STOCK PRICE

PREDICTION + BUY RECOMMENDATION

26 August 2023

CONTENTS

01

BACKGROUND

PROBLEM STATEMENT

- Formulate a model to recommend investors on purchase of specific stock

  “buy” if the predicted price (6 months later) is higher than current price

- NASDAQ market

PERFORMANCE METRICS

PRICE PREDICTION

Mean Absolute Percentage Error (MAPE)
Root Mean Square Error (RMSE)
Prediction Bias

BUY RECOMMENDATION

F1-score

02

DATA COLLECTION & CLEANING

DATA COLLECTION

EODDATA WEBSITE

- List of NASDAQ stock symbols

RAPIDAPI API

- OHLC (Open, High, Low, Close), Volume
- Company Profile
  Coy Name, Description, Sector, Industry
- Dividend Amount & Date

YAHOO FINANCE API

- Income Statements
- Balance Sheet
- Cash Flow

DATA CLEANING

EODDATA WEBSITE

- Assume all stocks are downloaded
  4653 symbols in total

RAPIDAPI API

- Missing stock symbols
- Stock symbol “NA”
- Missing stock profile
  e.g. Industry, Description
- Duplicated transaction dates
- Missing transaction volume

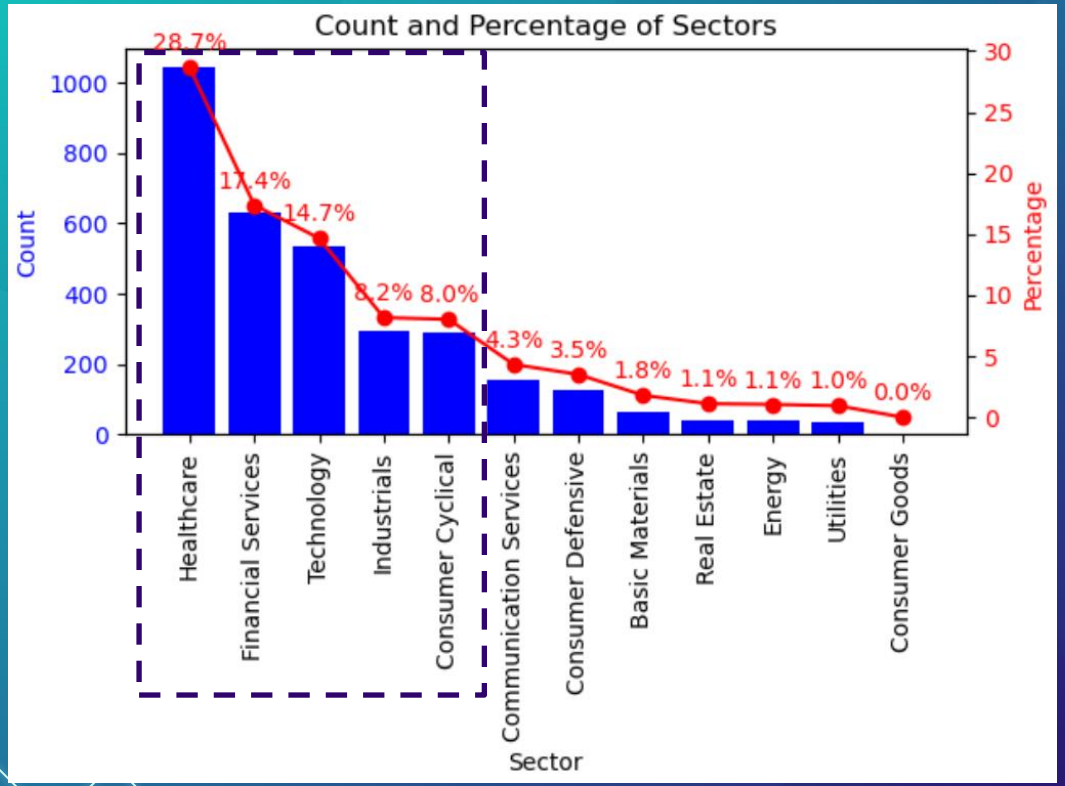
YAHOO FINANCE API

- A lot of missing data
  e.g. gain on sale of business, fixed assets revaluation reserve, cash from discontinued investing activities

