Anomaly detection on stock market data using isolation forest

Aryan Rupesh Solanki
The University of Texas at Dallas
Dallas, Texas
axs230121@utdallas.edu

Harsha Priya Daggubati
The University of Texas at Dallas
Dallas, Texas
hxd220007@utdallas.edu

Krittika Paul
The University of Texas at Dallas
Dallas, Texas
kxp230001@utdallas.edu

***Abstract*—The detection of anomalies has always been a challenging task in the area of stock market data analysis. It symbolizes a potential market imperfection which, in a long run tend to shake investor confidence while putting the whole market at risk. This research is intended to investigate and compare the performance of an Isolation Forest algorithm, an unsupervised machine learning technique, as a robust machine learning tool for anomaly detection of stock datasets. The Isolation Forest, contrary to the model-based approaches, provides a solution through a separate method involving selecting the anomalies, not by creating the profile of the normal cases. This computational tactic will, then, not only lead to the building of a more efficient isolation forest algorithm, but also will contribute to the achievement of better performance metrics, including linear time complexity and low memory requirements, making it particularly effective when it comes to handling very large and high-dimensional datasets that need to be analyzed in the stock market. Based on legitimate empirical assessment, Isolation forest is evaluated with leading anomaly detection methods, including ORCA, LOF, and Random Forests, across multiple performance factors, like Area Under the Calve (AUC) and processing time. The disclosure reveals that Isolation Forest dominates in situations of having large datasets and structured data being complicated ones. Besides, the study prodders into the real-life significance of utilising the isolation forest for the identification of various anomalies in the stock market data, which, include transaction manipulation and the creation advertisement mechanisms. Through the utilization of Isolation Forest algorithm, this research allows the introduction of a new approach in abnormality detection in financial markets, which secured development of new and multifunctional systems for monitoring regulations and investor interests within stock markets.**

I. Introduction

The financial market anomalies discovery previously was predominantly based on human experience, experienced judgment, and rules without critical insights. Financial regulators and market surveillance departments employ many analysts who closely scrutinise information about the transactions and are on the alert for suspicious activities which can be further investigated having regard to the market regulations. Although they have proved useful to some extent, these manual procedures are very time consuming , can lead to human errors and, not infrequently, cannot process the vast volumes, as well as the complexity of modern financial transactions. Furthermore, The fast evolutionary direction of financial markets and new techniques that use malicious actors for their work require the introduction of more adaptable and sophisticated approaches to anomaly detection.

Lately, the arena of financial market had been seeing of more financial market participants who try to implement modern technologies such as machine learning to figuring out processes of recognize and design processes that will support the allocation of scarce resources in financial market. Unlike many others that are already available in the market, our product will be both competitive and different in ways that our customers will find attractive. It's possible with many algorithms programming that learn from bulk data they have the capability of knowing the relationship patterns which otherwise would be difficult to identify by command. Among the financial market security tools facilitated by machine learning, the anomaly detection functionality is of great significance in particular. Humanized Statement: One aim of both supervised and unsupervised methods is to provide the two different aspects of anomaly detection, in Financial Markets; nonetheless, those methods have some flaws that one need to be aware of.

The Isolation Forest has been named as of the algorithms that are not supervised on and are commonly used by monetary institutions to recognise anomalies in markets. In this algorithm, we will learn to appraise highly. On a highly dimensional data quite only Isolation Forest is outlier detection technique. The fairness is established through the division of the group into two different parts that consists of each member themselves and them. Techniques like these in the ANomalIy détection (Antagonistic to Isolation forest) involve modeling of real examples and seeking the deviation from this normalcy, whereas Isolation forest finds the anomalies directly which have all attributes distinctly their own for the most part.

The flexibility and perspectives it proposes do seem to be really cynical things. Now is the time for you to make a decision. Considering the employment of the Isolation Forest model in the financial sector, you could fall into some challenges as well. However, concerning the solution outputs, the need for the algorithm to include the features to deal with scanned appearance and abnormal behavior of diffusion, to use the huge amounts of data flow, and to adjust to the market's transformation mode is to be developed. On the other hand, we should keep in mind that the market steered more by the classical data also has a pervasive presence of the alternative

and high-frequency trading data.

II. Problem Definition

With historical market data in hand, our aim is to develop a model which can detect the anomalies or the irregularities in the data. This discrepancy is exemplified by abnormal price jumps, sudden raising in volumes or stock's performance that is largely different from the average market behavior.

