Google Stock Prediction

Closing price

Presented by Group 5
supervised by: Dr.Mashael Aldayel

Google Stock Prediction

What to Know?

Introduction to Google Stock

Google is one of the largest most influential and successful technology companies in the world

Predicting the closing price of a stock is a complex problem because of several challenges but also can potentially help investors make more informed decisions about buying, selling, or holding stocks

Our class label is The closing price of a stock. It's one of the most commonly used prices to analyze a stock's performance

Google Stock Dataset sample

	symbol	date	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOpen	adjVolume	divCash	splitFactor
1	GOOG	2016-06-14 00:00:00+00:00	718.270	722.4700	713.1200	716.48	1306065	718.270	722.4700	713.1200	716.48	1306065	0	1
2	GOOG	2016-06-15 00:00:00+00:00	718.920	722.9800	717.3100	719.00	1214517	718.920	722.9800	717.3100	719.00	1214517	0	1
3	GOOG	2016-06-16 00:00:00+00:00	710.360	716.6500	703.2600	714.91	1982471	710.360	716.6500	703.2600	714.91	1982471	0	1
4	GOOG	2016-06-17 00:00:00+00:00	691.720	708.8200	688.4515	708.65	3402357	691.720	708.8200	688.4515	708.65	3402357	0	1
5	GOOG	2016-06-20 00:00:00+00:00	693.710	702.4800	693.4100	698.77	2082538	693.710	702.4800	693.4100	698.77	2082538	0	1

Our data consists of :
Number of Attributes: 14
Number of objects: 1258

Source :

