EXPLORING EUROPEAN CALL OPTION PRICING DATA ON S&P 500

CONTEXT OVERVIEW

EUROPEAN CALL OPTION

- Gives holder the right (but not the obligation) to purchase an asset at a given time for a given price
- Valuing such an option is tricky as it depends on the future value of the underlying asset.

1Y S&P Stock Index

5Y S&P Stock Index

Overview of Variables

Value: Current option value

S: Current asset value

K: strike price of an option

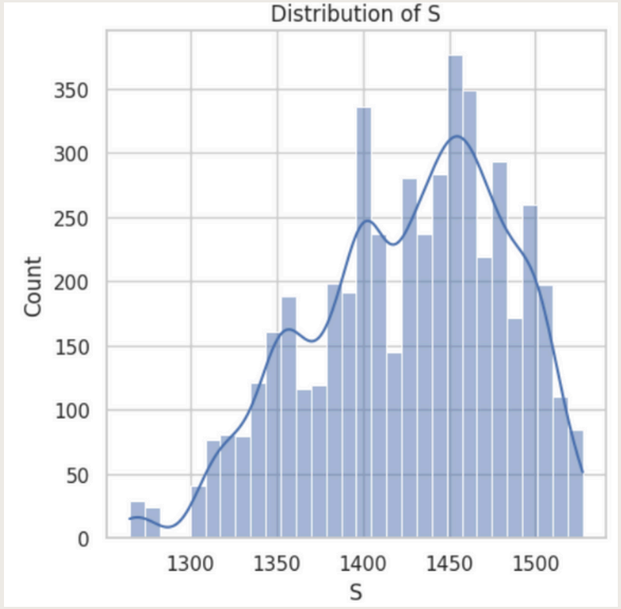
r: annual interest rate

tau: time to maturity in years

BS: The Black-Scholes formula was applied to this data to get C_pred. and If an option has C_pred – C > 0, i.e., the prediction over estimated the option value, we associate that option by (Over); otherwise, we associate that option with (Under).

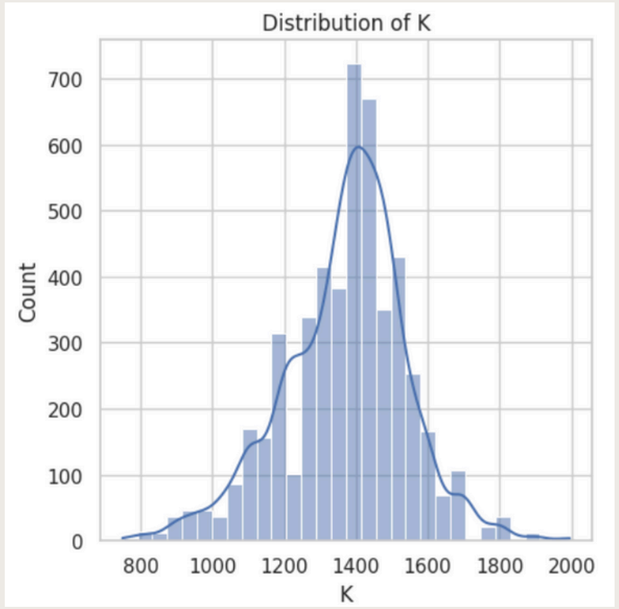
EXPLORATORY DATA ANALYSIS

Normal Distribution



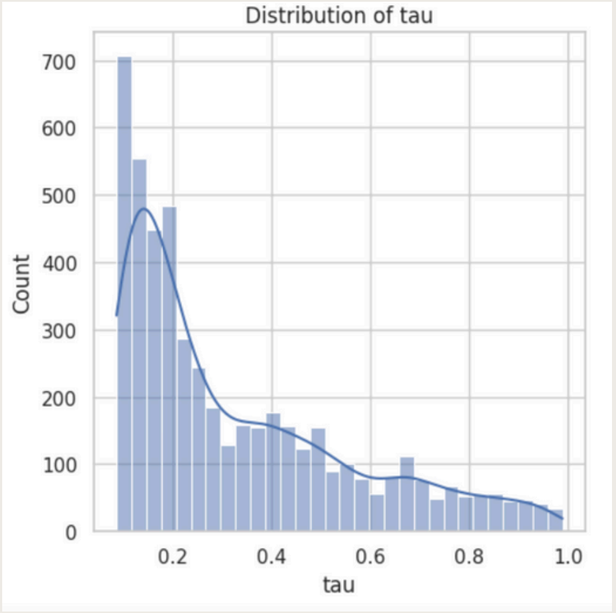
S

Slightly right skewed
Most assets are valued around the central range but with a few assets having higher values



K

Nearly normal distribution
Strike prices are around the mean value, and the likelihood of options withith prices far than the mean is symetrically decreasing

