"Predicting Stock Returns Using News Sentiment Analysis: A Comparative Study of Machine Learning Models during COVID-19"

Table of Contents

[1. Introduction 5](#_Toc144117192)

[1.1 Background 5](#_Toc144117193)

[1.2 Objective 7](#_Toc144117194)

[2. Literature Review 8](#_Toc144117195)

[2.1 Sentiment Analysis: A Brief Overview 8](#_Toc144117196)

[2.2 News Sentiment and Stock Market Movements 8](#_Toc144117197)

[2.3 Machine Learning Models in Financial Forecasting 9](#_Toc144117198)

[2.4 The COVID-19 Pandemic: A New Challenge 10](#_Toc144117199)

[2.5 Literature Review Summary 10](#_Toc144117200)

[3. Data 10](#_Toc144117201)

[3.1 Dataset Overview 10](#_Toc144117202)

[3.2 Data Sources 11](#_Toc144117203)

[3.3 Variables 12](#_Toc144117204)

[3.4 Dataset Specifications 13](#_Toc144117205)

[4. Preprocessing 14](#_Toc144117206)

[4.1 Handling Missing Values 14](#_Toc144117207)

[4.2 Visualization of Missing Data 15](#_Toc144117208)

[5. Feature Engineering 16](#_Toc144117209)

[5.1 Time-Based Features 16](#_Toc144117210)

[5.2 Pandemic Phase Indicator 17](#_Toc144117211)

[5.3 Return Calculations and Classification 17](#_Toc144117212)

[5.4 Volatility 18](#_Toc144117213)

[5.5 Correlation Matrix 18](#_Toc144117214)

[6. Exploratory Data Analysis (EDA) 20](#_Toc144117215)

[6.1 Descriptive Statistics 20](#_Toc144117216)

[6.2 Data Visualization 21](#_Toc144117217)

[6.2.1 Temporal Evolution of Average Sentiment 21](#_Toc144117218)

[6.2.2 Average Sentiment Score by Company 22](#_Toc144117219)

[6.2.3 Time Series of New COVID-19 Cases 25](#_Toc144117220)

[6.2.4 Time Series Plot of Close Prices 26](#_Toc144117221)

[7. Impact of Firm Size on News Sentiment Score 27](#_Toc144117222)

[7.1 Correlation and Regression 28](#_Toc144117223)

[7.2 Cluster Analysis of Firm Size and News Sentiment 31](#_Toc144117224)

[7.3 Cluster Analysis without Big Companies 33](#_Toc144117225)

[7. Feature Selection 36](#_Toc144117226)

[8. Model Evaluation 36](#_Toc144117227)

[8.1 Why AUC? 36](#_Toc144117228)

[8.1 Random Forest 37](#_Toc144117229)

[8.1.1 Theoretical Overview 37](#_Toc144117230)

[8.1.2 Methodology 38](#_Toc144117231)

[8.1.3 Results 38](#_Toc144117232)

[8.1.4 Discussion 41](#_Toc144117233)

[8.2 XGBoost 41](#_Toc144117234)

[8.2.1 Theoretical Overview 41](#_Toc144117235)

[8.2.2 Methodology 42](#_Toc144117236)

[8.2.3 Results 43](#_Toc144117237)

[8.2.4 Discussion 45](#_Toc144117238)

[8.3 LightGBM 45](#_Toc144117239)

[8.3.1 Theoretical Overview 45](#_Toc144117240)

[8.3.2 Methodology 46](#_Toc144117241)

[8.3.3 Results 46](#_Toc144117242)

[8.3.4 Discussion 48](#_Toc144117243)

[8.4 CatBoost 48](#_Toc144117244)

[8.4.1 Theoretical Overview 48](#_Toc144117245)

[8.4.2 Methodology 49](#_Toc144117246)

[8.4.3 Results 49](#_Toc144117247)

[8.4.4 Discussion 51](#_Toc144117248)

[8.5 AUC score in Each Model 51](#_Toc144117249)

[8.7 Comparative Analysis of Top 10 Firms by AUC Scores Across Models 53](#_Toc144117250)

[8.8 Comparative Analysis of Bottom 10 Firms by AUC Scores Across Models 55](#_Toc144117251)

[8.9 Feature Importance Across Models 58](#_Toc144117252)

[8.9.1 How Feature Importance is computed 58](#_Toc144117253)

[8.9.2 Feature Importance Across Models 60](#_Toc144117254)

[8.9.3 Analyzing the Implications 61](#_Toc144117255)

[9. Discussion 62](#_Toc144117256)

[9.2 Variability Among Models 63](#_Toc144117257)

[9.3 Limitations and Scope for Future Research 63](#_Toc144117258)

[10. Ethics 64](#_Toc144117259)

[10.1 Ethical Foundations and Risk Management 64](#_Toc144117260)

[10.2 Societal Implications and Ethical Risks 64](#_Toc144117261)

[10.3 Methodological Ethical Challenges 65](#_Toc144117262)

[10.4 Responsible Innovation and Future Directions 65](#_Toc144117263)

[11. Conclusion 66](#_Toc144117264)

[References 67](#_Toc144117265)

[Appendix 73](#_Toc144117266)

[Table 1 Variables 12](#_Toc144117267)

[Table 2 Sample of Companies with Their Tickers 13](#_Toc144117268)

[Table 3 Descriptive statistics 20](#_Toc144117269)

[Table 4 Results of AUC scores 51](#_Toc144117270)

[Table 5 Companies with Their Tickers 73](#_Toc144117271)

[Figure 1 Top 20 tickers with most missing in news sentiment score. 16](#_Toc144117272)

[Figure 2 Distribution of positive and negative returns 18](#_Toc144117273)

[Figure 3 Correlation matrix of features 19](#_Toc144117274)

[Figure 4 Temporal evolution of average sentiment score 22](#_Toc144117275)

[Figure 5 Average sentiment score by company 24](#_Toc144117276)

[Figure 6 Time series of new COVID-19 cases 26](#_Toc144117277)

[Figure 7 Trend of closing prices for all companies 27](#_Toc144117278)

[Figure 8 News sentiment score vs. MarketCap 29](#_Toc144117279)

[Figure 9 Regression line: Impact of market capitalization on news sentiment score. 30](#_Toc144117280)

[Figure 10 Clusters of companies 32](#_Toc144117281)

[Figure 11 KMeans clustering of firm (Exclusing Outliers) 34](#_Toc144117282)

[Figure 12 AUC scores for Random Forest by ticker 40](#_Toc144117283)

[Figure 13 AUC scores for XGBoost by ticker 44](#_Toc144117284)

[Figure 14 AUC scores for LightGBM by ticker 47](#_Toc144117285)

[Figure 15 AUC scores for CatBoost by ticker 50](#_Toc144117286)

[Figure 16 Top 10 frims and bottom 10 Firms in AUC score across models 56](#_Toc144117287)

[Figure 17 Feature importance across models 60](#_Toc144117288)

# 1. Introduction

## 1.1 Background

In today's digital era, the intertwining of the financial markets with the expanse of digital information has reached unprecedented levels. The advent of the digital age has seen a paradigm shift from traditional financial analysis rooted in stock price histories and fundamental analyses to more contemporary methods, such as sentiment analysis. This innovative approach leverages the prowess of algorithms to interpret and predict based on sentiments encapsulated within a wide array of textual data sources.

Sentiment analysis, an analytical technique that mines textual data to discern and quantify sentiments, emotions, and opinions, has become an invaluable tool in the financial realm. By systematically applying algorithms to extract sentiments from financial news, reports, and the burgeoning universe of social media, financial experts and researchers aim to predict stock market movements, striving to gain an edge in an ever-evolving and highly competitive landscape (Bollen, Mao, & Zeng, 2011).

The symbiotic relationship between news sentiment and stock market movements has been a focal point for many empirical studies. Research by Tetlock (2007), among others, has posited that stock prices are influenced by the release of news. The sentiments embedded in these news releases, either bullish or bearish, can act as catalysts for market volatility. This relationship was thrown into sharp relief with the onset of the COVID-19 pandemic, a period marked by heightened unpredictability and significant financial fluctuations (Mazur et al., 2021).

Despite the established interplay between news sentiment and stock returns, there remains a lacuna in the empirical exploration of the comparative efficacy of diverse machine learning models, especially during global events of the magnitude of the COVID-19 pandemic. The pandemic's wide-reaching effects across various sectors and geographies highlighted the need for sophisticated predictive models capable of capturing these nuances (Baker et al., 2020). Further, while existing literature often zeroes in on broad market indices or a select cluster of large-cap stocks, comprehensive studies evaluating the performance of these models across an extensive array of tickers, particularly those on NASDAQ, remain scant.

Furthermore, while existing literature has often concentrated on broad market indices such as the S&P 500 (Yousse, et al., 2022) or specific sectors like the Dow Jones Industrial Average (Sousa et al., 2019), there is a relative scarcity of in-depth studies focusing on the intricacies of NASDAQ-listed companies. For instance, the S&P 500, representing a broader market perspective, is frequently used as a benchmark for U.S. equities. The Dow, being price-weighted, is often seen as a reflection of the industrial sector's health. However, neither provides the granularity of technology-centric and growth-oriented firms that the NASDAQ offers.

The choice of NASDAQ as a focal point for stock market prediction using news sentiment analysis is underpinned by several compelling reasons. First and foremost, NASDAQ's competitive nature, especially for smaller stocks, makes it a dynamic and volatile market. Elliott and Warr (2003) observed that NASDAQ stocks exhibit pronounced price effects compared to NYSE stocks when added to indices like the S&P 500, suggesting that NASDAQ's dealer market system can lead to more pronounced price movements, making it an intriguing subject for sentiment analysis (Elliott & Warr, 2003).

Furthermore, NASDAQ is renowned for its diverse listings, encompassing a plethora of high-tech companies, startups, and firms with high market-to-book ratios. This diversity is not just a testament to the market's breadth but also its depth. Chen et al. (2019) emphasized that the effects of investor sentiment on stock price performance are particularly potent for small, young, and high market-to-book ratio firms, a demographic that is well-represented on NASDAQ.

In light of these factors, it becomes evident that while other indices like the S&P 500 and DJI offer valuable insights into broader market movements and blue-chip stock performance, NASDAQ's unique characteristics render it especially conducive for academic stock market prediction using news sentiment analysis.

## 1.2 Objective

This study seeks to bridge these gaps by embarking on a comparative evaluation of four cutting-edge machine learning models: Random Forest, XGBoost, LightGBM, and CatBoost, specifically focusing on their aptitude in predicting stock returns using sentiment scores during the COVID-19 era. I focus on these four models because they are among the most widely used and have demonstrated superior performance across various prediction tasks, including in finance, as evidenced by previous literature (Huang et al., 2020). More traditional models like linear regression and SVM have become outdated, while deep learning approaches tend to be complex and computationally expensive. In contrast, ensemble tree-based methods like Random Forest, XGBoost, LightGBM, and CatBoost provide a balance of interpretability, efficiency, and predictive accuracy.

