A Multifactorial Analysis of Microsoft Corporation's Stock Price Dynamics

CAPSTONE PROJECT WINTER 2024

SABABA AHMAD

Introduction

In the world of finance, understanding stock price behavior is a pivotal cornerstone for building a successful investment strategy. Given the breadth and range of factors that influence the dynamics of stock price movements, achieving a comprehensive insight into these variables presents a significant challenge. This project aims to conduct a detailed analysis of these factors, ranging from financial performance metrics, fundamental data analyses, the impact of news cycles, social media sentiment, implied volatility, and technical indicators, among others.

The objective of this study is to implement a bottom-up analysis of an individual stock – Microsoft Corporation (MSFT), and identify which factors exhibit the strongest association with the subsequent day's stock price. This study will then leverage these insights to make a one-day price prediction. The intention is to create a systematic workflow that can be replicated to any other publicly traded security under consideration and more generally informing short-term trading strategies.

Related Work

Speech emotion recognition and text sentiment analysis for financial distress prediction by Petr Hajek and Michal Munk used the finBERT financial sentiment analysis tool to analyze emotions in earnings call transcripts as a factor for predicting financial distress. The authors examined 1278 conference calls from the 40 largest US corporations and categorized emotions detected in managerial speech. This was combined with traditional machine learning and predictive models to find evidence of an improvement in prediction accuracy given the essential role of managerial emotions to predict financial distress. This study was instrumental in developing the conceptual framework for sentiment analysis in this project.

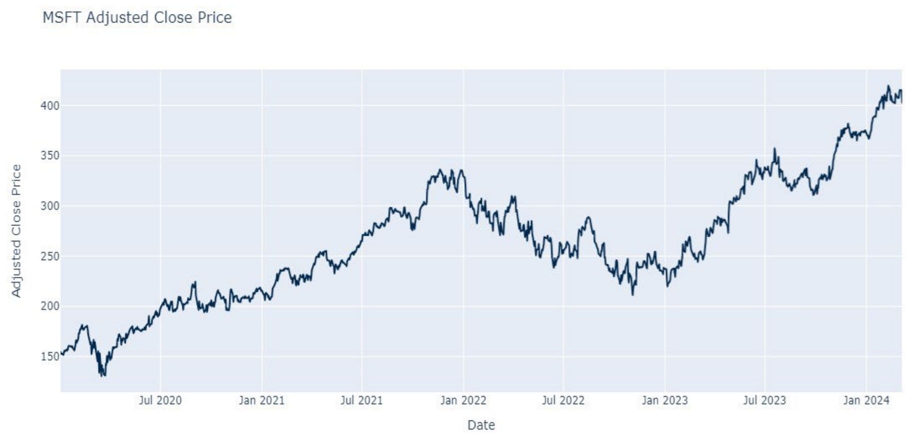
Implied volatility directional forecasting: a machine learning approach by Vrontos et al looks at the effectiveness of machine learning techniques in predicting the direction of implied volatility for the VIX index, the measure of market volatility based on options of the S&P500 stocks. Vrontos compares the performance of traditional econometric models with machine learning models to forecast the direction of implied volatility. This paper was the base on which was built the concepts to analyze the implied volatility of MSFT. Furthermore, the feature selection techniques employed by Vrontos informed the choice of relevant factors for this project.

The blog "WallStreetBets Impact on the Market" from Quiver Quantitative was also a relevant source of information in building the conceptual basis for the sentiment analysis component of this project. This article demonstrates the rapid rise of retail investors and online discussion forums like WallStreetBets being a source of inspiration for selectively focused, high-risk investment strategies. The frequency of stock mentions for less liquid stocks (lower trading volume) has the possibility of influencing their prices, suggesting a potential link between social media sentiment and stock price behavior. The article focuses on the correlation between activity in the discussion forums and stock price movement but does not explore causation in this phenomenon.

Data Sources, Factors, Features and Pre-Processing

Past Prices and Trading Volume

The adjusted closing price of MSFT was the target variable in this study. Price and volume data for the period January 1, 2020, to March 5, 2024 was obtained from Yahoo! Finance using the yfinance package. This data was of daily frequency so there were minimal pre-processing and cleaning steps involved.



CBOE Volatility Index (VIX)

The VIX index is a real-time indicator of the market's expectation for volatility over the coming time period. This index is used by investors to assess the level of risk or fear in the market when making investment decisions. This factor was included because it acts like a 'fear gauge', potentially indicating sentiment towards MSFT. 9-day, 30-day, 3-month and 6-month VIX indexes were used in this study, obtained from Yahoo! Finance using the yfinance library. This data was also of daily frequency so minimal cleaning and pre-processing were required.

