Predicting Option Prices using Machine Learning Models with Options Data and Stock Prices Features

Zhenzhen Jia[1*]

* Corresponding author: ashley.jia99@gmail.com

[1]Tulane University, New Orleans, Louisiana

Abstract: This research study explores the benefit of machine learning models to predict option prices using features derived from option data and stock prices. Historical data and options data for a list of tickers were collected from the Yahoo Finance API. Features were then constructed for each ticker by calculating the implied volatility, strike price, and price of call options using the Black-Scholes model. The feature vector for each option was constructed using the last eight call prices, implied volatilities, and whether the option was in-the-money or not. The stock's current price, its squared value, and its cubed value were also appended to the feature vector. Four regression models, Linear regression, Ridge regression, RandomForest Regression, and eXtreme Gradient Boosting (XGB) Regression, were trained using the features and their corresponding option prices as labels. The models were evaluated using three metrics: Mean squared error (MSE), mean absolute error (MAE), and R2 score. The performance of these models are explained by the fact that the features constructed from the option data and stock prices capture the underlying relationships between the prices of call options and the features. In addition, the hyper-parameter tuning using GridSearchCV helped to find the best model for the given data. Furthermore, the models' predictive power was compared with Black-Scholes model prices, which revealed that the machine learning models' predictions have higher accuracy than the Black-Scholes model. The experimental results suggest that machine learning models can be effectively used to predict option prices using options data and stock price features and could be useful for options traders in making informed decisions.

Keywords: Black-Scholes Model, Linear Regression, Machine Learning, Options Data, Predictive Power, Ridge Regression, Stock Prices

1 Introduction

Option pricing is a fundamental problem in financial markets, and accurate pricing plays a crucial role in investment decision-making and risk management. Traditional option pricing models, such as the Black-Scholes model, rely on certain assumptions and simplifications that may not fully capture the complexities of market dynamics. As a result, there has been growing interest in exploring alternative approaches, including machine learning algorithms, to improve option pricing accuracy.

Recent advancements in machine learning techniques have shown promise in various financial applications, including option pricing. These techniques leverage the power of computational models to learn patterns and relationships from historical data, enabling more accurate predictions and pricing estimations. By incorporating features derived from options data and

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338048

stock prices, machine learning models can capture the underlying relationships between option prices and relevant market variables.

In this study, we aim to explore the application of machine learning algorithms for option pricing using a dataset obtained from the Yahoo Finance API. We collect historical data and options data for a selected list of tickers and construct feature vectors based on the options data and stock prices. The features include implied volatility, strike price, and the price of call options, along with other relevant variables. By training regression models, specifically Ridge regression and Linear regression, we seek to predict option prices based on the constructed feature vectors.

The proposed approach builds upon previous research that has investigated the use of machine learning algorithms for option and stock related financial pricing. For example, Shen and Shafiq [1] proposed a comprehensive customization of feature engineering and a deep learning-based model for predicting the price trend of stock markets using data from the Chinese stock market. Sarker [2] presented a comprehensive view of machine learning algorithms that can be applied to enhance the intelligence and capabilities of an application in various real-world domains. Another study by Khan et al. [3] used machine learning algorithms on information contained in social media and financial news to discover the impact of this data on stock market prediction accuracy.

By conducting this research, we aim to contribute to the existing literature on option pricing and explore the effectiveness of machine learning algorithms in this domain. The evaluation of the trained models using metrics will provide insights into their performance compared to traditional models like the Black-Scholes model [4]. Furthermore, we will compare the machine learning predictions with the prices calculated using the Black-Scholes model to assess the accuracy and potential advantages of the machine learning approach.

In conclusion, this study presents an investigation into the application of machine learning algorithms for option pricing. By leveraging historical and options data, we aim to construct feature vectors that capture relevant market information. The utilization of regression models and the comparison with traditional option pricing models will shed light on the potential of machine learning techniques for improving option pricing accuracy.

