



Option Pricing Prediction Model

Group 29 Members:

Liang-Chi Liu

Yi-Ching Lin

Nattawut Kananusorn

Suwara Thianrungsot

Hsuan-Ting (Tommy) Wu



Table of contents

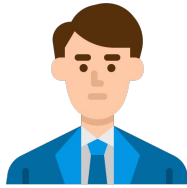
1. Business Problem
2. Data Exploration
3. Data Cleaning
4. Variables Selection
5. Model Selection
 - a. Regression problem
 - b. Classification problem
6. Conclusion
7. Appendix

1.

Business Problem



European Call Option on the S&P 500



Mr. Adam



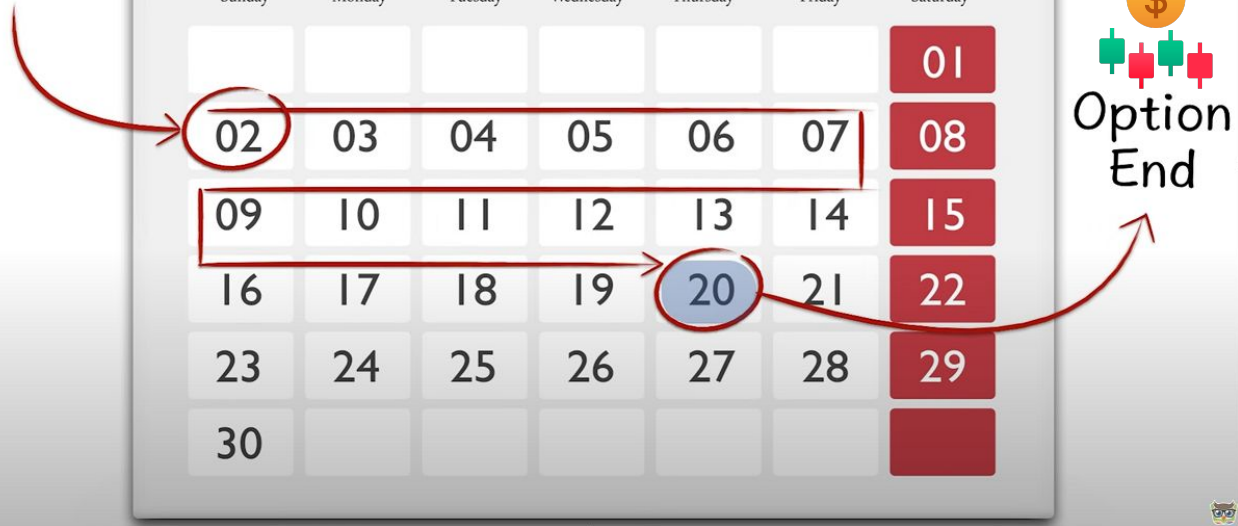
S&P 500



Mrs. Alice



Option
Start



Project Objective

*Make **a predictive model** for
current option value (Value)
and
BS (Over- or Undervalue)*

2.

Data Exploration



What does the data look like?

Data Set Shape: 1680 rows, 6 columns

Variable	Description	Data type
Value (C)	Current option value	Float
S	Current asset value	Float
K	Strike price of option	Float
r	Annual interest rate	Float
tau	Time to maturity (in years)	Float
BS	Over / Under based on whether (result of Black-Scholes formula - C) > 0	String

3.

