Pricing Indian Options with Advanced Tree-based Models

Thesis Project report submitted to

Indian Institute of Technology, Kharagpur

for partial fulfilment for the award of the degree of

MASTER OF SCIENCE

in

ECONOMICS

by

Abhishek Kumar Singh

(19HS20002)

Under the supervision of

Prof. Dripto Bakshi

DEPARTMENT OF HUMANITIES AND SOCIAL SCIENCES

Indian Institute of Technology, Kharagpur

Spring Semester, 2023-24

April 30, 2024

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: April 30, 2024 (Abhishek Kumar Singh)

Place: Kharagpur (19HS20002)

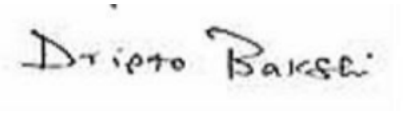
DEPARTMENT OF HUMANITIES AND SOCIAL SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Pricing Indian Options with Advanced Tree-based Models**" submitted by **Abhishek Kumar Singh** (Roll No. 19HS20002) to Indian Institute of Technology, Kharagpur towards partial fulfilment of requirements for the award of degree of MASTER OF SCIENCE in ECONOMICS is a record of bona fide work carried out by him under my supervision and guidance during Spring Semester, 2023-24.



Prof. Dripto Bakshi
Department of Humanities
and Social Sciences
Indian Institute of Technology,
Kharagpur
Kharagpur - 721302, India

Date: April 30, 2024

Place: Kharagpur

Abstract

Name of the student: **Abhishek Kumar Singh** Roll No: **19HS20002**

Degree for which submitted: **MASTER OF SCIENCE**

Department: **HUMANITIES AND SOCIAL SCIENCES**

Thesis title: **Pricing Indian Options with Advanced Tree-based Models**

Thesis supervisor: **Prof. Dripto Bakshi**

Month and year of thesis submission: **April 30, 2024**

This thesis explores the efficacy of advanced tree-based machine learning models in pricing options in the Indian financial market, specifically focusing on options based on the Nifty Index. Traditional option pricing methods, like the Black-Scholes model, have long served the financial community but often fall short under the dynamic and complex market conditions prevalent in emerging markets such as India due to rigid assumptions about market efficiency, volatility, and liquidity. This study evaluates and compares the performance of several tree-based models, including Random Forest, XGBoost, and LightGBM, to address these shortcomings. Using five years of historical option data from the National Stock Exchange (NSE) of India spanning, from April 1, 2019, to March 31, 2024, this research incorporates data fields such as open interest changes and market sentiment indicators, aiming to enhance the predictive accuracy of these models. By comparing these advanced models against the traditional Black-Scholes model, this study aims to provide a comprehensive understanding of their relative strengths and limitations in the context of the Indian options market.

Acknowledgements

Firstly, I would like to extend my sincere gratitude to my project supervisor and mentor Prof. Dripto Bakshi for his exceptional guidance and support, without which the project would not have progressed as it had over the semester. His invaluable guidance and advice carried me through all the stages of this project. Next, I would like to thank my parents and friends whose constant support kept me motivated and going throughout this project.

Contents

List of Figures

Chapter 1

Introduction

The pricing of options is a fundamental activity in financial markets, providing insights into future market behaviors and aiding in the formulation of effective trading and hedging strategies. Historically, the Black-Scholes model has been the cornerstone of option pricing due to its pioneering approach to estimating European option prices under a lognormal distribution assumption. However, the financial markets in emerging economies like India exhibit characteristics such as high volatility, market inefficiencies, and regulatory idiosyncrasies that challenge the applicability of traditional models.

The Nifty Index, which represents the aggregated performance of fifty major Indian companies, provides a unique yet complex landscape for option pricing due to its critical role in reflecting the broader market dynamics. The limitations of the Black-Scholes model, such as its assumptions of constant volatility and market efficiency, often lead to significant pricing errors when applied to such markets. This necessitates the exploration of alternative models that can incorporate more complex market phenomena and offer flexibility in adapting to rapid changes.

This thesis posits that advanced tree-based machine learning models, known for their robustness in handling nonlinear relationships and large datasets, can significantly improve the accuracy of option pricing in the Indian market. By integrating additional market data such as open interest, changes in open interest, and volatility indices, this research seeks to harness the predictive power of Random Forest, XGBoost, and LightGBM models.

Chapter 2

Literature Review

The main goal of this review is to analyse the methods currently used in pricing options, especially focusing on the tree-based models utilised in financial markets. This review is essential as it sets the stage for evaluating the potential improvements these models can offer over traditional methods, particularly within the Indian financial context.

2.1 Background Information

2.1.1 Basics of Option Pricing

Options are financial derivatives that provide the holder with the choice to buy or sell an underlying asset at a predetermined price before or on a specific date. There are two main types of options: calls, which give the holder the right to buy, and puts, which give the holder the right to sell. The strike price is the predetermined price at which the option can be exercised, while the expiration date refers to the date on which the option expires. Options are further categorized into two main styles: European and American. European options, which is the focus of this study, can only be exercised at the expiration date, not before, making them somewhat simpler to analyze than American options, which can be exercised at any time up to and including the expiration date. This flexibility of American options requires

more complex pricing methods and considerations of early exercise opportunities, especially in volatile markets. (Chen, 2024).

Pricing options is a crucial part of financial markets, where the fair value of the derivative is calculated based on various factors such as the underlying asset's price, strike price, time to expiration, volatility, and the risk-free interest rate. These factors play a role in the intrinsic and time value of the option, which ultimately determine its premium (Hull and Basu, 2016).

2.1.2 Importance of the Nifty Index as Underlying Asset

The Nifty Index, also known as the Nifty 50, is a prominent index on the National Stock Exchange of India (NSE), comprising a diverse collection of fifty major Indian companies. This indicator plays a crucial role in analysing the performance of the Indian stock market, making it a highly relevant subject to study in relation to option pricing. There are several reasons why it holds such significance.

1. Market Representation: The Nifty 50 encompasses key sectors of the Indian economy and serves as a benchmark index for assessing the overall performance of the Indian stock market and economy. This extensive representation makes it a favored option for investors seeking to hedge or speculate on the overall Indian market (NSE India, [n. d.]).

2. Liquidity: Options based on the Nifty Index are widely traded derivatives in India, indicating their significant liquidity. Having sufficient liquidity is essential for accurately pricing options because it guarantees enough trading activities to reflect the true market value of the options (Shah, 2023).

3. Volatility: A key component of option pricing is price volatility, which the Nifty Index, like many large indices, shows. The Nifty's inherent volatility offers traders numerous opportunities to utilise options for hedging and trading strategies. As a result, studying its option pricing becomes particularly significant (ICICI Direct, 2023).

This thesis attempts to tap into the rich dynamics provided by one of the most significant financial indexes in emerging nations by concentrating on options using the Nifty Index as the underlying asset, offering insights that are applicable both locally and worldwide.

2.2 Traditional Models for Option Pricing

2.2.1 Overview of Models

Option pricing models are fundamental to the strategies employed by financial institutions and individual investors in the options markets. The most influential traditional model:

Black-Scholes Model: Developed in the early 1970s by Fischer Black, Myron Scholes, and Robert Merton, the Black-Scholes Model assumes that the stock market follows a log-normal distribution and provides a formula for estimating the price of European-style options. The model is famous for being able to work in a continuous-time setting where stock returns are normally distributed and volatility stays the same over the life of the option (Black and Scholes, 1973)(Merton, 1973).

2.2.2 Limitations of Traditional Models

- Assumption of Market Efficiency: Both models operate under the assumption of highly efficient markets, where prices accurately reflect all available information. Nevertheless, the Indian market occasionally exhibits inefficiencies caused by lower participation rates and the influence of specific investor groups. These factors can result in price anomalies that are not considered by these models.
- Constant Volatility Assumption: The Black-Scholes model assumes that volatility is constant, but in real-world markets, volatility is often dynamic and influenced by various economic events. This holds especially true in emerging markets such as India, where economic news has the potential to cause substantial market fluctuations.

- Lack of Dividend Adjustment: Most of the time, traditional models don't take dividends into account properly. This can be a big problem when pricing options on stocks that are part of the Nifty Index, since dividends can change the prices of the underlying assets.

- Inapplicability to American Options: The models are mainly intended for European options, which can only be exercised at expiration. On the other hand, American options can be exercised at any time before and including the expiration date. These traditional models can sometimes result in pricing inaccuracies due to this discrepancy.

- Simplistic Risk-Free Rate Assumption: The models frequently rely on a fixed risk-free rate, which may not accurately reflect the dynamics of an emerging market where interest rates can fluctuate and be impacted by policy shifts.

In order to accurately capture the complexities of the Indian financial market, it is necessary to explore more sophisticated models that can overcome these limitations.

