Option Pricing Group Project

Group 27: Eloise Yu, Jessie Liu, Hee Jae Yoon, Scott Alexander, Leo Zeng

University of Southern California, Marshall School of Business

DSO 530: Applied Modern Statistical Learning Methods

Professor Xin Tong & Professor Paromita Dubey

April 19, 2024

Executive Summary

This report presents an analysis of European call option pricing data on the S&P 500, aiming to predict option value and BS (classification of overestimating or underestimating our option value) using a range of statistical and machine learning techniques. A thorough exploratory data analysis was conducted to understand dataset patterns, trends, and relationships. Data standardization, normalization, and outlier detection were performed to ensure data quality. Outliers, erroneous entries, and heavy correlations were identified and addressed, and binary identifiers for overvaluation and undervaluation were created. Using k-fold cross-validation, out-of-sample R-squared and classification errors were used to predict Value and BS accurately. Regression models were evaluated for predicting option value, including Linear Regression, KNN Regression, LASSO, RIDGE, Decision Tree, and Random Forest. Random Forest emerged as the optimal choice with a mean out-of-sample R-squared value of 99.69%. Logistic Regression, KNN Classification, Decision Tree, Random Forest, SVM, AdaBoost, XGBoost, and a neural network approach were considered for BS classification. XGBoost was selected due to achieving the lowest mean classification error of 5.95%. Ultimately, Random Forest and XGboost provide the most accurate approach for value and BS predictions on the 500 options of the test data set. Throughout this analysis, some key understandings to note are prediction accuracy was prioritized over interpretability and including all four predictor variables enhances model predictability. Additionally, machine learning models may outperform the Black-Scholes model by incorporating diverse variables, learning from historical data, and adapting to changing market conditions.

Review of Approaches Attempted

Exploratory Data Analysis (EDA) and Wrangling

Before formal modeling commences, understanding the patterns, trends, and relationships within a dataset is crucial. The team first gathered information regarding the variables in the training data set and ensured the data were on the same scale. After assessing the need for standardization and normalization, the members checked missing data in the training set and concluded that the data set had no missing data. From there, the members proceeded to find any erroneous entries in the training set by using the following methods: 1) ensure each data entry matches the expected data type, 2) plot boxplots and Interquartile Range (IQR) on X, K, tau, and r to check for outliers, 3) plot scatterplots between the predictor variables and the independent variable *Value* to prepare for regression analysis, 4) generate a correlation matrix to check for interactions between predictor variables. The study discovered that there are some outliers in each predictor and that *S* and *r* are heavily correlated. To address these issues, the team filtered out data outside 3 times the IQR range in case it becomes relevant in model selection. Lastly, the team converted the columns *Under* and *Over* into identifiers with 1 denoting overvaluation and vice versa.

Model Selection

The team utilizes k-fold cross-validation to assess the proposed predictive models' effectiveness. To accurately predict the variable *Value* and measure the effectiveness of the models, the team decided to compare the out-of-sample R-squared (CV Score) among Linear Regression, KNN Regression, LASSO, RIDGE, Decision Tree, and Random Forest models. The team started off with 4 different Linear Regression models: one using the unfiltered training data, one using the filtered training data, one excluding the variable *S*, and one excluding the variable *r* to account for the discoveries in the EDA process. Since the outputs indicated extremely similar CV Scores for the 4 models, the team decided to proceed with the original, unfiltered training data when assessing future models.

The team made the following assumptions for the regression models for the best possible result: 1) The team proceeded with 10 splits for k-fold CV; 2) Considering that the training set has only 5000 lines of data, the KNN Regression model was set to predict the output based on 5 nearest neighbors; 3) To facilitate the best possible CV Scores for LASSO and RIDGE, the team standardized the data and tuned for an optimal alpha; 4) While pruning would be ideal for

choosing the optimal alpha in Decision Tree, the calculation of the cost-complexity pruning path suggested 3683 possible alphas, which was not computationally possible on the team members' devices and might increase the risk of overfitting; Thus, the default alpha of 0 was used; 5) For the *max_features* parameter in Random Forest, the team experimented with 2, 3, and 4 features and decided that 3 features would product the best CV Score (the default of m = p/3 is too small).

