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

**1. Introduction**

**1.1 Background**

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

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

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

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

**1.4 Objective**

This study seeks to bridge these gaps by embarking on a comparative evaluation of four cutting-edge machine learning models: Random Forest, XGBoost, LightGBM, and CatBoost, specifically focusing on their aptitude in predicting stock returns using sentiment scores during the COVID-19 era. Central to this research are three pivotal questions:

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

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

**2. Literature Review**

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

**2.1 Sentiment Analysis: A Brief Overview**

Sentiment analysis, or opinion mining, involves discerning and quantifying sentiments, emotions, and opinions within textual data. Early work in the domain, such as that by Pang, Lee, and Vaithyanathan (2002), focused primarily on product reviews. Their seminal research illuminated the potential of using textual analysis to infer consumer sentiments. Recently, the application of sentiment analysis has expanded manifold, especially in the field of finance. A study by Loughran and McDonald (2011) stressed the importance of crafting specialized dictionaries to mine financial texts, noting that traditional dictionaries often misclassify financial terms, leading to inaccurate sentiment classifications.

**2.2 News Sentiment and Stock Market Movements**

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

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

**2.3 Machine Learning Models in Financial Forecasting**

Machine learning models have brought about a paradigm shift in financial forecasting. The traditional linear models, while valuable, often fell short in capturing the intricate nonlinear relationships inherent in financial data. The advent of machine learning heralded a new era, enabling more nuanced predictions.

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

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

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

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

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

**2.5 Literature Review Summary**

In summary, the confluence of sentiment analysis and financial forecasting is an exciting domain, rich with possibilities. The literature provides robust evidence of the deep-seated relationship between news sentiment and stock market movements. With the continued advancements in machine learning and the increasing ubiquity of digital textual data, the future holds immense promise for further exploration and innovation in this interdisciplinary field.

**3. Data**

**3.1 Dataset Overview**

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

**3.2 Data Sources**

In this paper, we 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.

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

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

**3.3 Variables**

The dataset encompasses the following variables:

|  |  |  |
| --- | --- | --- |
| **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). |  |
| **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. | Yahoo Finance |

**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 or outliers. Addressing these issues is paramount to ensure the robustness of any modeling efforts that follow.

**4.1 Handling Missing Values**

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

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

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

**4.2 Visualization of Missing Data**

A visualization was generated to highlight the top 20 tickers with the highest count of missing Sentiment\_Score values. This visualization can aid in understanding whether the missingness is random or if it correlates with specific tickers.

A graph of a number of missing values

Description automatically generated

**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, we aspire to enhance the predictive power of our machine learning models. In the context of this study, where we aim to predict stock returns through an amalgamation of stock price data, news sentiment, and pandemic-related metrics, the significance of feature engineering cannot be overstated.

**5.1 Time-Based Features**

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

**5.2 Pandemic Phase Indicator**

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

**5.3 Return Calculations and Classification**

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

A graph of a number of negative return

Description automatically generated

**5.3 Volatility**

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

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

**5.4 Correlation Matrix**

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

A screenshot of a graph

Description automatically generated

Figure 3 Correlation Matrix of Features

The heatmap unveils the correlation coefficients between different features:

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

The correlation matrix is paramount for two main reasons:

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

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

**6. Exploratory Data Analysis (EDA)**

**6.1 Descriptive Statistics**

First, let's 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. We commence by summarizing the numerical attributes within our dataset.

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

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

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

Sentiment\_Score: The average sentiment score is around 0.67. However, this metric is susceptible to significant fluctuations, spanning from -96.88 to 66.62. This wide range underscores the immense variability in daily news sentiment for different firms.

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

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

**6.2 Data Visualization**

**6.2.1 Distribution of Sentiment Score**

First, we plot the distribution of the Sentiment\_Score variable to understand how news sentiments are distributed over the entire dataset.