2.3 Introduction to Tree-Based Models

2.3.1 What are Tree-Based Models?

Tree-based models are a type of machine learning algorithms that rely on decision trees as their main building blocks. For both regression and classification problems, they are widely used. Three tree-based models that are commonly used in advanced analytics are:

- Random Forest: A Random Forest model is a machine learning technique that combines multiple decision trees to make predictions. By doing this, it can provide more accurate and reliable predictions by reducing the variability in the results. Every tree in the forest is constructed using a bootstrap sample, which is a sample selected from the training set with replacements. The Random Forest algorithm is known for its robust performance metrics and lower risk of

overfitting compared to a single decision tree. This is achieved by averaging the predictions of multiple trees (Breiman, 2001).

- XGBoost (Extreme Gradient Boosting): XGBoost is a machine learning algorithm that uses gradient boosting to create decision trees. It is specifically designed to be fast and efficient. The algorithm is known for its high flexibility and has consistently performed well in machine learning competitions. This is because it can effectively handle various types of data, relationships, and distributions. XGBoost enhances the performance and speed of the model by optimizing the utilization of computational resources and memory (Chen and Guestrin, 2016).

- LightGBM (Light Gradient Boosting Machine): LightGBM is a machine learning algorithm that builds on the gradient boosting framework. It does this by using a histogram-based approach to group continuous feature values into discrete bins. This feature allows for faster training and less memory usage, which makes it very efficient when working with large datasets and extensive feature spaces. LightGBM also has a unique way of dealing with sparse data and can handle categorical features naturally (Ke et al., 2017).

2.3.2 Advantages Over Traditional and Modern Models

Tree-based models provide several advantages over both traditional models such as Black-Scholes and modern deep learning models like neural networks, particularly in the context of pricing options:

- Interpretability: Tree-based models are different from deep learning models because they are not considered black boxes. Instead, they offer clear indications of which variables have the most impact on predictions. Transparency is really important in financial applications because it's just as important to know why a model is making certain predictions as it is to know how accurate those predictions are (James et al., 2013).

- Handling Non-linearity: Options pricing can be quite complex because it involves non-linear relationships. Factors like volatility and time decay play a

significant role in determining the price. Tree-based models have the ability to naturally capture non-linear relationships in the data, which means there is no need to transform the input data. This gives them a clear advantage over traditional linear models (Hand, 2007).

- Efficiency on Cross-Sectional Data: Tree-based models are great for dealing with large datasets that have many variables. They work especially well for cross-sectional studies commonly found in financial markets. Feature selection is performed by them, which helps in preventing overfitting. This is particularly beneficial in situations where the connection between the input variables and the target variable is intricate and non-linear (Hastie et al., 2009).

- Low Sensitivity to Outliers: Tree-based models tend to be less affected by outliers compared to neural networks. Extreme values can have a significant impact on neural networks. Financial datasets often contain outliers, which can be attributed to market shocks. This makes the ability to handle outliers particularly valuable (Hastie et al., 2009).

- Flexibility in Integration of Market Conditions: Tree-based models are different from models that need stationary input. They can adjust to changes in market conditions and incorporate real-time data updates. This allows them to offer more precise and timely pricing options (Garman and Klass, 1980).

2.4 Tree-Based Models in Financial Markets

The dynamic nature of financial markets demands advanced analytical methods to predict market movements and manage financial risk effectively. Recent developments in machine learning, particularly tree-based models, have significantly enhanced predictive accuracy and interpretability in financial applications.

2.4.1 Tree-Based Models in Financial Market Prediction

Tree-based models, including decision trees, random forests, and gradient boosting methods, have become foundational in financial predictions due to their ability to

model complex, non-linear relationships without extensive pre-processing of data. Gu et al. (2020) discuss the efficacy of machine learning in empirical asset pricing, noting that tree-based models provide robust predictive capabilities across various asset classes Gu et al. (2020).

Basak et al. (2019) further explore the application of tree-based models in financial market prediction. This study explores the efficacy of tree-based classifiers in forecasting the direction of stock market prices, presenting insights into their performance, methodologies, and implications within the realm of financial markets.(Basak et al., 2019).

2.4.2 Enhancements in Financial Distress Prediction

Tree-based models have become prominent tools for predicting financial distress in companies, appreciated for their ability to handle complex, nonlinear relationships without needing assumptions about data distributions. Gepp and Kumar (2015) demonstrated that decision trees provide clear, interpretable paths for decision-making, useful in identifying distressed companies. Advancing this, Qian et al. (2022) improved prediction accuracy using gradient boosted decision trees with a corrected feature selection measure, which helps in reducing overfitting and enhancing model generalization. Liu et al. (2023) further enriched the interpretability of these models by integrating explainable machine learning techniques, making the predictive process transparent and more accessible for stakeholders. Together, these studies underline the efficacy and clarity that tree-based models bring to financial distress prediction.

2.4.3 Advancing Explainability in Financial Models

Recent research underscores the importance of advancing explainability in financial models using tree-based machine learning techniques. Park et al. (2021) and Tran et al. (2022) both emphasize the need for models that are not only accurate but also transparent in their decision-making processes. Park et al. (2021) focus on the implementation of techniques like LIME and SHAP to enhance understanding

of bankruptcy predictions, revealing the influence of individual features in complex models. Tran et al. (2022), meanwhile, illustrate how explainable machine learning can demystify financial distress predictions in Vietnamese firms, aiding stakeholders in making informed decisions. Both studies highlight a critical shift towards making sophisticated predictive tools more interpretable and trustworthy in high-stakes financial environments.

Tree-based models have proven to be indispensable in the financial sector, offering a blend of predictive power and explainability that is crucial for modern financial analysis and decision-making. Future research should aim to bridge the gap between advanced predictive models and user-friendly interpretative frameworks to enhance both the utility and accessibility of financial machine learning applications.

2.5 Advances in Option Pricing Using ML

The application of machine learning in financial markets has revolutionized traditional methodologies, particularly in the domain of option pricing. Traditional models like Black-Scholes have been foundational but often fall short under the complex dynamics of modern financial markets. Recent advancements leverage machine learning to overcome these limitations, offering more adaptive and accurate pricing mechanisms.

2.5.1 Current Progress in ML for Option Pricing

1. **Data-Driven Models:** Jayaraman et al. (2022) highlight the efficacy of data-driven machine learning models, such as Random Forest, in pricing American style stock options. These models excel due to their minimal assumptions about market conditions, which allows for more flexible and accurate pricing under varied market scenarios (Jayaraman et al., 2022).

2. **Hybrid and Ensemble Models:** Wang et al. (2022) (2022) introduce an innovative approach that combines ensemble learning methods with network learning structures. Their model not only improves prediction accuracy by 36% compared to traditional methods but also demonstrates the potential of

hybrid models in enhancing the robustness and efficiency of predictions (Wang et al., 2022).

3. **Comparative Studies:** Research by Gan et al. (2020) (2021) explores the comparative advantages of machine learning models over conventional approaches. They find that models like XGBoost and LightGBM, known for handling large datasets and intricate variable interactions, provide significant improvements in option pricing (Gan et al., 2020).

4. **Deep Learning Applications:** Ruf and Wang (2019) review the use of neural networks in option pricing and hedging, noting the transition to complex deep learning networks. While these networks offer enhanced pattern recognition capabilities, they also introduce challenges related to overfitting and model interpretability (Ruf and Wang, 2019).

5. **Market Adaptability:** Ivaşcu (2021) investigates the adaptability of machine learning algorithms to dynamic market environments. The study confirms that machine learning models can outperform traditional models by adapting more effectively to market changes, particularly in volatile conditions (Ivaşcu, 2021).

2.6 Research Gaps

- **Lack of Research on Indian Markets:** Much of the existing research focuses on broadly recognized markets like the SP 500. There is a notable gap in studies specifically targeting the Indian options market, which has unique characteristics and regulatory environments.

- **Integration of Market Sentiments:** Current models seldom incorporate market sentiment indicators such as open interest, which can provide crucial insights into market trends and investor behavior.

- **Lack of research on the impact of Global Events:** The COVID-19 pandemic has introduced unprecedented volatility and economic shifts. Existing models often do not account for such global events, which can dramatically affect option pricing.

Chapter 3

Research Questions and Objectives

3.1 Research Questions

1. How effective are advanced tree-based machine learning models, such as Random Forest, XGBoost, and LightGBM, in pricing options on the Nifty Index compared to traditional models like Black-Scholes?

2. Can the integration of market sentiment indicators, such as open interest and trading volume, enhance the accuracy of option pricing models for the Indian market?

3. How do tree-based models adapt to the unique characteristics and dynamics of the Indian financial market?

3.2 Research Objectives

1. To evaluate and compare the performance of advanced tree-based machine learning models with traditional option pricing models in the context of the Indian financial market. This objective involves setting up experiments to benchmark the accuracy and computational efficiency of tree-based models against traditional methods like the Black-Scholes model using historical data from the Nifty Index.

2. To investigate the effectiveness of incorporating market sentiment indicators into tree-based models for enhancing the predictive accuracy of option pricing. This objective focuses on data engineering to integrate sentiment analysis and market indicators into the feature set of the pricing models, followed by an evaluation of their impact on pricing accuracy.

