MACHINE LEARNING IN BANKING AND FINANCE SECTOR

Ms. Priya P Nadar[1], Ms. Apeksha Khopkar[2]

[1]Auxillary Trainee,
BNP Paribas,
Infinity Building No.4,
Off Film City Road,
Malad(East),
Mumbai - 400 097

[2]Assistant Professor,
Navinchandra Mehta Institute of Technology and Development,
Dadar(West)

ABSTRACT

Finance and banking sector has several departments, every department has several products with policies and factors such as maintaining and retaining customers, fraudulent transaction identification. And these factors are vital for banks to survive in a long run. Before launching any new product or policies in the market they can have better automated decision making tools to know the audience reception and risk scale for investment. Having precise and accurate tools for identification of fraudulent transfer of money and take corrective actions can save lots of money and help in customer retention. So, varying on the type of observation and end results required, different types of machine learning algorithms can be used for decision making, identifying anomalies and taking actions. And banks and finance institutes are now focusing on automation by using AI and machine learning. Our case study will be focusing one such field which is quite vulnerable, that is, anomaly detection , best suited algorithms for fraudulent credit card detection is studied and presented in this research paper.

KEY WORDS

Machine learning, banks, anomaly detection, isolation forest algorithm, local outlier factor

I. INTRODUCTION

Machine learning algorithms are being rapidly adopted by varied applications in the financial services industry. As such, it is important to start considering the financial stability implications of such using machine learning. Because uses of this technology in finance are in developing and rapidly evolving phase, and data on usage are largely unavailable, any analysis must be

necessarily preliminary before making decisions, and developments in this area should be monitored closely.

Machine learning could be possibly better in following sections of banking: Fraud prevention, investment prediction, digital assistance, network security, algorithm trading, document interpretation, trade settlements, risk managements, customer service, marketing, loan underwriting, process automation or even money- laundering prevention. Our application focuses on one such domain that is anomaly detection in credit card transaction. Financial service providers have no greater responsibility than protecting their clients against fraudulent activity. But for every $1 lost to fraud, financial institutions pay $2.92 in recovery and associated cost. To win the war against financial fraud, financial companies must abandon outdated approaches. Identifying and preventing fraudulent transactions requires sophisticated solutions that can analyze high-volume data. Machine learning offers such a solution. By spotting patterns and using predictive analytics, machine learning algorithms can block fraudulent transactions with a degree of accuracy not even possible with stand-alone AI [1].

There are several algorithms that could be used for our application amongst which our application uses these two algorithms, namely, Local Outlier Factor and Isolation Forest Algorithm.

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors [2].

The Isolation Forest ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Since recursive partitioning can be represented by a tree structure, the number of splitting required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality and our decision function. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies [3].

II. REVIEW OF LITERATURE

PayPal uses three types of machine learning algorithms for risk management: linear, neural network, and deep learning. Experience has shown PayPal that in many cases, the most effective approach is to use all three at once. Dr. Hui Wang of PayPal, senior director of risk sciences says “So in order to get the best out of all [three], we “ensemble” them together. We have a “voting

committee." One is linear and one is nonlinear and we just ask them: What is your opinion on this file? Then we take their vote and eventually ensemble them together for our final assessment [4]".

Citibank: Citibank has their own startup accelerator, grouping multiple tech startups worldwide. Most of these companies develop products in the field of financial services and cybersecurity. One of their most notable moves was investing heavily in FeedzAI, the global enterprise that concentrates on using data science to identify and demolish fraudulent attempts in various avenues of financial activities, including online and mobile banking. FeedzAI uses machine learning algorithms to analyze huge volumes of Big Data real-time and alert the financial institutions of alleged fraud cases at once [5].

III. RESEARCH METHODOLOGY

Local Outlier Factor (LOF): Local Outlier Factor (LOF) is a score that tells how likely a certain data point is an outlier/anomaly.

$LOF \approx 1 \Rightarrow$ no outlier

$LOF \gg 1 \Rightarrow$ outlier

The LOF is a calculation that looks at the neighbors of a certain point to find out its density and compare this to the density of other points later on. While a small *k* has a more local focus, i.e. looks only at nearby points, it is more erroneous when having much noise in the data. With this *k* defined, we can introduce the *k-distance* which is the distance of a point to its *kth* neighbor. The *k-distance* is now used to calculate the reachability distance. This distance measure is simply the maximum of the distance of two points and the *k-distance* of the second point.
reach-dist(a,b) = max{k-distance(b), dist(a,b)}[6]

Basically, if point *a* is within the *k* neighbors of point *b*, the *reach-dist(a,b)* will be the *k-distance* of *b*. Otherwise, it will be the real distance of *a* and *b*. This is just a "smoothing factor". For simplicity, consider this the usual distance between two points.[6]

The *reach-dist* is then used to calculate still another concept—the local reachability density (lrd). To get the lrd for a point *a*, we will first calculate the reachability distance of *a* to all its *k* nearest neighbors and take the average of that number. The lrd is then simply the inverse of that average. Remember that we are talking about densities and, therefore, the longer the distance to the next neighbors, the sparser the area the respective point is located in. Hence, the less dense it is, the inverse.
lrd(a) = 1/(sum(reach-dist(a,n))/k)[6]

By intuition the local reachability density tells how far we have to travel from our point to reach the next point or cluster of points. The lower it is, the less dense it is, the longer we have to travel.

