Algorithmic Stock Price Forecasting

Samuel Hughes [1]
Muhammad Salman Asif [1]

Abstract

In this project, we implemented and compared multiple prediction algorithms to forecast stock prices 60 business days into the future. The models employed include gradient boosting frameworks such as XGBoost and LightGBM, dense neural networks, and long short-term memory (LSTM) networks. To enhance the predictive accuracy, the models were conditioned on additional macroeconomic inputs including congressional trading, providing insight into the influence of external financial decision-making on stock price trends. Our analysis found the LightGBM model as the best predictive option, offering a mean absolute error of 12.78 between the actual and predicted stock price, along with a 6.82% improvement over the baseline. Our evaluation highlights the strengths and limitations of each model, with a focus on their robustness, interpretability, and performance in leveraging auxiliary datasets for stock market prediction.

1. Introduction

Predicting stock market trends is a challenging yet essential task in the realm of financial decision-making. The dynamic nature of stock prices, influenced by a myriad of factors including economic indicators, market sentiment, and political developments, necessitates the use of advanced predictive techniques. Time series analysis has long been a cornerstone in this domain, offering tools to identify patterns and make forecasts based on historical data. However, traditional models often fall short when faced with the complex, nonlinear dependencies inherent in financial data.

This project aims to bridge this gap by leveraging a diverse suite of modern prediction algorithms. These include gradient boosting frameworks such as XGBoost and LightGBM, which are renowned for their ability to handle high-dimensional data, autoregressive models that exploit temporal dependencies, and deep learning architectures such as long short-term memory (LSTM) networks and dense neural networks, which are adept at capturing nonlinear patterns.

To ensure a robust evaluation, we employed multiple performance metrics, including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), to assess the predictive accuracy of each model. These metrics allow for a comprehensive comparison, highlighting the strengths and limitations of each approach.

Furthermore, we incorporate multiple sources of macroeconomic data as auxiliary features to enrich the predictive capabilities of our models, most notably congressional investment data. Congressional trading activity has drawn significant attention as a potential indicator of market trends, and its integration into our models seeks to uncover its predictive value. By conditioning our models on the external data, we aim to assess its influence on stock price movements and improve the robustness of our forecasts.

This paper presents a comprehensive analysis of the implemented models, their evaluation using diverse metrics, and the impact of incorporating external features. Through this work, we aim to contribute to the growing body of research exploring advanced methodologies for financial forecasting, offering insights into the interplay between market dynamics and external socio-economic factors.

2. Financial Data

The predictive models in this project were trained on a comprehensive dataset that integrates a range of financial and external features. These features were designed to capture the temporal, statistical, and macroeconomic dynamics influencing stock market trends. The data is categorized into three main types of features: traditional financial metrics, rolling window statistics, and external factors.

2.1. Problem Statement and Goals

The primary aim of this study is to leverage state-of-the-art machine learning techniques to forecast stock price movements and understand the key drivers of these changes. Our objectives are:

- Use advanced machine learning models to predict future stock prices.
- Identify critical factors that influence stock price

changes, both financial and external.

- Visualize and validate the accuracy of stock price predictions.

The motivation for this work stems from the potential for algorithmically validated predictions, which hold significant real-world applications in financial decision-making and investment strategies. The ability to provide accurate predictions offers not only financial advantages but also a deeper understanding of market dynamics.

2.2. Data Overview

The data used in this project is derived from a comprehensive set of financial and external metrics aimed at predicting stock price movements.

Target: The target variable for the models is to predict whether the price of a given stock will increase or decrease 60 business days into the future.

Scope: This study focuses on the Dow 30 Index, which comprises 30 high-profile companies in the United States, including Amazon, Apple, and Tesla. The dataset covers the period from October 2015 to September 2024, with the models trained and evaluated retroactively over this timeline.

2.3. Traditional Features

Traditional financial metrics serve as a foundational component of the dataset. These include:

- **Adjusted Closing Price:** A standardized measure of stock price after accounting for dividends, splits, and other corporate actions.
- **Trading Volume:** The total number of shares traded within a given period, reflecting market activity and liquidity.
- **Recent Stock Price Gains and Losses:** Indicators of short-term market performance, capturing momentum and volatility.
- **Aggregations of Stock Price in Past 60 Business Days:** Summary statistics such as mean and deviation of adjusted closing prices over a 60-day window to understand longer-term trends.

2.4. Rolling Window Features

Rolling window features were calculated to capture temporal trends and smooth fluctuations in stock prices. These include:

- **Mean Adjusted Close and Volume in 10-Day Windows:** Averages of adjusted closing prices and trading volumes over a 10-day moving window, providing localized trends.
- **Difference in Price of Adjacent Windows:** The change in the mean adjusted close and volume between consecutive rolling windows, highlighting shifts in stock price behavior.