3. To explore the interpretability of tree-based models in the context of option pricing, providing insights into the decision-making process of these models. This involves understanding which features are more important than others in predicting and influencing the final predicted prices.

Chapter 4

Dataset

4.1 Data Source and Description

Our study uses five-year historical data of Nifty index options, sourced from the National Stock Exchange (NSE) of India website. The dataset spans from April 1, 2019, to March 31, 2024, and includes the following fields: Date, Expiry, Strike Price, Settle Price, Close, Open Interest, Change in Open Interest (OI), and Underlying Value. Both put and call options are included, with identifiers set as 1 for calls and 0 for puts in the variable Option_type.

4.2 Data Fields and Their Importance

- **Settle Price**: This represents the closing price of the option on a particular trading day. It is vital as it reflects the market consensus of the option's value at the end of each trading session and serves as our **target variable** for modelling and comparing with model-predicted prices.
- **Open Interest (OI)**: Open Interest is a key metric that represents the total number of outstanding option contracts that remain open and have not been closed or delivered. OI is a direct indicator of the liquidity and depth of a particular option market. A higher OI indicates more activity and interest

in the option, suggesting better liquidity which facilitates easier entry and exit positions. For modelling purposes, variations in open interest can signal potential price movements.

- **Change in Open Interest (Change in OI)**: This variable measures the difference in open interest from the previous trading day. It is a crucial indicator for assessing the flow of money into or out of the market. A positive change (increase) in OI typically implies fresh buying or selling interest, indicating that the current trend may continue. Conversely, a decrease in OI might suggest that traders are closing their positions, which could either be due to profit bookings or loss cuttings. This feature helps in understanding market dynamics and predicting the directionality of option prices, as significant changes in OI can be correlated with pivotal movements in option pricing.

4.3 Additional Data Utilized

- **Risk-Free Rates**: Obtained from the Reserve Bank of India (RBI), specifically from "HBS Table No. 180 - Yield of SGL Transactions in Government Dated Securities for Various Maturities." These rates are interpolated based on the maturity periods of the options to calculate the theoretical prices of options.

- **Dividend Yield (Div_yield) of Nifty Index**: Sourced from the NSE website. This is crucial for pricing models as dividends impact the underlying asset's expected return, influencing the option's pricing.

- **Volatility**: Taken Implied volatility from INDIA VIX NSE website and Calculated as the historical volatility from the standard deviation of the Nifty's returns matching the option's maturity period. This measure is essential for estimating the option's sensitivity to market movements.

4.4 Analytical Enhancements

- **Inclusion of OI Metrics in Models**: In our tree-based modelling approach, both Open Interest and its changes are utilised as features to capture market sentiment and trader behaviour. These metrics help in enhancing the model's pricing accuracy by providing insights into market participants' commitment levels and potential future activity in the options market.
- **Temporal Features**: To examine potential temporal effects on option pricing, features such as date, month, year, and day of the week are incorporated into the tree-based models.

The dataset was segmented into training, validation, and testing subsets with ratios of approximately 70%, 15%, and 15% respectively. The final dataset comprised 2,642,153 observations after cleaning and preprocessing.

	Strike Price	Open Int	Change in OI	Underlying Value	time_to_maturity_days	riskfree_rates	Div_yield	Historical Volatility	Implied_Volatility	Settle Price
count	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06	2.646196e+06
mean	1.427481e+04	8.579579e+04	1.235613e+04	1.467294e+04	2.652344e+02	5.343650e+00	1.307733e+00	1.645139e-01	2.035787e+01	1.139616e+03
std	4.035007e+03	4.573058e+05	1.634155e+05	3.566297e+03	4.519343e+02	1.264069e+00	1.734360e-01	1.138512e-01	1.042809e+01	1.454735e+03
min	4.600000e+03	0.000000e+00	-9.214650e+06	7.610250e+03	1.000000e+00	3.461000e+00	9.500000e-01	1.000000e-05	1.013500e+01	0.000000e+00
25%	1.120000e+04	0.000000e+00	0.000000e+00	1.158820e+04	2.200000e+01	4.019050e+00	1.200000e+00	1.093354e-01	1.449000e+01	8.020000e+01
50%	1.415000e+04	0.000000e+00	0.000000e+00	1.495620e+04	4.800000e+01	5.441760e+00	1.290000e+00	1.437343e-01	1.748750e+01	6.037500e+02
75%	1.740000e+04	3.450000e+03	0.000000e+00	1.762225e+04	2.480000e+02	6.642900e+00	1.400000e+00	1.924576e-01	2.203250e+01	1.604700e+03
max	2.700000e+04	1.823400e+07	1.609445e+07	2.249355e+04	1.826000e+03	7.425009e+00	2.000000e+00	1.401401e+00	8.360750e+01	1.222010e+04

TABLE 4.1: Description of features

A high mean open interest (OI) and positive mean change in OI indicate that a large number of contracts are open and not yet settled and have increased over time. High std in OI and change in OI, indicating possibly volatile market conditions or a few contracts with very high OI. The risk-free rate seems to have a small standard deviation, which indicates stability in the risk-free interest rates during the time period covered by the data.

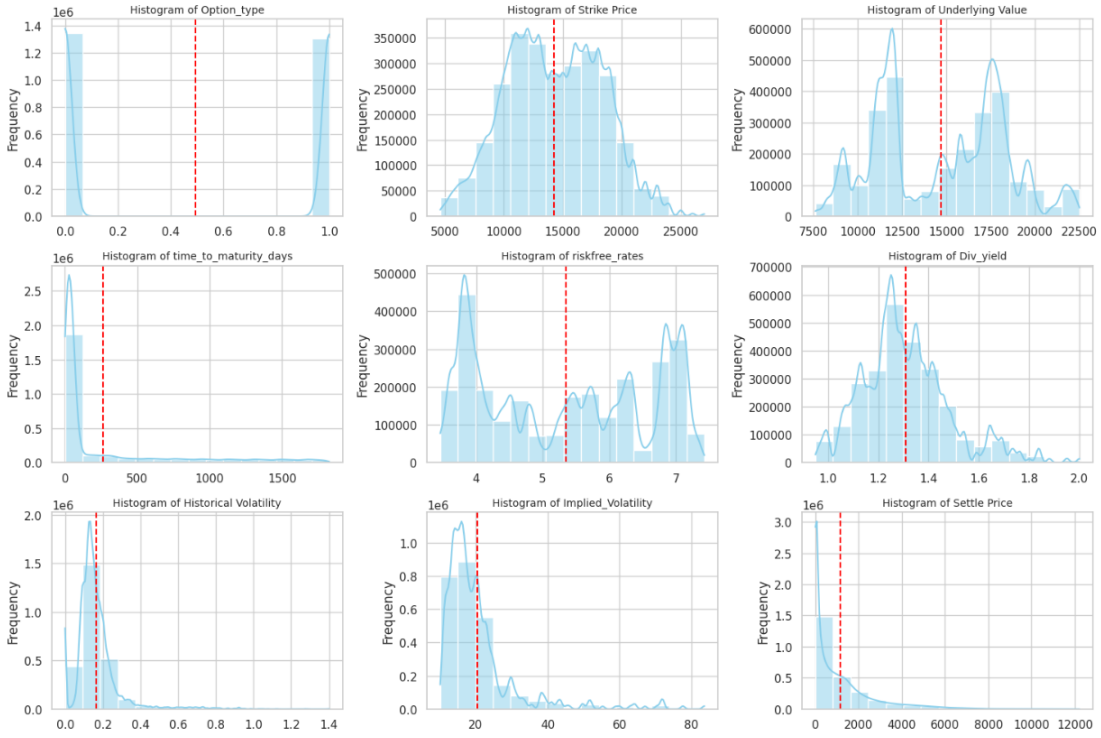


FIGURE 4.1: Histograms of features

The option type histogram shows a binary distribution, equally representing call and put options, suggesting a balanced market in terms of speculative directions. Time to maturity is predominantly short, indicating a market preference for short-term options, likely due to higher liquidity or trading strategies focused on near-term events. Strike prices and underlying values exhibit multimodal distributions, suggesting specific popular levels at which options are commonly struck, which could align with psychological pricing points or financial benchmarks. Risk-free rates and dividend yields vary but show central tendencies that indicate prevailing economic conditions or common market expectations. Historical volatility is widely spread, while implied volatility is right-skewed, implying that traders frequently anticipate higher future volatility than past trends suggest, a sign of a risk-averse or speculative market mindset. Finally, the settle price distribution is right-skewed with a long tail, showcasing that most options settle at lower prices with some significant outliers, representing high-value trades or particularly volatile market conditions. These insights collectively provide a nuanced understanding of the options market dynamics within the dataset.

