PREDICTING TESLA STOCK PRICE MOVEMENTS USING SELECTED MACHINE LEARNING TECHNIQUES, WITH A FOCUS ON INVESTORS SENTIMENT EXPRESSED IN COMMENTS SECTION OF YAHOO FINANCE

Lukasz Macias

Bern 08.11.2024

AGENDA

1) Why Tesla Stock
2) Data collection and sentiment scoring
3) Exploratory data analysis
4) Stock price prediction using:
    a) Logistic Regression
    b) Random Forest
    c) Long Short-Term Memory Model
5) From Prediction to Profit: Trading Simulation

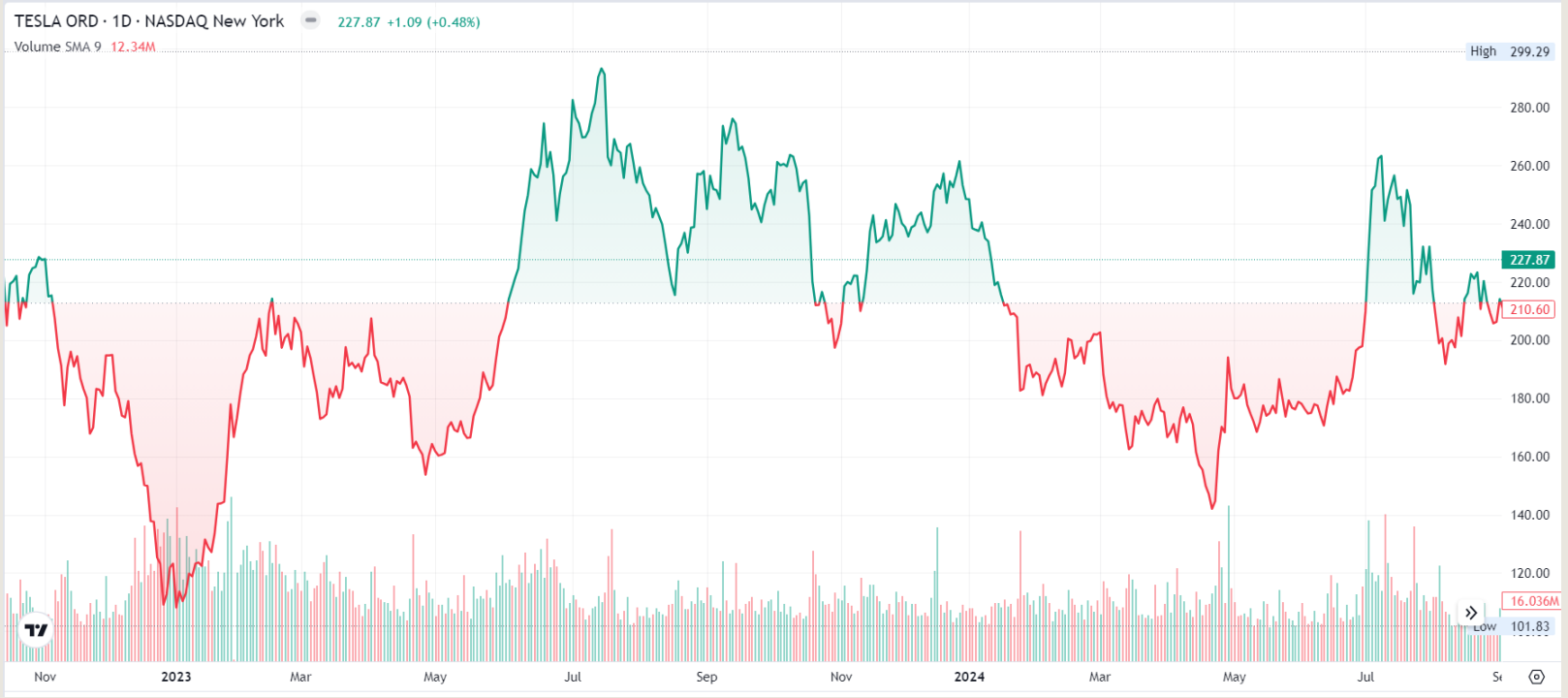
WHY TESLA?

The Magnificent 7 companies are among the seven dominant tech giants driving significant stock market growth through their innovation and influence, making them highly attractive to investors.

https://alaricsecurities.com/magnificent-7-stocks-from-nifty-fifty-to-market-dominance/

WHY WE PICKED TESLA?

- Characterized by frequent "bull runs" and sharp price drops
- Ideal for sentiment analysis due to its dynamic fluctuations which generate sentiment swings among the investors



DATA COLLECTION AND SENTIMENT SCORING

SENTIMENT DATA COLLECTION

- Web scrapped **648k** comments from Yahoo Finance: https://finance.yahoo.com/quote/TSLA/community/

yahoo!finance

THE #1 FINANCE SITE IN THE U.S.[1]

62% of adult Americans own stocks, reaching a 20-year high

Source: https://www.visualcapitalist.com/american-stock-ownership-by-share-of-financial-assets/

#1 platform for business news[1]

With 93M monthly unique visitors in the U.S.[1]

[1] Comscore Media Metrix Multi-Platform, Business/Finance News category, July 2022, U.S.

