Options Pricing Prediction

DSO 530: Team 11

OUTLINE

- The Problem Statement and Data Overview
- Data Preparation
- Exploratory Data Analysis
- Feature Engineering
- Value Prediction (Regression Analysis)
- BS Prediction (Classification Analysis)
- Final Recommendation / Model
- Project Questions
- Limitations
- References

PROJECT INTRODUCTION

The aim of the project is to build a Statistical/ Machine Learning model that will estimate the theoretical value of derivatives based on other investment instruments, also taking into consideration the impact of time and other risk factors.

PROBLEM STATEMENT and DATA OVERVIEW

Build a Statistical Machine Learning Model to predict the following:

- Value
- BS

FEATURES		RESPONSE	
S	Current Asset Value	VALUE	Current Option Value
K	Strike Price of Option	BS	Black-Scholes Predictor (Over/Under)
TAU	Time to maturity (in years)		
r	Annual interest rate		

DATA PREPARATION

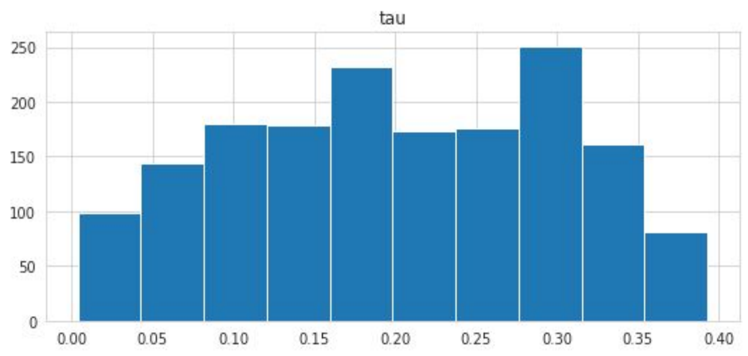
Null Values

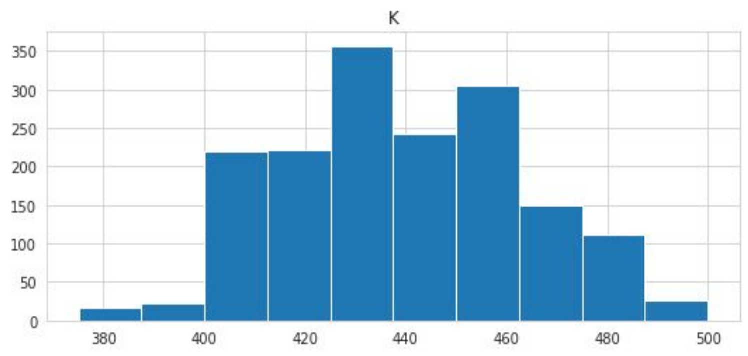
Rows Deleted: 2

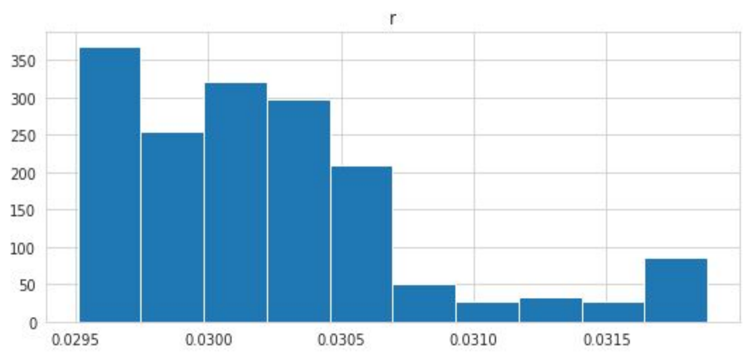
Outliers

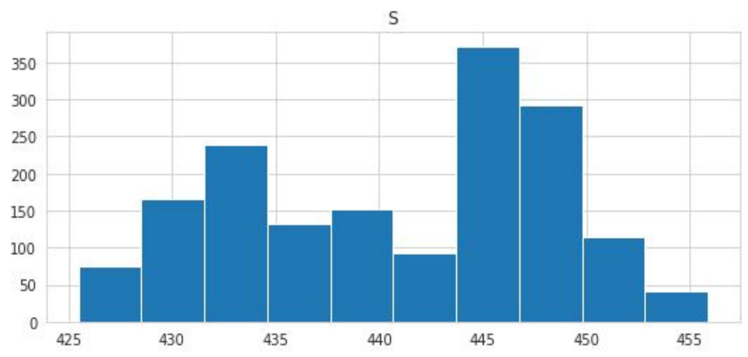
Rows Deleted: 3

EXPLORATORY DATA ANALYSIS









FEATURE ENGINEERING

SIMPLE OPERATIONS

Simple Operations	
$\sqrt{}$ TAU	Time to maturity (in years) under-root
TAU^2	Time to maturity (in years) squared
TAU^4	Time to maturity (in years) powered 4
1 + r	1 + Annual interest rate