Implied Volatility

Implied volatility is a measure used to estimate the degree of future price variability of a stock, based on the prices of its options. It reflects investors' predictions about the asset's potential movement but does not indicate the direction of the movement. This data was obtained from Nasdaq Data Link (formerly Quandl) using an API key. This dataset contained 64 features encompassing historical volatility, predicted historical volatility, implied volatility, and volatility skew. Once again, this dataset was of daily frequency so minimal cleaning and pre-processing were required.

Stock Grade, Rating and Analyst Recommendation

These datasets contain categorical data evaluating MSFT as an investment. The grades, ratings and recommendations are provided by analysts and ratings agencies, and reflect their views on the potential performance and creditworthiness of Microsoft. These datasets were obtained from Financial Modeling Prep with an API key and cover the period January 2020 to March 2024. The ratings and grades comprise of categories like “Buy”, “Sell”, “Hold”, or letter grades like “AAA”, “A”, “B”, etc. The datasets also contain the number of analysts and agencies that issued a certain grade or rating. During pre-processing, these two dimensions of data (number of ratings and the ratings themselves) were converted to a single dimension by assigning a number to each grade or rating, with 5 indicating a strong investment and 1 indicating a weak investment. This was then multiplied by the number of analysts that issued that grade to arrive at a weighted average. One issue with this dataset was the misalignment of the periodicity: unlike with stock price and volatility data, this dataset did not contain daily observations. Some dates contained multiple observations, for which the average weighted rating was used, while other dates did not contain any observations, for which a forward-fill method was used.

Financial Statements

An Income Statement illustrates a company’s finance performance and profitability over a period, a Balance Sheet provides a snapshot of a company’s financial position at a specific point in time and a Cash Flow Statement is a summary of a company’s cash inflows and outflows over a period of time. According to financial theory, these should be the strongest drivers of a company’s stock price. These datasets cover the period January 2020 to December, 2023 and were obtained from Financial Modeling Prep using an API key. The misaligned periodicity was the main challenge in working with this dataset because these were quarterly financial statements. Balance Sheet data could theoretically be linearly interpolated but Income Statement and Cash Flow data were more complex to pre-process because statement filing dates did not always fall on a stock trading date. The “Business” frequency of Pandas datetime format did not completely align with the stock trading dates so these datasets needed to be indexed using a custom indexing function. Furthermore, linear interpolation is not applicable for Income Statements and Cash Flow Statements and surprisingly no functions exist to convert quarterly financial performance to daily so custom functions needed to be created for this purpose also. One consideration that had to be made was whether to convert all daily data in this project to quarterly data to align them with the financial statements, which would mean losing information about daily behavior on other features, or whether to convert quarterly financial statements to daily but thereby introducing more noise and “manufacturing” smoothed data that is not completely “pure”. The period for January to March 2024 was extrapolated using the financial statement growth rates from December 2023.

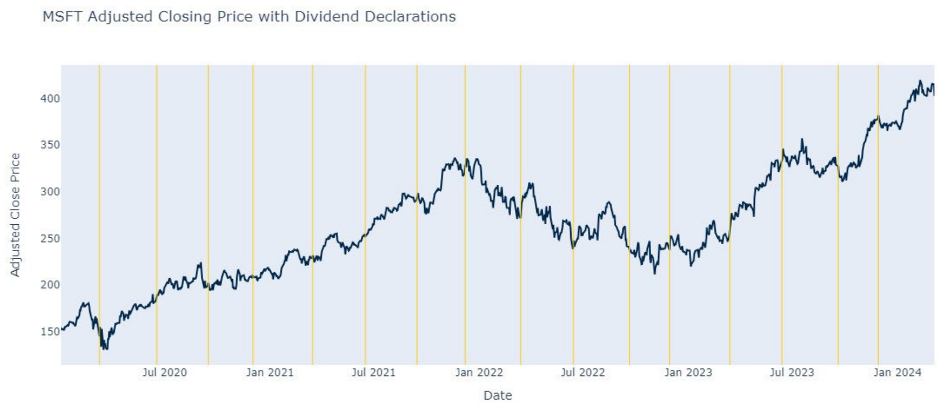
Earnings Surprise

This is the difference between expected and actual Earnings Per Share. This data also spans January 2020 to December, 2023 and was obtained from Financial Modeling Prep. Like the financial statement data, this dataset was also quarterly, but the remaining values were not interpolated and instead were filled with zeros because conceptually there were no “surprises” on the dates in between earnings releases.