Random Forest is chosen for its robust performance and inherent feature selection capabilities (Breiman, 2001). XGBoost has emerged as a dominant technique in Kaggle competitions and real-world predictions due to its speed and performance optimization (Chen & Guestrin, 2016). LightGBM utilizes gradient-based one-side sampling and exclusive feature bundling to achieve faster training speed and higher efficiency (Ke et al., 2017). CatBoost leverages categorical data and avoids overfitting through innovative ordered boosting and categorical features processing (Prokhorenkova et al., 2018). By comparing these leading-edge models, this study aims to provide valuable insights into their comparative strengths and suitability for stock market predictions. Central to this research are three pivotal questions:

* Comparative analysis: In the backdrop of the COVID-19 pandemic, how do machine learning models like XGBoost, Random Forest, LightGBM, and CatBoost fare in leveraging news sentiment for stock return predictions?
* Feature Importance: Which attributes play a crucial role in influencing the predictions of these models? Is news sentiment consistently paramount, or do other factors such as volume or market cap wield a more significant influence?
* Stock Specificity: Does stocks’ intrinsic factors (i.e., business model or industry sector) affects the predictability of the stock returns.

By addressing these research questions, this investigation aspires to furnish seminal insights into the nuances of financial forecasting, especially in an epoch marked by volatility and rapid change.

# 2. Literature Review

The study of the relationship between news sentiment and stock market movements is a burgeoning field in financial research, enriched by the intersection of computational linguistics, finance, and machine learning. The increasing volume and accessibility of digital textual data, coupled with rapid advancements in computational techniques, have breathed new life into sentiment analysis and its applications in the realm of financial forecasting.

## 2.1 Sentiment Analysis: A Brief Overview

Sentiment analysis, or opinion mining, involves discerning and quantifying sentiments, emotions, and opinions within textual data. Early work in the domain, such as that by Pang, Lee, and Vaithyanathan (2002), focused primarily on product reviews. Their seminal research illuminated the potential of using textual analysis to infer consumer sentiments. Recently, the application of sentiment analysis has expanded manifold, especially in the field of finance. For example, Valle-Cruz et al. (2020) emphasized its significance as a forecasting tool for the stock market, with polarity being a widely adopted technique. Moreover, Kumar et al. (2021) highlighted the benefits of sentiment analysis in enhancing data accuracy in financial contexts. A study by Loughran and McDonald (2011) stressed the importance of crafting specialized dictionaries to mine financial texts, noting that traditional dictionaries often misclassify financial terms, leading to inaccurate sentiment classifications.

## 2.2 News Sentiment and Stock Market Movements

The link between stock market performance and news sentiment has been an area of keen academic interest. Tetlock (2007) demonstrated that negative words in financial news were strongly correlated with downward stock price movements. Further, Engelberg and Parsons (2011) illustrated how front-page news articles could induce stock market volatility. These findings underscore the market's sensitivity to news sentiment, with investors often reacting, either positively or negatively, to the sentiments embedded in news articles.

The influence of social media platforms, particularly Twitter, has also been explored in depth. Bollen, Mao, and Zeng's (2011) pioneering work revealed that Twitter mood could indeed serve as a predictor for stock market movements. Their research posited that the collective mood derived from Twitter feeds was indicative of subsequent changes in the Dow Jones Industrial Average.

## 2.3 Machine Learning Models in Financial Forecasting

Machine learning models have brought about a paradigm shift in financial forecasting. Traditional statistical models like multivariate linear regression have been extensively used for prediction tasks in finance (Brooks, 2014). However, these linear models often fail to capture complex nonlinear relationships and interactions in financial data (Atsalakis & Valavanis, 2009).The advent of machine learning heralded a new era, enabling more nuanced predictions.

Various machine learning algorithms have been deployed in the financial sector. XGBoost, for instance, is renowned for its robustness and has been employed extensively in financial forecasting, as highlighted by Chen and Guestrin (2016). On the other hand, Random Forest, a versatile ensemble learning method, has been lauded for its accuracy in financial applications (Liaw and Wiener, 2002).

Recent studies have also shed light on the efficacy of more advanced algorithms such as LightGBM and CatBoost. Ke, Meng, Finley, and Wang (2017) extolled the virtues of LightGBM, emphasizing its scalability and efficiency. Similarly, Prokhorenkova et al. (2018) explored the capabilities of CatBoost, highlighting its prowess in handling categorical features, a common occurrence in financial datasets.

## 2.4 The COVID-19 Pandemic: A New Challenge

The onset of the COVID-19 pandemic brought about unprecedented challenges for financial forecasting. Traditional models were often found wanting in the face of the pandemic's volatile impact on global stock markets. Baker et al. (2020) detailed the heightened economic uncertainty induced by COVID-19, emphasizing the need for more adaptable forecasting models.

The pandemic also underscored the significance of news sentiment. With news cycles dominated by pandemic-related updates, investor sentiment was highly influenced by the tone and content of news articles. This phenomenon accentuated the symbiotic relationship between news sentiment and stock market movements, with studies like Zhang, Xie, Wang, and Chen's (2020) work highlighting the potential of sentiment analysis in predicting stock market volatility during the pandemic.

## 2.5 Literature Review Summary

In summary, the evolving relationship between news sentiment and stock market movements stands at the intersection of computational linguistics, finance, and machine learning. Initially, sentiment analysis, with its origins in product reviews, transitioned to significantly impact financial forecasting. The interplay between news sentiment and stock market dynamics has been particularly intriguing, with evidence suggesting pronounced market reactions based on news sentiment. The advent of machine learning has ushered in a transformative phase in financial forecasting, introducing tools that adeptly unravel complex data patterns. The challenges posed by the COVID-19 pandemic brought to the fore the criticality of nimble forecasting models and magnified the role of news sentiment during times of global upheaval.

# 3. Data

## 3.1 Dataset Overview

The dataset consolidates daily stock data with sentiment scores derived from news analytics, covering a wide range of firms listed on NASDAQ. The study spans from April 30, 2013, to April 26, 2023, allowing me to not only capture the intricacies of the COVID-19 pandemic era but also the years leading up to it. This comprehensive timeframe facilitates a deeper understanding of stock returns in relation to news sentiment.

## 3.2 Data Sources

In this paper, I utilized three types of datasets, in particular, stock data, news sentiment data, and COVID-19 data.

Stock Data: The stock data is at daily level and was sourced using the Yahoo Finance Python library, which records information regarding stock indicators including opening price, closing price, high, low, adjusted closing price, and trading volume. The significance of stock data, especially from Yahoo Finance, in predicting stock behaviors has been emphasized in previous research. For instance, Jagwani, Gupta, Sachdeva, and Singhal (2018) explored stock price forecasting using data from Yahoo Finance and highlighted the importance of analyzing both seasonal and nonseasonal trends. Furthermore, Bordino, Kourtellis, Laptev, and Billawala (2014) underscored the predictive power of Yahoo Finance user browsing behavior in determining stock trade volumes.

Sentiment Data: RavenPack Analytics, a leading provider of real-time data services for financial professionals, supplied the sentiment scores. They aggregate data from numerous sources, including but not limited to Dow Jones Financial Wires, Wall Street Journal, Barron’s, and MarketWatch.

COVID-19 Data: The data detailing new daily COVID-19 cases in the US was procured from Our World in Data. This dataset stands out for its comprehensiveness and timely updates. Data is regularly sourced from esteemed organizations such as the World Health Organization (WHO), Johns Hopkins University, and the European Centre for Disease Prevention and Control (ECDC).

## 3.3 Variables

The dataset encompasses the following variables:

Table Variables

|  |  |  |
| --- | --- | --- |
| **Name of Variables** | **Explanation** | **Source** |
| **Date** | Highlights the weekdays, strictly adhering to stock market operational days, spanning from April 30, 2013, to April 26, 2023. | Yahoo Finance |
| **Open, High, Low, Close, Adj Close** | These are conventional stock price indicators illustrating daily stock dynamics. | Yahoo Finance |
| **Volume** | Represents the quantity of shares exchanged on a particular day. | Yahoo Finance |
| **Ticker** | Identifiers for NASDAQ-listed firms, encapsulating a broad array of companies from diverse sectors. | Yahoo Finance |
| **Sentiment\_Score** | Quantifies the sentiment, derived as an average from all pertinent news articles associated with the respective company on a specific day. The score oscillates between -1 (denoting negative sentiment) and 1 (representing positive sentiment). | RavenPack Analytics |
| **New\_Covid\_Cases** | Chronicles the daily count of fresh COVID-19 cases reported in the US. | Our World in Data |
| **MarketCap** | Portrays the market capitalization of the corresponding company on the given date. | Yahoo Finance |
| **PandemicPhase** | A binary delineation demarcating the periods before and after the World Health Organization's official declaration of the pandemic on March 11, 2020. | World Health Organization |
| **Volatility** | Evaluated as the rolling standard deviation across the past 5 days of returns. | Yahoo Finance |
| **Year, Month, Weekday** | Time-oriented features extracted from the Date column. | Yahoo Finance |
| **PositiveReturn** | A binary flag signifying whether the stock registered a positive return on that day.If the value is set to one, it indicates that the stock registered a positive return on that particular day; otherwise, it suggests a non-positive return. | Yahoo Finance |

## 3.4 Dataset Specifications

The dataset utilized for this research is a meticulously curated amalgamation of financial metrics, sentiment scores, and pandemic-related data, specifically curated to decipher the intricate dynamics of the stock market in the wake of the COVID-19 pandemic. Derived from the NASDAQ-100 index, the dataset encapsulates observations from 102 companies, representing a diverse spectrum of industries and market capitalizations. Some notable firms included are tech behemoths like Apple, NVIDIA, and Microsoft, along with other significant players across different sectors.

In total, the dataset boasts a substantial 233,135 observations, offering a comprehensive view of the market's trajectory during the stipulated period. The sentiment scores, a pivotal component of this dataset, exhibit a range from a minimum of -96.88 to a maximum of 66.62, with a median sentiment value of 0.06. These scores provide insights into the market's sentiment, reflecting investor perceptions and reactions to daily news and events.

Table Sample of Companies with Their Tickers

|  |  |
| --- | --- |
| **Company name** | **Ticker** |
| **Microsoft Corp** | MSFT |
| **Apple Inc** | AAPL |
| **NVIDIA Corp** | NVDA |
| **…** | … |

(Note: The table represents a sample of the companies included in the dataset. The complete list can be found in the appendix.)

# 4. Preprocessing

The primary objective of preprocessing is to ensure that data is in an appropriate state for subsequent analysis, making it a critical stage in the data pipeline. In finance, data often comes from multiple sources, each with its unique characteristics and potential pitfalls, such as missing values. Addressing these issues is paramount to ensure the robustness of any modeling efforts that follow.

## 4.1 Handling Missing Values

One of the most common challenges in financial datasets is the occurrence of missing values. An initial assessment revealed that our dataset had 30,397 missing values in the Sentiment\_Score column and 150,833 in the New\_Covid\_Cases column.

For the New\_Covid\_Cases column, the missing values were imputed with zeros. This choice is justified given the nature of the data: on days where no new cases are reported or data isn't available, it's reasonable to consider the new cases as zero.