NON-TRIVIAL

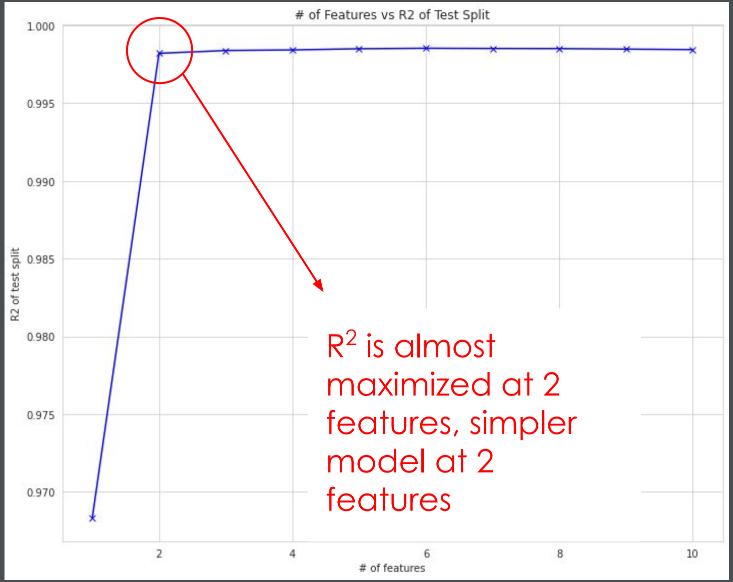
Scaling	
S/K	Current Asset Value/Strike Price of Option
$\|S-K\|$	Absolute difference between C.A.V. and Strike Price

01

VALUE PREDICTION

Regression Analysis

FEATURE SELECTION and MODEL OVERVIEW



R^2 is almost maximized at 2 features, simpler model at 2 features

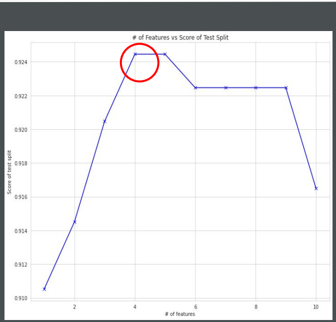
Method	**Forward Stepwise** **(Mean Square Error)**
Model	**Random Forest Regressor**
Top Features	**S/K,√ TAU**
R^2 on Test Split	**99.82%**

