Machine Learning and European Call Option Pricing Beyond Black-Scholes

Course: DSO 530
Group 19
Members: Minjoo Sung_8386157617, Shun Yeung Ng_7272502833, Daniel Wong Ortiz_2870959937, Yongzhu Liang_5122143519, Yichen Zhang_8581720671, Kumail Abbas _8169806771
Contact Email: shunyeun@usc.edu

Executive Summary

Our project explores the viability of machine learning (ML) methods over the traditional Black-Scholes model for predicting European call option prices on the S&P 500. The Black-Scholes formula, a cornerstone of financial valuation, relies on assumptions often at odds with the realities of market behavior. We set out to determine whether a more nuanced ML approach could provide greater accuracy in the valuation of these options, which are critical for financial strategy and risk management.

Within the scope of our analysis, we processed two datasets—the training set comprising 5,000 options with known outcomes, and the test set featuring 500 options for predictive validation. Our methodological framework was two-pronged: a regression analysis aimed at predicting the current option value (Value), and a classification analysis evaluating the pricing accuracy of the Black-Scholes model (BS).

Our findings reveal that ML models have a strong potential to surpass the Black-Scholes formula in predictive accuracy. This is particularly evident in volatile market conditions, where the adaptability of ML models to new data proves advantageous. Moreover, the ability to interpret ML model predictions can offer valuable insights into the factors driving option prices, which is of paramount importance for strategic decision-making in finance.

The only dataset we are using for this analysis is option_train. It has four features (S, K, tau, and r) and two response (value, BS) variables. S is the current asset value. K represents the strike price of an option. R is the annual risk-free interest rate. Tau is the time to maturity in years. For two responses, value is the current option value, and BS is the Black-Scholes formula. We used value as the response variable and all four features for regression analysis. For the classification problem, we first convert Black-Scholes to a binary variable; ‘1’ means over and ‘0’ otherwise. We also checked for missing data by using the ‘data.isnull() function’. However, there is no missing data.

In conclusion, our exploration into the use of statistical and machine learning methods revealed a significant potential for ML applications in option pricing—a domain traditionally dominated by mathematical models. This project illustrates the evolving landscape of financial analytics, where data-driven models are poised to enhance and perhaps redefine the standards of option valuation.

Regression Analysis

In our project, we explored six regression models to predict the current value of European call options on the S&P 500, using 70% of the dataset for training and 30% for testing, supplemented by 5-fold cross-validation. This approach ensured robustness and helped mitigate overfitting.

The regression models we examined included:

1). Logistic Regression
2). K-Nearest Neighbors (KNN)
3). Decision Tree
4). Random Forest
5). Gradient Boosting
6). XGBoosting

As shown from the table below, all of our regression models exhibit high R-squared values above 0.9, indicating that each model generally predicts the target variable effectively. Thus, we focus on RMSE values for our final selection, identifying the model with the least prediction error.

Linear Regression is not suitable for our complex and non-linear datasets as evidenced by the highest RMSE of 34.30. K-Nearest Neighbors (KNN) and Decision Tree offer improvements in RMSE but also face challenges. KNN struggles with computational efficiency in large datasets, and Decision Trees are prone to overfitting, limiting their practical application without rigorous tuning. While Random Forest and XGBoost effectively address the issues of overfitting and scalability, and generally achieve lower RMSEs, Gradient Boosting still emerges as the best choice. With the lowest RMSE of 6.62, it clearly excels at handling complex predictive tasks, demonstrating its superiority in precision and error reduction. This consistent effectiveness makes Gradient Boosting the preferred choice for our project, especially when tackling intricate datasets like those involving European call options on the S&P 500. Additionally, we decided against using LASSO and Ridge regression since our dataset contains only four features, making these models less appropriate due to their preference for larger datasets.

In conclusion, Gradient Boosting is the optimal model for our project predicting European call options on the S&P 500. Its superior accuracy and robust error correction make it ideally suited for our data, leading us to choose it for its reliability in complex scenarios and alignment with our high standards for financial modeling. This choice confirms Gradient Boosting as the best fit for our analytical needs.