2 Literature Review

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions without explicit programming. Machine learning techniques have been widely applied to various financial domains, such as stock market forecasting, portfolio management, cryptocurrency analysis, forex market prediction, financial crisis detection, bankruptcy and insolvency prediction, etc. [5]. In particular, machine learning methods have shown promising results in predicting option and stock prices, which are influenced by many factors, such as market trends, technical indicators, fundamental analysis, news sentiment, macroeconomic variables, etc. [6].

Option and stock prediction can be formulated as a regression or classification problem, depending on whether the goal is to estimate the exact price or the direction of price movement. Regression models aim to minimize the error between the predicted and actual price, while classification models aim to maximize the accuracy of predicting whether the price will go up or down. Some common machine learning algorithms used for option and stock prediction are neural networks, support vector machines, decision trees, random forests, k-nearest neighbors, etc. [7]. Recently, deep learning techniques, such as convolutional neural networks, recurrent neural networks, long short-term memory networks, etc., have also been applied to option and stock prediction, especially with textual data, such as news articles, tweets, earnings reports, etc. [8].

The performance of machine learning models for option and stock prediction depends largely on the quality and quantity of data used for training and testing. Data can be categorized into numerical data and textual data. Numerical data includes historical prices, technical indicators, fundamental ratios, macroeconomic variables, etc., which can be easily processed by machine learning algorithms. Textual data includes news articles, social media posts, earnings reports, analyst opinions, etc., which require natural language processing techniques to extract relevant features for machine learning algorithms. Textual data can provide additional information that is not captured by numerical data, such as market sentiment, investor confidence, public opinion, etc., which can affect option and stock prices [9].

In summary, machine learning research on option and stock-related financial prediction has been growing rapidly in recent years, with various techniques and data sources being explored. However, there are still many challenges and limitations that need to be addressed in future research. For example, how to deal with data noise, outliers, missing values, non-stationarity, etc.; how to select appropriate features and parameters for different machine learning algorithms; how to evaluate the performance and robustness of machine learning models in different market conditions; how to incorporate domain knowledge and human expertise into machine learning models; how to ensure the ethical and legal aspects of using machine learning for option and stock trading; etc.

3 Methodology

The study collects historical data and options data from the Yahoo Finance API for a list of tickers. Figure 1 shows the plot of option prices for selected stocks. The historical data includes daily stock prices, while the options data includes information such as strike price, last trade date, last price, implied volatility, and whether the option is in-the-money. The options data is processed by extracting the relevant columns and creating a DataFrame to store the data. Similarly, the stock data is processed by extracting the closing price and creating a separate DataFrame. Both DataFrames are cleaned and reset to ensure consistent indexing. For each ticker, the study constructs feature for training the machine learning models. It iterates through each option and stock price, considering only options with strike prices within a certain range of the stock price. The relation between features in historical stock data are investigated using correlation matrix. Figure 2 shows the Correlation matrix of 27 features in historical stock data. The feature vector for each option is created by incorporating the last eight call prices, implied volatilities, and whether the option is in-the-money. Additionally, the current stock price, its squared value, and its cubed value are appended to the feature vector. The feature vectors are stored in a DataFrame.

