Group 6: Sean Atkinson, Sameer Khan, Parker Davis

Assignment 4.1: Fraud Assessment Methodology

Abstract

The following report provides insight into the utilization of anomaly detection to identify fraudulent sales based on the quantity and value of the products sold. Using two unsupervised learning techniques, a comparative analysis is performed to recommend the best method for Group 6 retailers to detect fraud. Our analysis uses sales transaction data to train a model against the testing data set to determine a F1 score for each unsupervised method. An F1 score will combine the precision/recall metrics for each model and determine the best method of fraud detection in the sales department.

Keywords: *Unsupervised learning, F1 Score, Density-Based Clustering of Applications with Noise* (*DCSCAN), Local Outlier Factor (LOF), Isolation Forest*

Introduction

The data provided as inference of suspected fraudulent activity amongst sales personnel with respect to the classification of “OK” and “FRAUD " requires a statistical model in order to automate the process of determining transaction validity. The application of anomaly detection through various methods can correctly identify cases of fraud while maintaining a low false positive rate. Density-Based Clustering of Applications with Noise (DBScan) is an unsupervised learning non-linear algorithm which utilizes the density of a cluster to determine outliers in the data. A data point is classified as normal or an anomaly using a local outlier factor (LOF) to see if observations fit into the composed clusters. The data will be characterized and similarly aligned data points will create the cluster. As data is added to the model those clusters can grow or specific data points will become apparent as not being within a specific range of reachability or connectivity with respective clusters formed in the model.

A secondary method that the team will evaluate is the Isolation Forest technique. This method is conceptualized on the basis that anomalies are few relative to normal data points and also different from these points such that they can be identified by the model. The tree concept is based on how the model builds the branches of a tree, via a random selection of features. These features are used to process data points and align them to other similar points represented by branches of the tree. We would expect branches comprised of typical points to traverse deeper into the tree structure. The concept of anomalies is the branches built for outliers are less dense and so a branch is easily identified as it is not aligned to similar features so they branch out on their own.

Using the output of the models above, the team believes a single model can be chosen from the analysis to both provide accurate results but repeatable processes in determining the classification of both “OK” and “FRAUD” transactions. Training and test models will be created for DBScan and Isolation Forest to see which one produces more accurate results in terms of the correct classification of both normal points and outliers.

Literature Review

Credit card fraud detection has long been an important need. In many cases unsupervised learning methods prove much more successful at identifying fraud than supervised learning. This is because unsupervised learning models can often form their own relationships between variables and not have to follow the strict format of a supervised method. Much work has been done in using DBScan and Isolation Forest as fraud detection methods. DBScan is an unsupervised method that performs density clustering with multidimensional data. An effective model was developed to look at outlier and fraud detection where the two most important factors for a good model are the epsilon and minimum sample values (Alam 2020). These parameters set the standard for how effective the model will be. The epsilon determines the minimum distance between points and the minimum samples determine the minimum number of neighbors a point should have. These numbers take some trial and error to make a model operate effectively.

Vijayakumar, Divya, Sarojini & Sonika utilized a similar dataset with known fraud detections flagged to help train their isolation forest model. Through preprocessing with PCA it was discovered that their data was highly skewed with a vast majority of purchases being non-fraudulent. This is expected as there are vastly more "typical" transactions when compared to fraudulent ones. They first tried identifying outliers in the data using LOF, which compares a data point to its neighbors with regards to local cluster density. Then, the isolation forest method was used as an alternative method to find outliers. This proved highly successful due to its ability to detect anomalies not by using simple distance measures like LOF, but by starting from a random feature and setting a random threshold to compare data and create branches (2020). This methodology was their most successful approach.

Methods

Using the current assessment results and a dataset for training the unsupervised learning model we have the following characteristics to provide as input to our model:

1. The training data consist of 133,731 unlabeled sales transactions across 798 products.

2. Test data include 15,732 inspected sales transactions across these same products
3. Manual inspection by the analysis team shows the test data found 14,462 transactions to be normal and 1,270 fraudulent.

