Report for

Stock Sentiment Analysis using Machine Learning

Shoilayee Chaudhuri- 21112102

Problem Overview:

The aim of the project is to develop a sentiment analysis model to predict the movement of stock prices based on textual data from news articles, social media posts, and other sources of financial news and opinions. By analysing the sentiment expressed in these texts, the model will seek to uncover insights into investor sentiment and market sentiment, which can be valuable indicators for making informed trading decisions.

Solution Overview:

Data for 4 stocks, namely – Apple, Microsoft, Google and Amazon were collected. Data consisted of historical data for stock open, close, high, low etc and news related to stocks that was used to score the sentiment.

Flow of solution:

1. Collect Historical Data:

- Historical price data for each stock in the portfolio was collected
- Data is taken from Yahoo finance for the last 5 years.
- Data includes open, close, high, low, volume, dividends and stock splits according to dates.
- Two new columns, Movement and Previous Close were calculated and added to this.
- Movement indicated if the stock close price increased/decreased.

Out[19]:

	Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Ticker	Previous Close	Movement
0	2019-06-19 00:00:00-04:00	48.242993	48.291316	47.670398	47.805695	84496800	0.0	0.0	AAPL	NaN	Decrease
1	2019-06-20 00:00:00-04:00	48.409701	48.467686	47.844354	48.189846	86056000	0.0	0.0	AAPL	47.805695	Increase
2	2019-06-21 00:00:00-04:00	48.030392	48.525675	47.873348	48.025558	191202400	0.0	0.0	AAPL	48.189846	Decrease
3	2019-06-24 00:00:00-04:00	47.967572	48.358969	47.878180	47.977238	72881600	0.0	0.0	AAPL	48.025558	Decrease
4	2019-06-25 00:00:00-04:00	47.941000	48.141530	47.182371	47.250023	84281200	0.0	0.0	AAPL	47.977238	Decrease

2. Collect Text Data:

- News headlines related to the stocks on different dates were collected from Google news.
- Beautiful-soup and requests were used for this data.
- Vader sentiment analysis was used to get a sentiment score of the headlines.

Out[69]:

	Date	Sentiment	Stock Symbol
0	19/06/2024	0.9257	AAPL
1	18/06/2024	0.9804	AAPL
2	17/06/2024	0.9895	AAPL
3	16/06/2024	0.9179	AAPL
4	15/06/2024	0.8807	AAPL
...	...	...	...
1825	21/06/2019	0.9075	AMZN
1826	20/06/2019	0.9793	AMZN
1827	19/06/2019	-0.6391	AMZN
1828	18/06/2019	0.9931	AMZN
1829	17/06/2019	0.9947	AMZN

3. Label the stocks for supervised learning:

- Sentiment score was mapped according to the dates and stock symbol to the data-frame containing historical price data.

Out[122...

	Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Ticker	Previous Close	Movement	Sentiment
0	2019-06-20	48.409701	48.467686	47.844354	48.189846	86056000	0.0	0	AAPL	47.805695	Increase	0.0000
1	2019-06-21	48.030392	48.525675	47.873348	48.025558	191202400	0.0	0	AAPL	48.189846	Decrease	0.2500
2	2019-06-24	47.967572	48.358969	47.878180	47.977238	72881600	0.0	0	AAPL	48.025558	Decrease	0.3182
3	2019-06-25	47.941000	48.141530	47.182371	47.250023	84281200	0.0	0	AAPL	47.977238	Decrease	0.9381
4	2019-06-26	47.781548	48.559506	47.680076	48.271999	104270000	0.0	0	AAPL	47.250023	Increase	0.2023

4. Model and training:

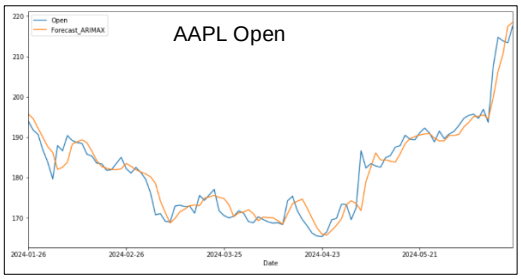
- Due to the sequential nature of the data, we use time-series forecasting model – ARIMA (Autoregressive integrated moving average). It predicts future values based on past values. ARIMA makes use of lagged moving averages to smooth time series data.
- Since we need to predict the price of the stock for a day, we cannot use the feature values of the same day since they will be unavailable at actual inference time.
- Thus, we need to use statistics like mean, standard deviation of their lagged values.
- We will use three sets of lagged values, one previous day, one looking back 7 days and another looking back 30 days as a proxy for last week and last month metrics.

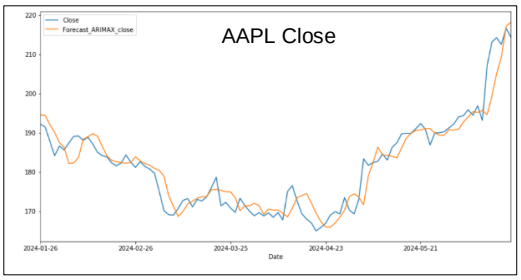
Out[7]:

