Data Science Capstone Project Accompanying Document

Problem Statement

The stock market is an ever-evolving market where the only constant is change. An array of factors affect the price of a stock on a daily basis and it is impossible to predict when the market is going to pick up or drop based off just a few variables. Nobody would have predicted the extent to which the the covid-19 pandemic would have caused the market to crash & hence many people were unable to minimise their losses as a result. Hence, we can utilise Machine Learning algorithms to make predicitions related to future stock prices.

My hypothesis is that we are able to utilise machinese learning models effectively to predict to the price of the stock in the near future. In this project, I will be using the Linear Regression model to predict future stock prices based on date. Afterwhich, I will utilise the logistic regression model to predict whether the stocks price will increase or decrease before the market opens for that particular day.

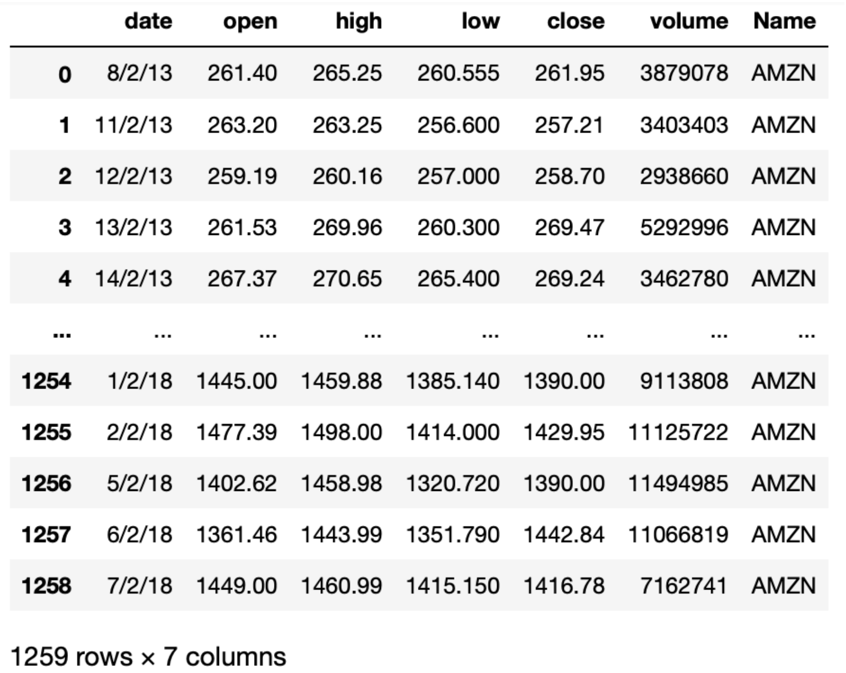
The linear regression model will be classified a success if the line of best fit produced fits the data accurately which means the model can be used to accurately predict the future price of the stock. The logistic regression model will be classified a success if the model accurately predicts if the price of the stock rose/fell the next day. This is so that if tomorrow’s closing price is higher than today’s closing price, we can plan to buy the stock, else we will plan to sell it.

Overview Of Data

1. Dataset was obtained from Kaggle.com: Amazon Stock Price data

2. Original data source was 1259 rows x 7 columns. Columns include “date”, “open”, “high”, “low”, “close”, “volume”, “name”

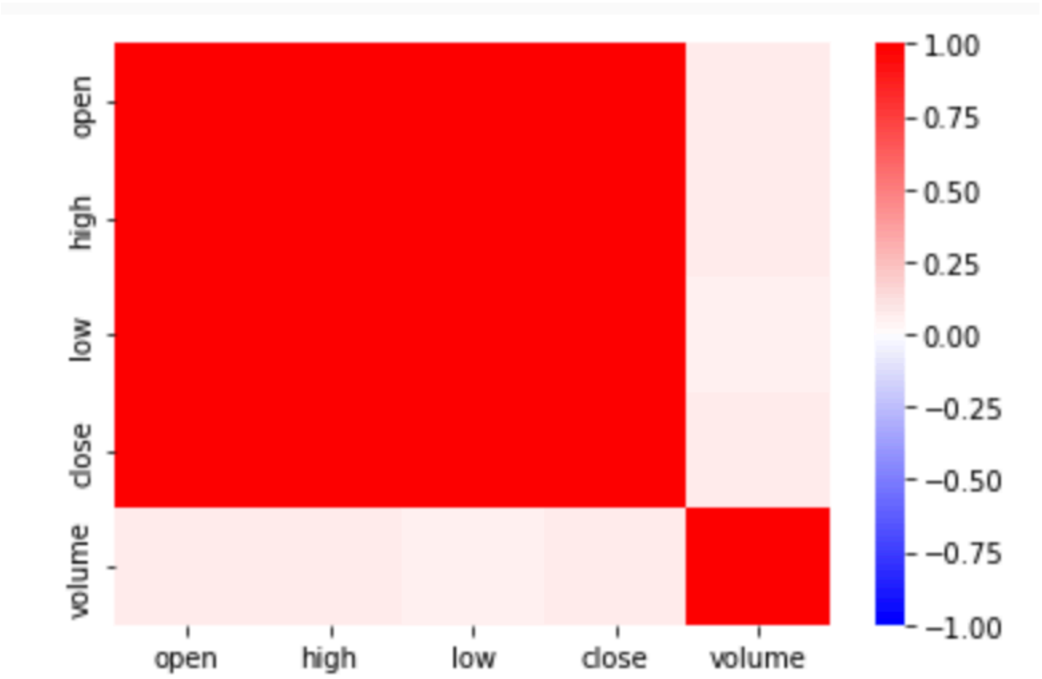
3. Data was checked for any null values but fortunately the dataset had no null values as identified by the functions .info() & .isna.sum().