02

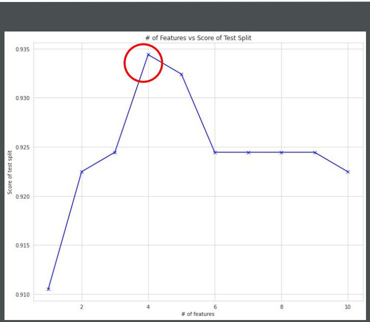
BS PREDICTION

Classification Analysis

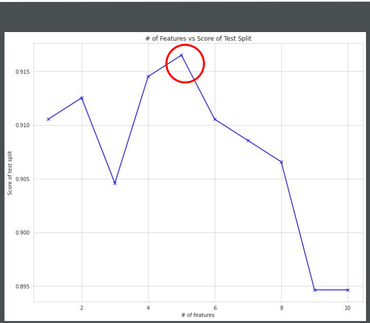
FEATURE SELECTION



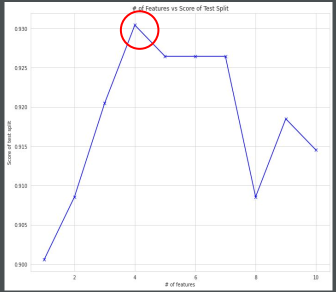
Logistic Regression



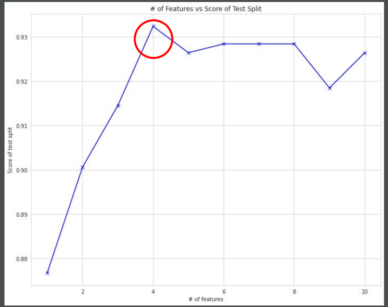
SVM



Naive Bayes



KNN



Random Forest

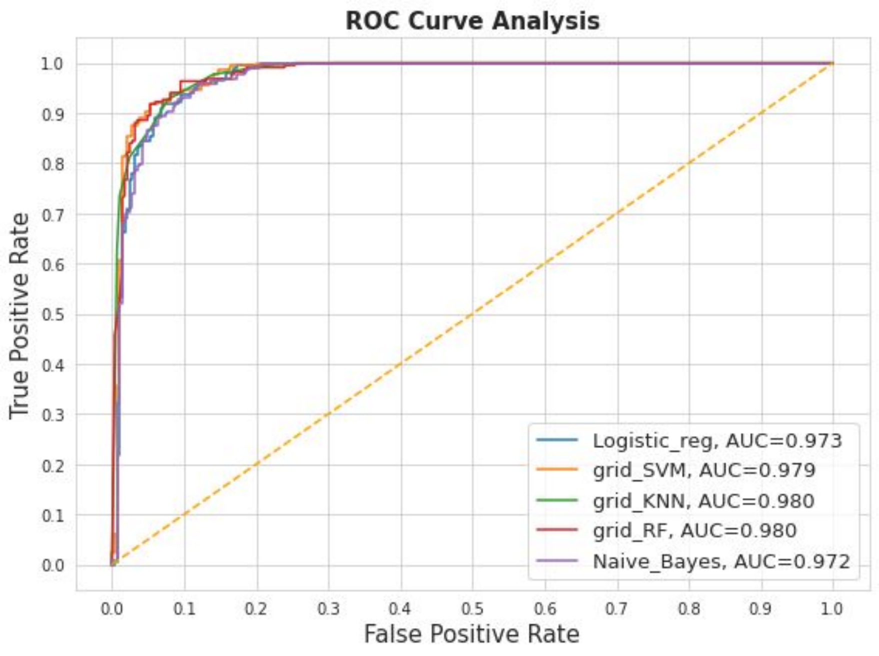
MODEL OVERVIEW

MODEL	FEATURE COUNT (best accuracy)	TOP FEATURES
Logistic Regression	4	r, S/K,√ TAU, TAU^4
SVM	4	S, r, S/K,√ TAU
KNN	4	S, S/K,√ TAU, TAU^2
Random Forest	4	S, r, S/K, √ TAU
Naive Bayes	5	S, r, S/K, |S-K|, TAU^4

HYPERPARAMETER TUNING

MODEL	TOP FEATURES	HYPERPARAMETERS (roc-auc)
Logistic Regression	r, S/K, $\sqrt{\text{TAU}}$, TAU^4	-
SVM	S, r, S/K, $\sqrt{\text{TAU}}$	Kernel = ‘rbf’, Gamma = 0.01, C = 100
KNN	S, TAU, S/K, $\sqrt{\text{TAU}}$, TAU^2, TAU^4	Leaf Size = 7, N-Neighbours = 8, p = 1
Random Forest	S/K, S, K, TAU, r	N-Estimators = 188, Min Samples Split = 2, Min Samples Leaf = 1, Max Features = ‘sqrt’, Max Depth = 12, Bootstrap = True
Naive Bayes	S, r, S/K, $\|S-K\|$, TAU^4	-

ROC - AUC CURVE and BEST THRESHOLD



MODEL	BEST THRESHOLD	ACCURACY ON TEST SPLIT
Logistic Regression	0.508195	92.25%
SVM	0.559499	93.24%
KNN	0.500000	91.25%
Random Forest	0.372405	92.84%
Naive Bayes	0.349611	91.25%

MAJORITY PREDICTION is the SOLUTION

RECORD	PREDICTION AT BEST THRESHOLD					FINAL PREDICTION	ACTUAL	TYPE
	L.R.	SVM	KNN	R.F.	N.B.			
1	1	1	1	1	0	1 (4 out of 5)	1	TP
2	1	0	0	0	0	0 (4 out of 5)	0	TN
3	1	1	1	1	1	1 (5 out of 5)	1	TP
4	0	1	0	1	0	0 (3 out of 5)	1	FN
5	0	1	1	1	0	1 (3 out of 5)	0	FP

