Predicting Short-Term Stock Price Actions Using Artificial Intelligence

Sri Rithish Palani

Abstract

Predicting short-term price actions of stocks is a significant aspect of stock market investing. Numerous methods exist, but many lack efficiency and accuracy. With recent advancements in Artificial Intelligence (AI), research has explored its potential in predicting stock markets. However, limited research focuses specifically on short-term price action prediction. This paper discusses the methodology and findings of our AI models in predicting short-term price actions of stocks.We analyzed key metrics from a sample of 80 stocks using various AI models to assess their accuracy in predicting short-term stock prices. The results suggest that AI models can be effective in forecasting short-term stock movements, but their accuracy is highly dependent on the preprocessing techniques and the features used. Notably, the Orthogonal Matching Pursuit CV model achieved the highest accuracy, reaching up to 80%, particularly when previous quarter prices were included in the analysis.

Introduction

The stock market is a cornerstone of the global economy, offering companies a way to raise capital and investors the opportunity to buy and sell ownership stakes. Stock prices are driven by a complex mix of factors, including economic indicators, company performance, market sentiment, and geopolitical events. Investors are always on the lookout for methods to predict stock price movements, aiming to maximize returns and manage risks. Traditional approaches like technical analysis and fundamental analysis have been widely used. Technical analysis focuses on studying historical price charts and utilizing indicators like moving averages and the relative strength index (RSI) to forecast future price trends. However, these methods can often struggle to keep pace with the fast-evolving financial markets. While this method can be effective in stable market conditions, it may not account for sudden market shifts or anomalies. Fundamental analysis, by contrast, involves assessing a company's financial health through a detailed examination of its balance sheet, income statement, and cash flow statement. This approach aims to uncover the intrinsic value of a stock, providing a solid foundation for investment decisions. However, it can be time-consuming and may not quickly adapt to market changes driven by investor sentiment or external factors. This project aimed to leverage AI to predict short-term stock price movements, enhancing market dynamics understanding and generating monetary gains. Although stock prediction has been extensively explored, AI-based research has emerged only recently, focusing primarily on long-term predictions. The lack of research on AI for short-term predictions motivated our study. We analyzed current stock metrics and developed a novel metric, momentum, which calculates the average change of a stock metric over the past year. These metrics were incorporated into various AI models to identify the most effective approach.

The goal of predicting short-term stock price movements is not just a theoretical exercise but a practical necessity for investors seeking to maximize their returns. Traditional methods like technical analysis and fundamental analysis have their strengths, but they often struggle to keep up with the fast-paced changes in financial markets. By incorporating AI into these processes, we're introducing a new level of precision and adaptability in predicting stock prices. Our approach is rooted in the belief that AI can identify patterns and correlations that are too complex for human analysts to detect.

Methods

To achieve our objective, we compiled a dataset of 80 stocks, carefully selecting them to ensure a representative sample across various industries and sectors. This broad spectrum of data allows us to capture a wide range of market behaviors and company profiles, making our findings more applicable to different market conditions.

Our dataset includes the following columns:

- **Symbol**: The ticker symbol of the stock, which is a unique series of letters assigned to a security for trading purposes.
- **Description**: A brief description of the company, providing context for its market activities.
- **Industry**: The industry in which the company operates, helping to compare performance against industry peers.
- **Sector**: The broader sector classification of the company, such as Technology, Healthcare, etc.
- **Current Price**: The latest trading price of the stock, providing a snapshot of its current market value.
- **Price Momentum**: The rate of change in the stock price over a specified period, indicating the strength of the price trend.
- **Beta**: A measure of the stock's volatility relative to the overall market. A beta greater than 1 suggests higher volatility, while less than 1 suggests lower volatility.
- **Market Cap**: The total market value of the company's outstanding shares, used to classify the size of the company.
- **P/E (Price to Earnings Ratio)**: A valuation ratio comparing the company's current share price to its per-share earnings, used to assess valuation.

- **P/E Momentum**: The rate of change in the P/E ratio over a specified period, providing insight into valuation trends.
- **P/B (Price to Book Ratio)**: A ratio comparing the company's market value to its book value, indicating whether it's over or undervalued.
- **P/S (Price to Sales Ratio)**: A ratio comparing the company's market value to its total sales, offering insight into valuation.
- **P/S Momentum**: Tracks changes in the P/S ratio over time, reflecting shifts in market valuation.
- **Net Income**: The company's total profit, indicating overall profitability.
- **Net Income Momentum**: Tracks changes in net income over time, showing profit trends.
- **Total Revenue Annual**: The total revenue generated by the company in a year, reflecting its scale of operations.
- **FCF (Free Cash Flow)**: The cash generated after capital expenditures, indicating the company's financial health and ability to generate cash.
- **Dividend Yield**: The dividend expressed as a percentage of the stock price, reflecting the return on investment through dividends.
- **Total Current Assets Quarterly**: Total assets expected to be converted into cash within a quarter, indicating liquidity.
- **Total Current Liabilities Quarterly**: Total liabilities due within a quarter, showing short-term financial obligations.
- **Total Debt Quarterly**: The total amount of debt the company holds, indicating financial leverage.
- **EBITDA**: A measure of overall financial performance, indicating earnings before interest, taxes, depreciation, and amortization.

