Executive Summary

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

Student ID : 720034950
Name : Kota Kobayashi

Introduction

In the contemporary financial environment, news sentiment's role in predicting stock market trends has been highlighted. This relationship became even more pronounced amidst the global disruption caused by the COVID-19 pandemic. With the increased dependency on digital news platforms and algorithmic trading during this period, news sentiment analysis has solidified its place as an essential tool in shaping investment decisions.

Problem Statement & Background:

Historically, numerous empirical studies have delved into the dynamics between news sentiment and stock returns. However, the COVID-19 pandemic has brought about unique challenges that haven't been previously tackled. Many existing research methodologies, tailored for the pre-pandemic world, now appear outdated. Past research, such as those by Cakra and Trisedya (2015), leaned on regression methodologies, while the current context sees the rise of more advanced models like CatBoost. This discrepancy in methodologies has led to an inconsistent research landscape. Moreover, the narrow scope of many studies, often focusing on specific sectors or broad indices, further exacerbates the fragmentation in the literature. This has left stakeholders, from financial experts to individual investors, facing a multitude of questions regarding model selection, feature importance, and the general applicability of sentiment analysis.

Adding another layer of complexity is the narrow focus of many studies, either targeting specific sectors, like big tech firms, or broad indices like the S&P 500. While a few have ventured into analyzing the NASDAQ, their limited numbers contribute to the existing research void (Lee, 2022). This patchwork of fragmented and sometimes

outdated research leaves stakeholders, from financial experts to individual investors, navigating a maze of confusion. They are left grappling with pivotal questions about optimal model selection, feature prioritization, and the broader applicability of sentiment analysis across a varied spectrum of firms.

Research Objectives:

To address these challenges and provide actionable insights, this research embarked on answering three pivotal questions:

- *Comparative Analysis:* How do present-day machine learning models perform in leveraging news sentiment for stock return predictions during the pandemic?

- *Feature Significance:* In these models, which variables are pivotal? Does news sentiment hold its ground against other influential factors like volume or the influence of COVID-19?

- *Company-Specific Variations:* How does sentiment analysis fare across different firms and sectors?

Key Findings:

- *Model Efficacy & Sectoral Specificity:* As shown in the table, XGBoost emerged as the standout performer. However, deeper exploration highlighted performance variations across sectors. Biotechnology firms, like Biogen Inc and Gilead Sciences Inc, exhibited heightened prediction accuracy. Their extended R&D cycles, strict regulatory environments, and continuous innovation resulted in predictable financial trajectories, making them ripe for predictive modeling. This granularity in results underscores the necessity of sector-specific modeling for nuanced predictions.

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

(author's own work)

- *Feature Importance & Sentiment Reliability:* The study emphasized the dominance of traditional stock indicators, especially Volume. Interestingly, sentiment score, despite its relevance, was found less influential than expected. This suggests that investment strategies should not be overly reliant on sentiment analysis alone. A fusion of sentiment scores with traditional stock indicators is paramount for a holistic prediction model.

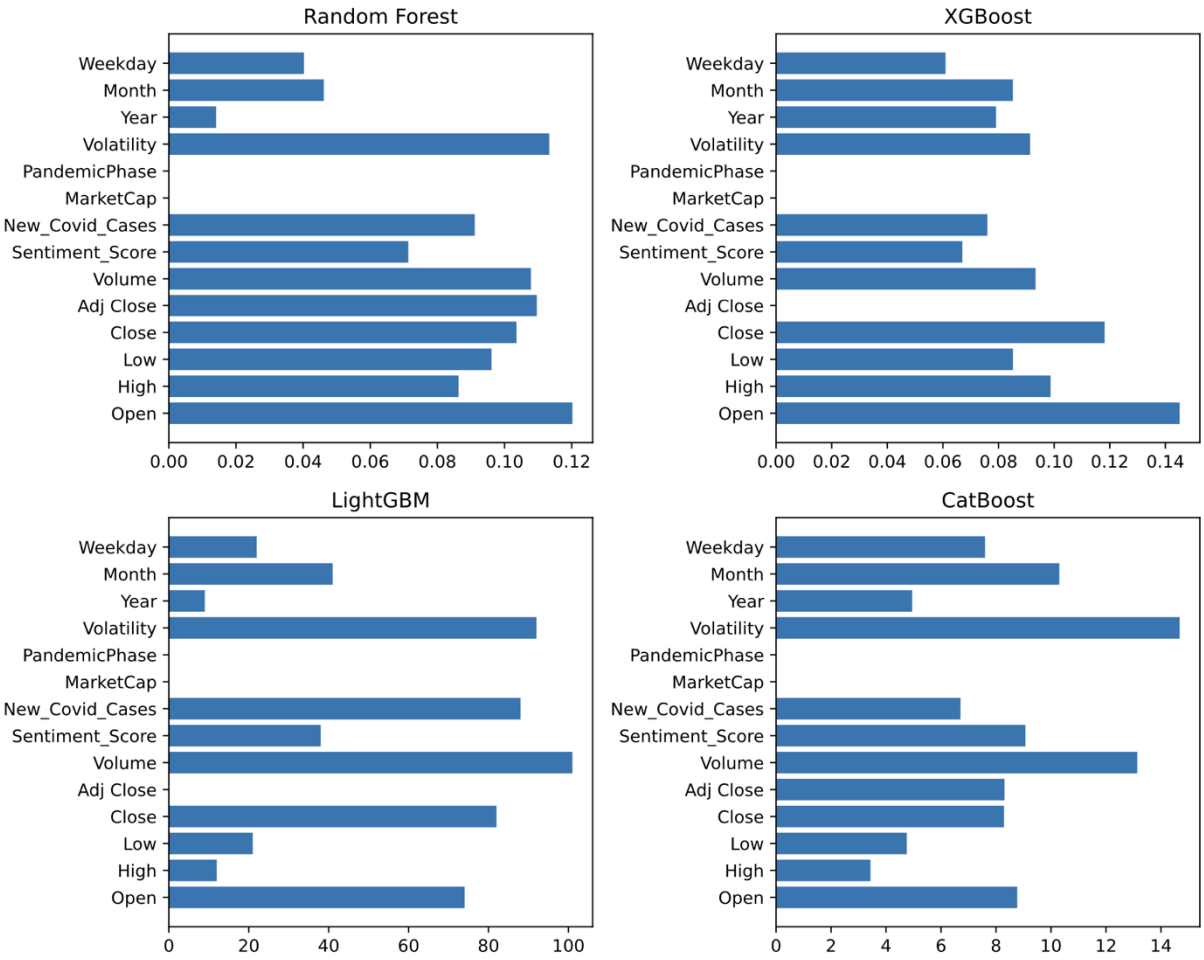


Figure Feature importance across models

(author's own work)

- *Pandemic Data & Predictions:* The importance of granular pandemic data like new COVID-19 cases was accentuated, while binary indicators like pandemic phase were less influential. This finding accentuates the need for detailed, continuous data that reflects dynamic external events, aiding in capturing the intricacies of stock returns during tumultuous periods.

- *Company-Specific Sentiment Variations:* A pronounced correlation between firm size and news sentiment was observed. However, there were anomalies with some firms diverging significantly from general trends, emphasizing the need for a bespoke approach when leveraging sentiment scores in financial strategies.

Recommendations:

Drawing from these insights, the following recommendations are proposed:

- **Model Selection for Predictive Analysis:** Financial stakeholders, including analysts and hedge fund managers, should prioritize the XGBoost model. Its superior performance, especially when analyzing biotech firms, makes it a prime candidate. A move towards sector-specific models, like XGBoost tailored for biotechnology, can optimize prediction accuracy.

- **Feature Importance and Sentiment Analysis:** Financial researchers and investors should adopt a balanced stance. Sentiment analysis, while invaluable, should be integrated with traditional stock indicators for a comprehensive prediction model. The study emphasizes the importance of not overly relying on sentiment analysis. A hybrid approach, which melds sentiment scores with traditional stock metrics, is the way forward.

- **Incorporating Pandemic Data:** For those crafting financial models, there's a pronounced need to emphasize granular pandemic data. The study's findings accentuate the importance of including data like new COVID-19 casesover broader indicators. Such granularity offers a dynamic understanding of external influences, especially pivotal during events like the pandemic.

- **Future Research Directions:** The realm of sentiment analysis in stock predictions is vast, holding myriad avenues for exploration. Researchers are poised to delve deeper, either by building on this study's foundation or by charting fresh territories. The pronounced accuracy in biotech firms, for instance, begs for further exploration. Additionally, the consistent success of the XGBoost model could be a launchpad for research aiming to refine its application or merge it with other innovative methodologies.

Concluding Reflections:

This study offers a comprehensive insight into the multifaceted relationship between news sentiment and stock returns during a globally disruptive period. It stands as a guiding light for a broad spectrum of stakeholders, providing a roadmap to navigate the intricate terrains of modern financial landscapes. As the world of finance continues to evolve, research endeavors like this will be instrumental in shaping future strategies and investment decisions.

Ethical Considerations:

In the realm of computational research, upholding ethical standards is paramount. As this study delves into the intricate interplay of news sentiment and stock returns, it's imperative to ensure that the interpretation and insights derived do not inadvertently propagate misleading or detrimental financial advice. Additionally, leveraging sentiment data from news sources necessitates awareness of potential biases, misrepresentations, or oversimplifications. Throughout this research, a conscientious approach was adopted, anchored firmly in the AREA (Anticipate, Reflect, Engage, Act) framework, ensuring a comprehensive and ethically sound exploration of the topic at hand (EPSRC, n.d.).

References

Cakra, Y. E., & Trisedya, B. D. (2015, October). Stock price prediction using linear regression based on sentiment analysis. In *2015 international conference on advanced computer science and information systems (ICACSIS)* (pp. 147-154). IEEE.

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/

Lee, K. P., & Song, S. (2022). Informational Content of CEO Tweets and Stock Market Predictability. *Available at SSRN 4228651*.

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

Selected style: Dissertation style

Email : kk528@exeter.ac.uk

Student ID : 720034950

Name : Kota Kobayashi

Github: https://github.com/kotaaaaaaa/Stock-market-prediction

Table of Contents

1. Introduction

1.1 Background

In the rapidly evolving landscape of financial markets, the role of digital information, particularly news sentiment, has become increasingly crucial. This dissertation delves into the predictive power of news sentiment analysis in forecasting stock returns, focusing on the tumultuous period marked by the COVID-19 pandemic. By adopting a dissertation-style project, I aim to bridge the existing gaps in literature, providing a comprehensive enquiry into overlapping literatures, adopting an evidence-based approach, and underpinning the methodology with data science principles. The choice of this style is necessitated by the interdisciplinary nature of the topic, allowing for a thorough exploration that culminates in findings of profound significance.

The confluence of the digital age with financial markets has reshaped traditional methods of stock analysis. With the rise of the digital era, sentiment analysis has emerged as a pivotal tool, tapping into the vast reserves of online textual data to discern market directions (Wankhade et al., 2022). This transition is further accentuated by global upheavals, notably the COVID-19 pandemic. The pandemic, characterized by its unpredictability and profound economic implications, accentuated the symbiotic relationship between news sentiment and stock market movements (Biswas et al., 2020).

Understanding the dynamics of news sentiment is of paramount importance to financial experts, hedge fund managers, and individual investors (Dickinson et al., 2015). Predicting stock movements based on sentiment can offer a competitive edge, potentially leading to better investment decisions and strategies. Moreover, with the proliferation of digital news sources and the surge in algorithmic trading, the ability to quickly analyze and act upon news sentiment has become a strategic asset in the financial realm (Dickinson et al., 2015).