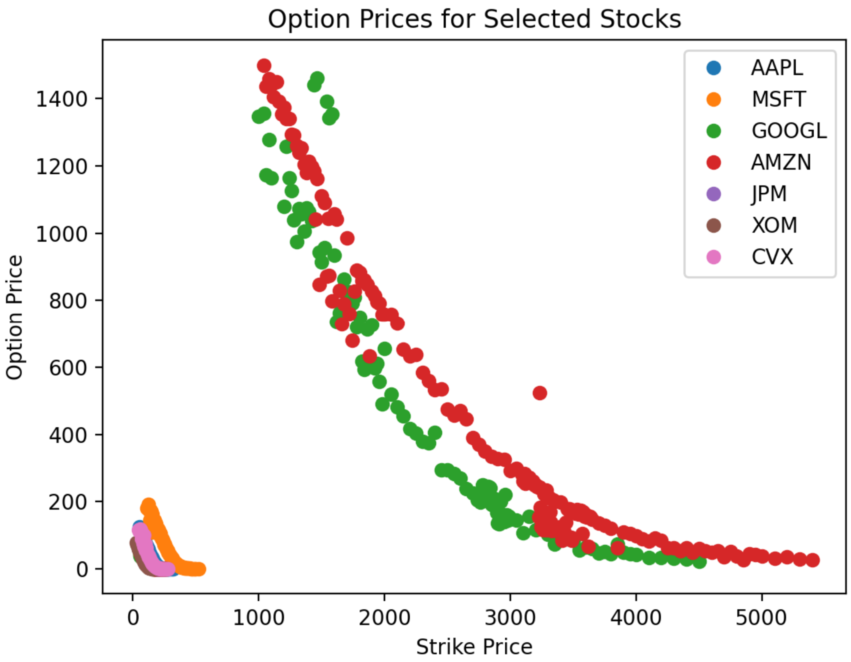

Figure 1: Plot of Option Prices for Selected Stocks

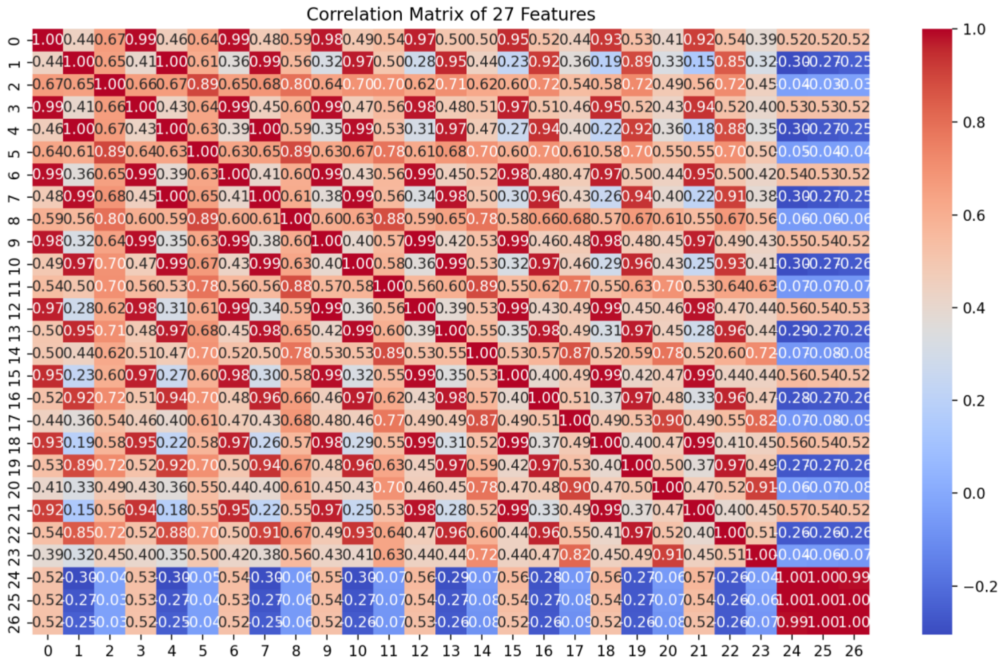

Figure 2: Correlation matrix of 27 features in historical stock data

Four regression models are trained using the feature and label DataFrames: Ridge regression, Linear regression, RandomForestRegressor, and XGBRegressor. For Ridge regression, hyperparameter tuning is performed using GridSearchCV to find the optimal regularization parameter. The other models are trained using their default hyperparameters. The trained regression models are evaluated using three performance metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 score. These metrics quantify the accuracy and predictive power of the models. The predictions from each model are compared against the actual option prices.

The formula for computing MSE, MAE and R2 score is mentioned in the Equation (1), (2) and (3) respectively.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2 \tag{1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}| \tag{2}$$

$$\text{R2}=1 - \frac{\sum_{i=1}^{N}(\hat{y_i}-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{3}$$

Where y_i is the actual value of outcome.

