PROJECT SUMMARY

DATA QUALITY REPORT

Data Source

The data been worked on is financial times series data set, to be explored for insight extraction as well as model prediction. The data sets include; company information, stock prices, price, and market indices.

Data overview

Upon loading the data sets, we can view the various columns in each set and their respective data types. The following was observed;

a) DATA TYPES MISMATCH OVERVIEW

- Ticker column appear as object rather than strings
- Date column also appears as object rather than a date datatype
- ipo_date column is also an object rather than a date datatype¶

b) MISSING VALUES OVERVIEW

- Stock prices data set, had presence of missing values. All other data sets were free of missing values.

c) DUPLICATES OVERVIEW

- There were no observed duplicates in the data sets

DATA ISSUES ADDRESSED

- Data type corrections needed = Ticker and date columns would be corrected to ensure they reflect their true data types (i.e string and date type)

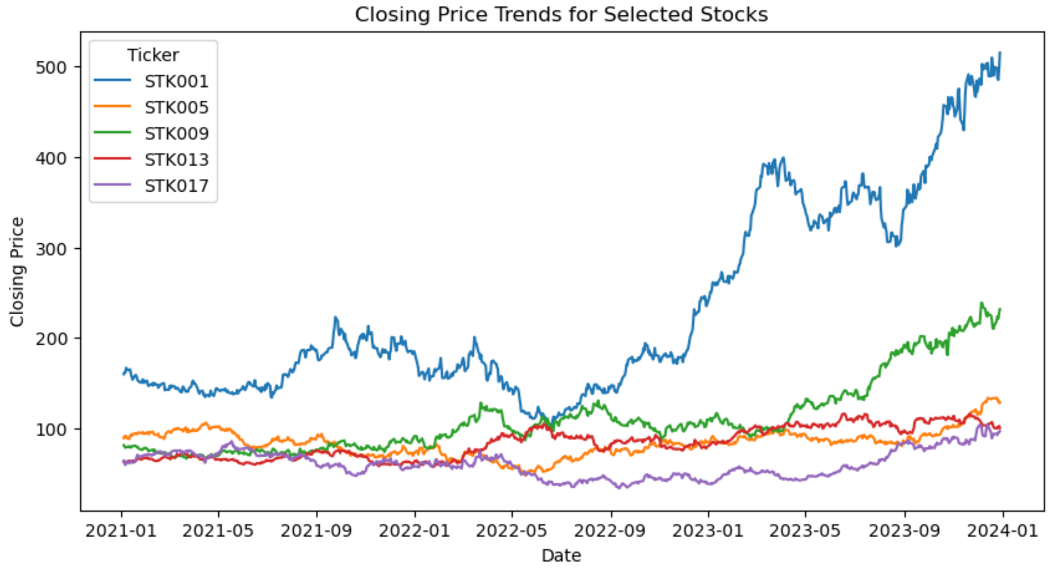
- Missing dates discovered: - i.e either applying Forward fill (ffill) which carries last observation forward (common for financial data). or - Backward fill (bfill) method which fills with next available value.

Data was cleaned and ready for further use.

EXPLORATORY DATA ANALYSIS SUMMARY

After data cleaning was done, an exploratory data analysis was carried on the data sets to draw brief insights.

1. **Price Trend Analysis**: Random selection of five sample stocks from five sectors was done to observe price moving over the years. Focus was more on the closing price (i.e the final prices at market close for each period).



Sector	**Ticker**
Consumer	STK013
Energy	STK017
Finance	STK009
Healthcare	STK005
Technology	STK001

INSIGHT:

- Stock with ticker **STK001(Technology)** shows the highest upward trend followed by the stock **STK009(Finance).**

- The ticker **STK013(Consumer)** exhibits a sideways trend. same as the **Healthcare stock(STK005).**
- The **Energy stock(STK017),** however, exhibits more of a downward trend over the years.¶

2. SUMMARY STATISTICS ON SAMPLE STOCKS (i.e DAILY RETURNS USING STOCK PRICE DATA SET)

S/N	sector	Sample ticker	mean	median	std
0	Consumer	STK013	0.08	0.11	1.89
1	Energy	STK017	0.10	0.18	2.96
2	Finance	STK009	0.16	0.16	2.34
3	Healthcare	STK005	0.07	-0.01	2.15
4	Technology	STK001	0.18	0.18	2.45

*INSIGHT:

Mean : Shows the average daily return (%) i.e shows the typical daily gain/loss.