Data Cleaning



Data Cleaning

After data cleaning: 1675 rows, 11 columns



Drop Missing Values

Remove 2 rows due to missing values



Drop Erroneous Entries

3 rows removed due to outliers

- $S = 0$
- $\tau = 146$
- $\tau = 250$



Convert BS binary class

1 column created

- Under = 0
- Over = 1



Standardize variables

4 columns created

- Z scores of S, K, r, τ

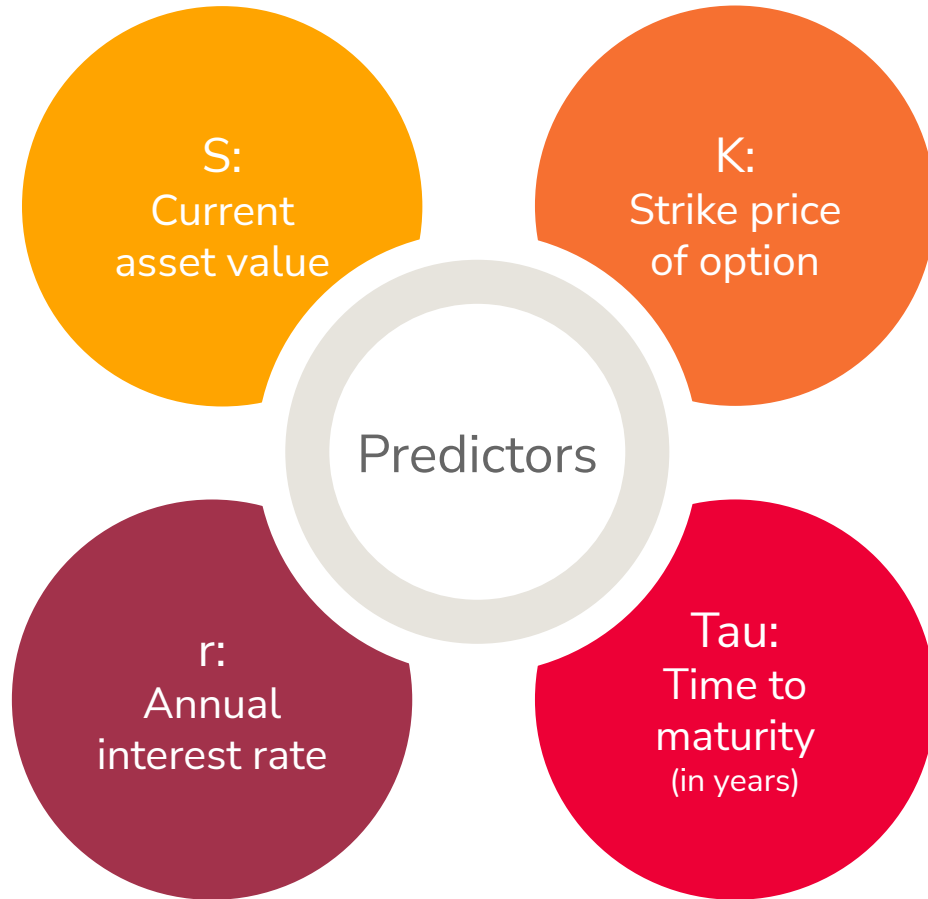
4.

Variables Selection



We include all
four variables
in the model.

Why?



5.

