Analyzing the Correlation Between Retail Traders’ Sentiments and Equity Market Movements

Haozhe Zeng | Cornell University | [hz657@cornell.edu](mailto:hz657@cornell.edu)

Zixiao Wang | Cornell University | [zw699@cornell.edu](mailto:zw699@cornell.edu)

## Abstract

This research seeks to explore the impact of retail traders' sentiments, primarily from forums like Twitter (X) and Reddit, on equity market movements. The investigation will discern the duration of this correlation, whether it's short-term or extends to mid-long term. It will also ascertain if the correlation is more pronounced in specific stock categories like penny stocks or tech giants or if such a correlation might be absent altogether. Upson determining the relationship, the project aim to develop a machine learning model that can detect potential trading signal.

## Introduction

The financial arena has experienced a profound metamorphosis in the past few years, predominantly propelled by the digitization of trading platforms. Such innovations have democratized financial market access, resulting in an influx of retail traders actively engaging in stock trading. Characterized by their agility and swift mobilization capacity, these traders have ascended as a potent force in the equity market, contesting the dominance of traditional institutional entities.

One of the most prominent platforms that has come to symbolize this new wave of retail trading is the WallStreetBets forum on Reddit. Serving as a discussion hub, WallStreetBets has become a focal point for retail traders to share insights, strategies, and sentiments about various stocks. The power of such collective sentiment became glaringly evident during events like the [GameStop short squeeze](https://en.wikipedia.org/wiki/GameStop_short_squeeze), where concerted buying actions driven by discussions and emotion on the forum led to unprecedented stock price surges, catching many institutional investors off guard.

However, while events like the GameStop incident have made headlines, a holistic analysis scrutinizing such forums' overarching influence on the equity market is yet to be undertaken.

This initiative intends to meticulously examine the sentiments reverberating within these platforms and ascertain their potential linkage with equity market oscillations. The endeavor is not just about identifying superficial correlations; it seeks to fathom the extent and intensity of such influences. Inevitably, questions emerge: Is there consistency in these correlations across varied stock sectors? Is the impact of these sentiments more pronounced for specific stock types, be it penny stocks or industry giants? And crucially, can the sentiments serve as a predictive tool for forecasting market trajectories?

## This study's ambition is to offer an exhaustive insight into the dynamic interplay between retail trader sentiments and the intricacies of equity market behavior in this digital era. The ultimate aspiration is to harness the insights garnered from sentiment analysis to gauge midterm market fluctuations. Adding an intriguing dimension, the project also explores the potential of crafting a machine learning model aimed at identifying trading signals.

### Why this approach?

Previous research has studied the relationship between retail sentiment and equity market movements, with many findings showing a positive correlation. There's also been interest in using machine learning to predict stock prices. However, the integration of sentiment analysis with machine learning to predict stock movements is less common in the literature.

Most prevailing research models tend to fall into one of two categories: they either employ machine learning for long-term stock predictions, sidelining sentiment analysis, or they entirely overlook the sentiment component. This presents a significant oversight. The stock market is inherently dynamic, continuously shaped by a myriad of factors. To solely rely on a monolithic prediction model, as many current studies opt to, poses limitations.

There emerges an undeniable imperative for more adaptive techniques, such as the rolling window method. By ensuring periodic model training and consistent recalibrations in line with fresh data, this approach promises a model that evolves in tandem with market changes, ensuring a more robust prediction mechanism.

## Challenges and what did not work

### Data

Data Acquisition: Gathering relevant and high-quality data presented significant obstacles. We initially aimed to source real-time data from platforms like Reddit and Twitter. However, our efforts were hampered by API rate limits, reducing our collection efficiency. The premium versions of these APIs come at a steep price, while the free versions have many limitations. We had to explore alternative methods to gather sufficient stock-related posts; otherwise, the scarcity of data could severely hinder our model's performance.