	date	open	high	low	close	volume	Name
0	8/2/13	261.40	265.25	260.555	261.95	3879078	AMZN
1	11/2/13	263.20	263.25	256.600	257.21	3403403	AMZN
2	12/2/13	259.19	260.16	257.000	258.70	2938660	AMZN
3	13/2/13	261.53	269.96	260.300	269.47	5292996	AMZN
4	14/2/13	267.37	270.65	265.400	269.24	3462780	AMZN
...	...	...	...	...	...	...	...
1254	1/2/18	1445.00	1459.88	1385.140	1390.00	9113808	AMZN
1255	2/2/18	1477.39	1498.00	1414.000	1429.95	11125722	AMZN
1256	5/2/18	1402.62	1458.98	1320.720	1390.00	11494985	AMZN
1257	6/2/18	1361.46	1443.99	1351.790	1442.84	11066819	AMZN
1258	7/2/18	1449.00	1460.99	1415.150	1416.78	7162741	AMZN

1259 rows × 7 columns

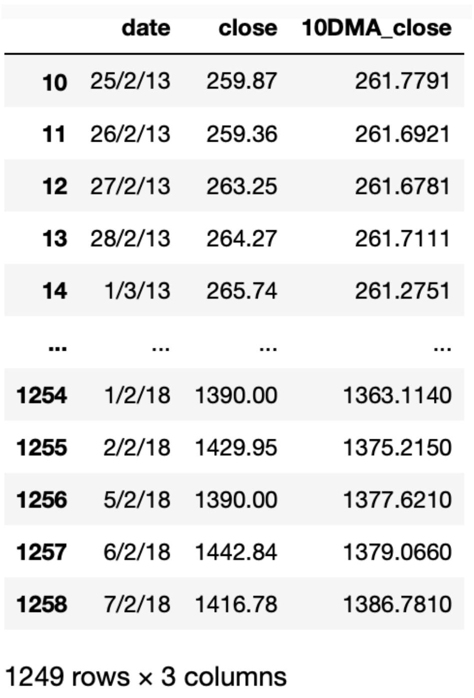
Findings From Correlation Matrix



From the heatmap's results, it can be deduced that there isn't any correlations between any of the given data.

Updated Dataset

To better predict the price of the stock in the future using a Linear Regression model, I will be utilising the explanatory variable of Moving Averages. To make the data set more precise, I will be dropping all the columns except for the "date" & "close" column which reflects the closing price of the stock for a particular day. I will also be adding the the column named "10MDA_close" which represents the 10 day moving average price based on the close price of the stock for a particular day. I also dropped the first 10 rows of null values as it is not possible for a 10 day moving average for the first 10 days of data.