03

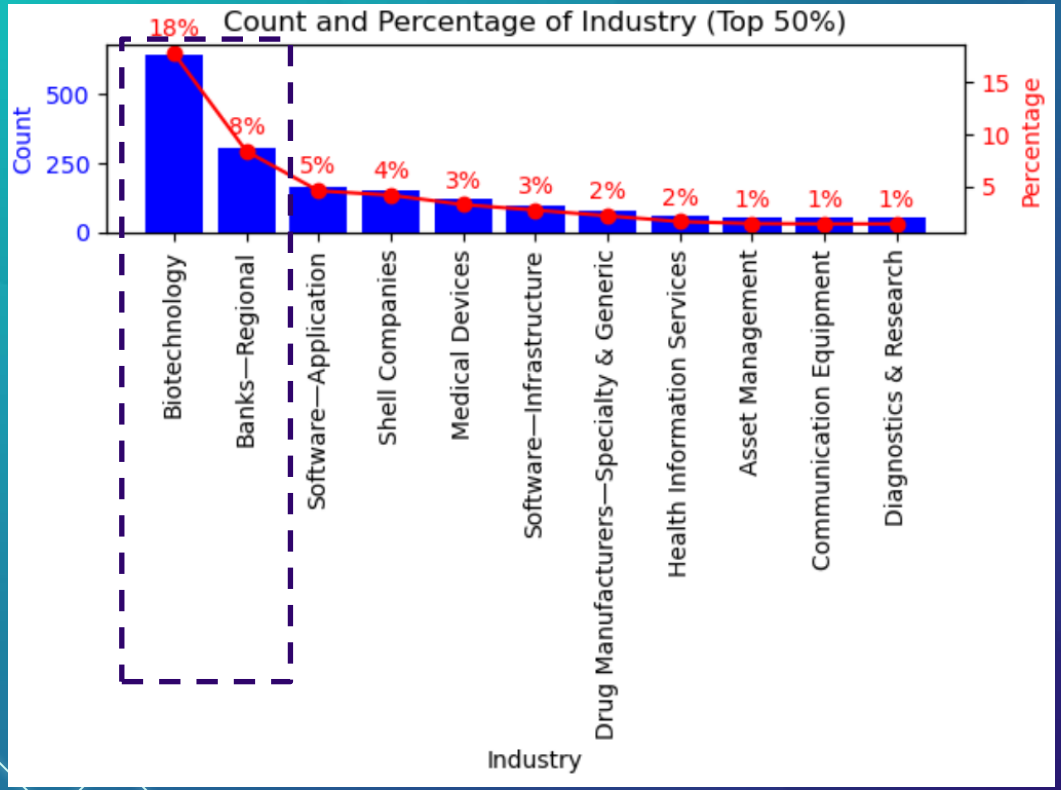
EDA

COMPANY PROFILE (3633 SYMBOLS EXTRACTED)



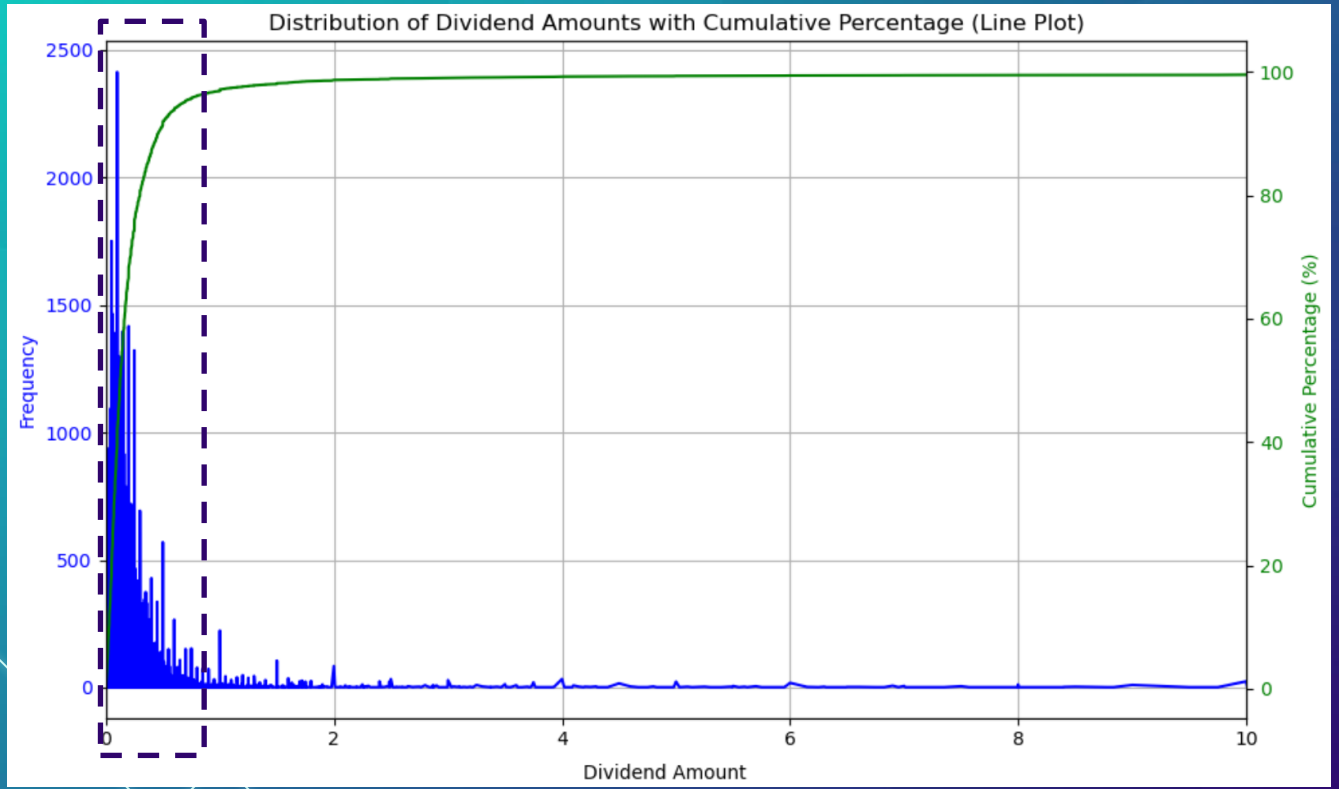
- 12 sectors in total
- Top 5 counts of sectors add up to more than 60% of the stocks

COMPANY PROFILE (3633 SYMBOLS EXTRACTED)



- 137 industries in total
- “Biotechnology” has the most number of stock symbols follow by “Banks – Regional”

DIVIDEND PROFILE (1904 SYMBOLS EXTRACTED)