### Sentiment Analysis

Noise Filtering: We sourced a dataset containing over 1.6 million Twitter posts. However, this dataset wasn't exclusively about the equity market; instead, it was a broader collection of general tweets. Other datasets we identified that were specific to the equity market were either unlabeled or contained a limited number of posts, typically around 8,000 entries. Currently, we are using the 1.6 million post dataset as our training set and the smaller, equity-specific datasets as validation sets. Given the non-specific nature of the larger dataset (with many irrelevant posts), our model's training set accuracy stands at 80%. In contrast, its accuracy on the test set drops to approximately 60%. This is visualized in the confusion matrices shown below:



**Steven: what you have done to solve this problem. Include everything you have tried (success and failure)**

**Steven: what is your current state**

**Steven: what is your future plan**

### Stock

Several studies propose using a singular model to forecast stock returns for an extended period, sometimes spanning up to a hundred days. I find this approach potentially limiting. Given the dynamic nature of the market, relying on one model to predict returns over multiple days seems unrealistic.

In contrast, I advocate for a model that is recalibrated daily, leveraging fresh data for each day's prediction. After forecasting the next day's or even the next week's return, the model can then assimilate the actual return data for that day. This iterative approach allows the model to continually refine its predictions based on the latest market conditions. Termed the "rolling window" method, this strategy emphasizes daily predictions while updating the dataset after each forecast. Such an approach is more attuned to the market's dynamic, enhancing the accuracy and relevance of predictions. 

Two critical components define a rolling window model: the window size and the duration of the return you're predicting. While this model excels in capturing market dynamics, it can be computationally demanding due to its iterative training nature. Determining the optimal window size poses a challenge, as it can range from a short span of 100 days to several thousand days. Naturally, larger window sizes intensify the computational burden. When working with intricate deep learning models like Long Short-Term Memory (LSTM), it might be more reasonable to set a threshold for deciding when to update the model, rather than retraining it at every iteration. This can balance the need for updated information with the practicalities of computational efficiency.

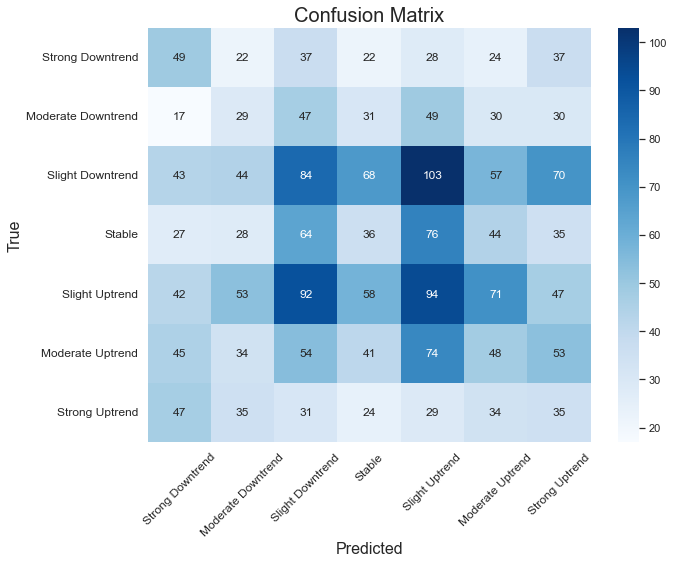
Moreover, determining the precise aspect to predict brings its own set of challenges. The main objective of this research is to identify the correlation between retail sentiment and stock movement. Given this aim, it initially seemed fitting to treat it as a classification challenge, aiming to predict if the stock movement for the next day would be positive or negative.

To achieve this, I used an XGBoost classifier. The methodology applied to categorize the next day's return was as follows:

1. If the return value fluctuated between -0.002 and 0.002, it was categorized as 'Stable'.
2. A slight increase between 0.002 and 0.01 was labeled as 'Slight Uptrend', while a slight decrease between -0.01 and -0.002 was termed as 'Slight Downtrend'.
3. If the rise was between 0.01 and 0.02, it indicated a 'Moderate Uptrend', and a fall between -0.02 and -0.01 indicated a 'Moderate Downtrend'.
4. Any return value above 0.02 was classified as a 'Strong Uptrend', while any value below -0.02 was termed as a 'Strong Downtrend'.