2.5. External Factors

To complement the traditional financial metrics, external factors were incorporated to account for macroeconomic and political influences:

- **Macroeconomic Indicators:** Key economic metrics, including the unemployment rate, gross domestic product (GDP), and the 10-year treasury yield rate, which provide a broad economic context for stock price movements.
- **Congressional Trading Activity:** Congressional stock purchases and sales reported within the past 90 calendar days. This data offers insights into potential market-altering activities based on privileged information or broader economic expectations.

These diverse financial and external features form the foundation of the predictive models, enabling them to capture the multifaceted dynamics of the stock market and improve forecasting accuracy.

3. Methods

This section describes the models implemented for stock price prediction and the evaluation methodology used to measure their performance. The selected models include gradient boosting techniques and neural networks, which leverage different aspects of the dataset's structure and temporal characteristics.

3.1. Models

3.1.1. Gradient Boosting Models

Gradient boosting is an ensemble learning technique that builds predictive models incrementally by optimizing a loss function. Two gradient boosting models were implemented in this study:

XGBoost (Extreme Gradient Boosting): XGBoost improves upon traditional gradient boosting by incorporating advanced regularization (L1 and L2 penalties) and optimizing memory usage. The prediction at time t is modeled as a

sum of M decision trees:

$$\hat{y}_t = \sum_{m=1}^{M} f_m(x_t), \quad f_m \in \mathcal{F},$$

where $\mathcal{F}$ is the space of regression trees. The objective function includes a loss function $\mathcal{L}$ and a regularization term Ω for model complexity:

$$\text{Obj} = \sum_{t=1}^{N} \mathcal{L}(y_t, \hat{y}_t) + \sum_{m=1}^{M} \Omega(f_m).$$

XGBoost is known for its scalability and ability to handle missing values effectively (1).

LightGBM: LightGBM is another gradient boosting framework optimized for speed and efficiency. It uses a histogram-based learning algorithm to reduce computational overhead and splits trees leaf-wise instead of level-wise, which minimizes loss more effectively at each iteration. The leaf-wise approach is particularly beneficial for large datasets with high-dimensional features (3).

3.1.2. Neural Networks

Neural networks are powerful tools for learning complex patterns and nonlinear relationships in data. Two types of neural networks were implemented in this study:

Dense Neural Network (DNN): The DNN is a fully connected feed-forward network where each neuron in one layer is connected to all neurons in the subsequent layer. Given input features x, the output of a dense layer is:

$$h_i = \sigma(W_i x + b_i),$$

where W_i and b_i are the weights and biases, and σ is the activation function. The network is trained using backpropagation to minimize the loss between predicted and actual stock prices (5).

Long Short-Term Memory (LSTM): LSTMs are specialized recurrent neural networks designed to model sequential data by addressing the vanishing gradient problem. The LSTM cell maintains a memory vector c_t and a hidden state h_t at each time step, updated as follows:

$$\begin{aligned}
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f), \\
i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i), \\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o), \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c), \\
h_t &= o_t \odot \tanh(c_t).
\end{aligned}$$

where f_t, i_t, and o_t are the forget, input, and output gates, respectively. LSTMs are particularly effective for capturing long-term dependencies in time-series data (2).

3.2. Evaluation Methods

The models were evaluated using a rigorous process that involved hyperparameter tuning, validation, and performance assessment on unseen test data.

3.2.1. Hyperparameter Tuning

Optimization techniques were used to tune the hyperparameters of each model. For XGBoost and LightGBM, parameters such as learning rate, maximum depth, and the number of estimators were optimized. For neural networks, the number of layers, hidden units, regularization, and learning rates were adjusted.

3.2.2. Training and Validation

The dataset was divided into three subsets using a cross-sectional splitting strategy:

- **Training Set:** Used to train the models on 14 stocks from the Dow 30 Index.
- **Validation Set:** Used to fine-tune hyperparameters and prevent overfitting on 8 stocks separate from training.
- **Test Set:** Evaluated on 8 stocks unseen in the training and validation phases to ensure generalization.

3.2.3. Evaluation Metrics

To quantitatively assess model performance, several evaluation metrics were used:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual stock prices:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|.$$

- **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily, providing a sense of prediction reliability:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}.$$

- **Comparison Against Baseline:** Model predictions were benchmarked against a baseline naive forecast, which assumes no price change.

This combination of advanced machine learning models and robust evaluation metrics ensures that the predictions are accurate and interpretable, providing insights into the factors influencing stock price movements.