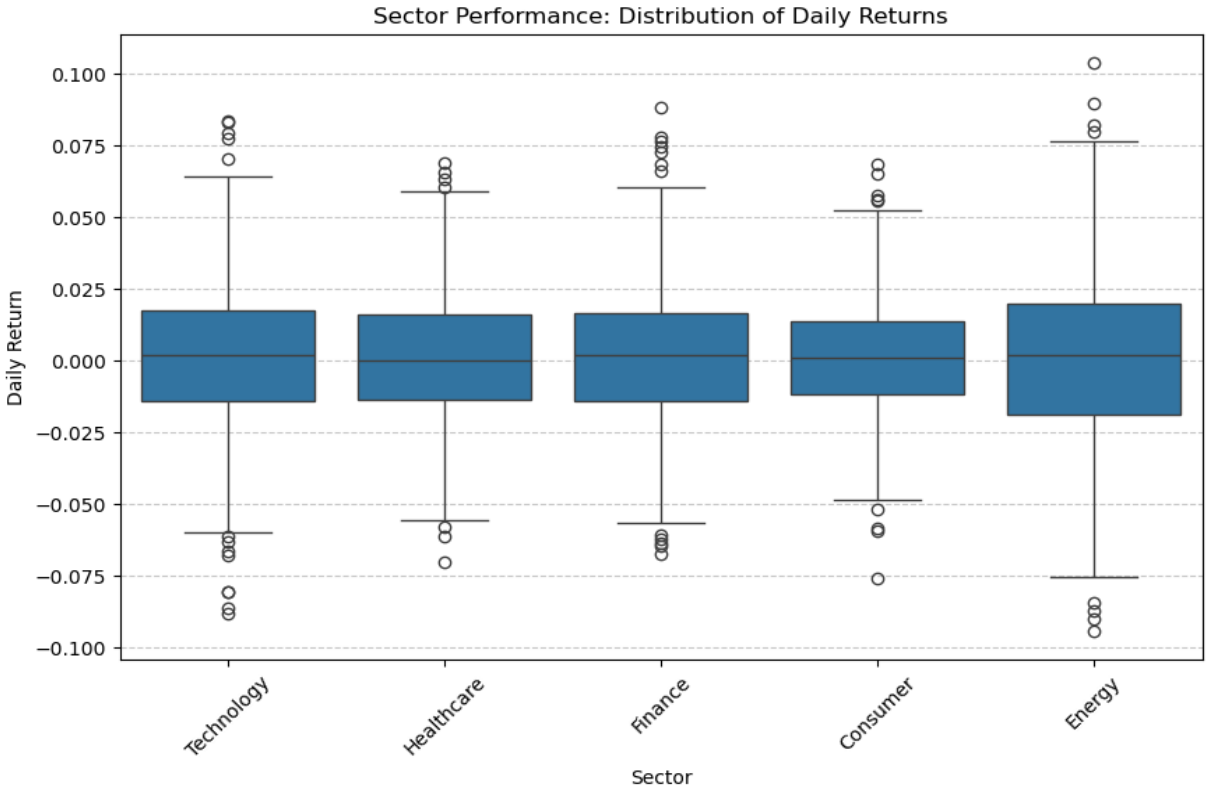
Median: shows the middle daily return (%).

Standard deviation of daily returns (%): i.e measures the volatility — how much the stock price fluctuates daily. Higher values mean more risk.

- Highest average daily return: Technology (STK001) and Finance (STK009).
- Most volatile stock: Energy (STK017) → highest std dev at 2.96%.
- Least volatile stock: Consumer (STK013) → lowest std dev at 1.89%.
- Healthcare (STK005) had slightly negative median daily returns → suggests more frequent small losses than gains.¶

3. SECTOR COMPARISON

NB: In the finance space, extreme values (i.e outlier) can are often meaningful signals rather than errors, hence not treated in the data cleaning stage.