The initial approach to understanding stock movements involved a detailed categorization ranging from stable periods to pronounced uptrends and downtrends. The confusion matrix presented showcased the performance of this classification method. The methodology took approximately 20 minutes for execution, which may be deemed lengthy for real-time analysis. When using only the stock price as a predictor, the accuracy was marginally better than a random guess. This suggests a need for a more comprehensive and efficient methodology.

Recognizing the limitations of the classification approach, a shift towards regression was considered. The rationale was that predicting a continuous outcome (the stock's future price or return) might be more effective. Once the future price is estimated, it can then be discretized into categories. The revised strategy employed an XGBoost Regressor, aiming to predict the next day's return. The predicted return was then converted into a categorical representation of stock movement. The XGBoost Regressor was notably more efficient, completing its run in about 4 minutes. This model achieved an accuracy of approximately 20%. While this is an improvement over the classifier approach, there remains room for enhancement.



### Current State, Stock

In the face of consistent challenges, it became evident that our model's focus on predicting next-day returns might not be the optimal approach. A deeper dive into the methodology and its implications illuminated several key insights.

1. Temporal Dynamics of Sentiment: Leveraging sentiment analysis in our model highlighted that the effects of retail sentiment on stock prices aren't instantaneous. Rather, there's a lagged impact, reflecting a more gradual influence on stock movements.
2. Uncertainty of Short-Term Predictions: The attempt to predict single-day returns proved fraught with uncertainties. Factors like daily news events, global market movements, and institutional trading decisions can cause significant price fluctuations in the short term.
3. Broadening the Timeframe: Our objective isn't about chasing daily fluctuations but understanding broader market dynamics. Adopting a swing trading perspective, which focuses on capturing gains in a stock (or any financial instrument) over a period of several days to weeks, aligns more closely with our goals.

In light of these insights, it's clear that a shift in strategy towards predicting mid to long-term stock movements, taking into account the more subtle and prolonged impacts of retail sentiment, could provide a more accurate and actionable framework for our endeavors.

Consequently, I shifted the focus of my model to forecast the returns for the upcoming week. My primary interest transitioned from pinpointing stock movements to uncovering viable trading strategies, which I deem to be more pragmatic. As it stands, I employ a rolling window time series model. Each day, the model predicts the stock price for five days ahead and undergoes daily retraining to assimilate the latest information.

The objective now is to predict the stock price of APPLE for a given time frame. Various features and methodologies were experimented with, to improve the model's performance.

1. Dataset: Stock price of APPLE from 2010/01/01 to 2023/01/01.
2. Initial Metric: The model started with a Mean Absolute Percentage Error (MAPE) of approximately 4%.
3. Model Optimization:
   1. The model was adjusted to run on only 10% of the original time series data for feature selection, which optimized processing and ensured a more streamlined approach.
4. Feature Selection:
   1. Several features were experimented with, including volume data, open price and its lags, highs and lows of a day, and economic indicators.
   2. Inclusion of moving averages (Mas) and SPY brought a significant increase in the model’s performance.
   3. Several other features were added and tested such as RSI, WVAD, MACD, CCI, BOLL, and others. However, not all added significant value to the model's predictive capability.
5. Model Performance:
   1. After multiple iterations, feature additions, and adjustments, the model achieved a MAPE of 1.87%. This is a notable improvement from the initial 4%.
6. Processing Time:
   1. The model takes approximately 2.46 minutes to run on the entire dataset, demonstrating efficiency in processing.
7. Window Size:
   1. A window size of 200 was used for the model, typically representing 1 year of stock history.