#1 video platform in finance category

With 6.2M monthly unique viewers.

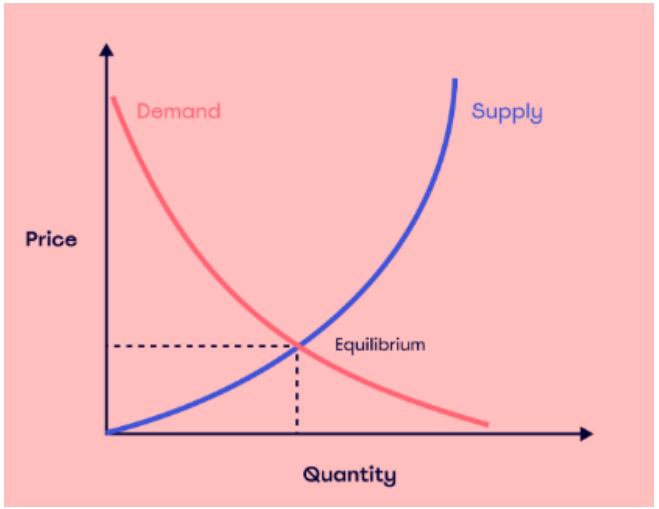
[2] Comscore Video Metrix, Desktop Content Only, July 2022, U.S.

Source: https://www.advertising.yahooinc.com/about/our-brands/yahoo-finance

SENTIMENT ASSUMPTIONS

Comment makers represent actual or potential TESLA stock owners and express the general sentiment of the market regarding the asset

- **Positive sentiment** drives **demand** for TSLA shares, as market participants anticipate price increases based on optimistic commentary. This increased demand **pushes the stock price higher.**
- **Negative sentiment** instills fear among market participants, creating the perception that the stock price will fall. This encourages shareholders to sell, increasing **supply** and **driving the price down.**



BULL AND BEAR MARKET

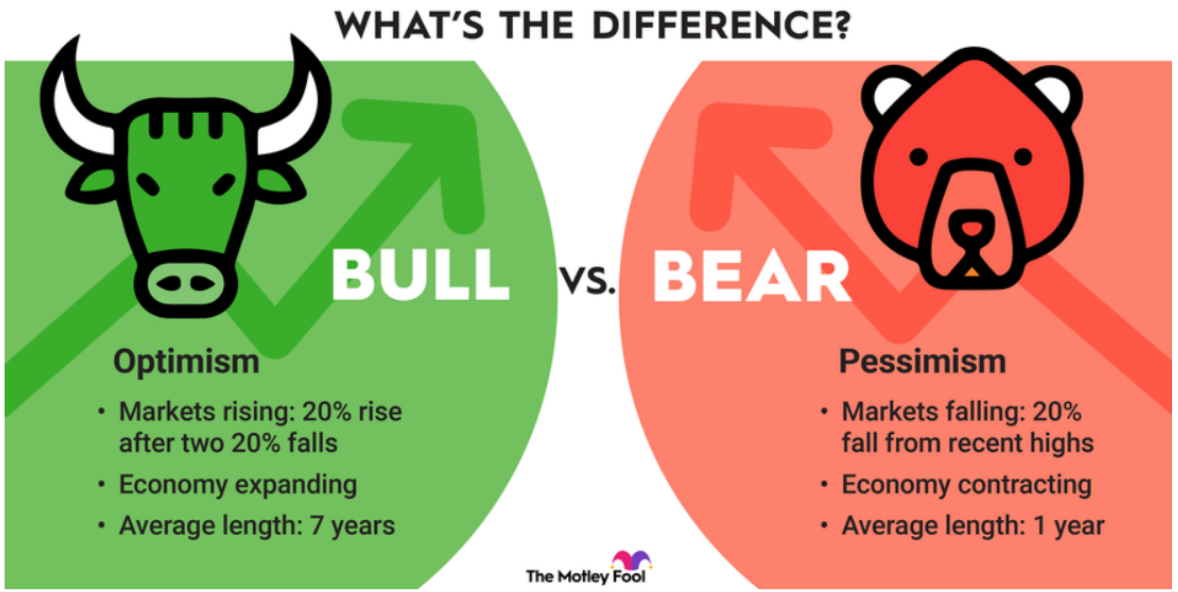
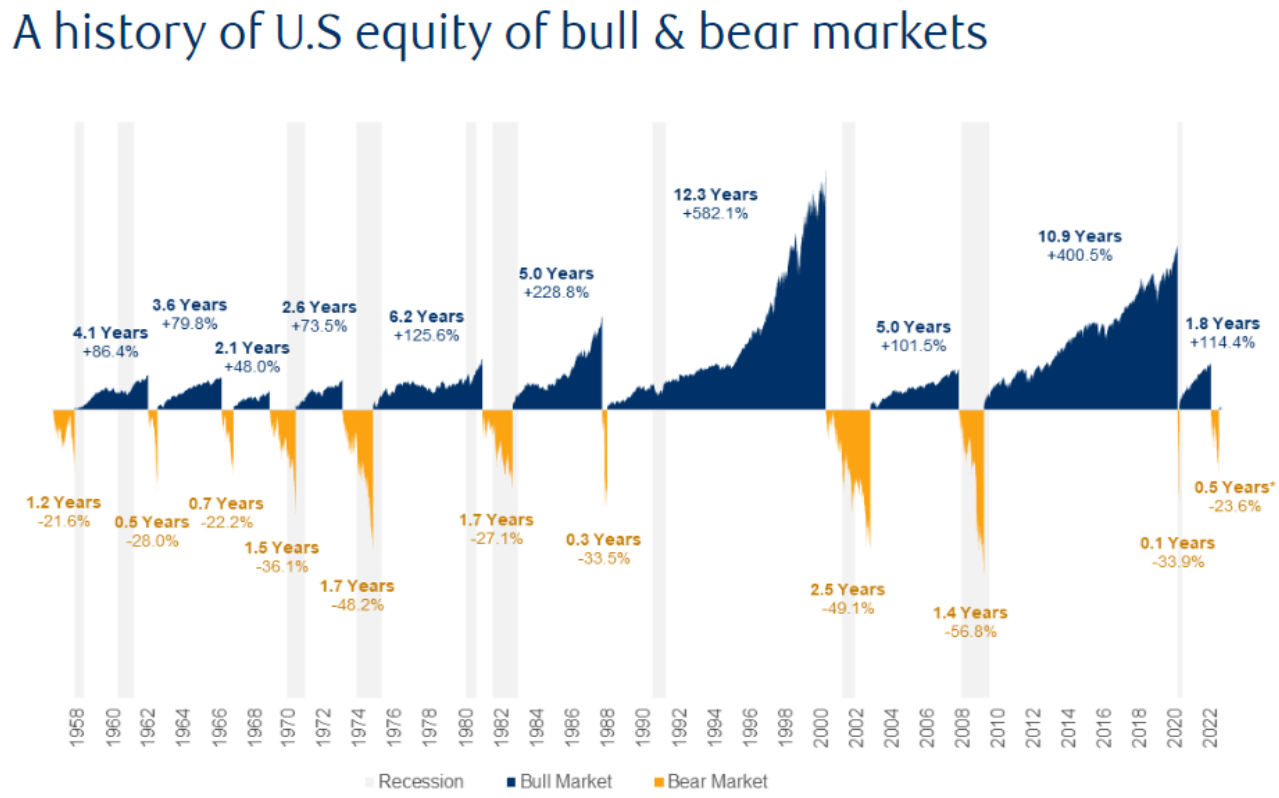


