearest neighbors. The 2nd closest point is said to be the 2nd nearest neighbor to the point. The following image shows the cluster of a point with various neighbors and their k-distances:



The reachability distance is calculated using this distance. It is defined as the maximum of the distance between two points and the k-distance of that point. In the following equation, where B is the point in the center and A is a point near to it.

$$\mathrm{lrd}_k(A) := 1 / \left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|} \right)$$

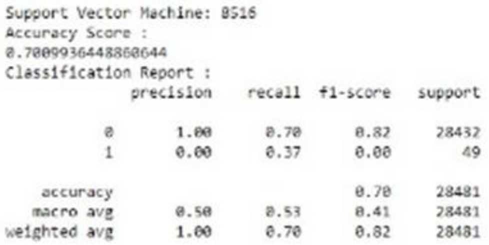
The calculation of Local outlier factor (LOR) is done by taking the ratio of the average of the lrds of k number of neighbors of a point and the lrd of that point. Here is the equation for LOR:

$$\mathrm{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\mathrm{lrd}_k(B)}{\mathrm{lrd}_k(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \mathrm{lrd}_k(B)}{|N_k(A)| \cdot \mathrm{lrd}_k(A)}$$

So, in the equation, if the density of the neighbors and the point are almost equal we can say they are quite similar; if the density of the point is greater than the density of the neighbor, then the point lies inside the cluster, and if the density of the neighbors is more than the density of the point we can say that the point is an outlier.

3.Support Vector Machine

Support Vector Machine is a linear model that creates a hyper plane which classify the data into classes and their corresponding category labels. The parameters predicted by SVM model shows 8516 fraudulent cases which is false. SVM model is not performing well because of highly unbalanced dataset with only 70% accuracy.



```
Support Vector Machine: 8516
Accuracy Score :
0.7009936448860644
Classification Report :
              precision    recall  f1-score   support

           0       1.00      0.70      0.82     28432
           1       0.00      0.37      0.00        49

    accuracy                           0.70     28481
   macro avg       0.50      0.53      0.41     28481
weighted avg       1.00      0.70      0.82     28481
```

CONCLUSION

The Credit Card fraud detection project implemented is a Machine learning Anomaly detection technique of identifying rare events or observation which can raise suspicions from the rest of the observations.
Credit card fraud is a act of criminal dishonesty. So there is a great demand to identify such criminal activities and take necessary steps to avoid them for maintaining harmony in the society. Accuracy score, precision, recall are being used as the performance parameters. If we clearly compare the models that we implemented the local outlier factor algorithm and isolation forest algorithm clearly out performs the Support Vector Machine model.

REFERENCES:

[1] Nitu Kumari, S. Kannan and A. Muthukumaravel, “Credit Card Fraud Detection Using Genetic-A Survey” published by Middle-East Journal of Scientific Research , IDOSI Publications, 2014

[2] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, “A Tool for Effective Detection of Fraud in Credit Card System”, published in International Journal of Communication Network Security ISSN: 2231 – 1882, Volume-2, Issue-1, 2013.

[3] Rinky D. Patel and Dheeraj Kumar Singh, “Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm”, published by International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[4] M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik, “Improving a credit card fraud detection system using genetic algorithm”, published by International conference on

Networking and information technology, 2010.

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

CONTRIBUTION:

Hrithik Diwakar(PES2201800666) – Local Outlier Factor Algorithm

Karthik M(PES2201800410) – Isolation Forest Algorithm

Vinyas Vittal(PES2201800641) – Pre Processing

Varshit Singh Yadav(PES2201800608)– Support Vector Machine Algorithm

----