The stock price prediction model for APPLE has undergone extensive fine-tuning and experimentation. The emphasis on feature engineering and model adjustments has led to a significant improvement in prediction accuracy, as evidenced by the reduction in Mean absolute percentage error (MAPE) from 4% to 1.74% for the training set. The inclusion of moving averages (MAs) and SPY as features was especially beneficial, highlighting the importance of these variables in predicting APPLE's stock price.

The current model is a rich compilation of various columns, each presenting a unique facet of stock market information. The depth and variety of these columns allow for in-depth analysis and the crafting of sophisticated trading strategies. Here's a succinct breakdown of each column:

1. **Date**: Represents the specific day for the data point, giving chronological context to the observations.

**Price & Volume Columns**:

1. **Close**: The price at which the stock settled at the day's end.
2. **Close\_lag\_i**: A historic reference, this reflects the closing price from 'i' days ago, aiding in drawing comparisons over time. The current data set includes a 10-day lag.
3. **Volume**: Represents the sheer volume of shares that exchanged hands on that day, indicating the day's trading intensity.

**Moving Averages**: These offer a smoothed version of the price data, revealing underlying trends by averaging out short-term fluctuations:

1. **MA5**: Reflects short-term trends using a 5-day period.
2. **MA10 & MA20**: Capture medium-term movements.
3. **MA50 & MA200**: Provide insights into longer-term trends and are particularly watched by traders.

**Indicators**: These are a mix of momentum, volume, and volatility metrics that traders often utilize to decipher market sentiments:

1. **WVAD**: This indicates the flow of money, revealing the balance between buying and selling pressure.
2. **MACD**: Illustrates the relationship between two moving averages of a stock's price. It's accompanied by:
3. **macd\_line**: The main line indicating the trend.
4. **signal\_line**: The trigger for buy and sell signals.
5. **RSI**: Measures the speed and change of price movements, often used to identify overbought or oversold conditions.
6. **CCI**: Helps in determining cyclical trends.
7. **BB\_Upper, BB\_Lower, Buy\_Signal & Sell\_Signal**: These boundaries of the Bollinger + RSI, Double Strategy serve as volatility indicators.
8. **WVF, WVF\_color, upperBand & rangeHigh**: Relates to the Williams Vix Fix, identifying bottoms in stock advancements.
9. **VPT**: Combines volume and price to spotlight changes in trend direction.
10. **AD**: Shows the flow of money, offering insights into the accumulation or distribution state of the stock.

The current model performance:

Processing: 100%|██████████| 2866/2866 [02:46<00:00, 17.20it/s]