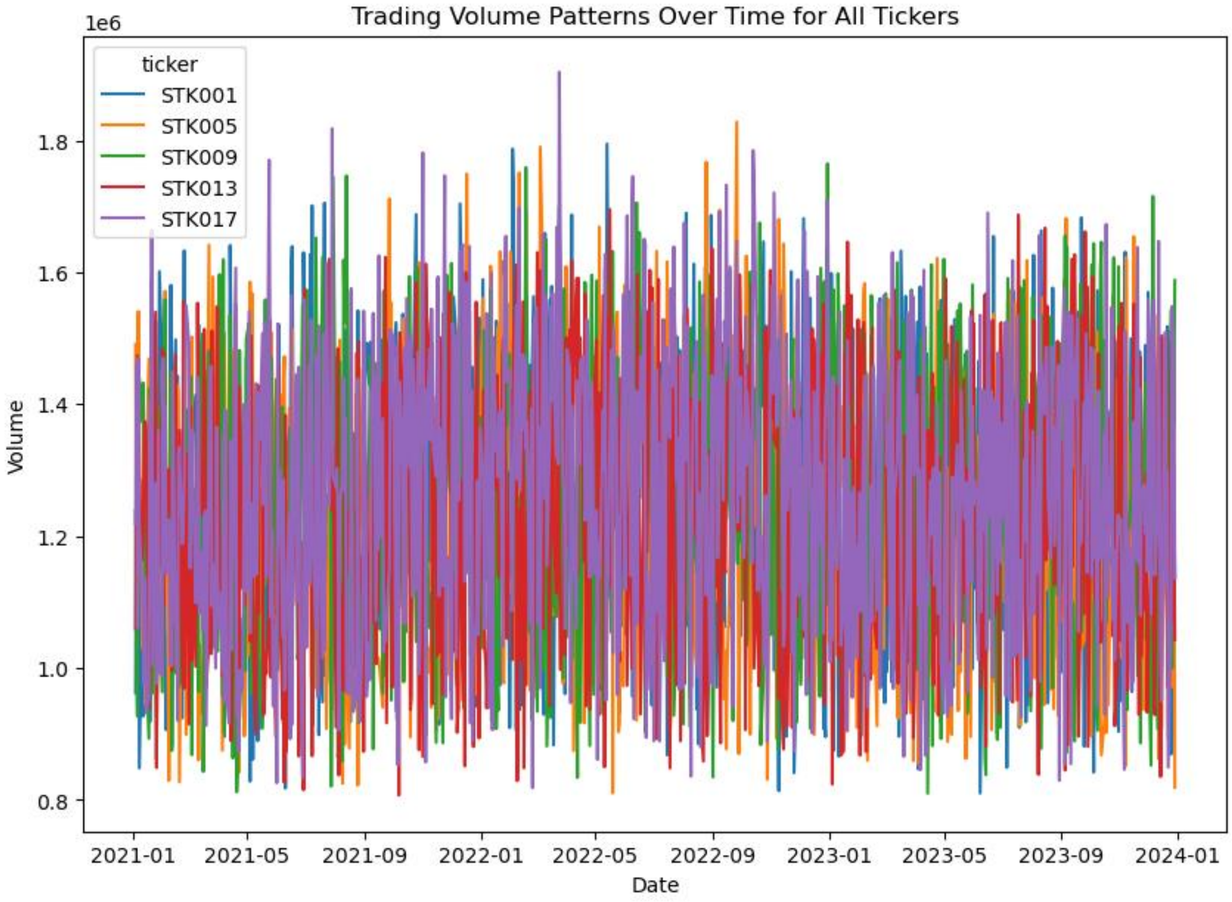
INSIGHT: Outliers represent stocks whose returns are far from the typical range in that sector.

- The Outliers in each sector reveal that some stocks perform better, giving higher returns, while some perform badly, giving negative returns.

- The Technology sector, for example, as seen above, has most of its stocks giving returns of 2.5%, but a few stocks give returns of above 7.5%, and negative returns of -7.5%.¶

4. TRADE VOLUME ANALYSIS

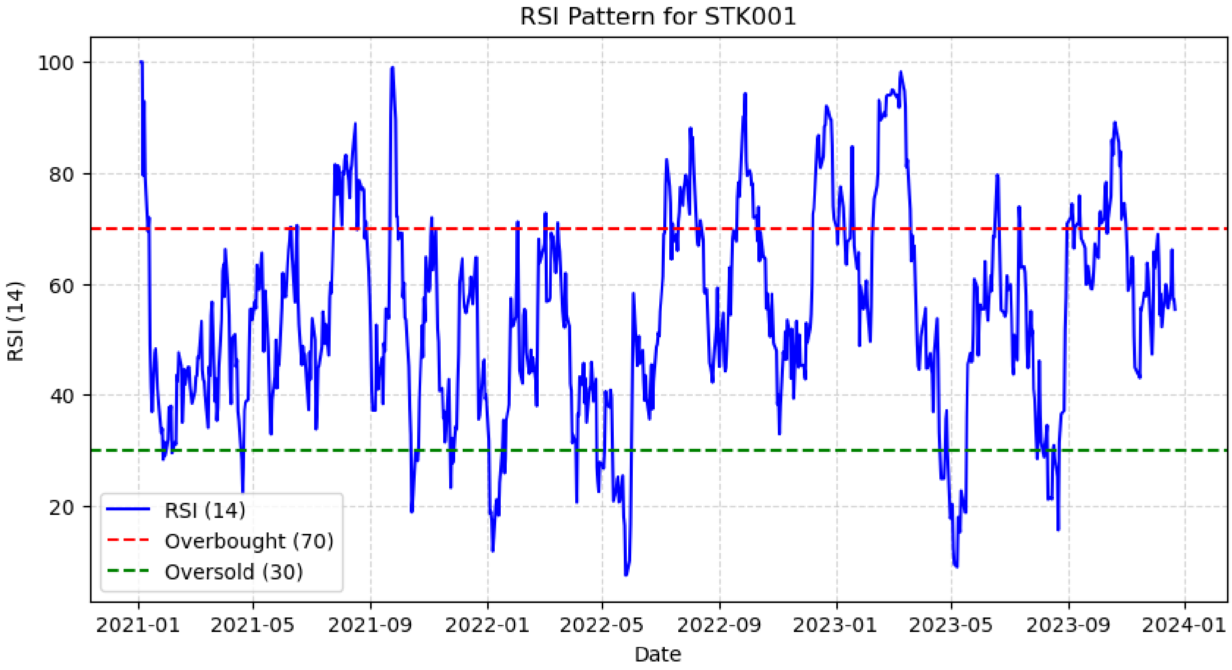
i.e Trading volume patterns over time.