These metrics were chosen because they provide a comprehensive understanding of each company, covering aspects of profitability, valuation, market behavior, and financial health.

We organized this data into a spreadsheet, ensuring it was clean and free of inconsistencies. Using the Lazy Regressor extension, we tested various AI algorithms for accuracy. The Lazy Regressor was chosen for its ability to quickly and efficiently compare multiple machine learning models with minimal manual coding. This tool automates the process of training and evaluating models, allowing us to focus on analyzing results rather than on model implementation. Initial tests identified the Gradient Boosted Regressor as the best model. However, to improve accuracy, we added a previous quarter price column, providing a historical baseline for the models. Additionally, we refined the dataset by scaling large numbers, such as market capitalization, to ensure uniformity and improve model performance.

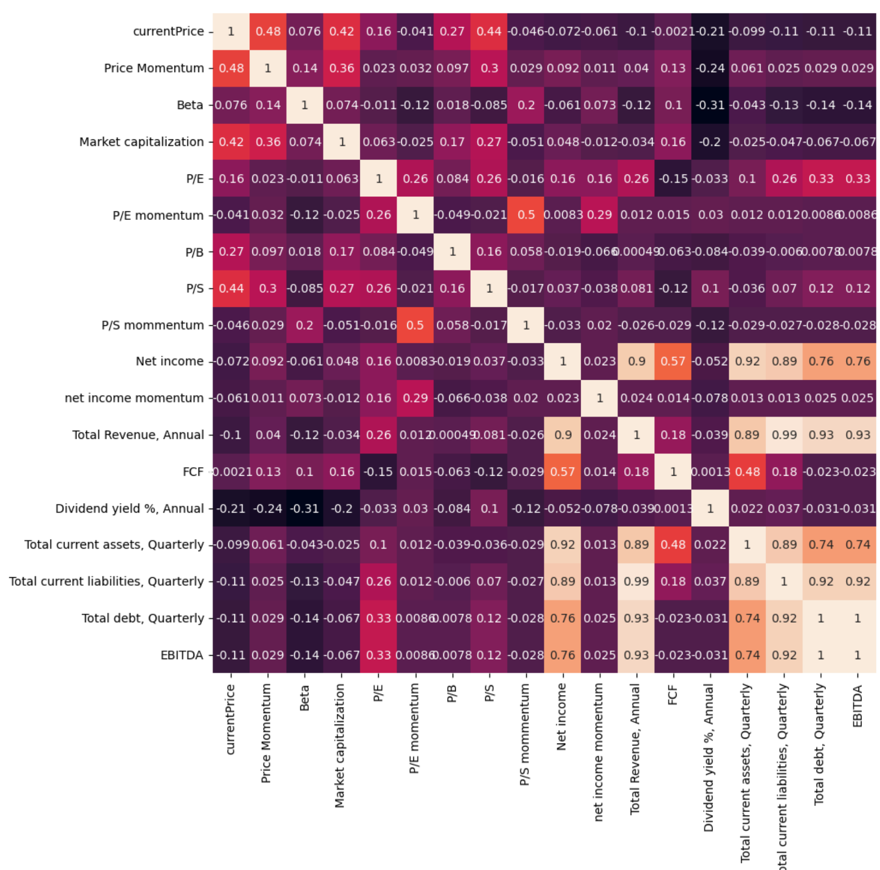
Results/Analysis

Firstly, we visualized our initial dataset using various graphs and charts, including correlation matrices, box and whisker plots, bar charts, and heat maps. These visualizations helped us gauge the correlation between different features in our data.

Initial Dataset Visualization:

- **Correlation Matrix**: This heat map shows the correlations between different features.

Figure 1



The correlation matrix revealed several significant relationships between features. For instance, market capitalization and total revenue showed a strong positive correlation, suggesting that larger companies tend to have higher revenues. Conversely, features like beta and dividend yield showed little to no correlation, indicating that these metrics may not be directly related in our dataset.

After understanding the data correlations, we created our features and target variables, preparing them for input into AI models. We split our data into training and testing sets to ensure efficient model performance.

Next, we used the Lazy Regressor extension to test multiple AI algorithms and determine the best model. Initially, the Gradient Boosted Regressor showed the best performance, so we subsequently applied it to our dataset, as well as other models like Logistic Regression, SVC, and XGBClassifier.