- Most stock symbols have very low dividend ~ less than $1

OHLC PROFILE (4557 SYMBOLS EXTRACTED)

1%

HIGH VALUED

48 symbols
consistently more than $100

6%

PENNY STOCK

282 symbols
consistently less than $1

65%

"OTHERS"

2937 Symbols

28%

LOW VALUED

1290 symbols
consistently less than $10

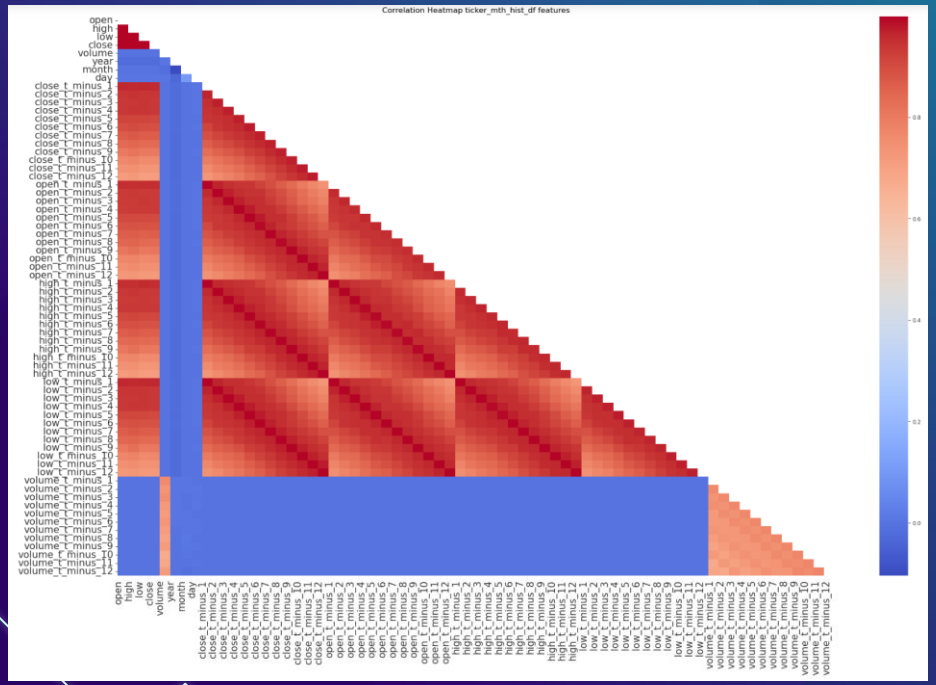
No observable trend amongst the 4 categories

04

MODEL

PRE-PROCESSING

CREATE TIME LAGGED OHLC FOR T-6 TO T-12

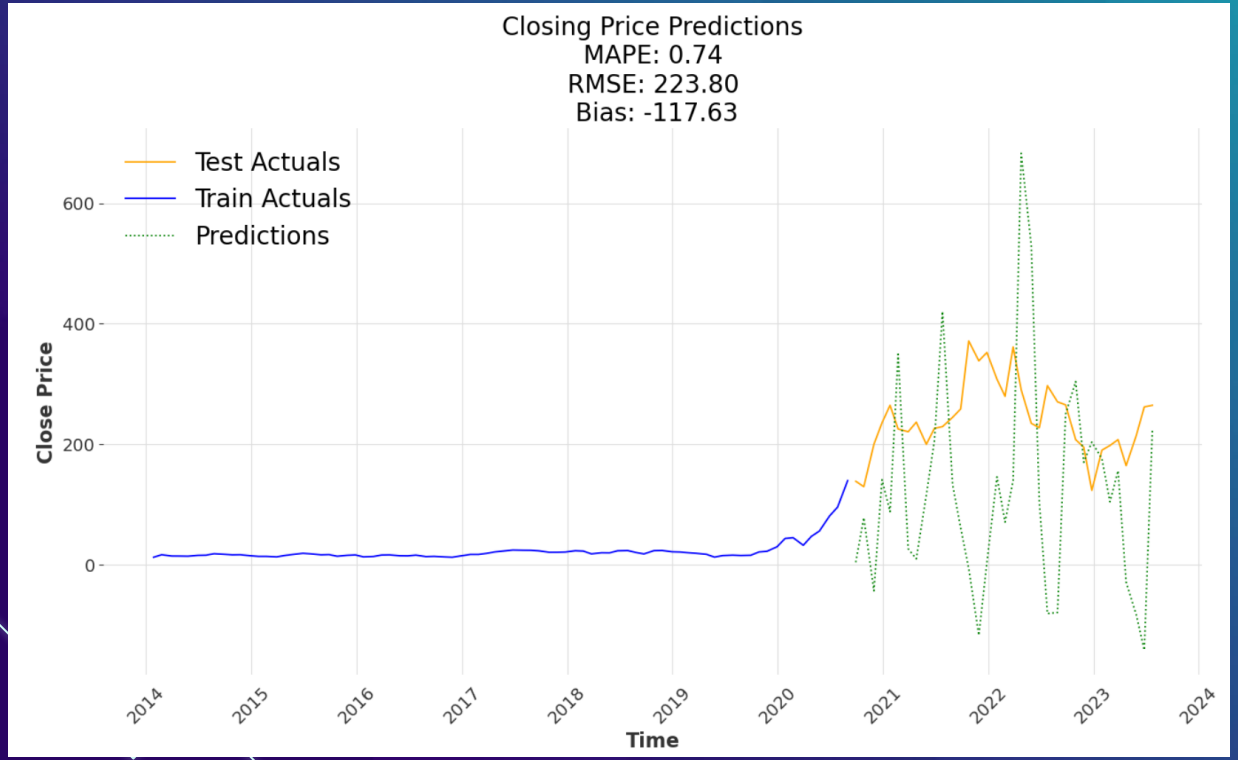


- OHLC values of the same day are closely related with each other
- Values with nearer time period have greater correlationship (e.g. T-1 is closely related to T-0, T-12 is closely related to T-11)