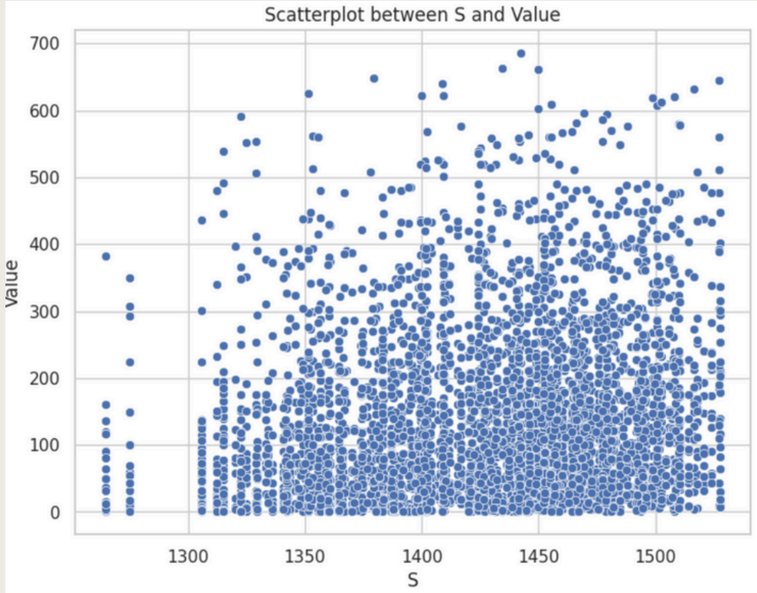


tau

Right skewed
A majority of the options have shorter times until maturity

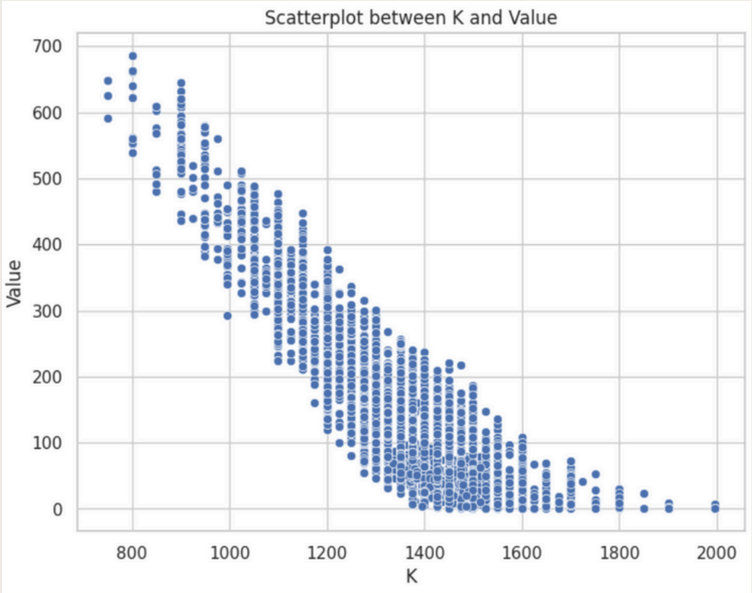
EXPLORATORY DATA ANALYSIS

Scatterplot



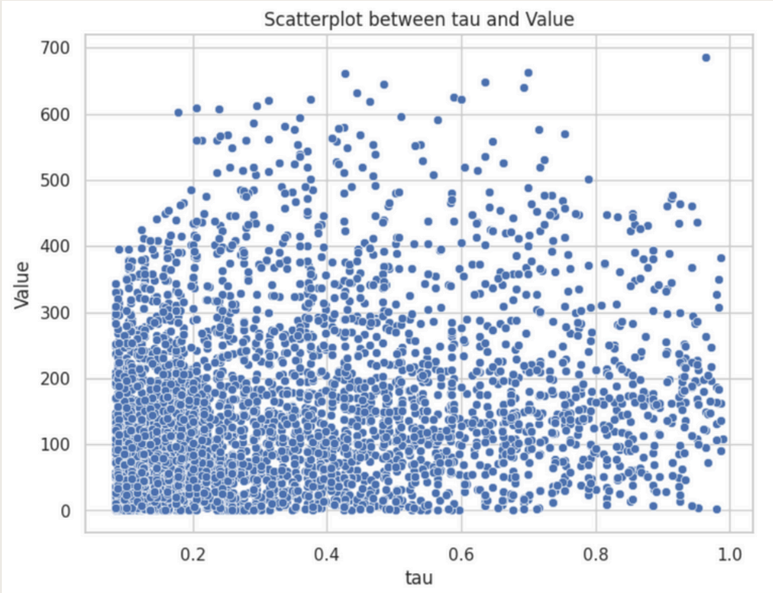
S

No simple linear relationship
Slight indication that as S increases, the range of Value also increases, especially after an S value of around 1450.



K

Non-linear, possibly exponential decrease in Value as K increases.
Value decreases sharply when K is low and then levels off as K increases.

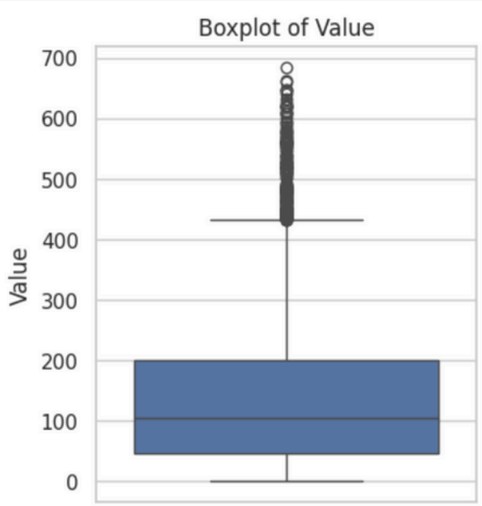


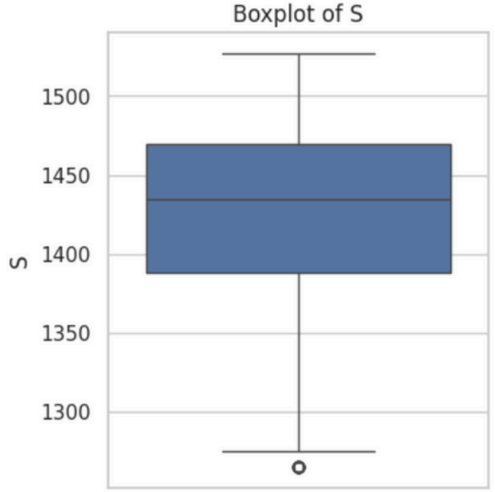
tau

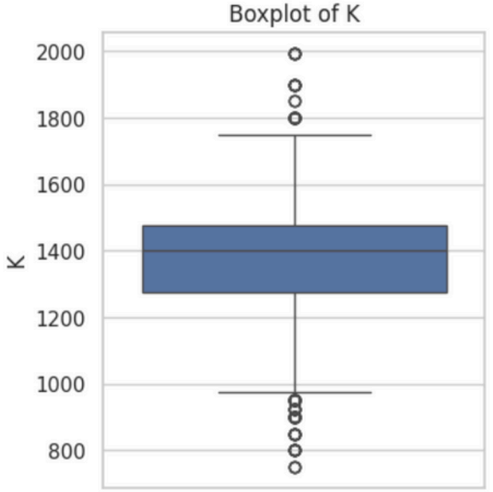
Right skewed
A majority of the options have shorter times until maturity

