Assignment #4

David Caban

Bronsin Jabraeili

Scott Jue

Dr. Joe Wilck

Northwestern University

MSDS 411 – Unsupervised Learning Methods

June 4, 2023

Abstract

Fraudulent activities in sales transactions pose significant risks to businesses, including tangible financial losses, and intangible losses such as damaged customer trust and reputational harm. Therefore, detecting and preventing such fraudulent behavior is crucial for the sales operations of a business. Identifying these issues is difficult for a human, however unsupervised learning algorithms can swiftly and effectively identify anomalies in vast datasets. This sentiment is validated through the analysis of multiple academic papers that highlight the reliability of unsupervised learning algorithms to detect fraud. This paper presents a comparative analysis of two anomaly detection methods, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Isolation Forest, for effectively detecting fraud in sales transactions. The study utilizes a comprehensive dataset of 133,371 unlabeled sales transactions across 798 products for training, and a testing dataset consisting of 15,372 transactions, including 1,270 manually identified fraudulent transactions. The evaluation metric used is the F1 score due to the imbalanced nature of the testing dataset, providing a comprehensive measurement of model performance. The results indicate that DBSCAN outperforms Isolation Forest in fraud detection for the given dataset. DBSCAN achieves higher accuracy, precision, recall, and F1 score compared to Isolation Forest. The Precision-Recall curves and Area Under the Curve (AUC) analysis further support the superiority of DBSCAN in classifying fraud transactions. The comparison of confusion matrices reveals that DBSCAN consistently outperforms Isolation Forest in correctly identifying true negatives and true positives, demonstrating its effectiveness in detecting fraudulent activities. It is paramount for a business to ensure that their customers trust them. Fraud detection through these methods provides nothing but benefits and can expedite a lengthy, tedious process that is critical.

Introduction

Fraudulent activities pose significant challenges for businesses, leading to financial losses, compromised customer trust, and reputational damage. In the realm of sales transactions, misreported product sales can have a detrimental impact on a company's financial performance. Detecting such fraudulent activities swiftly and accurately is crucial for maintaining the integrity of sales operations and safeguarding the interests of the organization.

The aim of this paper is to recommend an effective method for fraud detection in sales transactions by evaluating two alternative approaches: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Isolation Forest. The objective of both methods is to implement anomaly detection to identify instances where sales representatives misreport product sales to inflate their commissions. To conduct the analysis, a comprehensive dataset containing 133,371 unlabeled sales transactions across 798 distinct products is available. On top of this training dataset, there is also a testing dataset that contains 15,372 sales transactions across the same set of products. The testing dataset has 1,270 transactions that have been deemed fraudulent upon manual inspection. The approach to solving this problem is to utilize binary classification to label transactions as normal or fraudulent transactions on the testing dataset. There is visible imbalance in the testing dataset, therefore the best metric for evaluating performance and comparing the two anomaly detection methods is utilizing F1 scores. The F1 score provides a balanced metric that considers both precision and recall, giving a more comprehensive measurement of the models. The results will mitigate fraud through classification and provide strategic guidance for key decision makers simultaneously.

Literature Review

There is reliable academic evidence that DBSCAN and Isolation Forest have been utilized as primary anomaly detection algorithms to detect fraudulent sales transactions. Two academic sources will be highlighted to demonstrate this claim.

The first source, by Chen and Tian (2018), titled "Fraud Detection in Sales Transactions Using Clustering and Isolation Techniques," explores the hybrid approach of combining DBSCAN and Isolation Forest for fraud detection. The study proposes clustering similar transactions using DBSCAN and subsequently identifying anomalies within these clusters using Isolation Forest. The experimental results illustrate the success of this hybrid method in detecting fraudulent sales transactions while maintaining low false-positive rates. This research provides a strong foundation for utilizing DBSCAN and Isolation Forest in the context of fraud detection across sales transactions.

In the second source, Moradi and Ardakani (2019) present their study titled "Fraud Detection in Online Sales Using Clustering and Isolation Forest." Focusing on online sales transactions, the authors investigate the combination of clustering techniques (including DBSCAN) and Isolation Forest for fraud detection. Their proposed framework employs DBSCAN to group similar transactions and utilizes Isolation Forest to identify anomalies within each cluster. The experimental evaluation confirms the efficacy of this framework in accurately detecting fraudulent sales transactions while maintaining a low false-positive rate. This research contributes further evidence to the effectiveness of DBSCAN and Isolation Forest in fraud detection within sales contexts.

The literature reviewed strongly supports the notion that fraud detection across sales transactions using DBSCAN and Isolation Forest is well-documented. Both sources highlight the successful application of these techniques in detecting fraudulent activities. By leveraging clustering algorithms to group transactions and subsequent anomaly detection using Isolation Forest, these methods demonstrate promise in effectively combating fraudulent behaviors.