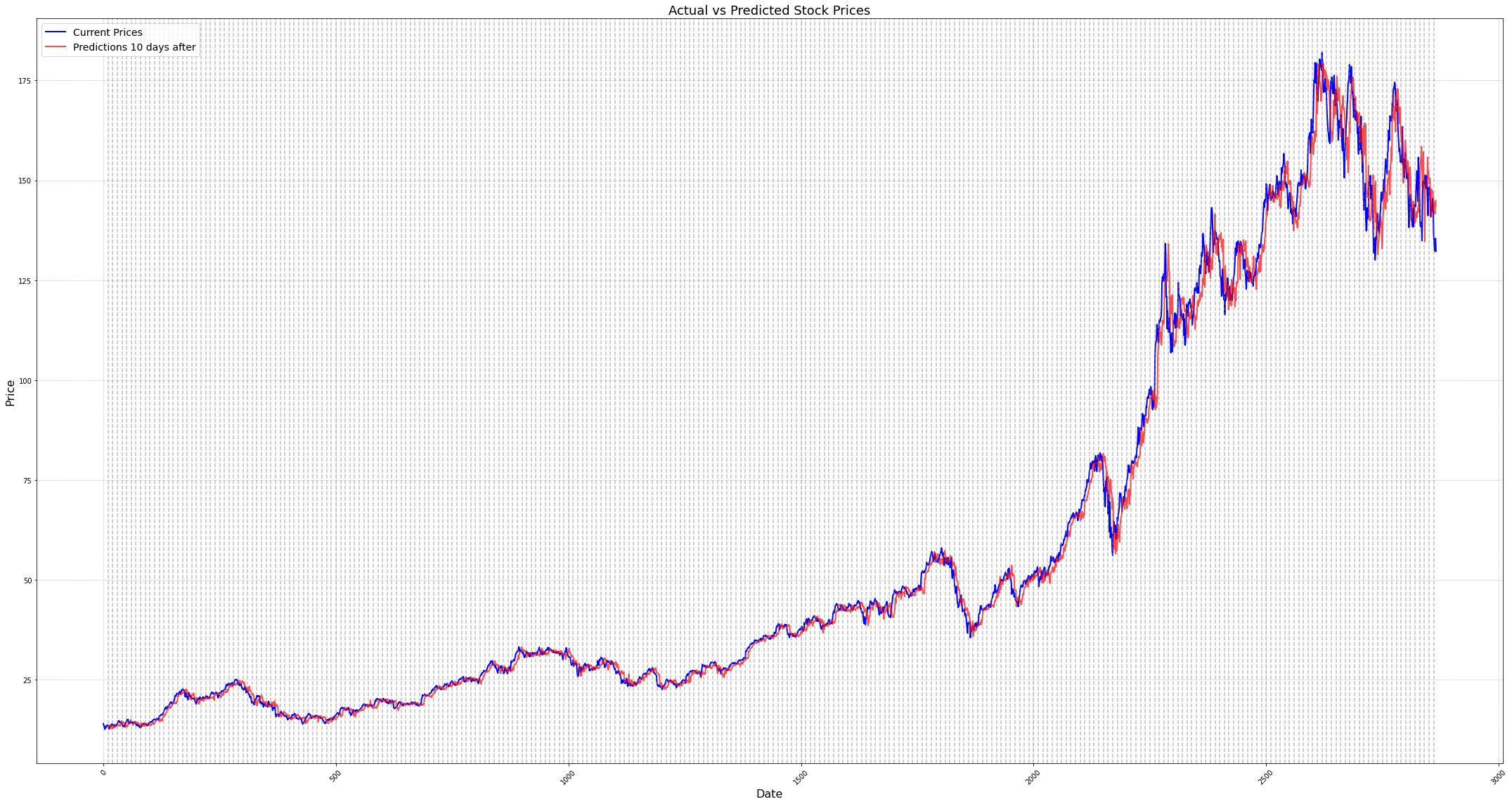
Mean Squared Error (MSE): 4.568191920142539

Mean Absolute Percentage Error (MAPE): 1.87%

Root Mean Squared Error (RMSE): 2.137332898764846







The displayed residual plot showcases the differences between observed and predicted values for a given model. The data points, represented as blue dots, seem to be scattered randomly around the horizontal red-dotted line, which signifies zero residual or perfect prediction. The random dispersion indicates that the model has a good fit for the data, as there's no discernible pattern or trend in the residuals. This suggests that the model's assumptions, particularly those regarding linearity, independence, and homoscedasticity, are likely met. However, there are a few notable outliers, which might require further investigation to understand if they result from specific external factors or data anomalies. Overall, the residual plot suggests a well-performing model, but attention should be given to the few outliers present.

### Trading Strategy:

In the realm of financial forecasting, possessing merely a model that predicts weekly outcomes falls short of the comprehensive approach needed. What truly matters is the development of a sturdy methodology that seamlessly translates these projections into concrete, actionable measures, ultimately leading to a sophisticated trading strategy. To this end, I have architected a straightforward yet effective strategy that seamlessly integrates predictive return analytics with in-depth historical stock price information.