However, the Sentiment\_Score column presents a more intricate scenario. A deeper dive revealed two companies (PepsiCo and Activision Blizzard) lacked sentiment scores across the sample period. It's plausible that these companies might not have had significant news coverage or that the data source did not track their sentiment for the given period. Given the complete absence of sentiment data for these tickers, these two companies were excluded from the dataset to avoid biasing the models with potentially misleading zeros. For the remaining entries, missing values in the Sentiment\_Score column were filled with zeros, operating under the assumption that a missing value might imply neutral sentiment or lack of significant news on that particular day.

## 4.2 Visualization of Missing Data

A visualization was generated to highlight the top 20 tickers with the highest count of missing Sentiment\_Score values. This visualization can aid in understanding whether the missingness is random or if it correlates with specific tickers. Both PepsiCo and Activision Blizzard exhibited a complete absence of sentiment scores throughout the sample period. This conspicuous absence suggests that the missing values for these firms are not random. One potential explanation is that these companies might not have garnered substantial news coverage during this period, or the data source might have excluded their sentiment for the duration under study.

In contrast, firms including Atlassian Corp(TEAM), CoStar Group Inc(CSGP), and others demonstrated occasional occurrences of missing sentiment scores. This observation implies that while these companies had sentiment data on specific days, there were instances where sentiment information was absent, potentially due to a lack of significant news coverage or other external factors. Such intermittent missing values could be indicative of the fact that certain companies, despite being prominent in the NASDAQ-100, might not always be the focal point of news cycles. This could be attributed to various reasons, ranging from the sector they operate in, the magnitude of their financial operations on specific days, or the overshadowing presence of other major financial or global events that divert media attention.

Given the patterns discerned, it is clear that addressing missing values is of paramount importance. The decision to fill these gaps with zeros, especially for firms with occasional missing data, might imply days with neutral sentiment or a lack of meaningful news. This decision, while practical, is a pivotal preprocessing step that sets the stage for the subsequent analysis, underscoring the importance of thorough data handling in financial research.

A graph of a number of missing values

Description automatically generated

Figure Top 20 tickers with most missing in news sentiment score.

# 5. Feature Engineering

Feature engineering is a pivotal phase in the data science pipeline, especially in the realm of financial machine learning, where the intricacies of the financial world meld with the mathematical rigor of machine learning. By transforming, aggregating, or creating new variables from existing data, I aspire to enhance the predictive power of our machine learning models. In the context of this study, where I aim to predict stock returns through an amalgamation of stock price data, news sentiment, and pandemic-related metrics, the significance of feature engineering cannot be overstated.

## 5.1 Time-Based Features

Time, inherently sequential, holds paramount importance in financial datasets. Extracting finer granularities from the date, such as year, month, and weekday, can provide the model with valuable temporal context. This aids in discerning seasonality, trends, or specific day-of-week effects that might be latent in the stock returns. *Year, Month, Weekday* variables are constructed used to capture the time-relevant information.

## 5.2 Pandemic Phase Indicator

The COVID-19 pandemic has indubitably left an indelible mark on the financial markets. To encapsulate this transformative event, a binary variable, *'PandemicPhase'*, was introduced. This variable bifurcates the dataset into two phases – before and after the World Health Organization officially declared the pandemic on March 11, 2020.

## 5.3 Return Calculations and Classification

Stock returns, calculated as the percentage change in closing prices from one day to the next, serve as a foundational metric in financial analyses. This continuous variable was then transformed into a binary format, *'PositiveReturn'*, indicating whether the stock witnessed a positive return on a given day.

A graph of a number of negative return

Description automatically generated

Figure Distribution of positive and negative returns

## 5.4 Volatility

Volatility, the statistical measure of the dispersion of returns, is a critical risk metric in finance. In this study, I computed the rolling standard deviation of the past 5 days' returns to capture short-term volatility. Such a feature can be instrumental in understanding the risk-reward dynamics and gauging market uncertainty.

By integrating these engineered features, our dataset is now equipped with a richer set of variables that encapsulate the multifaceted world of financial markets, especially during the tumultuous period of the COVID-19 pandemic.

## 5.5 Correlation Matrix

Understanding inter-variable relationships is paramount in financial modeling. A correlation matrix can provide a snapshot of how different features in our dataset relate to each other. This can be instrumental in identifying potential multicollinearity or variables that might be of particular interest in predicting stock returns. The correlation matrix is presented below

A screenshot of a graph

Description automatically generated

Figure Correlation matrix of features

The heatmap unveils the correlation coefficients between different features:

* The diagonal, as anticipated, shows a correlation of 1 since it represents the correlation of each variable with itself.
* Open, High, Low, Close, and Adj Close showcase high mutual correlations, which is expected since these stock indicators are often intertwined.
* Interestingly, New\_Covid\_Cases does not exhibit strong correlations with other features, suggesting that its influence on stock returns is likely nuanced and mediated by other external factors.
* Sentiment\_Score has a relatively low correlation with PositiveReturn, reinforcing the complexity of the relationship between news sentiment and stock performance.

The correlation matrix is paramount for two main reasons:

* It assists in recognizing potential multicollinearity, which might complicate our modeling efforts.
* It aids in discerning which features might be especially pertinent or redundant in predicting stock returns.

In summary, the feature engineering phase has augmented the dataset by infusing it with variables that encapsulate the nuanced dynamics of the financial domain. By imbuing the model with temporal context, pandemic-induced market shifts, stock return patterns, and volatility measures, I aim to provide a holistic perspective, enhancing the model's predictive acumen.

# 6. Exploratory Data Analysis (EDA)

## 6.1 Descriptive Statistics

First, I provide a summary of the dataset's main characteristics using descriptive statistics. This includes measures such as mean, standard deviation, minimum and maximum values for each column. I commence by summarizing the numerical attributes within our dataset.

Table Descriptive statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Open** | 228099.0 | 135.13 | 219.70 | 0.70 | 38.23 | 71.41 | 149.69 | 2697.75 |
| **High** | 228099.0 | 136.92 | 222.53 | 0.71 | 38.66 | 72.30 | 151.75 | 2721.85 |
| **Low** | 228099.0 | 133.29 | 216.77 | 0.65 | 37.74 | 70.54 | 147.60 | 2687.81 |
| **Close** | 228099.0 | 135.15 | 219.67 | 0.70 | 38.21 | 71.46 | 149.69 | 2703.26 |
| **Adj Close** | 228099.0 | 131.23 | 220.08 | 0.70 | 34.34 | 65.79 | 144.80 | 2703.26 |
| **Volume** | 228099.0 | 11554842.24 | 28560594.65 | 0.00 | 1472750.00 | 3154700.00 | 8661400.00 | 1065523200.00 |
| **Sentiment\_Score** | 228099.0 | 0.67 | 2.38 | -96.88 | 0.00 | 0.08 | 0.38 | 66.62 |
| **New\_Covid\_Cases** | 228099.0 | 28569.96 | 83851.28 | 0.00 | 0.00 | 0.00 | 20673.00 | 1265520.00 |
| **MarketCap** | 228099.0 | 214808803513.93 | 478458310487.78 | 16346392576.00 | 39679574016.00 | 62220886016.00 | 131073245184.00 | 2828020482048.00 |

Stock Prices (Open, High, Low, Close, Adj Close): The minimum price is 0.70 dollar while the maximum is 2,703.26 dollar. The distribution of prices appears to be right-skewed, as the mean is significantly greater than the median.

Volume: The trading volume displays substantial variability, ranging from days with zero traded volume to those exceeding a billion.

Sentiment\_Score: The average sentiment score is around 0.67. However, this metric is susceptible to significant fluctuations, spanning from -96.88 to 66.62. This wide range underscores the immense variability in daily news sentiment for different firms. A major portion of the sentiment scores is clustered around the 0 mark, indicating many days have neutral or minimal sentiment. The distribution shows slight positive skewness, with a significant number of days having positive sentiment scores. There are fewer days with strongly negative sentiment scores compared to those with positive scores.

New\_Covid\_Cases: The mean daily count of new COVID-19 cases in the US stands at approximately 28,570. With the maximum number of cases recorded in a single day being over 1.26 million, it underscores the magnitude of the pandemic's peak phases.

MarketCap: The companies' market capitalization exhibits a wide range, spanning from roughly 15 billion dollar to a staggering 2.8 trillion dollar.

## 6.2 Data Visualization

### 6.2.1 Temporal Evolution of Average Sentiment

The temporal fluctuations of sentiment, averaged across stocks, reflect overarching market moods from 2013 to 2023. The "Temporal Evolution of Average Sentiment Score" figure (see below) reveals a general upward trend in sentiment over the years, interspersed with occasional dips and spikes. The pronounced troughs, especially around early 2020, likely correspond to the global uncertainties presented by the COVID-19 pandemic. As the sentiment subsequently rebounds, it could be indicative of positive developments like vaccine announcements or economic recovery signals.

Such rapid transitions underline the stock market's sensitivity to real-time events and the prevailing reactive nature. For a comprehensive understanding, cross-referencing these sentiment shifts with significant global events on those dates would be pivotal. Established studies have emphasized the interconnectedness of market sentiment and major global occurrences, further underscoring the importance of this analysis (Tetlock, 2007).

A graph showing a blue line

Description automatically generated

Figure Temporal evolution of average sentiment score

### 6.2.2 Average Sentiment Score by Company

The systematic analysis of average sentiment scores across diverse companies provides insights into possible biases, patterns, or tendencies in news coverage. Companies such as META often emerge in a positive light, possibly due to favorable strategic decisions, groundbreaking innovations, or robust corporate governance. On the other hand, certain firms may frequently find themselves portrayed negatively, perhaps due to market challenges, controversies, or adverse market conditions.

Before delving into specific observations from the data, it's essential to understand the methodology behind these sentiment scores. The 'Sentiment Score' column values typically range from -1 to 1. A score near -1 suggests a prevailing negative sentiment, whereas a score closer to 1 reflects a predominantly positive sentiment. These scores are aggregated values, capturing the essence of all news articles related to the company for a given day. Importantly, while the sentiment score of an individual news item remains confined between -1 and 1, the aggregated score isn't bound by this range. For instance, on a day when Microsoft is covered in three news items with sentiment scores of 0.8, 0.7, and 0.6, the cumulative sentiment score amounts to 2.1, exceeding the typical -1 to 1 threshold. This means that while the sentiment of a single news item may be negative, the aggregated or average score tends to be neutral or positive, reflecting a more balanced overall portrayal in the media.

With this understanding, key takeaways from the Figure 6 depicting the average sentiment scores of various companies throughout the dataset include:

* Companies manifest a wide spectrum of average sentiment scores, with some near neutrality, while others exhibit pronounced positive or negative tendencies. Such variability points to the nuanced representation these firms receive in news outlets.
* Notably, even the companies with the least favorable average sentiment, such as ODFL, have scores just above zero, suggesting that their news coverage is largely neutral with a slight negative inclination.
* A significant portion of companies cluster around a neutral average sentiment, indicating a balanced news portrayal over the studied duration.