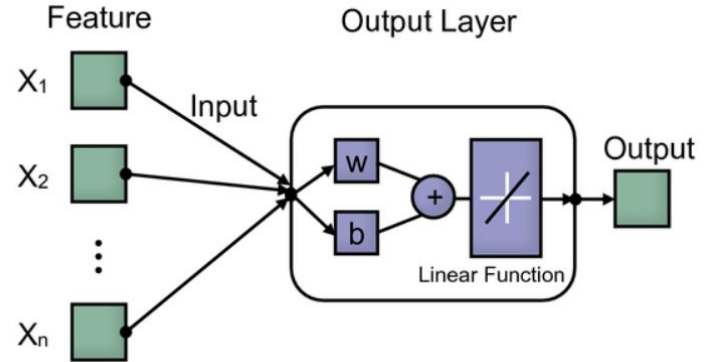
Model Selection



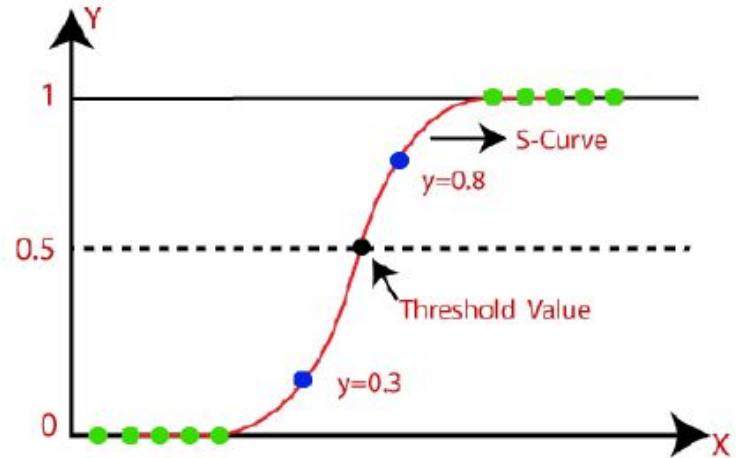
Let's review some concepts



Linear Regression



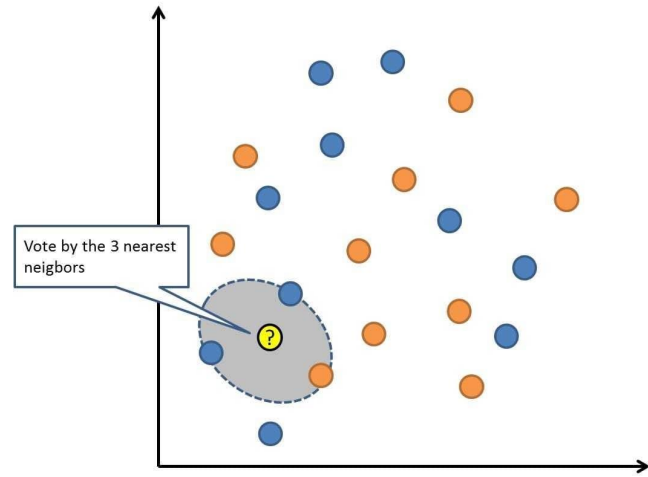
Logistic Regression



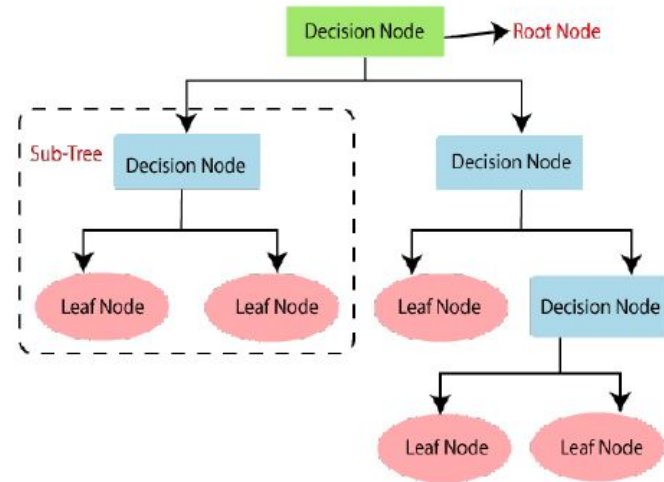
Let's review some concepts



K-Nearest Neighbors