LOF of a point tells the density of this point compared to the density of its neighbors. If the density of a point is much smaller than the densities of its neighbors (LOF ≫1), the point is far from dense areas and, hence, an outlier. [6]

Isolation Forest: Isolation Forest explicitly identifies anomalies instead of profiling normal data points. Isolation Forest, like any tree ensemble method, is built on the basis of decision trees. In these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature [7].

In principle, outliers are less frequent than regular observations and are different from them in terms of values. That is why by using such random partitioning they should be identified closer to the root of the tree (shorter average path length, i.e., the number of edges an observation must pass in the tree going from the root to the terminal node), with fewer splits necessary.The idea of identifying a normal vs. abnormal observation can be observed in Figure 1 from [8]. A normal point (on the left) requires more partitions to be identified than an abnormal point (right).

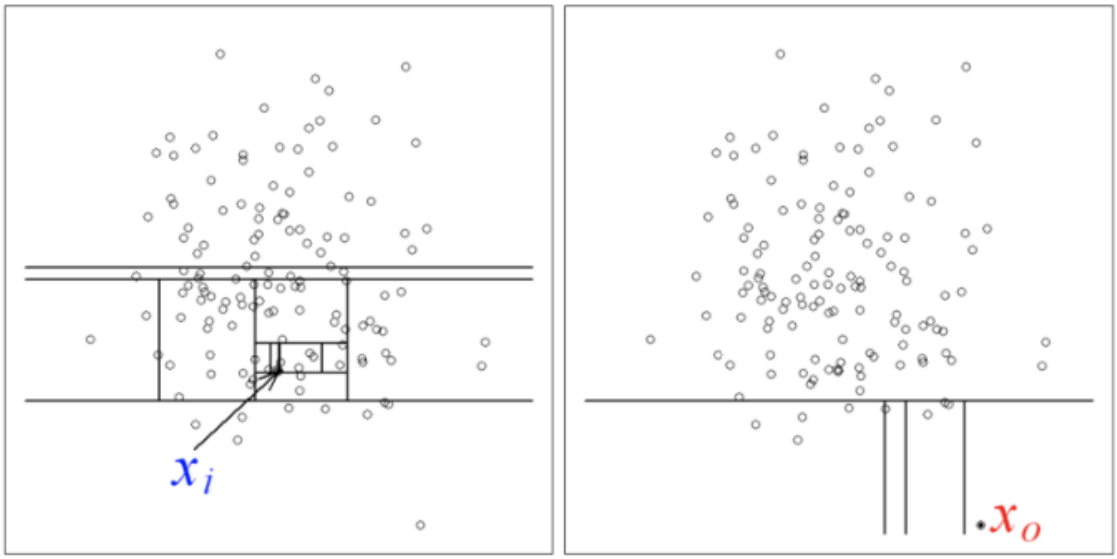

Figure 1 Identifying normal vs. abnormal observations from [8].

As with other outlier detection methods, an anomaly score is required for decision making. In case of Isolation Forest it is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where h(x) is the path length of observation x, c(n) is the average path length of unsuccessful search in a Binary Search Tree and n is the number of external nodes.

Each observation is given an anomaly score and the following decision can be made on its basis:

Score close to 1 indicates anomalies

Score much smaller than 0.5 indicates normal observations

If all scores are close to 0.5 than the entire sample does not seem to have clearly distinct anomalies. [7]

IV. DATA INTERPRETATION/ANALYSIS

We have used python with sklearn for training the model with dataset. The dataset is obtained from kaggle's official website [9] which contains European Dataset containing credit card transactions in a .csv file.

```
Fraud = data[data['Class'] == 1]
Valid = data[data['Class'] == 0]

outlier_fraction = len(Fraud)/float(len(Valid))
print(outlier_fraction)

print('Fraud Cases: {}'.format(len(data[data['Class'] == 1])))
print('Valid Transactions: {}'.format(len(data[data['Class'] == 0])))
```

```
0.0015966014193575613
Fraud Cases: 227
Valid Transactions: 142177
```

Figure 2 Outlier ratios in dataset, Fraud Cases and Valid Transaction

The columns used in dataset with its range of values displayed in histogram are as follows:

One important column to be observed is “Class” which has maximum values towards 0, indicating normal transactions and very few towards 1 indicating fraudulent transaction. But it’s obvious, because these data are collected from real world transactions.