4. Performance

4.1. Performance Comparison

The table below compares the performance of various models based on Mean Absolute Error (MAE) and improvement over the baseline:

Model	MAE	RMSE	Improvement
LightGBM	12.78	21.12	6.82%
XGBoost	13.14	22.03	4.21%
Dense Neural Network	13.21	22.20	3.70%
LSTM	13.40	22.10	2.32%

4.2. Visualization of Results

The scatter plot below visualizes the actual vs. predicted stock prices for the best-performing model:

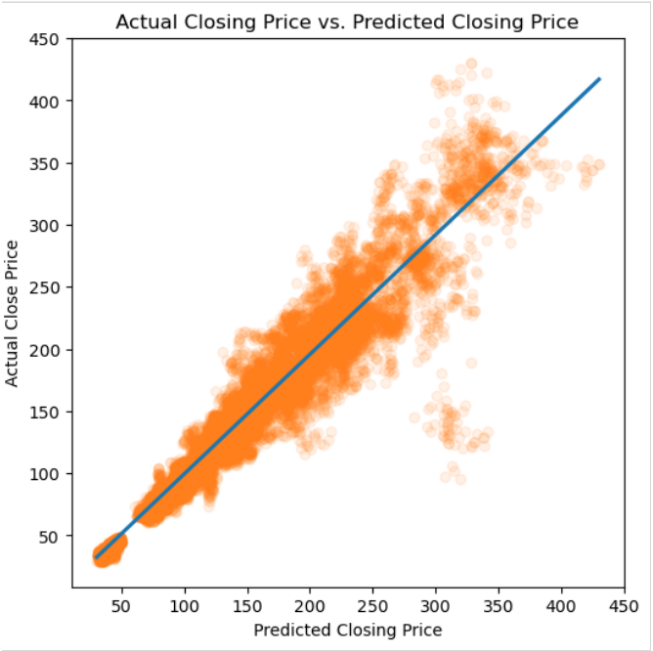

Figure 1. Actual Closing Price vs. Predicted Closing Price

5. Feature Importance

5.1. Insights

From the analysis, the following insights are derived:

- **Unemployment Rate:** A strong indicator of stock price movement.
- **Recent Price Decrease:** May signal a rebound, indicating that a market correction is imminent.
- **Congressional Influence:** Shows mixed effects, not consistently significant as a predictor.

Figure 2. SHAP Plot of Feature Importance (4)

6. Probabilistic Forecast

Probabilistic forecasts were generated to provide a range of possible outcomes for stock prices. These forecasts give investors a confidence interval around predicted values.

6.1. Visualization of Forecast Ranges

The fan chart below illustrates the range of predicted stock prices with confidence intervals:

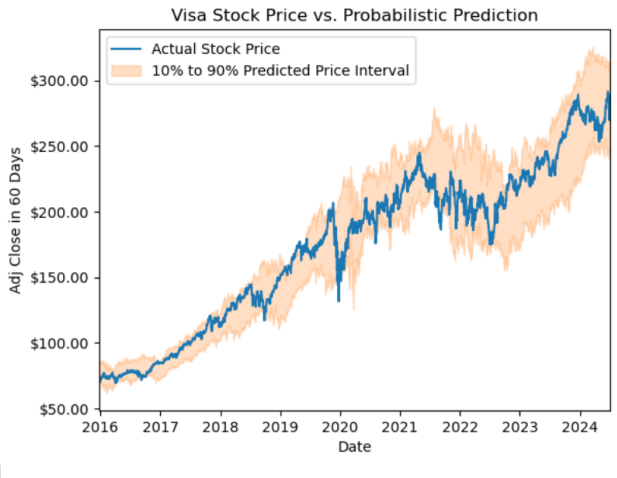

Figure 3. Probabilistic Forecast with Confidence Intervals

6.2. Benefits of Probabilistic Forecasts

- **Risk Assessment:** Quantifies uncertainty in predictions for better decision-making.
- **Tail Risk Awareness:** Highlights best-case and worst-case scenarios using upper and lower bounds.
- **Enhanced Robustness:** Facilitates incorporation of unexpected external shocks.

7. Discussion

This project has made significant contributions to the field of stock market prediction by applying advanced machine learning techniques to forecast stock price movements. Through the integration of a variety of data sources and state-of-the-art supervised learning models, we have achieved noteworthy success in both model performance and interpretability. This section discusses the key achievements, the novelty of our approach, and outlines promising directions for future work.