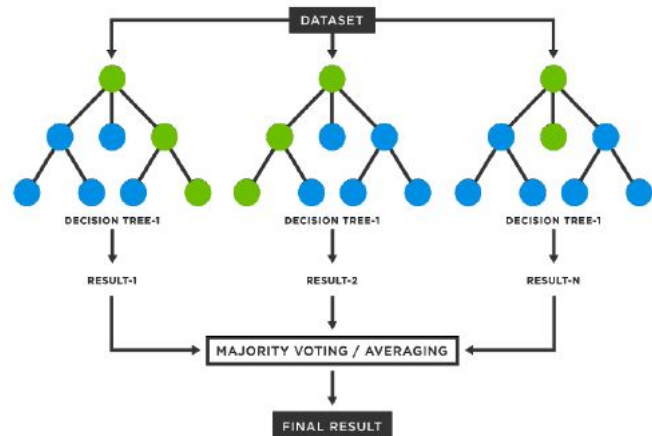
Decision Trees



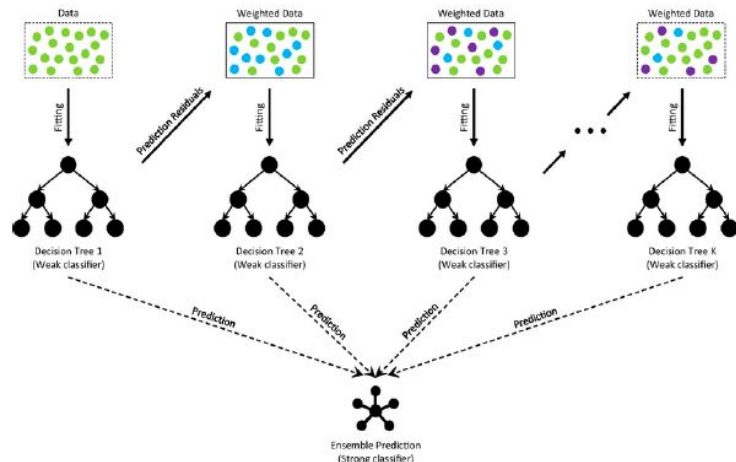
Let's review some concepts



Random Forest



Gradient Boosting Trees

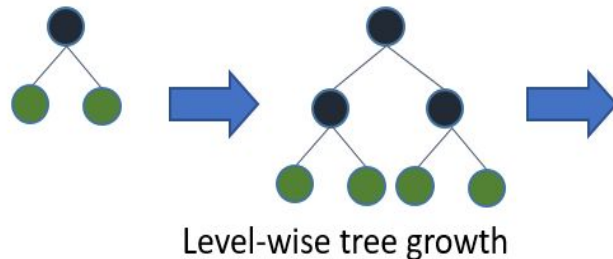


Let's review some concepts

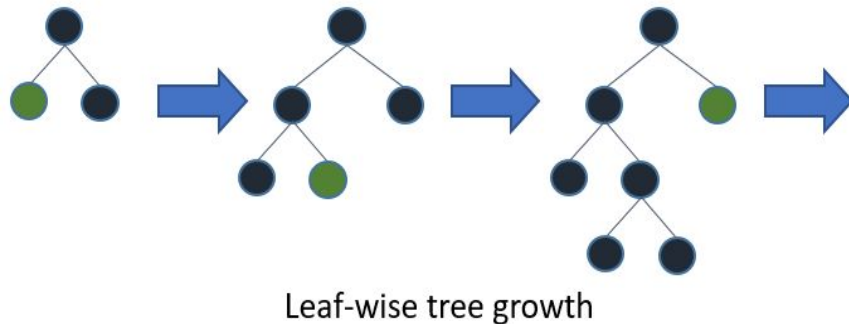


XGBoost

(Extreme Gradient
Boosting Tree)