Image source: The Motley Fool



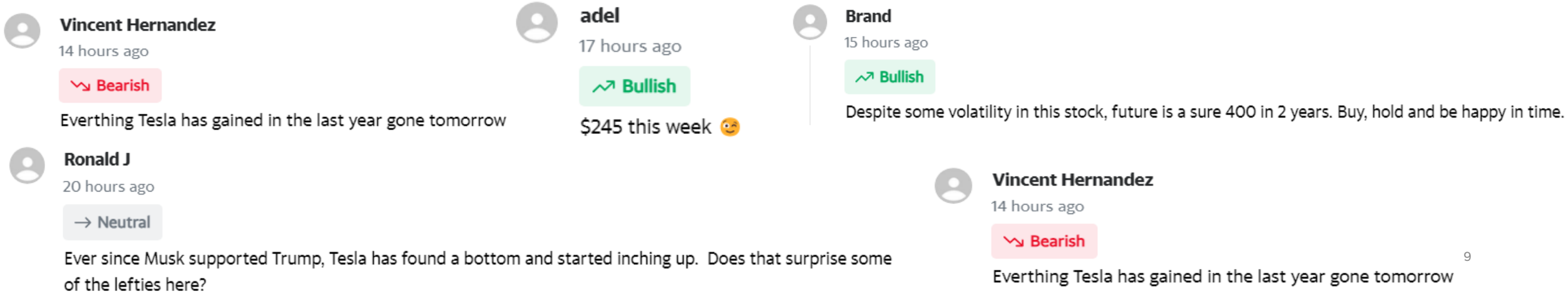
Source: https://www.rbcgam.com/en/ca/learn-plan/investment-basics/the-bulls-the-bears/detail

SENTIMENT SCORING

Each comment receives a sentiment score ranging from –1 (highly negative) to 1 (highly positive), based on a two-tier approach:

- **44.4% of comments** have a sentiment tag where users **explicitly** express their outlook on Tesla's future stock price, the score is assigned as follows:
  - **BULLISH** – "price will raise" - sentiment Score: **1**
  - **NEUTRAL** – "no change in price" - sentiment Score: **0**
  - **BEARISH** – "price will fall" - sentiment Score: **–1**