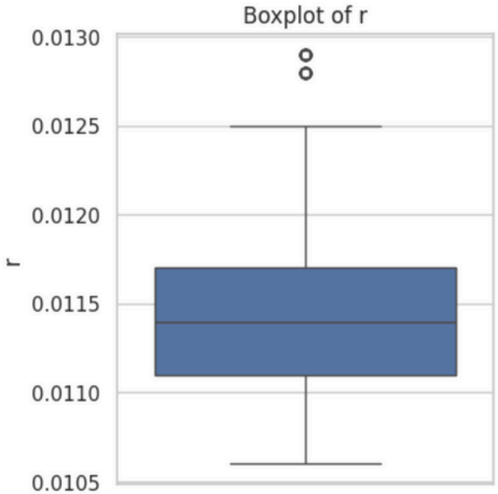
EXPLORATORY DATA ANALYSIS

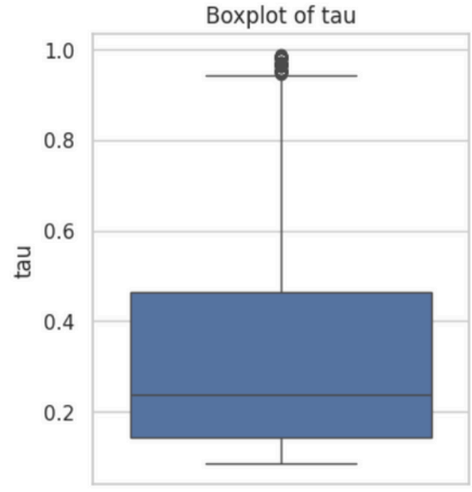
Boxplots











Value

Right skewed distribution
Significant numbers of outliers above the upper whisker, meaning unusually high Option values

S

Relatively symmetric distribution around the median
More outliers on the lower value

K

Left skewed distribution
More outliers on the lower value, meaning there are more lower strike price value

r

Mostly concentrated around the median value with relatively low variance
Few outliers above the upper whisker

tau

Skewness towards lower values
More notable outliers on the upper end, indicating more options with higher time to maturity

INSIGHTS FROM OUR EXPLORATORY DATA ANALYSIS

1. **Significant Outliers in higher end of Value (C)**
Given that **Value** is the actual current option value and **C_pred** is the predicted option value by Black-Scholes formula, these outliers may indicate instances where the Black-Scholes formula may not be accurate, or there are anomalies in the market data.

2. **Distribution of S and K**
The variables **S (current asset value)** and **K (strike price of the option)** have a considerable impact on the **Value** and will be used as predictors in your ML models. The boxplot analysis showed outliers on both ends for S and a left-skewed distribution for **K**, which could affect the predictive models. Therefore, we implemented transformation methods to mitigate the influence of these outliers.

3. **Annual interest rate (r)**
The distribution for the **annual interest rate r** showed a tight interquartile range but with some significant outliers. Since interest rates generally have less variability, these outliers might be errors or rare events which need to be investigated during that year.

Transformation of Variables

TRANSFORMATION OF VARIABLES

- To normalize distribution of variables
- Stabilization of Variance
- Interpretability

	Unnamed: 0	Value	S	K	tau	r	BS	S_log	K_log	tau_log	r_log	S_ln	K_ln	tau_ln	r_ln
0	1	348.500	1394.46	1050	0.128767	0.0116	Under	7.240263	6.956545	-2.049750	-4.456750	7.240979	6.957497	0.121126	0.011533
1	2	149.375	1432.25	1400	0.679452	0.0113	Under	7.267002	7.244228	-0.386469	-4.482953	7.267700	7.244942	0.518468	0.011237
2	3	294.500	1478.90	1225	0.443836	0.0112	Under	7.299054	7.110696	-0.812301	-4.491842	7.299730	7.111512	0.367303	0.011138
3	4	3.375	1369.89	1500	0.117808	0.0119	Over	7.222486	7.313220	-2.138697	-4.431217	7.223215	7.313887	0.111370	0.011830
4	5	84.000	1366.42	1350	0.298630	0.0119	Under	7.219949	7.207860	-1.208549	-4.431217	7.220681	7.208600	0.261310	0.011830
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	4996	325.250	1465.15	1175	0.424658	0.0111	Under	7.289713	7.069023	-0.856472	-4.500810	7.290395	7.069874	0.353931	0.011039
4996	4997	36.000	1480.87	1480	0.101370	0.0111	Over	7.300385	7.299797	-2.288979	-4.500810	7.301060	7.300473	0.096555	0.011039
4997	4998	90.000	1356.56	1500	0.673973	0.0120	Under	7.212707	7.313220	-0.394566	-4.422849	7.213444	7.313887	0.515200	0.011929
4998	4999	175.875	1333.36	1200	0.309589	0.0122	Under	7.195457	7.090077	-1.172510	-4.406319	7.196207	7.090910	0.269713	0.012126
4999	5000	106.375	1480.87	1475	0.504110	0.0111	Under	7.300385	7.296413	-0.684962	-4.500810	7.301060	7.297091	0.408201	0.011039