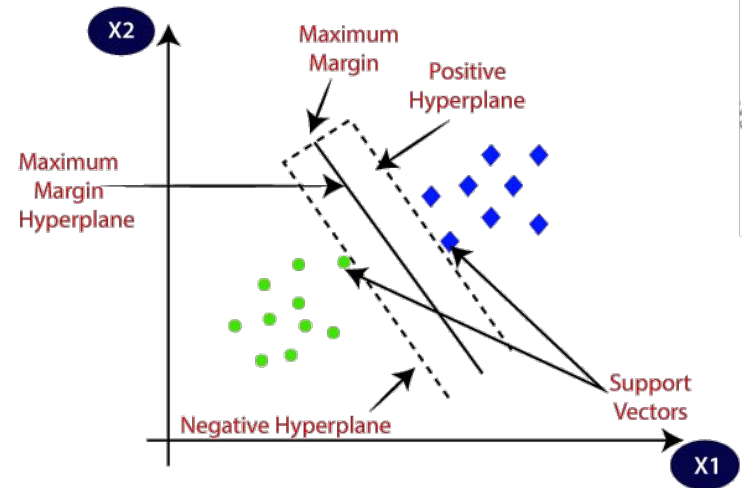
LightGBM



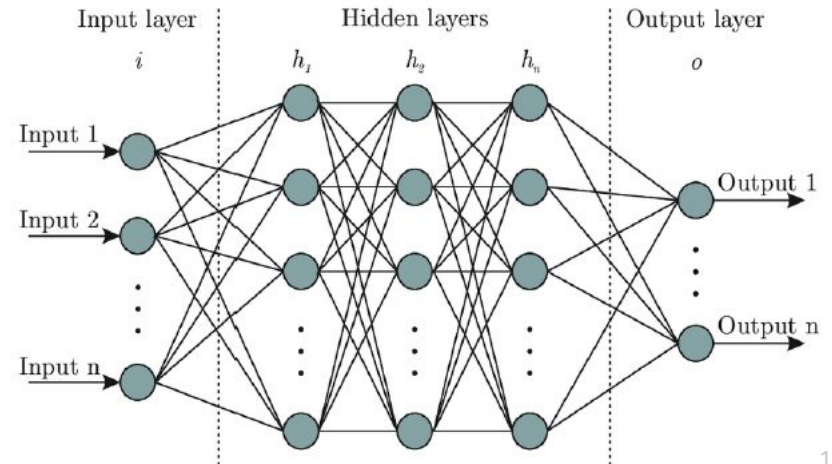
Let's review some concepts



Support Vector
Machine



Neural Network



Steps

1

Split dataset into training and testing sets (80:20)

2

Use training set to find the best parameters for each model utilizing GridSearchCV

3

Validate tuned models with testing set and compare accuracy scores

Final Model (Regression)

Model	Parameters						R2	
Linear Regression							Train	Test
							91.083%	91.140%
Decision Tree	Criterion	max_depth	min_leaf	min_split	splitter		Train	Test
	absolute_error	18	1	2	random		98.632%	98.859%
Random Forest	Criterion	max_depth	min_leaf	min_split	n_estimators		Train	Test
	mse	None	1	2	100		99.460%	99.552%
Gradient Boosting Tree	Criterion	max_depth	min_leaf	min_split	n_estimators	learning_rate	Train	Test
	mse	4	1	10	800	0.1	99.874%	99.902%
XGBoost	learning_rate	max_depth	n_estimators				Train	Test
	0.1	4	800				99.858%	99.898%
LightGBM	learning_rate	max_depth	n_estimators				Train	Test
	0.1	-1	150				99.346%	99.713%
Neural Network	activation	# of nodes	# of layer	learning_rate			Train	Test
	relu	10	2	adaptive			99.102%	98.998%

Final Model (Classification)

Model	Parameters					Accuracy	
Logistic Regression	penalty	solver	C			Train	Test
	l2	newton-cg	1			91.493%	91.642%
Decision Tree	max_depth	min_leaf	min_split			Train	Test
	9	2	2			90.746%	92.239%
Random Forest	max_depth	min_leaf	min_split	n_estimators		Train	Test
	10	2	5	50		92.239%	92.537%
Gradient Boosting Tree	max_depth	min_leaf	min_split	n_estimators	learning_rate	Train	Test
	5	4	2	100	0.1	92.239%	92.537%
XGBoost	colsample_bytree	gamma	max_depth	min_child_weight	subsample	Train	Test
	0.8	1	4	1	0.6	93.284%	93.134%
LightGBM	lambda_l1	lambda_l2	max_depth	n_estimators		Train	Test
	1	1	4	150		92.836%	93.731%
Neural Network	activation	hidden_layer_sizes	learning_rate			Train	Test
	relu	(100, 3)	constant			93.209%	93.134%
SVM	kernel	gamma	C			Train	Test
	rbf	0.1	100			92.687%	93.433%
KNN	n_neighbors					Train	Test
	9					92.015%	92.537%

Accuracy vs. Interpretation

Investor or Business Users

✓ Accuracy

Data Analyst or Data Scientist

✓ Both



6.

Conclusion



Conclusion

Regression

Gradient Boosting Trees



Value

Classification

Light Gradient Boosting
(LGB)



BS



- ★ **Black-scholes Model VS Machine Learning?**
- ★ **Predict option values for Tesla stocks by this model?**