	# of comments	% of comments
BEARISH	126'007	19.4%
BULLISH	147'633	22.8%
NEUTRAL	14'370	2.2%
null	360'077	55.6%
TOTAL	**648'087**	**100.0%**



SENTIMENT SCORING

- For the remaining **55.6% of comments** without a sentiment tag, we assign a sentiment score using the **VADER** (Valence Aware Dictionary and sEntiment Reasoner) lexicon.
- **VADER** is a pre-trained sentiment analysis tool specifically designed for social media text.
- It uses a lexicon of words with pre-labeled sentiment scores and adjusts based on factors like punctuation, capitalization, and negation.
- The sentiment score range:
  - **-1**: Extremely Negative
  - **0**: Neutral
  - **1**: Extremely Positive

Mark

23 hours ago

It is difficult to believe that people are buying or even trading this stock anymore.

```
{'neg': 0.152, 'neu': 0.848, 'pos': 0.0, 'compound': -0.3612}
```

Henry Schneider

6 hours ago

Kamala harris knows what is AI and can support tech stocks

```
{'neg': 0.0, 'neu': 0.787, 'pos': 0.213, 'compound': 0.4019}
```

AlexD

1 day ago

Tesla holding strong today.

```
{'neg': 0.0, 'neu': 0.476, 'pos': 0.524, 'compound': 0.5106}
```

jc

22 hours ago

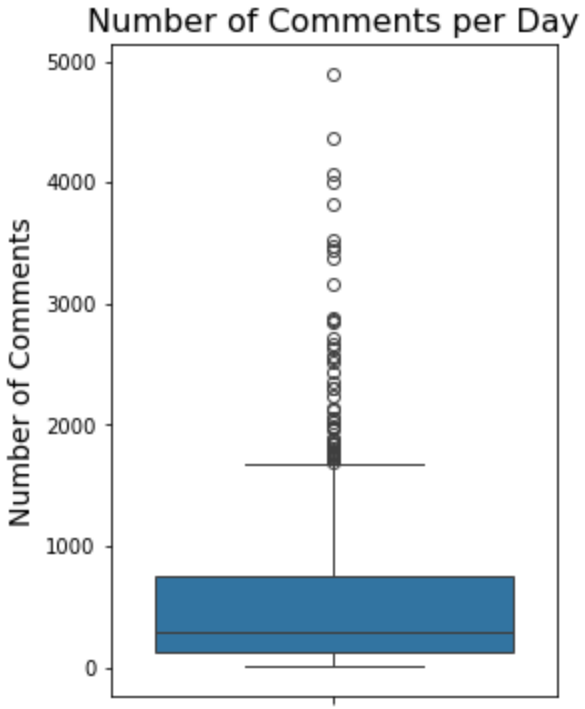
Cybertrucks are so ugly.

```
{'neg': 0.586, 'neu': 0.414, 'pos': 0.0, 'compound': -0.6418}
```

