CSE 575

Anomaly Detection Final Project Report

Team Members:

Zaid Buni

Luis Perez Lemus

Chun-Wei Hu

Avinda De Silva

Table of contents

Introduction.....2
Problem Description.....2
Methodology..... 3
Results.....7
Conclusion..... 10
Future Work.....11
References..... 12

Introduction

The stock market is a vital part of the worldwide economy. Millions of shares are traded daily and prices are continuously changing. The primary objective of this initiative is to increase the ability of anomaly detection based on stock market analysis. By customizing features to build training models for specific industries, the aim is to bolster precision in identifying and understanding anomalies that permeate the market. This approach will not only enhance anomaly detection but also enable a more nuanced correlation between real-world events and fluctuating stock prices, thereby providing deeper insights into the dynamics of the market.

The focal benefits derived from this enhanced anomaly detection methodology are twofold. Firstly, the precision achieved in detecting anomalies within various industrial features will allow us to have a more detailed understanding about the abnormal events happening in the stock market. Secondly, the ability to find the relation between these anomalies with real-world events will enable stakeholders to connect the real-world factors and the market fluctuations together, thus providing invaluable insights for strategic decision-making.

Problem Description

The challenge at hand involves the comprehensive analysis of stock data to discern anomalies within the dynamic stock market landscape. These anomalies manifest as significant price fluctuations, abnormal trading volumes, or potentially fraudulent transactions. They require a more detailed examination of time-series data across diverse industrial sectors. To achieve this, advanced machine learning models such as Facebook Prophet are created for not only unsupervised anomaly detection techniques, but also predicting the future flow. In instances where a predefined dataset delineating anomalies isn't available, the model is trained iteratively using available data.

The primary goal is to identify anomalies occurring over a specified duration and subsequently delve into auxiliary data sources to unravel their root causes. For instance, when anomalies arise within the agricultural stock sector, the analysis extends beyond stock data to scrutinize weather patterns. This multifaceted approach aims to ascertain whether weather conditions have influenced crop yields, thereby impacting agricultural stock prices. Additionally, leveraging news datasets is integral in identifying similar occurrences, enabling the correlation of external events with market irregularities for a comprehensive understanding of these anomalies.

In addition, enhanced anomaly detection systems include industry segmentation, allowing for a sector-specific analysis that caters to the unique characteristics of each industry. This segmentation facilitates a more focused examination, improving the accuracy of anomaly detection within particular sectors. Additionally, the system incorporates sophisticated anomaly identification and visualization tools to present anomalies in a comprehensible format, aiding in their interpretation and understanding. Moreover, the utilization of custom Prophet models tailored to the specific nuances of different industries ensures a more refined and accurate analysis, enhancing the overall efficacy of anomaly detection within the stock market.

Methodology

In order to understand the intricate dynamics of the stock market, a meticulous approach to data collection, preprocessing and analysis is required. In this study we present a detailed methodology which was employed to handle the NASDAQ time-series stock dataset that spanned from 1960 to 2020. The dataset encompasses various attributes including opening price, closing price, high price, low price, volume and more.

Data Processing

To streamline the data for meaningful and important analysis, a systematic pre-processing pipeline was implemented.This pre-processing can be broadly divided into three parts:

Trimming Data: The original dataset had a vast amount of data and information to choose from. In this project we chose to focus solely on the date and closing price attributes. Additionally, we also selected a smaller data range, in order to predict anomalies better. The original dataset ranged from 1960 to 2020, for this project it was decided to choose a set between 1990 to 2020.

This simplification allowed for a more straightforward analysis while retaining the essential linear features.

Industry Segmentation: To enhance granularity, the data was segmented into their respective industries. This facilitates a more nuanced understanding of trends and anomalies that could be detected within specific sectors.

Averaging Normalization: Closing stock prices were averaged within each industry and provided a representative value. Normalization was then applied to standardize the values among all industries, which greatly reduced the error and mitigated the impact of varying scales amongst different companies.

In order to successfully post-process the data. The intended goal was aimed to identify the anomalies from the given data and extract valuable insights as follows:

Graphical Analysis: Using visual representation, in the form of graphs, we were able to use the detected anomalies data and visualize when and how these occurred. These graphs provided a comprehensive overview of the trends of the certain industry and highlight any deviations from the norm.

Anomaly Clustering: Clusters of anomalies were identified which prompted a detailed investigation as to the causes of these deviations. Each clustering of anomalies was examined to discern patterns and outliers.