INSIGHT: This reveals a large clustering around 1.0 to 1.4 million shares, implying stocks typically trade around this volume. **Stocks typically trade around 1 million shares to 1.4 million shares.**

5. **TECHNICAL INDICATOR EXPLORATION (Relative Strength Index)**
i.e Visualizing Relative Strength Index (RSI) patterns and identifying overbought/oversold conditions on a sample stock.

- RSI above 70 = Stock may be "too expensive" (overbought)
- RSI below 30 = Stock may be "too cheap" (oversold)
- RSI between 30 -70 = Neutral, thus trend is stable

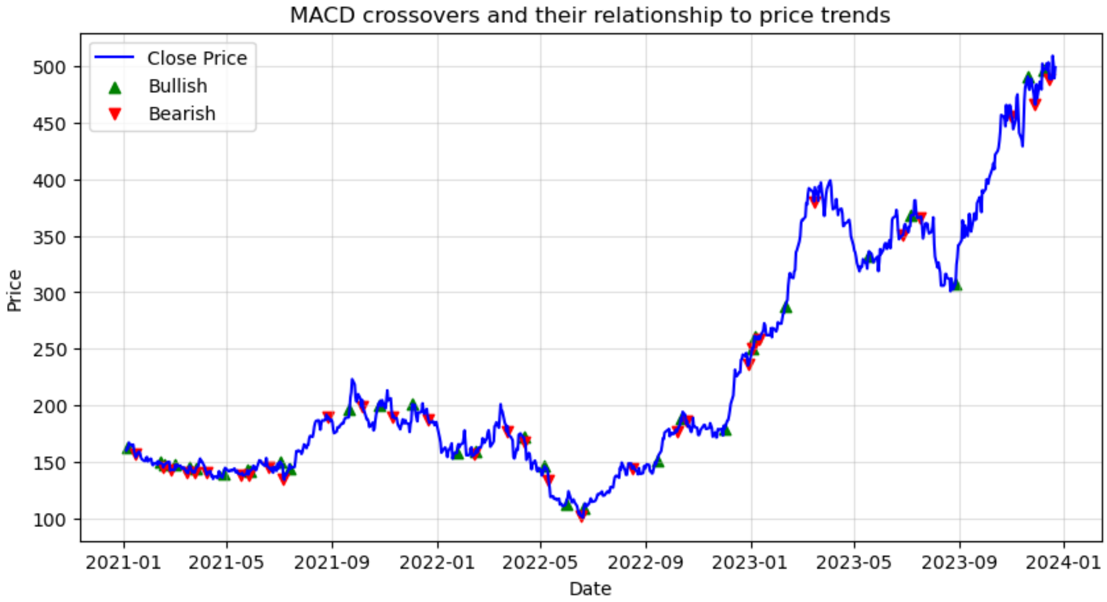


INSIGHT: for this sample stock (i.e STK001)

❖ The RSI (blue line) above 70 suggests the stock might be overbought — i.e too many investors have been buying it, pushing the price up to a level that may be unsustainable in the near term.

❖ The RSI below 30 suggests the stock might be oversold — i.e too many investors have been selling it, pushing the price below its fair value or creating an opportunity for a potential rebound.¶