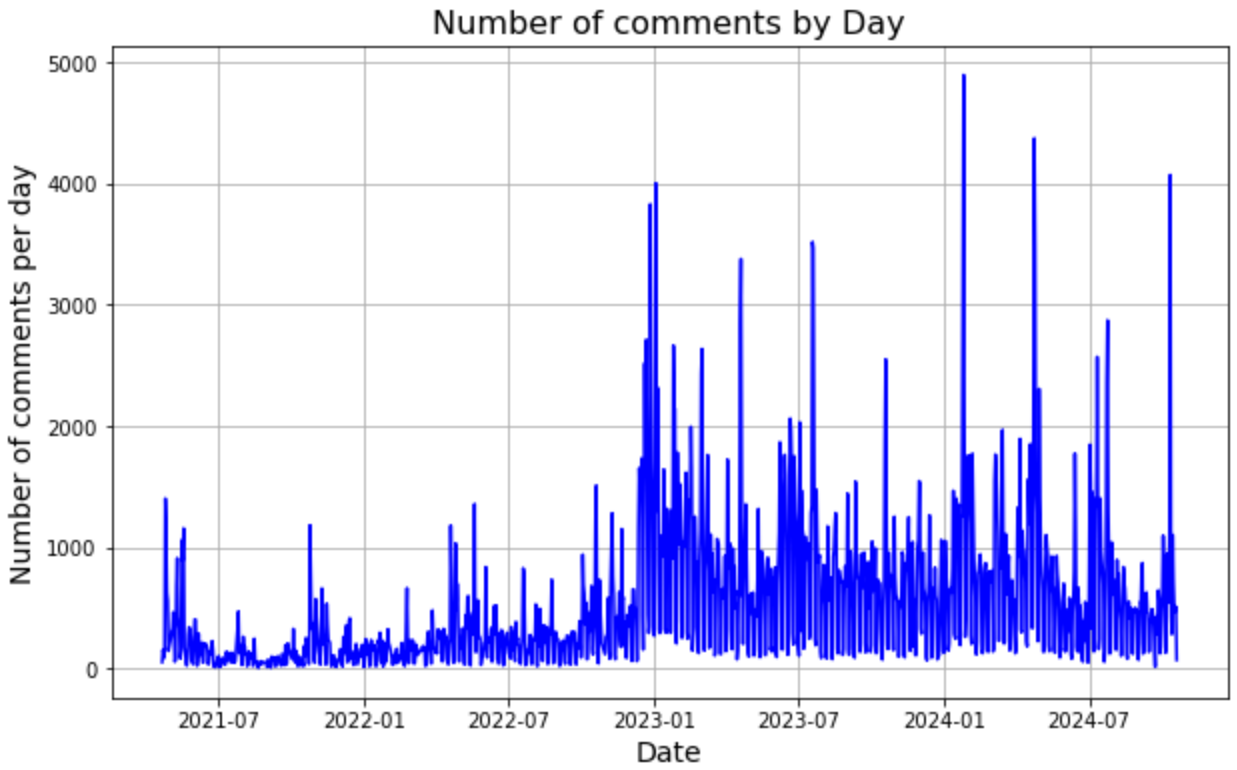
EXPLORATORY DATA ANALYSIS

COMMENTS DATA COLLECTION

- Analysis period: 01.05.2021 - 30.09.2024, 3 years and 5 months
- Web scrapped **648k** comments from Yahoo Finance: https://finance.yahoo.com/quote/TSLA/community/



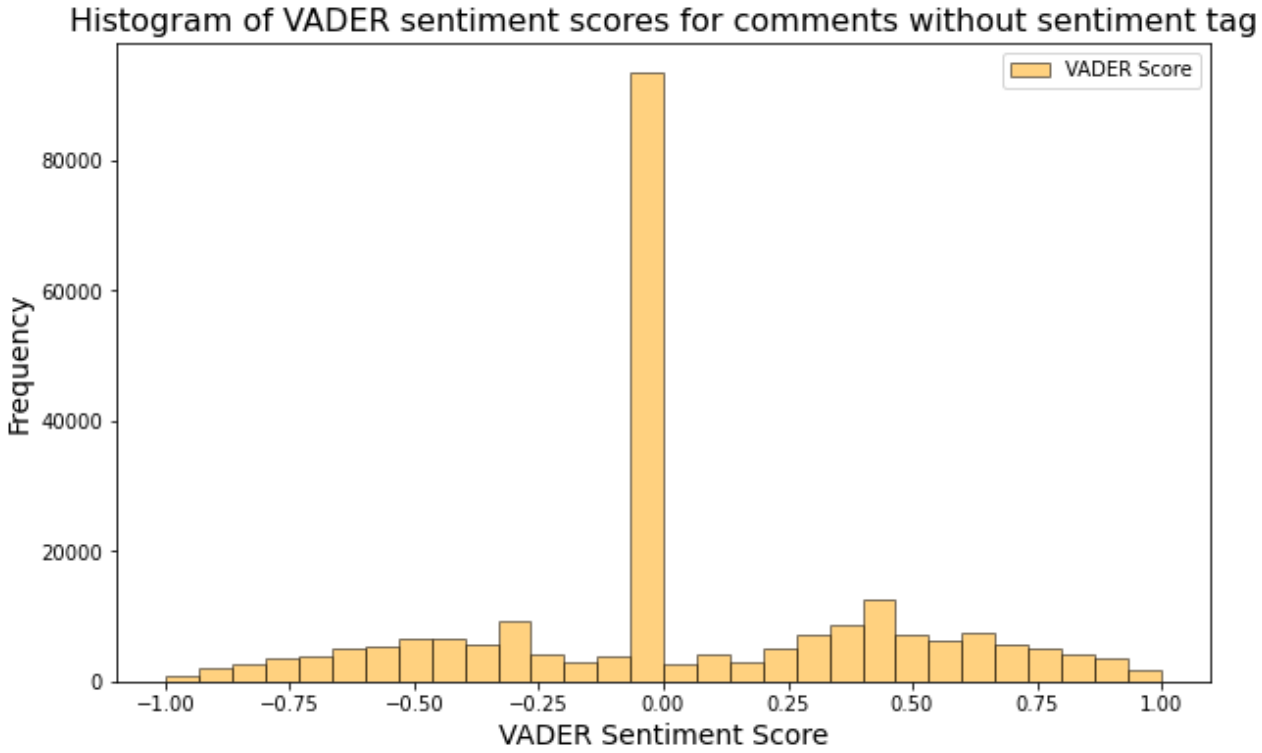
	comment_count
count	1240.000000
mean	522.650806
std	583.815412
min	1.000000
25%	128.000000
50%	293.500000
75%	749.500000
max	4900.000000