STOCK PRICE PREDICTION

BASE MODEL (TSLA)

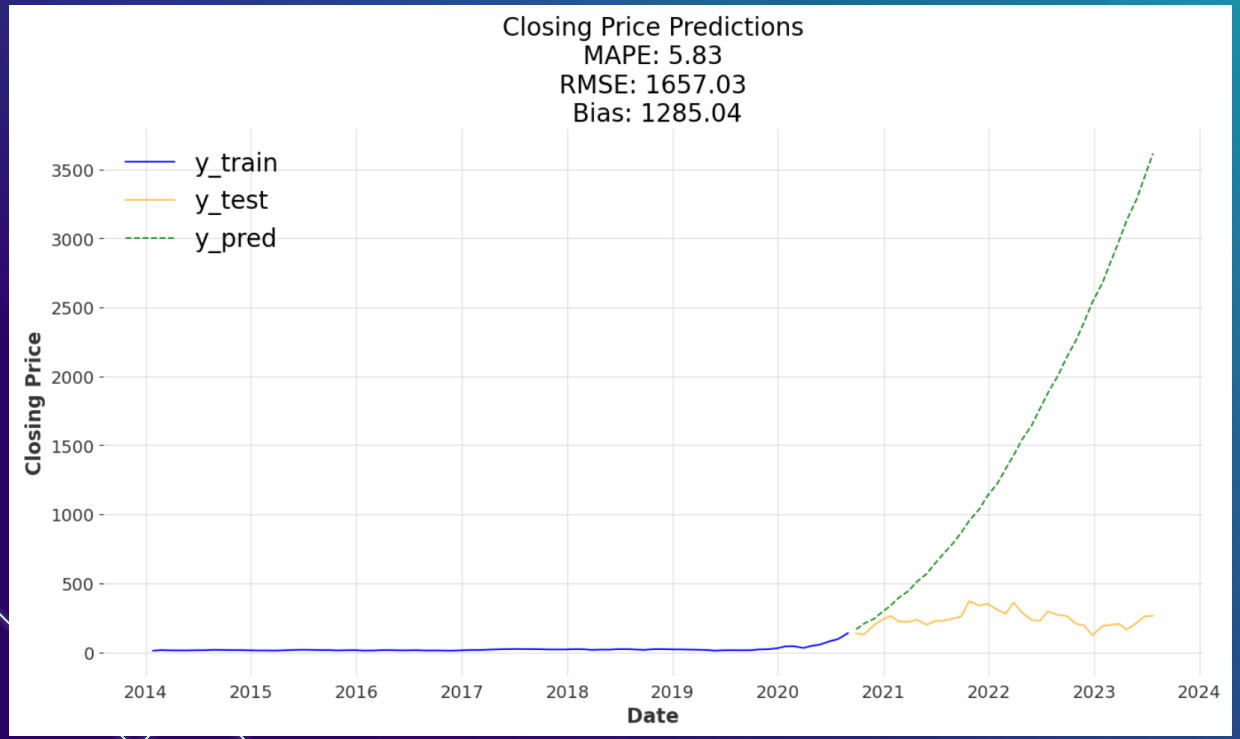
LINEAR REGRESSION



- Fit y_train (Close Price) and X_train (volume, OHLC T-6 to T-12)
- Predictions can be <$0 (not possible)
- Accuracy of model is very low: high MAPE

TIME SERIES (TSLA)

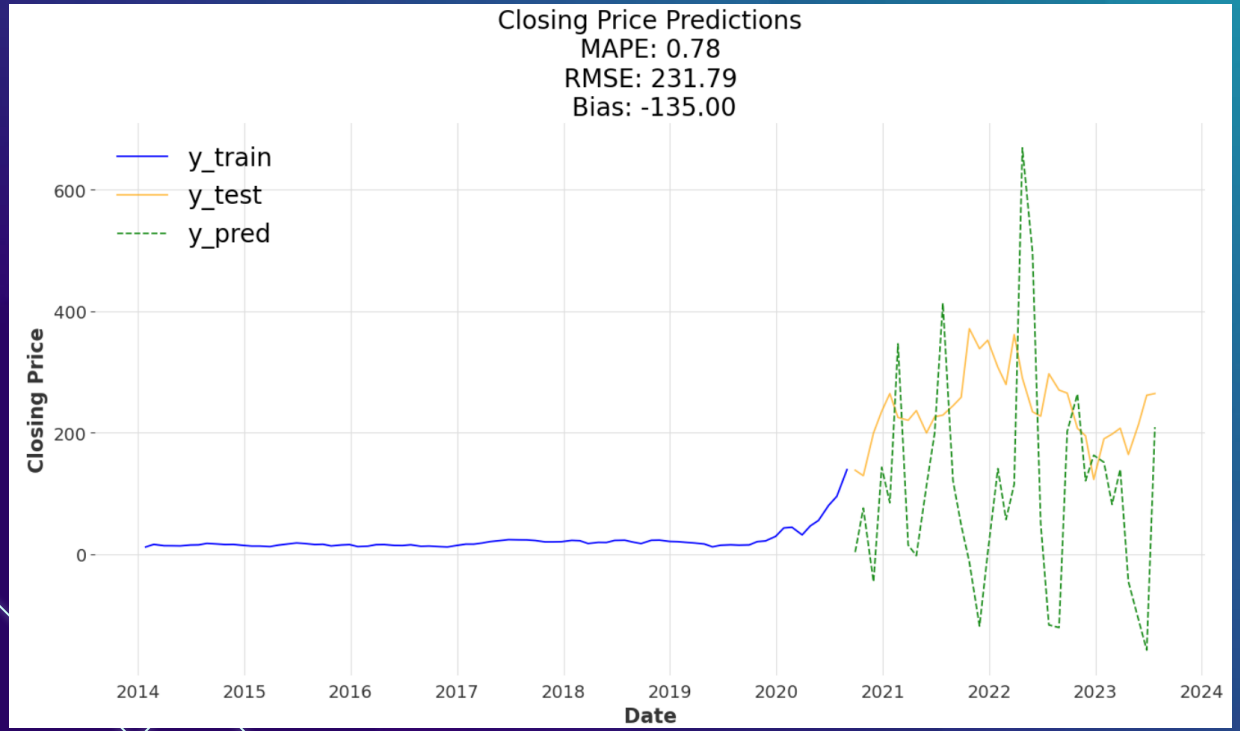
ARIMA (p=1, d=3, q=3)