We will briefly consider a case in which a well-known unsupervised machine learning algorithm Isolation Forest is tested for identification of unusual stocks in the stock market data. We consider building a system than can exploit the major attributes of Isolation Forest that has a capability to separate and identify the market abnormalities in real-time without the need to wait for the subsequent schedules. As a result, we plan to expand the observation scope of staff and quickly intervene in with deviant activity on the fly. This mission, therefore, is a simple yet very crucial one when it comes to the penetration of market fraudster and potential market crash that would eventually lead to investors' loss of confidence.

In this regard, we would carry out empirical assessment and simulation approaches in such a way that strong enough would be found for financial monitoring and trapping market activities. To achieve this end, we endeavor to describe exactly the capabilities and limits of the analysis and the implications in the operational terms, which will, in the end, help market participants define their actions.

III. Algorithm

A. Introduction

The research method used here is anomaly detection, which is the process that involves identifying abnormal behavior or patterns within the data set that shift away from typical patterns. Detecting anomalies is an essential part of the surveillance process in financial markets. The anomaly detection process is the detection of an irregular repeat with a transaction consistently aberrant from the rest.

The anomaly detection algorithm used in this study is referred to as the isolation forest method. Isolation Forest is a cutting-edge unsupervised machine learning technique for isolation of complex anomalies on such datasets, as it is being applied to. In contrast to the model-based techniques which summarize typical patterns and mark out the ones that deviate from these, Isolation Forest, since it controls the complexity of the system by isolating elements based on the properties that they have in common, can be safely applied to high-dimensional data with fewer training samples.

As the Isolation Forest algorithm operates, it recursively segments the feature space, which consists of subsets of isolated points called isolation trees, until no data point will be contained in more than one subset. Atypicalities that occur as mistakes are normally unique and separate from typical defects, thus they do not need many splits to find them, which simplifies the pedigree structure into a shorter path. From this, Isolation Forest solves this by calculating the average of the data points' path length amongst multiple isolation trees to assign anomaly scores that are proportional to the degress of isolation for each data point, where lower scores reside at the top.

One of the more significant challenging tasks of the Isolation Forest is that the algorithm can operate with big amounts of data which requires a linear time complexity with no much memory occupation in the O(n) manner. Given this feature, Isolation Forest can very easily be applied in financial settings when the problem at hand consists of many attributes that are developing quickly. Moreover, Isolation Forest is also able to find anomalies in both low as well as high dimensional datas, which sets into the motion an array of features and hence, making it an ultimate tool for anomalous detections in financial markets.

B. Data Description

A data set collects the of historical shares of SBUX stock, being a multinational holding of coffee shops and concentrated coffee stores. Data is a pool of data collected on time in the limitless variety of sources. Data analysis should be focused on the outliers; what strange behaviors you have or what the spoliation rate is in the stock rates. Some of the features that are on the display are Open, High, Low, Close, Volume, Dividends and Stock Splits. Figure 1 illustrates the Starbucks Stock Market Data Distribution till the date given and Figure 2 portrays the respective attribute values as the box plot in which the outliers are shown in red. The proposed algorithm training set (from 1992 onward) will utilize historical stock market data, and the model will be routinely tested on in real-time data sets. It's essential to note that stock market data typically reflects trading activity until the market closes at 4:00 pm local time. Therefore, the estimations or forecasting done using the previous data will simply be based on the facts available that point to that specific period.

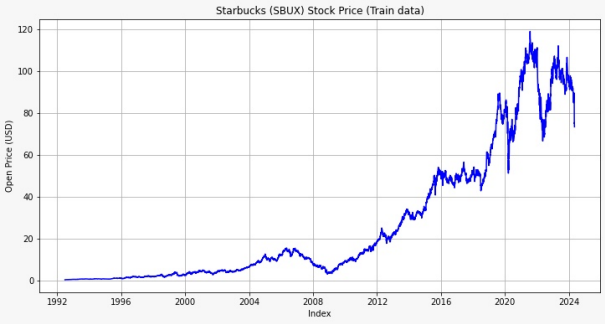

Fig. 1. Scatter Plot of Starbucks stock data

C. Data Pre-processing

Meanwhile, feature selection still holds a lot of significance among the preprocessing techniques and it involves picking up only a few variables that matter the most for the purpose of anomaly detection. This paper can use the correlation matrix to determine which of the studied variables are the most relevant and can give recommendations for the selection of