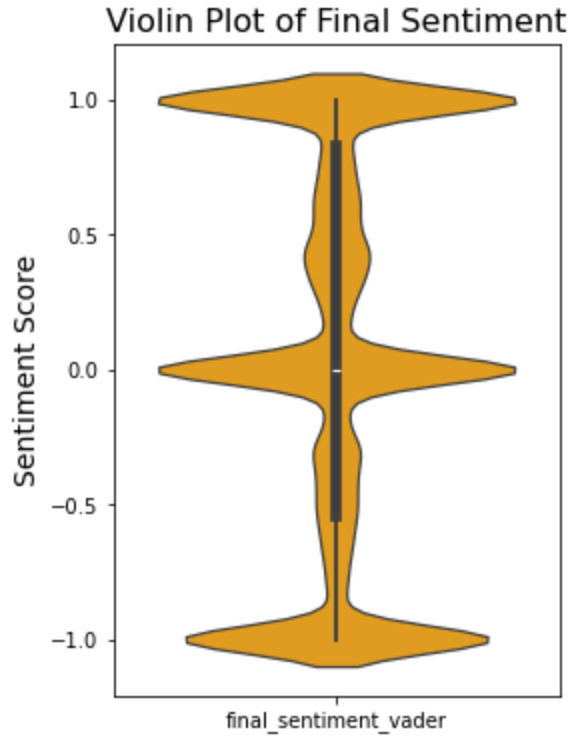


SENTIMENT SCORES

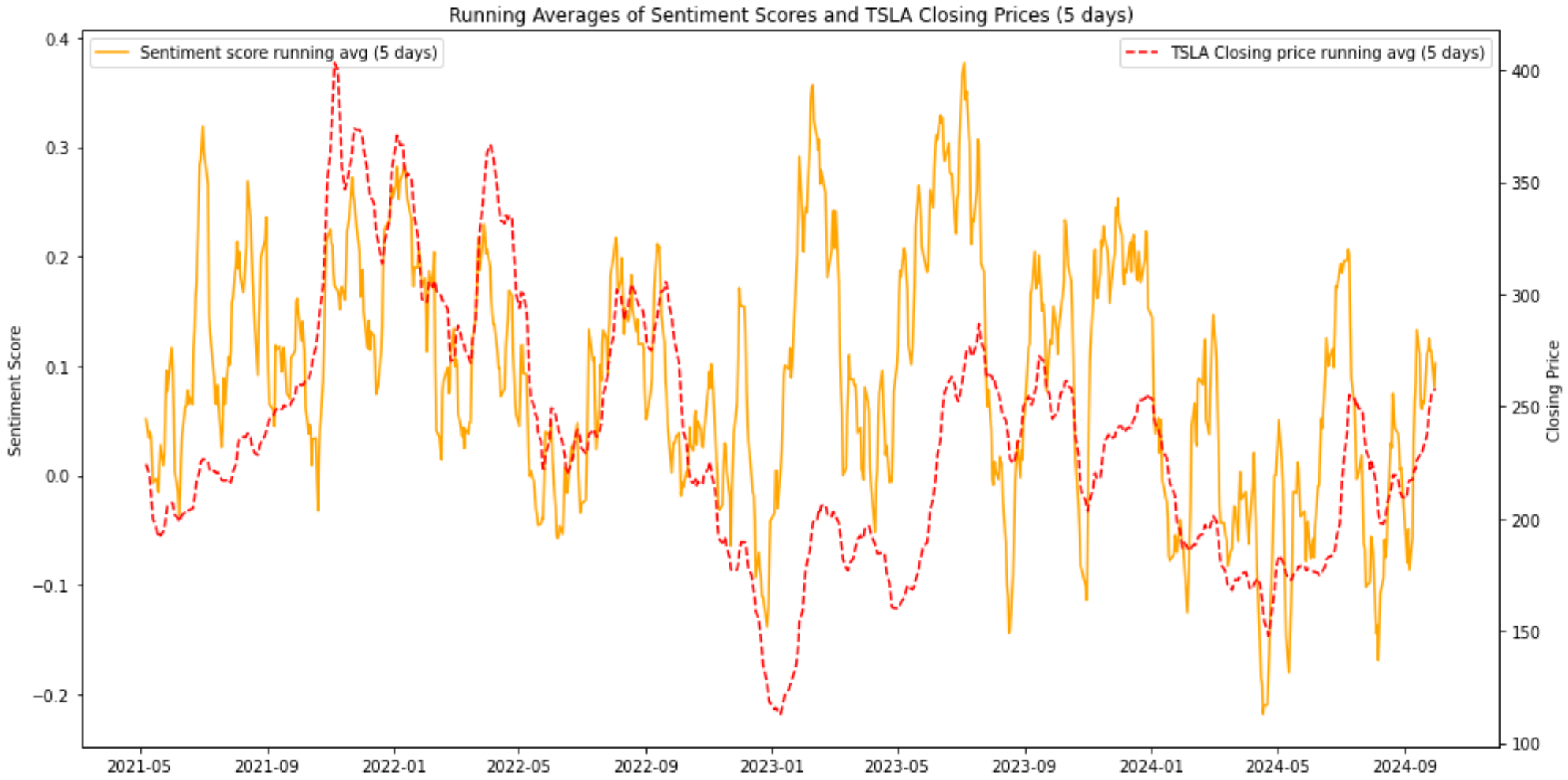
- For **44.4%** of comments with BULLISH/BEARISH/NEUTRAL sentiment tag we assign a value: –1, 0 or 1 based on the tag
- For the outstanding **55.6%** of comments without the tag we derive the sentiment score based on the VADER sentiment