- Fit y_train (Close Price)
- Optimization based on AIC gives p=1, d=3, q=3
- Prediction shows an ever exponential increasing close price (not possible)
- Results is worse than the Base Model

TIME SERIES (TSLA)

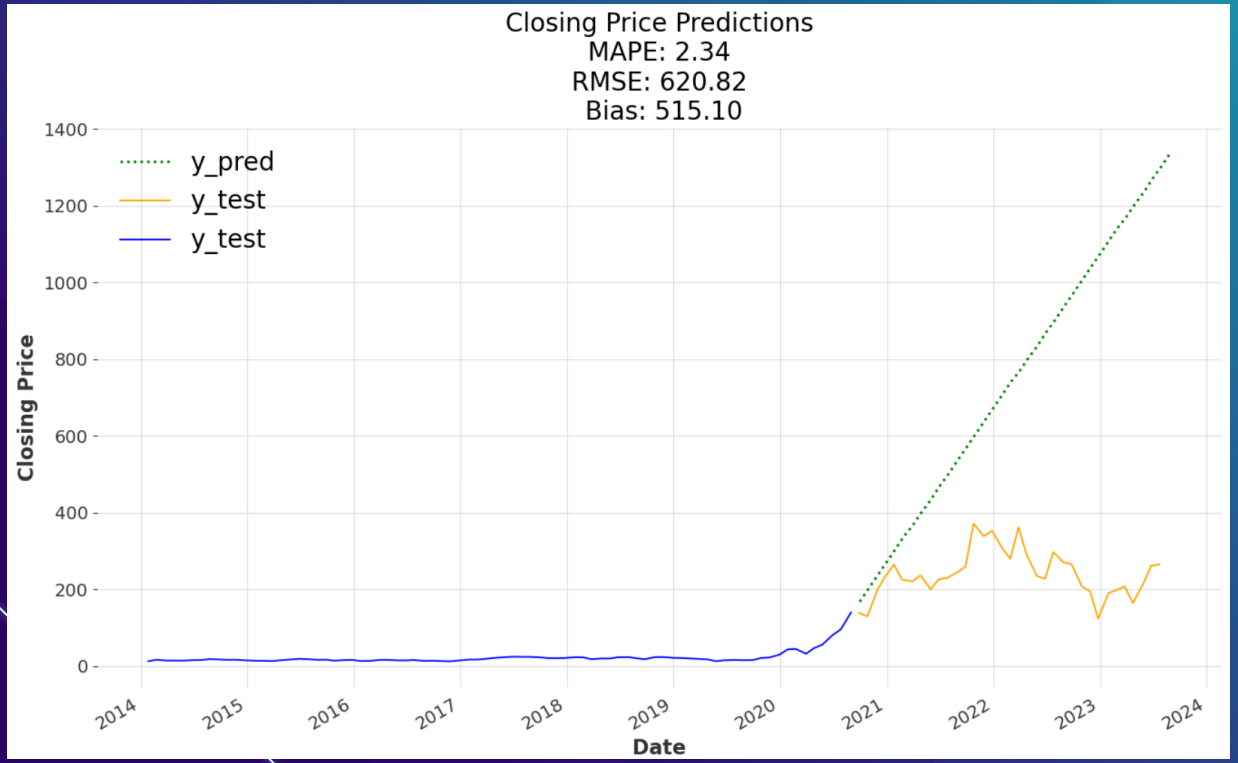
ARIMA WITH EXOGENOUS FEATURES (p=1, d=0, q=0)



- Fit y_train (Close Price) and X_train (volume, OHLC T-6 to T-12)
- Optimization based on AIC gives p=1, d=0, q=0
- Prediction is better with exogenous features but some predictions can be <$0 (not possible)
- Results is close to the Base Model

TIME SERIES (TSLA)