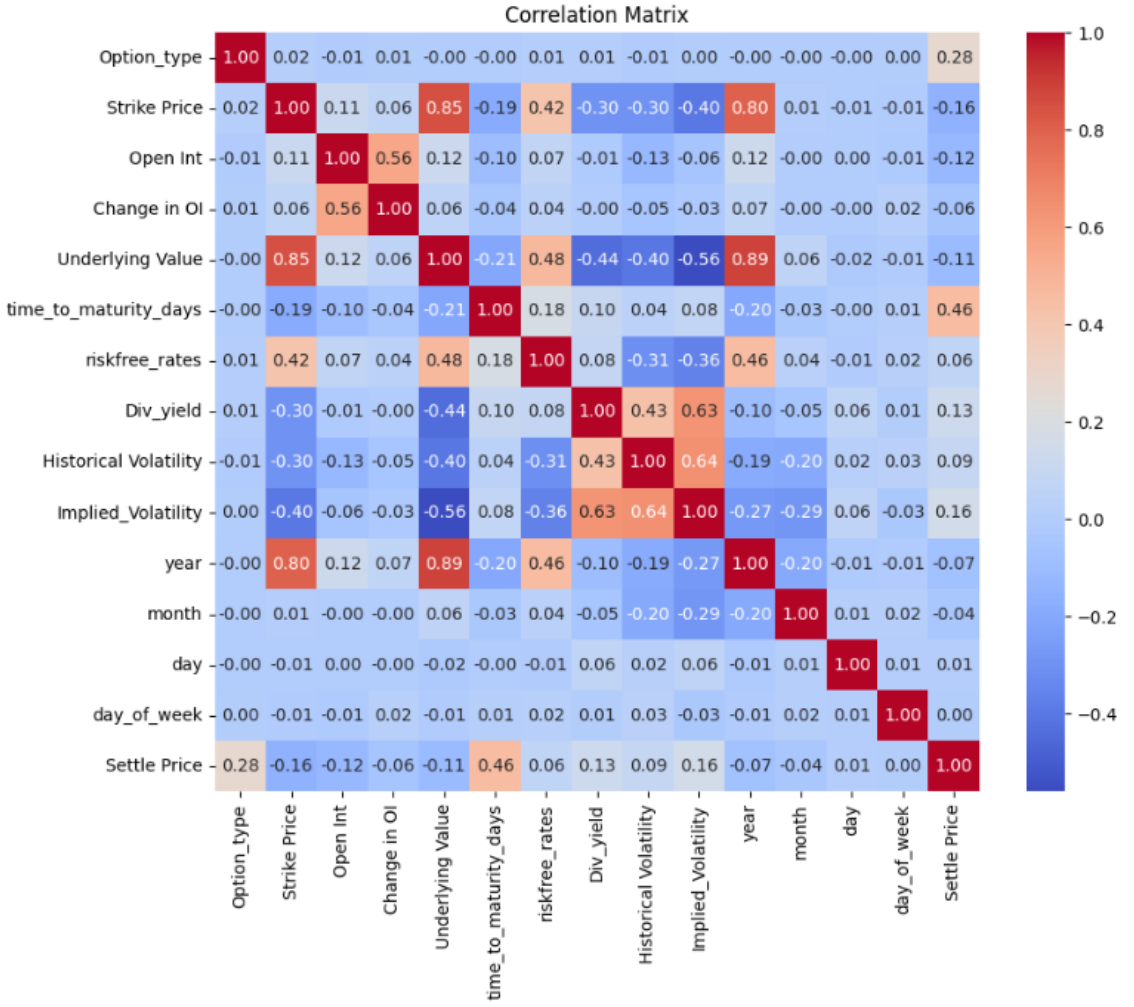


FIGURE 4.2: correlation of features

The correlation matrix for various option trading metrics reveals several key relationships. Strike price and underlying value are highly correlated (0.85), suggesting that options are often priced closely to the underlying asset's current market value, reflecting market efficiency in option pricing. There's a notable positive correlation between open interest and change in open interest (0.56), indicating that increases in trading volume are often accompanied by new contracts being added to the market, possibly signaling increased market activity or speculative interest. Both historical and implied volatilities show a strong correlation (0.64), which illustrates that market expectations of future volatility (implied) are closely aligned with past volatility trends, impacting option pricing strategies. Additionally, dividend yield shows a moderate negative correlation with implied volatility (-0.44) and settle price (-0.30), suggesting that higher dividend yields might lead to lower perceived risk and consequently lower volatility and option premiums. These correlations provide insights into the dynamics of option pricing and trading behavior, reflecting how market variables interact to shape option valuation.

Chapter 5

Models

This study employs six different models to price options, each implemented in Python. These include the classical Black-Scholes Model for theoretical pricing of European options and various advanced tree-based machine learning models which offer robust predictive capabilities and handle non-linear relationships effectively. The tree-based models explored are Random Forest with unlimited max depth (max_depth=None), allowing the nodes to expand until all leaves are pure, two configurations of Gradient Boosted Decision Trees using XGBoost, implemented with a maximum depth of 5 and with a more complex structure, having a maximum depth of 15, and two configurations of LightGBM models utilised with a maximum depth of 5 and 15. XGboost and LightGBM models were trained on a Kaggle Notebook's P100 GPU, providing substantial computational power to manage the extensive dataset.

5.1 Black-Scholes Model

The Black-Scholes model serves as the foundational theoretical model for pricing European-style options, which is used as a benchmark for other models. The model assumes a lognormal distribution for stock prices, which is a common assumption in the financial markets. The formulas for the Black-Scholes model, taking into account several key factors, are given by:

$$C = Se^{-qT}N(d_1) - Ke^{-rT}N(d_2),$$
$$P = Ke^{-rT}N(-d_2) - Se^{-qT}N(-d_1),$$

where

$$d_1 = \frac{\ln(S/K) + (r - q + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}},$$
$$d_2 = d_1 - \sigma\sqrt{T}.$$

Here, C represents the call option price, P denotes the put option price, S is the current price of the underlying security, K is the option's strike price, T refers to the time to the option's expiry, r is the risk-free interest rate, q is the dividend yield, σ is the implied volatility of the underlying security, and $N(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

This model gives us a benchmark to analyse the performance of other models.

5.2 Random Forest

The Random Forest model is employed for predicting the Settle Price of options using a comprehensive set of features extracted from the dataset. The Random Forest model was trained using the `RandomForestRegressor` class from the `sklearn.ensemble` library.

5.2.1 Model Training

The dataset was preprocessed to separate the target variable, Settle Price, from the feature columns. The data was then split into training, validation, and testing sets. The Random Forest model was configured with the following parameters:

- `n_estimators=100`: This parameter specifies the number of trees in the forest. A higher number of trees increases the robustness of the model but also the computational load.

- `max_depth=None`: By setting this to `None`, each tree in the forest is allowed to expand until all leaves are pure or until all leaves contain less than the minimum split samples.

The selected features included the Option Type (binary indicator for Call or Put), Strike Price, Open Interest, Change in Open Interest (reflecting market sentiment and potential price movements), Underlying Value (market price of the Nifty index), Time to Maturity, Risk-Free Rates (based on government securities), Dividend Yield of the Nifty index, and Historical Volatility (measured as the standard deviation of past price movements matching the option's maturity). Additionally, temporal features such as Year, Month, Day, Day of the Week, and YearMonth were incorporated to examine the effects of temporal dynamics and the specific impact of the COVID-19 pandemic on option pricing.

5.3 XGBoost

The study leverages the XGBoost framework to train two distinct models of Gradient Boosted Decision Trees, varying in complexity with maximum depths of 5 and 15. These models were trained on the same dataset used for the Random Forest, facilitating a direct comparison of their predictive capabilities under identical conditions.

5.3.1 Model Configuration

XGBoost is renowned for its efficiency and flexibility in handling various types of predictive modeling. For this study, two models were constructed:

- **XGBoost Model (Max Depth 5)**: This model, with a relatively shallow tree depth, is designed to capture essential patterns in the data without fitting excessively to the training dataset, thus reducing the risk of overfitting.

- **XGBoost Model (Max Depth 15)**: With a greater depth, this model can learn more detailed data specifics, potentially improving performance on complex option pricing scenarios but at the risk of overfitting.

Both models are configured to utilize an adaptive learning rate. This advanced feature adjusts the learning rate dynamically as boosting rounds progress. Initially, a higher learning rate is employed to make significant improvements quickly. As the number of boosting rounds increases, the learning rate decreases, allowing for finer adjustments to the model. This strategy aims to optimize the balance between learning speed and accuracy.

