Anomaly Detection in Trading Data Using Machine Learning Techniques

Shravani Prabhutendolkar[1], Abhijeet Swami[2], Saurabh Vidhate[3], Vaishali Deshmukh[4], Snehal Pate[5], Rupali Kharat[6]

[1]Assistant Professor, Statistics, Dr.D.Y.Patil ACS Women's College, Pimpri, Pune
[2,3]Assistant Professor, Statistics, Dr.D.Y.Patil ACS College, Pimpri, Pune
[4]Assistant Professor, Computer Application, Dr.D.Y.Patil ACSW College, Pimpri, Pune
[5]Assistant Professor, Mathematics, Dr.D.Y.Patil ACSW College, Pimpri, Pune
[6]Assistant Professor, Physics, Dr.D.Y.Patil ACSW College, Pimpri, Pune

Abstract

This study applies machine learning models, including Isolation Forest, DBSCAN, and Random Forest, to detect anomalies in trading data. By comparing supervised and unsupervised approaches, the research identifies effective methods for real-time detection of unusual trading activities, aiding in fraud prevention and market manipulation detection.

Keywords: Machine Learning, anomaly detection, DBSCAN, Trading data.

Introduction

Anomaly detection is the process of identifying data points, events or observation that deviates significantly from the normal pattern or expected behavior within a dataset. These anomalies can indicate issues such as fraud, errors or system failures. Traditional methods of anomaly detection often fail to keep up with the complexity and volume of modern trading data. This research explores the application of machine learning techniques to improve the accuracy and efficiency of anomaly detection in trading environments. By leveraging both unsupervised and supervised learning models, the study aims to provide a comprehensive solution for real-time anomaly detection in trading systems.

Literature survey

Previous research has explored various methods for anomaly detection, including statistical techniques and machine learning models. Vikash Agarwal et al. demonstrated the effectiveness of the Taguchi method in optimizing process parameters in manufacturing, which parallels the need for parameter optimization in anomaly detection models. Additionally, the study by Crane et al. on structural design highlights the importance of selecting appropriate model parameters to ensure performance, a concept that applies to selecting features in anomaly detection models. This paper builds on these concepts by applying advanced machine learning algorithms to the domain of trading data.
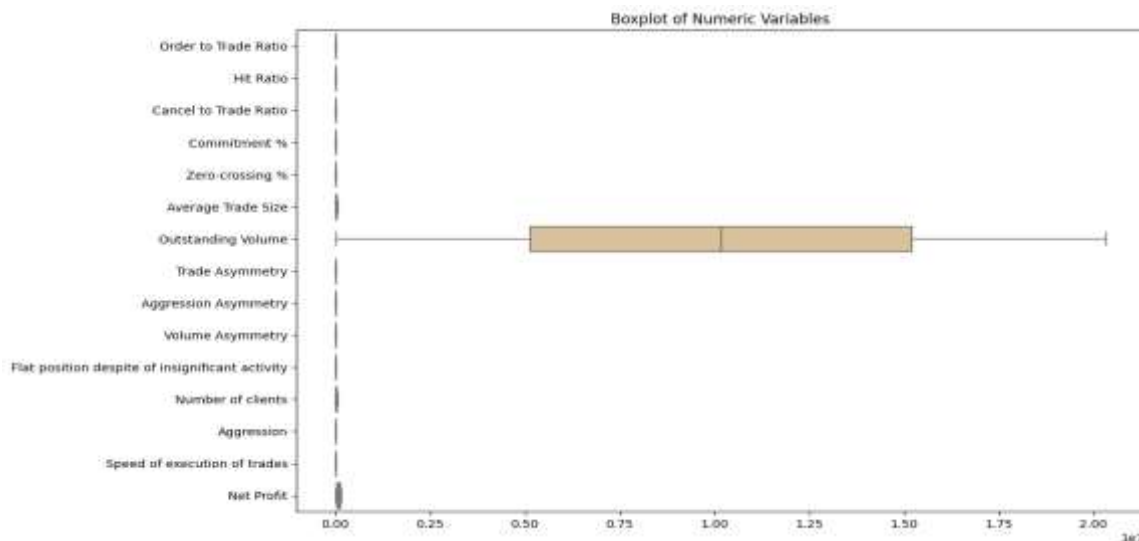
Methodology