Figure Average sentiment score by company

* Most companies display a proclivity for positive sentiment, hinting at consistently favorable media coverage. Conversely, a few veer towards the negative end, suggesting a more critical or adverse representation.
* Variations in sentiment scores across companies could be ascribed to factors like sectoral affiliations, corporate strategies, financial performance, and market standing. However, these reasons are still somewhat vague. A deeper dive into these variations will be conducted in the following sections to provide a clearer understanding.
* Variations in sentiment scores across companies could be ascribed to factors like sectoral affiliations, corporate strategies, financial performance, and market standing. However, these reasons are still somewhat vague. A deeper dive into these variations will be conducted in *7. Impact of Firm Size on News Sentiment Score* to provide a clearer understanding.

Recognizing these patterns and intricacies is pivotal, as it offers a detailed perspective on the dynamic relationship between news sentiment, stock returns, and intrinsic company attributes.

### 6.2.3 Time Series of New COVID-19 Cases

The time series plot illustrates the progression of new COVID-19 cases over the study period:

A graph of a number of cases

Description automatically generated

Figure Time series of new COVID-19 cases

* Initially, there were minimal or no reported cases until early 2020, marking the onset of the pandemic.
* The cases surged in waves, peaking in late 2020 and early 2021, followed by another spike towards the end of 2021.
* After the peaks, there were periods of decline, indicating efforts to control the pandemic, potentially through lockdowns, vaccination drives, and other preventive measures.
* By the end of the study period, the number of new cases had decreased, but periodic fluctuations were still evident.

### 6.2.4 Time Series Plot of Close Prices

We will observe the trend in closing prices over time for all companies in our dataset. This visualization will provide insights into the overall market trend and highlight any significant market movements.

A graph of a stock market

Description automatically generated with medium confidence

Figure Trend of closing prices for all companies

The plot visualizes the trends in closing prices over the entire duration for the myriad of companies in our dataset:

* The plot captures the multifarious trajectories of different companies, underscoring the disparate impacts of global and company-specific events on stock prices.
* Some stocks demonstrate pronounced growth, while others remain relatively stable. There are also instances of stocks exhibiting volatile fluctuations.
* The dense overlapping of lines, especially during the pandemic era, encapsulates the increased market volatility and the synchronous movement of multiple stocks, which is characteristic of global events with widespread financial implications.

# 7. Impact of Firm Size on News Sentiment Score

Numerous literatures have emphasized the intricate link between firm size and news sentiment analysis. For instance, Gillam, Guerard, and Cahan (2015) asserted that media's influence on firms is profound, with firm size playing a pivotal role in shaping news sentiment. With the burgeoning significance of social media platforms, Peng, Zhang, and Gopal (2022) observed that investors' reactions to news, combined with specific firm characteristics such as firm size, profoundly influence the sentiment of financial discussions online. Furthermore, Li and Yang (2017) posited that sentiment associated with individual stocks, which could be influenced by firm size, has a notable impact on stock returns, underscoring the criticality of understanding the relationship between these variables. Additionally, Xu and Zhou (2018) associated the predictive power of sentiment indices with various firm characteristics, identifying firm size as a potential determinant in this interplay.

Despite these extensive studies highlighting the interrelation between firm size and news sentiment, no literature explicitly delves into the direct impact of firm size on news sentiment score. This section, therefore, seeks to bridge this gap by analyzing sentiment scores across firms of varying sizes, aiming to discern patterns and dependencies that might provide insights into potential variations in media representation based on firm size.

## 7.1 Correlation and Regression

The relationship between a company's market capitalization and its news sentiment score is explored through a scatter plot of average sentiment scores against market capitalization.

Upon quantifying this relationship, the Pearson correlation coefficient was calculated to be approximately 0.702, indicating a strong positive correlation between market capitalization and news sentiment score. This suggests that larger companies, in terms of market capitalization, tend to receive a more favorable news sentiment on average.

A graph with numbers and a line

Description automatically generated with medium confidence

Figure News sentiment score vs. MarketCap

To delve deeper into the relationship between firm size (measured by market capitalization) and news sentiment score, a simple linear regression model was employed.

A graph with a red line

Description automatically generated

Figure Regression line: Impact of market capitalization on news sentiment score.

The model can be represented as:

Where:

* is the intercept.
* represents the change in sentiment score for a one-unit change in market capitalization.
* is the error term.

Regression Results:

* Coefficient for MarketCap: The coefficient of 2.597e-12 suggests that for every one-unit increase in market capitalization, the sentiment score increases by approximately This positive coefficient indicates a direct relationship between firm size and news sentiment score.
* R-squared: The R-squared value of 0.492 indicates that approximately 49.2% of the variability in news sentiment scores can be explained by the market capitalization of the firms. This is a moderate explanatory power, suggesting that while firm size does play a role in influencing news sentiment, other factors might also be at play.
* P-value for MarketCap: The p-value is less than 0.05, suggesting that market capitalization is a statistically significant predictor of news sentiment score.
* Intercept: The intercept value of 0.1082 suggests the expected sentiment score when the market capitalization is zero. However, in the context of this analysis, the intercept doesn't have a practical interpretation as companies can't have a market capitalization of zero.

The regression analysis reveals that firm size, as measured by market capitalization, has a statistically significant positive effect on news sentiment scores. This suggests that larger firms might be more likely to receive positive news coverage compared to their smaller counterparts. This could be attributed to various reasons, including the larger firms' visibility, their strategic decisions, or their ability to influence media narratives. However, it's important to note that this is based on a simple linear regression, which might not capture the complexities and nuances of the relationship fully.

## 7.2 Cluster Analysis of Firm Size and News Sentiment

The clustering visualization provides an intricate snapshot of companies based on their market capitalization and average news sentiment scores. Through a KMeans clustering approach, we've segmented the companies into four distinct groups, each potentially representing a unique interplay between firm size and media portrayal.

A graph with red and blue dots

Description automatically generated

Figure Clusters of companies

Note: Cluster 1 Companies: Broadcom Inc, Adobe Inc, Netflix Inc, Intel Corp, Starbucks Corp, PayPal Holdings Inc, Airbnb Inc, Moderna Inc, American Electric Power Co Inc, Electronic Arts Inc, Dollar Tree Inc, eBay Inc, Zoom Video Communications Inc, and many others. Cluster 2

Companies: Microsoft Corp, Meta Platforms Inc, Alphabet Inc. Cluster 3 Company: Apple Inc. Cluster 4 Companies: NVIDIA Corp, Amazon.com Inc, Tesla Inc.

* Cluster 1 (Red): This cluster houses companies with a relatively high market capitalization and neutral to positive news sentiment scores. It's conceivable that these are large-cap firms that generally receive balanced media coverage, with occasional positive news highlights.
* Cluster 2 (Blue): Firms in this cluster have a moderate market capitalization and predominantly neutral news sentiment scores. These might represent mid-cap companies that receive consistent, yet balanced, media attention.
* Cluster 3 (Green): This cluster encompasses companies with lower market capitalization but relatively positive sentiment scores. These could be smaller firms that, despite their size, have garnered favorable media coverage, possibly due to innovative products, strategies, or positive financial performance.
* Cluster 4 (Cyan): Companies in this cluster exhibit both low market capitalization and neutral to slightly positive sentiment scores. It's possible that these are smaller firms with minimal media coverage or those that have been neither exceptionally praised nor critically assessed in the news.

The centroids (marked with yellow 'X') signify the mean values of the clusters, guiding the distinction between the groupings.

This cluster analysis underscores the intricate relationship between a company's market standing and its media portrayal. While larger companies might often find themselves in the limelight, the nature of that attention—be it positive, negative, or neutral—can be influenced by a myriad of factors beyond just size.

In the subsequent sections, I will delve deeper into cluster 1 (Red), analyzing the specific characteristics and nuances that define each grouping without big companies.

## 7.3 Cluster Analysis without Big Companies

In this section, because the previous clustering included big companies, Cluster 2

(Microsoft Corp, Meta Platforms Inc, Alphabet Inc), Cluster 3 (Apple Inc) and Cluster 4 (NVIDIA Corp, Amazon.com Inc, Tesla Inc), they are excluded to delve deeper into the rest of the companies. Likewise, through a KMeans clustering approach, I have segmented the companies into four distinct groups.

**A graph with colored crosses

Description automatically generated with medium confidence**

Figure KMeans clustering of firm (Exclusing Outliers)

Cluster 0 (21 firms):

Companies in this cluster include:

* PDD Holdings Inc ADR
* Booking Holdings Inc
* Illumina Inc
* ... and 18 others.

Cluster 0 predominantly comprises technology and internet-based companies. These firms, despite their differences in market capitalization, exhibit similar sentiment scores, signifying a shared news sentiment trend. Notably, Cluster 0 tends to have sentiment scores higher than those in Cluster 3.

Cluster 1 (2 firms):

Companies in this cluster:

* Cisco Systems Inc
* Comcast Corp

Cluster 1 is minimal, consisting of two major tech companies. Both are renowned for their vast consumer and enterprise networks. Their influential positions in the tech industry might be the reason they share similar market sentiments.

Cluster 2 (6 firms):

Companies in this cluster include:

* ASML Holding NV
* Adobe Inc
* Broadcom Inc
* ... and 3 others.

Primarily involved in technology, software, and hardware development, these firms' shared industry focus might explain their grouping in terms of market capitalization and sentiment score.

Cluster 3 (62 firms):

Companies in this cluster include:

* Lucid Group Inc
* Atlassian Corp
* Constellation Energy Corp
* ... and 59 others.

Cluster 3, the most extensive cluster, encompasses a mix of companies from various sectors, including technology, energy, finance, and more. Despite the diversity, this cluster has the lowest sentiment scores among all clusters, suggesting specific underlying factors influencing their media portrayal.

In summary, the clustering results suggest that while industry focus plays a role in determining shared sentiment trends among companies, market capitalization also significantly influences how companies are grouped based on their news sentiment scores.

# 7. Feature Selection

(Still thinking whether I should do or not)

# 8. Model Evaluation

The predictive modeling of stock returns, especially in the volatile financial climate shaped by the COVID-19 pandemic, necessitates the utilization of robust machine learning algorithms. This section will elucidate the methodologies and results of four chosen models: Random Forest, XGBoost, LightGBM, and CatBoost.

## 8.1 Why AUC?

AUC (Area Under the Curve) is suitable in the financial area because it provides a comprehensive measure of the performance of classification models, particularly in the context of imbalanced datasets, which are common in finance. AUC is a metric that evaluates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across different classification thresholds. This makes it a valuable tool for assessing the performance of models in predicting financial risks, such as credit default, fraud, and stock performance.

In a study by Gou et al. (2022), the authors used random forest and XGBoost models to evaluate the default risk of personal financial loans and found that both models had high AUC values, indicating good learning and prediction capabilities.

Another study by Kim et al. (2023) proposed a Geometric Mean-based Boosting (GMBoost) algorithm to resolve class imbalance problems in finance, such as bankruptcy, card insolvency, and card fraud. The authors found that GMBoost outperformed conventional boosting algorithms in terms of AUC, demonstrating its effectiveness in handling imbalanced datasets.