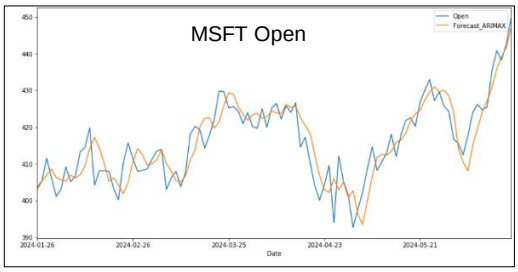
Low_std_lag3	Low_std_lag7	Low_std_lag30	Volume_mean_lag3	Volume_mean_lag7	Volume_mean_lag30	Volume_std_lag3	Volume_std_la
1.489431	2.336112	4.924371	96304672.0	96288360.0	96565608.0	18478110.0	2327646
1.489431	2.336112	4.924371	86056000.0	86056000.0	86056000.0	18478110.0	2327646
0.020502	0.020502	0.020502	138629200.0	138629200.0	138629200.0	74349736.0	7434973
0.018295	0.018295	0.018295	116713336.0	116713336.0	116713336.0	64844868.0	6484486
0.400338	0.341788	0.341788	116121736.0	108605296.0	108605296.0	65271108.0	5537326

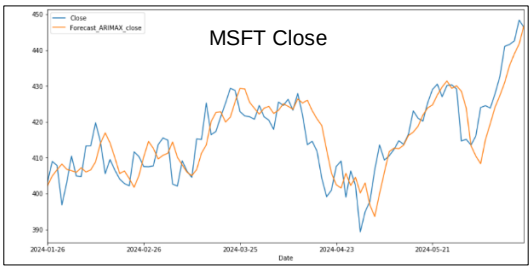
- Next we use auto_arima to separately train on the open and close data stock-wise. It is then used to forecast for the last 100 datapoints of the dataset.

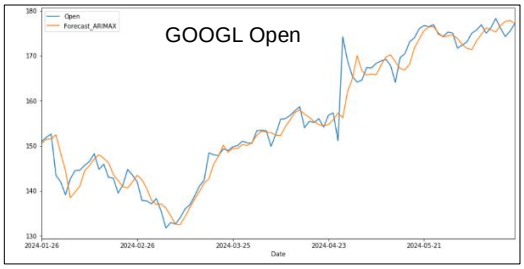
- Visualization helps to give a better insight in the predicted data.
- Matplotlib is used to plot charts wherever required.
- Below plots show the actual and predicted open and close price for each of the stocks.

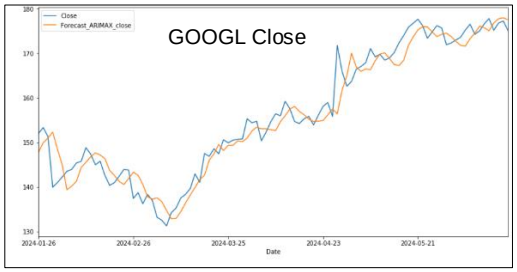


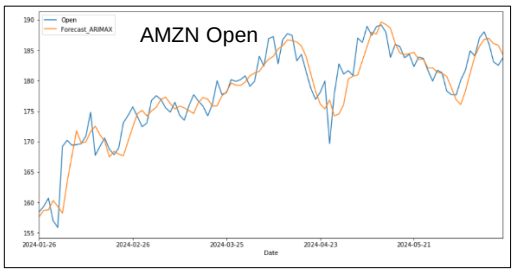


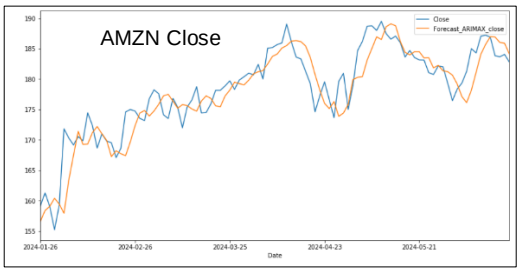












5. Metrics:

1. Sharpe Ratio
   ~ The Sharpe ratio measures the risk-adjusted return of an investment or trading strategy. It's calculated as the ratio of the excess return of the portfolio over the risk-free rate to the standard deviation of those returns.
2. Maximum Drawdown
   ~ Maximum drawdown measures the largest single drop from peak to trough in the value of a portfolio or asset.
3. Number of Trades
   ~ This metric simply counts the total number of buy and sell trades executed over the evaluation period.
4. Win Ratio
   ~ The win ratio measures the proportion of profitable trades out of the total number of trades.

5. RMSE (Root Mean Squared Error)
   ~ A measure of the differences between predicted and actual values, calculated as the square root of the average of squared differences.
6. MAE (Mean Absolute Error)
   ~ A measure of the average magnitude of errors in a set of predictions, calculated as the average of the absolute differences between predicted and actual values.

```
Sharpe Ratio: 1.83
Maximum Drawdown: 10.25%
Number of Trades: 208
Win Ratio: 55.00%
```

```
RMSE of Auto ARIMAX open: 3.611684255796013

MAE of Auto ARIMAX open: 2.5657522167943903
```

```
RMSE of Auto ARIMAX close: 4.340052499816127

MAE of Auto ARIMAX close: 3.1557381018739297
```

Conclusion:

The project successfully developed a stock sentiment analysis model to predict the movement of stock prices based on textual data from news articles and financial news sources.

Challenges and Limitations:

Due to computational resource constraints, the dataset could not be expanded further. This limitation restricted the model's training and testing on a more extensive dataset, which might have led to even more precise results. Despite this, the model demonstrated decent performance.

Future Work:

With additional computational resources, future iterations of this project could involve expanding the dataset to include more stocks and a longer historical period. This expansion would likely enhance the precision and reliability of the model's predictions, further supporting informed trading decisions.