Customized Learning Rate Scheduler

The pseudo code for the learning rate scheduler is as follows:

```
FUNCTION eta_decay(iteration):
    max_iter = 40000
    x = iteration + 1
    eta_base = 0.1
    eta_min = 0.01
    decay_rate = -(x / 8000)^2 / max_iter
    RETURN eta_min + (eta_base - eta_min) * exp(decay_rate)
```

The code used to train the XGBoost Model with a maximum depth of five written in Python is shown below:

```
max_iter = 40000
eta_decays = ARRAY[SIZE=max_iter]
FOR iteration FROM 0 TO max_iter-1:
    eta_decays[iteration] = eta_decay(iteration)

CALLBACK LearningRateScheduler(iteration):
    IF iteration < len(eta_decays):
        RETURN eta_decays[iteration]
```

```
    ELSE:
        RETURN eta_decays[-1]

PARAMS = {
    'booster': 'gbtree',
    'eval_metric': 'mae',
    'max_depth': 5,
    'tree_method': 'hist',
    'eta': 0.5,
    'device': 'cuda',
    'reg_lambda': 25,
    'reg_alpha': 0.2
}

model = xgb.train(
    params=PARAMS,
    dtrain=dtrain,
    num_boost_round=max_iter,
    early_stopping_rounds=100,
    evals=[(dtrain, 'train'), (dval, 'validation')],
    evals_result=progress,
    verbose_eval=100,
    callbacks=[
        LearningRateScheduler(lambda iteration: eta_decays[iteration])
    ]
)
```

The Python code used to train the XGBoost Model with a maximum depth of fifteen is identical to the code used to train the model with a maximum depth of five, except for the difference in setting the `'max_depth'` parameter to 15.

5.4 LightGBM Models

LightGBM, short for Light Gradient Boosting Machine, is a highly efficient gradient boosting framework known for its fast training speed and high efficiency. Utilizing the LightGBM framework, two distinct models were trained to predict the Settle Price of options, each with varying levels of complexity based on their maximum depths. LightGBM, known for its high efficiency and lower memory usage, was chosen for its capability to handle large-scale data like ours.

5.4.1 Model Configuration

The models are configured with different tree depths to evaluate how depth influences the model's ability to capture underlying patterns and complexities in the data:

- **LightGBM Model (Max Depth 5)**: This model is designed with a shallow tree depth, aiming to prevent overfitting while ensuring the model remains general enough to apply across different data scenarios.
- **LightGBM Model (Max Depth 15)**: This deeper model allows for more complex interactions and finer segmentation of the data, potentially leading to higher accuracy but at an increased risk of fitting to noise in the training data.

Both models were trained on the same dataset as used for the Random Forest and XGBoost models, ensuring consistency in data inputs across all models. Adaptive learning rates were applied to both LightGBM models, similar to the approach used with XGBoost.

The pseudo code for LightGBM Model with a maximum depth of fifteen is as follows:

```
FUNCTION eta_decay(iteration, num_iteration):
    max_iter = num_iteration
    x = iteration + 1
    eta_base = 0.1
```

```
    eta_min = 0.01
    decay_rate = -(x / 8000)^2 / max_iter
    RETURN eta_min + (eta_base - eta_min) * exp(decay_rate)

train_data = LightGBM_Dataset(X_train, y_train)
valid_data = LightGBM_Dataset(X_val, y_val, reference=train_data)

params = {
    'objective': 'regression',
    'metric': 'mae',
    'boosting_type': 'gbdt',  # Can be adjusted to 'goss' or 'dart' if required
    'num_leaves': 2^15 - 1,
    'max_depth': 15,
    'lambda_l1': 0.2,
    'lambda_l2': 25,
    'verbose': 3,
    'device_type': 'gpu'
}

num_round = 40000
bst = LightGBM_train(
    params,
    train_data,
    num_round,
    valid_sets=[valid_data],
    callbacks=[
        reset_parameter(learning_rate=lambda iter: eta_decay(iter, num_round)),
        early_stopping(stopping_rounds=50, verbose=3)
    ]
)
```

The Python code used to train the XGBoost Model with a maximum depth of five is identical to the code used to train the model with a maximum depth of fifteen, except for the difference in setting the `'max_depth'` parameter to five.

Chapter 6

Results

The testing results for all the models are summarized in the table below.

Table 6.1: Models MAE and RMSE

Model	MAE	RMSE
XGBoost 15	19.72	53.18
Random Forest	20.13	56.50
XGBoost 5	23.72	48.03
LightGBM 15	16.05	37.32
LightGBM 5	21.61	40.66
Black Scholes	102.76	225.74

In evaluating the performance of various models for option pricing, the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) were used as the primary metrics to measure accuracy. The results, as outlined in Table 6.1, indicate significant differences in performance across the models.

6.1 Model Performance Analysis

6.1.1 XGBoost 15:

The XGBoost model with a maximum depth of 15 achieved an MAE of 19.72 and an RMSE of 53.18. This indicates that allowing the model to build deeper trees has

indeed captured more complex patterns, which has improved the accuracy compared to the Random Forest and shallower XGBoost model.

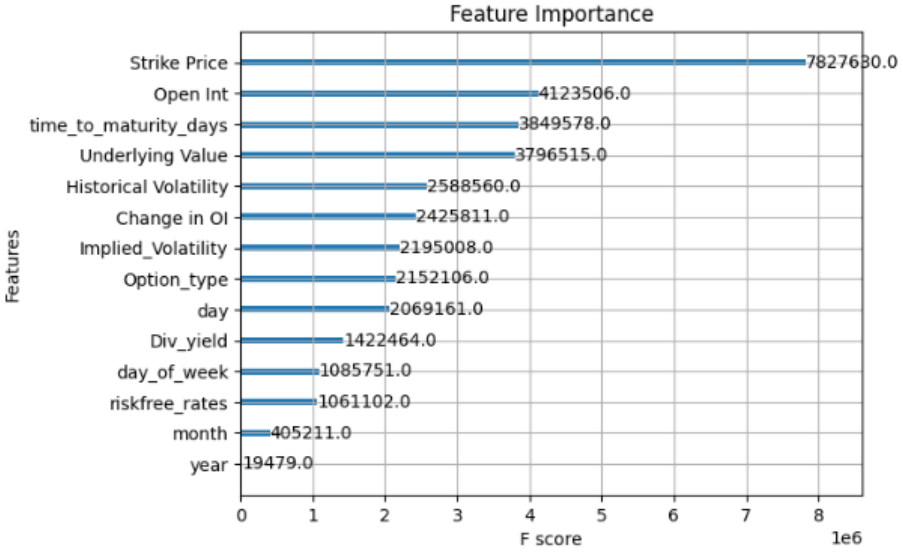


FIGURE 6.1: XGBoost Feature Importance

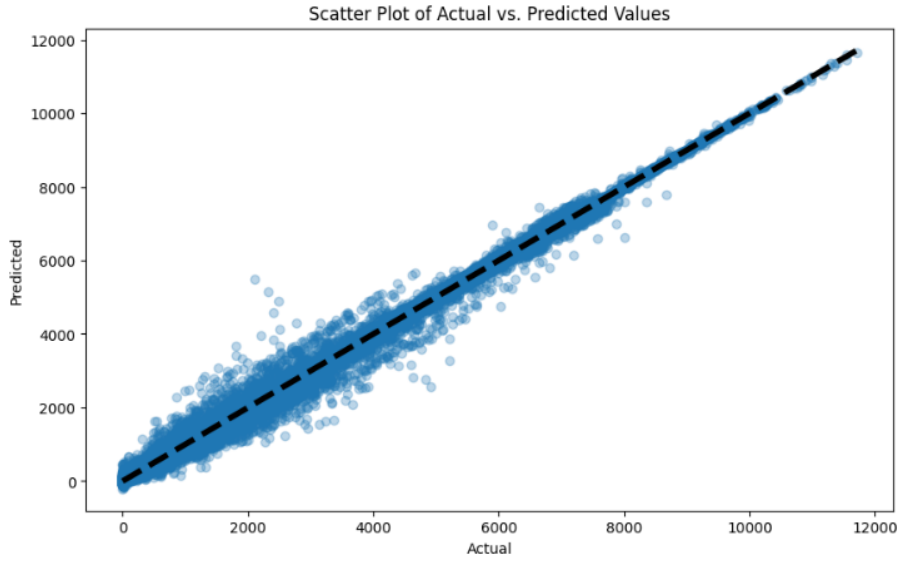


FIGURE 6.2: XGBoost Scatter Plot

In the XGBoost model, the *Strike Price* is an important factor. However, the *Open Int* and *Underlying Value* also have a significant impact. This means that the model is sensitive to both the market conditions of the option's underlying asset and the trading volume, which reflects market sentiment. *Historical Volatility* and *Change in OI* are not as important as *time_to_maturity_days*, but they still play a crucial role in

the model by taking into account the changes in the market over time. It seems that the *Option_type* is not very important, which could mean that the option's specific characteristics are used in a more detailed way when determining its price. Notably, the feature *day* also carries importance, pointing to the model's attentiveness to temporal effects beyond mere contract specifics.

6.1.2 Random Forest:

This model scored an MAE of 20.13 and an RMSE of 56.50. While it performed reasonably well, it was slightly outperformed by the XGBoost 15 model. The results suggest that the ensemble method of Random Forest is effective but may not capture the intricacies of option pricing as well as the more sophisticated XGBoost with a deeper tree.