With the insight provided by the analytics team, the creation of a model will require elements of the initial dataset to be altered. The following alterations have been made:

1. Rows with missing elements for either quantity or val have been removed
2. Categorical columns (ID, Insp) have been removed
3. The elements in the Quant and Val columns have been scaled to reflect normalized values

Missing data would not provide the inference needed and given the number of complete transactions this is seen as a required first step. Scaling provides an opportunity to explore the interconnections between the “normal” data elements and the outliers. Using such an approach provides both a measure to decrease the issue of model development with heavily skewed data (Figure 1). Development of a model with such a differential in the variables is often unstable, meaning the model may suffer from poor performance during learning and sensitivity to input values resulting in higher generalization error.

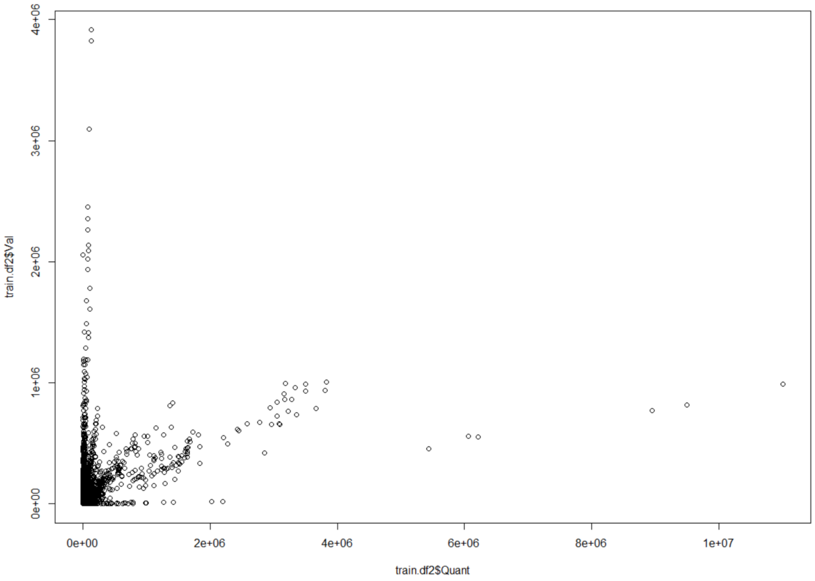

Figure 1 Unscaled data visualization

Density-Based Clustering of Applications with Noise (DBSCAN)

In order to accurately model the data with DBSCAN two critical parameters need to be set. The minPts which defines the minimum amount of data points allowed to make up a cluster, and epsilon (eps) which is the minimum distance between points to be counted in the same cluster. The minPts input is selected through trial and error based on domain knowledge; there is not an effective way to algorithmically determine an appropriate value. Eps can be determined using a K nearest neighbors method as seen below in Figure 2. This method calculates the average distance between points and a specified number of their nearest neighbors. The ideal eps value would be located at the elbow of the plot.

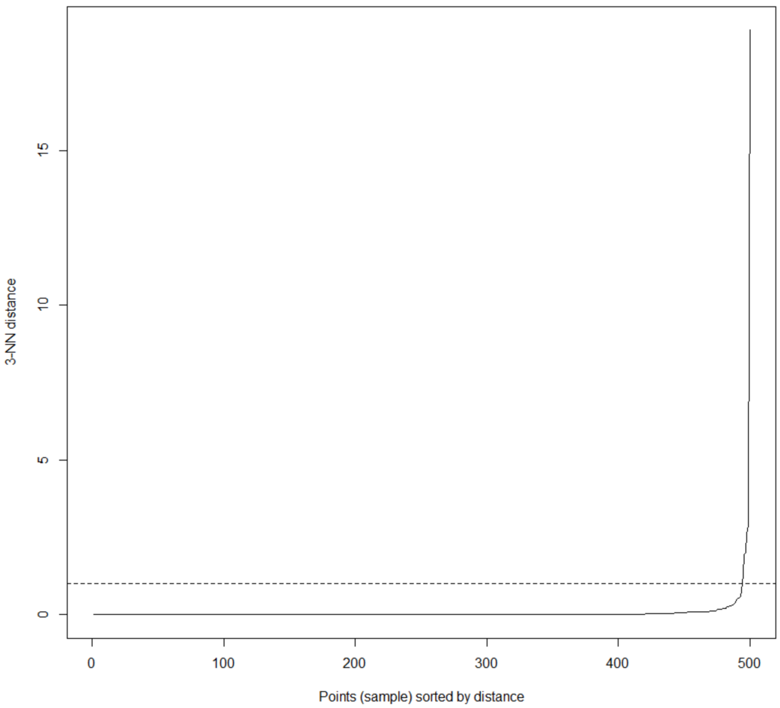

Figure 2 Defining the EPS value for DBSCAN

Even using a sample data set the elbow identification was not apparent and so an estimated 0.1 eps value was selected based on the output shown in Figure 2. In performance of the DBSCAN training process the following data was obtained and identified in Figure 3. Given the output and interpretation of the model the utility of using this method has some limitations for both interpretation and contextualization of fraudulent transaction identification.

```
DBSCAN clustering for 129091 objects.
Parameters: eps = 5, minPts = 4
The clustering contains 3 cluster(s) and 8 noise points.
```