1.2 The Situation and the Challenge

While the interplay between news sentiment and stock returns has been a focal point for many empirical studies, the events of the COVID-19 pandemic have presented unprecedented challenges in this domain. Much of the existing literature addresses

the period prior to the pandemic, with methodologies that are now considered outdated or limited in their scope. For instance, while Cakra and Trisedya (2015) used regression methodologies to predict stock price, recent advances have introduced more sophisticated models like CatBoost, which remain underexplored in the context of news sentiment analysis (Yeo, 2021). Disparate models are employed across papers, with a notable lack of comparative studies consolidating their findings.

Furthermore, there's a noted inconsistency in the literature regarding the subjects of these studies. Some studies are narrowly focused on specific firms or sectors, such as big tech firms or broad market indices like the S&P 500 (Biktimirov, 2021). Even though a few studies have targeted the NASDAQ, they are comparatively few in number, which limits the comprehensiveness of the research landscape (Lee, 2022). This fragmented and occasionally outdated research landscape has left financial experts, hedge fund managers, and individual investors in a quandary. They grapple with critical questions about the best models to employ, the features to prioritize in these models, and the applicability of sentiment analysis across a diverse range of firms, including small-scale enterprises.

1.3 Methodology

In response to these challenges, this study undertakes a comparative evaluation of four state-of-the-art machine learning models: Random Forest, XGBoost, LightGBM, and CatBoost. The focus is sharpened on their capability to predict stock returns using sentiment scores during the unprecedented COVID-19 era. The selection of these models is underpinned by their prominence and demonstrable success across a spectrum of prediction tasks, notably in finance, as corroborated by preceding studies like Huang et al. (2020). While erstwhile mainstays such as linear regression and SVM have witnessed diminishing relevance, deep learning models, despite their powerful capabilities, often come across as intricate and resource-intensive. Ensemble tree-based methodologies like the ones chosen for this study emerge as the golden mean, harmonizing interpretability, operational efficiency, and predictive prowess.

1.4 Research Questions

To bridge these identified gaps in the literature and provide clarity to stakeholders in the financial sector, this research seeks to address three main questions:

- Comparative Analysis:
  In light of the COVID-19 pandemic, how do contemporary machine learning models, specifically XGBoost, Random Forest, LightGBM, and CatBoost, compare in their effectiveness at leveraging news sentiment to predict stock returns?

- Feature Importance:
  Given the myriad of variables that could influence stock predictions, which attributes emerge as most crucial for these models? Is news sentiment consistently paramount, or do other factors such as volume or COVID-19 wield a more significant influence?

- Company-Specific Variations:
  Considering the vast landscape of the financial market, how do sentiment scores differ among companies?

Through a meticulous exploration of these questions, this study aims to offer valuable insights and guidance, particularly in the context of the challenges presented by the COVID-19 pandemic.

2. Literature Review

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 et al. (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 et al. (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 (Renault, 2020).

Recent studies have also shed light on the efficacy of more advanced algorithms such as LightGBM and CatBoost. Ke et al. (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.

Post-COVID-19, this study highlights four pieces of literature that apply sentiment analysis to forecast stock returns.. Huynh et al. (2021) examined global equity markets, indicating a nuanced relationship between sentiment and stock returns. Duan et al. (2020) analyzed the Chinese stock market using textual data, highlighting positive correlations between sentiment and stock metrics. Costola et al. (2020) discerned a positive link between sentiment scores and US market returns, while Ren et al. (2019) achieved a high forecasting accuracy for China's SSE 50 Index through sentiment analysis.

However, these studies overlooked the US's NASDAQ market, with its tech-centric composition. The diverse methodologies employed across these works, without comparative analysis of multiple models, leave an ambiguity in model selection for investors. Moreover, the predominant focus on large tech firms raises questions about the applicability of findings to firms of varying sizes. Such gaps underscore unresolved queries for financial stakeholders: the optimal model choice, the key features for analysis, and the efficacy of sentiment analysis for smaller firms.

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.

Among global indices, NASDAQ stands out as the prime choice for sentiment analysis for several key reasons. First, its heavy concentration of tech and innovative firms generates substantial public discourse on platforms like social media and news outlets, offering a rich dataset for analysis. Second, companies listed on NASDAQ display heightened sensitivity to news and events, amplifying the impact of sentiment shifts. For instance, research on the Michigan Index of Consumer Sentiment revealed that NASDAQ stocks react more swiftly to news insights, underscoring their susceptibility. This unique blend of tech-centric listings and heightened news reactivity makes NASDAQ especially conducive for sentiment-driven market predictions (Wu et al., 2019).

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 et al. (2018) explored stock price forecasting using data from Yahoo Finance and highlighted the importance of analyzing both seasonal and nonseasonal trends. Furthermore, Bordino et al. (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. According to Shi and Ho (2021) and Ho et al. (2020), RavenPack Analytics is suitable for sentiment scores because it offers a comprehensive database that captures firm-specific news releases and their sentiment scores at high frequencies, enabling researchers to examine correlations between news sentiment and stock return volatility.

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 1 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	Yahoo Finance

(author’s own work)

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 2 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. This is author’s own work.)

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*, 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* 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. 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.

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.

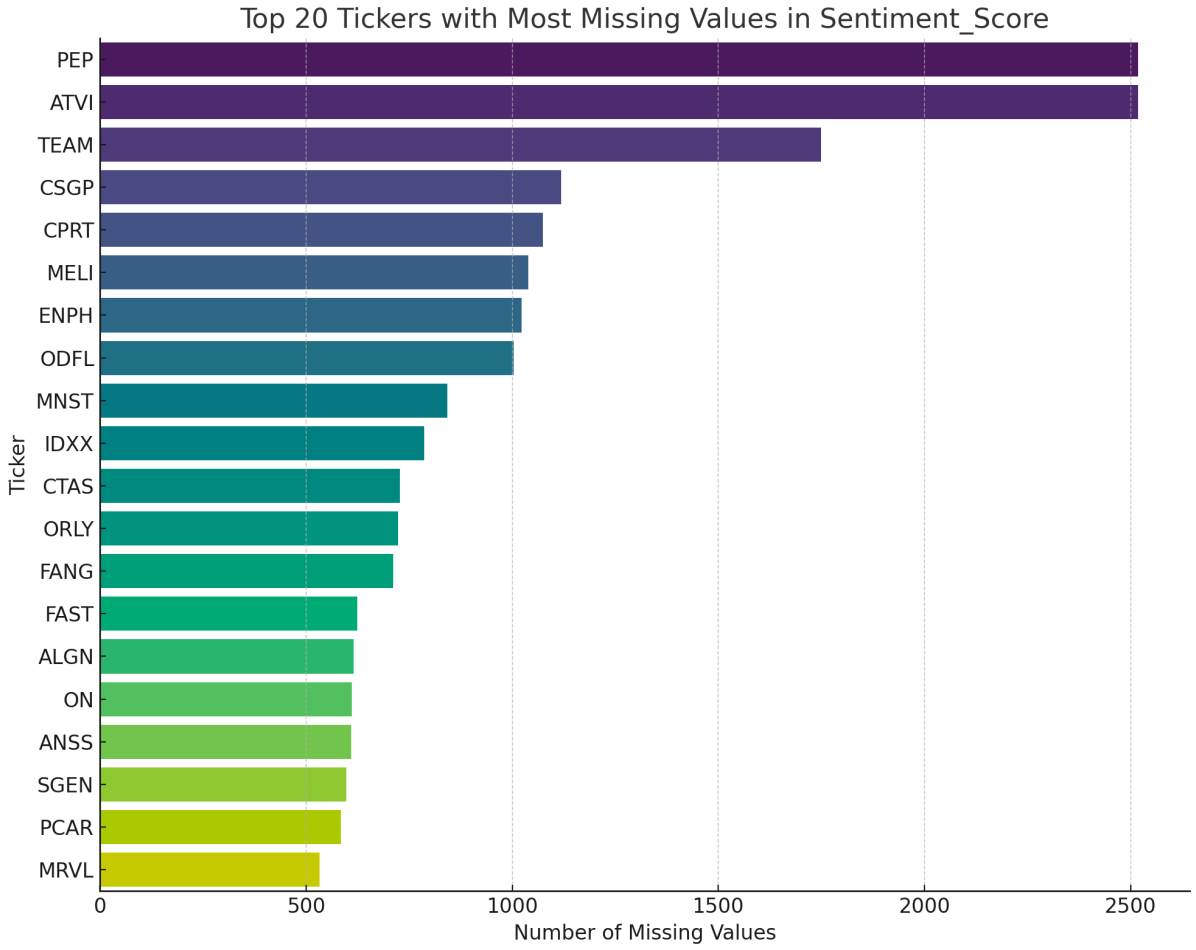


Figure 1 Top 20 tickers with most missing in news sentiment score.

(author's own work)

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.

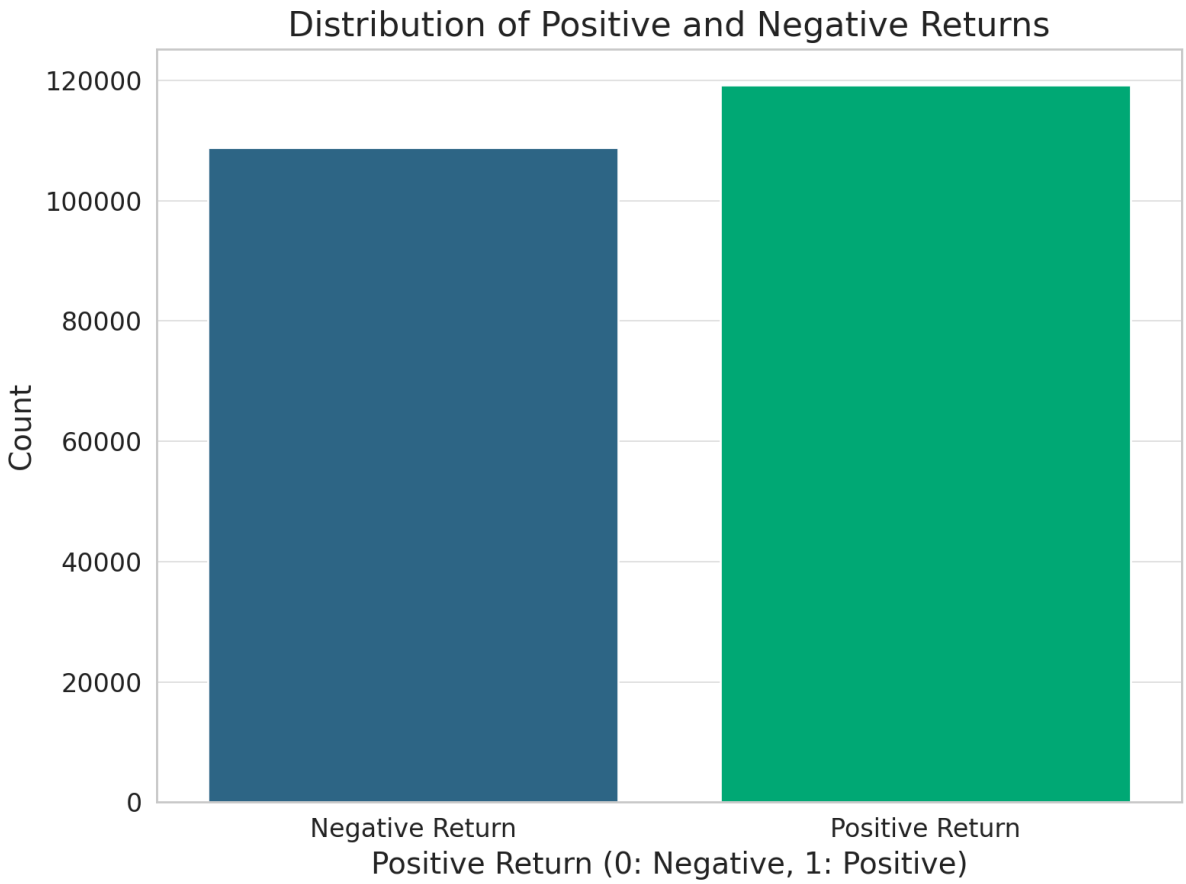


Figure 2 Distribution of positive and negative returns

(author's own work)

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.