7.1. Key Achievements

Our research has resulted in several important outcomes that have advanced our understanding of stock market prediction and demonstrated the potential of advanced machine learning techniques:

- **Outperformance of Benchmark:** One of the most significant achievements of this work is the 6.8% improvement in model performance compared to traditional benchmark models. This outperformance demonstrates the effectiveness of advanced predictive algorithms, specifically the LightGBM model, in capturing complex, non-linear patterns in stock price data. By outperforming the benchmark, our model proves its ability to deliver more accurate predictions, offering a powerful tool for financial forecasting.
- **Identification of Driving Factors:** A key insight from this research is the identification of the primary factors that drive stock market movements. Through feature analysis, we were able to pinpoint specific variables that significantly influence stock price fluctuations. This step not only adds transparency to the model but also enhances its interpretability, allowing stakeholders to understand the logic behind the predictions. It opens the door to more informed decision-making in stock market investments.
- **Probabilistic Forecasting:** Another major contribution of this work is the incorporation of probabilistic forecasting. Unlike traditional models that provide single-point predictions, our model delivers a range of possible outcomes, complete with uncertainty estimates. This probabilistic approach offers a more robust view of the future, providing investors with a better understanding of potential risks and rewards. By quantifying uncertainty, we empower decision-makers to make more informed and less risky investment choices.

7.2. Novelty of the Approach

This project stands out due to its innovative combination of several elements that have not been widely explored in the context of stock market prediction:

- **Use of Congressional Data:** A distinguishing feature of our approach is the integration of congressional data as a feature in stock price prediction. While most stock market prediction models focus on technical indicators and historical price data, we introduced a novel element by including political data. This could potentially capture the impact of political events, legislative decisions, and government actions on market movements. Our work suggests that incorporating such non-traditional data sources can provide additional predictive power, highlighting the broader scope of factors influencing market behavior.
- **State-of-the-Art Supervised Learning Models:** The focus on utilizing advanced supervised learning models places our work at the cutting edge of stock market prediction. The LightGBM algorithm proved to be an excellent choice for capturing the intricate dynamics of stock price movements. Our implementation demonstrated how sophisticated machine learning models could outperform more conventional techniques, offering new insights and more precise predictions.

7.3. Future Work and Directions

Although our work has demonstrated promising results, there remain several avenues for future exploration that could further enhance the model's performance and broaden its scope:

- **Incorporation of Financial and Company News Sentiment:** An exciting opportunity for future development is the incorporation of sentiment analysis from financial news and company reports. Daily news articles, investor sentiment, and market commentary often drive stock price movements, and by extracting sentiment from such sources, we can better capture market reactions to breaking news. By integrating these sentiment scores into our model, we could enhance the model's responsiveness to real-time events and improve the accuracy of predictions. This will enable the model to adapt dynamically to the shifting sentiment in the market, increasing the robustness of the forecasts.

- **Expanding Beyond Dow 30 Stocks:** While our analysis focused on the Dow 30 stocks, there is a significant opportunity to expand the model to cover a broader range of companies. This could include stocks from other indices, such as the S&P 500 or emerging markets. A more diverse dataset would help test the generalizability of the model and allow it to make predictions across a wider array of sectors, potentially improving its accuracy and applicability in real-world scenarios.

- **Incorporating Macroeconomic and Global Factors:** Beyond financial and political data, the inclusion of macroeconomic indicators such as GDP growth, inflation rates, interest rates, and global economic conditions could offer a more comprehensive understanding of market movements. These factors often play a crucial role in influencing stock prices, and their inclusion could lead to even more precise predictions.

- **Leveraging Large Language Models (LLMs):** As a future project, we plan to explore the use of Large Language Models (LLMs) to embed a wider range of context into the prediction process. LLMs, such as GPT-based models, can process vast amounts of unstructured data, such as daily news articles, social media feeds, and corporate earnings reports. By leveraging LLMs, we can incorporate a broader range of context—such as global events, economic trends, and investor sentiment—into our stock market forecasts. This could make our predictions more robust, precise, and accurate by accounting for the complexities of the broader news landscape and real-time developments that influence stock prices.

- **Real-Time Deployment and Market Integration:** An important next step is the real-time deployment of our model for live stock market prediction. This would involve integrating our model with financial data APIs to deliver real-time forecasts and allow for continuous learning. With the ability to adapt in real-time to market conditions, the model could become a valuable tool for investment firms and financial analysts.

In conclusion, this work represents an important step forward in stock market prediction, showcasing the potential of deep learning models and the integration of unconventional data sources, such as congressional data, in forecasting stock movements. Our achievements in outperforming traditional benchmarks, identifying key driving factors, and providing probabilistic forecasts set the stage for future advancements. With the integration of sentiment analysis, broader datasets, and the use of cutting-edge models like LLMs, we are confident that future iterations of this model will yield even more precise, robust, and context-aware stock market forecasts.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[2] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[4] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[5] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.