Metrics and Ratios

These datasets contain information like Earnings Per Share, Net Profit Margin, EBITDA, liquidity and profitability ratios, Return on Equity and other important indicators that provide critical insights into a company’s health, profitability, and operational efficiency. This data spans January 2020 to December, 2023 and was obtained from Financial Modeling Prep. This data is derived from financial statements so was originally quarterly in frequency but was linearly interpolated to align with stock price data.

Dividends



Dividends declared and paid are a sign of a company's health, influencing investor perception as they react to these signals of confidence and profitability, thus affecting the stock price. This data also spans January 2020 to December, 2023 and was obtained from Financial Modeling Prep. Like with earnings surprise data, this dataset was not interpolated and dates in between dividend declarations and payment were filled with zeros.

Earnings Calls Sentiment

Earnings calls themselves offer deep insight into a company's financial health, strategic direction, and management outlook. But as outlined in the study by Hajek and Munk, sentiment analysis combined with traditional machine learning and price prediction frameworks can yield a more robust model. Rachel Tatman in her post "The Trouble with Sentiment Analysis" argues that sentiment analysis is ineffective due to the misalignment between measured text sentiment and the actual feelings or intentions behind the text. Furthermore, the word-list approach does not capture nuance, context, sarcasm, etc. FinBERT, a variant of BERT (Bidirectional Encoder Representations from Transformers), the sentiment analysis tool used by Hajek and Munk, and the one employed in this project, is fine-tuned for financial text, and is significantly more advanced than simpler models as it captures contextual nuances. It is designed to understand financial language specifically and potentially address Tatman's concerns about context and nuance. To give a few examples, finBERT labels the following text "Why microsoft shares are falling. microsoft corporation msft shares are trading lower by 3.69% to $256.80 tuesday morning" negative with a score of 0.97. "Arista stock is rising. profits topped estimates.. arista networks stock is gaining ground in late trading Monday" is labelled positive and scored 0.95. "Microsoft slows some hiring amid economic uncertainty." is labelled negative and scored 0.93. Given these examples, sentiment analysis score does not refer to emotions or the number of positive and negative words in a sentence. Rather, the label and score combination is reflective of what the text reveals about the profitability, operations, and general investment outlook for the company.

Earnings calls transcripts were downloaded from AlphaVantage using an API key, for the period March 2022 to March 2024 in JSON format. This was converted to a Pandas DataFrame where, along with additional cleaning and preprocessing, each sentence was processed by finBERT and given a label and score. Neutral sentences were removed, the label and scores were combined to a one-dimensional score, with scores of negative labels being multiplied by -1 and scores of positive labels remaining unchanged.

News and Social Media Sentiment

News data sentiment was included in the factors because it can impact investor perception and behavior, and thus influence stock price behavior. Social media sentiment can offer real-time insight into public perception of MSFT, with the collective reactions of investors to various company activities. This sentiment can function as a leading indicator for stock price movements. News data covers August 2022 to March 2024 and was obtained from Financial Modeling Prep and AlphaVantage. Social media sentiment contains content from StockTwits and Twitter from April 2023 to March 2024 and was obtained through Financial Modeling Prep. Data from Reddit for the pages r/WallStreetBets and r/Stocks was downloaded from Reddit using an instance. This was put into a DataFrame, with the title and content for each post joined into one string and then processed with finBERT to acquire the label and sentiment score.

Congress Trades

Congress members and legislators have access to non-public information and an understanding of upcoming policy changes that could affect companies and industries. Based on this information, their trades can offer signals about future market trends.

This chart shows a comparison of stock returns following the purchase of MSFT by Congress members against average returns over the same period without such a purchase. The blue line indicates the excess return, which is the difference between the return observed post-Congress members' purchase and the typical return for that time frame. Approximately 63 trading days after such a purchase, there is a marked increase where returns begin to exceed the norm. This outperformance reaches a peak around 170 trading days post-purchase, where the return is observed to be 12% higher compared to an average period of the same duration, suggesting a considerable influence of Congressional trading activity on the stock's performance. This dataset spans January 2020 to March 2024 and was obtained through the website Quiver Quantitative, which is a repository of alternative data for investing and trading. Only the columns for activity date, amount, and activity type were kept. Sale amounts were multiplied by -1 and purchase amounts were unchanged, reducing this to a single dimension.