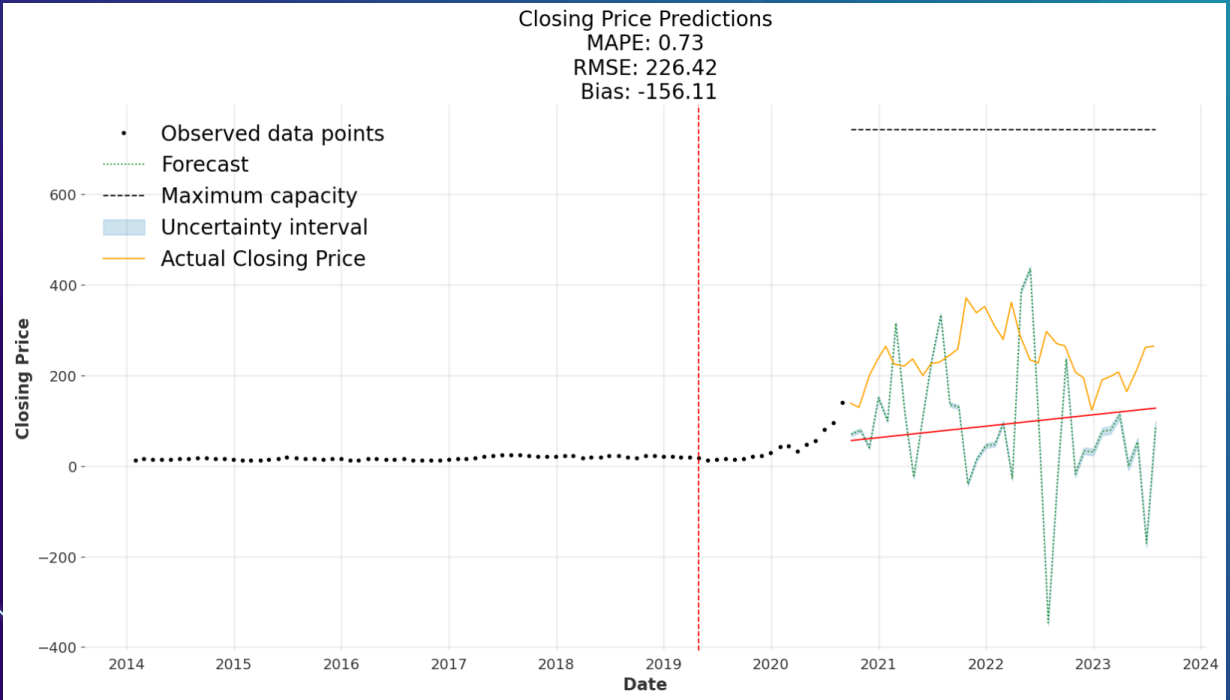
SIMPLE EXPONENTIAL SMOOTHING



- Fit y_train (Close Price)
- Prediction shows an ever linear increasing close price (not possible)

TIME SERIES (TSLA)

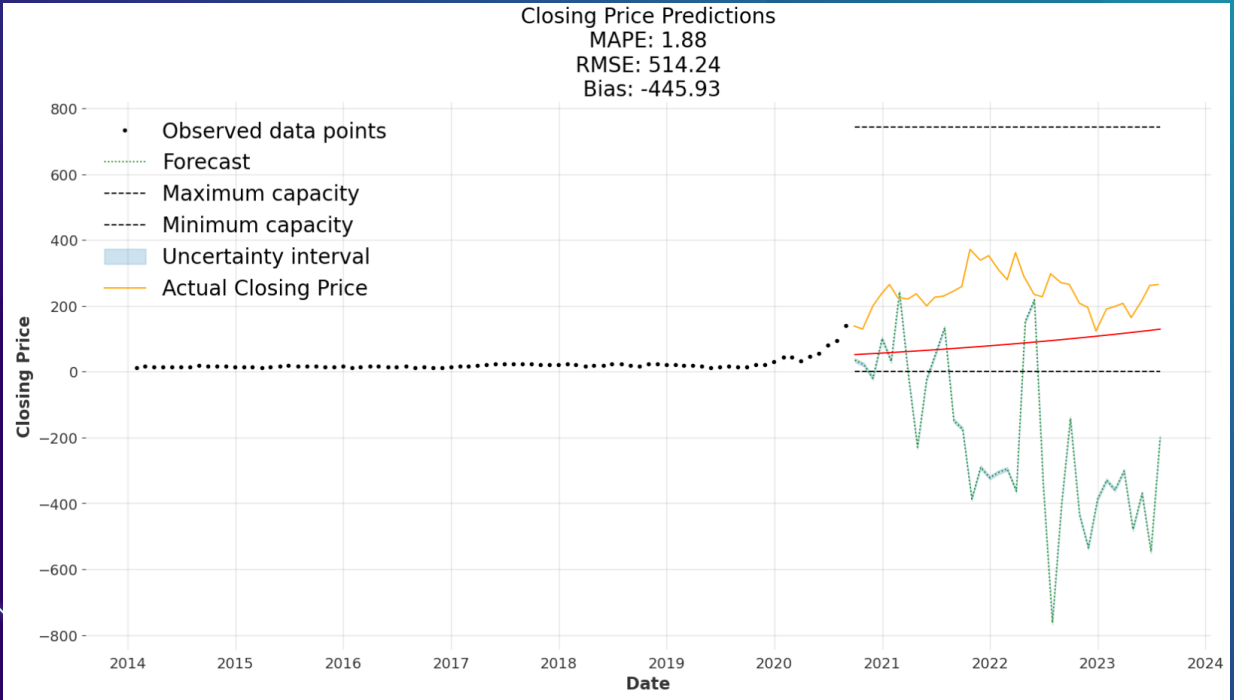
PROPHET LINEAR MODEL



- Fit y_train (Close Price) and X_train (OHLC T-6 to T-12)
- Predictions can be <$0 (not possible)
- Results is close to the Base Model