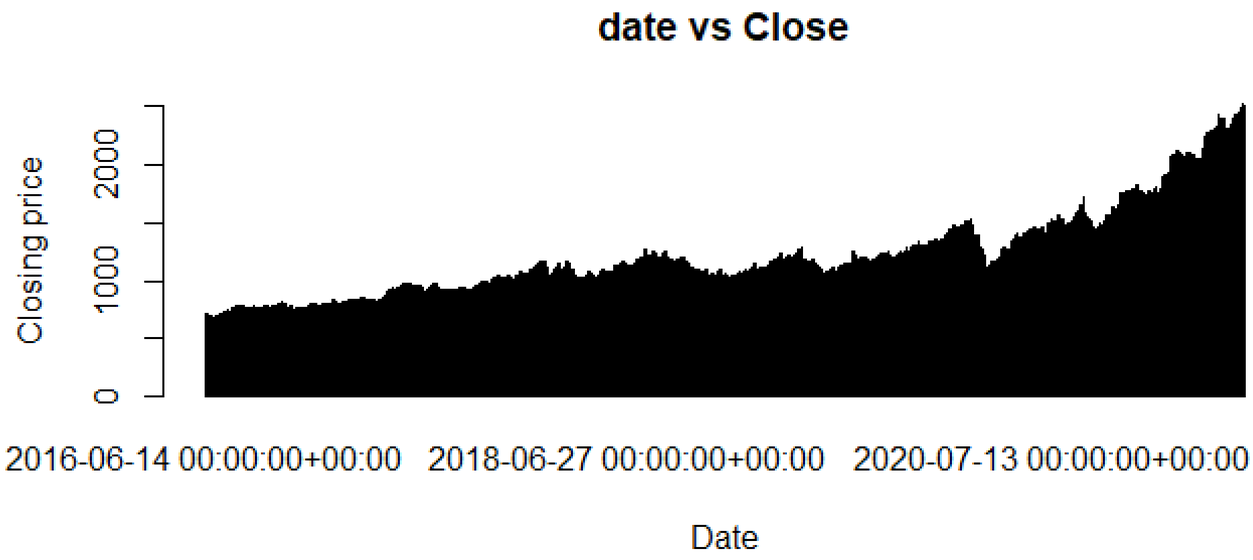
Google Stock Prediction

Regression, LSTM, Keras, Tensorflow

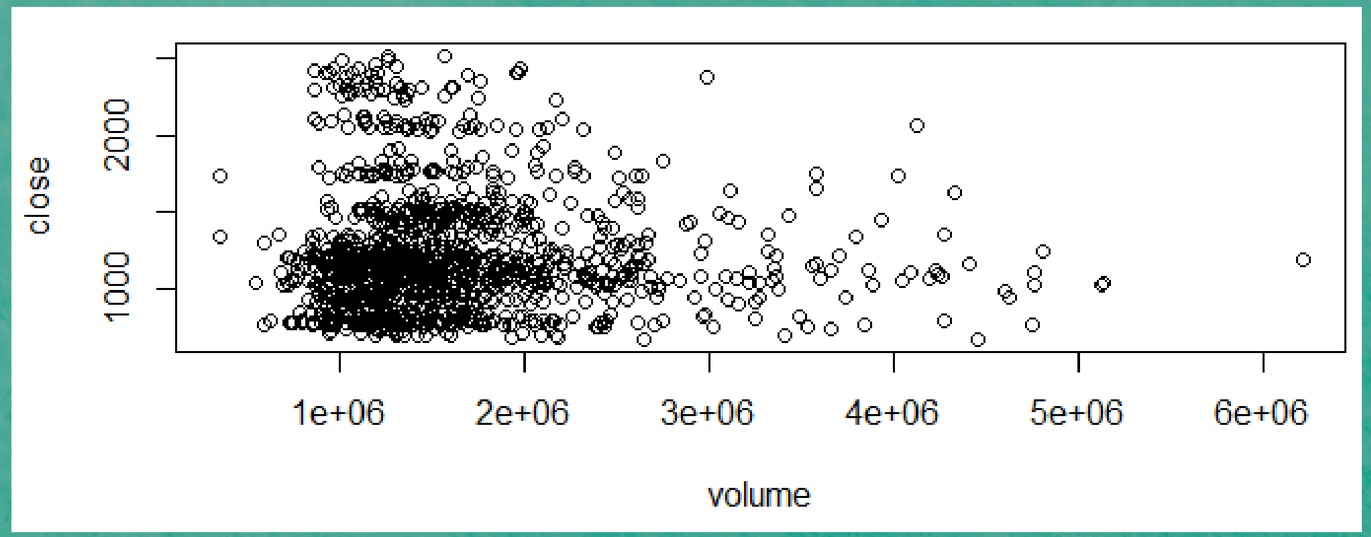
k kaggle.com

Graphs codes

- histogram
- Scatter Plot
- Barplot

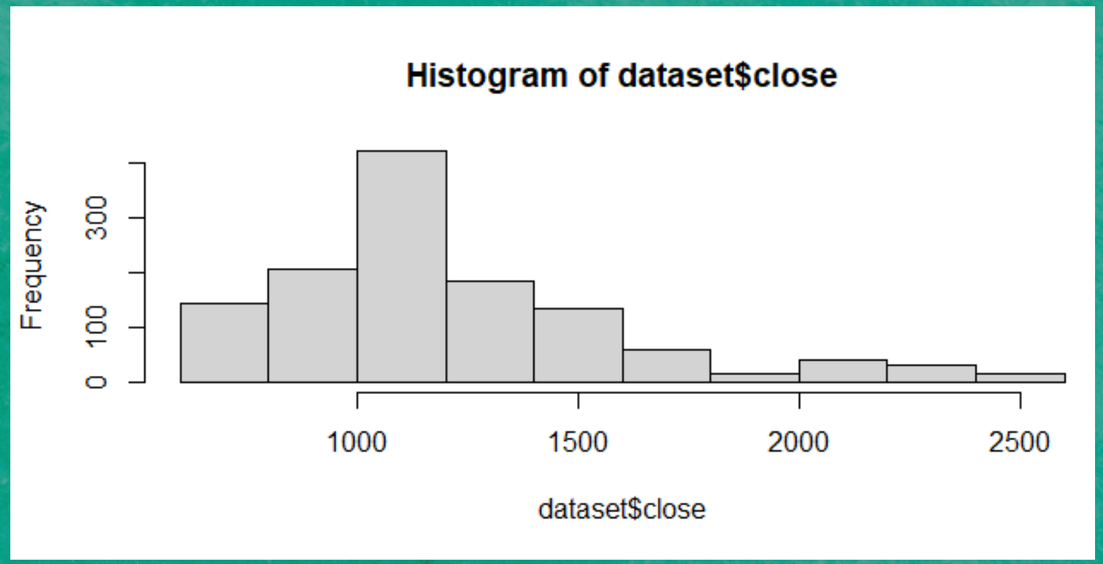


Graphs codes



Scatter Plot

Histogram



Data Summarization and Preprocessing

We applied several preprocessing techniques to improve the accuracy and efficency of the data which is :