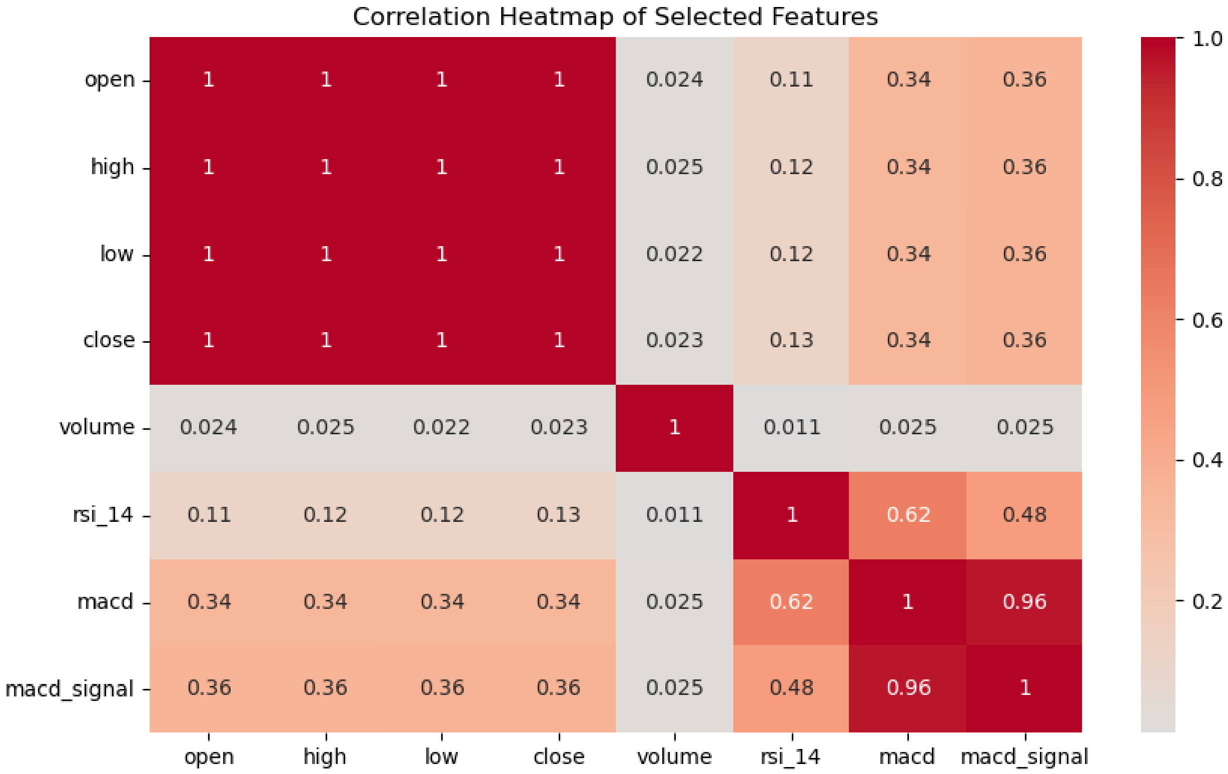
6. MOVING AVERAGE CONVERGENCE DIVERGENCE (MACD) TECHNICAL INDICATOR



INSIGHT:

- ❖ Momentum is not price direction but rather measures how fast prices are moving. So a rise in momentum (the line), implies prices are rising quickly – a strong bullish trend.
- ❖ As a momentum tool, it helps traders time trades effectively, i.e knowing at what points to buy or sell stocks. At bullish signals, they might buy, and at bearish signal, they might sell.

7. CORRELATION ANALYSIS



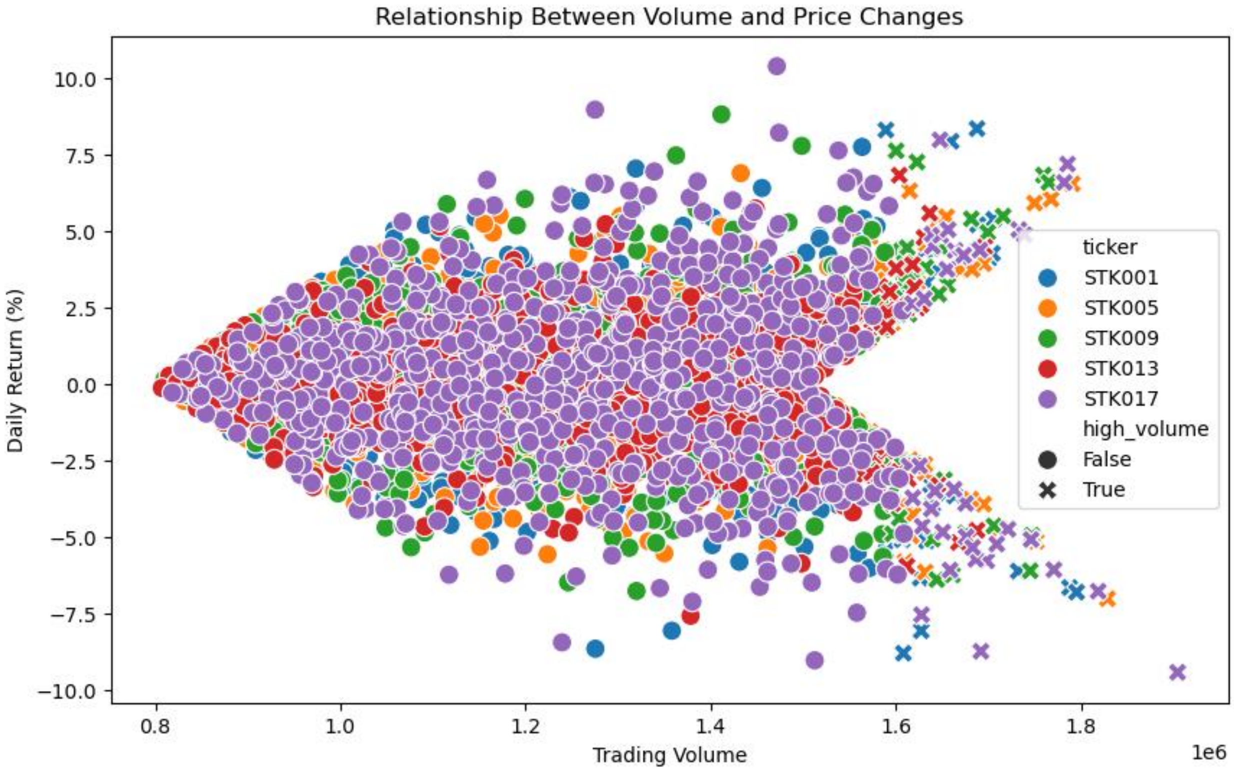
	open	high	low	close	volume	rsi_14	macd	macd_signal
open	1	1	1	1	0.024	0.11	0.34	0.36
high	1	1	1	1	0.025	0.12	0.34	0.36
low	1	1	1	1	0.022	0.12	0.34	0.36
close	1	1	1	1	0.023	0.13	0.34	0.36
volume	0.024	0.025	0.022	0.023	1	0.011	0.025	0.025
rsi_14	0.11	0.12	0.12	0.13	0.011	1	0.62	0.48
macd	0.34	0.34	0.34	0.34	0.025	0.62	1	0.96
macd_signal	0.36	0.36	0.36	0.36	0.025	0.48	0.96	1