0	1	2	3

8	129072	7	4

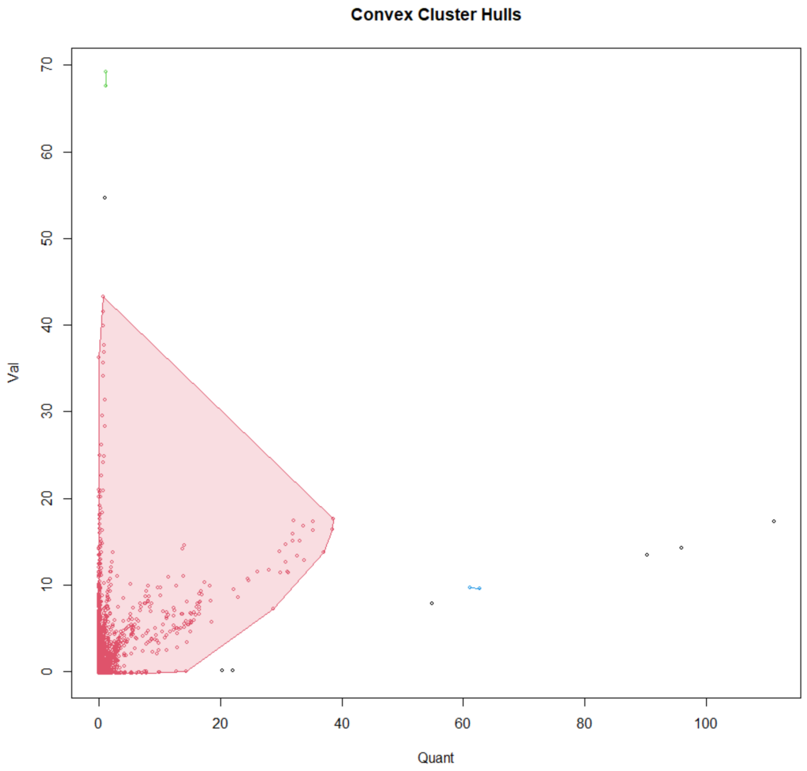

Figure 3 DSCAN training output and associated table

DBSCAN clustering for 15349 objects.

Parameters: eps = 5, minPts = 4

The clustering contains 10 unique cluster(s) and 7500 noise points.