Moreover, a study by Phongmekin and Jarumaneeroj (2018) used financial ratios and company's industry data of stocks in the finance sector of the Stock Exchange of Thailand to construct classification models predicting stock performance. The authors explored various classification techniques, including Logistic Regression (LR), Decision Tree (DT), Linear Discriminant Analysis (LDA), and K-Nearest Neighbor (KNN), and evaluated their performances using AUC. They concluded that all methods were comparatively good, with LR and LDA being the most useful classifiers for risk-averse investors.

In summary, AUC is a suitable metric in the financial area because it provides a comprehensive evaluation of classification models, particularly in the context of imbalanced datasets, which are common in finance. AUC has been successfully applied in various financial studies, demonstrating its effectiveness in assessing the performance of models in predicting financial risks and stock performance.

## 8.1 Random Forest

### 8.1.1 Theoretical Overview

Random Forest is an ensemble learning method that combines multiple decision trees to produce a more generalized and accurate prediction. The underlying mechanism is rooted in the "bagging" approach, where multiple weak learners (in this case, decision trees) combine to form a robust model.

Given a training set with responses , the algorithm for Random Forest is:

1. For to (where is the number of trees):

* Draw a bootstrap sample of size from the training data.
* Grow a decision tree​ to the bootstrapped data by recursively repeating the following process for each terminal node of the tree, until the node size is less than the threshold:
  + Select variables randomly from the predictors.
  + Split the node using the variable that provides the best split, according to some objective function (e.g., minimizing the Gini coefficient).

The forest's output for a test observation is the mode of the outputs of the trees.

### 8.1.2 Methodology

In the context of this study:

* Features: All columns except 'Date', 'PositiveReturn', and 'Ticker'.
* Target Variable: 'PositiveReturn'.
* Data Split: 80% training, 20% test.
* Hyperparameters: Number of trees () = 100.

In machine learning models, parameters refer to the internal variables optimized during training. On the other hand, hyperparameters encompass external configurations preset by a researcher before training commences (Bergstra & Bengio, 2012). These hyperparameters determine the model's structure and behavior and remain unaltered during the training phase. The significance of aptly chosen hyperparameters is highlighted by their considerable impact on the model's performance, underscoring the importance of hyperparameter tuning in machine learning processes (Snoek, Larochelle, & Adams, 2012).

A quintessential hyperparameter in tree-based ensemble models, such as Random Forests, is the 'number of trees', commonly represented as . This parameter delineates the count of individual decision trees constructed within the ensemble. Every tree within this ensemble furnishes a prediction. In classification tasks, the mode of these predictions, or the most frequent class, culminates as the ensemble's final prediction. An augmented number of trees generally bolsters the robustness and precision of predictions, particularly in scenarios with abundant features. This augmentation diminishes the model's variance, thereby mitigating overfitting (Breiman, 2001). Nevertheless, an escalation in trees also amplifies computational demands and may result in diminishing performance returns. In the context of this study, the ensemble was configured with 100 trees, harmonizing computational efficiency with predictive accuracy.

### 8.1.3 Results

To quantitatively assess the performance, we utilized the Area Under the Curve (AUC) score, given by:

where:

* is the true positive rate at threshold *.*
* is the false positive rate at threshold *.*

A visual representation of the AUC scores for Random Forest across different tickers is given below:

A colorful lines on a white background

Description automatically generated

Figure AUC scores for Random Forest by ticker

From the results, it's evident that while certain tickers like 'NFLX', 'GILD', and 'VRTX' yielded promising AUC scores, others, particularly 'META', 'ASML', and 'AZN', were less satisfactory. The diverse AUC scores across tickers suggest that the predictability of stock returns might be intrinsically linked to individual tickers.

### 8.1.4 Discussion

Random Forest's mechanism of using multiple trees and bootstrapped samples inherently makes it resistant to overfitting. Each tree's vulnerability to overfitting due to noise is mitigated by the averaging process. However, its performance can vary based on the depth of the trees, number of trees, and the number of features considered at each split.

In our analysis, the AUC scores were variable across tickers, suggesting that while Random Forest was adept at discerning patterns for certain stocks, it was less effective for others. This could be attributed to the inherent noise in financial data, external market factors not captured in our features, or the non-stationarity of stock return dynamics.

## 8.2 XGBoost

### 8.2.1 Theoretical Overview

XGBoost, short for eXtreme Gradient Boosting, is an advanced implementation of the gradient boosting algorithm. It has gained immense popularity due to its computational efficiency and capability to achieve superior results.

Gradient boosting is an ensemble technique wherein new models are trained to predict the residuals or errors of prior models. The XGBoost algorithm specifically utilizes decision trees as its base models.

Given a differentiable loss function where is the true label and is the prediction, the objective in each iteration is to find a function that minimizes:c

where:

* is the regularization term.
* is the number of iterations.

XGBoost incorporates both *L1* (Lasso regression) and *L2* (Ridge regression) regularization components in its objective function. This regularization helps in reducing overfitting, making XGBoost more robust than simple gradient boosting.

### 8.2.2 Methodology

For our analysis:

* Features: All columns except 'Date', 'PositiveReturn', and 'Ticker'.
* Target Variable: 'PositiveReturn'.
* Data Split: 80% training, 20% test.
* Hyperparameters: use\_label\_encoder=False, evaluation metric = 'auc', early stopping rounds = 10.

Within the XGBoost methodology, several hyperparameters are noteworthy:

The use\_label\_encoder=False configuration indicates that XGBoost will not employ label encoding for the target variable, suggesting the target is pre-encoded.

The evaluation metric = 'auc' denotes that the model optimizes for the Area Under the Receiver Operating Characteristic (ROC) Curve. This metric gauges binary classification performance, where a value closer to 1 suggests superior classification. An AUC of 0.5 denotes the model's predictions are equivalent to random guessing.

With early stopping rounds = 10, the model's training ceases if there's no improvement in the 'auc' metric after 10 iterations. This serves as a guard against overfitting, halting the model training when performance on validation data starts to wane.

These hyperparameters collectively enhance the model's training regimen, striking a balance between performance and the risk of overfitting.

### 8.2.3 Results

The efficacy of XGBoost was gauged using the AUC score, as previously discussed.

A screen shot of a screen

Description automatically generated

Figure AUC scores for XGBoost by ticker

The resultant AUC scores for XGBoost consistently outperformed those of Random Forest for a majority of the tickers. It was observed that the model performed exceptionally well for tickers like 'NFLX', 'CMCSA', 'BIIB', and 'SIRI', showcasing its proficiency in diverse stock scenarios.

### 8.2.4 Discussion

XGBoost's success can be attributed to several factors. The integration of regularization in its objective function inherently penalizes complex models, thereby reducing overfitting. Additionally, its ability to handle missing data, its efficient implementation of the boosting algorithm, and its flexibility in defining custom optimization objectives and evaluation criteria make it apt for diverse datasets and challenges.

For our dataset, XGBoost's consistent superior performance suggests that its capability to model non-linear relationships and intricacies in the data is well-suited to the task of predicting stock returns based on news sentiment during the COVID-19 era.

## 8.3 LightGBM

### 8.3.1 Theoretical Overview

LightGBM, or Light Gradient Boosting Machine, is a gradient boosting framework that employs tree-based learning algorithms. It has been designed to be more efficient than its counterparts, achieving faster training speeds and consuming lower memory.

A unique property of LightGBM is its utilization of histogram-based algorithms, which divide continuous feature values into discrete bins. This method speeds up the training process and reduces memory usage.

The objective function for LightGBM is similar to XGBoost, incorporating both the loss function and a regularization term:

where:

* represents the regularization term.
* denotes the number of iterations.

### 8.3.2 Methodology

For the LightGBM model:

* Features: All columns with the exception of 'Date', 'PositiveReturn', and 'Ticker'.
* Target Variable: 'PositiveReturn'.
* Data Split: 80% for training, 20% for testing.
* Hyperparameters: Boosting type set to "goss", a max depth of 5, and random state set to 0.

The boosting type "goss" refers to Gradient-based One-Side Sampling. In this method, LightGBM maintains the data distribution by using all the negative gradients but performs random sampling on the instances with positive gradients.

### 8.3.3 Results

The model's performance was gauged using the AUC score, the same metric employed for the previous models.

A colorful striped chart with a black border

Description automatically generated with medium confidence

Figure AUC scores for LightGBM by ticker

From our results, LightGBM showed competitive AUC scores, being particularly adept with tickers like 'GILD', 'BIIB', and 'EXC'. However, for some tickers, such as 'META', 'ASML', and 'GEHC', it didn't fare as well as expected.

### 8.3.4 Discussion

LightGBM's histogram-based approach provides an edge in computational efficiency, allowing it to handle larger datasets with ease. Moreover, its ability to manage categorical features natively, coupled with its advanced regularization, helps in curbing overfitting.

For our dataset, the model showed considerable prowess but had specific areas of weakness. This might indicate that while LightGBM's approach is generally effective, certain stock return dynamics influenced by news sentiment during the COVID-19 period might be better captured by other models.

## 8.4 CatBoost

### 8.4.1 Theoretical Overview

CatBoost, an acronym for "Category" and "Boosting", is a gradient boosting library developed by Yandex. It is specifically designed to handle categorical features without the need for extensive preprocessing such as one-hot encoding. This makes it particularly suitable for datasets with a high number of categorical attributes.

The objective function of CatBoost is an iterative refinement, employing a loss function and a regularization term:

where:

* is the regularization term.
* is the number of iterations.
* is a regularization coefficient.

One of CatBoost's standout features is its treatment of categorical variables. It employs a technique known as "ordered boosting", wherein it uses statistics from previously seen data in the learning process and avoids potential target leakage.

### 8.4.2 Methodology

For the CatBoost model:

* Features: All columns excluding 'Date', 'PositiveReturn', and 'Ticker'.
* Target Variable: 'PositiveReturn'.
* Data Split: 80% for training and 20% for testing.
* Hyperparameters: Iterations set to 1000, learning rate at 0.01, and verbose mode turned off for cleaner output.

### 8.4.3 Results

The performance of the CatBoost model was gauged using the AUC score, consistent with the evaluation of previous models.



Figure AUC scores for CatBoost by ticker

The results indicate that CatBoost displayed strong proficiency with tickers like 'GILD', 'BIIB', and 'SIRI'. On the contrary, it didn't achieve high scores for tickers such as 'PANW', 'GEHC', and 'DDOG'.

### 8.4.4 Discussion

CatBoost's unique approach to handling categorical data allows it to often outperform other gradient boosting models, especially when the data contains a significant number of categorical attributes. Its in-built regularization, efficient handling of overfitting, and the ordered boosting technique make it a robust choice for various applications.

In the context of our dataset, CatBoost demonstrated promising results. However, its performance varied across tickers. This suggests that while CatBoost's unique techniques offer distinct advantages, the intricate interplay between stock returns and news sentiment during the pandemic era might be better grasped by some models over others.

## 8.5 AUC score in Each Model

To provide a comprehensive understanding of the relative strengths and weaknesses of the four models in the context of our dataset, I employed a comparative analysis. This involved juxtaposing the AUC scores of the models across all tickers, thereby offering a bird's-eye view of their performances.