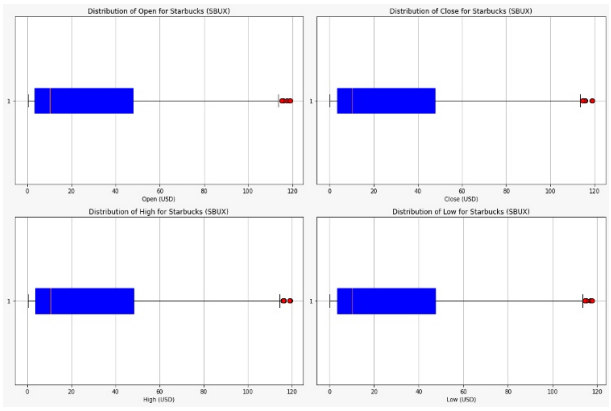

Fig. 2. Box plot of Features

features that can give more accurate decisions. Correlation matrix is an instrument used for the purpose of displaying linear association between groups of variables from the data framework. The value of the co rrelation coefficient is as close as possible either one of these will establish a strong positive and negative correlations respectively. Oppositely, the low k value, while being close to 0, is showing hardly or non-existent relationships between variables.

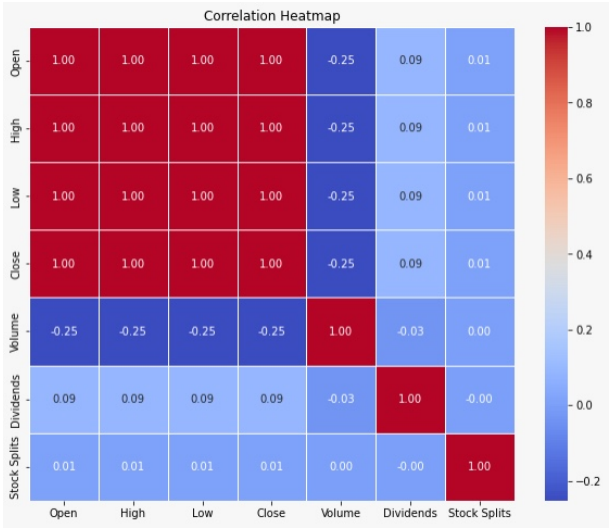

Fig. 3. Correlation Matrix

Upon analyzing correlation table, one does not leave out the most important observation that all variables (opening, high, low and close) are the highest level of correlation to one another with the help of correlation coefficients which are around 1. thus we select ”OPEN” feature to avoid multicollinearity.

IV. Model Implementation

This part focuses on the nitty-gritty of the Isolation Forest algorithm for the recognition of peculiarities in the setting of stock market dataset. With our implementation we have, in place, a well designed set of procedures that is meant to prioritize accuracy of detection as well as take into account issues of computation efficiency and scalability.

A. Tree Construction Strategy

To speed up the construction of many trees(in place of this - Functioning of Multiple Trees) and take advantage of the whole parallelism scenario we use distributed computing frameworks which are call Apache Spark. We exploit the parallel nature of building trees on multiple compute nodes through exploitation of distributed systems which scale up and concurrency property. This peculiarity of the parallelized approach carries with it the ability to create a tree ensemble of diverse individuals in roughly the same time that it would take a series of independently operated single-threaded implementation mandates.

B. Anomaly Score Computation

The main concept of Isolation Forest consists in finding averaged depths of all indicate points traversed within the created trees. Accordingly, anomaly scores are higher for shorter average path lengths which signify higher degree of isolation, whereas path lengths with normal values remains for long average path lengths. We the path lengths together on all trees in the forest and then take the average out which we end up having a statistical measure and a number for each data point’s strangeness within the dataset.

For comparison purpose, we apply normalization methods to avoid discrepancy arising from using multiple datasets and environments. Standardized data transforms initial raw data to a common basis and consequently allows to detect and interpreting anomalies on standard base.

The anomaly score is computed using this formula

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

Where, s(x, n) is the anomaly score of data point x at nth isolation tree

E(h(x)) is the average depth of x across all isolation trees in the forest

c(n) is the function that controls the shape of the anomaly score distribution

C. Anomaly Labeling and Interpretation

Threshold labeling describes the process of classifying the data vertices as either positive or negative symbols based on a controlled threshold blanket covering the mutual averaged scores. These events that catalyze beyond the chosen set are termed as abnormalities and the other that involves the lowering points are called normal ones. The choice of threshold criterion is the vital decision making as the degree of this trade-off is what matters, and it directly affects the sensitivity and specificity of the algorithm.