- Data cleaning
- Data transformation
- Feature selection

Data cleaning

- Handling wrong values
- Handling all outliers

Data transformation

- Normlization
  1-min-max
  2-encoding

- Discretization

Data Summarization and Preprocessing

summary

Min , 1st Qu , Median
Mean , 3rd Qu , Max

mean &variance

```
mean(dataset$close)
[1] 1216.317

 var(dataset$close)
[1] 146944.5
```

Data cleaning

```
#Checking NULL
is.na(dataset)

#find the total null values
sum(is.na(dataset))
```

Detec the outliers

```
OutClose = outlier(dataset$close, logical =TRUE)
sum(OutClose)
Find_outlier = which(OutClose ==TRUE, arr.ind = TRUE)
OutClose
Find_outlier

#Remove outlier
dataset= dataset[-Find_outlier,]
```

Data transformation

Feature selection

we used 2 feature selection methods provided by the caret R package which is :

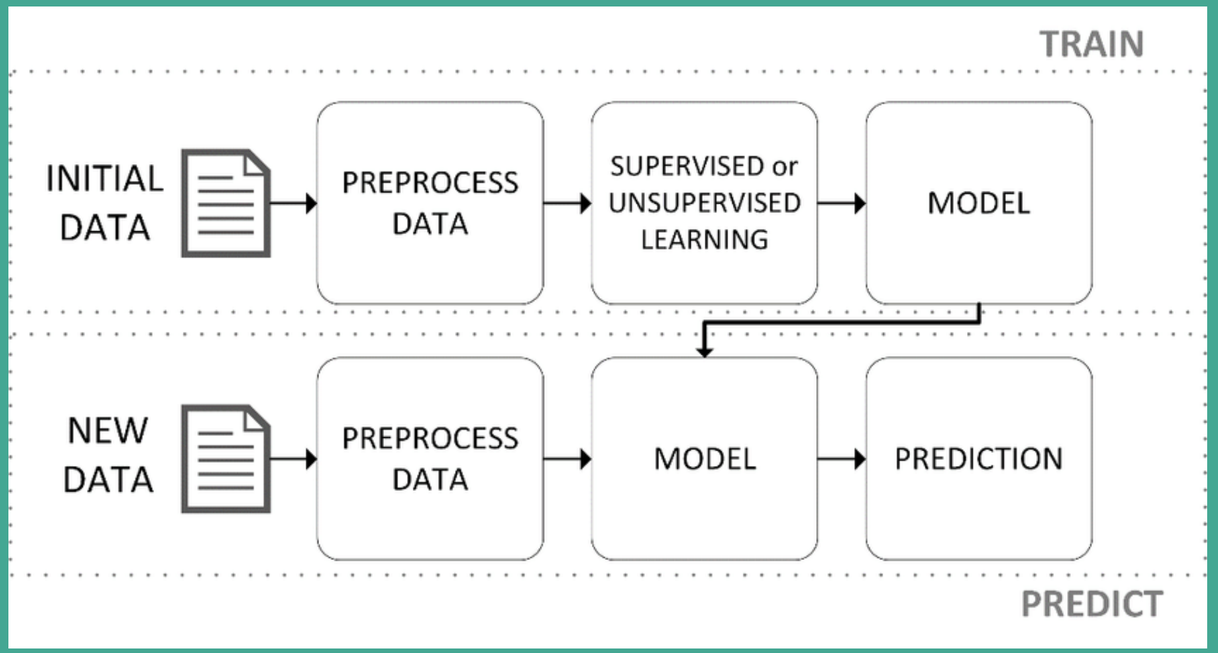
1- using Recursive Feature Elimination or RFE

2- Remove Redundant Features

Classification

The goal of classification is to assign predefined labels or categories to input data based on its characteristics or features.

classification is a supervised ,



Splitting criteria and algorithms

we implemented 3 algorithms on each split to evaluate 9 models in total.