Table Results of AUC scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Max AUC** | **Min AUC** | **Mean AUC** | **Median AUC** |
| **Random Forest** | 0.8119747899159660 | 0.5270964691046660 | 0.6773262130929950 | 0.677745245825603 |
| **XGBoost** | 0.9107142857142860 | 0.5805366483897400 | 0.7526194751681700 | 0.7519644741193330 |
| **LightGBM** | 0.8396358543417370 | 0.4333333333333330 | 0.6481785020287460 | 0.6392358095839580 |
| **CatBoost** | 0.8416610132054340 | 0.470279086625589 | 0.6209455609252810 | 0.6113474506838090 |

This table presents a succinct comparison of AUC scores obtained from each model for every ticker. Key insights are:

Overall Performance:

XGBoost consistently stands out, achieving superior AUC scores for a majority of the tickers. LightGBM and Random Forest display a more varied performance, with their efficacy differing noticeably across tickers. CatBoost's results closely mirror those of Random Forest and LightGBM but are generally outperformed by XGBoost.

Consistency:

XGBoost showcases a relatively stable performance across various tickers, suggesting its robustness in capturing the nuances of the dataset. In contrast, Random Forest and LightGBM demonstrate more variability in their results.

Top-Performing Tickers:

Certain tickers like 'NFLX', 'GILD', 'BIIB', and 'SIRI' consistently achieve commendable AUC scores across most models. This could indicate that the features related to these tickers are particularly predictive, or the data for them is inherently more distinguishable.

Underperforming Tickers:

Some tickers, such as 'META', 'ASML', 'INTU', and 'AZN', regularly register lower AUC scores irrespective of the model employed. This might be indicative of less predictive features or potentially noisy data associated with these tickers.

The comparative analysis illuminates the intricate dynamics between stock return predictability, news sentiment, and the inherent characteristics of the machine learning models. It is evident that while certain models excel in specific contexts, there isn't a one-size-fits-all solution. The diversity in AUC scores across tickers for all models underscores the notion that stock return predictability could be deeply intertwined with individual tickers. This revelation further bolsters the strategy of devising ticker-specific models.

From a broader perspective, this comparative study underscores the importance of model selection in financial machine learning tasks. While certain algorithms may be theoretically sound, their practical application requires careful consideration of the dataset's characteristics and the specific problem context.

## 8.7 Comparative Analysis of Top 10 Firms by AUC Scores Across Models

The in-depth assessment of the top 10 companies, as measured by Area Under the Curve (AUC) scores across Random Forest, XGBoost, LightGBM, and CatBoost models, offers insightful perspectives into the commonalities and variances between these models in terms of specific companies they rank highly.

Commonalities:

* GLOBALFOUNDRIES Inc consistently ranks high in the Random Forest, XGBoost, and LightGBM models. This suggests that the company's financial or operational dynamics might be inherently stable and discernible.
* Biogen Inc's consistent high rank across all models suggests that its financial patterns might have aspects that are universally recognized and effectively captured by various modeling techniques.
* Sirius XM Holdings Inc, Exelon Corp, Gilead Sciences Inc, Warner Bros Discovery Inc, and Seagen Inc are other firms that consistently appear in the top 10 lists of most models, emphasizing their stable representation across modeling techniques.

Deep Dive into Commonalities:

The consistent high ranking of specific companies across various models can be attributed to both industry dynamics and inherent financial stability. For instance, firms such as Biogen Inc and Gilead Sciences Inc operate within the pharmaceutical and biotech industries (Schiraldi, 2014). These sectors stand out due to their long and extensive research and development phases, often culminating in several years of work before a product is ready for the market. The predictability in these sectors is further augmented by a rigorous regulatory environment. With stringent protocols for clinical trials and product approvals, there's an inherent stability in their operational dynamics (Choudhury, 2016). Once a product gains regulatory approval, it typically enjoys a period of market exclusivity, courtesy of patent protections. Furthermore, the pharmaceutical and biotech sectors are known for their substantial reinvestments into R&D, ensuring a steady flow of innovations. The combination of continuous innovation and the extended lifecycle of their products often translates into stable financial trajectories, making them more amenable to predictive modeling (Ottoo, 2018).

Differences:

* Firms like Xcel Energy Inc and American Electric Power Co Inc are unique to the XGBoost model, while Autodesk Inc is exclusive to LightGBM's top 10. This suggests potential model-specific sensitivities or feature importance variations.
* Dollar Tree Inc's presence solely in the Random Forest's top 10 might hint at unique trends or patterns that this specific model captures.
* Booking Holdings Inc and Walgreens Boots Alliance Inc are unique to the CatBoost model's top 10, indicating a different feature weighting or recognition pattern inherent to the CatBoost algorithm.

Deep Dive into Differences:

Autodesk Inc, a leader in software solutions for architecture and construction, is uniquely highlighted in LightGBM's top ten. The software industry, with its rapid innovation cycles and shifting consumer preferences, might present specific trends that the LightGBM algorithm identifies more effectively, perhaps due to its focus on leaf-wise tree growth and dealing with categorical features.

Dollar Tree Inc's solitary presence in Random Forest's top ten is intriguing. As a prominent player in the discount store industry, its financial dynamics, influenced by consumer spending patterns, inventory management, and real estate decisions, might have unique patterns that the ensemble nature of Random Forest can capture more effectively (Khare et al.).

Lastly, the inclusion of companies like Booking Holdings Inc, a digital travel industry leader, and Walgreens Boots Alliance Inc, a pharmacy and retail giant, solely in CatBoost's top ten points towards the model's unique handling of categorical data and iterative refinement. These firms, operating in sectors marked by intense competition and rapidly evolving business models, might exhibit data patterns that are more coherently deciphered by the CatBoost algorithm, given its ability to focus on feature interactions.

## 8.8 Comparative Analysis of Bottom 10 Firms by AUC Scores Across Models

A screenshot of a graph

Description automatically generated

Figure Top 10 frims and bottom 10 Firms in AUC score across models

Commonalities:

* Firms such as ASML Holding NV and AstraZeneca PLC ADR consistently appear in the bottom 10 across multiple models, suggesting challenges in predicting their financial patterns.
* Tech and e-commerce companies like PDD Holdings Inc ADR and MercadoLibre Inc frequently rank in the bottom across models, indicating potential complexities in their financial dynamics.

Deep Dive into Commonalities:

The repeated appearance of companies like ASML Holding NV and AstraZeneca PLC ADR in the bottom 10 across various models suggests inherent complexities in these firms' financial metrics or operations that make them challenging to predict. ASML Holding NV, a leader in semiconductor manufacturing, operates in an industry characterized by rapid technological advancements and intense competition, which can introduce unpredictability in financial metrics. Similarly, AstraZeneca, a pharmaceutical giant, might face unpredictability due to the outcome-based nature of drug trials and the shifting regulatory landscape (Lexchin, 2012).

Moreover, the tech and e-commerce sectors, represented by firms like PDD Holdings Inc ADR and MercadoLibre Inc, are known for their swift pace of innovation and changing market dynamics. The models might struggle to keep up with the rapid changes in these sectors, or the features used might not adequately capture the intricate dynamics of these industries.

Differences:

* Companies like Datadog Inc and Palo Alto Networks Inc are uniquely positioned in the bottom 10 of the CatBoost model, whereas they don't appear in the other models' lists.
* Some firms, like GE HealthCare Technologies Inc, are specific to the bottom rankings of one model, in this case, LightGBM, suggesting model-specific sensitivities.

Deep Dive into Differences:

The unique positioning of Datadog Inc, a cloud monitoring service, and Palo Alto Networks Inc, a cybersecurity solution provider, in CatBoost's bottom 10, indicates potential sensitivities or feature weightings specific to the CatBoost algorithm. The intricate dynamics of cloud services and cybersecurity, which evolve rapidly with technological advancements, might be challenging for CatBoost to decipher, given its approach to handling categorical data and iterative refinement.

On the other hand, the exclusive appearance of GE HealthCare Technologies Inc in LightGBM's bottom 10 suggests that this model might have specific challenges in capturing the dynamics of the healthcare tech sector. There could be nuances in the data of this company that don't resonate well with LightGBM's algorithm. This is consistent with findings by Mahussin et al. (2021) who confirmed significant differences in the volatility of healthcare and technology sectors before and during the Covid-19 outbreak. Their study suggests that the healthcare sector might be particularly challenging to predict in terms of stock market return, especially when considering the period influenced by COVID-19.

## 8.9 Feature Importance Across Models

The predictive ability of machine learning models is significantly governed by the importance of features used in the model. The study, leveraging models like Random Forest, XGBoost, LightGBM, and CatBoost, seeks to understand the salience of features like stock indicators, sentiment scores, and COVID-19 data in forecasting positive stock returns.

### 8.9.1 How Feature Importance is computed

The importance of a feature is typically computed based on the frequency, depth, or improvement it brings when used in trees. In the context of tree-based models, such as Random Forest, XGBoost, LightGBM, and CatBoost, feature importance is often gauged by how frequently a feature is used to split the data, the depth at which the feature is used, or the improvement in purity it brings about. In this study, *feature\_importances\_* attribute of Python, common to these models, provides a normalized value indicating the relative importance of each feature when making a prediction.

* Random Forest: This model computes feature importance by looking at how much each feature decreases the impurity. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.
* XGBoost: The importance is computed as the average gain of the feature when it is used in trees.
* LightGBM: This gradient-boosting framework computes importance as the number of times a feature is used to split the data across all trees.
* CatBoost: Importance in CatBoost is calculated as the average difference between the prediction values with and without the feature.

In the subsequent sections, Figure 17 shows the results to delve into the specific importance scores of features as derived from each of these models.

A group of blue and white bars

Description automatically generated

Figure Feature importance across models

### 8.9.2 Feature Importance Across Models

This section is about the specific importance scores of features across models based on the Figure 17.

Random Forest:

Random Forest, being an ensemble of decision trees, offers insights into the importance of features based on their frequency and position in the trees. The model attributes significant importance to stock indicators such as *Open*, *Close*, *Adj* *Close*, and *Volume*, with importance scores of approximately 0.120, 0.104, 0.110, and 0.108, respectively. The sentiment score, a reflection of daily news sentiment, holds a significance of 0.071. *New\_Covid\_Cases*, a representative of pandemic dynamics, has been valued at 0.091. Strikingly, *MarketCap* and *PandemicPhase* are bereft of any importance in this model.

XGBoost:

XGBoost operates by assigning and optimizing weights to features, offering a gradient-boosting mechanism. The model prioritizes stock indicators such as *Open*, *Close*, and *Volume* with importance scores of 0.145, 0.118, and 0.093, respectively. *Sentiment\_Score* carries a weight of 0.067, while *New\_Covid\_Cases* has an importance of 0.076. Once again, *MarketCap* and *PandemicPhase* are rendered inconsequential.

LightGBM:

LightGBM, a gradient boosting framework, returns feature importance as the number of times a feature is used to split the data. *Volume*, with a score of 101, and *New\_Covid\_Cases*, with a score of 88, emerge as dominant features. *Sentiment\_Score*, with a count of 38, underscores its significance in the model. Consistent with the other models, *MarketCap* and *PandemicPhase* fail to influence the model.