Regression

Regression

LINEAR REGRESSION

Using Ordinary Least Squares Method

OLS Regression Results

Dep. Variable:	Value	R-squared:	0.958
Model:	OLS	Adj. R-squared:	0.958
Method:	Least Squares	F-statistic:	3.795e+04
Date:	Wed, 17 Apr 2024	Prob (F-statistic):	0.00
Time:	20:10:29	Log-Likelihood:	-23319.
No. Observations:	5000	AIC:	4.665e+04
Df Residuals:	4996	BIC:	4.667e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5614.3591	20.947	268.025	0.000	5573.294	5655.425
S	0.6136	0.007	93.935	0.000	0.601	0.626
K_log	-887.7109	2.769	-320.546	0.000	-893.140	-882.282
tau_ln	203.1969	2.225	91.323	0.000	198.835	207.559

Omnibus:	3266.020	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	57056.225
Skew:	2.866	Prob(JB):	0.00
Kurtosis:	18.525	Cond. No.	8.30e+04

Methodology

Minimizing the sum of the squared differences between the observed and predicted values of the dependent variable.

Significant Variables

Choose variables which has the most significance by examining their p-values which in this case is S, K, K_ln, K_log and tau_ln

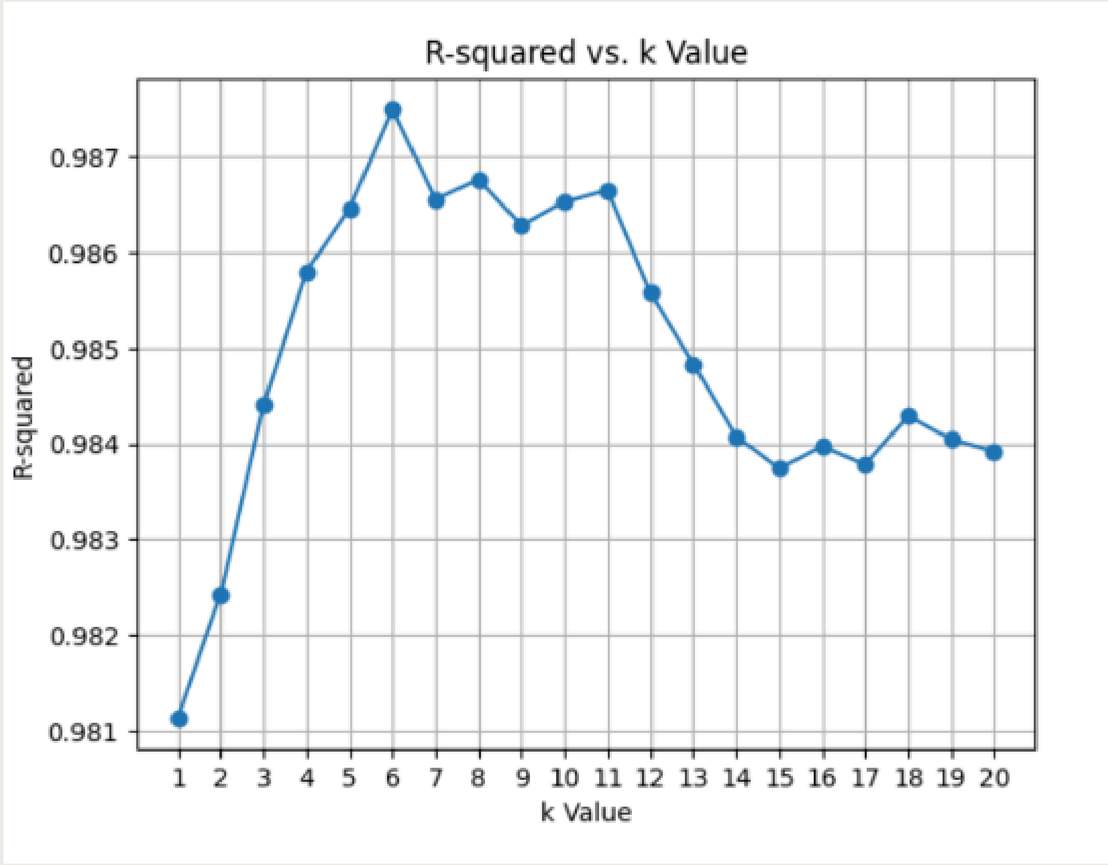
The more the variables, the higher the r-squared. However to prevent multicollinearity we didn't include redundant (i.e, K_ln and K) variables in the model.