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 3 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	2148088035137.93	4784583104877.78	16346392576.00	39679574016.00	62220886016.00	131073245184.00	2828020482048.00

(author's own work)

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

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.

6.2.1 Temporal Evolution of Average Sentiment

The temporal fluctuations of sentiment, averaged across stocks, reflect overarching market moods from 2013 to 2023. Figure 3 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).

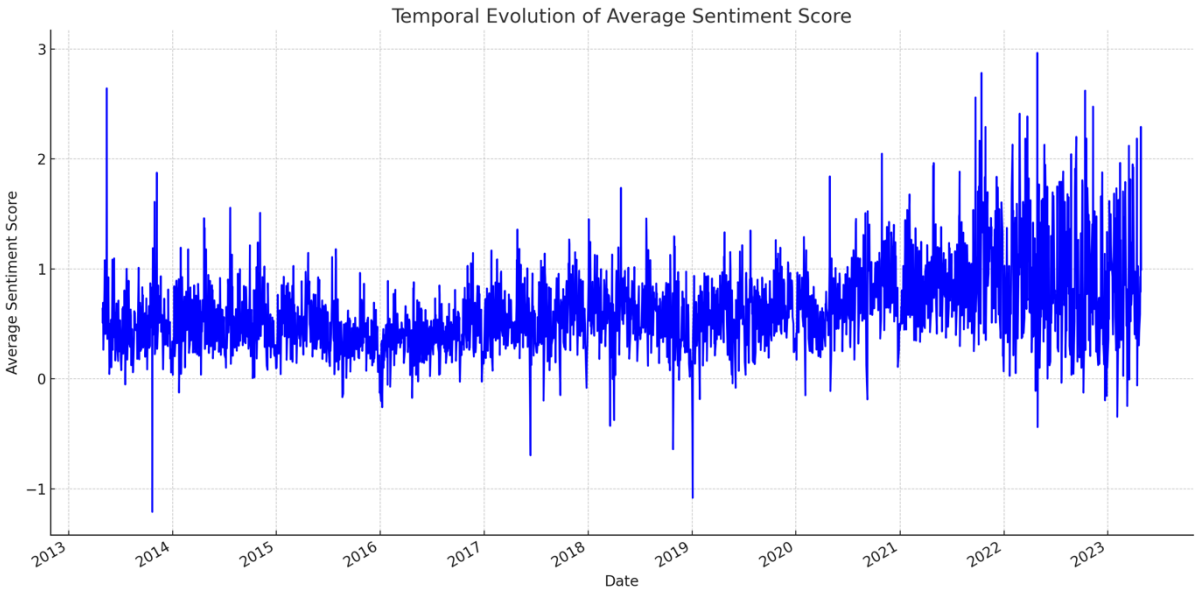


Figure 3 Temporal evolution of average sentiment score

(author's own work)

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.

Figure 4 Average sentiment score by company

(author's own work)

Key takeaways from the Figure 4 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 Old Dominion Freight Line Inc (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.
- 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 between firms are observed. The variations could be ascribed to firm size (Gillam et al., 2015; Li and Yang, 2017; Peng et al., 2022) However, that is still somewhat vague. A deeper dive into these variations will be conducted in *7. The relationship between Firm Size and 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:

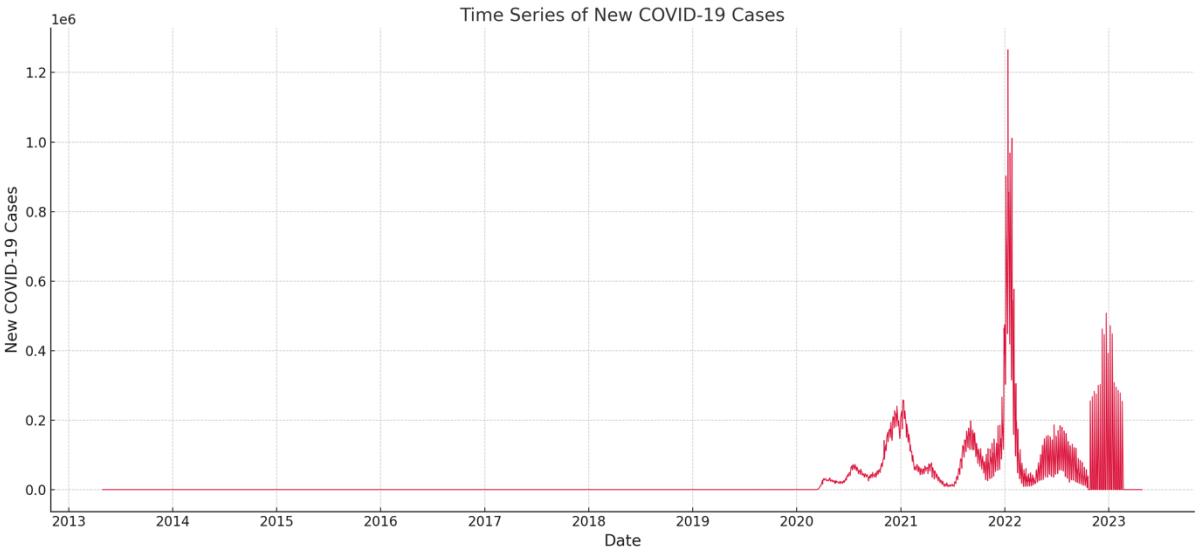


Figure 5 Time series of new COVID-19 cases

(author's own work)

- 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.

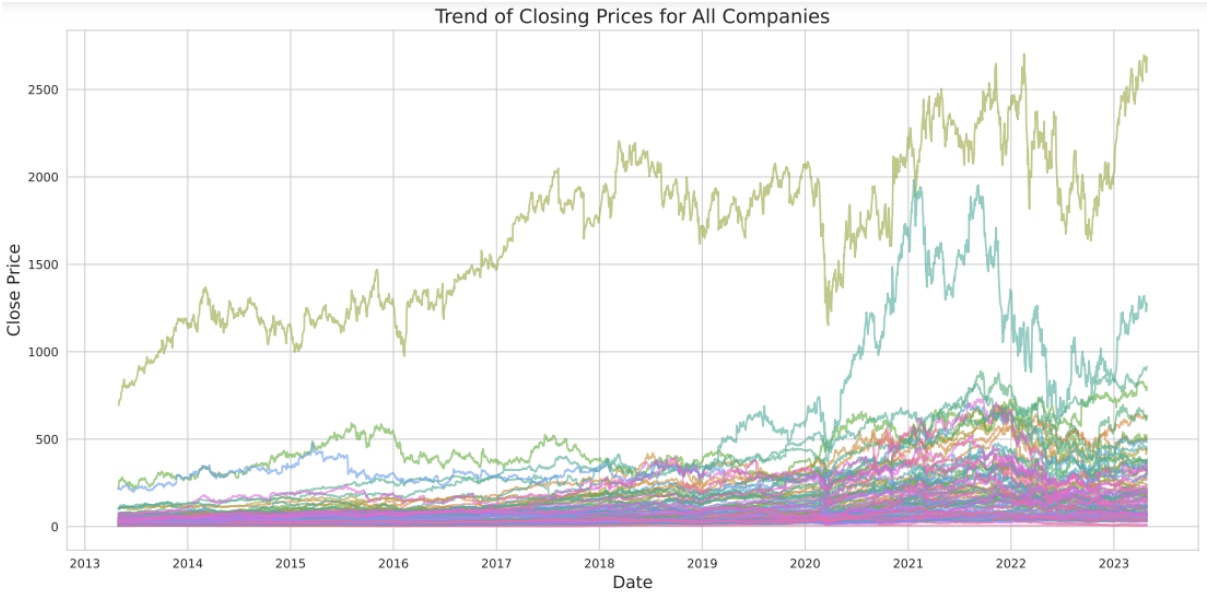


Figure 6 Trend of closing prices for all companies

(author's own work)

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.

6.3 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

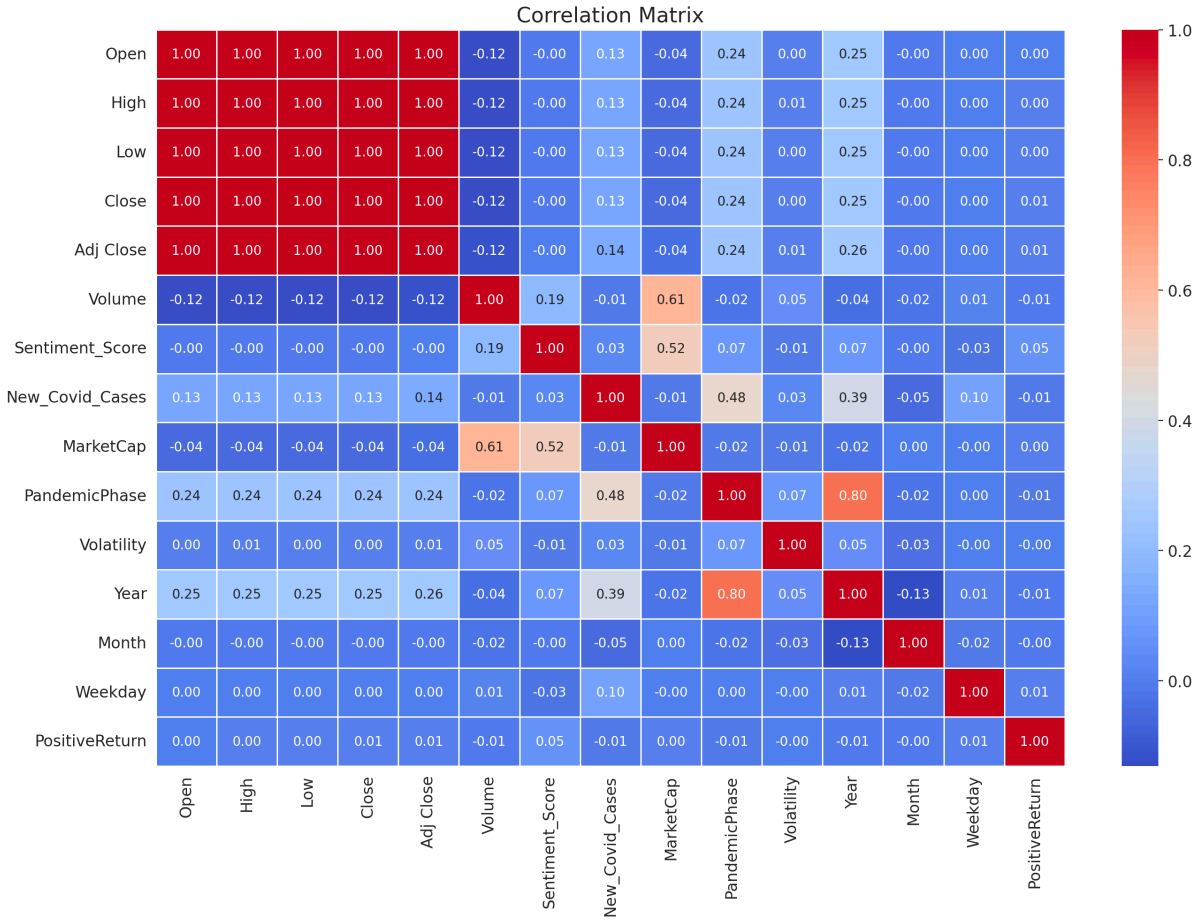


Figure 7 Correlation matrix of features

(author's own work)

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 high correlation with *MarketCap*. This might be reflective of larger firms receiving more media attention, which subsequently influences the sentiment metrics. Analysing these two features might help to answer one of the research questions, Company-Specific Variations. Therefore, a detailed analysis between them will be conducted in the next section.