All these features were then combined into one master DataFrame which was used for the models.

Random Forest Models

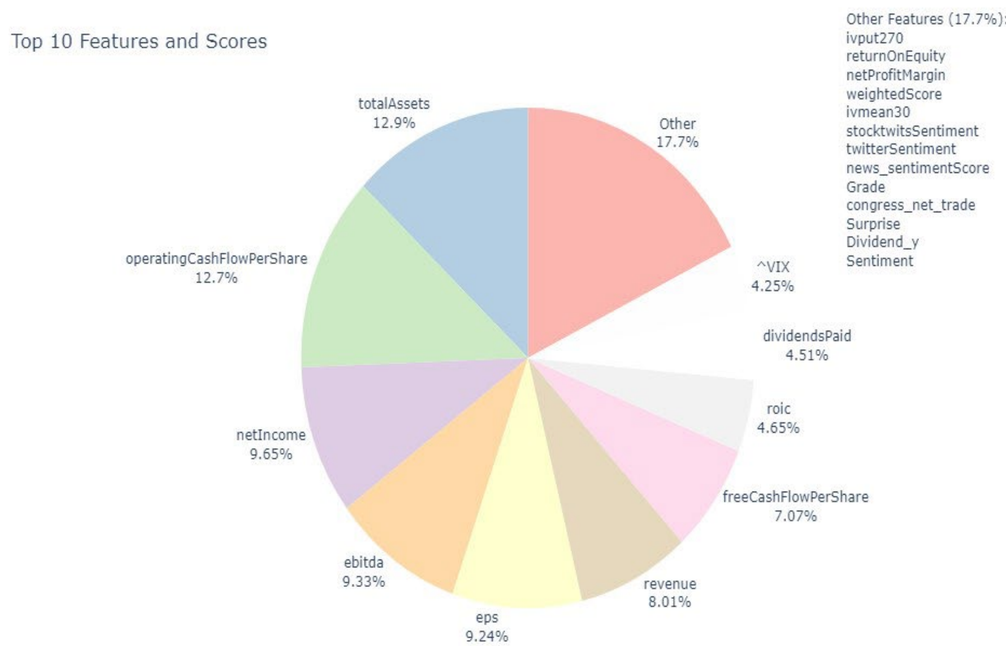
With the master DataFrame containing 283 features in place, the next task was to analyze which features are most associated with the stock price. The target variable was shifted back by 1 to assess the impact of the features on the following days' price. Correlation and regression models were not chosen because the relationships between the variables may not necessarily follow a linear or logarithmic relationship. Furthermore, many features are multicollinear, which is problematic for regression models. A random forest model was chosen for this part of the project due to its composition of many decision trees, which perform feature selection during the training process by choosing the most informative features at each split. Multicollinearity is less of an issue for this type of model since individual trees in the forest are trained on different subsets of features and data points. This reduces the variance and overfitting caused by multicollinearity seen in simpler models. Random forest models aggregate predictions from multiple trees, reducing the impact that any single, potentially collinear feature may have on the overall prediction.

The random forest models were separated into two sets of models: one set with all 283 features and one set with 23 selected features based on domain knowledge. For both sets, a simple random forest model was implemented to obtain the top 10 features. Then a new model was trained iteratively with these top features and evaluated based on RMSE scores. Next, a model using Grid Search CV was implemented that optimized the number of trees and minimum number of samples required to split a node. Additionally, a version with forward selection and a version with backward elimination were also implemented. Below are the results:

Random Forest Model	**Test RMSE**
All Features: Basic Model	35.335
Selected Features: Basic Model	43.164
All Features: Iteratively Trained on Top Features	31.994
Selected Features: Iteratively Trained on Top Features	45.869
All Features: Grid Search CV	35.236
Selected Features: Grid Search CV	**31.213**
All Features: Forward Selection	113.434
Selected Features: Forward Selection	39.57
All Features: Backward Elimination	34.147
Selected Features: Backward Elimination	32.46

5 fold cross-validation using TimeSeriesSplit() was used on the dataset. RMSE scores were derived from performance on test data and most of the models yielded RMSEs in a similar range. These results indicate the complexity of stock price movement: these models are capturing some patterns in the data but are likely missing other predictive signals. The Grid Search CV model using selected features yielded the lowest test RMSE. The top feature importances from this model are illustrated below:

The chart indicates that the top features are predominantly key financial indicators. Total Assets represents the size of the company while Net Income, EBITDA and EPS are profitability measures. Operating cash flow and free cash flow signify operational efficiency. VIX and implied volatility (ivput270) shows the sensitivity to market risk. Sentiment scores are ranked



lower in importance in comparison to the key financial metrics. The results of the model's feature importance are congruent with established financial theories and empirical evidence in the domain of finance and investment analysis.