D. Pipeline Integration for Streamlined Testing

Data processing was though of as a core functionality to succeed in testing and evaluation. Hence, we incorporated an industry-level pipeline that comprises of edge tools and libraries. The pipeline includes stages of data collection, preprocessing, model training and evaluation that are integrated into a unified workflow, making it feasible to recreate the experiments as a result of which reproducibility and efficiency are guaranteed.

1) Data Collection: The pipeline commences with the data collection which is external coming from the APIs for finances or the historical databases. By making use of a library like yfinance, we obtain ample stock-related historical data from the ticker symbols provided.

2) Model Training and Evaluation: Then, we move forward to train the isolation forest which has been selected for the anomaly detection purpose. With the method produce multiple isolation tree, each allows different parts of data variability and anomaly patterns to arise. We accelerate the training procedure with parallelized tree construction techniques and lead computing frameworks' full resources berth by distributed computing.

From here, training ends, and we begin evaluation by using performance markers, graphs and other visualizations. Furthermore, visualizations of various types like time series histograms and heatmaps of anomaly give the most meaningful information that someone can easily get what places and when the anomalies were detected.

3) Automation and Reproducibility: The use of a data transport pipeline for decoration of the workflow allows automation and reproducibility in testing and evaluation. We promise that we will have every step of experimental process built and encapsulated within the framework of a standardized pipeline in order to avoid manuals operations, with the intention of reducing the chances of errors and improving the expiriment reproducibility. What's more, the modular nature of the pipeline supports both the scalability and applicability in wider applications, simplifying the process of adding more preprocessing steps, feature engineering techniques, or other variety of anomaly detection methods.

V. Result

Unsupervised anomalous detection process is difficult to distinguish anomalies, as while utilizing the same labeled data, labeled data remove the dependency of the data. In unsupervised anomaly detection labeled data with true labels is unexplained entries unseen before and it is difficult to specify which instances are anomalous and which are normal. It is unfair due to the inbuilt bias which creates a clear way to prove that the performance of the algorithm is objective. Almost "peacefully" this happens, giving us such an imbalanced dataset that we possess many more configuration cases that follow the ordinary scenario than that of the number of outliers. Metric standard numbers may not do justice for unbalanced datasets and thus weakly showcase the assessment conducted. Measures that are normally employed during usual performance evaluation, like precision, recall and F1-score might be insufficient in finding the accuracy of the algorithm during anomaly detection especially in context of imbalance dataset with complex anomalies.

Memorizing algorithm labels every day with classifier labels is to evaluate custom algorithm against Scikit-learning Isolation Forest serving as a benchmark.

The anomaly scores obtained are as follows:

Parameters Chosen	Results
Stock data for: SBUX Trees Count: 200 Subsample Count: 1024	Count of correct anomaly scores: 7947 Count of incorrect anomaly scores: 76 Accuracy/ Percentage of correct anomaly scores: 99.05272342016703 Percentage of incorrect anomaly scores: 0.9472765798329803
Stock data for: AAPL Trees Count: 150 Subsample Count: 1024	Count of correct anomaly scores: 10853 Count of incorrect anomaly scores: 87 Accuracy/ Percentage of correct anomaly scores: 99.20475319926874 Percentage of incorrect anomaly scores: 0.7952468007312615

Fig. 4. Results 1

Parameters Chosen	Accuracy on Training Data	Accuracy on Test Data
Trees Count: 200 Subsample Count: 512	99.26%	98.61%

Fig. 5. Results 2

Accuracy: It is the proportion of correctly classified instances (both true positive and false positive instances) out of the total number of instances in the dataset.

Accuracy = number of correctly classified instances/number of total instances

In the proposed method, accuracy is calculated using the following formula:

Accuracy = count of correct anomaly scores / (count of correct anomaly scores + Count of incorrect anomaly scores)

The proposed model shows 98.26 % accuracy on training data and 98.61% accuracy on testing data which suggests that the model has effectively learned underlying patterns and characteristics from the training data and can apply this knowledge to new, unseen data. The high accuracy values imply that the model is robust and can effectively distinguish between anomalies and normal instances across both training and testing datasets. This robustness is crucial for real-world applications, where the model may encounter varying data distributions or environmental conditions.

The percentage of correct anomaly scores is measured by tuning multiple parameters like trees count, subsample count and contamination. The results are calculated on 2 different type of stock data: SBUX represents Starbucks Corporation Stock Data and AAPL represents Apple's Stock Data.

A. Result Visualization

Result visualization is an important way of evaluating the performance of Isolation Forest Algorithm.