7. The relationship between Firm Size and News Sentiment Score

As Figure 7 shows that there is a moderate positive correlation between firm size and sentiment score, the analysis will do a deep dive into the link between them. Numerous literatures have emphasized the intricate link between firm size and news sentiment analysis. For instance, Gillam et al. (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 et al. (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.

Therefore, it is important to make the relationship between them clear before making models and predict stock market returns. This section will analyse 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. This analysis will also be helpful for answering one of the research questions, Company-Specific Variations.

7.1 Correlation and Regression

To avoid outliers, the average sentiment scores by company are utilized instead of original sentiment score in this section, similar to Section 6.2.2. The Figure 8 shows 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 positive correlation between market capitalization and news sentiment score.

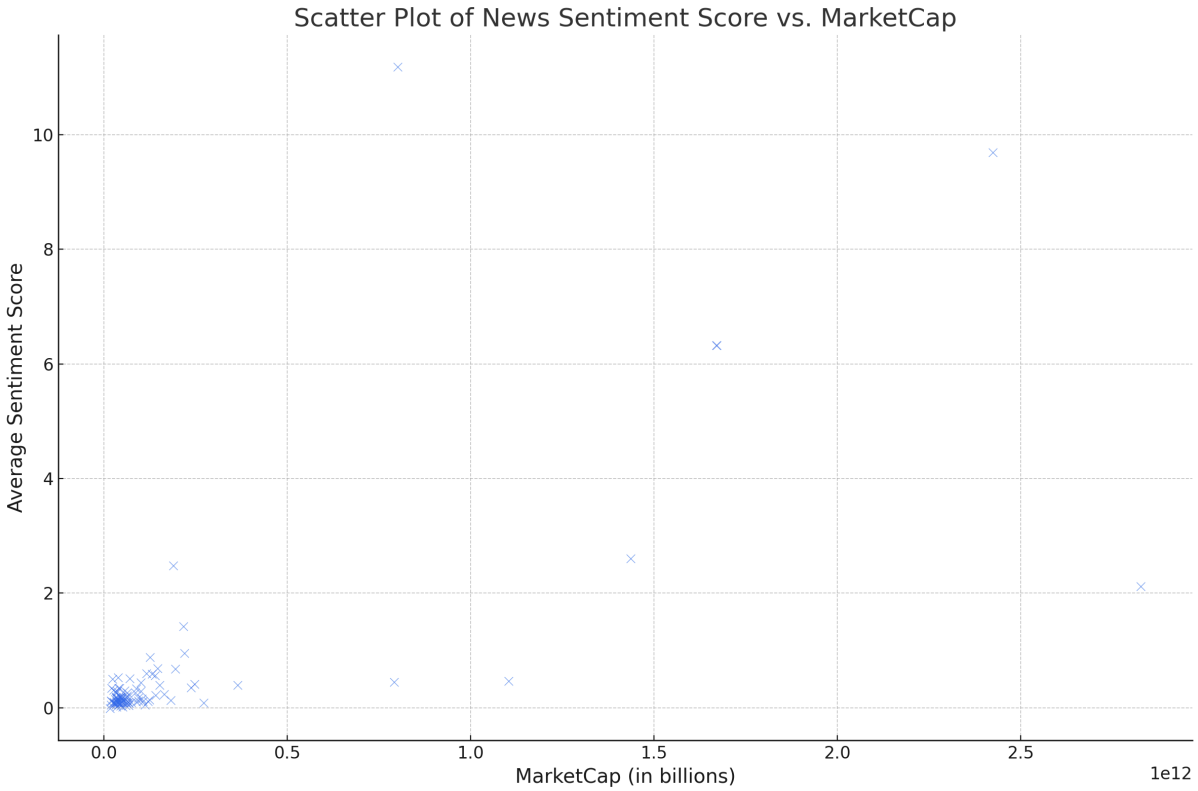


Figure 8 News sentiment score vs. MarketCap

(author's own work)

To delve deeper into the relationship between firm size (measured by market capitalization) and news sentiment score, a simple linear regression model was employed in Figure 9.

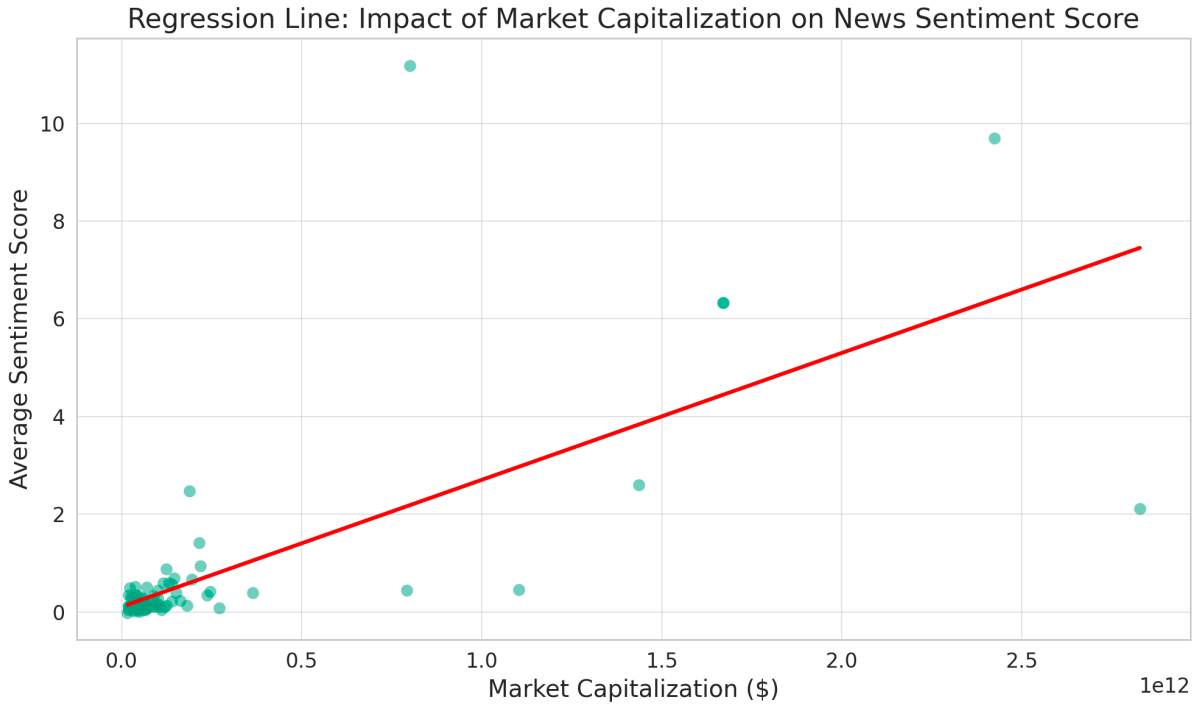


Figure 9 Regression line: Impact of market capitalization on news sentiment score.

(author's own work)

The model can be represented as:

$$SentimentScore = \beta_0 + \beta_1 \times MarketCap + \varepsilon$$

Where:

- β_0 is the intercept.
- β_1 represents the change in sentiment score for a one-unit change in market capitalization.
- ε 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 that larger companies, in terms of market capitalization, tend to receive a more favorable news sentiment on average.
- 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.

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. The KMeans clustering approach is a widely-used unsupervised machine learning technique that partitions a dataset into k distinct, non-overlapping subsets based on the inherent similarities among data points. It achieves this by iteratively reassigning data points to clusters and recalculating cluster centroids until the within-cluster sum of squares is minimized (Cui, 2020).

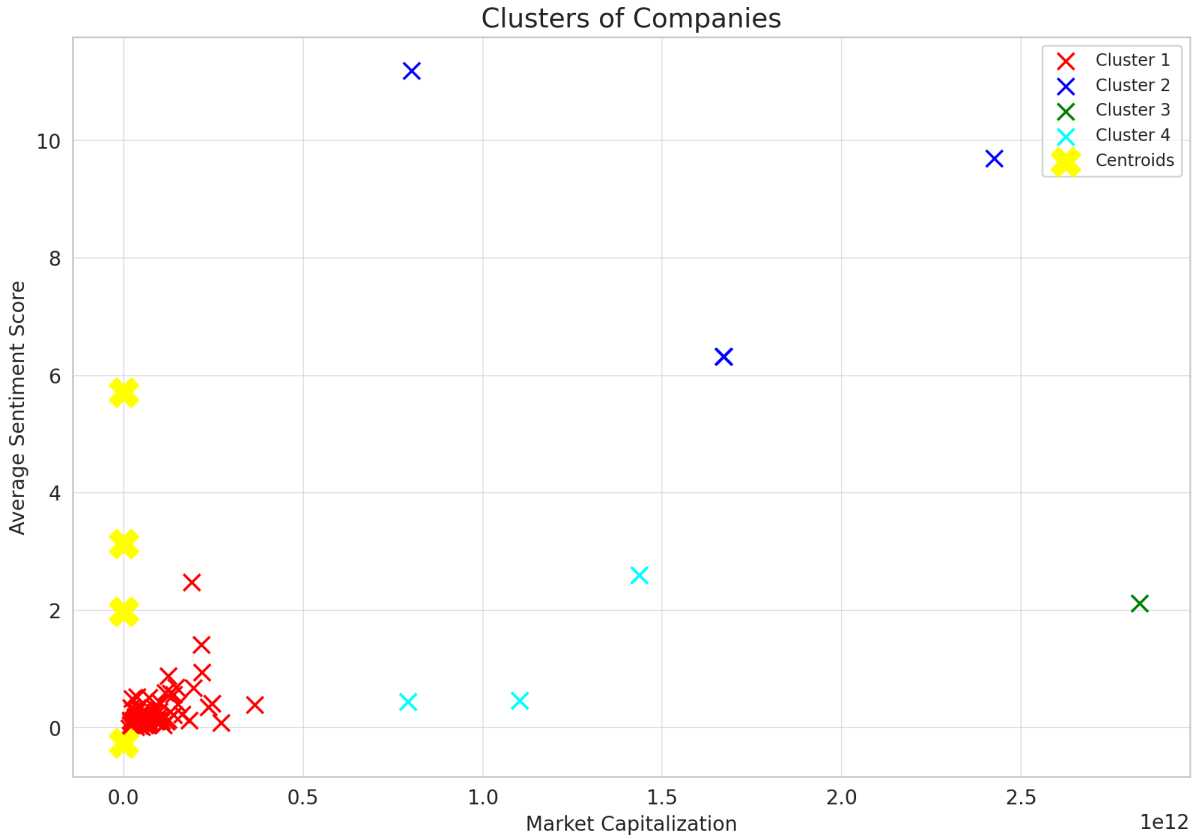


Figure 10 Clusters of companies

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.

(author's own work)

- 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.

Addressing the research question of how sentiment scores vary among companies and the challenges certain firms or sectors may pose in sentiment analysis, our study found a pronounced correlation between firm size and news sentiment. Larger firms, as denoted by their market capitalization, typically attract more favorable media sentiment. However, what the cluster analysis revealed is that while there were overall correlations, a small number of companies had exceptionally high market capitalizations or sentiment scores that were outliers. Though broad relationships existed, particular corporations exhibited anomalous financial valuations or attitudinal metrics distinct from the general trends. Therefore, for financial experts, hedge fund managers, and individual investors leveraging sentiment scores, it's imperative to discern if the firm under consideration belongs to outlier-like clusters (2, 3, 4), as these categories present unique sentiment dynamics.