Regression	Out-of-Sample R square	RMSE
Linear Regression	0.924	34.30
KNN	0.947	28.38
Decision Tree	0.991	11.43

Random Forest	0.996	7.66
Gradient Boosting	0.997	6.62
XGBoosting	0.983	16.10

*As highlighted in the table, Gradient Boosting is the best model we choose

Classification Analysis

We conducted an extensive analysis of the Black-Scholes model for valuing European call options on the S&P 500. We employed a 70% training, 30% testing split, and 5-fold cross-validation to ensure robust model evaluation and prevent overfitting.

The classification models we assessed are:

1). Logistic Regression
2). K-Nearest Neighbors (KNN)
3). Decision Tree
4). Random Forest
5). Gradient Boosting
6). XGBoosting

As suggested from our table below, we observed that Logistic Regression and K-Nearest Neighbors (KNN) exhibited higher classification errors of 0.109 and 0.089 respectively, showing limitations in accurately detecting true positives. Decision Trees improved slightly with an error of 0.086 but were prone to overfitting. Random Forest reduced the classification error to 0.078 by mitigating overfitting through ensemble averaging, although it still did not achieve the best error balance.

Among the models, both Gradient Boosting and XGBoost emerged as the top performers with the lowest classification errors of 0.067, effectively balancing false positives and negatives. Given the choice between the two, Gradient Boosting was selected due to its superior interpretability and computational efficiency, making it the optimal choice for our project on modeling European call options on the S&P 500. This decision underscores our commitment to leveraging sophisticated analytical techniques that provide both high accuracy and practical usability in financial modeling, ensuring a robust and efficient approach.

Classification Model	Classification Error	Confusion Matrix
Logistic Regression	0.109	[[1109, 65], [100, 226]]
KNN	0.089	[[1094, 80], [55, 271]]

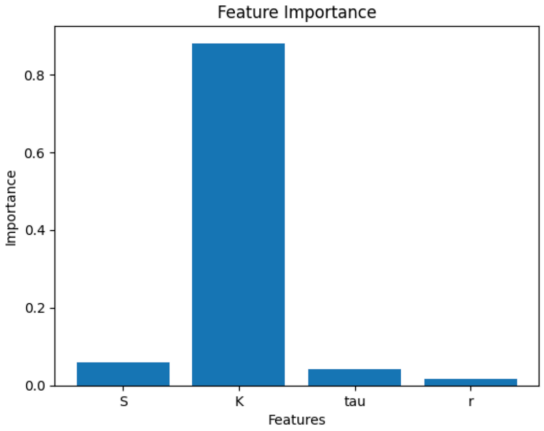
Decision Tree	0.086	[[1095, 79], [50, 276]]
Random Forest	0.078	[[1120, 54], [63, 263]]
Gradient Boosting	0.067	[[1125, 49], [52, 274]]
XGBoosting	0.067	[[1118, 56], [47, 279]]

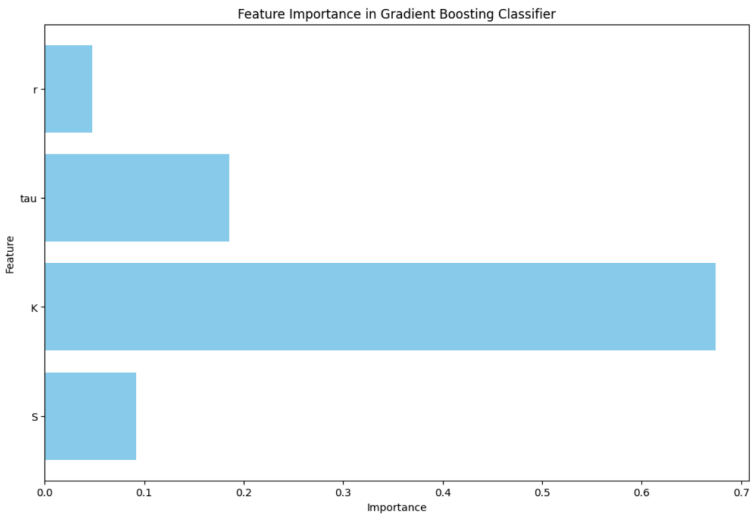
*As highlighted in the table, Gradient Boosting is the best model we choose