INSIGHT

This reveals that every change in one column, for instance ('open', 'high', 'low', 'close), is exactly proportional to changes in the others. For example, if open goes up by 1 unit, high, low, and close also go up by a fixed amount.¶

- ❖ It helps investors spot leading/lagging indicators. For instance rsi_14 vs macd suggest a strong correlation hence help investors know these indicators move in sync with stock prices. They can they use it to time entry or exit points.

8. VOLUME PRICE CHANGE ANALYSIS



INSIGHT: Most points are clustered around low to moderate trading volumes and small daily returns, implying the “normal” trading days, and small price changes with the trading volume.¶

MODEL PERFORMANCE REPORT

The models been trained on our data sets were the;

-Logistic Regression Model

- Random Forest Classifier Model.

Goal/Aim: Was to train these model to gain some level of accuracy in predicting our target variable [Thus, the **future return**: Uptrend, downtrend, or sideways]

1. Logistic Regression Model:

Logistic Regression Model's classification report

```
             precision    recall  f1-score   support

           0       0.45      0.01      0.02      1356
           1       0.39      0.17      0.24      1583
           2       0.37      0.84      0.51      1712

    accuracy                           0.37      4651
   macro avg       0.40      0.34      0.26      4651
weighted avg       0.40      0.37      0.27      4651
```

Insight:

Class 0 (downtrend):

- Precision 0.45 → only 43% of predicted class 0 were correct.
- Recall 0.01 → the model almost never correctly identifies true class 0
- F1-score 0.02 → extremely poor performance

Class 1 (sideways):

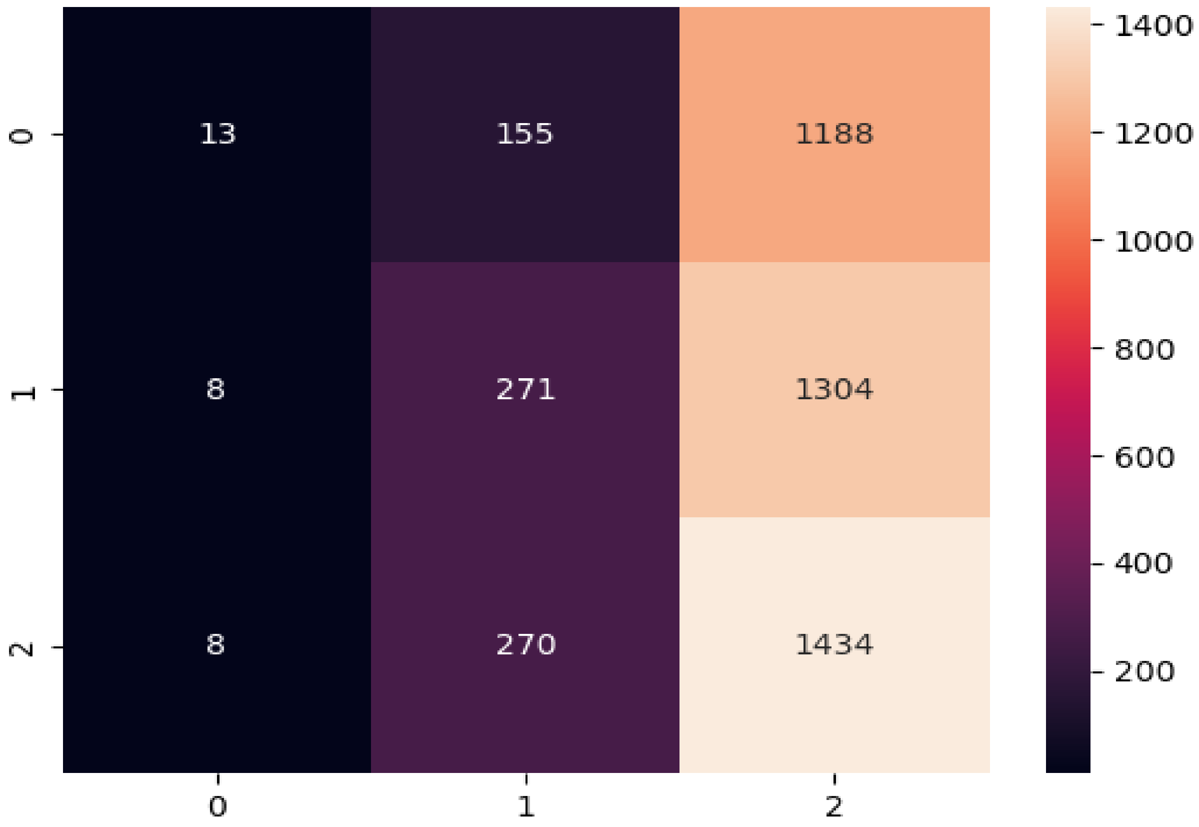
- Precision 0.39 → 39% of predictions for class 1 were correct
- Recall 0.17 → the model slightly identifies actual class 1 better, but still misses most instances
- F1-score 0.24 → low
-