8. Model Evaluation

While other models such as Support Vector Machines (SVM) or Neural Networks could have been considered, they often require more intricate tuning and can be computationally intensive (Cortes and Vapnik, 1995; Schmidhuber, 2015). For

instance, SVM, though powerful, might not scale efficiently with the vast volume of our dataset. Neural Networks, while versatile, can sometimes be seen as a black box, making the interpretation of results challenging in a domain where understanding the underlying dynamics is crucial. Therefore, 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.

8.2 Random Forest

8.2.1 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 $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, the algorithm for Random Forest is:

1. For $b = 1$ to B (where B is the number of trees):
    - Draw a bootstrap sample Z of size n from the training data.
    - Grow a decision tree T_b 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 m variables randomly from the p 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 B trees.

8.2.2 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 ($n_{estimators}$) = 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 $n_{estimators}$. 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.

8.2.3 Results

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

$$AUC = \int_0^1 TPR(t) dFPR(t)$$

where:

- $TPR(t)$ is the true positive rate at threshold t.
- $dFPR(t)$ is the false positive rate at threshold t.

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

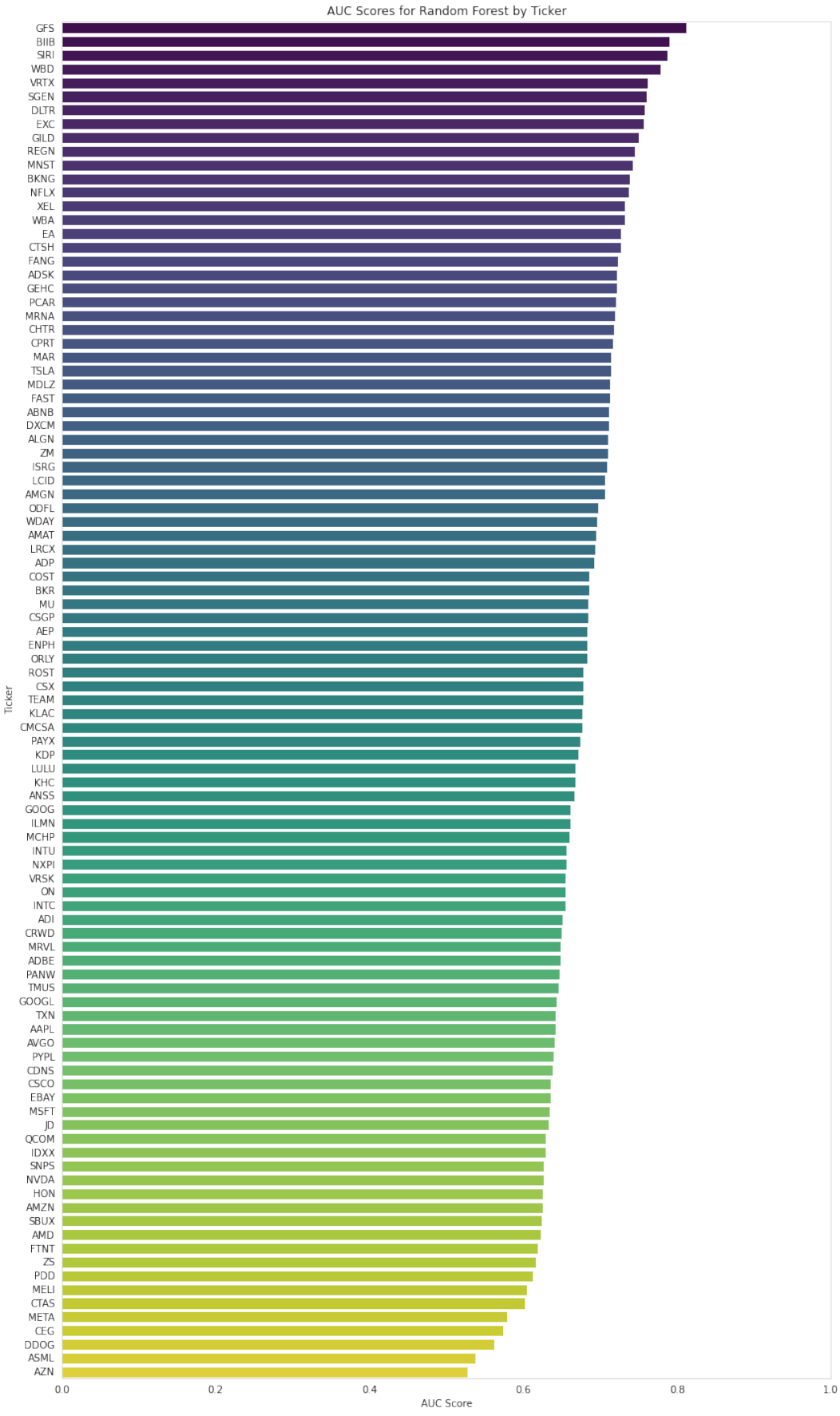


Figure 11 AUC scores for Random Forest by ticker

(author's own work)

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.

8.2.4 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.

8.3 XGBoost

8.3.1 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 $L(y, F(x))$ where y is the true label and $F(x)$ is the prediction, the objective in each iteration is to find a function $h(x)$ that minimizes:c

$$Obj = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{i=1}^{t} \Omega(f_i)$$

where:

- $\Omega(f_i)$ is the regularization term.
- t 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.

8.3.2 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.

8.3.3 Results

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

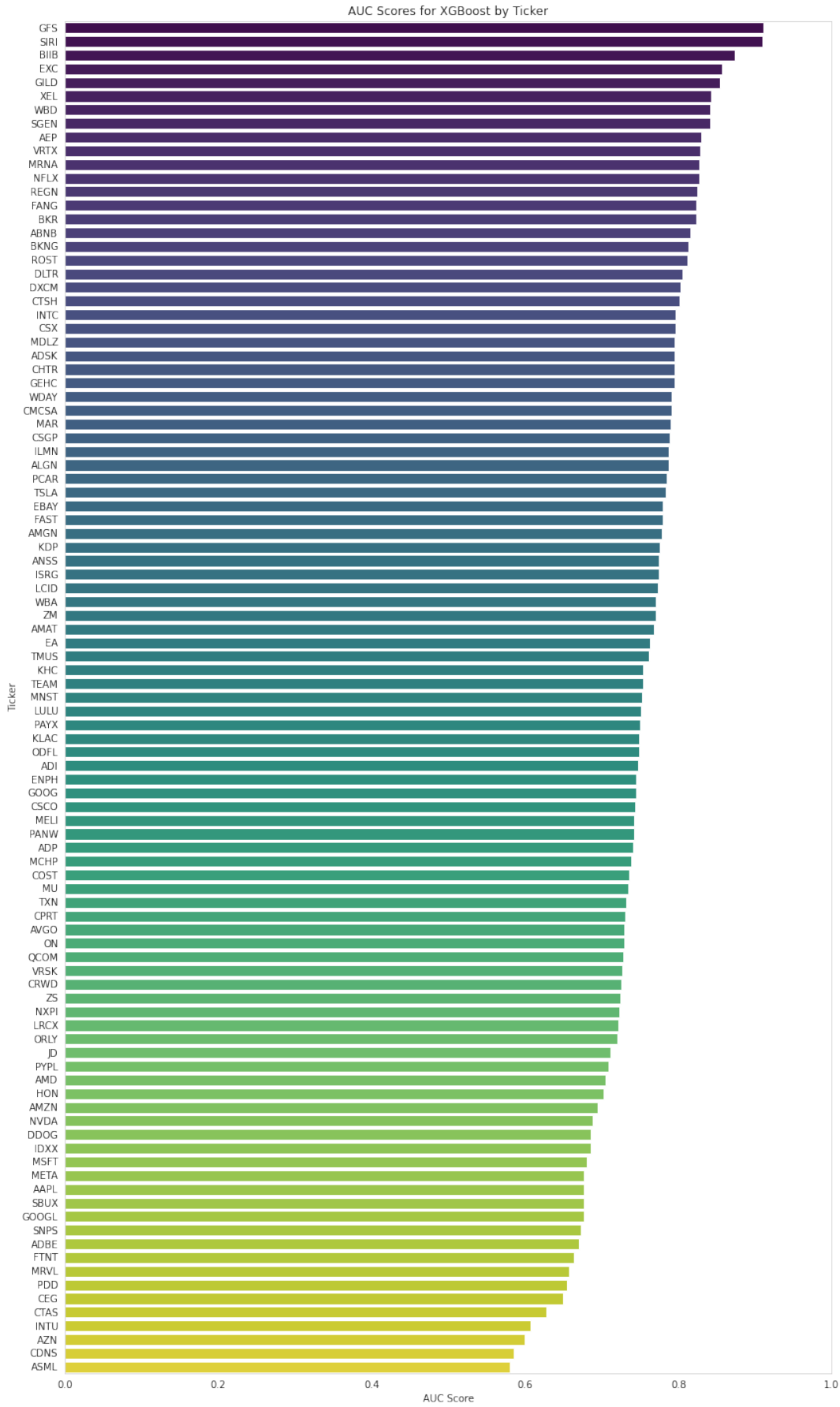


Figure 12 AUC scores for XGBoost by ticker

(author's own work)

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.

8.3.4 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.

8.4 LightGBM

8.4.1 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:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{i=1}^{t} \Omega(f_i)$$

where:

- $\Omega(f_i)$ represents the regularization term.
- t denotes the number of iterations.

8.4.2 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.

8.4.3 Results

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

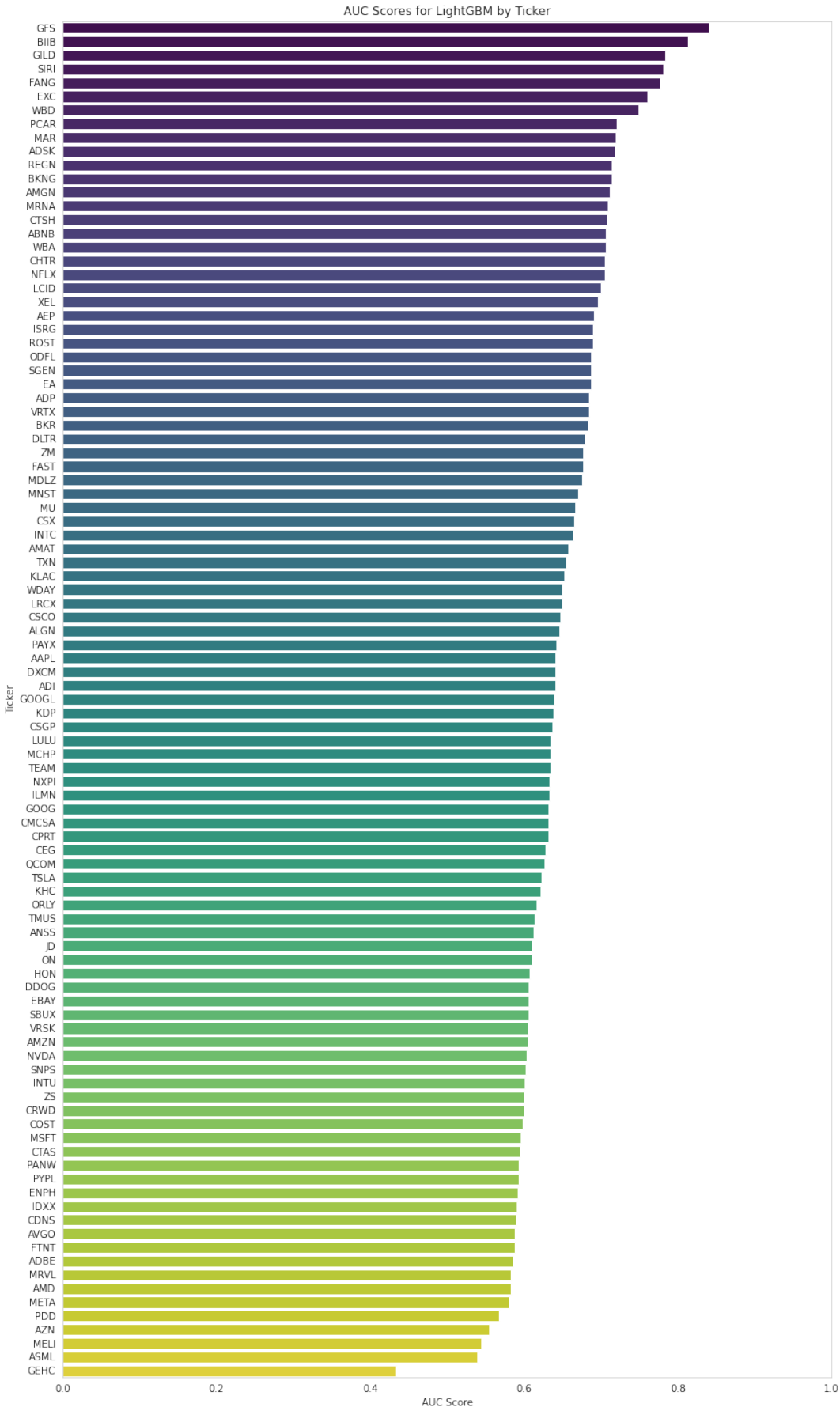


Figure 13 AUC scores for LightGBM by ticker

(author's own work)

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.