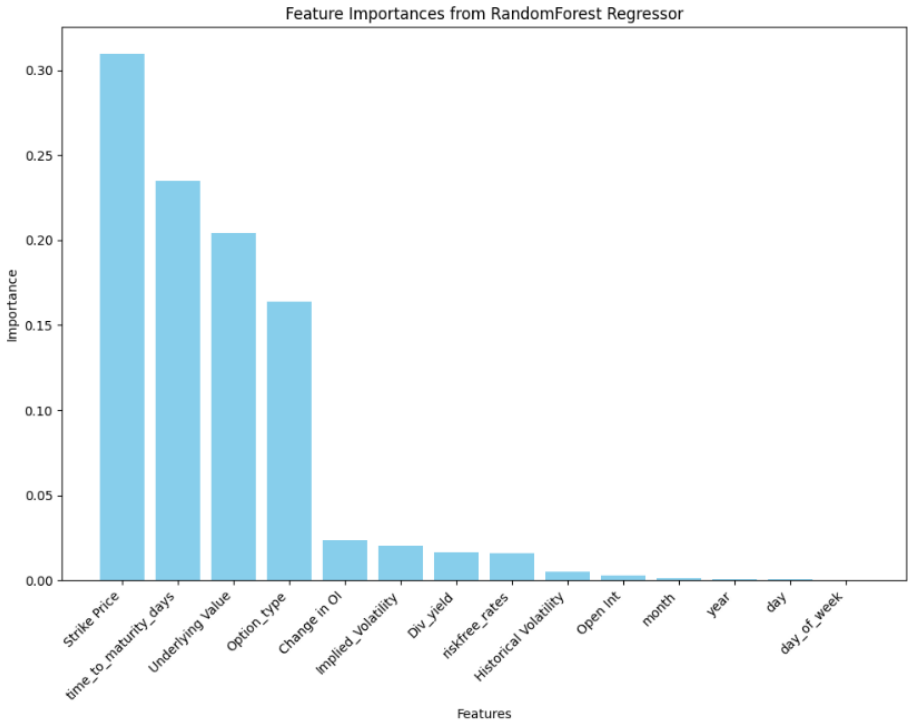


FIGURE 6.3: Random Forest Feature Importance

In the Random Forest model, *Strike Price* is the most important factor. *Time to maturity in days* and *Underlying Value* are also given significant importance, and it makes sense because these factors are fundamental in determining the premium of

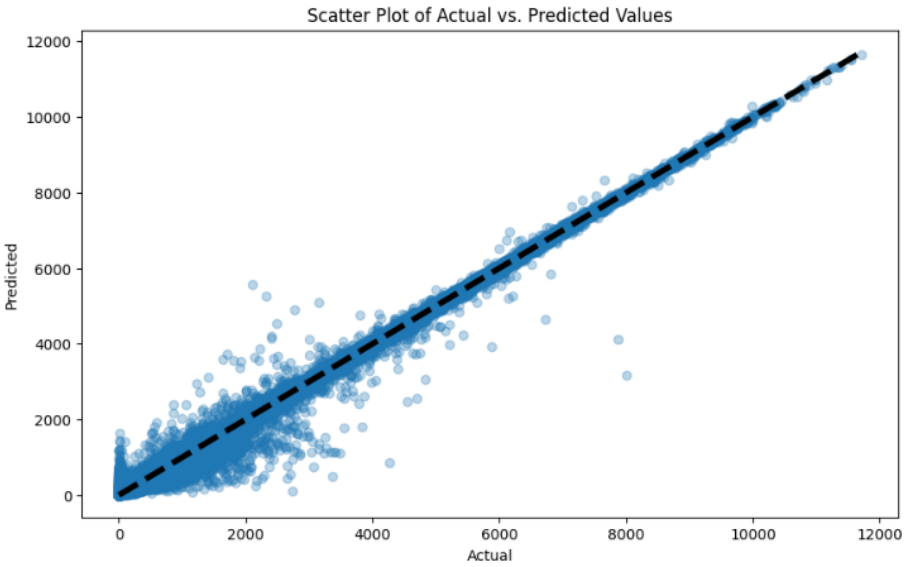


FIGURE 6.4: Random Forest Scatter Plot

an option. *Option_type* has a significant impact, indicating that the model is able to effectively distinguish between calls and puts. However, it seems that the model doesn't give as much importance to *Div_yield* and *riskfree_rates*. This suggests that the model is more sensitive to market and contract-specific factors rather than these economic factors.

6.1.3 XGBoost 5:

The shallower XGBoost model, with a maximum depth of 5, had an MAE of 23.72 and an RMSE of 48.03. Despite a higher MAE than its deeper counterpart and the Random Forest model, it achieved a lower RMSE, suggesting a better handling of outliers or extreme values in the dataset.

6.1.4 LightGBM 15:

Achieving the best performance among the models with an MAE of 16.05 and an RMSE of 37.32, LightGBM with a maximum depth of 15 excels in handling large datasets and complex features. Its superior performance can be attributed to its efficient handling of categorical features and gradient-based learning.

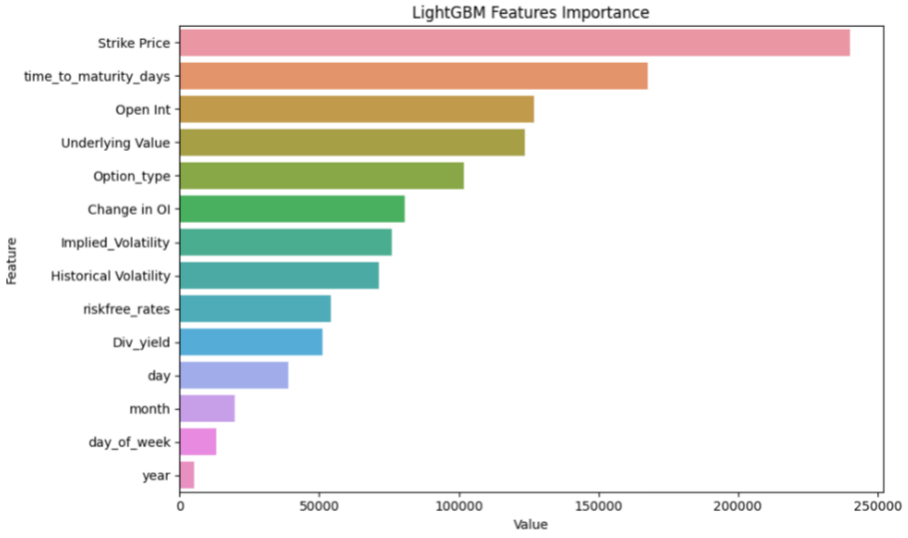


FIGURE 6.5: LightGBM Feature Importance

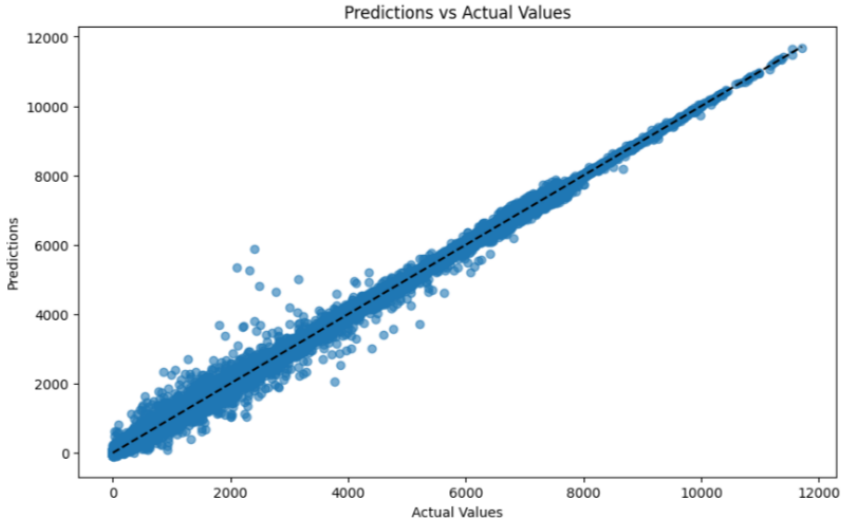


FIGURE 6.6: LightGBM Scatter Plot

The LightGBM model gives the highest importance to `Strike Price` and `time_to_maturity_days`, which is consistent with the pattern observed in the other models. The model seems to prioritize `Open Int` more, which is in line with its focus on market liquidity and investor sentiment. It's interesting to note that `Historical Volatility` is given more importance than the Random Forest in LightGBM's predictions. This suggests that market turbulence is taken into account more seriously. `Div_yield` and `riskfree_rates` have less influence, just like what we saw in the Random Forest model. This suggests that short-term market factors might be given

more importance than long-term economic indicators in this particular modeling scenario.

6.1.5 LightGBM 5:

The LightGBM model with a maximum depth of 5 showed an MAE of 21.61 and an RMSE of 40.66. It outperformed its XGBoost counterpart with the same depth in terms of RMSE but had a slightly worse MAE. The results show that while it may not be as accurate on average (MAE), it handles extreme errors (RMSE) better.