A graph of a distribution of sentiment scores

Description automatically generated

The histogram illustrates the distribution of the sentiment scores. A few observations can be made:

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. However, the existence of both positive and negative extremes indicates significant news events that have affected the stock market.

**6.2.2 Temporal Evolution of Average Sentiment**

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

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

A graph showing a blue line

Description automatically generated

**6.2.3 Average Sentiment Score by Company**

By examining the average sentiment scores across various companies, we can discern patterns or biases in news coverage. Notably, certain companies (e.g., META) might consistently receive positive news coverage due to their strategic decisions, innovations, or corporate governance, while others might often be portrayed in a negative light.



The horizontal bar chart illuminates the average sentiment scores for various companies over the dataset's duration:

* Companies exhibit a broad spectrum of average sentiment scores, ranging from slightly negative to notably positive. This diversity in sentiment underscores the differential media portrayal and news coverage these entities receive.
* A majority of the firms cluster around a neutral average sentiment, indicative of balanced news coverage over the period.
* Certain companies like 'ZS', 'WBA', and 'WBD' tend towards the positive end, suggesting consistent favorable news coverage. Conversely, firms like 'DDOG', 'LCID', and 'GEHC' lean towards the negative end, implying that they might have been subjects of more critical or unfavorable news.
* The disparities in sentiment scores across companies might be attributed to a myriad of factors – from sectoral influences, corporate strategies, financial performance, to market positioning.

Unearthing these company-specific sentiment patterns offers valuable insights, enabling a nuanced understanding of the intricate interplay between news sentiment, stock returns, and company attributes.

**6.2.4 Time Series of New COVID-19 Cases**

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



Figure XXX (figure number): XXX (figure name)

Notes: (please add additional info, such as the definition of the lines here)

* 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.

**6.2.5 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.



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

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

**7. Feature Selection**

(Still thinking whether I should do or not)

**8. Model Evaluation**

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

**8.1 Random Forest**

**8.1.1 Theoretical Overview**

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

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

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

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

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

**8.1.2 Methodology**

In the context of this study:

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

**8.1.3 Results**

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

where:

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

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

A colorful lines on a white background

Description automatically generated

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

**8.1.4 Discussion**

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

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

**8.2 XGBoost**

**8.2.1 Theoretical Overview**

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

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

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

where:

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

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

**8.2.2 Methodology**

For our analysis:

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

where evaluation metric represents…

The early stopping feature is particularly beneficial in preventing overfitting. The model's training stops if the performance metric on a validation set doesn't improve for a set number of rounds.

**8.2.3 Results**

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

A screen shot of a screen

Description automatically generated

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

**8.2.4 Discussion**

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

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

**8.3 LightGBM**

**8.3.1 Theoretical Overview**

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

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

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

where:

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

**8.3.2 Methodology**

For the LightGBM model:

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

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

**8.3.3 Results**

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

A colorful striped chart with a black border

Description automatically generated with medium confidence

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

**8.3.4 Discussion**

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

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

**8.4 CatBoost**

**8.4.1 Theoretical Overview**

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

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

where:

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

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

**8.4.2 Methodology**

For the CatBoost model:

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

**8.4.3 Results**

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

A screen shot of a screen

Description automatically generated

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

**8.4.4 Discussion**

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

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

**8.5.1 Methodology**

To provide a comprehensive understanding of the relative strengths and weaknesses of the four models in the context of our dataset, we 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.

**8.5.2 Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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.

**9. Discussion**

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

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

**9.2 Variability Among Models**

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

**9.3 Limitations and Scope for Future Research**

Every analytical endeavor, while shedding light on certain aspects, also comes with its set of constraints. Our 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 our models are trained on a decade of data, the inherent unpredictability of stock markets means that past performance is not always indicative of future results. This principle is well-embedded in financial literature and serves as a reminder of the tentative nature of stock return predictions.

**10. Conclusion**

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

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

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

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

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

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

References

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

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

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

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).

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.

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.

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*.

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.

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.

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

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

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*.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*.

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*.