Business Implementations

From a business perspective, we concluded that regression model accuracy is more important since financial decisions, which involve large sums of money, are based on strong and supportive predictions, so accuracy will be a key part of them. As for classification models, interpretation is more important for decision makers to understand why the model deviates, and under which conditions can give us insights on how to improve the model and adjust it contextually.

We believe machine learning models will theoretically outperform Black-Scholes since they can capture non-linear relationships from financial market factors that interact in complex ways. They also make flexible assumptions, they do not assume constant factors and do learn from previous data, therefore identifying unusual conditions such as, for example, outliers. These models can also make use of extra features, parameter tuning, and cross validation to better generalize unseen data. In an ideal scenario, machine learning would unequivocally excel as the superior model for financial analysis; however, practical limitations persist. These include the necessity for substantial investments to develop and implement such models, along with the considerable challenge of making their complex methodologies comprehensible and transparent to the financial community.





We would suggest keeping four variables, which are current asset value, strike price, annual risk-free interest rate, and time to maturity in years, on both of our classification and regression models, even though ML model performance might not be affected significantly as long as K - strike price is secured as the key predictor (scored 0.85 for the GB regressor and 0.67 for the GB classifier with sklearn feature_importance) in terms of feature's purity and frequency. The fact that all four features are accessible and almost free to extract would not make any business benefit from collecting any fewer of them.

In order to use our trained model to confidently make predictions on Tesla stock options, we think that it might be feasible, but it will require users to adopt certain procedures to generate business outcomes and be statistically convincing. First, testing data needs to be reassessed through testing confidence intervals with a broader S&P 500/high-tech sector population to reduce overfitting risks on the model, since the Tesla option can be extremely volatile. We recommend collecting sophisticated training data that can better represent TSLA movement and momentum. Second, more features have to be considered in order to live up to the sophisticated and dynamic investing environment. Other trading features like trading volume, stock volatility, and media news can heavily lift or punish options pricing through supply and demand, which our current model did not include, thus failing to capture and explain when those features actually make a difference in option pricing. Third, design a pipeline that ensures the model can be evaluated, tuned, and updated properly, ideally on a daily basis. The unbeaten champion (model) does not exist in the real financial market world because, at the end, humans made their decisions (on purchase and sell) and game theory to compete with each other, which is something that is extremely hard to quantify in statistical models; therefore, the pricing will go off from what the machine previously predicted. But updating the model daily can help you better react, adjust to the most current market dynamics, and make better predictions.

Conclusion

Our study finds that machine learning techniques excel in predicting the price of European call options on the S&P 500 index compared to Black-Scholes models. Among all the ML models evaluated, the Gradient Boosting model stood out, showing accuracy and adaptability under different market conditions. It outperforms all other attempted models in terms of regression and classification metrics. These findings suggest that it is providing a more accurate option valuation. ML has considerable potential to reshape financial analysis, and option valuation is critical for informed financial strategy and risk management. However, the complexity of these ML models requires continuous efforts to improve their interpretability and rigorous validation to ensure their effective integration into financial practice.

Appendix

Regression Analysis:
https://drive.google.com/file/d/1bmvUvOGFARymkVud0cfK7cYY-kTW9ZPw/view?usp=sharing

Classification Analysis:
https://drive.google.com/file/d/18KdrsWBGxZuISpCTK2jOygirqh9wdp42/view?usp=sharing

Prediction:
https://colab.research.google.com/drive/1qNLxi-7R7lGSLbmgRRRcDEVV_UnGBLYe?usp=sharing