0	1	2	13	21	22	24	26	28	40	47
7500	8015	1	3	1	1	7	6	1	1	7

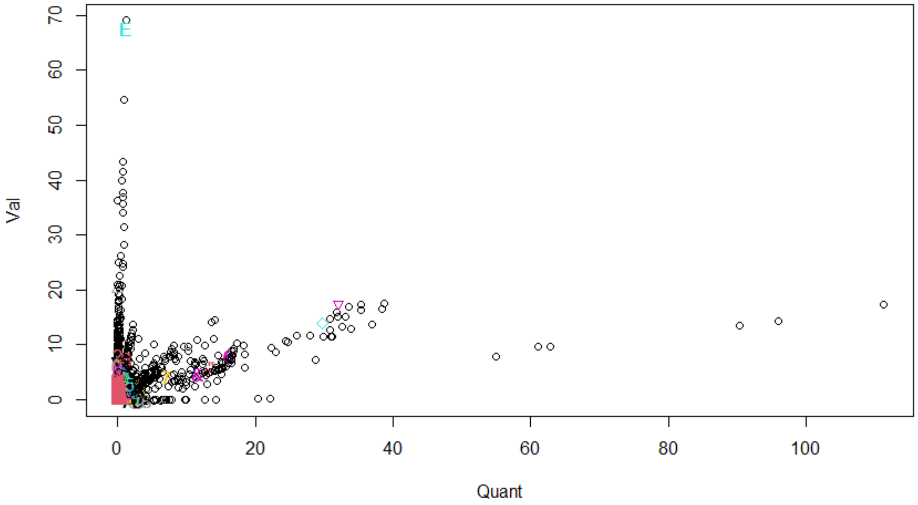

Figure 4 Prediction using DBSCAN and associated cluster identification table

From the figures, it can be determined that DBScan might not be the optimal method for outlier detection in this scenario. Figure 4 shows some clustering where a majority of transactions lie in the low quantity and low cost region. Outliers exist on the extreme ends of value and quantity. When using DBScan the memory requirement became a problem with the dataset. Although this dataset is large it is still relatively small which makes it difficult to scale up a DBScan model. From the test data table above it can be seen that the model identifies 7500 noise points, nearly half of the available transactions, which makes it difficult to draw any conclusions regarding outliers in the dataset. It also identifies one main cluster of 8,500 points and a bunch of other smaller clusters.

Local Outlier Factor (LOF)

Local outlier factor (LOF) was applied on the dataset to help determine fraudulent transactions. LOF takes a similar approach to DBScan in the sense that it uses a data point’s distance to a number of its nearest neighbors. LOF classifies a point as an outlier if it has a significantly lower density than its neighbors. For this reason, LOF yielded similar and difficult to interpret results. Due to the skew in the data the outliers are so few that they may get counted as noise or incorporated into the cluster. Since there are so many observations in the space signifying low quant and low val, it is difficult to capture local outliers due to the high density area (Figure 5).

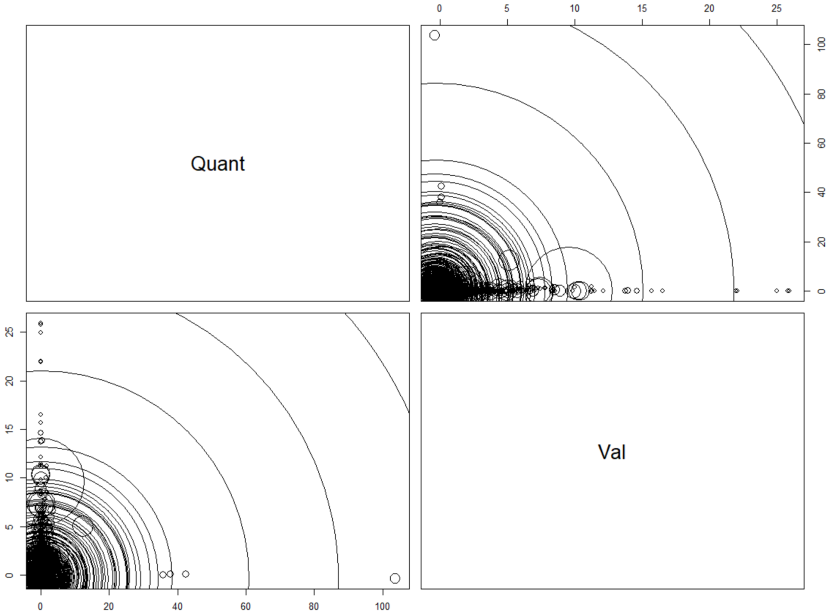

Figure 5 LOF output

Isolation Forest

The Isolation Forest method is used next to see if we can get a better picture of the outlier distribution without having to rely on distance/density measures. An isolation forest model is fitted to the training data which generates an anomaly score and average depth for each data point. Using a histogram of anomaly (Figure 6) scores, a trial and error evaluation was used to determine the cutoff point for whether an observation is an outlier or not.