Historical Event Attribution: For each identified anomaly cluster, we would analyze anomalies and find historical events that coincided with them. This was done by cross-referencing external sources and market reports to attribute anomalous behavior to specific occurrences such as economic crises, company/industry specific events or regulatory changes.

As the landscape of data analytics evolves, the applications such as advanced forecasting models become imperative when it comes to the extraction of meaningful insights. One model on the forefront is Facebook Prophet's forecasting and detection model. This versatile tool is important not only for accurate prediction but also identifying anomalies within time-series data.

Facebook Prophet Model

The core of Prophet's methodology derives in its use of an additive linear model on non-linear data. This allows the data to be broken into separate components such as seasonality or trends as separate entities. This approach is a specific application of project pursuit allowing the model to handle complex, non-linear patterns in real-world data.

$$f(X_1, \dots, X_p) = \sum_{j=1}^{p} f_j(X_j).$$

Additive Linear Model Equation (Projection Pursuit)

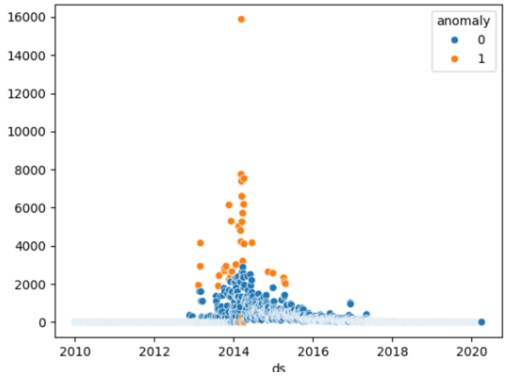
Prophet employs the regression model stated above to predict trends and skews in real-time data. This is done by separating the time-series data into various components and by studying historical patterns it predicts future trajectory.

The anomaly detection capability of Prophet is seen from its ability to recognize deviations in the data from the expected patterns. The model compares predicted values with the observed data and flags instances where the data diverges. This approach to anomaly detection enables timely responses to unexpected events.

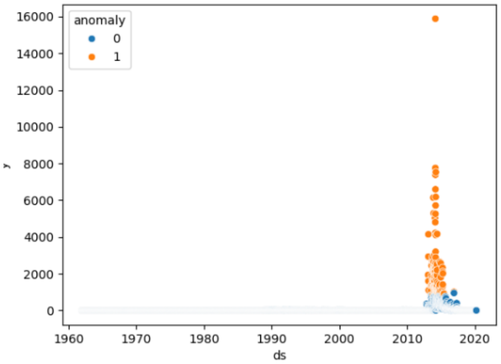
Error Reduction

In order to reduce the overall error of our dataset and predictions, we implemented multiple adjustments such as data normalization and a smaller data range to reduce error rates.

Refining Data Range: In order to reduce the source of errors, the data range was initially narrowed down from 1960-2020 to 2010-2020. This adjustment allowed for a smaller Mean Absolute Percentage Error (MAPE) however, was still high and certain outliers were still skewing anomaly detection. The next pre-processing step we took was to normalize the data set and increase the date range to 1990-2020.

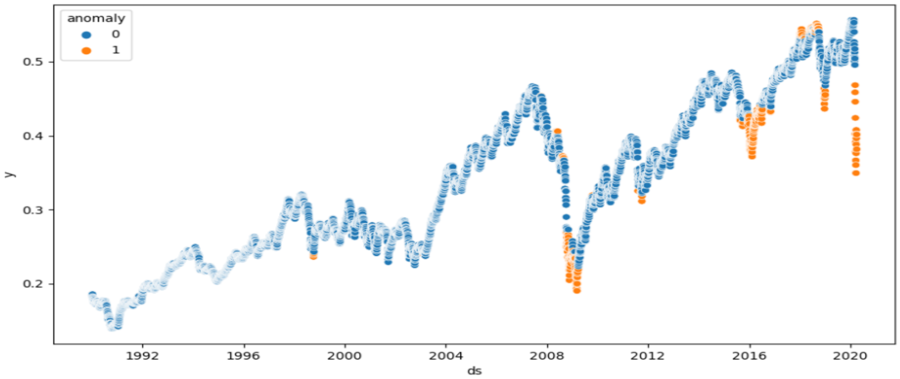


By running anomaly detection on the whole dataset, we get a Mean Absolute Percentage Error of 9000% error



By reducing the date range from 1960-2020 to 2010-2020, we still have a Mean Absolute Percentage Error of 5600%

Data Normalization: After reducing the range, it was still obvious that the MAPE was still significantly higher than intended and more modification needed to be implemented in order to properly identify anomalies. The combination of both techniques is what allowed for the greatest reduction in error. Additionally, with these modifications the prophet model was now successfully able to detect anomalies within the time-series data.