CatBoost:

CatBoost, designed to handle categorical data efficiently, reveals *Volume* and *Volatility* as paramount features, with scores of approximately 13.14 and 14.68, respectively. *Sentiment\_Score*, too, is substantial, with a weight of 9.07. The pattern of *MarketCap* and *PandemicPhase* being non-influential persists.

### 8.9.3 Analyzing the Implications

Stock indicators consistently emerge as dominant features across all models, emphasizing their foundational role in determining stock dynamics. Features such as *Open*, *Close*, and *Volume* not only offer insights into daily stock behavior but also provide a snapshot of the stock's financial health, investor sentiment, and market liquidity. This finding aligns with Aouadi et al. (2013), who demonstrated that investor attention is strongly correlated with trading volume.

The *Sentiment\_Score*, although capturing the daily sentiment towards a stock, does not exhibit as pronounced an influence as some other features. While it quantifies investor perceptions and market reactions to news and events, its relative importance is overshadowed by the more traditional stock indicators. However, in the grander scheme of stock return predictions, its impact, albeit present, isn't as substantial as one might anticipate, especially when juxtaposed with attributes like *Open* or *Volume*. This finding differs from Shi and Ho (2020) and other academic papers that only measure the impact of news sentiment scores on stock market returns.

The glaring non-importance of *MarketCap* across models suggests that a company's size, often seen as a reflection of its stability and resilience to market shocks, isn't a significant predictor in this context. This could be attributed to the fact that during tumultuous times, especially like those during the pandemic, even large-cap companies faced unprecedented challenges, rendering their market capitalization less indicative of stock return behavior.

Similarly, the PandemicPhase, despite its binary delineation of pre and post-pandemic periods, fails to make a mark in the predictive models. This could be because the granular and continuous data provided by *New\_Covid\_Cases* captures the pandemic's influence on stock returns more effectively. The daily count of fresh COVID-19 cases offers a more dynamic representation of the pandemic's progression and its potential impact on investor sentiment and market dynamics.

In conclusion, while various features play a role in predicting positive stock returns, some emerge more influential than others. The most prominent feature across models is the *Volume*, emphasizing the importance of liquidity and trading activity in forecasting stock returns. This is closely followed by stock indicators such as *Open* and *Close*. *New\_Covid\_Cases*, representing the pandemic's daily dynamics, also stands out as a critical feature. On the other hand, attributes like *MarketCap* and *PandemicPhase*, despite their conceptual significance, do not wield substantial influence in the models used for this analysis.

# 9. Discussion

**9.1 Model Efficacy and XGBoost's Superiority**

The comparative evaluation across the machine learning models accentuated the distinct prominence of XGBoost in terms of consistent superior performance. At the heart of XGBoost's efficiency is its gradient boosting framework, which optimizes for both system performance and algorithmic speed, as expounded by Chen & Guestrin (2016). Its capacity to handle sparse data, missing values, and its inherent regularization makes it aptly suited for financial datasets that often encompass intricate non-linear relationships. In the financial realm, where markets are inherently volatile, the adaptability of XGBoost offers a significant advantage. Its gradient boosting mechanism accentuates the effects of new additions (trees in this instance), which in essence adapts to the errors of the preceding sequence of trees. This iterative refinement underlines its superior efficacy, especially in the context of stock returns during the tumultuous phase of the COVID-19 pandemic.

## 9.2 Variability Among Models

While XGBoost stood out distinctly, it's pivotal to understand the nuanced performances of the other models: Random Forest, LightGBM, and CatBoost. These models, though effective in their capacities, displayed distinct variability across different stock tickers. Breiman (2001) in his seminal work on Random Forests, elucidated the model's ability to reduce overfitting through the creation of an "ensemble" of decision trees. However, in the intricate domain of stock predictions, where multiple factors interplay, even a robust ensemble can miss out on capturing some nuanced relationships. LightGBM and CatBoost, newer entrants into the machine learning realm, bring advanced gradient boosting methodologies. Their variability in performance underscores a core principle in financial econometrics: the idiosyncratic nature of stocks. Each stock, governed not just by market dynamics but also sectoral and firm-specific news, can resonate differently with algorithmic structures, leading to these observed variabilities.

## 9.3 Limitations and Scope for Future Research

Every analytical endeavor, while shedding light on certain aspects, also comes with its set of constraints. The study, though comprehensive, focuses primarily on NASDAQ-listed companies. This geographical and exchange-centric concentration might not capture the global intricacies of stock behaviors, especially during a worldwide event like the COVID-19 pandemic. The sentiment scores, albeit aggregated from a range of sources, provide a macro view of sentiment, potentially overlooking more granular sentiment fluctuations on a day-to-day basis. Furthermore, while the models are trained on a decade of data, the inherent unpredictability of stock markets means that past performance is not always indicative of future results. This principle is well-embedded in financial literature and serves as a reminder of the tentative nature of stock return predictions.

# 10. Ethics

When dealing with financial data, the ethical considerations are paramount. As we delve deeper into the intricate tapestry of stock market reactions to news sentiment, particularly during tumultuous times such as the COVID-19 pandemic, we must critically evaluate the ethical dimensions of our work. This section aims to address these concerns, providing a holistic understanding of the ethical challenges and methodological ramifications, and suggesting pathways for responsible innovation.

## 10.1 Ethical Foundations and Risk Management

An in-depth appreciation of computer ethics is imperative for any computational study. The digital landscape, replete with vast datasets, is rife with potential pitfalls. While the present study aims to provide insights into stock market dynamics, there's an underlying responsibility to ensure that data interpretation doesn't lead to misleading or harmful financial advice. Furthermore, the utilization of sentiment data from news outlets raises concerns about the potential for bias, misrepresentation, or oversimplification. Recognizing these challenges, a comprehensive plan has been laid out to manage and mitigate potential ethical risks. Central to this strategy is the adherence to the AREA (Anticipate, Reflect, Engage, Act) framework, which provides a dynamic approach to navigating ethical considerations (EPSRC, n.d.).

## 10.2 Societal Implications and Ethical Risks

Beyond the immediate realm of financial markets, the study's findings have broader societal implications. In an age where news is rapidly disseminated and consumed, understanding its impact on market sentiments is crucial. However, there's a risk that such analyses might inadvertently prioritize or devalue certain news sources, potentially leading to monopolization or undermining of certain media outlets. Moreover, the ethical dimensions extend to the potential for creating feedback loops, where market reactions to news sentiment might influence subsequent reporting, leading to cyclical biases or self-fulfilling prophecies.

## 10.3 Methodological Ethical Challenges

The intricate interplay between news sentiment and stock market reactions presents unique methodological challenges. The potential for confounding variables, especially during unprecedented global events like the COVID-19 pandemic, is high. There's a risk of oversimplifying complex relationships or drawing premature conclusions. Additionally, the reliance on sentiment scores, which are inherently subjective, introduces potential biases. These methodological concerns not only impact the study's validity but also raise ethical issues regarding the responsible interpretation and presentation of findings.

## 10.4 Responsible Innovation and Future Directions

Responsible innovation is the cornerstone of any research endeavor, especially in areas with profound societal implications. It emphasizes the need to consider the broader impacts of research and innovation, striving for positive societal and economic benefits while mitigating unintended negative consequences (UK Research and Innovation, n.d.). Moreover, the UK government framework emphasizes the importance of continuous evaluation, transparency in methodologies, and effective public engagement, thereby reinforcing the principles of the AREA framework and ensuring that research remains accountable, transparent, and ethically grounded (UK Government, 2020). To achieve this responsible innovation, the following four ethical recommendations are proposed for future research in this domain based on AREA framework and the UK government framework:

1. Transparency: Ensure that methodologies, particularly sentiment analysis algorithms, are transparent and open to scrutiny. This will foster trust and facilitate peer evaluations.
2. Engagement: Engage with a diverse array of stakeholders, including financial experts, media representatives, and the general public, to gather varied perspectives and address potential biases.
3. Reflection: Continuously reassess the research's ethical dimensions, staying vigilant to emerging challenges and adapting methodologies accordingly.
4. Action: Prioritize responsible dissemination of findings, emphasizing the preliminary nature of insights and potential limitations. Furthermore, collaborate with media outlets and financial institutions to ensure that the research's implications are understood and applied responsibly.

# 11. Conclusion

The complexity of stock markets, amplified by unpredictable events such as the COVID-19 pandemic, demands rigorous and adaptive modeling techniques. The study, grounded in the realm of financial machine learning, embarked on an exploration to assess the efficacy of four prominent machine learning models in predicting stock returns using news sentiment analysis, especially during the challenging times of the COVID-19 pandemic.

Addressing the first research question regarding the comparative performance of machine learning models during the pandemic, the results clearly identified XGBoost as the standout model. This model not only consistently outperformed its counterparts - Random Forest, LightGBM, and CatBoost - but also showcased remarkable resilience to the uncertainties of the pandemic. The model's robustness is evident in its ability to capture nonlinear relationships and handle the multifaceted influences of the COVID-19 period.

The second research question delved into feature importance, aiming to discern which attributes held the most sway in model predictions. Unsurprisingly, news sentiment emerged as a potent predictor across all models, reinforcing the idea that investor sentiment, shaped by the news, plays a crucial role in stock market dynamics. However, it's worth noting that other features, such as trading volume and market capitalization, also exhibited significant influence, suggesting that while sentiment is vital, it's the interplay of various factors that truly drives stock returns.

The third research question sought to ascertain the predictability of specific stocks. The analysis spotlighted stocks like 'NFLX', 'GILD', 'BIIB', and 'SIRI' as particularly predictable, hinting at intrinsic factors related to their business model or industry that might make them more susceptible to sentiment analysis predictions. While this discovery is noteworthy, it also emphasizes the necessity for further research into the nuances of individual stocks and sectors.

In conclusion, the interrelationship between news sentiment and stock market movements, while always significant, has gained paramount importance in the wake of events that dominate global news cycles, such as the COVID-19 pandemic. Traditional models, though valuable, have shown limitations in their predictive capabilities in such scenarios. Machine learning models, especially ones like XGBoost, offer a promising avenue for financial forecasting in these tumultuous times. They harness the power of news sentiment, coupled with other influential factors, to make more accurate predictions.

However, while the study provides valuable insights, it is imperative to approach the results with measured optimism. Financial markets are complex entities, influenced by a myriad of factors. While machine learning models offer an advanced toolset for predictions, they are not infallible. Future research should aim to refine these models further, incorporate more diverse data sources, and explore the integration of deep learning techniques for enhanced accuracy.

# References

Aouadi, A., Arouri, M., & Teulon, F. (2013). Investor attention and stock market activity: Evidence from France. Economic Modelling, 35, 674-681.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. Expert Systems with Applications, 36(3), 5932-5941.

Baker, S. R., Bloom, N., Davis, S. J., & Terry, S. J. (2020). COVID-induced economic uncertainty. *National Bureau of Economic Research.*

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. J*ournal of Computational Science*, 2(1), 1-8.