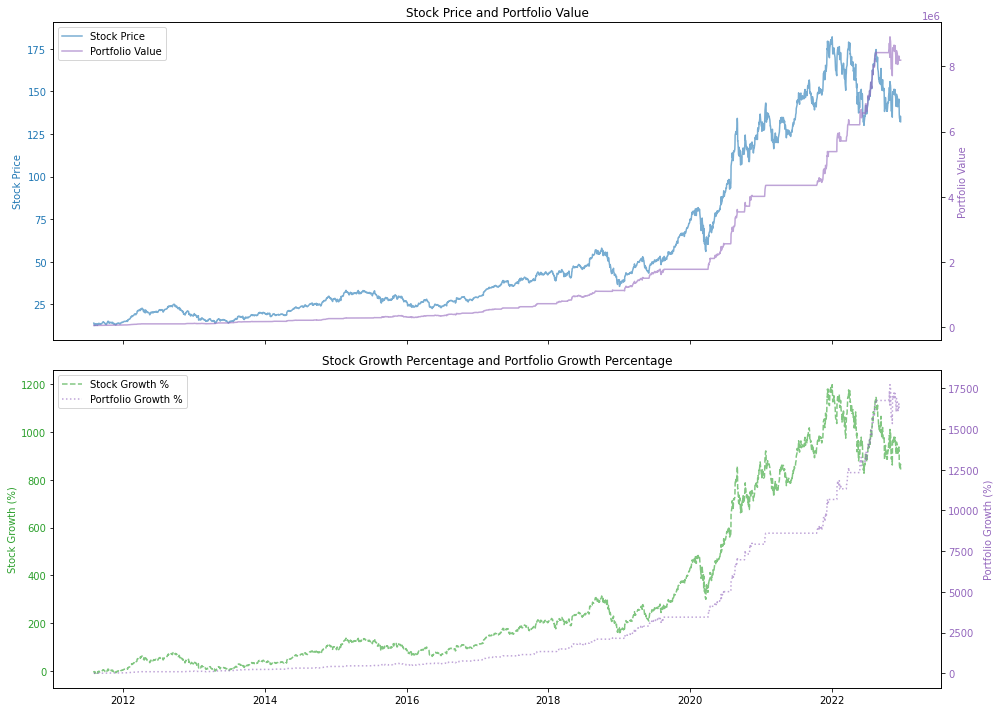
**Forecast Generation:** Predictions are meticulously crafted for each stock data entry following the stipulated window size. Concurrently, both the predicted and actual returns spanning a 5-day period are discerned.