To improve the Mean Absolute Percentage Error, we found that normalizing the data as well as reducing the date range to be from 1990-2020 greatly helped.

Results

After preprocessing our data and feeding it into our model. We were able to successfully detect anomalies. But to ensure its accuracy we have to look at historical events that occurred during that time period to evaluate whether the model did detect a significant event in the dataset. In order to ensure further accuracy, we fragmented the stock market dataset into its respective stock sectors i.e. technology sector stocks, financial sector stocks, real estate sector stocks, etc. This fragmentation allowed us to discover more anomalies and attribute said anomalies to their historical event.

We selected four sectors for the purpose of avoiding redundancy. The blue points on the graph depict normal stock data, whereas the orange points depict data identified as anomalies by our machine learning model. We then decided to circle clusters of anomalous data points and try to determine the cause of the anomaly cluster. Our model is tuned in such a way that only detects significant anomalies since our dataset spans for 30 years, so we use historical events to discover if the anomaly detection was accurate.

Unique Anomalies:

Basic Materials Sector:

- Commodity bubble (pre-2008): The basic materials sector experienced a commodity bubble, where prices soared due to increased demand and speculation, followed by a significant decline.
- 2011 US Debt Ceiling Crisis: This event did not directly impact the sector, but it created uncertainty and lowered confidence in the global economy, impacting demand for basic materials.
- Chinese market turbulence (2015-2016): The volatility in the Chinese stock market had an adverse effect on the sector, mainly in oil.

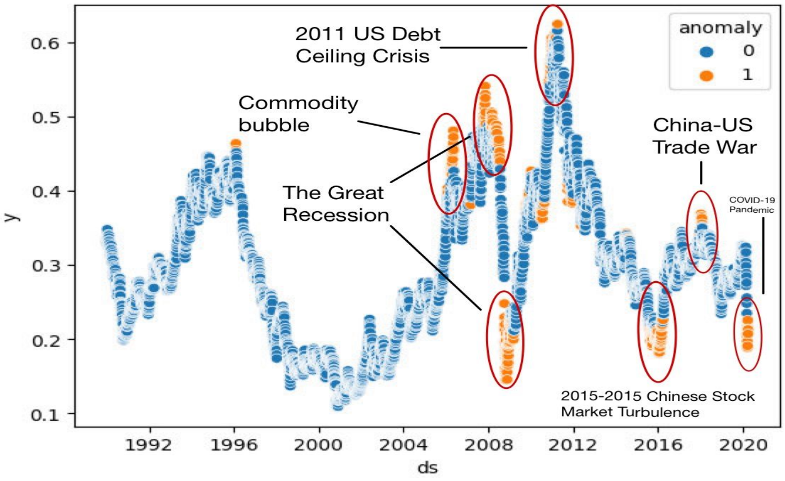


Figure 1: Basic Materials Sector from 1990-2020

Real Estate Sector:

- Increased mortgage rates (2016): The rise in mortgage rates impacted affordability, with the 30-year fixed-rate average jumping to 4.3 percent. This impacted the market trend negatively.
- Housing shortage (2020): The model identified a unique anomaly in the form of a housing shortage in 2020. This refers to the housing shortage in early 2020, around the cutoff of this dataset.

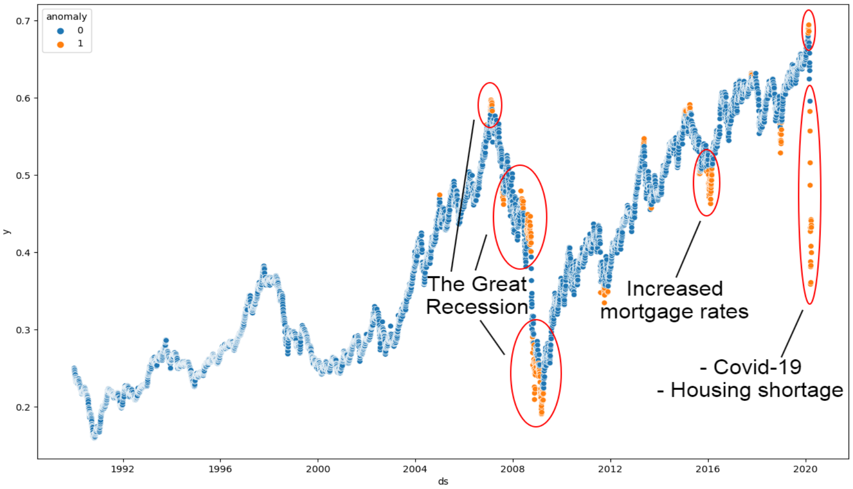


Figure 2: Real-Estate Sector from 1990-2020

Financial Sector:

- Brexit vote (2016): The decision of the United Kingdom to leave the European Union led to significant market volatility, with people immediately selling off their risk assets.
- Interest rate hikes (2018): The Federal Reserve's decision to raise interest rates four times in 2018 alone caused market volatility, specifically affecting the financial sector.