8.4.4 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.

8.5 CatBoost

8.5.1 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:

$$2 = \sum_{i=1}^{n} L(y_i, F(x_i)) + \lambda \sum_{i=1}^{t} \Omega(f_i)$$

where:

- $\Omega(f_i)$ is the regularization term.
- t is the number of iterations.
- λ 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.

8.5.2 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.

8.5.3 Results

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

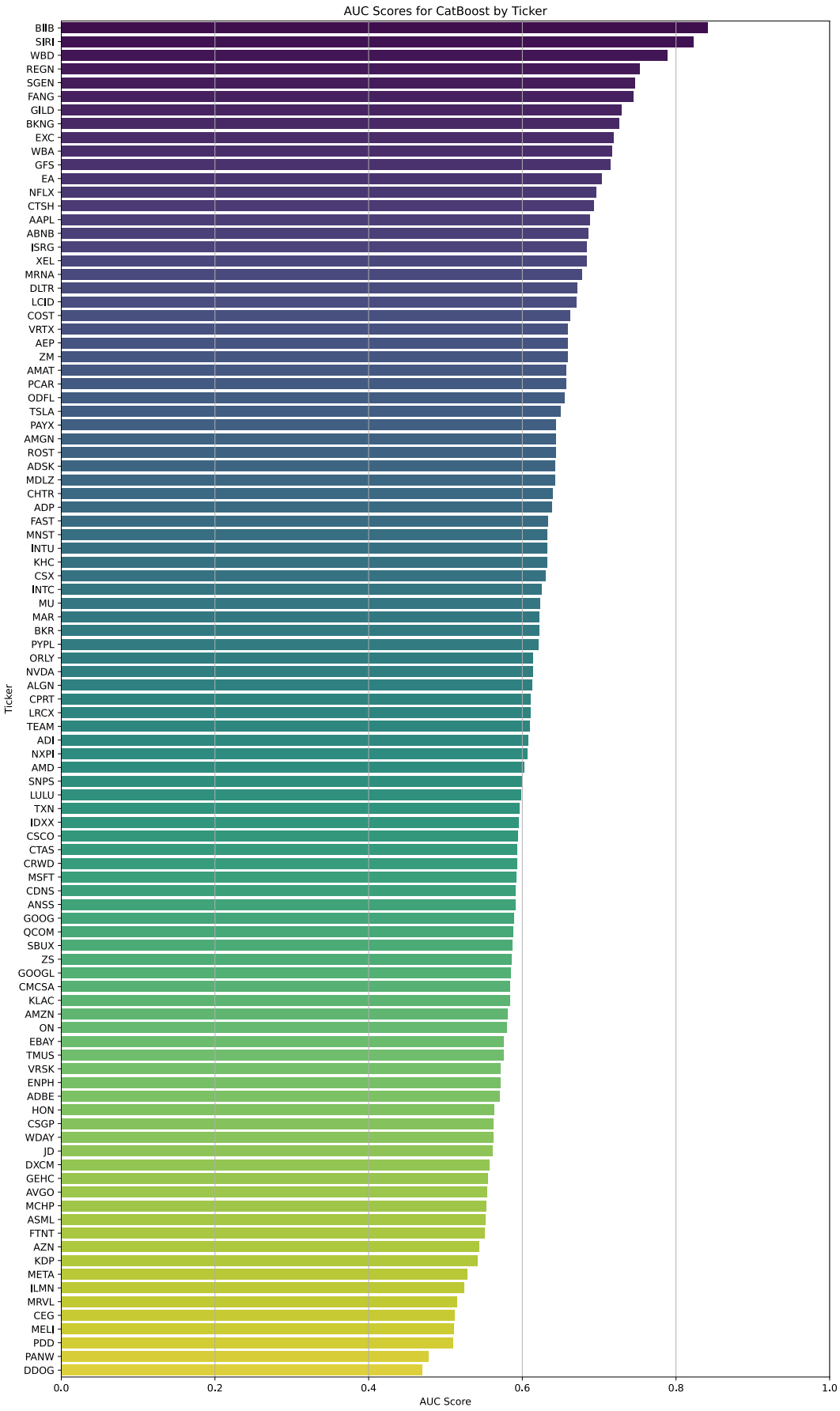


Figure 14 AUC scores for CatBoost by ticker

(author's own work)

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'.

8.5.4 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.

8.6 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 4 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

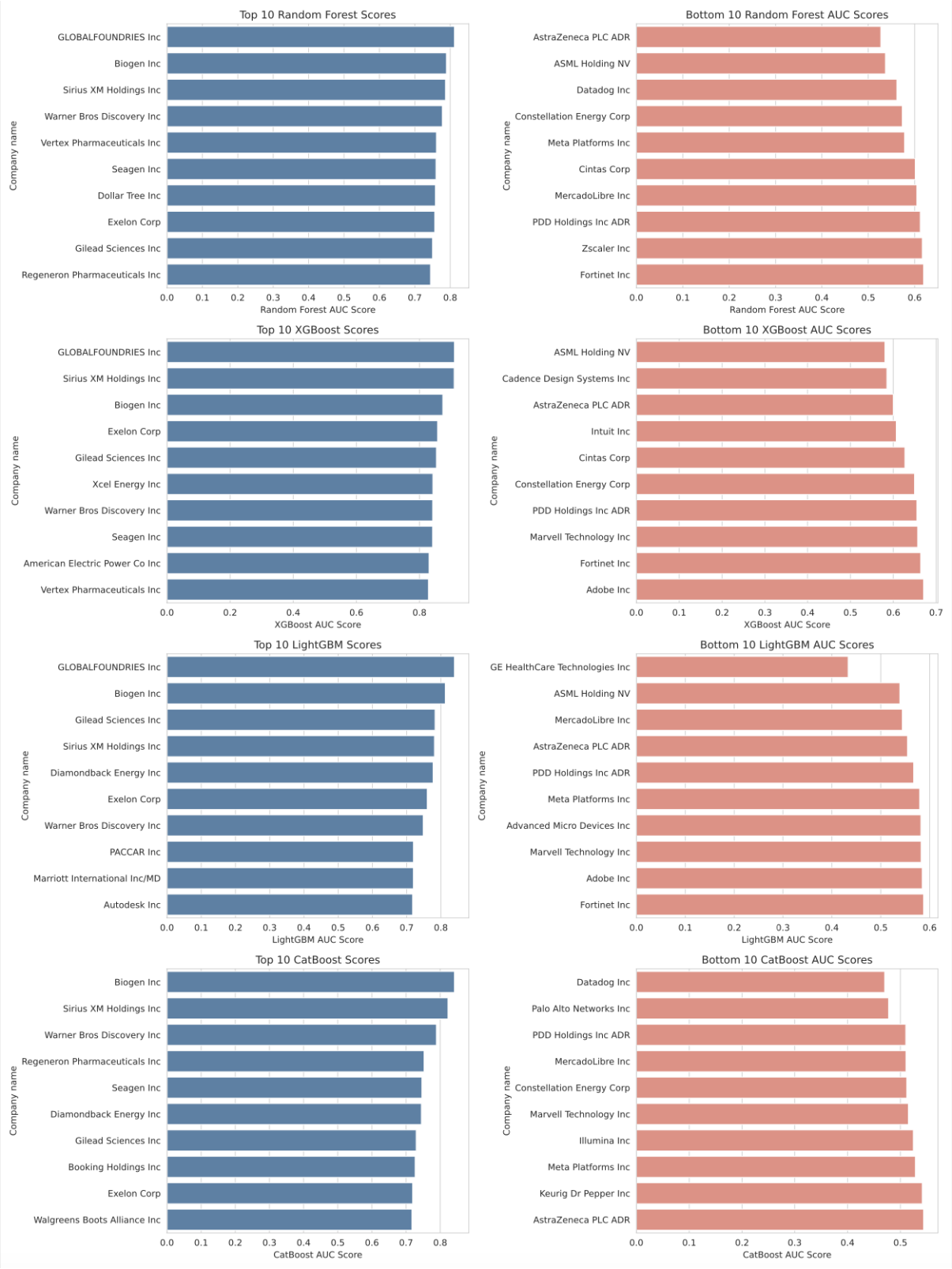


Figure 15 Top 10 firms and bottom 10 Firms in AUC score across models

(author's own work)

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

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 16 shows the results to delve into the specific importance scores of features as derived from each of these models.