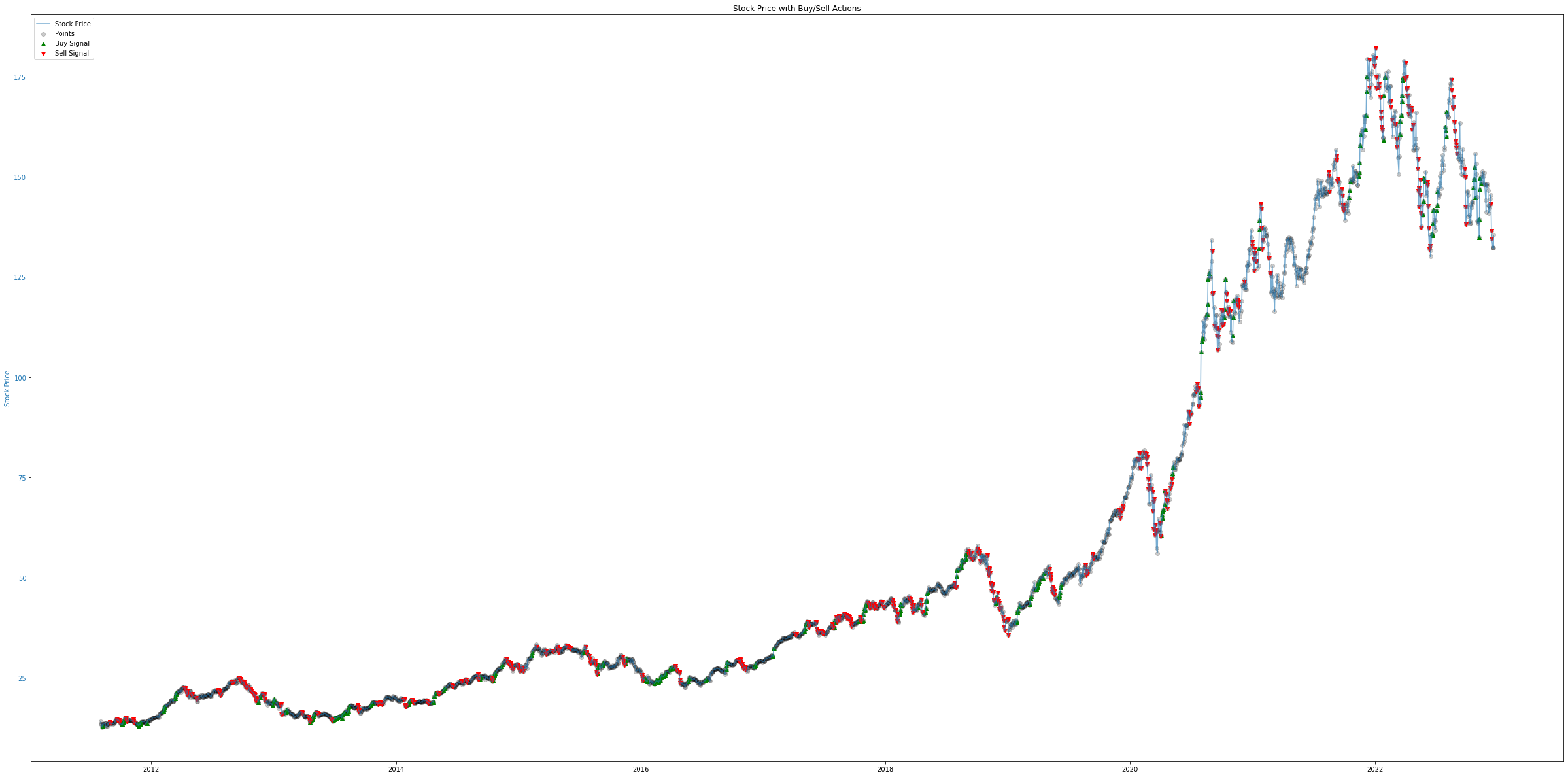
**Portfolio Initialization:** The strategy springs to life with a pre-allotted capital and without any initial stock engagements. As it unfolds, the cumulative value of the portfolio, synthesizing both available capital and the value of held stocks, is consistently monitored and documented.

**Trading Logic—The Pivotal Mechanism:** At the heart of our strategic architecture lies the adaptive trading logic. This mechanism sets buy and sell benchmarks anchored on prior window\_size real returns, employing the 75th and 25th percentiles as guiding metrics. For every predictive interval, signals that either breach the buying criteria or fall below the selling criteria are registered. Decisions flow organically from these cues:

1. **Acquisition Strategy:** Should there be at least three robust buy prompts within the forecasted range and provided there's ample capital on hand, a stock purchase is greenlit.
2. **Divestment Strategy:** On the flip side, if three or more sell prompts surface and stocks are held, a divestiture move is undertaken.

**Strategy Visualization:** The tangible outcomes of the trading strategy are vividly portrayed through two primary graphical illustrations: The premier chart contrasts the stock's market performance with the trajectory of the investor's portfolio. In tandem, the following chart illuminates the parallel growth dynamics of the stock and the portfolio. Complementing these, an exhaustive visual narrative plots out each buy/sell maneuver against the backdrop of the stock's price timeline.





Currently, our model and trading strategy outperform the stock's return by 10x, demonstrating the ability to yield profits even amidst bearish market trends. While our present framework adeptly identifies selling cues, it requires further refinement in effectively discerning buying signals.