The next section will examine predictive performance. These are models relevant to short-term strategies; hence the prediction time frame is a one-day period. Longer time-frame predictions require different types of models with appropriate assumptions and parameters.

Simple Baseline Model

Finance theory posits that stock prices evolve randomly but over time exhibit a general trend, or drift, which reflects the average return. This phenomenon is called a Random Walk with a Drift. Given a starting price and a drift term, this model will add a random error term and output an estimate of the following day's price. This is the baseline model against which to judge the performance of a predictive model.

LSTM Model Past Prices

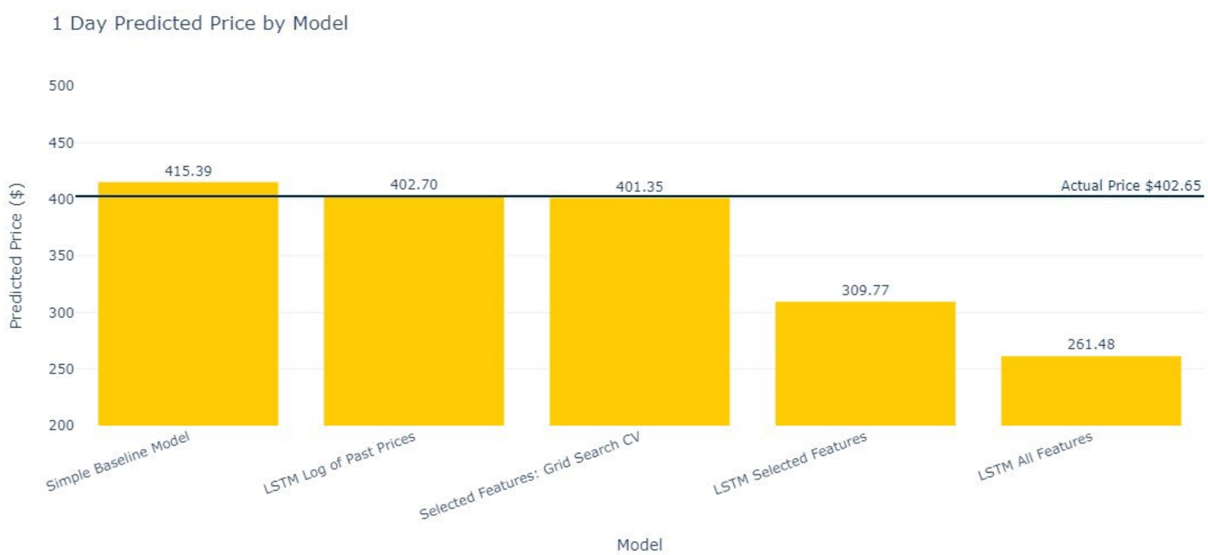
A characteristic of financial data is its long-term dependencies and patterns. This makes LSTMs well-suited to model this type of data as they can learn from the sequential order of past prices and identify trends and cycles indicative of future behavior. This LSTM model contains only the log of MSFT price data and focuses on the temporal aspect of price behavior. It assumes that historical prices contain recurring patterns. Utilizing 5-fold cross-validation ensures robust learning not tailored to a specific subset of data and enhanced generalizability to unseen data.

LSTM Models with All Features and Selected Features

The LSTM model with all features aims to leverage the breadth of data, examining complex, multi-faceted patterns that could influence the target variable. In contrast, the model with 23 selected features utilizes domain knowledge to streamline the analysis, reduce noise and computational complexity while keeping the most significant features for forecasting.

Predictive Models Results

Model	Test RMSE	Predicted Price	Validation RMSE
Simple Baseline Model	3.659	415.394	12.744
LSTM Log of Past Prices	0.017	402.701	0.051
LSTM All Features	76.523	261.475	141.175
LSTM Selected Features	35.884	309.768	92.882
Selected Features: Grid Search CV	31.213	401.352	1.298



The table and the chart above present the outcome of the various predictive models used to make a one-day forecast of MSFT stock price on unseen data. Except for the baseline model, the models were trained and evaluated based on test RMSE scores. The LSTM model with all features shows the poorest predictive performance, suggesting that including all features without a selection process introduces noise and hinders predictive ability. The LSTM model with selected features shows a slight improvement indicating that that feature selection improves predictive ability, but not enough to capture nuances required for an accurate prediction. The simple baseline mode, whose data was not split or cross-validated, shows a better test RMSE but an overestimation in the predicted price. The Random Forest model with the Grid Search CV and selected features has a relatively high test-RMSE but a low validation RMSE, suggesting that the model generalizes well to unseen data. The most accurate prediction is from the LSTM Past Prices model with an exceptionally low test and validation RMSE. The predicted price of $402.70 is close to the actual price of $402.65, highlighting its effectiveness in capturing the stock’s price movement.