R-squared Value

We have obtained an R-squared value of 0.958 using the 3 significant variables

MODEL BUILDING - KNN REGRESSION

- Best k value: 6
- Best R-squared: 0.987
- Mean Squared Error: 199.677

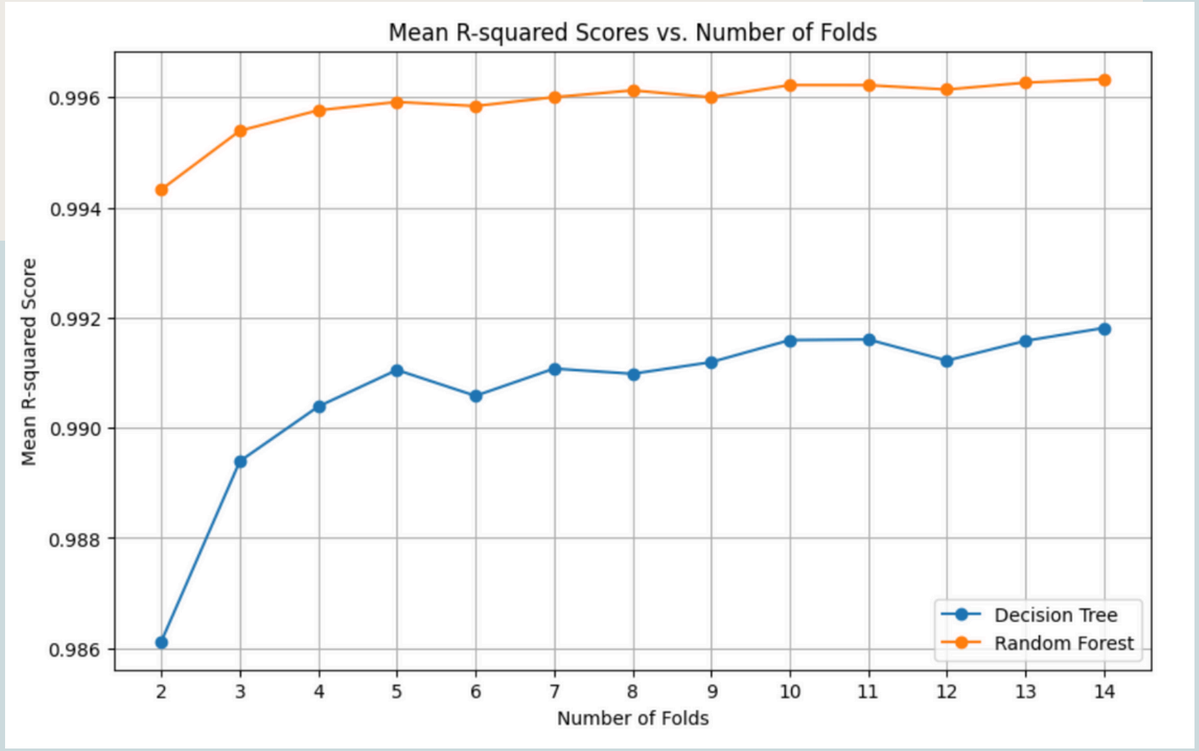


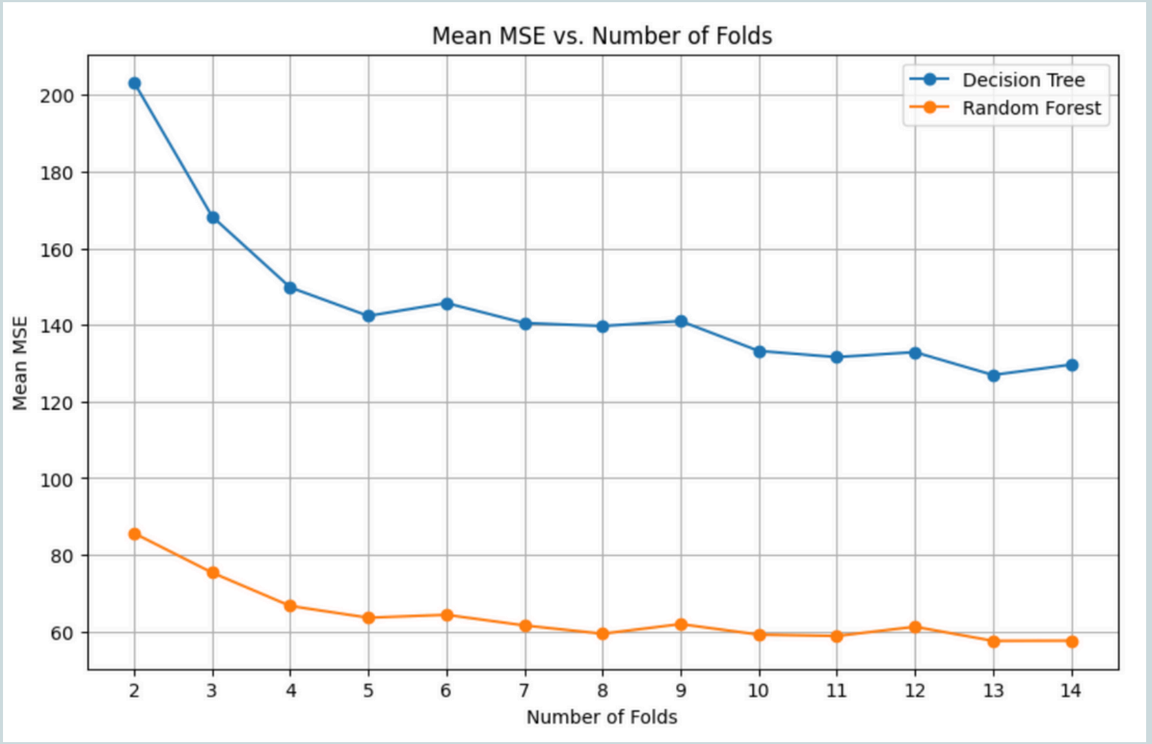
MODEL BUILDING

- Decision Trees
- Random Forest

Random Forest

- # of folds: 14
- Highest R-squared: ~0.996
- Lowest MSE: ~58.38

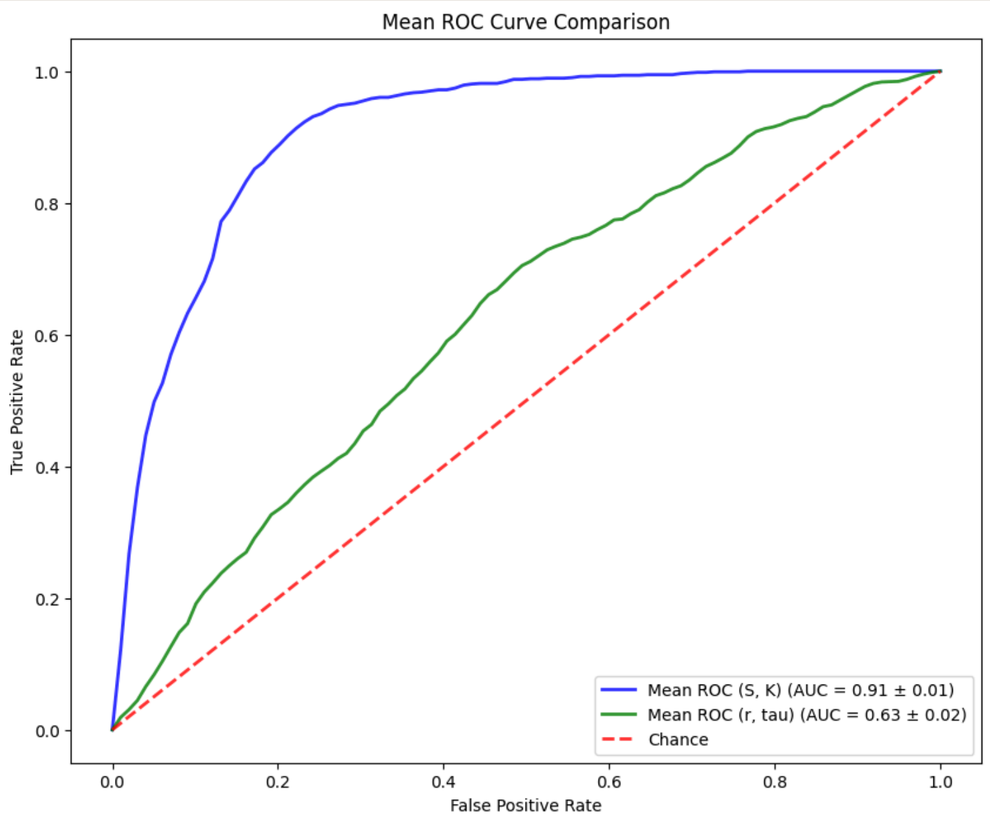




Classification

Classification

CROSS VALIDATION




```
predictor varible: S, K; response variable: BS
mean of classification errors using 10-fold CV: 0.1524
mean of auc using 10-fold CV: 0.9091936968479676

predictor varible: r, tau; response variable: BS
mean of classification errors using 10-fold CV: 0.22640000000000002
mean of auc using 10-fold CV: 0.6251935476881723
```

Methodology