Thank You

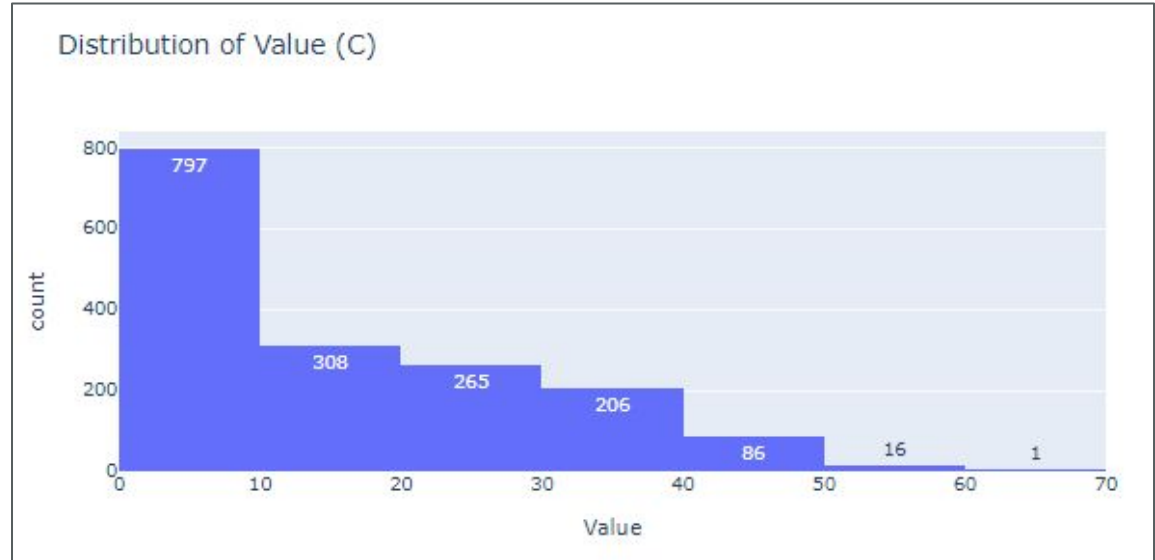
7.

Appendix



Variable: Current Option Value (C)

- 1 row with null value
- Majority of values are around 0-10



Variables:
S (Current asset
value),
K (Option Strike
Price)

- K values are more centered, around the (420, 460) range
- S values are more spread open, around the (375, 490) range
- 1 row has null values for both K and S
- 1 row has S value of 0



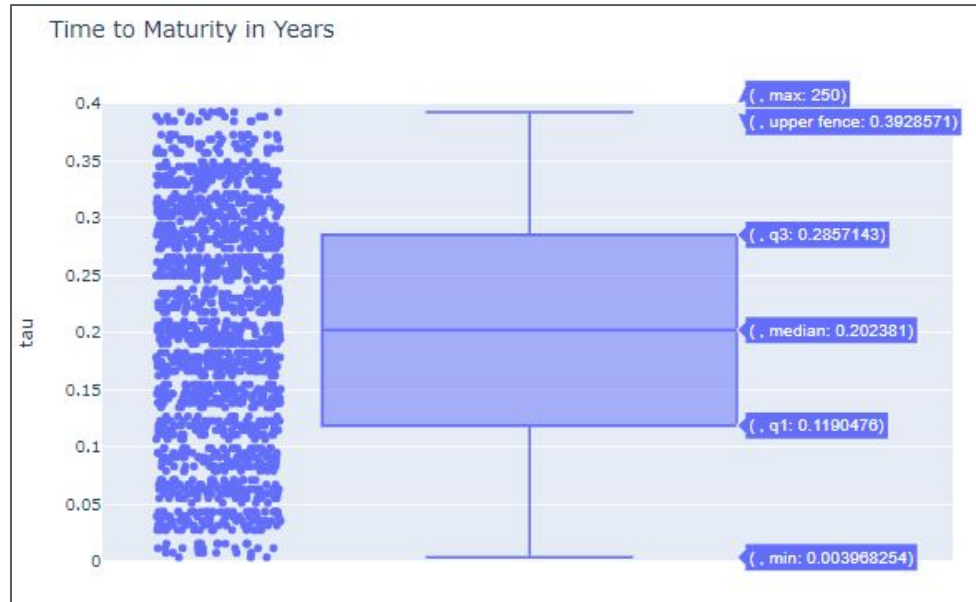
Variable: r (Annual Interest Rate)

- Rates range from 0.029 to 0.032



Variable: tau (Time to Maturity in Years)

- 2 extremely high values (146, 250)
- 1 row with null value
- Majority of values range from 10 days to 4 months



Variable: BS

Value	Count of values	Percentage
Over	946	56%
Uder	734	44%