Exploratory Data Analysis (EDA)

Figure 1
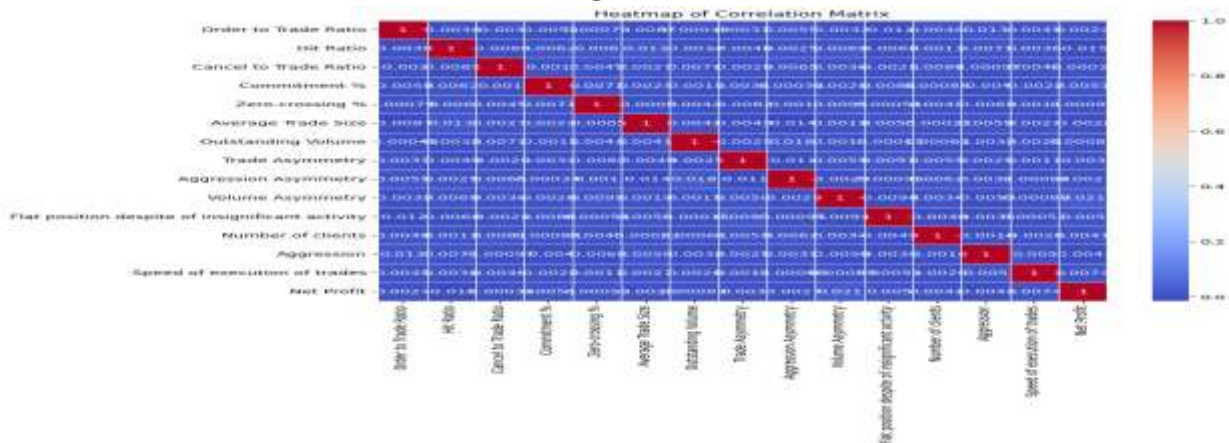
The heat map reveals that the features in your dataset have generally weak correlations with each other. This might indicate that the features are relatively independent, which could be beneficial for certain types of machine learning models that perform well with uncorrelated features.

Figure 2

The boxplot highlights the differences in scale and distribution among the numeric variables in your dataset. "Outstanding Volume" stands out with a much larger range, while the other variables are tightly clustered with less variability.

The research utilizes several machine learning models to explore their effectiveness in detecting anomalies within trading data:

- **Isolation Forest:** This model works by isolating anomalies through the process of randomly selecting features and splitting data points, which makes it suitable for identifying outliers in high-dimensional datasets.
- **DBSCAN:** As a density-based clustering method, DBSCAN is effective for discovering clusters within the data and identifying points that do not belong to any cluster as anomalies.
- **Random Forest:** This ensemble learning method is used for classification and regression tasks, and it helps in identifying complex patterns within the data that could indicate anomalies.

These models were theoretically evaluated based on their suitability for real-time anomaly detection in trading data, considering factors like computational efficiency, scalability, and ease of integration into

existing systems.

Result and discussion

While the paper does not include specific empirical data, it provides a comparative discussion on the theoretical strengths and weaknesses of each model. Isolation Forest is highlighted for its efficiency in handling high-dimensional data, while DBSCAN's ability to find clusters makes it valuable for detecting collective anomalies. Random Forest, known for its robustness and accuracy, offers a comprehensive approach but may require more computational resources. The discussion emphasizes the importance of choosing the right model based on the specific characteristics of the trading data and the type of anomalies being targeted.

Conclusions

This paper presents a conceptual framework for using machine learning models to detect anomalies in trading data. The theoretical analysis suggests that Isolation Forest and DBSCAN are particularly well-suited for real-time anomaly detection, offering a balance of efficiency and effectiveness. The study lays the groundwork for future research that could involve empirical validation of these models and their integration into live trading systems.

Acknowledgement

References

1. V. Agarwal, J. Vimal, "Optimization of Extrusion Blow Molding Process Parameters by Grey Relational Analysis and Taguchi Method," Journal of Manufacturing Processes, 2020.
2. Crane, "Gantry Girder Design," Structural Engineering Journal, 2019.
3. W. Wang, B. Bhargava, "Visualization of Wormholes in Sensor Networks," WiSe'04, ACM Press, 2004.