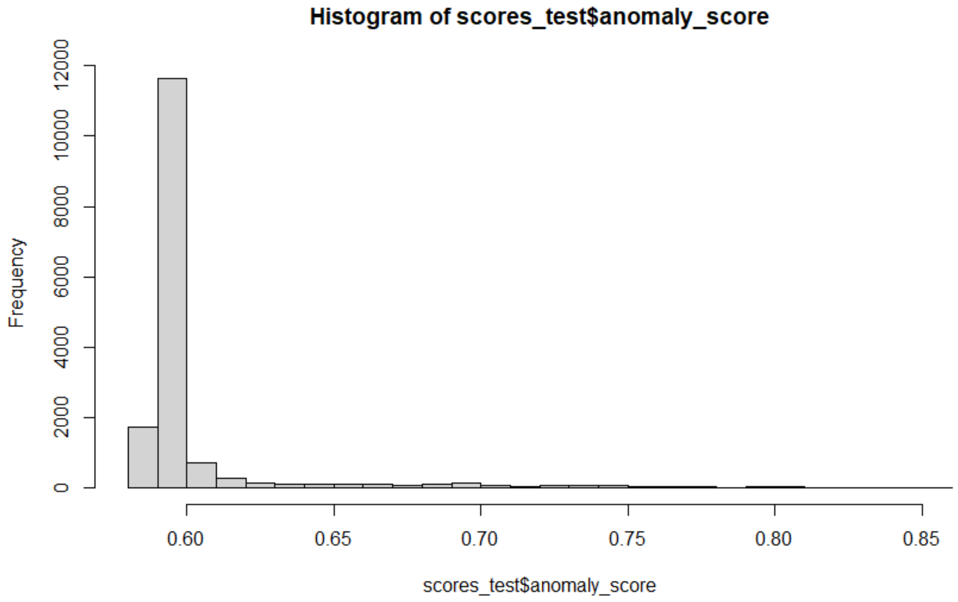

Figure 6 Histogram of resulting anomalies

In order to determine optimal cutoff score the following table was created to assess the overall fit of the predictive capabilities of the model all anomaly scores above 0.62 per Figure 6 are considered for determining the outliers within this analysis dataset. Using multiple iterations of the assessment model Figure 7 provides the cut off ranges and to lead us to determine the optimal approach for the underlying identification of those sales associates misreporting product sales. It is the resulting analysis where we want to reduce the number of sales that are fraudulent and are classified as legitimate transactions.

<table>
<tr><th>Fit</th><th>0.62</th><th>0.70</th><th>0.75</th><th>0.80</th></tr>
<tr><td><b>Confusion matrices</b></td>
<td><table><tr><th></th><th>0</th><th>1</th></tr><tr><td><b>0</b></td><td>13274</td><td>1059</td></tr><tr><td><b>1</b></td><td>1073</td><td>140</td></tr></table></td>
<td><table><tr><th></th><th>0</th><th>1</th></tr><tr><td><b>0</b></td><td>13999</td><td>1132</td></tr><tr><td><b>1</b></td><td>348</td><td>67</td></tr></table></td>
<td><table><tr><th></th><th>0</th><th>1</th></tr><tr><td>0</td><td>14029</td><td>1162</td></tr><tr><td>1</td><td>138</td><td>37</td></tr></table></td>
<td><table><tr><th></th><th>0</th><th>1</th></tr><tr><td><b>0</b></td><td>14317</td><td>1191</td></tr><tr><td><b>1</b></td><td>30</td><td>8</td></tr></table></td>
</tr>
</table>

Figure 7 Assessment of anomaly detection range

Addressing the element of reducing the false negative results from the analysis we can determine the utility of addressing the resulting false negative result for each of the cut off points. Using 0.80 as the cut off value provides the best FN result and limits the number of fraudulent transactions that are identified as ‘OK’. The tradeoff is the increased number of elements identified as fraud (false positives) but are actually legitimate transactions. Management will need to review these considerations in order to determine the trade off between these two factors of the analysis for identifying with greater accuracy fraud transactions but also the time to review transactions that will be legitimate but identified as fraud.