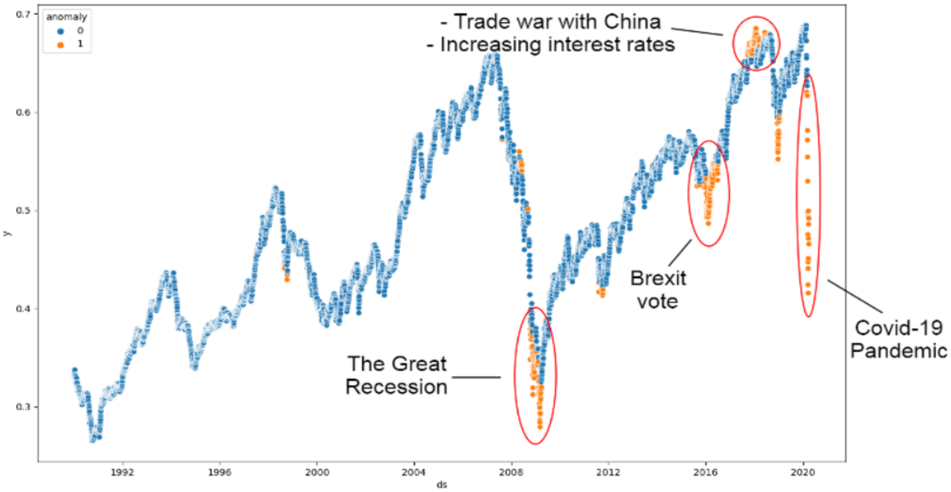


Figure 3: Financial Sector from 1990-2020

Technology Sector:

- Dot-com bubble (2000): The dot-com bubble led to a market crash as the technology stocks were overvalued from huge growth in the late 1990s.
- Crashing oil prices and market volatility in China (2016-2017): The tech sector was impacted due to a reliance on the global market, and its exposure to the Chinese market.

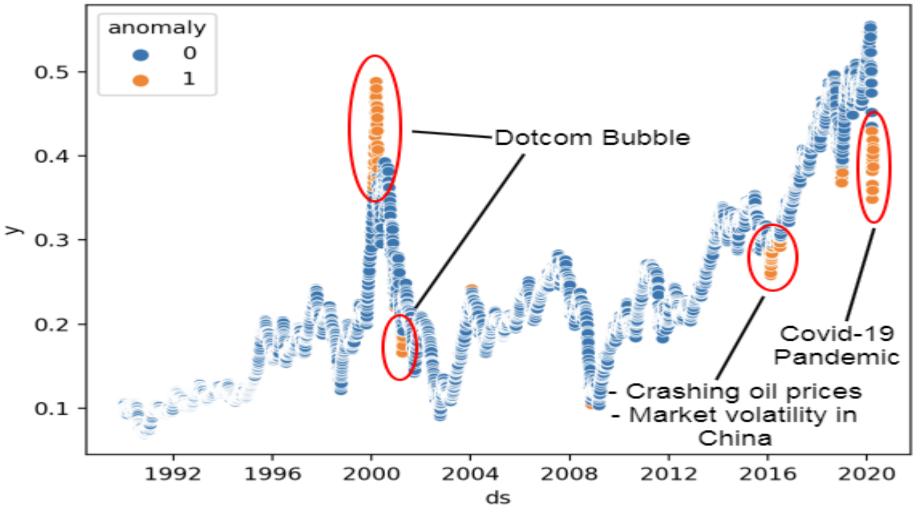


Figure 4: Technology Sector from 1990-2020

Shared Anomalies:

- Great Recession (2007-2009): The Great Session impacted most of these sectors, causing the Nasdaq Composite to drop 40% Although all other sectors were affected, the technology sector did not take as huge a hit. The demand for PCs was strong along with a rise in the use of smartphones and social media. It took a hit but was not nearly as hard as the other sectors.
- Trade War (2018): The trade war between the United States and China affected the financial and materials sector the most as a result of tariffs imposed on Chinese products.
- COVID-19 pandemic (2020): The global pandemic disrupted the world, impacting all sectors. The magnitude of the effect varies per industry but the entire Nasdaq Composite dropped 7.29%

Conclusion

The identification of anomalies within the NASDAQ time series stock data allows for a heightened understanding of historical trends that affect the stock market. By leveraging multiple techniques, this study not only discloses patterns and irregularities but also helps expand upon what events shape market movements. This new perspective allows for better understanding of decision-making processes within the overall market and understanding the interplay between historical events and stock price deviations.