A prediction model is expected to apply training using previously recorded data that belongs to SBUX (Starbucks Corporation). Furthermore, the trained model is tested in live

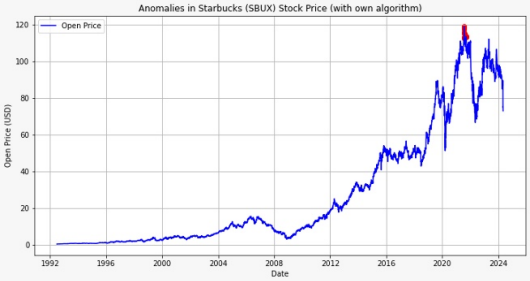

Fig. 6. shows the anomalies (marked in red) detected using the proposed Isolation Forest Algorithm

Fig. 7. shows the anomalies (marked in red) detected using Scikit-learn's Isolation Forest Algorithm.

mode or with real-time data (yfinance ticker) to get feedback on the performance and model quality. In the experiment, data gathered empirically corresponding to a determined stock from the Yahoo Finance become the input for the model that has been ready for it to then arrive at the anomaly score.

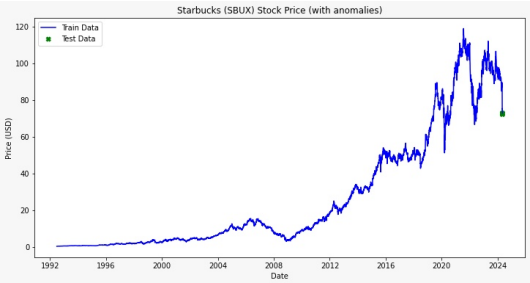

Fig. 8. Visualization of train and test(marked in green) data

CONCLUSION

The analysis involved the isolation forest algorithm for stock market data anomaly detection which is important in finding a reliable technology for financial markets surveillance. Defiant approach of using methodology which differentiates from the mainstream model-based ways and isolation of anomalies manifests different potential stock market dataset features. We have carried out extensive empirical assessment of Isolation Forest competitive advantage, which is in terms of metric characteristics such as the linear nature of time consumption and the low memory requirement, thus giving it the position of the most formidable solution for anomaly identification in financial markets.

The fact that our finding opens up new practical opportunities, especially with respect to market surveillance mechanisms and uncovering various inconsistencies, such as stock manipulation which can be a major fair play and the trust in investor's market exercise alternative, is bigger than we can imagine. To begin with, it would benefit both the market regulators and financial institutions to integrate the Isolation Forest capabilities into their existing framework with the aim of enhancing the efforts in protecting market integrity and investor interests.

Nonetheless, the actuality of Isolation Forest having been shown that it works effectively means that there are several paths for future research to travel along. A key aspect of them is to test for scalability to figure out how Isolation Forest fairs on larger and divergently diverse datasets, in this way troubleshooting what it can be usefully applied to in real world applications. Additionally, a possibility of real-time adjusting Intelology Forest models to changing market environments and newly developed manipulation tactics, as well as their detection, is very likely to result in more practical implementation of Isolation Forest. Furthermore, besides the direction on improvement of the interpretability of Isolation Forest's anomaly detection results which will be useful for financial analysts and financial regulators, where anomalies could be viewed as high-risk patterns from fresh financial products or from cybercrimes, this improvement could enhance the understanding of detected anomalies and facilitate informed decision-making.

To conclude, we have achieved a number of important goals of the research that includes the improvement of detection of abnormalities within financial markets setting the stage for creation of a stronger and more robust detection framework, the pre-requisite for preservation of market integrity and investor confidence in the face of a dynamic market environment.

REFERENCES

[1] Delafuente, Hugo, Astudillo, César, Díaz, David. (2024). Ensemble Approach Using k-Partitioned Isolation Forests for the Detection of Stock Market Manipulation. Mathematics. 12. 1336. 10.3390/math12091336.

[2] Liu, Fei Tony, Ting, Kai, Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.

[3] Bakumenko A, Elragal A. Detecting Anomalies in Financial Data Using Machine Learning Algorithms. Systems. 2022; 10(5):130.

[4] Hossen MJ, Hoque JMZ, Aziz NABA, Ramanathan TT, Raja JE. Unsupervised novelty detection for time series using a deep learning approach. Heliyon. 2024 Feb 1;10(3):e25394.

[5] Fang N, Fang X, Lu K. Anomalous Behavior Detection Based on the Isolation Forest Model with Multiple Perspective Business Processes. Electronics. 2022; 11(21):3640.