	# of comments	% of comments
BEARISH	126'007	19.4%
BULLISH	147'633	22.8%
NEUTRAL	14'370	2.2%
null	360'077	55.6%
TOTAL	**648'087**	**100.0%**





SENTIMENT VS TSLA CLOSING PRICE



TSLA STOCK PRICE PREDICTION – LOGISTIC REGRESSION

LOGISTIC REGRESSION

- Logistic regression, despite its name, is a linear model for classification rather than regression
- ML technique used to predict the binary outcomes that are categorical and mutually exclusive.
- In this model, one class is assigned a target value of 1 (positive class), and the other class is assigned a target value of 0 (negative class).
- To predict the probability p of a point belonging to the positive class, logistic regression applies a *sigmoid function* to a linear combination of input features. The sigmoid function transforms this linear output into a probability range between 0 and 1, where:

$$p = \sigma(w_0 + w_1 x_1 + ... + w_p x_p)$$

Where $\sigma(x) = \frac{1}{1+e^{-x}}$.

- The model's parameters w are optimized to minimize a specific *loss function*, which, in binary classification, is the *log loss*. This loss function measures the difference between the predicted probabilities and the actual target values, helping adjust w to improve prediction accuracy.

LOGISTIC REGRESSION

Goal: Predict whether Tesla's closing price on the next trading day will be **higher** or **lower** than today's closing price.

Data Preprocessing

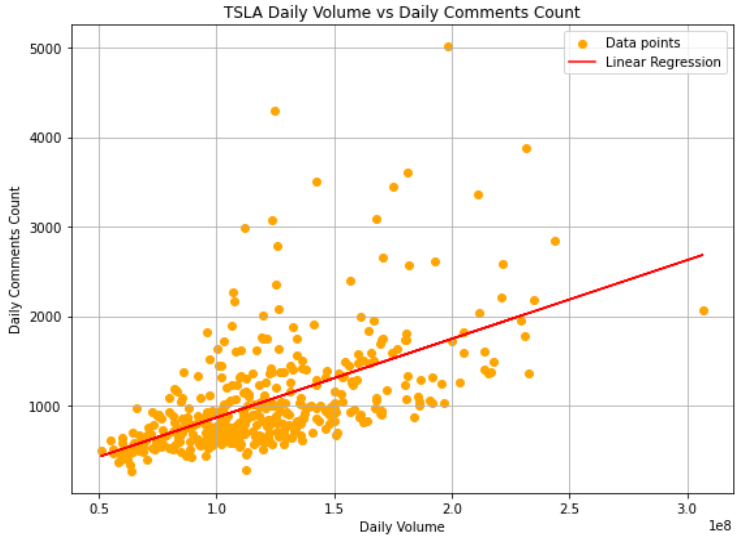
- Aggregate by day, add a column with T+1 closing price, and a target column: 1 if price went up or 0 if down

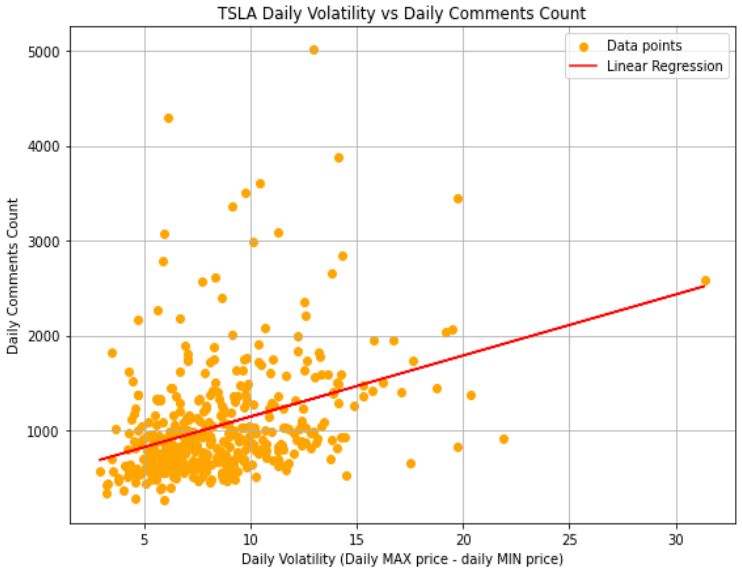
Features selection:

1) 5 days running average of the sentiment score
2) 5 days running average of the TSLA closing price
3) 5 days running average of TSLA daily volume
4) 5 days running average of TSLA daily volatility
5) Closing price today