Bordino, I., Kourtellis, N., Laptev, N., & Billawala, Y. (2014, March). Stock trade volume prediction with yahoo finance user browsing behavior. In *2014 IEEE 30th International Conference on Data Engineering* (pp. 1168-1173). IEEE.

Breiman, L. (2001). *Random forests. Machine learning, 45*(1), 5-32.

Brooks, C. (2014). Introductory econometrics for finance. Cambridge university press.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chen, Y. W., Chou, R. K., & Lin, C. B. (2019). Investor sentiment, SEO market timing, and stock price performance. *Journal of Empirical Finance*, *51*, 28-43.

Choudhury, A. (2016). Architecture of an Integrated Regulatory Information Management Platform for Clinical Trials: A Case Study in Regulatory Information Management System Implementation. In *Software Innovations in Clinical Drug Development and Safety* (pp. 163-201). IGI Global.

Elliott, W. B., & Warr, R. S. (2003). Price Pressure on the NYSE and Nasdaq: Evidence from S&P 500 Index Changes. *Financial Management*, *32*(3), 85–99.

Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, 66(1), 67-97.

Engineering and Physical Sciences Research Council (EPSRC). (n.d.). Research integrity in healthcare technologies. Health Technologies Impact and Translation Toolkit. Retrieved August 27, 2023, from <https://www.ukri.org/councils/epsrc/guidance-for-applicants/what-to-include-in-your-proposal/health-technologies-impact-and-translation-toolkit/research-integrity-in-healthcare-technologies/responsible-research-and-innovation/>

Fahlenbrach, R., Rageth, K., & Stulz, R. M. (2020). How valuable is financial flexibility when revenue stops? Evidence from the COVID-19 crisis. *National Bureau of Economic Research*.

Gillam, R. A., Guerard Jr, J. B., & Cahan, R. (2015). News volume information: Beyond earnings forecasting in a global stock selection model*. International Journal of Forecasting, 31*(2), 575-581.

Gou, Q., Niu, H., Gao, W., Qin, B., & Li, J.Y. (2022). Research on financial loan default risk prediction based on integrated model. *International Conference on Cloud Computing, Internet of Things, and Computer Applications*.

Huang, J. (2018). The customer knows best: The investment value of consumer opinions. *Journal of Financial Economics, 128*(1), 164-182.

Jagwani, J., Gupta, M., Sachdeva, H., & Singhal, A. (2018, June). Stock price forecasting using data from Yahoo finance and analysing seasonal and nonseasonal trend. In *2018 Second international conference on intelligent computing and control systems (ICICCS)* (pp. 462-467). IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*.

Khare, R.R., Khurshid, A.A., Jain, A., & Shukla, S. (2022). Predictive Sensor for Biological Oxygen Demand in water using Active Learning based Random Forest Algorithm. *NeuroQuantology*.

Kim, M., Ahn, J., & Kim, Y. (2023). Performance optimization-based boosting algorithm for resolving class imbalance problems in finance. *The Korean Data Analysis Society*.

Kumar, R. S., Saviour Devaraj, A. F., Rajeswari, M., Julie, E. G., Robinson, Y. H., & Shanmuganathan, V. (2021). Exploration of sentiment analysis and legitimate artistry for opinion mining. *Multimedia Tools and Applications*, 1-16.

Lexchin, J. (2012). Those Who Have the Gold Make the Evidence: How the Pharmaceutical Industry Biases the Outcomes of Clinical Trials of Medications. *Science and Engineering Ethics*, 18, 247-261.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

Li, J., & Yang, C. (2017). The cross-section and time-series effects of individual stock sentiment on stock prices. *Applied Economics*, *49*(47), 4806-4815.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. *The Journal of Finance*, 66(1), 35-65.

Mahussin, N., Jaapar, A., & Mustafa, L. (2021). Volatility of Technology and Healthcare Sectors Before and During Covid-19 Pandemic. *Malaysian Journal of Science Health & Technology*.

Mazur, M., Dang, M., & Vega, M. (2021). COVID-19 and the march 2020 stock market crash. Evidence from S&P1500. *Finance research letters*, *38*, 101690.

Ottoo, R.E. (2018). Valuation of Corporate Innovation and the Pricing of Risk in the Biopharmaceutical Industry: The Case of Gilead. *Risk Management eJournal*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Peng, J., Zhang, J., & Gopal, R. (2022). The good, the bad, and the social media: financial implications of social media reactions to firm-related news. *Journal of Management Information Systems*, *39*(3), 706-732.

Qu, Z., & Perron, P. (2013). A Stochastic Volatility Model with Random Level Shifts and its Applications to S&P 500 and NASDAQ Return Indices. Capital Markets: Market Efficiency eJournal.

Phongmekin, A., & Jarumaneeroj, P. (2018). Classification Models for Stock's Performance Prediction: A Case Study of Finance Sector in the Stock Exchange of Thailand. *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 1-4.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Schiraldi, M.M. (2014). Innovations in Pharmaceutical Industry Innovations in the Pharmaceutical Industry Guest Editorial.

Shi, Y., & Ho, K. (2020). News sentiment and states of stock return volatility: Evidence from long memory and discrete choice models. *Finance Research Letters.*

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

Sousa, M. G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P. H., Fernandes, E. R., & Matsubara, E. T. (2019, November). BERT for stock market sentiment analysis. In 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI) (pp. 1597-1601). IEEE.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance, 62*(3), 1139-1168.

UK Government. (2020, September 16). Data Ethics Framework. GOV.UK. <https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020>

UK Research and Innovation (UKRI). (n.d.). Responsible innovation. Good research resource hub. Retrieved August 27, 2023, from <https://www.ukri.org/manage-your-award/good-research-resource-hub/responsible-innovation/>

Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Sandoval-Almazán, R. (2022). Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods. *Cognitive computation*, 1-16.

Xu, H. C., & Zhou, W. X. (2018). A weekly sentiment index and the cross-section of stock returns. *Finance Research Letters*, *27*, 135-139.

Yousaf, I., Youssef, M., & Goodell, J. W. (2022). Quantile connectedness between sentiment and financial markets: Evidence from the S&P 500 twitter sentiment index. *International Review of Financial Analysis, 83*, 102322.

# Appendix

Table 5 Companies with Their Tickers

|  |  |
| --- | --- |
| **Company name** | **Ticker** |
| **Microsoft Corp** | MSFT |
| **Apple Inc** | AAPL |
| **NVIDIA Corp** | NVDA |
| [**Amazon.com**](http://Amazon.com) **Inc** | AMZN |
| **Meta Platforms Inc** | META |
| **Tesla Inc** | TSLA |
| **Alphabet Inc** | GOOGL |
| **Alphabet Inc** | GOOG |
| **Broadcom Inc** | AVGO |
| **PepsiCo Inc** | PEP |
| **Costco Wholesale Corp** | COST |
| **Adobe Inc** | ADBE |
| **Cisco Systems Inc** | CSCO |
| **Netflix Inc** | NFLX |
| **Advanced Micro Devices Inc** | AMD |
| **Comcast Corp** | CMCSA |
| **T-Mobile US Inc** | TMUS |
| **Texas Instruments Inc** | TXN |
| **Intel Corp** | INTC |
| **Honeywell International Inc** | HON |
| **Intuit Inc** | INTU |
| **QUALCOMM Inc** | QCOM |
| **Intuitive Surgical Inc** | ISRG |
| **Amgen Inc** | AMGN |
| **Applied Materials Inc** | AMAT |
| **Starbucks Corp** | SBUX |
| **Booking Holdings Inc** | BKNG |
| **Analog Devices Inc** | ADI |
| **Mondelez International Inc** | MDLZ |
| **Gilead Sciences Inc** | GILD |
| **Automatic Data Processing Inc** | ADP |
| **Vertex Pharmaceuticals Inc** | VRTX |
| **Lam Research Corp** | LRCX |
| **PayPal Holdings Inc** | PYPL |
| **Regeneron Pharmaceuticals Inc** | REGN |
| **Palo Alto Networks Inc** | PANW |
| **Activision Blizzard Inc** | ATVI |
| **Micron Technology Inc** | MU |
| **CSX Corp** | CSX |
| **Synopsys Inc** | SNPS |
| **KLA Corp** | KLAC |
| **ASML Holding NV** | ASML |
| **Cadence Design Systems Inc** | CDNS |
| **Fortinet Inc** | FTNT |
| **O'Reilly Automotive Inc** | ORLY |
| **Monster Beverage Corp** | MNST |
| **Marriott International Inc/MD** | MAR |
| **Charter Communications Inc** | CHTR |
| **MercadoLibre Inc** | MELI |
| **Airbnb Inc** | ABNB |
| **NXP Semiconductors NV** | NXPI |
| **Marvell Technology Inc** | MRVL |
| **Dexcom Inc** | DXCM |
| **Cintas Corp** | CTAS |
| **Microchip Technology Inc** | MCHP |
| **Moderna Inc** | MRNA |
| **Lululemon Athletica Inc** | LULU |
| **Autodesk Inc** | ADSK |
| **PDD Holdings Inc ADR** | PDD |
| **Workday Inc** | WDAY |
| **PACCAR Inc** | PCAR |
| **American Electric Power Co Inc** | AEP |
| **Keurig Dr Pepper Inc** | KDP |
| **Kraft Heinz Co/The** | KHC |
| **IDEXX Laboratories Inc** | IDXX |
| **Copart Inc** | CPRT |
| **Paychex Inc** | PAYX |
| **ON Semiconductor Corp** | ON |
| **Exelon Corp** | EXC |
| **Old Dominion Freight Line Inc** | ODFL |
| **Biogen Inc** | BIIB |
| **AstraZeneca PLC ADR** | AZN |
| **Ross Stores Inc** | ROST |
| **GE HealthCare Technologies Inc** | GEHC |
| **Electronic Arts Inc** | EA |
| **Seagen Inc** | SGEN |
| **CoStar Group Inc** | CSGP |
| **GLOBALFOUNDRIES Inc** | GFS |
| **Xcel Energy Inc** | XEL |
| **Baker Hughes Co** | BKR |
| **Cognizant Technology Solutions Corp** | CTSH |
| **Fastenal Co** | FAST |
| **Verisk Analytics Inc** | VRSK |
| **Crowdstrike Holdings Inc** | CRWD |
| **Dollar Tree Inc** | DLTR |
| **Warner Bros Discovery Inc** | WBD |
| **Datadog Inc** | DDOG |
| **Constellation Energy Corp** | CEG |
| **Illumina Inc** | ILMN |
| **ANSYS Inc** | ANSS |
| **Align Technology Inc** | ALGN |
| **Atlassian Corp** | TEAM |
| **Walgreens Boots Alliance Inc** | WBA |
| **Diamondback Energy Inc** | FANG |
| **Enphase Energy Inc** | ENPH |
| **eBay Inc** | EBAY |
| **Zscaler Inc** | ZS |
| **Sirius XM Holdings Inc** | SIRI |
| **Zoom Video Communications Inc** | ZM |
| [**JD.com**](http://JD.com) **Inc ADR** | JD |
| **Lucid Group Inc** | LCID |