Here, we employed logistic regression for binary classification which predicts the probability that an option is over/under valued and uses that to predict the outcome. We implemented the 10-fold CV which divides the dataset into 10 parts and uses each fold as a test set once

Analysis

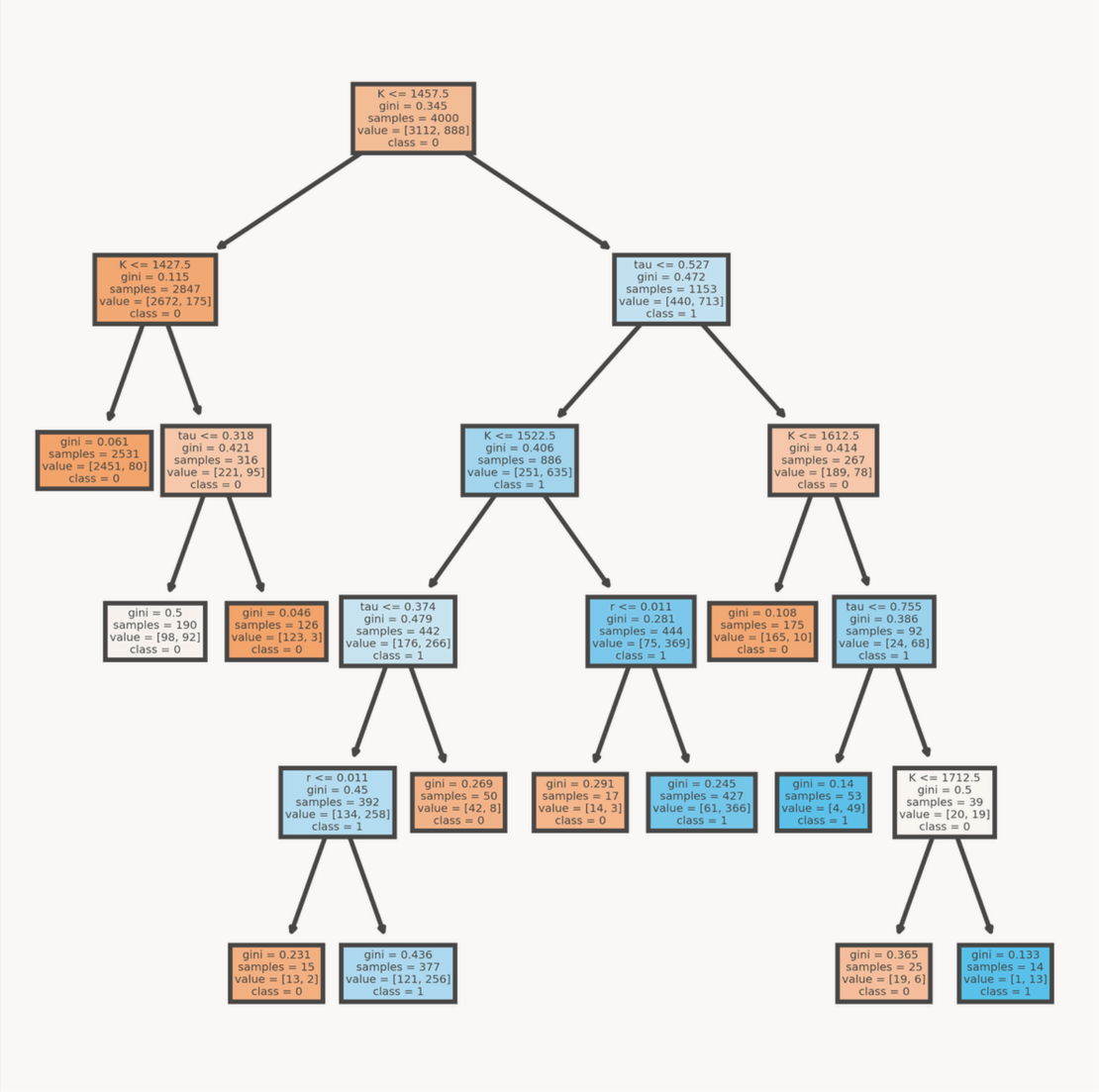
We evaluated 2 predictor sets: (S,K) and (r, tau) to reduce overfitting and assess the importance of each feature. We generated probability predictions on the test subset of each fold and then used that to classify each option as undervalued or overvalued. We then evaluated the accuracy of these predictions, represented in the ROC

Result

AUC of CV method with predictors S,K outperforms r,tau significantly with AUC of 0.91. This means that the underlying stock's price and strike price are a much greater predictor of over/under valueness of an option. While AUC is >90%, this classification algorithm is not the best-performing one.