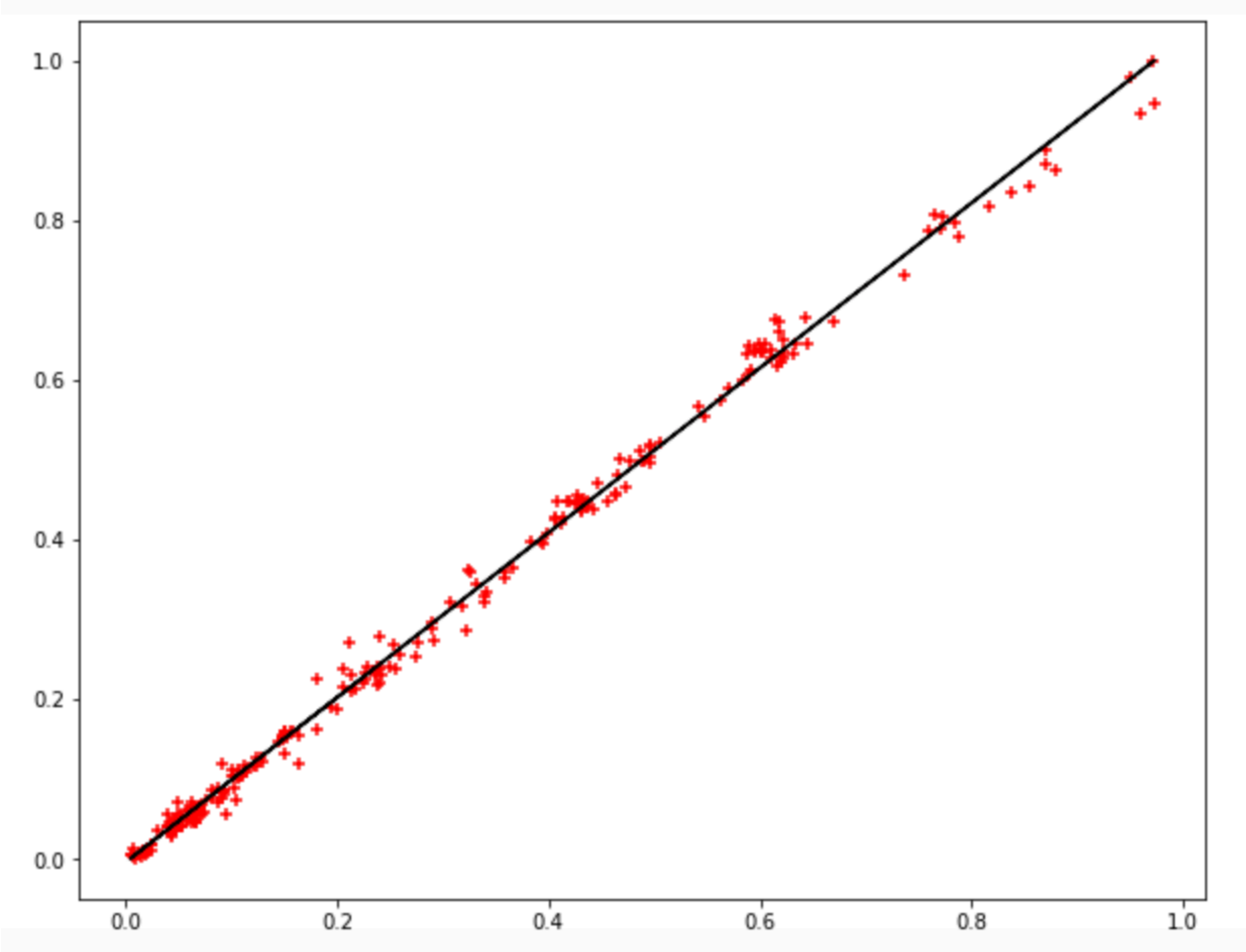


	date	close	10DMA_close
10	25/2/13	259.87	261.7791
11	26/2/13	259.36	261.6921
12	27/2/13	263.25	261.6781
13	28/2/13	264.27	261.7111
14	1/3/13	265.74	261.2751
...	...	...	...
1254	1/2/18	1390.00	1363.1140
1255	2/2/18	1429.95	1375.2150
1256	5/2/18	1390.00	1377.6210
1257	6/2/18	1442.84	1379.0660
1258	7/2/18	1416.78	1386.7810