The algorithms:

- Information Gain(ID3)
- Gain Ratio(C5.0)
- Gini index(CART)

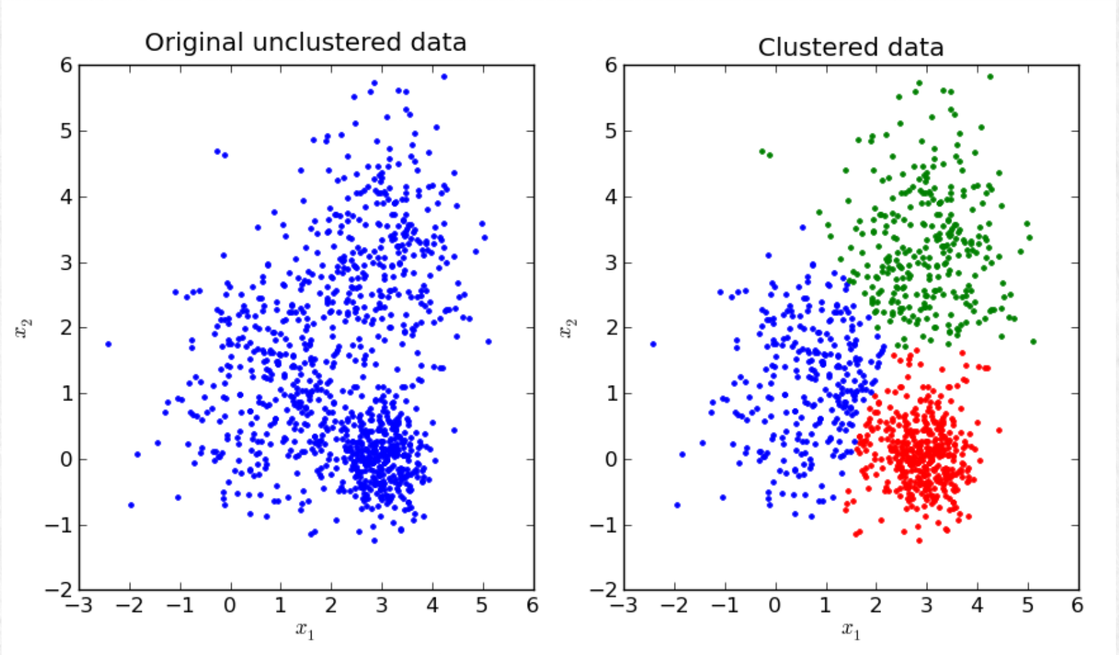
Evaluation and Comparison

	70% Training and 30% Testing data			80% Training and 20% Testing data			60% Training and 40% Testing data		
	IG	IG ratio	Gini andex	IG	IG ratio	Gini andex	IG	IG ratio	Gini andex
Accuracy	%98.6	%99	%98.6	%98.4	%99.5	%98.4	%98.8	%98.8	%98.8
precision	%97.3	99	%97.3	%97.3	1	%97.3	%96	%97.3	%96
sensitivity	%99.5	%98.2	%99.5	%99.4	%98.6	%99.4	%99.4	%99.3	%99.4
specificity	%96.8	%99.5	%96.8	%96.5	1	%96.5	%96.5	%98.6	%96.5

Clustring

The goal of clustering is to group a set of objects or data points into subsets or clusters based on the similarity between them.

Unlike classification , Clustering is an unsupervised.



Apply k-means clustering for different values of K:

We used K-means technique that represents the clusters by the center of cluster

- The average silhouette
- Total within-cluster sum of square
- The BCubed (precision and recall)

k= 2

k= 3

k= 4

Evaluation and Comparison

	K=4	K=3	K=2
average silhouette width	0.43	0.37	0.45
total withn-cluster sum of squares	1900.127	2908.955	4126
BCubed precision	70.30498	80.95238	100
BCubed recall	79.63636	9.272727	0.1818182