Class 2 (uptrend):

- Precision 0.37 → 37% of predictions for class 2 were correct
- Recall 0.84 → the model successfully captures most actual class 2 samples
- F1-score 0.51 → moderate performance

Observation: Using class weight='balanced' helped the model detect the minority class (1) better. Class 0 remains very difficult to predict. Overall, the model is still biased toward class 2, but the macro and weighted F1-scores improved slightly. Logistic Regression may be limited for this dataset — more advanced models are likely needed.

Confusion Matrix



The confusion matrix helps see how well the model predicted the various classes accurately. In this for instance, the Logistic regression model accurately predicted:

- class 0, 13 times
- Class 1, 271times
- Class 3, 1434 times

2. Random Forest Classifier Model

Random Forest Model's classification report

```
Random Forest Model Accuracy after tuning: 0.35

Classification Report:
              precision    recall  f1-score   support

           0       0.28      0.14      0.19      1356
           1       0.38      0.38      0.38      1583
           2       0.36      0.50      0.42      1712

    accuracy                           0.35      4651
   macro avg       0.34      0.34      0.33      4651
weighted avg       0.34      0.35      0.34      4651
```

Class 0 (downtrend):

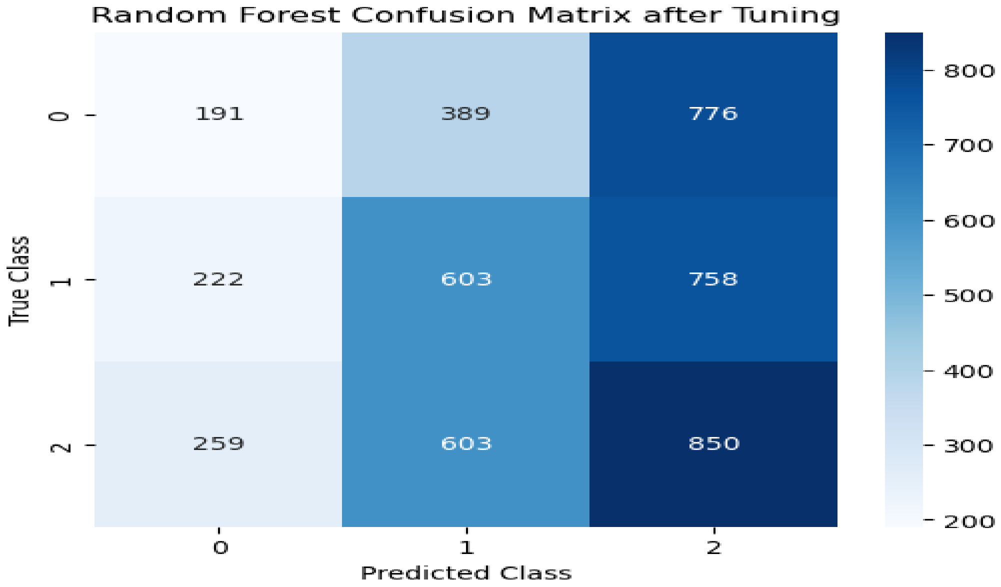
- Precision 0.28 → only 28% of predicted class 0 were correct.
- Recall 0.14 → the model slightly identifies true class 0
- F1-score 0.19 → slightly moderate performance

Class 1 (sideways):