6.1.6 Black-Scholes Model:

As a traditional model, the Black-Scholes had the highest MAE and RMSE of 102.76 and 225.74, respectively. This significant difference in error metrics indicates that the Black-Scholes model, while foundational for option pricing theory, does not compete well with machine learning models when applied to actual market data, particularly in the complex Indian market environment.

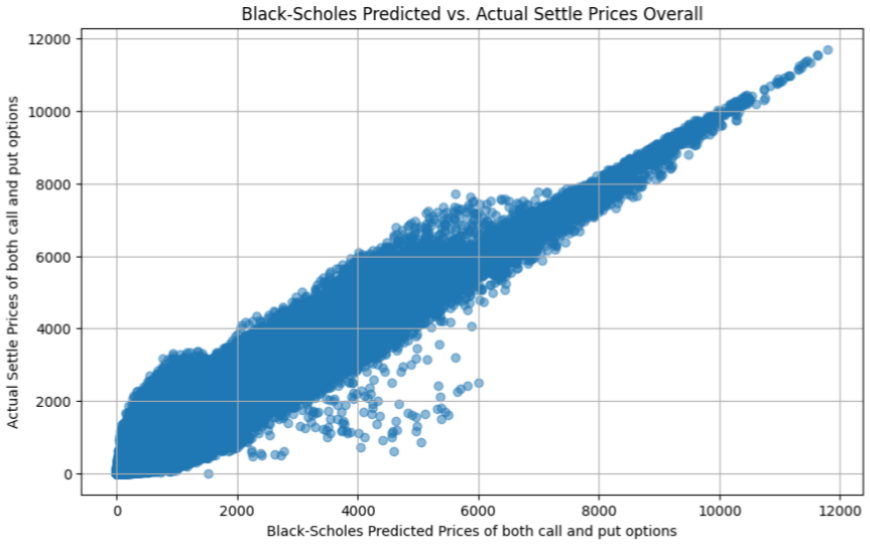


FIGURE 6.7: Black-Scholes Predicted vs Actual Price

The scatter plots depicting Black-Scholes model predictions versus actual prices for call and put options demonstrate several trends.