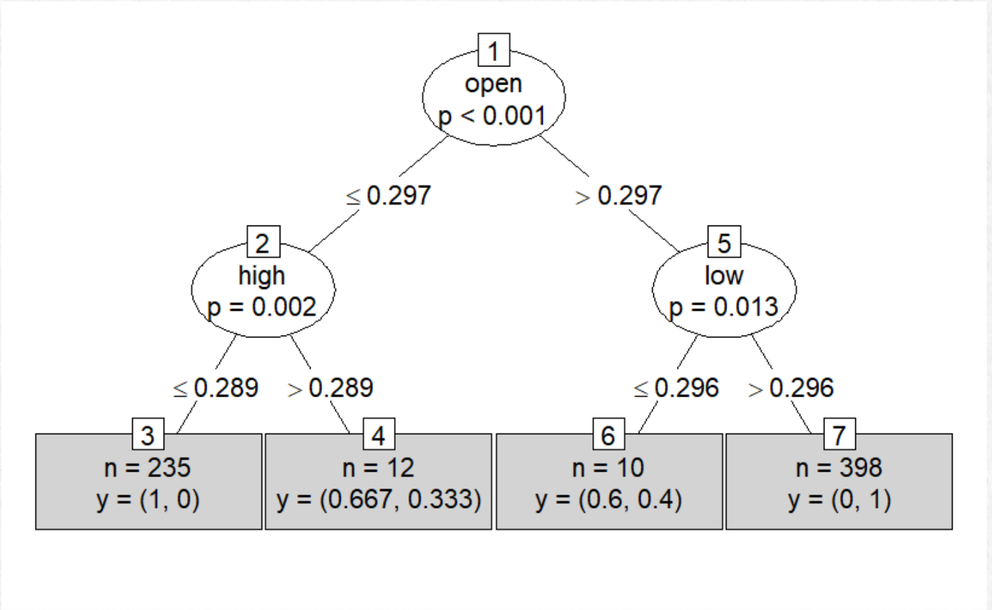
Findings

we found that

- the best decision tree was IG with 60-40 split

- Gain ratio tree has highest accuracy in genral

- classification is not directly applicable for predicting continuous variables

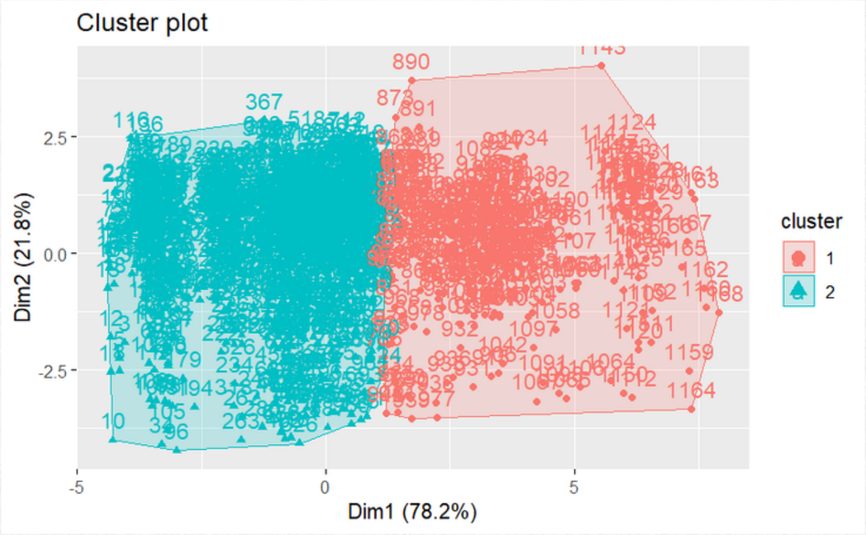


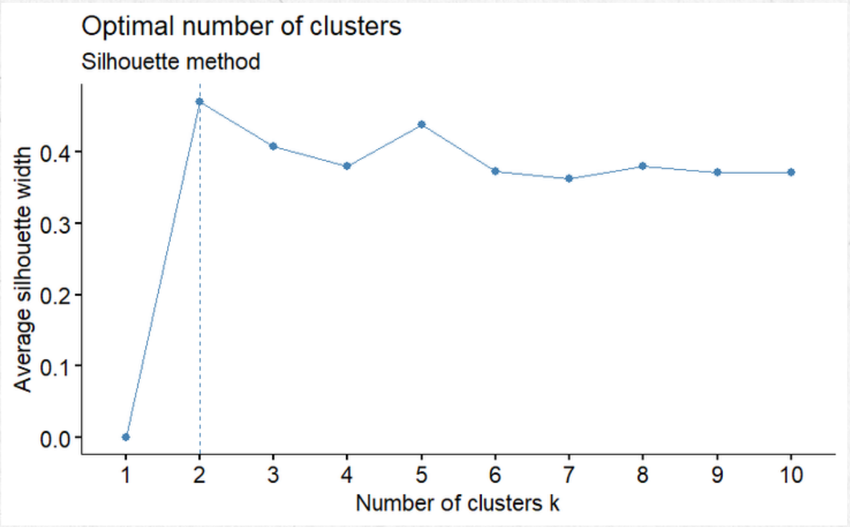
Accuracy	%99.5	%99	%98.8
precision	1	%99	%97.3
sensitivity	%98.6	%98.2	%99.3
specificity	1	%99.5	%98.6

Findings

we found that clustering with k=2 is better suited for our Dataset

has highest average silhouette width





Thank you very much!

Wijdan Alhashim - 443200530
Shaden Alturki - 443203057
Reem Alnaseer - 443200497