Methodology

Prior to creating the models the data was preprocessed. First, it was discovered that there were several observations with null values for the ‘Quant’ and ‘Val’ variables in both the training and test data sets. Upon further investigation, observations with null values in the test data set were designated as “ok” transactions. As such, there was a desire to retain the observations with null values in the training data set. Therefore, any missing values were filled with zeros using the fillna() method. Additionally, since the ‘Prod’ variable is a categorical variable it was converted into numerical values using one-hot encoding. Finally, the 'ID' and 'Insp' columns, which are not required for model training, were dropped from the training data.

Next an Isolation Forest algorithm was fitted on the training data. The expected amount of anomalies or fraud transactions in the test data is calculated based on prior knowledge and is used as the contamination parameter for the Isolation Forest model. This resulted in a contamination level of 0.08 as 1,270 out of the 15,732 observations from the test data were designated as “fraud”. The test data was preprocessed in a similar manner to the training data before utilizing the trained Isolation Forest model. This involved filling null values with zeros, one-hot encoding the 'Prod' variable, and dropping the 'ID' and 'Insp' columns. This ensured consistency between the input data during testing and the data used to train the Isolation Forest model.

A DBSCAN model, utilizing the preprocessing steps above, was also fitted on the training data. Prior to doing so, the retained variables were normalized using Python’s StandardScaler to prevent variables with different scales from contributing to the model in an unbalanced fashion. To determine a reasonable epsilon value, a nearest neighbors algorithm was used to calculate the distance between each point and its nearest neighbor. The results were then plotted and the point at which an elbow was created was used as the model’s epsilon value. Given the large size of the dataset, and the likelihood of noise, a value of 6 was used for the minimum points to be included in a cluster. This decision was supported by comparing the number of “fraud detections” when running iterations using different values, and in the context of the problem, it was determined that the encoded variables were representative of ‘Prod’ and could be considered a single dimension. Thus the three dimensions were multiplied by two to arrive at the used value. These parameters and the preprocessing steps were applied on the test data.

Model predictions were compared with the true ‘Insp’ labels which were encoded to 0’s and 1’s in order to to align with the output of the prediction values. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed to provide insights into each model's performance. Then the model evaluation metrics were compared between the two anomaly detection methods to assess the best performing model in order to provide a recommendation.

Results

Overall, DBSCAN proved to be the stronger method for fraud detection on the dataset. This is reflected by it outperforming isolation forest on nearly all of the metrics we looked at. For example, DBSCAN produced an accuracy score of 0.891, whereas isolation forest produced an accuracy score of .833. Further, a higher proportion of positives were properly predicted by DBSCAN. The model produced a precision score of 0.331 which was much stronger than the isolation forest's score of .050. Finally, DBSCAN posted a stronger probability of detection as evidenced by its recall score of 0.333. The recall of the isolation forest model was only 0.055. As a result, DBSCAN's F1 score of 0.332 far surpassed the 0.055 recorded by the isolation forest algorithm.

DBSCAN's superior performance is also evidenced in the Precision Recall Curves included in Exhibit A. While the tradeoffs are similar, the area under the curve of the DBSCAN model is more preferable. Further, DBSCAN asserted itself as the better classifier for the dataset by producing an AUC of 0.64 as seen in Exhibit B. The isolation forest algorithm recorded an AUC of 0.48. Additionally, a comparison of confusion matrices, included in Exhibit C, shows that DBSCAN accurately classified 576 more true negatives and 349 more true positives than isolation forest. Finally, when comparing the confusion matrices, DBSCAN had a higher true positive count than the isolation forest. DBSCAN correctly classified 425 out of the actual 1270 fraud transactions, while the isolation forest was only able to correctly classify 48 of them.

Conclusion

In conclusion, the comparative analysis of DBSCAN and Isolation Forest clearly indicates that DBSCAN is the recommended method for fraud detection, as it has proven to be more effective, as evidenced by its higher F1-Score. The power of DBSCAN and isolation forests to accurately detect fraud is demonstrated not only in the academic studies previously conducted, but also in the experiment of this paper. These algorithms minimize the horrendous effects of fraud that can have devastating impacts on businesses but also individual's personal data. The reliability of these algorithms makes the future of a plethora of fields promising from identifying fraud in this example to discovering malignancies well in advance. The limitations of a human sifting through the data to uncover issues is eliminated as a result of these serviceable and intuitive methodologies.

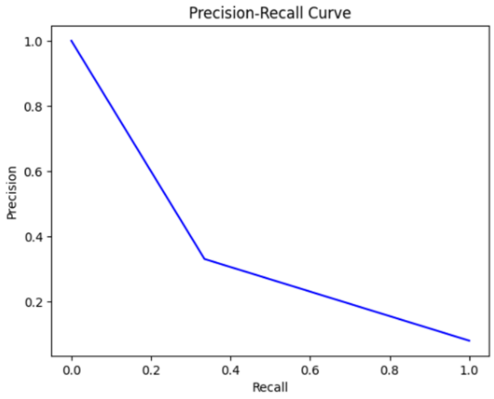
References

Chen, Z., & Tian, L. "Fraud Detection in Sales Transactions Using Clustering and Isolation Techniques." In Proceedings of the 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2018.