Classification

DECISION TREE CLASSIFIER



VALIDATION ACCURACY: 0.907

TEST ACCURACY: 0.887

Why Decision Tree Classifier

- Chosen for its interpretability and ease of understanding; decision trees mimic human decision-making processes.
- Capable of capturing non-linear relationships between features and the target variable.

Interpretation

- Most decisions are made based on the 'K' feature (strike price of option), which is a primary driver in option valuation
- The depth of the tree is optimized to prevent overfitting

Classification

SUPPORT VECTOR MACHINE

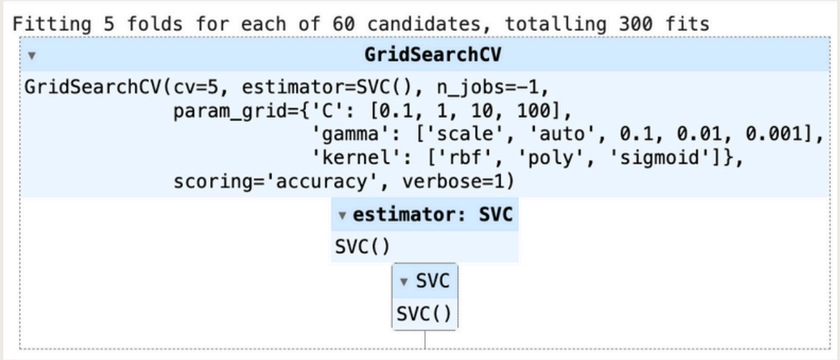
BEST VALIDATION ACCURACY: 0.920
TEST ACCURACY: 0.911

Why Support Vector Machine

- Effectiveness in high-dimensional spaces and its capacity for complex classification boundaries.
- Excels in situations where the relationship between class labels and features is not linearly separable, thanks to the kernel trick.
- Robust to overfitting, especially in high-dimensional space.

Interpretation

- Implemented an exhaustive search over specified parameter values for an SVM.
- Grid-search approach ensures the selection of the most optimal parameters, such as 'C' for regularization strength and 'gamma' for kernel coefficient, which significantly influence model performance.