The “time” column indicates the gap between each transaction performed. Several other columns are masked by kaggle [9], for not disclosing customer’s personal details such as name, address etc.

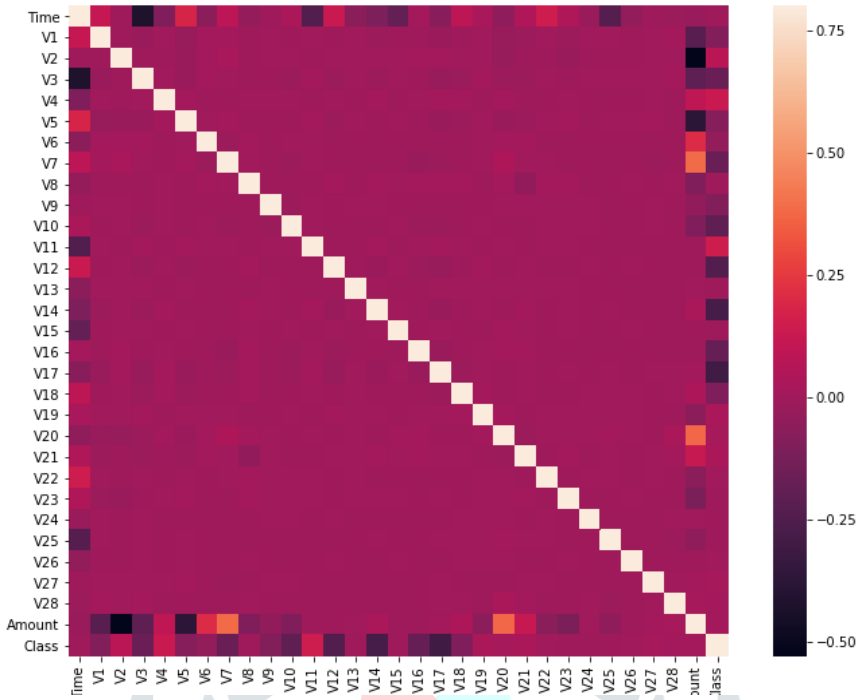

Figure 3 Correlation matrix between columns

Values in Figure 3 are not strongly related because the values lie closer to zero. There are several greyish areas as well indicating less strong relations between them. For example, V17 could be stronger negative relation whereas V4 could be stronger positive relation. And this dataset hereby suffice our application needs.

So the next step is to feed our algorithm with the dataset.

So for the outcome would predict the normal transactions and outliers as fraudulent transactions

V. RESULTS

After training our model with dataset, the outcomes obtained is good.

```
Isolation Forest: 327
0.9977037161877476
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    142177
           1       0.28      0.28      0.28       227

   micro avg       1.00      1.00      1.00    142404
   macro avg       0.64      0.64      0.64    142404
weighted avg       1.00      1.00      1.00    142404

Local Outlier Factor: 439
0.9969172214263644
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    142177
           1       0.04      0.04      0.04       227

   micro avg       1.00      1.00      1.00    142404
   macro avg       0.52      0.52      0.52    142404
weighted avg       1.00      1.00      1.00    142404
```

Figure 4 (i) Output of Isolation Forest Algorithm and (ii) Local Outlier Factor

In Figure4, “0” indicates valid transactions and “1” indicates fraudulent transactions. The models used, calculated the precision, recall, f1-score and support. Most crucial deciding columns are precision and recall. Precision indicates false positives and recall indicates false negative.

So the ratio between false positive and false negative is far from each other, because the dataset that we used has approximate 0.17% of fraud activities. So depending on the dataset values we are training the model’s response to it changes.

VI. CONCLUSION

The model could be implemented in small scale firms with limited dataset as in our application. The more real world transactions we feed to the model gradually it will adapt to changes and respond properly. The isolation forest algorithm is best suited for our application because our dataset has several columns which may have led to overfitting in other algorithms. Thus, these two unsupervised learning algorithm is best suited for our applications.

VII. SCOPE FOR FURTHER RESEARCH

Amongst several areas of banking where machine learning could be used we have only produced solution for fraud detection for credit card transactions. There are many areas to explore in banking and finance sector where varied machine learning algorithms could be used.

VIII. REFERENCES

1. AVIRAM EISENBERG, 7 Ways Fintechs Use Machine Learning to Outsmart the Competition , NOVEMBER 30, 2018

2. scikit-learn developers, https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html

3. scikit-learn developers ,https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

4. Eric Knorr, Editor in Chief, InfoWorld, How PayPal beats the bad guys with machine learning

5. Vladimir Fedak, 5 use cases of Machine Learning in the banking industry, Jan 22, 2018

6. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, №2, pp. 93–104). ACM.

7. Eryk Lewinson, Jul 3, 2018, Outlier Detection with Isolation Forest

8. [1] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 413–422). IEEE.

9. https://www.kaggle.com/mlg-ulb/creditcardfraud