Detecting Anomalous Market Activity Using Outlier Detection and Time-Series Analysis on S&P 500 Data

Kyle Mollard - 301560261

1. Project Overview

The objective of this project is to identify anomalous trading patterns within the S&P 500 dataset by applying data mining techniques in conjunction with time-series analysis. We aim to leverage clustering and outlier detection methods to detect irregular movements in stock prices and trading volumes, potentially indicating significant financial events or irregular market behavior. This analysis will use historical daily price and volume data from individual S&P 500 stocks and compare these patterns against broader market movements. Lastly, we will compare the results of methods introduced in this class, time-series methods applied to stocks independently, and a combination of both if time permits to find the best method(s).

2. Problem Definition

Financial markets experience periods of volatility due to various economic, political, and internal factors, leading to sudden and atypical price changes. This project aims to detect these anomalies within the S&P 500 dataset by using clustering and outlier detection to pinpoint specific days or stocks with abnormal trading behaviors. Additionally, the project will explore whether these anomalies align with or deviate from broader market trends.

3. Dataset Description and Justification

Dataset Source: The Kaggle dataset includes historical data on S&P 500 stocks, providing daily records for individual stocks, as well as the overall S&P 500 index.
https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks

Dataset Components:

Stock Data (***sp500_stocks.csv***): Contains daily price, volume, and other metrics for individual stocks, enabling time-series analysis on each stock.

Company Data (***sp500_companies.csv***): Includes sector, industry, and descriptive details for each company, which will aid in clustering and contextualizing anomalies within sectors.

Index Data (***sp500_index.csv***): Contains daily S&P 500 index values, facilitating a comparison between individual stock anomalies and market-wide behavior.

This dataset combination supports a comprehensive analysis through clustering, outlier detection, and time-series modeling, allowing for a detailed exploration of individual and market trends.

4. Project Tasks and Methodology

4.1 Data Preprocessing

Handling Missing Values: Use interpolation or removal techniques to manage any gaps in stock prices or volume data.

Normalization: Standardize the price and volume data to ensure consistency in clustering and outlier detection.

Feature Engineering: Generate new features such as daily returns, moving averages, and volatility measures to enrich the time-series analysis.

4.2 Exploratory Data Analysis (EDA)

Data Visualization: Visualize time-series data for each stock, analyzing patterns, trends, and seasonal effects.

Correlation Analysis: Examine correlations between features like volume and price changes to understand their relationships and potential indicators of anomalies.

Sector Comparison: Use the ***sp500_companies.csv*** data to assess sector-based differences in price and volume patterns, identifying industries with higher volatility.

4.3 Clustering

Clustering Algorithms: Apply K-Means and DBSCAN to group stocks by sector or similar behavior patterns over time.

Dimensionality Reduction: Use PCA or t-SNE for 2D visualization of clusters, making it easier to identify unique groups within the data.

Performance Evaluation: Evaluate clustering quality using Silhouette Score and Calinski-Harabasz Index to determine the most appropriate algorithm.

4.4 Outlier Detection

Methods: Implement Isolation Forest and Local Outlier Factor (LOF) to detect anomalous days with extreme price or volume changes.

Visualization: Plot outliers over time to observe when these anomalies occur, highlighting patterns in the dataset.

Analysis: Evaluate the nature of these outliers to determine if they align with significant market events or individual stock behavior.

4.5 Time-Series Analysis

Stationarity Check: Apply the Augmented Dickey-Fuller test to check for stationarity and transform data as necessary.

ARIMA Modeling: Develop ARIMA models to forecast stock prices and residual analysis, identifying time-series anomalies not captured in other techniques.

Comparison with Index Data: Use the sp500_index.csv data to compare stock-specific anomalies with broader market trends.

4.6 Model Evaluation and Interpretation

Anomaly Comparison: Compare anomalies detected through clustering and outlier methods with ARIMA residuals and broader market movements.

Effectiveness Assessment: Determine the robustness of each method in detecting meaningful anomalies, considering limitations due to the absence of real-time external factors.

Conclusion: Summarize the results, highlighting insights on market trends and limitations of the dataset for predictive analysis.

5. Hyperparameter Tuning

Optimization: Use Grid Search for tuning Isolation Forest and LOF parameters to maximize the accuracy of outlier detection.

Evaluation: Document the performance improvements before and after tuning, reflecting on their impact on the model's ability to detect anomalies.

6. Conclusion

Project Insights: Provide insights into the types of anomalies detected within the S&P 500 stocks and how they relate to sector behavior and broader market trends.

Methodological Reflections: Discuss the effectiveness of clustering, outlier detection, and time-series methods for financial market analysis.

Future Work: Suggest potential enhancements, such as incorporating real-time news data or sentiment analysis, to further contextualize market anomalies.

7. Deliverables

Final GitHub Report: A comprehensive report detailing each project task, from data preprocessing to model evaluation, including visualizations and metric results.

Code Submission: Documented Python code within a GitHub repository.

Presentation: A poster summarizing the project, key methods, and insights, with visualizations of the clustering and anomaly detection findings.