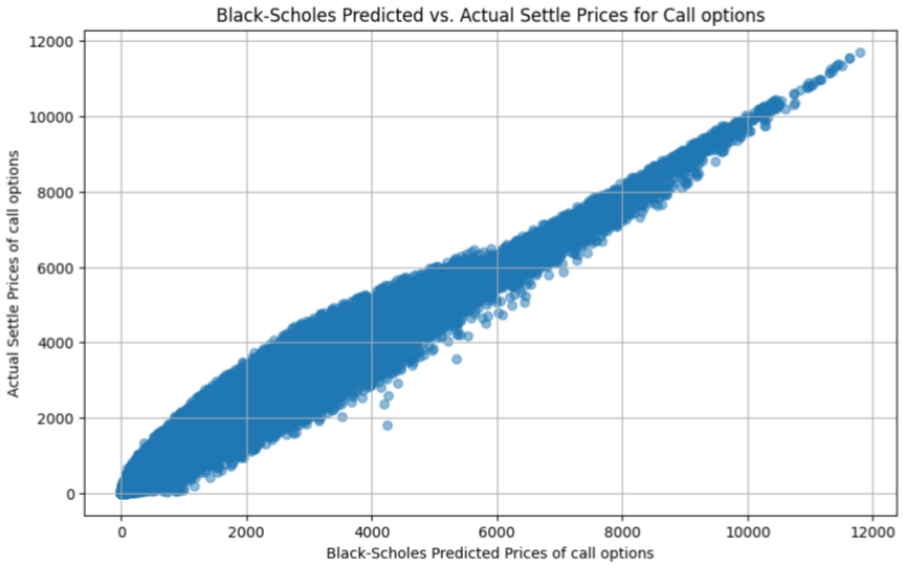


FIGURE 6.8: For Call Option

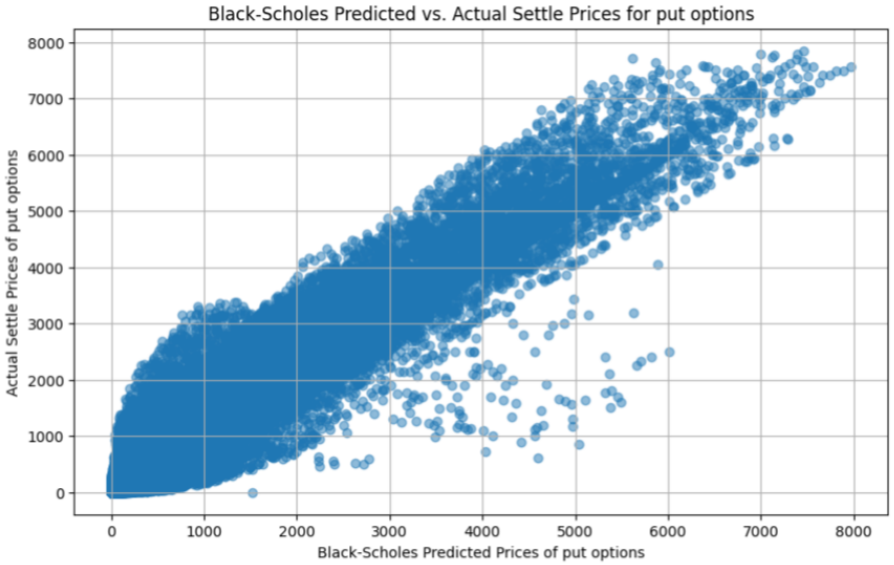


FIGURE 6.9: For Put Options

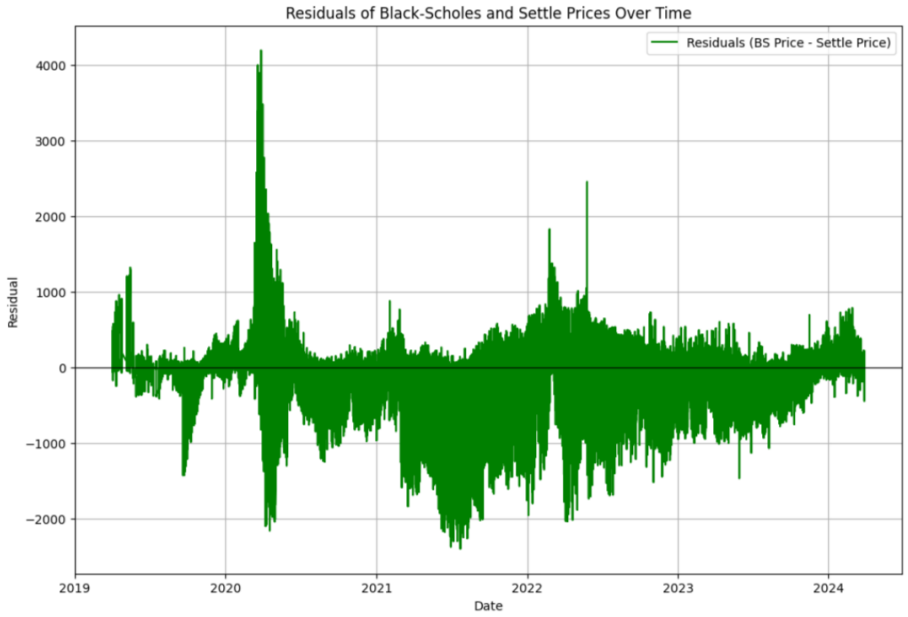


FIGURE 6.10: Residual Plot for both Options

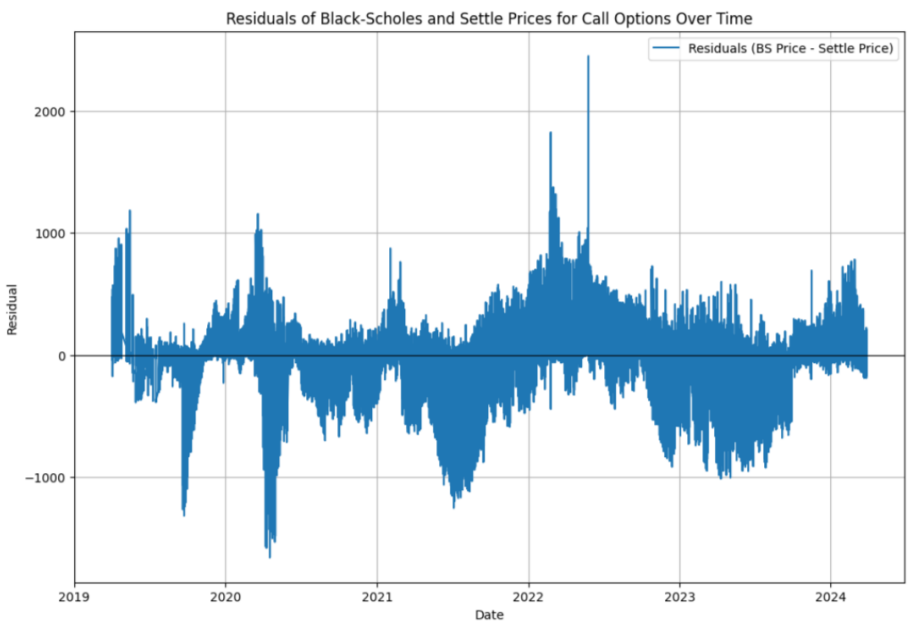


FIGURE 6.11: Residual Plot for Call Options

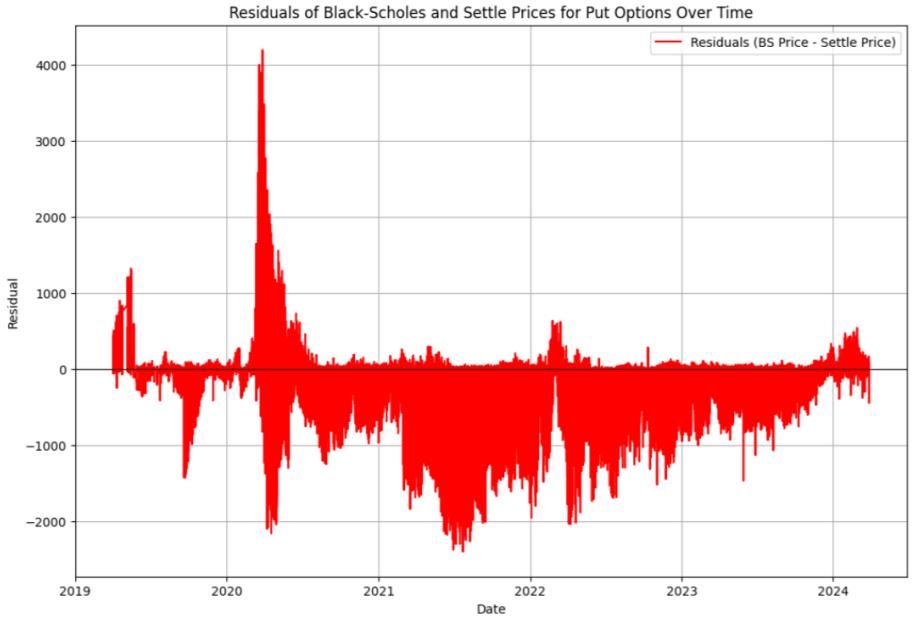


FIGURE 6.12: Residual Plot for Put Options

As the option prices increase, the number of options decreases, indicating fewer transactions at higher price levels but with better predictive accuracy by the Black-Scholes model. Particularly, the model shows a more accurate prediction for call options, where the points closely align along a near 45-degree line, suggesting a strong correlation between predicted and actual prices. In contrast, the plot for put options displays greater scatter and deviation from this line, indicating less predictive accuracy, especially at lower prices. Overall, the model's performance improves with higher-priced options, reflecting that while the Black-Scholes model generally underestimates option prices, it does so more uniformly for calls than puts, and its predictive consistency improves as the price of the options increases.

The residual plots for the Black-Scholes model illustrate significant limitations in pricing options, particularly under the volatile conditions of the Indian market. The model exhibits considerable bias and volatility in residuals, with put options showing particularly large underestimations, especially during market downturns like those observed in early 2020. The call options, while slightly better predicted, still reflect substantial inaccuracies around major market events, underscoring the model's failure to adapt to real-time market dynamics. These findings emphasize the need for more adaptive and flexible pricing models that can dynamically incorporate actual market conditions to improve accuracy and risk management.

Chapter 7

Conclusions

The analysis of the results clearly demonstrates the superiority of machine learning models over the traditional Black-Scholes model in predicting the prices of Indian options. The LightGBM 15 model stands out as the most effective model, suggesting that its algorithm is particularly well-suited for this application. The adaptive learning rate and tree-specific parameters likely contributed to its performance. These findings have important implications for practitioners and researchers in financial markets, advocating for the adoption of advanced machine learning techniques in option pricing models. The comparative performance also suggests that while depth of learning is a critical factor in model accuracy, the choice of algorithm and its inherent efficiency in handling the data are equally important.

Based on all three models, it is consistently found that the 'Strike Price' and 'time_to_maturity_days' are the most influential features. This emphasizes how important they are in any option pricing model. However, there are subtle differences that become apparent when looking at the secondary features. The Random Forest model seems to focus more on contract specifics and fundamental option characteristics. On the other hand, XGBoost pays attention to real-time market conditions in addition to foundational option attributes. LightGBM is a machine learning model that pays close attention to 'Open Int' and appears to be well-suited to understanding the current market sentiment. This could be because it is good at analyzing transactional data.

7.1 Future Research

The current study opens several avenues for future research in the realm of option pricing using machine learning models. Key areas of interest could include:

- **Feature Engineering**: Investigating additional features, such as macroeconomic indicators, market sentiment analysis from news articles and social media, could potentially enhance model performance.
- **Hyperparameter Optimization**: Further tuning of hyperparameters using methods such as grid search, random search, or Bayesian optimization with the increase of computational resources may yield improvements in model accuracy and robustness.
- **Ensemble Techniques**: Combining the predictions of various models through ensemble techniques like stacking, blending, or boosting could leverage the strengths of individual models to improve overall predictive power.
- **Deep Learning Approaches**: Exploring deep learning architectures like recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, could be beneficial for capturing the temporal dependencies in option pricing.

Bibliography

Suryoday Basak, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, and Sudeepa Roy Dey. 2019. Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance* 47 (2019), 552–567.

Fischer Black and Myron Scholes. 1973. The pricing of options and corporate liabilities. *Journal of political economy* 81, 3 (1973), 637–654.

Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

J.James Chen. 2024. Introduction to Options Trading. Website. `https://www.investopedia.com/terms/o/option.asp` Accessed: April 29, 2024.

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

Lirong Gan, Huamao Wang, and Zhaojun Yang. 2020. Machine learning solutions to challenges in finance: An application to the pricing of financial products. *Technological Forecasting and Social Change* 153 (2020), 119928.

Mark B Garman and Michael J Klass. 1980. On the estimation of security price volatilities from historical data. *Journal of business* (1980), 67–78.

Adrian Gepp and Kuldeep Kumar. 2015. Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science* 54 (2015), 396–404.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33, 5 (2020), 2223–2273.

David J Hand. 2007. Principles of data mining. *Drug safety* 30 (2007), 621–622.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

John C Hull and Sankarshan Basu. 2016. *Options, futures, and other derivatives*. Pearson Education India.

ICICI Direct. 2023. Volatility Index (VIX) Effect on Options Pricing. https://www.icicidirect.com/ilearn/futures-and-options/articles/what-is-volatility-index-how-it-impacts-options-pricing. [Online; accessed 2024-04-30].

Codruț-Florin Ivașcu. 2021. Option pricing using machine learning. *Expert Systems with Applications* 163 (2021), 113799.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

JD Jayaraman, Xiaodi Zhu, and Ahmad A Rabaa'i. 2022. An Evaluation of Data Driven Machine Learning Approaches to Option Pricing. *Proceedings of the Northeast Business & Economics Association* (2022), 45–48.

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

Jiaming Liu, Chengzhang Li, Peng Ouyang, Jiajia Liu, and Chong Wu. 2023. Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. *Journal of Forecasting* 42, 5 (2023), 1112–1137.

Robert C Merton. 1973. Theory of rational option pricing. *The Bell Journal of economics and management science* (1973), 141–183.

NSE India. [n. d.]. NIFTY 50 Live | NSE Nifty 50 Index Today - NSE India. https://www.nseindia.com/products-services/indices-nifty50-index. [Online; accessed 2024-04-30].

Min Sue Park, Hwijae Son, Chongseok Hyun, and Hyung Ju Hwang. 2021. Explainability of machine learning models for bankruptcy prediction. *Ieee Access* 9 (2021), 124887–124899.

Hongyi Qian, Baohui Wang, Minghe Yuan, Songfeng Gao, and You Song. 2022. Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. *Expert Systems with Applications* 190 (2022), 116202.

Johannes Ruf and Weiguan Wang. 2019. Neural networks for option pricing and hedging: a literature review. *arXiv preprint arXiv:1911.05620* (2019).

Y. Shah. 2023. Trading Index Options. https://www.samco.in/knowledge-center/articles/trading-index-options/.

Kim Long Tran, Hoang Anh Le, Thanh Hien Nguyen, and Duc Trung Nguyen. 2022. Explainable machine learning for financial distress prediction: evidence from Vietnam. *Data* 7, 11 (2022), 160.

Miao Wang, Yunfeng Zhang, Chao Qin, Peipei Liu, Qiuyue Zhang, et al. 2022. Option pricing model combining ensemble learning methods and network learning structure. *Mathematical Problems in Engineering* 2022 (2022).