Initial Model Performance:

- **Gradient Boosted Regressor**: Mean Squared Error (MSE) ~ 6000
- **Logistic Regression, SVC, XGBClassifier**: MSE ~ 10,000

Gradient Boost regressor results

Symbol	currentPrice	Predicted Price
ASMLF	1039	340.485934
TMUS	180.01	228.862603
SCL	84.54	15.659611
NVO	143.24	436.503046
BE	14.89	77.654146
AMBP	3.91	26.026309
INTU	573.68	406.547432
COVTY	25.61	15.133065
SKFRY	21.77	12.482684
CSCO	45.93	207.693672
OGFGF	6.4	7.2844
NKE	97	76.091362

TREX	80.07	144.352632
GE	162.1	219.933662
XPRO	20.37	18.033478
NVDA	1207.5	854.086074
ERNXY	19.59	9.599931

The initial models demonstrated limited predictive accuracy, with the mean squared error (MSE) values being in the range of 10,000. To enhance accuracy, we incorporated a previous quarter price column into our dataset, providing a historical reference point crucial for forecasting future price movements. We then re-visualized this updated data and repeated the analysis process. This time, the Lazy Regressor tool identified OrthogonalMatchingPursuitCV as the most effective model, indicating improved accuracy with a lower MSE and higher correlation between predicted and actual stock prices.

Updated Dataset Visualization:

- **Correlation Matrix**: Shows updated feature correlations.

Figure 2

The updated correlation matrix highlighted some shifts in relationships between features, likely due to the inclusion of the previous quarter price. This historical data added another layer of complexity and depth to the dataset, enhancing the models' ability to detect trends.

Updated Model Performance:

- **OrthogonalMatchingPursuitCV**: MSE ~ 6000, Accuracy ~ 80%
- **Other models**: MSE ~ 10,000

OrthogonalMatchingPursuitCV results:

Symbol	Current Price	Predicted Price
ASMLF	1039	937.769188
TMUS	180.01	177.642815
SCL	84.54	144.793037
NVO	143.24	84.005725
BE	14.89	89.921674
AMBP	3.91	-11.905522
INTU	573.68	639.34006
COVTY	25.61	34.180276
SKFRY	21.77	9.564862
CSCO	45.93	61.717932
OGFGF	6.4	-5.304051
NKE	97	136.617333
TREX	80.07	78.779936
GE	162.1	97.320397
XPRO	20.37	26.683213
NVDA	1207.5	596.815002
ERNXY	19.59	18.567147

The addition of the previous quarter price column significantly improved our model's accuracy. Finally, we attempted to improve our dataset by scaling down columns with large numbers, such as market capitalization. We repeated the analysis, and Lazy Regressor suggested Kneighbors as the best model. However, this resulted in a decline in performance.

Scaled Dataset Visualization:

- **Correlation Matrix**: Shows scaled feature correlations.

Figure 3

Scaling the dataset altered the relationships between some features, as shown in the updated correlation matrix. For example, before scaling, market capitalization had a relatively strong positive correlation with total revenue. However, after scaling, this correlation weakened, likely due to the disproportionate scaling of market capitalization compared to other features. Similarly, the correlation between EBITDA and total debt increased post-scaling, indicating that the scaling process impacted the relative importance and relationships of these features within the dataset.

Scaled Dataset Model Performance:

- **Kneighbors**: MSE ~ 11,000, Accuracy ~ 60%
- **Other models**: MSE ~ 10,000

Kneighbors Results:

Symbol	currentPrice	Predicted Price
ASMLF	1039	552.1333333
TMUS	180.01	147.2311111
SCL	84.54	18.8344444
NVO	143.24	153.3844444
BE	14.89	23.30888889
AMBP	3.91	11.86888889
INTU	573.68	352.4177778
COVTY	25.61	25.81
SKFRY	21.77	19.82444444
CSCO	45.93	88.19888889
OGFGF	6.4	13.61333333
NKE	97	118.1944444
TREX	80.07	43.53222222
GE	162.1	98.23444444
XPRO	20.37	19.26555556
NVDA	1207.5	359.2066667
ERNXY	19.59	15.28222222

The decline in performance with the scaled dataset can be attributed to only partially scaling down some columns, suggesting that all columns should have been scaled to improve accuracy fully. This finding underscores the importance of consistent preprocessing steps across all features.

Next Steps

Our research on leveraging AI to predict short-term stock price movements has yielded insightful results, yet there remains ample room for improvement and further exploration. To enhance the performance of our models, we will undertake a more systematic approach to data normalization and standardization. By refining data preprocessing techniques, we aim to ensure that all features contribute optimally to the model's performance, addressing the mixed results observed when scaling specific columns.