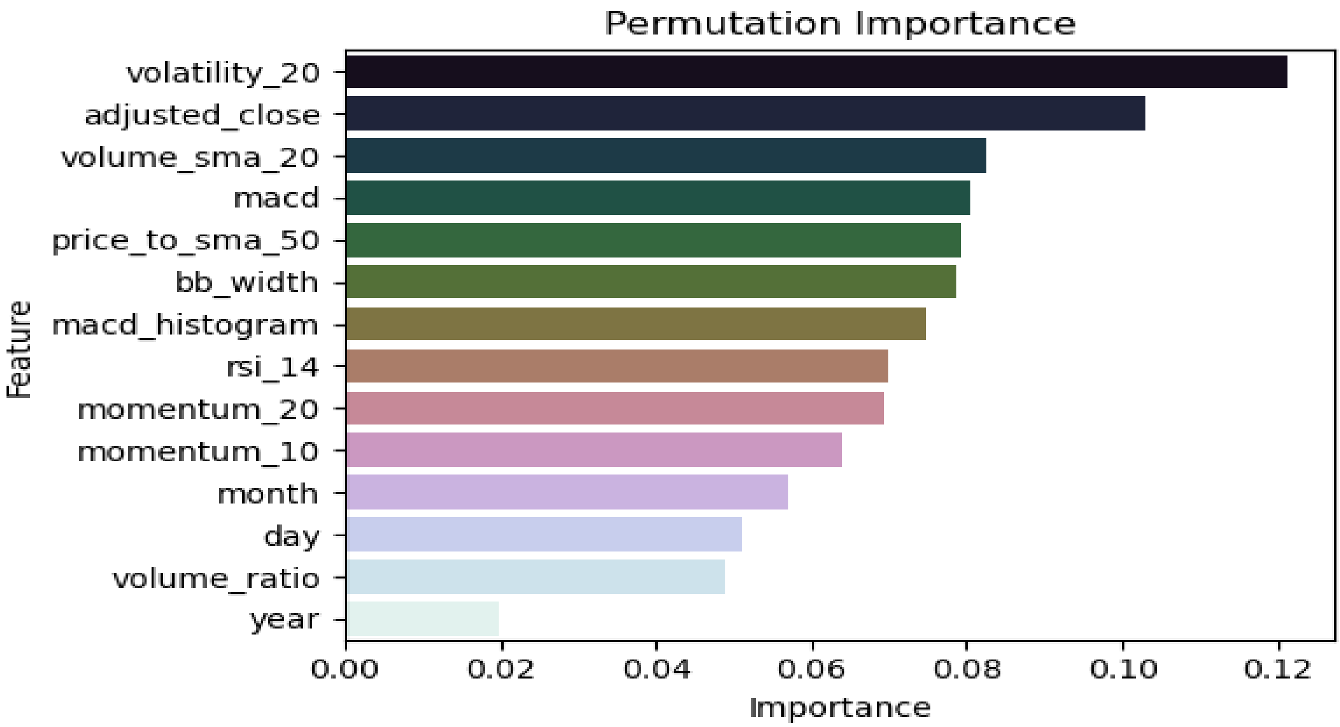
- Precision 0.38 → 38% of predictions for class 1 were correct
- Recall 0.38 → The model moderately identifies actual class 1 better, but still misses most instances
- F1-score 0.38 → slightly moderate
-

Class 2 (uptrend):

- Precision 0.36 → 36% of predictions for class 2 were correct
- Recall 0.50 → the model averagely captures most actual class 2 samples
- F1-score 0.42 → moderate performance


Random Forest Confusion Matrix after Tuning
True Class
0
1
2
191
389
776
222
603
758
259
603
850
0
1
2
Predicted Class
800
700
600
500
400
300
200

FEATURE IMPORTANCE



INSIGHT:
The feature importance reveals the variables which had the most influence on model predictions.

ESSENTIAL BUSINESS INSIGHT

- **Markets are noisy hence expected low accuracy of models**
  Financial markets contain a high level of randomness caused by news events, macroeconomic shifts, investor sentiment, and unexpected shocks. Because of this unpredictability, no model — especially one based purely on historical price data — can achieve very high accuracy.

  Business insight: Forecasts should be used as *decision-support*, not as guaranteed predictions. Trading or risk strategies must incorporate uncertainty, buffers, and scenario planning.

- Technical indicators lag price → model struggles in turning points
  Indicators like RSI and MACD react only *after* price movements occur. This makes it difficult for the model to detect rapid reversals or early trends.

  Business insight: Relying solely on lagging indicators can cause delayed entry/exit signals. Businesses may need complementary leading indicators (e.g., volume spikes, news sentiment, macro triggers) to improve timing.

- **Random Forest Classifier model** performs better than baseline but still limited

Clear next steps for future improvement

The results highlight specific opportunities to strengthen prediction performance:

- Incorporate more informative features (e.g., sentiment data, macroeconomic variables)
- Experiment with advanced models such as gradient boosting
- Use longer time windows or engineered features to reduce noise
- Explore ensemble blending or probabilistic forecasting

Business insight: The project provides a roadmap. With richer data and more sophisticated modeling, the system could evolve into a more reliable trend-prediction or risk-management tool.