Discussion and Next Steps

The random forest models show that financial metrics like Total Assets, Cash Flow, Net Income and EBITDA are the most important in terms of their association with the following days stock price. This aligns with financial theory, which states that these indicators contain valuable information about a company’s performance and future prospects. Additionally, feature importance scores of implied volatility, sentiment factors like VIX and social sentiment reinforces the complex nature of stock price behavior, where investor sentiment and perception play vital roles. Amongst the predictive models, the LSTM model with the log of past prices produced exceptional results, demonstrating its effectiveness in capturing temporal dependencies and trends within the time-series. In contrast, models with more features displayed higher RMSE possibly because of the challenge of effectively incorporating different types of data and the presence of multicollinearity in the features.

It must be noted that due to ethical concerns, this study is to be used for educational purposes only and may not be used for investment advice. Individuals without the proper training and background knowledge of investment management may fail to consider the risks of an asset without being fully aware of its investment characteristics and impact on the portfolio as a whole.

As a next step in this line of research, this project could be enhanced by exploring other types of models like deep neural networks or ensemble methods. Different feature selection methods could also be investigated. Furthermore, dimensionality reduction techniques such as Principal Component Analysis could be employed to create composite indicators that encapsulate underlying trends without redundancy. After further model refinement, deploying this to a real-time trading environment could be considered, as well as assessing performance across different market conditions as new data becomes available. This study sheds light on the multi-faceted dynamics of financial markets and paves the way for further innovations in predictive modeling in the field of finance and investments.

Statement of Work

This project was completed by Sababa Ahmad, who was responsible for the research, including data collection, model development, analysis, and writing of the report. Sababa's comprehensive efforts encompassed the conceptualization of the study, implementation of various machine learning models, analysis of results, and formulation of conclusions and future work directions.

Bibliography

Hajek, Petr ; Munk, Michal. “Speech Emotion Recognition and Text Sentiment Analysis for Financial Distress Prediction.” *Neural Computing & Applications*, vol. 35, no. 29, London: Springer London, 2023, pp. 21463–77, doi:10.1007/s00521-023-08470-8.

Lang, Erik. "Reddit API Lab - Create." SIADS682 Social Media Analytics, submitted 12 March, 2024, University of Michigan. Unpublished course assignment.

Myers, Greg. "Assignment 4 - Tree-based classification & Synthesis Project." SIADS542 Supervised Learning, submitted 17 September, 2022, University of Michigan. Unpublished course assignment.

Tatman, Rachel. "The Trouble with Sentiment Analysis." *Making Noise and Hearing Things*, 19 Apr. 2022, https://makingnoiseandhearingthings.com/2022/04/19/the-trouble-with-sentiment-analysis/.

Vrontos, Spyridon D.; Galakis, John; Vrontos, Ioannis D. “Implied Volatility Directional Forecasting: A Machine Learning Approach.” *Quantitative Finance*, vol. 21, no. 10, Bristol: Routledge, 2021, pp. 1687–706, doi:10.1080/14697688.2021.1905869.

"WallStreetBets Impact on the Market." *Quiver Quantitative*, Quiver Quantitative, January 20, 2021, https://www.quiverquant.com/blog/?p=wallstreetbets.

Appendix

Below is a table of results for the LSTM Past Prices model tested on other stocks. Judging by the validation RMSE, it performed better on some stocks than others.

	Test RMSE	**1 Day Actual Price**	**1 Day Prediction**	**Validation RMSE**
AAPL	0.019	185.4	192.28	6.885
EL	0.019	144.8	145.60	0.800
KR	0.017	45.98	45.42	0.561
RL	0.024	146.03	144.25	1.778
HAS	0.020	49.21	50.36	1.147
APA	0.032	35.82	35.59	0.231
CZR	0.039	47.78	46.88	0.895
MKTX	0.026	285.51	291.90	6.386
BR	0.014	199.89	205.75	5.861