Classification Model Performance

CONFUSION MATRIX

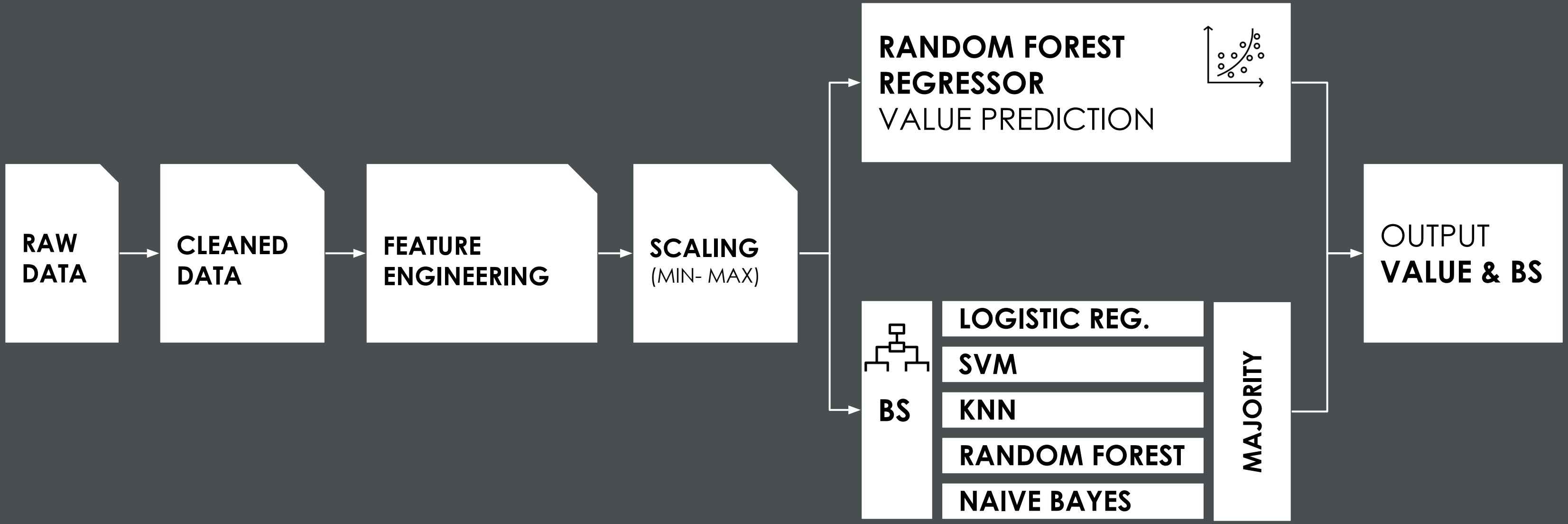
	Predicted Under	**Predicted Over**
Actual Under	265 **TN**	19 **FP**
Actual Over	16 **FN**	203 **TP**

Metrics:

- **Accuracy: 93.04%**
- **Classification Error: 6.96%**

- By using **Majority criteria**, we are eliminating the bias and dependency on dataset.
- If an individual model is considered, it's accuracy depends on the dataset and random state

Final Recommendation / Model



RAW DATA

CLEANED DATA

FEATURE ENGINEERING

SCALING (MIN- MAX)

RANDOM FOREST REGRESSOR VALUE PREDICTION

BS

LOGISTIC REG.

SVM

KNN

RANDOM FOREST

NAIVE BAYES

MAJORITY

OUTPUT VALUE & BS

Project Questions?

(1) In both prediction problems, would you argue if prediction accuracy or interpretation is more important? Why?
This business case deals with the prediction of stock option prices, and classification based on that stock option prices. Hence it becomes obvious that we need to focus more on prediction accuracy rather than the interpretation of the model. The higher the accuracy of the model the better are the chances of correct classification of BS.

(2) Why do you think machine learning models might outperform Black-Scholes in terms of predicting option values?
The traditional black scholes money model is considered to be a weak predictor of option pricing for deep in the money or deep out of the money state. ANN models through back-propagation reduces the squared errors and through validation set training avoid over-fitting

Project Questions?

(3) Can you argue from a business perspective that all four predictor variables should be included in your prediction (i.e., no variable selection is necessary)?
Black scholes utilizes 6 parameters for option pricing prediction i.e strike price, price of the underlying, risk-free rate, dividend yield, volatility and time-to-maturity. In our case, all four predictor variables are necessary since they will be used as proxies for historical volatility, implied volatility, and volatility index i.e VIX.

(4) Are you comfortable about directly using your trained model to predict option values for Tesla stocks? Why?
We will need to perform a validation testing on the Tesla stocks in order to understand the extent of overfitting/underfitting. Looking at the stock price pattern, we can clearly see that the stock has a lot of volatility Feb, 2020 onwards. It would be meaningful to use this data set for calculating the prediction accuracy of our model before suggesting the use of our model. We can then use the model to predict TSLA stock prices. Other factors to consider are: MAE, R^2, tuning the model.

Limitations

- By using **Majority criteria for Classification Prediction**, we are increasing the computational time. This may result in delayed prediction

- Our models are built on ~1600 rows of data, and therefore have high bias. To reduce the bias in the model performance it is recommended to introduce more data

- The model result depends on the feature engineering we did, for better performance we may end up using complicated features which might not even be required, so the guidance of a subject matter expert in creating features would be desirable.

Please send us questions and suggestions!

chadhaa@usc.edu

Appendix

Future Scope

Since we have already created the model on the 70% of the data, and validated its performance on the remaining 30%, it is considered a good practice to re-train the model on the complete 100% of the data. However this was out of the scope of our project and so we did not do it.

References

- Evaluating Machine Learning models using Hyperparameter Tuning
- Scoring Methodology in Feature Selection
- Hyperparameter Optimization with Random Search and Grid Search
- Sequential Feature Selection