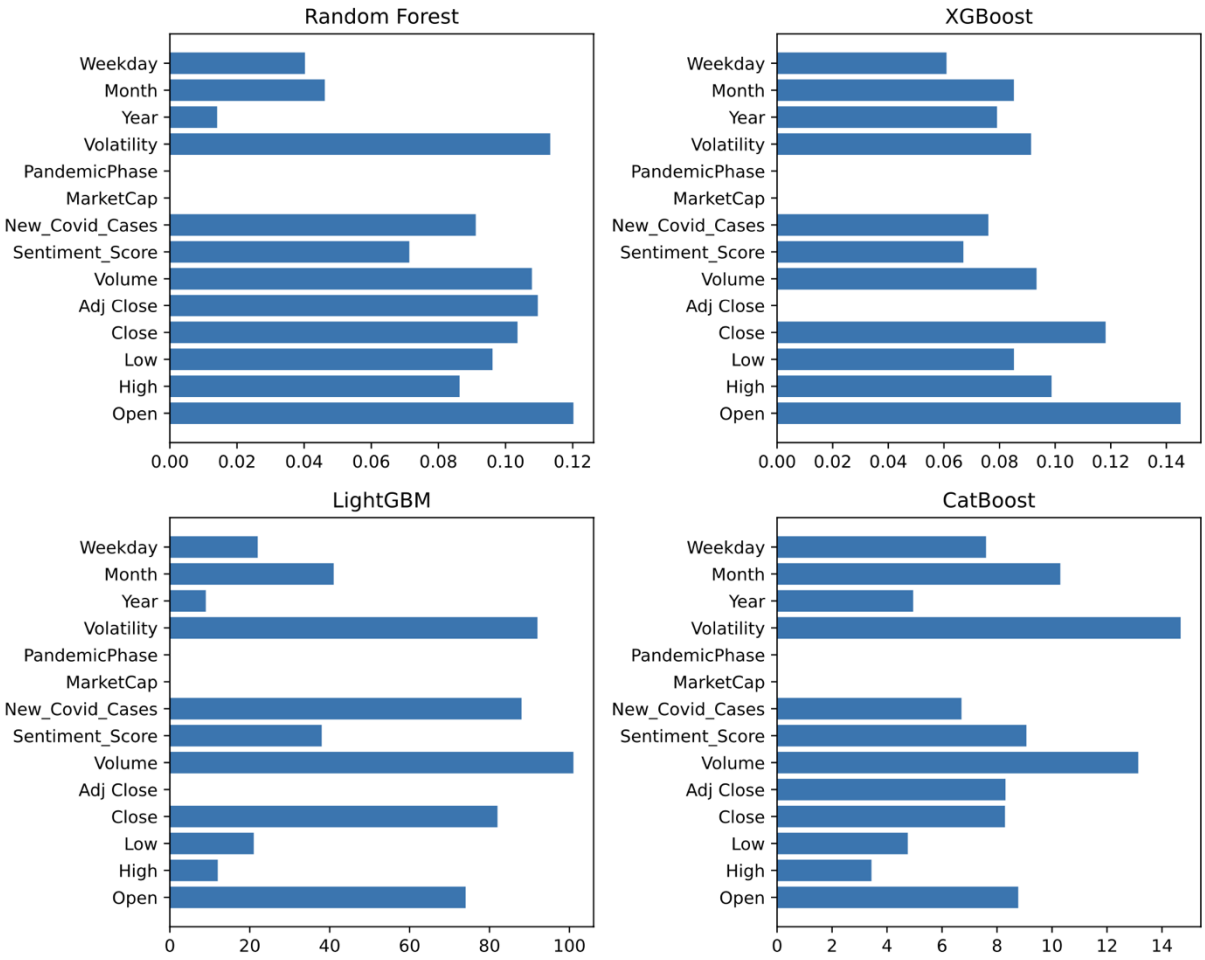


Figure 16 Feature importance across models

(author's own work)

8.9.2 Feature Importance Across Models

This section is about the specific importance scores of features across models based on the Figure 16.

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 (2021) 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.

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 contemporary financial landscape has witnessed an upsurge in the importance of news sentiment, notably against the backdrop of the COVID-19 pandemic. The

primary objective of this research was to delve into the predictive power of news sentiment analysis in forecasting stock returns during this tumultuous period. Through an evidence-based approach, this study aimed to bridge existing gaps in the literature, providing clarity and offering valuable insights to stakeholders in the financial sector.

To achieve the objectives, three pivotal research questions were posed:

- How do contemporary machine learning models, specifically XGBoost, Random Forest, LightGBM, and CatBoost, compare in their effectiveness at leveraging news sentiment to predict stock returns during the COVID-19 era?
- Which features emerge as the most crucial for these models in stock predictions?
- How do sentiment scores differ among companies?

Addressing the first research question, our analysis showcased XGBoost's consistent superiority in predicting stock returns using news sentiment. Notably, biotech firms such as Biogen Inc and Gilead Sciences Inc achieved higher accuracy across models. These sectors, characterized by extended R&D phases and rigorous regulatory environments, often enjoy market exclusivity once a product gains approval. Their continuous innovation and the durability of their products typically lead to predictable financial patterns, further enhancing their suitability for predictive modeling. Given these insights, the model's varied results across tickers underscore the importance of customizing models for specific companies or sectors. This emphasizes the need for stakeholders to carefully select models, considering the unique characteristics of the data.

Regarding the second research question on feature importance, our study emphasized the dominance of traditional stock indicators, particularly Volume, in shaping predictions. The *Sentiment_Score*, which captures daily sentiment towards stocks, was found to have a less pronounced influence compared to attributes like Open or Volume. Additionally, the *PandemicPhase*'s binary delineation, representing pre and post-pandemic periods, lacked significant influence in the models. This could be due to the more granular data provided by *New_Covid_Cases*, which offers a dynamic representation of the pandemic's influence on stock returns.

Lastly, in exploring the third research question on company-specific sentiment variations, we discerned a significant correlation between firm size and news sentiment. While larger firms generally attracted more favorable media sentiment, the cluster analysis identified specific firms that deviated from the general trend. This nuanced understanding underscores the importance of considering individual firm characteristics when leveraging sentiment scores. Financial experts, hedge fund managers, and individual investors must be especially vigilant when analyzing firms belonging to outlier-like clusters, as these present unique sentiment dynamics that may not align with broader market trends.

12. 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.

The realm of sentiment analysis in stock predictions is vast, holding myriad avenues for exploration. Researchers are poised to delve deeper, either by building on this study's foundation or by charting fresh territories. The pronounced accuracy in biotech firms, for instance, begs for further exploration. Additionally, the consistent success of the XGBoost model could be a launchpad for research aiming to refine its application or merge it with other innovative methodologies.

13. Recommendations

This research on leveraging news sentiment to predict stock returns, especially amidst the challenges of the COVID-19 pandemic, has provided significant insights. Based on these findings, we present a set of structured recommendations for various stakeholders:

1. Model Selection for Predictive Analysis:
   Audience: Financial stakeholders including analysts, hedge fund managers, and individual investors.
   Recommendation: Given that the XGBoost model consistently outperformed others in leveraging news sentiment for stock return predictions, it's logical for financial stakeholders to prioritize this model. Its efficacy becomes especially pronounced when analyzing biotech firms like Biogen Inc and Gilead Sciences Inc. For a more grounded real-world application, there's potential in shifting towards sector-specific models, like XGBoost tailored for biotechnology, to achieve heightened prediction accuracy.

2. Feature Importance and Sentiment Analysis:
   Audience: Financial researchers and investors focusing on feature selection for predictive models.
   Recommendation: While sentiment scores capture valuable insights, their influence is often overshadowed by traditional stock indicators like Volume. Therefore, it's advisable not to rely excessively on sentiment analysis alone; a hybrid approach, integrating both sentiment scores and traditional stock indicators, is desirable for a comprehensive and accurate stock prediction.

3. Incorporating Pandemic Data:
   Audience: Data scientists and financial modelers.
   Recommendation: Models should prioritize the inclusion of granular pandemic data, like New_Covid_Cases, over binary indicators like PandemicPhase. The more detailed and continuous nature of such data offers a dynamic representation of external events like the pandemic's progression, which can be pivotal in capturing the subtleties influencing stock returns during such unprecedented periods.

4. Future Research Directions:

   Audience: Academic researchers and financial scholars.

   Recommendation: The nuances of sentiment analysis in stock return prediction offer vast avenues for future exploration. Researchers are advised to approach this topic by building upon the existing foundation laid by this study. For instance, the sector-specific variations in sentiment analysis, particularly the pronounced accuracy observed in biotech firms, can be further explored to understand the underlying dynamics. Moreover, given the consistent performance of the XGBoost model, future research could delve into refining its application, potentially combining it with other methodologies or integrating additional features to optimize its predictive capabilities. The aim would be to either replicate the current findings in different contexts or expand upon them, ensuring that the methodologies remain relevant and adaptive to evolving market dynamics.

To conclude, these recommendations offer a comprehensive guide to harnessing the potential of news sentiment in stock return predictions, with a focus on the intricacies of model selection, feature integration, and the dynamic influence of external events.

14. 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).

Biktimirov, E. N., Sokolyk, T., & Ayanso, A. (2021). Sentiment and hype of business media topics and stock market returns during the COVID-19 pandemic. *Journal of Behavioral and Experimental Finance*, *31*, 100542.

Biswas, S., Sarkar, I., Das, P., Bose, R., & Roy, S. (2020). Examining the effects of pandemics on stock market trends through sentiment analysis. *J Xidian Univ*, *14*(6), 1163-76.

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.

Cakra, Y. E., & Trisedya, B. D. (2015, October). Stock price prediction using linear regression based on sentiment analysis. In *2015 international conference on advanced computer science and information systems (ICACSIS)* (pp. 147-154). IEEE.

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.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2020). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Research in International Business and Finance*, 64, 101881 - 101881.

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), 5-8.

Dickinson, B., & Hu, W. (2015). Sentiment analysis of investor opinions on twitter. *Social Networking*, *4*(03), 62.

Duan, Y., Liu, L., & Wang, Z. (2020). COVID-19 Sentiment and Chinese Stock Market: Official Media News and Sina Weibo. *Social Science Research Network.*

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*.

Ho, K. Y., Shi, Y., & Zhang, Z. (2020). News and return volatility of Chinese bank stocks. *International Review of Economics & Finance*, *69*, 1095-1105.

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

Huynh, T., Foglia, M., Nasir, M., & Angelini, E. (2021). Feverish sentiment and global equity markets during the COVID-19 pandemic. *Journal of Economic Behavior & Organization*, 188, 1088 - 1108.

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.

Lee, K. P., & Song, S. (2022). Informational Content of CEO Tweets and Stock Market Predictability. *Available at SSRN 4228651*.

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.

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*.

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.

Ren, R., Wu, D., & Liu, T. (2019). Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Systems Journal*, 13, 760-770.

Renault, T. (2020). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, *2*(1-2), 1-13.

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.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

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

Shi, Y., & Ho, K. Y. (2021). News sentiment and states of stock return volatility: Evidence from long memory and discrete choice models. *Finance Research Letters*, *38*, 101446.

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.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780.

Yeo, J. S. W. (2021). Predicting stock market index with gradient boosting machine ensemble, bayesian optimization, temporal consistency analysis, market sentiment analysis, game theory and novel holdout method.

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.

15. Appendix

Table 5 Companies with Their Tickers

Company name	Ticker
Microsoft Corp	MSFT
Apple Inc	AAPL
NVIDIA Corp	NVDA
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	**Cognizant Technology Solutions Corp**	CTSH	**Align Technology Inc**	ALGN
Biogen Inc	BIIB	**Fastenal Co**	FAST	**Atlassian Corp**	TEAM
AstraZeneca PLC ADR	AZN	**Verisk Analytics Inc**	VRSK	**Walgreens Boots Alliance Inc**	WBA
Ross Stores Inc	ROST	**Crowdstrike Holdings Inc**	CRWD	**Diamondback Energy Inc**	FANG
GE HealthCare Technologies Inc	GEHC	**Dollar Tree Inc**	DLTR	**Enphase Energy Inc**	ENPH
Electronic Arts Inc	EA	**Warner Bros Discovery Inc**	WBD	**eBay Inc**	EBAY
Seagen Inc	SGEN	**Datadog Inc**	DDOG	**Zscaler Inc**	ZS
CoStar Group Inc	CSGP	**Constellation Energy Corp**	CEG	**Sirius XM Holdings Inc**	SIRI
GLOBALFOUNDRIES Inc	GFS	**Illumina Inc**	ILMN	**Zoom Video Communications Inc**	ZM
Xcel Energy Inc	XEL	**ANSYS Inc**	ANSS	**JD.com Inc ADR**	JD
Baker Hughes Co	BKR			**Lucid Group Inc**	LCID

(author's own work)

Research Proposal

Predicting Stock Returns Using News Sentiment Analysis during COVID-19

Aims and Objectives:

The primary objective of this research is to evaluate the potential of news sentiment analysis as a predictor of stock returns, especially against the tumultuous backdrop of the COVID-19 pandemic. By leveraging advanced machine learning models, the study aims to:

- Compare the efficacy of contemporary machine learning models, notably XGBoost, Random Forest, LightGBM, and CatBoost, in utilizing news sentiment for predicting stock returns during the COVID-19 era.
- Identify the most influential features within these models for stock predictions.
- Investigate variations in sentiment scores across different companies.

Problem Statement and Background:

Problem Statement and Background:

The intricate dance between news sentiment and stock market movements has been an area of interest for financial analysts for years. The advent of the COVID-19 pandemic, however, threw a wrench in the works, turning predictable patterns into chaos. With the world gripped by unprecedented events, the traditional reliance on historical data and previous models came under scrutiny. Pre-pandemic studies, though rich in insights, suddenly seemed to lack the dynamism to factor in the pandemic's upheavals.