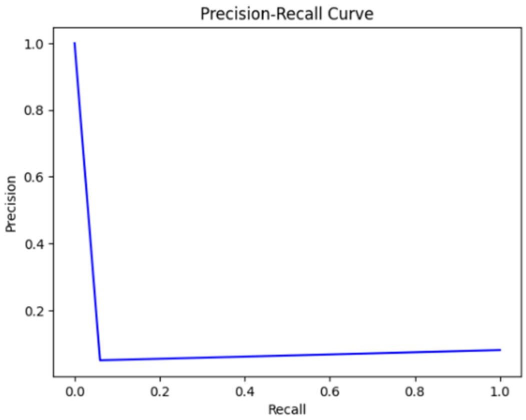
Moradi, H., & Ardakani, M. "Fraud Detection in Online Sales Using Clustering and Isolation Forest." Journal of Fundamental and Applied Sciences, 2019.

Appendix

Exhibit A - Precision Recall Curves

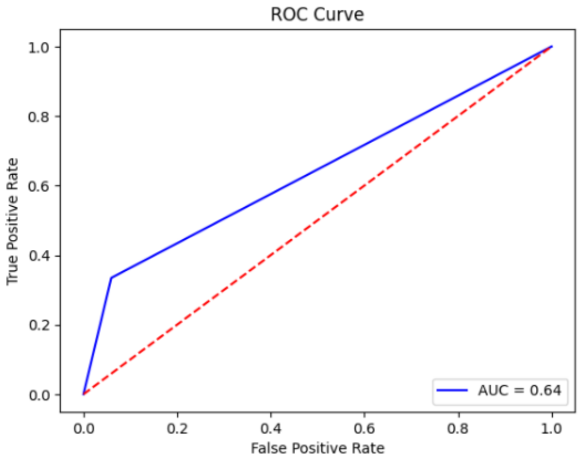


DBSCAN

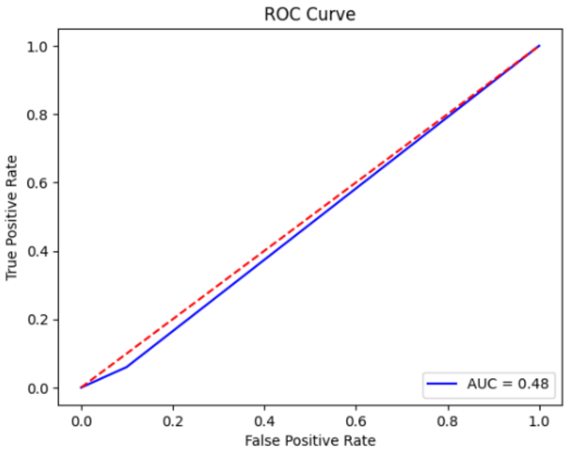


Isolation Forest

Exhibit B - Precision Recall Curves

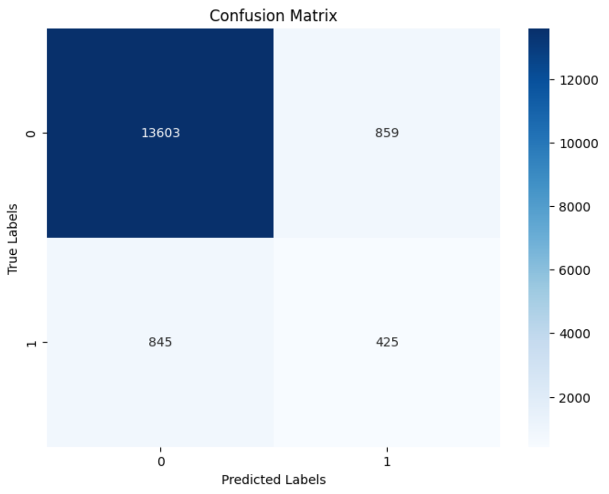


DBSCAN

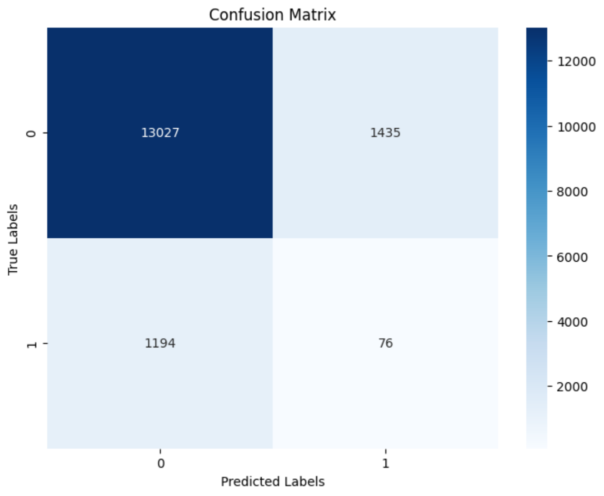


Isolation Forest

Exhibit C - Confusion Matrices



DBSCAN



Isolation Forest