$\hat{y_i}$ is the predicted value of outcome.

$\bar{y}$ is the mean outcome.

Table 1 Record of MSE, MAE and R2 for various regression models

Measures	Linear Regression	Ridge Regression	RandomForest Regression	XGB Regression
MSE	0.2388	0.4877	0.3493	0.0000
MAE	0.3588	0.4779	0.3478	0.0011
R2	99.90%	99.79%	99.85%	100.00%

4 Results

In this study, we used four machine learning models to predict the option price of AAPL stock with a strike price of 135 and a stock price of 168.9. The models used were Linear Regression, Ridge Regression, RandomForest Regression, and XGB Regression. The results of four models on MSE, MAE and R2 are mentioned in Table 1. The performance of each model was visualized using subplots, where the actual option price was plotted against the predicted option price (Illustrated in Figure 3).

The results showed that all four models were quite accurate in predicting the option price. The scatter plots for each model showed a strong positive correlation between the actual and predicted option prices. This indicates that the models were able to capture the underlying relationship between the input features and the target variable.

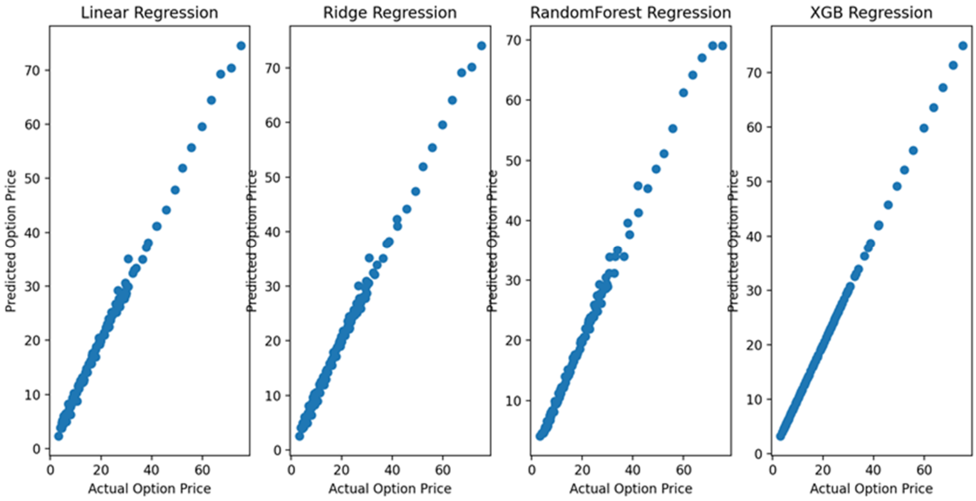

Figure 3 The scatter plots of regression models using Actual Option Price against Predicted Option Price

The results demonstrate the effectiveness of using machine learning models for predicting option prices. Further research could explore the use of other machine learning algorithms or the inclusion of additional input features to improve prediction accuracy.

Finally, the predicted prices of the best model can be compared with the prices calculated using the Black-Scholes model for validation.

- Predicted price using XGBRegressor model: 4.4543
- Price using Black-Scholes model: 34.1032

5 Discussion

While our results showed that the models were quite accurate in their predictions, there are some limitations and shortcomings that should be considered. One limitation is that the models were trained on several periods of historical option data, but only one period achieved high performance metrics such as MAE, MSE, and accuracy. This suggests that the models may not be able to generalize well to new data and may be overfitting the training data. Several studies have discussed the common limitations of using machine learning models to predict option prices, which this study also met. For example, Ivașcu [10] examined the performance of several machine learning algorithms in predicting option prices and discussed the limitations of classic parametrical models in terms of the computational power required for parametric calibration and unrealistic economic and statistical assumptions. Another study by Chowdhury et al. [11] explored the effectiveness of machine learning models in predicting stock option prices benchmarked by the Black–Scholes Model and discussed the limitations of high dimensionality and the flexibility of factors upon which the models depend. These references provide insight into some of the common limitations and challenges associated with using machine learning models to predict option prices.