TIME SERIES (TSLA)

PROPHET LOGISTIC MODEL



- Fit y_train (Close Price) and X_train (OHLC T-6 to T-12)
- Most predictions are <$0 (not possible)
- Results is worse than Prophet Linear Model

TIME SERIES (TSLA)

VARIMA (p=1, d=1, num_samples=1)

- Fit y_train (Close Price) and X_train (OHLC T-6 to T-12)
- Best model so far

MODEL SUMMARY (TSLA)

Model No.	Model Tried	MAPE	RMSE	Bias	Remarks
1	Linear Regression	1.82	495.81	-429.38	Predictions of close price can go below $0. Model is not practical
2	ARIMA	5.83	1657.03	1285.04	p=1, d=3, q=3
3	ARIMAX (with exogenous features)	0.75	224.51	-120.9	p=1, d=0, q=0
4	SARIMAX (with exogenous features & seasonality)	0.75	224.51	-120.9	p=1, d=0, q=0
5	Simple Exponential Smoothing	2.34	620.82	515.1	Almost a linear prediction, not practical
6	Prophet Linear Model	0.73	226.42	-156.11	Predictions of close price can go below $0. Model is not practical
7	Prophet Logistic Model	1.88	514.24	-445.93	Predictions of close price can go below $0. Model is not practical
8	VARIMA (Vector Autoregressive Integrated Moving Average)	0.22	62.71	2.92	p=1, d=1, q=0, num_samples=1
9	VARIMA (Vector Autoregressive Integrated Moving Average)	0.32	109.66	-78.03	p=3, d=1, q=3, num_samples=3
10	VARIMA (Vector Autoregressive Integrated Moving Average)	0.23	66.47	-19.89	p=1, d=1, q=3, num_samples=10
11	VARIMA (Vector Autoregressive Integrated Moving Average)	0.63	175.60	121.19	p=1, d=1, q=2, num_samples=3

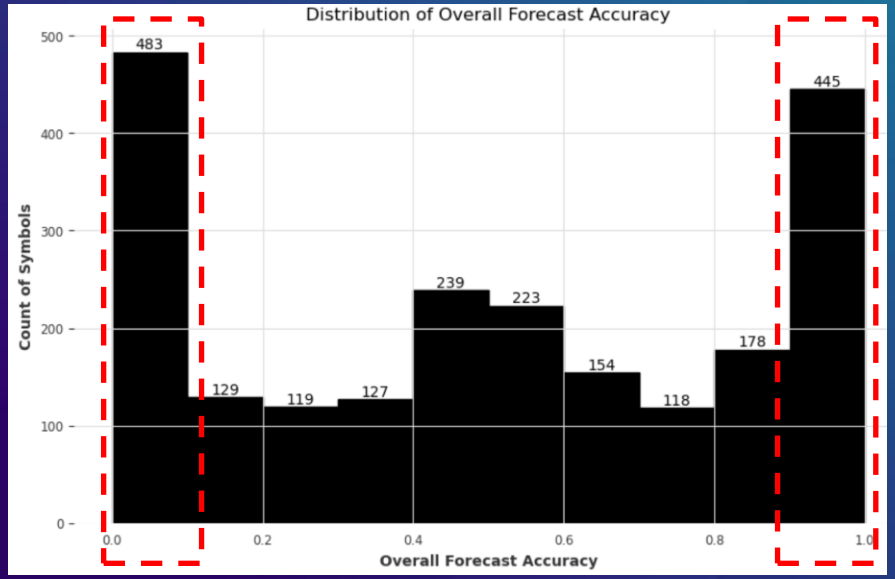
Best Model: VARIMA (p=1, d=1, num_samples=1)

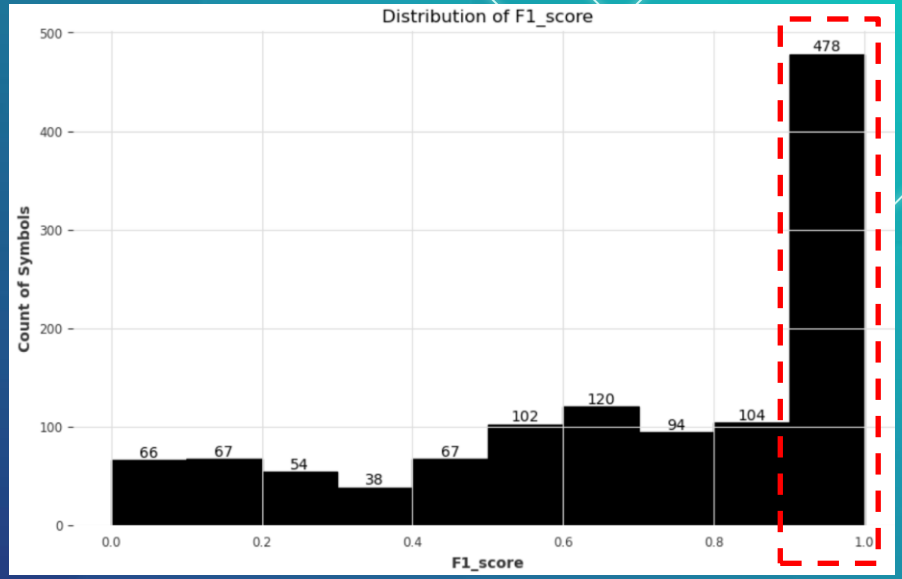
STOCK PURCHASE RECOMMENDATION

METHODOLOGY

- Run VARIMA (p=1, d=1, num_samples=1) for all stocks
- Compare predicted monthly close price from 2020-09-28 to 2023-03-27 against close price as of 2020-08-31
- For each month, if predicted price is greater, recommend “buy”. Else, recommend otherwise.