1249 rows × 3 columns

Model Performance

Linear Regression was used in this project

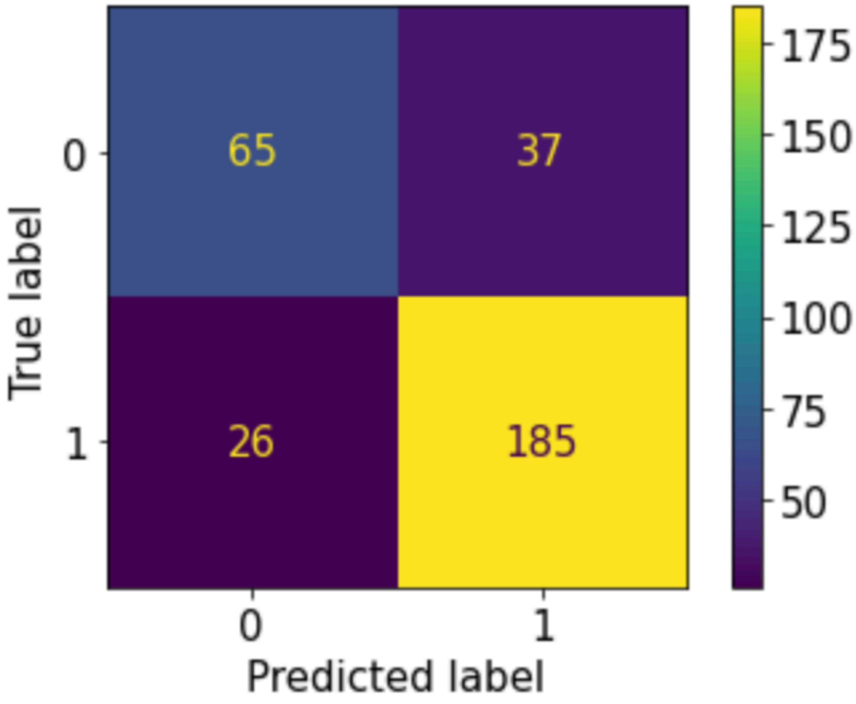


- Since the line cuts through majority of the points on the scatter plot, we can deduce that the model is accurate & the linear regression model can be used to predict the closing price of the stock on a particular day.

```
- mean squared error: 0.00024087323075501444
  r2: 0.9963463053024904
```

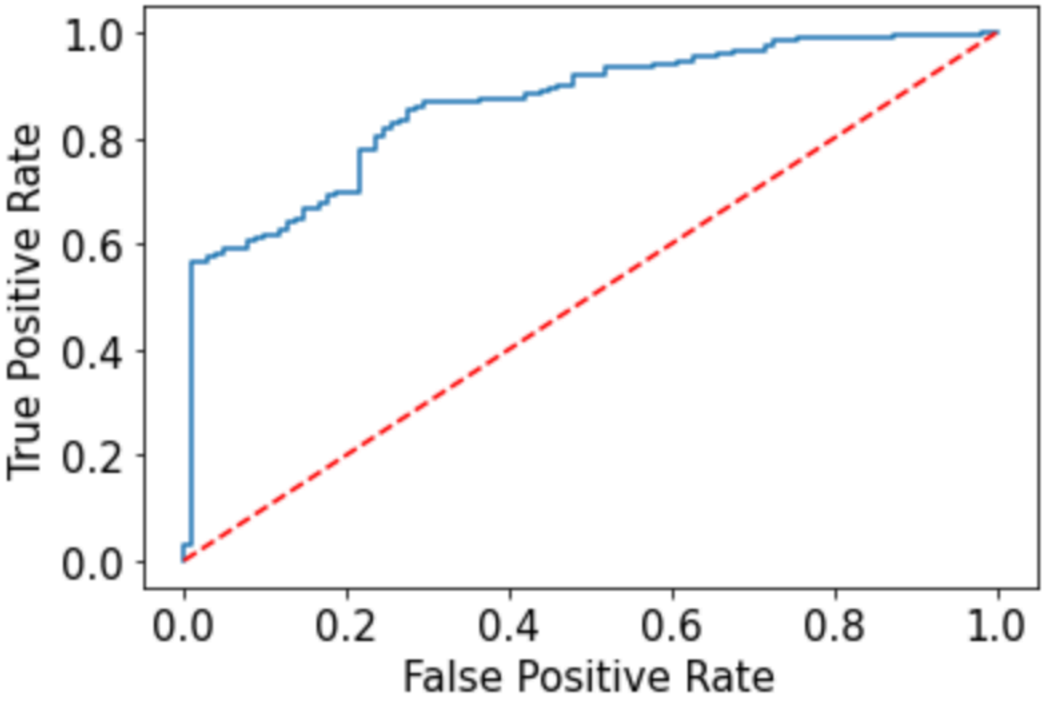
– Since the mse value is relatively low & the r2 value is very close to 1, it highlights that the data is a good fit for the regression model & the model will be able to predict the price of the stock accurately.

Logistic Regression was also used in this project



- The confusion matrix above is displayed for the train and test data with Logistic Regression

- The accuracy score was 0.7987220447284346



- The accuracy score was 0.8622339931233157
- Since the ROC curve is not close to the diagonal line & relatively close to the perfect classifier, we can deduce that the model fits the dataset well & can be used to predict if the price of the stock will rise or fall the next day.

Summary

1. Used the Linear Regression model to predict the future stock prices of the AMZN stock.
2. Used the Logistic Regression model to predict if the price of the stock will rise or fall the next day.
3. Both models yielded relatively accurate results that were determined by the line of fit for the Linear regression model & the AUC score for the Logistic regression model.
4. Decision Tree model & KNN model can be applied alongside the Logistic Regression model to determine which one of the 3 classifier model return the most accurate result.
5. Limitation of the dataset is that it contain irrelevant features such as the daily high & the daily low price of the stock that does not have much relevance to my prediction.
6. Hyperparameter tuning could be utilised to improve the accuracy score of the model.
7. Since the observed values in our dataset are recorded in sequence, it could possibily result in autocorrelation. We expect a dataset to not have autocorrelation for the Linear Regression model to be fully accurate, hence the accuracy could have been compromised as a result.
8. We can attain more information regarding the stock by including commonly used technical indicators such as the Relative Strength Index alongside Moving Averages to attain a higher prediction accuracy.