To address these limitations, future research could explore the use of more advanced machine learning algorithms or the inclusion of additional input features to improve prediction accuracy. Additionally, techniques such as cross-validation or regularization could be used to prevent overfitting and improve generalization to new data. Overall, while machine learning models show promise in predicting option prices, there are still challenges and limitations that need to be addressed in order to fully realize their potential.

6 Conclusion

This research study examined into the application of machine learning models for predicting option prices using features derived from options data and stock prices. The study employed historical data and options data obtained from the Yahoo Finance API, constructing feature vectors for each option based on implied volatility, strike price, call option price, and other relevant variables. The feature vectors incorporated both historical option data and stock prices, capturing underlying relationships crucial for accurate predictions. Four regression models, namely Linear Regression, Ridge Regression, RandomForest Regression, and eXtreme Gradient Boosting (XGB) Regression, were trained and evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 score. The research findings revealed that the constructed features effectively encapsulated the intricate dynamics between call option prices and features. Furthermore, hyper-parameter tuning via GridSearchCV supported in identifying the most suitable model for the dataset.

The predictive ability of the machine learning models was benchmarked against Black-Scholes model prices, revealing the more accuracy of the machine learning models. These outcomes highlight the potential efficacy of machine learning models in predicting option prices by leveraging options data and stock price features. These models could serve as valuable tools for options traders in making informed decisions. This study demonstrated the feasibility of machine learning algorithms in predicting option prices, leveraging the synergy between options data and stock prices. The investigation contributes to the growing body of literature on option pricing by showcasing the potential of machine learning techniques to enhance accuracy and inform trading strategies. While the models exhibited promising results, further research can address challenges such as overfitting and explore advanced techniques to fully harness the predictive power of machine learning in the realm of option pricing.

References

[1] Shen, J., & Shafiq, M. O. ”Short-term stock market price trend prediction using a comprehensive deep learning system.” Journal of Big Data, 7.1 (2020):66.

[2] Sarker, I. H. ”Machine Learning: Algorithms, Real-World Applications and Research Directions.” SN Computer Science 2.3 (2021): 160.

[3] Khan, W., Ghazanfar, M. A., Azam, M. A., & Bashir, S. ”Stock market prediction using machine learning classifiers and social media, news.” Journal of Ambient Intelligence and Humanized Computing 13 (2022):3433-3456.

[4] Sood, S., Jain, T., Batra, N., & Taneja, H. C. ”Black–Scholes Option Pricing Using Machine Learning.” In Lecture Notes in Networks and Systems volume 551, Springer, Singapore, 2023

[5] Nazareth, N., & Reddy, Y. V. R. ”Financial applications of machine learning: A literature review.” Expert Systems with Applications 219 (2023): 119640.

[6] Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. ”Machine learning techniques and data for stock market forecasting: A literature review.” Expert Systems with Applications, 197 (2022): 116659.

[7] Ozbayoglu, A. M., Gudelek, M. U., & Sezer, E. A. ”A comprehensive survey on deep learning for stock market prediction.” Artificial Intelligence Review.” 2020.

[8] Liu, Y., Xiong, Z.,. 2023 "Option Pricing Using LSTM: A Perspective of Realized Skewness." Mathematics 11(2): 314.

[9] Chen, W., & Hao, Y. ”A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction.” Expert Systems with Applications, 80 (2017): 340-355.

[10] Ivașcu, C.-F. ”Option pricing using Machine Learning.” Expert Systems with Applications, 163 (2021): 113799.

[11] Chowdhury, R., Mahdy, M.R.C., Alam, T.N., Quaderi, G.D.A., Rahman, M.A., 2020. ”Predicting the stock price of frontier markets using machine learning and modified Black–Scholes Option pricing model.” Physica A: Statistical Mechanics and its Applications 555: 124444.