Data Preprocessing and Feature Engineering

One of the primary areas for improvement is data preprocessing. Inconsistent scaling and normalization can lead to suboptimal model performance, as observed in our study. Future work will involve a more rigorous and systematic approach to data preprocessing. This includes ensuring that all numerical features are standardized and scaled appropriately. Additionally, we will explore advanced preprocessing techniques such as feature selection and dimensionality reduction to identify and retain only the most relevant features.Previous studies, including those by Han et al. (2011) and Kuhn and Johnson (2013), have underscored the critical importance of consistent data preprocessing in achieving accurate model predictions.

In addition to refining preprocessing techniques, we're planning to explore additional features that could further enhance our predictive power. This includes the incorporation of technical indicators such as moving averages, the Relative Strength Index (RSI), and the Moving Average Convergence Divergence (MACD), which are widely used in financial analysis. Moving averages help smooth out price data to reveal trends over specific periods, RSI measures the speed and change of price movements to identify overbought or oversold conditions, and MACD, a trend-following indicator, illustrates the relationship between two moving averages of a stock's price. Furthermore, sentiment analysis from news articles and social media will be integrated to capture market sentiment and investor behavior. Macroeconomic factors, such as interest rates, inflation, and Gross Domestic Product (GDP) growth, will also be considered to provide a more comprehensive picture of market dynamics.

Ensemble Methods and Advanced Models

To further improve model accuracy, we will experiment with ensemble methods such as stacking, bagging, and boosting, which combine the strengths of multiple AI models. Ensemble methods have been shown to enhance predictive performance by reducing overfitting and improving generalization. Techniques like Random Forests, Gradient Boosting Machines, and Voting Classifiers will be explored. We will also investigate the use of more advanced models, such as deep learning techniques including LSTM (Long Short-Term Memory) networks, which are well-suited for time-series prediction.

Incorporating Real-Time Data

Incorporating real-time data is another crucial step in our future work. Developing a pipeline that allows our models to receive and process real-time stock data will potentially improve prediction accuracy by providing more timely predictions. This involves setting up real-time data feeds and ensuring that our models can process and update predictions continuously. Alongside real-time data integration, we will focus on hyperparameter tuning and cross-validation to optimize model performance, utilizing techniques such as grid search and randomized search. Hyperparameter optimization is essential for fine-tuning model parameters to achieve the best possible performance.

Expanding the Dataset

We also plan to expand our dataset by increasing the sample size of stocks and incorporating data from different market conditions and time periods. This expansion will help generalize our findings and improve the robustness of our models. By including a diverse set of stocks from various sectors and market environments, we can ensure that our models are not biased towards specific market conditions. This will involve gathering historical data from multiple market cycles, including bull and bear markets, to train and test our models under different scenarios.

Additionally, we will consider a broader range of evaluation metrics, such as mean absolute error (MAE), R-squared, and directional accuracy, to gain a more comprehensive understanding of model performance. These metrics will provide a more nuanced view of model accuracy and reliability, helping us to identify areas for further improvement.

Automated Trading Strategies

Finally, we will explore the practical application of our models by developing automated trading strategies based on the predictions. This will involve backtesting and paper trading to evaluate the financial viability of AI-driven predictions. Backtesting will allow us to simulate trading strategies on historical data to assess their performance and risk. Paper trading, on the other hand, will enable us to test strategies in real-time without financial risk, providing valuable insights into their practical applicability.

By refining our methodologies, expanding our dataset, and integrating real-time data, we aim to develop more accurate and reliable AI models for short-term stock prediction. These efforts will contribute to the ongoing advancement of AI in financial markets, providing investors and researchers with powerful tools for decision-making and analysis. Our goal is to create a robust framework that can adapt to changing market conditions and deliver consistent, high-quality predictions, ultimately enhancing the capabilities of investors and researchers alike.

Conclusion

Our research demonstrates the potential of AI models to predict short-term stock price movements, albeit with varying degrees of success. The Gradient Boosted Regressor and Orthogonal Matching Pursuit CV models showed promise, with the latter achieving up to 80% accuracy following the inclusion of a previous quarter price column. However, challenges remain, particularly in data preprocessing and feature scaling.

The exploration of different AI models and the incorporation of additional features significantly improved prediction accuracy, highlighting the importance of comprehensive data analysis and model selection. Our findings pave the way for future research and enhancements in this domain.

Our ongoing work promises to contribute valuable insights to the field of financial prediction and AI, ultimately advancing the capabilities of investors and researchers alike.

Acknowledgments

We would like to extend our sincere gratitude to Polygene for their support and collaboration. The resources and data provided by Polygene were instrumental in the successful completion of this research. Special thanks go to Vijay Krishna for his exceptional contributions to our project. His technical expertise and insightful discussions significantly enhanced the quality of our research. Thank you to everyone who contributed to this project.

References

scikit-learn. (n.d.). ElasticNetCV. In *scikit-learn 1.5.1 documentation*. Retrieved July 9, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.