The team initially compared the classification error (CV Score) among Logistic Regression, KNN Classification, Decision Tree, Random Forest, and Support Vector Machine (SVM) models. The following assumptions were made for the classification models for the best possible result: 1) The team proceeded with 10 splits for stratified k-fold CV; 2) The Logistic Regression model uses 10,000 as iteration; 3) The KNN Classification model was based on 5 nearest neighbors; 4) Pruning for the Decision Tree model was conducted here for an optimal alpha; 5) The SVM uses linear kernel. Since the above models covered in class did not fully meet the team's expectation for classification error, two boosting methods, AdaBoost and XGBoost, and one deep learning method, Neural Network, were further explored for classification.

Final Approaches

After conducting cross-validation on multiple statistical models, the group members evaluated out-of-sample R squared value and classification error to choose the two final approaches to predict Value and the BS. Among regression models, the members opted for random forests, which are bagged decision tree models that split on a random subset of features on each split. This randomness helps to de-correlate the trees within the ensemble, thereby reducing overfitting and improving generalization performance. Consequently, random forests achieved a mean out-of-sample R-squared value of 99.69%, the highest accuracy in predicting Value among all other methods.

As for classification models, the members chose XGBoost (Extreme Gradient Boosting), an advanced implementation of boosting algorithms designed to optimize model performance. XGBoost employs a gradient-boosting framework to train weak predictive models and minimize the loss function sequentially. As a result of this approach, it reached the lowest mean classification error of 5.95% in predicting BS.

Conclusion

In addressing prediction problems, both accuracy and interpretability hold significant importance. However, in this analysis, a prioritization was given to prediction accuracy. We

opted for Random Forest for regression, primarily because it offers a superior balance between high prediction accuracy and relatively good interpretability compared to other models like AdaBoost and XGBoost. Furthermore, it showed only a slight difference in the R-squared value. This decision exemplifies the inherent trade-off between prediction accuracy and interpretability, where enhancing one aspect can often lead to compromises in the other. Similarly, XGBoost was utilized for classification since it yields a lower classification error, providing better prediction accuracy.

Machine learning models may outperform the Black-Scholes model in predicting option values due to their ability to process and learn from historical data and adapt to new information, therefore continuously improving their predictive accuracy. Unlike Black-Scholes, which is based on fixed assumptions such as constant volatility and a log-normal distribution of stock prices, machine learning models can incorporate a much wider array of variables, including not only the historical price data, but also factors that influence market dynamics such as economic conditions, political events, and changes in market sentiment. This adaptability and enhanced accuracy make machine learning models particularly effective in financial markets, where conditions are constantly changing and precision in option valuation is crucial for effective risk management and financial decision-making.

Including all four predictor variables in our prediction models is beneficial for several reasons. First, they comprehensively represent the market, with each variable capturing essential aspects of financial theory and option pricing. For instance, the current asset value (S) directly impacts the option's intrinsic value, while the strike price (K) determines its money status, influencing pricing significantly. The annual interest rate (r) affects the cost of money and the option's discounted strike price at expiry, and time to maturity (T) represents the time value of money and the potential for favorable price movements. Secondly, using all variables enhances model predictability and reliability, capturing more variability and improving forecast accuracy. Omitting any variable risks critical information loss, potentially leading to underfitting and less effective stakeholder communication.

When considering using a trained model to predict Tesla stock options, it is important to acknowledge the implications of using historical data that may not accurately reflect current economic conditions or market compositions. The inability to adjust for inflation over time means that financial models might not maintain relevance or accuracy as economic conditions

evolve. Moreover, the fact that Tesla did not exist during the original data collection period highlights a significant issue in data suitability and model applicability. Lastly, the study is limited by the fact that it does not utilize time-series data, which restricts the accuracy of predictions based on the analysis. Market conditions are known to fluctuate frequently and unpredictably, which may further diminish the reliability of our predictive outcomes. Thus, caution should be exercised when applying these findings to real-world scenarios where temporal dynamics are crucial.