While researchers like Cakra and Trisedya (2015) have previously employed regression techniques, the modern world, with its rapid technological advances, has seen the rise of potent machine learning models like CatBoost. Yet, the exploration of such models, especially in the context of sentiment analysis during a global crisis, remains scant. Additionally, the existing literature's focus varies widely—some delve deep into specific sectors, while others cast a broader net, analyzing extensive indices

like the S&P 500. The NASDAQ, despite its significance, has seen limited research focus, further muddying the waters (Lee, 2022).

This fragmented approach to research has given rise to multiple challenges. Stakeholders, from hedge fund managers to individual investors, find themselves at crossroads. With no consolidated research guiding them, critical decisions about model adoption, the relevance of features, and the application of sentiment analysis across various sectors remain clouded in uncertainty. This prevailing ambiguity underscores the pressing need for a comprehensive study, one that not only bridges existing gaps but also paves the way for future research endeavors.

Literature Overview:

The intricate relationship between news sentiment and stock returns has been a subject of interest for financial researchers for many years. While several studies have shed light on this dynamic, there remains a notable gap in understanding its nuances, especially in the modern digital age. Cakra and Trisedya (2015), in their pioneering work, employed regression techniques to fathom this relationship. However, as the technological landscape evolved, newer machine learning models such as CatBoost have emerged, which, despite their potential, have not been extensively explored in the context of news sentiment analysis.

A review of the literature reveals that the majority of studies have adopted a rather narrow lens, focusing either on specific sectors, such as big tech firms, or on broad market indices. Lee (2022) is among the few who ventured into the realm of the NASDAQ, yet even such studies are sparse. The contemporary financial landscape, particularly with the advent of algorithmic trading and real-time news dissemination, mandates a deeper exploration. The significance of news sentiment, notably against the backdrop of the COVID-19 pandemic, has been accentuated, with digital news platforms playing a pivotal role in influencing investment decisions.

Furthermore, the inconsistencies in research methodologies and the lack of a unified approach have further complicated matters. Some studies have emphasized the role of positive news sentiment in driving stock prices up, while others have highlighted the adverse impact of negative news on stock performance. The role of neutral news,

however, remains largely underexplored. Additionally, while large-cap stocks and their sensitivity to news sentiment have been extensively studied, there's a dearth of research focusing on mid-cap or small-cap stocks.

This fragmented approach to research, coupled with the rapid advancements in machine learning and data analytics, underscores the pressing need for a comprehensive, comparative study. Such a study would not only bridge the existing gaps in the literature but also provide actionable insights to stakeholders, enabling them to make informed investment decisions in today's volatile financial markets.

Proposed Methodology:

The methodology for this research is designed to be robust, comprehensive, and adaptable to the intricacies of the modern financial landscape. At its core, it champions a data-driven approach, harnessing the power of state-of-the-art machine learning models to decode the relationship between news sentiment and stock returns.

Data Collection:

- Primary Data: We will gather real-time news sentiment data from reputed financial news aggregators and media outlets, focusing on articles, editorials, and reports that discuss stock market trends, company-specific news, and broader economic indicators.
- Secondary Data: Historical stock prices, trading volumes, and other relevant financial indicators will be sourced from established financial databases. This will help in understanding the past performance and establishing a baseline for our predictions.

Pre-processing:

Before feeding the data into our machine learning models, it will undergo rigorous preprocessing. This includes:

- Cleaning: Removing any inconsistencies, missing values, and outliers.

- Transformation: Converting textual news data into quantifiable sentiment scores using natural language processing techniques.
- Normalization: Standardizing the data to ensure that all variables are on a similar scale.

Model Training and Validation:
Four primary machine learning models – XGBoost, Random Forest, LightGBM, and CatBoost – have been chosen due to their prominence and proven efficacy in similar research contexts. Each model will be:

- Trained on a subset of the data, wherein the algorithm learns the relationship between news sentiment and stock returns.
- Validated on a separate subset, ensuring that the model's predictions are accurate and generalizable.

Comparative Analysis:
A side-by-side comparison of the models' performance metrics, such as Mean Squared Error (MSE), R-squared values, and prediction accuracy, will be conducted. This will provide insights into which model is most adept at leveraging news sentiment for stock return predictions.

Feedback Loop and Model Refinement:
Post-initial analysis, a feedback mechanism will be established. Any insights or anomalies observed will be looped back into the model training phase, ensuring continuous refinement and optimization of the models.

By embracing this holistic methodology, the research aims to not only shed light on the nexus between news sentiment and stock returns but also offer a replicable blueprint for future studies in this domain.

Analytical Approach:

Post data collection, the research will focus on a rigorous analysis phase. The efficacy of each model will be evaluated based on its prediction accuracy, with a keen

emphasis on understanding feature importance. Furthermore, sector-specific performance variations will be explored, providing a nuanced understanding of the models' adaptability across different sectors.

Limitations:

Potential limitations include the dynamic nature of news sentiment, which may vary based on external global events, and the possibility that some machine learning models may not be optimally suited for this specific type of analysis. Data availability and the inherent biases in news sources might also pose challenges.

Ethical Challenges and Risks:

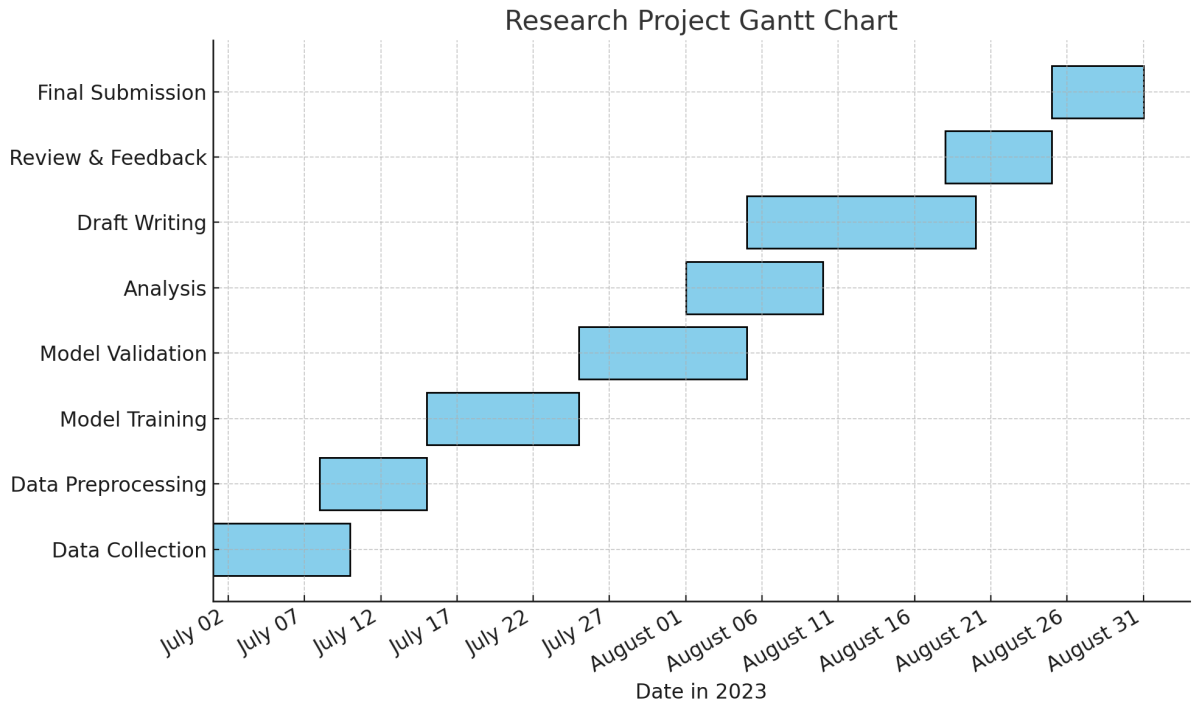
A paramount consideration will be ensuring that the interpretation of data does not inadvertently propagate misleading financial advice. Utilizing sentiment data from diverse news sources necessitates awareness of potential biases or misrepresentations. Adherence to the AREA (Anticipate, Reflect, Engage, Act) framework will be central to navigate ethical considerations.

Executive Summary Presentation:

The executive summary, encapsulating the research's salient findings, will be provided as a comprehensive written statement, ensuring clarity and accessibility for a diverse audience of stakeholders. This format is chosen for its traditional yet effective means of conveying complex information in a digestible manner.

Timeline and Project Plan:

A detailed Gantt chart will be developed, outlining the research's timeline, from initial data collection to final analysis and report drafting. This will ensure systematic progress tracking and timely completion of each research phase.



Data Acquisition and Analysis Approach:

Harnessing the right data is pivotal for the success of this research. We'll source data from reputable financial databases and news sentiment aggregators, ensuring a comprehensive dataset that encompasses a wide range of companies, sectors, and timeframes. The data's granularity will be essential, especially in capturing the nuances of stock returns during the pandemic.

Post-acquisition, the data will undergo rigorous preprocessing – handling missing values, removing outliers, and ensuring normalization where necessary. This ensures the data's integrity and readiness for modeling. The machine learning models, including XGBoost, Random Forest, LightGBM, and CatBoost, will be trained on a subset of this data and validated on another. The evaluation metrics, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), will provide insights into each model's predictive accuracy.

Stakeholder Relevance and Impact:

The research holds immense value for a broad spectrum of stakeholders, ranging from individual investors to institutional financial bodies. By elucidating the intricacies of predicting stock returns using news sentiment, especially during the tumultuous

pandemic era, the study offers a roadmap for investment strategies. Financial analysts can harness the insights to refine their models, while individual investors gain clarity on the weightage of news sentiment in their investment decisions.

Moreover, companies can glean insights into how news sentiments affect their stock performance, enabling them to strategize their public relations and media engagements more effectively. Policymakers, too, can harness the findings to understand the broader market dynamics during global crises, informing future policy decisions.

Conceptual and Theoretical Relevance:

Beyond its practical implications, this research also contributes conceptually to the evolving discourse on news sentiment's influence on stock markets. It builds on existing literature while introducing novel insights, especially in the context of the unprecedented COVID-19 era. The study's findings can serve as a foundation for future research endeavors, promoting a deeper understanding of the intricate interplay between news, public sentiment, and financial markets.

Ethical Considerations and Risk Management:

Beyond the previously mentioned ethical challenges, there's also the imperative to ensure data privacy. The datasets utilized will be scrutinized to ensure they don't contain personally identifiable information. While the AREA framework will guide the overarching ethical approach, specific measures will include transparent data sourcing, ensuring unbiased analysis, and avoiding overgeneralizations that might lead to financial misadventures for those relying on the research's insights.

In terms of risks, there's the inherent unpredictability of stock markets, which means that while models can predict with a certain accuracy, there's always the potential for unforeseen global events that can drastically shift market dynamics. Stakeholders must be made aware of this intrinsic uncertainty.

Executive Summary Approach:

To maximize reach and impact, the executive summary will be presented in both written and visual formats. A comprehensive written document will detail the findings,

while infographics will visually represent the study's key insights, ensuring accessibility for diverse audiences. These visual aids will be especially valuable for stakeholders who prefer a quick, yet comprehensive, overview of the research outcomes.

Closing Remarks:

This research proposal offers a detailed roadmap for a study poised to make significant contributions to the field of financial analytics. By delving deep into the relationship between news sentiment and stock returns during a challenging period, it aims to offer insights that are both practically relevant and conceptually enriching. The methodologies chosen, the ethical considerations, and the anticipated outcomes all align to ensure a research endeavor that's robust, relevant, and ready to make a meaningful impact.