Classification

GRADIENT BOOSTING CLASSIFIER

BEST VALIDATION ACCURACY: 0.930
TEST ACCURACY: 0.930

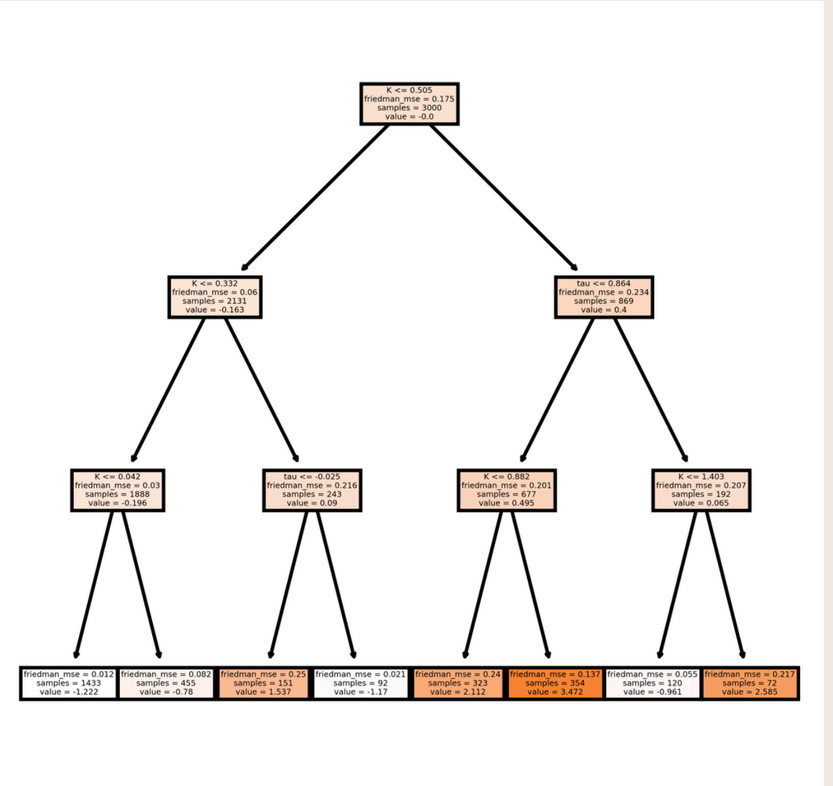
Why Gradient Boosting

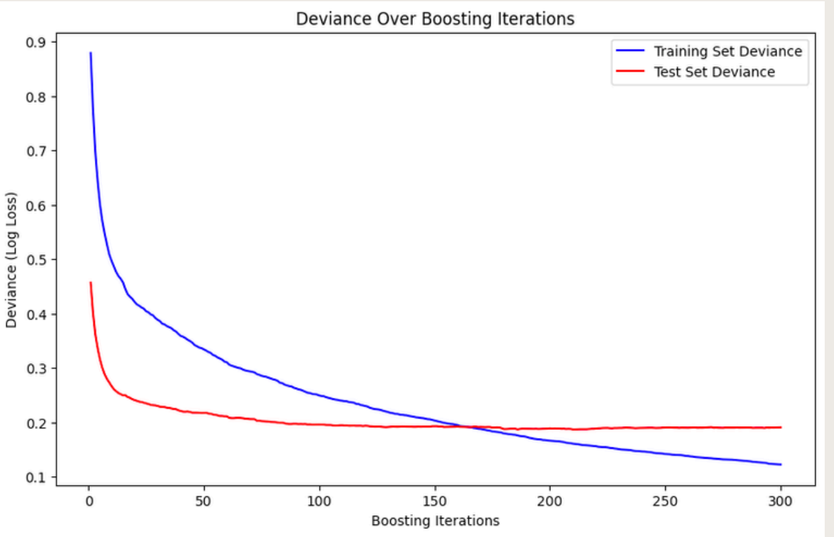
- Proficiency in handling complex datasets with intricate feature interactions and non-linear relationships.
- Combines weak learners to form a strong learner, iteratively improving upon areas where the previous models had high error.

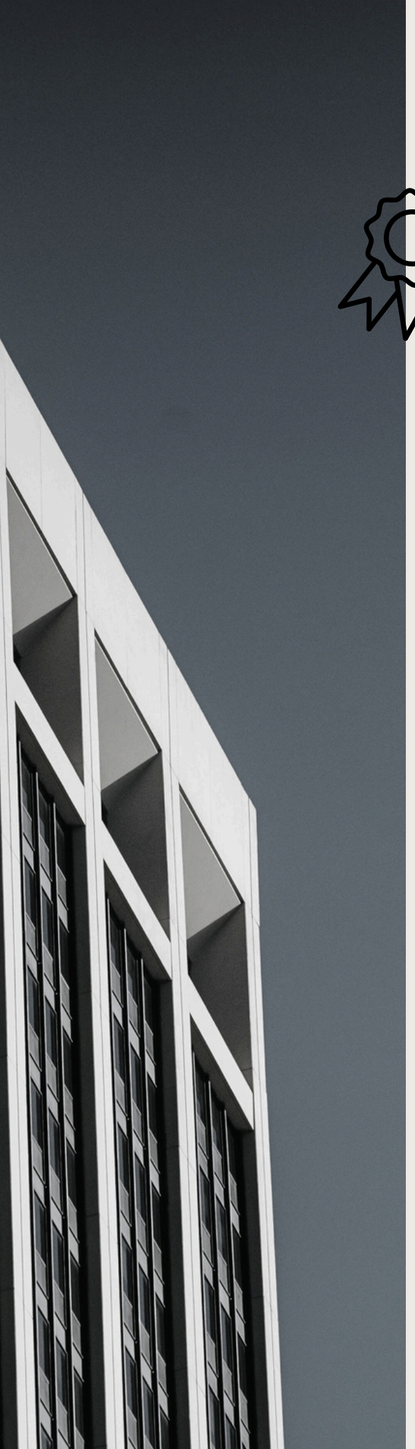
Interpretation

- The tree diagram represents an individual decision tree within the Gradient Boosting ensemble, showcasing the sequential improvement process.
- The tree splits indicate the hierarchical importance of features, with splits higher in the tree carrying more weight in the decision-making process.

- The graph shows a rapid initial decrease in deviance is observed, indicating strong learning from the model. As iterations continue, the decrease plateaus, indicating diminishing returns on learning with each additional tree.







RANDOM FOREST CLASSIFIER

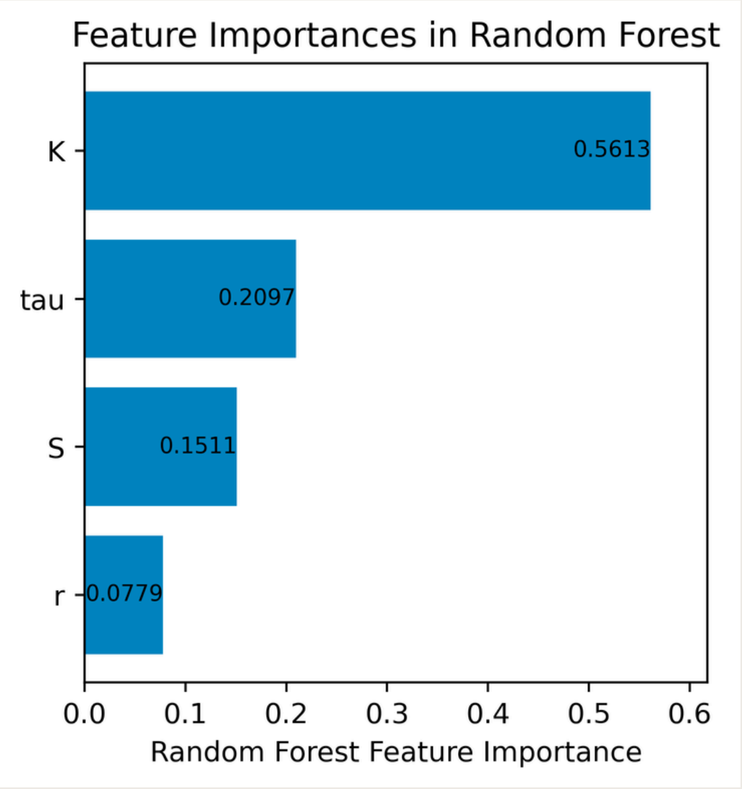
BEST VALIDATION ACCURACY: 0.935
TEST ACCURACY: 0.930

Why Random Forest Classifier

- Random Forest's ability to handle high-dimensional datasets and its inherent method of feature selection makes it a robust choice for financial modeling

Interpretation

- The model's emphasis on 'K', 'tau', and 'S' aligns with financial theories on the factors that influence option pricing.
- These insights not only bolster confidence in the model's predictions but also corroborate the well-established financial principles governing option markets.



Thank you for listening!