Facebook's Prophet model use of additive models and regression techniques provided fast and accurate anomaly detection and forecasting features. However, the original dataset posed challenges from its extensive range and its scale. Refining the scope of the data along with adopting the data normalization technique proved to be the most effective. The methodologies used in error reduction highlights the importance of making adjustments to achieve accuracy and reliability.

In conclusion, the methodologies we use include the evolution of machine learning practices. By adapting the model for each industry, users can compare the diversity of each dataset and have a brief view of the complexities, hidden patterns and make specific decisions based on the results.

As the technology evolves, the most important part is still pursuing precision, this will lead the way for the bright future of data-driven insight innovation and success.

Future Work

For future work that can be applied to this project, there are many opportunities that can be leveraged from anomaly detection and stock market forecasting. Firstly, we can use the current stock market anomalies that exist to predict future stock market prices. Secondly, we can use the forecasting feature available in our model and run anomaly detection of the returned data in order to foresee upcoming global events. Finally, we can also tune our model to use sentiment analysis from web scraping companies to improve the forecasting and anomaly detection features of our project.

References

Facebook Prophet: https://facebook.github.io/prophet/docs/quick_start.html#python-api

Stock Market Dataset: https://www.kaggle.com/datasets/paultimothymooney/stock-market-data

Additive Models: https://rafalab.dfci.harvard.edu/pages/649/section-10.pdf

2014's Biggest Moments in Tech, Business Insider:
https://www.businessinsider.com/2014s-biggest-moments-in-tech-2014-12

Timeline of U.S. Stock Market Crashes:
https://www.investopedia.com/timeline-of-stock-market-crashes-5217820

Nasdaq Stock Screener: https://www.nasdaq.com/market-activity/stocks/screener

Mortgage rates' rise continues, reaching highs not seen in more than two years:
https://www.washingtonpost.com/business/economy/2016/12/22/335b27ac-c888-11e6-8bee-54e800ef2a63_story.html

Brexit market impact: https://russellinvestments.com/us/blog/brexit-market-impact

Fed's interest rate history: The federal funds rate from 1981 to the present (2018):
https://www.bankrate.com/banking/federal-reserve/history-of-federal-funds-rate/#:~:text=Rate%20cuts%202019%2D2020&text=That%20is%2C%20until%202017%2C%20when,peaked%20at%202.25%2D2.5%20percent.

The State of the Nation's Housing 2020:
https://www.jchs.harvard.edu/sites/default/files/reports/files/Harvard_JCHS_The_State_of_the_Nations_Housing_2020_Report_Revised_120720.pdf

The Late 1990s Dot-Com Bubble Implodes in 2000:
https://www.goldmansachs.com/our-firm/history/moments/2000-dot-com-bubble.html

What Was the COVID-19 Stock Market Crash of 2020? Causes & Effects:
https://www.thestreet.com/dictionary/c/covid-19-stock-market-crash-of-2020

The 2015-16 Chinese Market Crash:
https://www.avatrade.com/blog/trading-history/the-2015-16-chinese-market-crash

What Happened to Oil Prices in 2016?:
https://www.fool.com/investing/2016/12/17/what-happened-to-oil-prices-in-2016.aspx

Stocks fall back on commodities slump:
https://www.washingtonpost.com/business/2011/05/03/AFjbBQLG_story.html

Dow plunges 531 points, CNN:
https://money.cnn.com/2015/08/21/investing/stocks-market-lookahead-august-21/index.html?iid=hp-toplead-dom

The Light At The End Of The Tech Downturn:
https://www.forbes.com/sites/timbajarin/2023/02/03/the-light-at-the-end-of-the-tech-downturn/?sh=5b50a971465a

Trump's Trade War with China Is Officially Underway:
https://www.nytimes.com/2018/07/05/business/china-us-trade-war-trump-tariffs.html

Stock Market Crash 2020: Everything You Need to Know:
https://www.nasdaq.com/articles/stock-market-crash-2020%3A-everything-you-need-to-know-2020-03-10