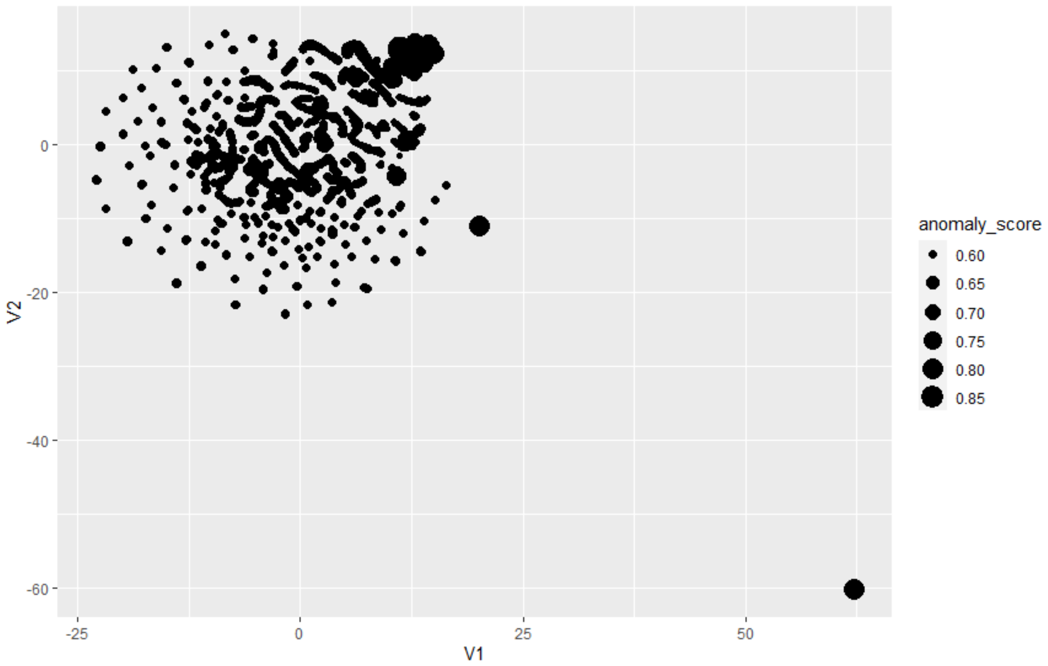

Figure 8 Isolation forest cluster analysis

The graphical interpretation of the resulting 0.80 result shows some clearly identified anomalies within the dataset and allows for identification of activity classified above the 0.80 cutoff value. The format of the graphic may not provide a method to determine the accuracy and utility of the underlying isolation forest predictive model so the utilization of confusion matrices are provided in the results section.

Results

In order to assess the predictive value of the isolation forest model the utilization of a confusion matrices and specific metrics for Recall, Precision and F1 are use to determine the

	Actual (OK)	Actual (Fraud)
Predict (OK)	14317 (TP)	1191 (FP)
Predict (Fraud)	30 (FN)	8 (TN)

$$\text{Recall} = TP/TP+FN = 14317/14317+30 = 13274/14327 = \mathbf{0.997}$$

$$\text{Precision} = TP/TP+FP = 143174\ /14317+1191 = 13274/14333 = \mathbf{0.93}$$

$$F1\ = 2/(1/\text{Precision}) + (1/\text{Recall}) = 2/(1/0.93)+(1/0.997)\ 2/\ 1.08+1.003 =\ 2/2.083 = \mathbf{0.96}$$

Figure 9 Confusion Matrix and associated metrics

Figure 9 provides the overall assessment of the model with the resulting cutoff score of 0.80. Elements for consideration when fitting the model and utility in building predictive capability is to understand the data used for training in order to assess the contamination of data to build and utilize the model. It is also important to understand the inherent bias based on the model's method of branching to create the iolations forest imputation and assessment of anomalies. Given the F1 score of 96% for the binary classification we have performed and show the account of the precision and recall of the model created. Given the management need to catch fraudulent transactions a Recall score is critical for building a method of identification.

Conclusion

The analysis has provided the organization insight into the utility of two unsupervised learning methods to detect fraud. The recommendation from Group 6 is the use of isolation forests. The analysis has shown the number and density of data will not be appropriately applied with the use of a density based scan approach and the resulting predictive analysis indicates the benefit and predictive capability of isolation forest technique. DBSCAN provided difficult to understand and less accurate modeling. On top of that DBSCAN also requires a tremendous amount of memory to operate. With this dataset many memory allocation errors were received. To scale DBSCAN up to a larger dataset would take a lot of time and memory space which may not be practical. Isolation forest yielded more reliable results. After experimentation with cutoff values a final F1 score of 96% was achieved. This in combination with the smaller memory requirement makes speed and scalability more plausible. For further research it is suggested that we proceed with isolation forest.

References

Alam, M. (2020, October 10). *DBSCAN-a density-based unsupervised algorithm for fraud detection*. Medium. Retrieved March 10, 2022, from https://towardsdatascience.com/dbscan-a-density-based-unsupervised-algorithm-for-fraud-detection-887c0f1016e9

Ajiboye, A. R., Akintola, A. G., & Ameen, A. O. (2015). Anomaly detection in dataset for improved model accuracy using DBSCAN clustering algorithm.

Vijayakumar, V., Divya, N. S., Sarojini, P., & Sonika, K. (2020). Isolation forest and local outlier factor for credit card fraud detection system. *International Journal of Engineering and Advanced Technology (IJEAT), 9*, 261-265.