### Next Step, Stock

Moving forward, there are several avenues to explore to enhance the robustness and efficacy of our stock prediction model:

1. **Feature Enrichment:** Dive deeper into the inclusion of potential predictors that encapsulate global economic shifts and overarching market dynamics. Such features can be pivotal in capturing exogenous shocks and external factors influencing stock prices.
2. **Addressing Multicollinearity:** A systematic evaluation of the features is essential to ascertain any collinearity present. Multicollinearity can undermine the model's interpretability and diminish its predictive prowess. Utilizing techniques like Variance Inflation Factor (VIF) can assist in detecting and mitigating these issues.
3. **Advanced Trading Strategies:** Expand the trading strategy's scope to encompass more sophisticated tactics such as short-selling. This would allow capitalization on both upward and downward market movements, offering a more holistic trading approach.
4. **Refining Buy-Signal Identification:** Given the current model's shortcoming in accurately pinpointing buying signals, targeted efforts should be made to optimize this aspect. This might involve recalibrating threshold values or integrating alternative algorithms.
5. **Incorporating Sentiment Analysis:** A key dimension that's often overlooked is the sentiment prevailing among retail investors. Once a reliable sentiment analysis model is in place, merging it with the current framework could provide a more rounded perspective on market movements. Analyzing chatter on social media platforms, financial forums, or news outlets can be instrumental in this regard.
6. **Model Evaluation and Continuous Feedback:** It would be prudent to establish a feedback loop where the model's predictions are constantly compared with actual outcomes. Such a mechanism would be invaluable for ongoing model refinement. Also, other models other than XGBoost are left to be experimented.
7. **Stress Testing:** Given the unpredictable nature of financial markets, stress-testing the model under various hypothetical adverse scenarios can provide insights into its resilience and areas of potential vulnerability.

By adopting these strategies and continually iterating based on real-world outcomes, we can aspire to achieve a state-of-the-art stock prediction model that's both adaptive and predictive in an ever-evolving market landscape.

## Current State

### Data

### NLP

### Stock Prediction

## Future Scope (Next Month)

# Citations:

Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE, 12(7): e0180944. <https://doi.org/10.1371/journal.pone.0180944>

Relation: Bao, Yue, and Rao present a deep learning framework utilizing stacked autoencoders and long-short term memory for analyzing financial time series. Notably, they introduce concepts of buy and sell signals based on predicted prices, resonating with our exploration into machine learning-driven financial predictions.

Differentiation: While they lay the groundwork in understanding financial time series through deep learning, our research extends this by incorporating contemporary machine learning methodologies to forecast stock returns over shorter durations. Additionally, we delve into portfolio management through our trading strategy, a topic not explored in their paper.

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. The Journal of Finance and Data Science, 2(1), 42-57. <https://doi.org/10.1016/j.jfds.2016.03.002>

Relation: This paper delves into trading signals and the intricacies of implementing a comprehensive trading strategy. Its content is rich in explaining how trading decisions can be informed and executed.

Differentiation: Unlike the paper's emphasis on broader sectors like SPY, our approach zeroes in on individual stocks. Our research also capitalizes on a myriad of indicators, dedicating significant effort to feature selection and engineering, aspects that weren't as extensively addressed in the referenced paper.

Pezim, B. (2018). How To Swing Trade. Preface by A. Aziz. ISBN: 9781726631754.

Relation: The book provides an extensive overview of swing trading strategies and market dynamics, setting the stage for our exploration of stock market behaviors.

Differentiation: Our project enhances these basic principles with state-of-the-art machine learning techniques to forecast stock market returns, delivering a modern, technology-enhanced viewpoint.

### Online Social Media and Stock Market:

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

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. Journal of Economic Behavior & Organization, 107, 730-743.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies, 27(5), 1367-1403.

### Sentiment Analysis and Opinion Mining:

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

Kumar, A., & Lee, C. M. (2016). Retail investor sentiment and return comovements. The Journal of Finance, 61(5), 2451-2486.

### NLP Techniques for Financial Markets:

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. The Journal of Finance, 66(1), 35-6

## GitHub

<https://github.com/howie-zeng/Analyzing-the-Correlation-Between-Retail-Traders--Sentiments-and-Equity-Market-Movements>