RECOMMENDATION PERFORMANCE



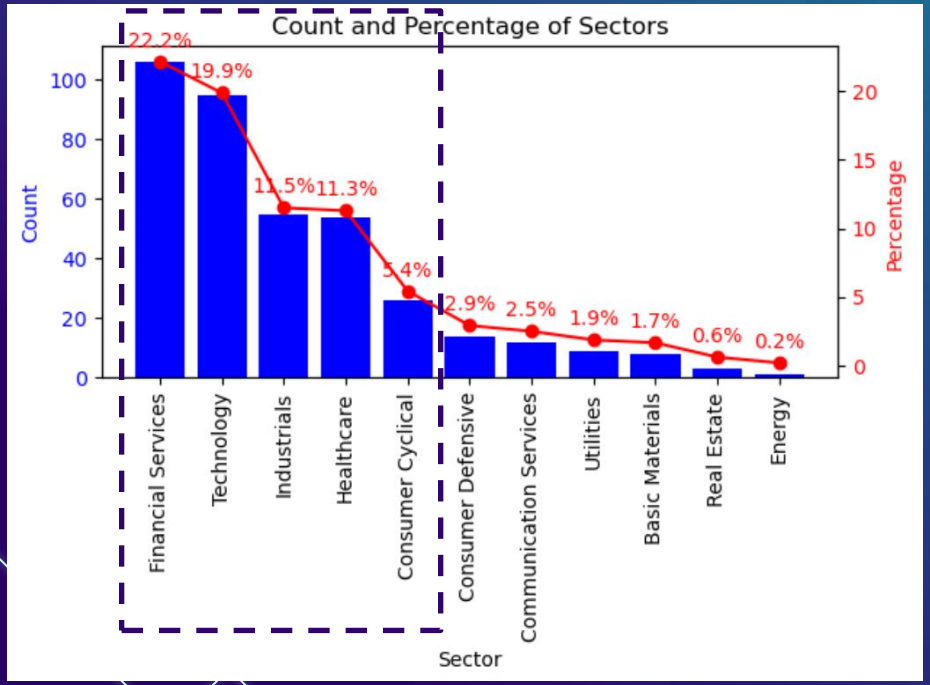


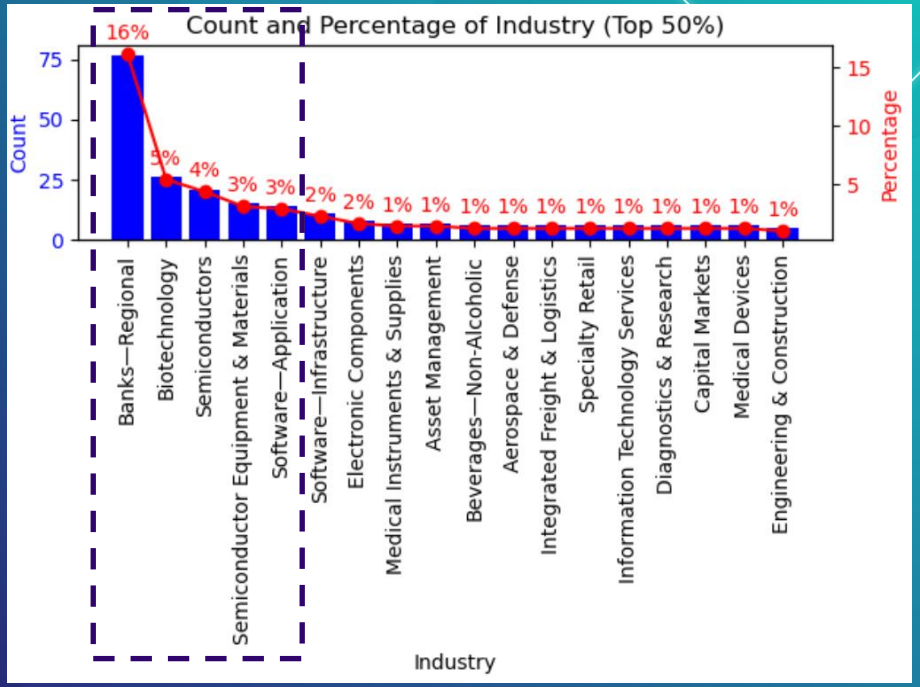
Forecast accuracy distribution has peaks at 2 extreme ends

F1-score distribution is skewed towards left. 478 of the symbols have very good F1_score

RECOMMENDATION

Identified 478 stocks with good F1_score (>= 0.9) to recommend “buy”/”not buy” to investors





05

CONCLUSION

MODEL SUMMARY

RUN VARIMA MODEL

With approximately past 6 years of OHLC

COMPARE FORECAST

Against current stock price

RECOMMEND "BUY"

If prediction > current stock price

LIMITATION

- limited parameters used for training model
- unable to run model on stock with missing information
- assume VARIMA is the best model for all stocks

NEXT STEP

- Ingest more parameters such as 'sector', 'industry’)
- Develop more robust model to account for missing information
- Understand the various finance measurements to consider for modelling.

THANK YOU