Model Training

- Train/test split = 20%
- Random state = 1





LOGISTIC REGRESSION - EVALUATION

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

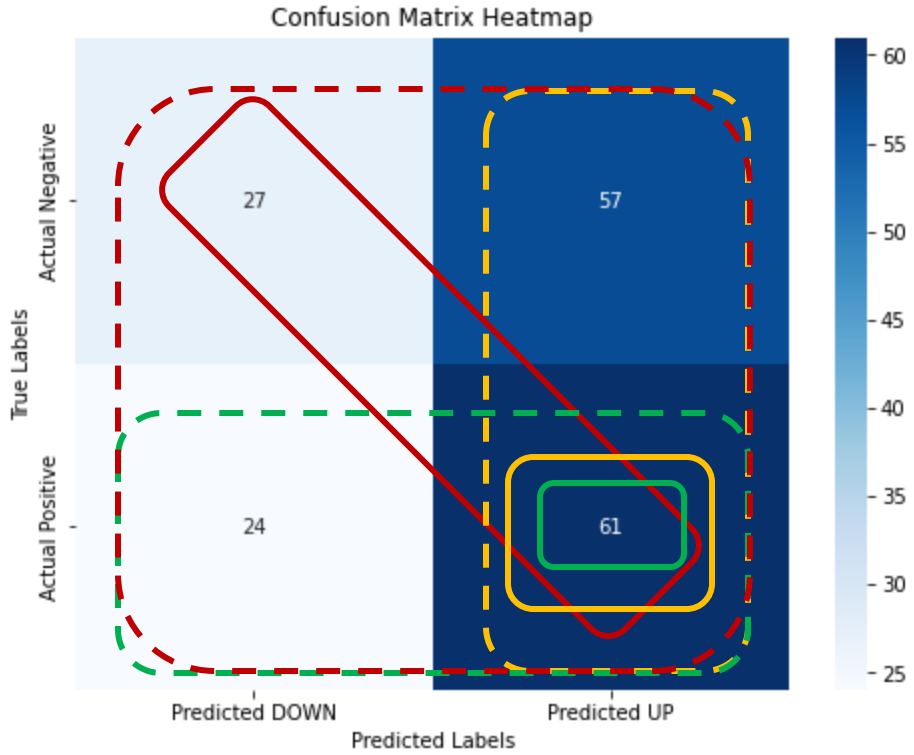
52,1%

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

51,6%

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

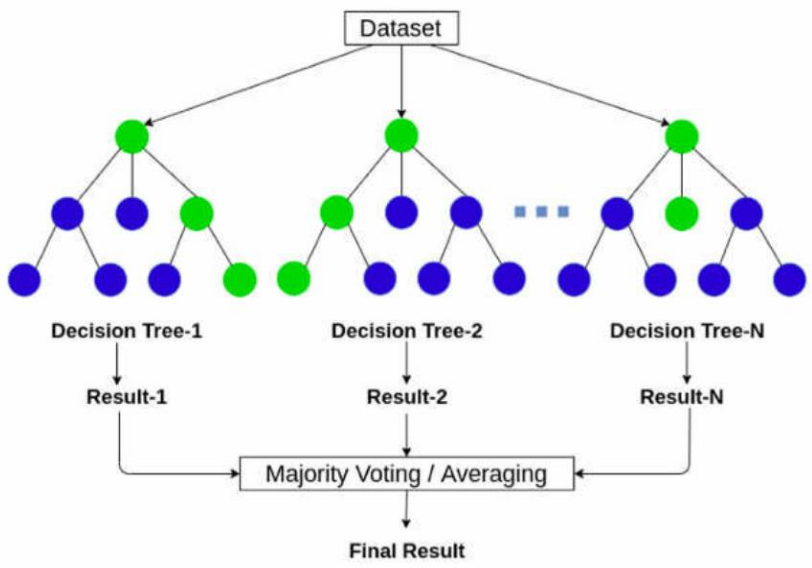
71,6%



TSLA STOCK PRICE PREDICTION – RANDOM FOREST

RANDOM FOREST

- Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
- It can be applied to classification or regression problems.
  .
- Instead of relying on a single decision tree, the Random Forest model combines the output of each tree, using the average of their predictions to produce a final, more reliable output.



Source: https://www.gofar.ai/p/predicting-market-risk-using-machine

RANDOM FOREST

Goal: Predict TSLA Closing price on T+1

Data Preprocessing

- Aggregate by day, add a target column with T+1 closing price

Features selection:

1) 5 days running average of the sentiment score
2) 5 days running average of the TSLA closing price
3) 5 days running average of TSLA daily volume
4) 5 days running average of TSLA daily volatility
5) Closing Price today

Model Training

- Train/test split = 20%
- Random state = 333
- Random Forest Regressor with n_estimators = 100 (# of decission trees)

RANDOM FOREST – EVALUATION

MSE - Mean Squared Error

- Measure accuracy of predictions in regression tasks
- Avg. Of the squared differences between actual and predicted values
- Squaring to emphasize larger errors, hard to interpret because the results are squared as well

MSE: 103

R^2 – R-squared

- a metric used in regression analysis to measure how well a model's predictions match the actual data.
- Specifically, it indicates the proportion of the variance in the dependent variable that is predictable from the independent variables
- **0** means the model explains none of the variance (no predictive power).
- **1** means the model explains all the variance (perfect predictive power).

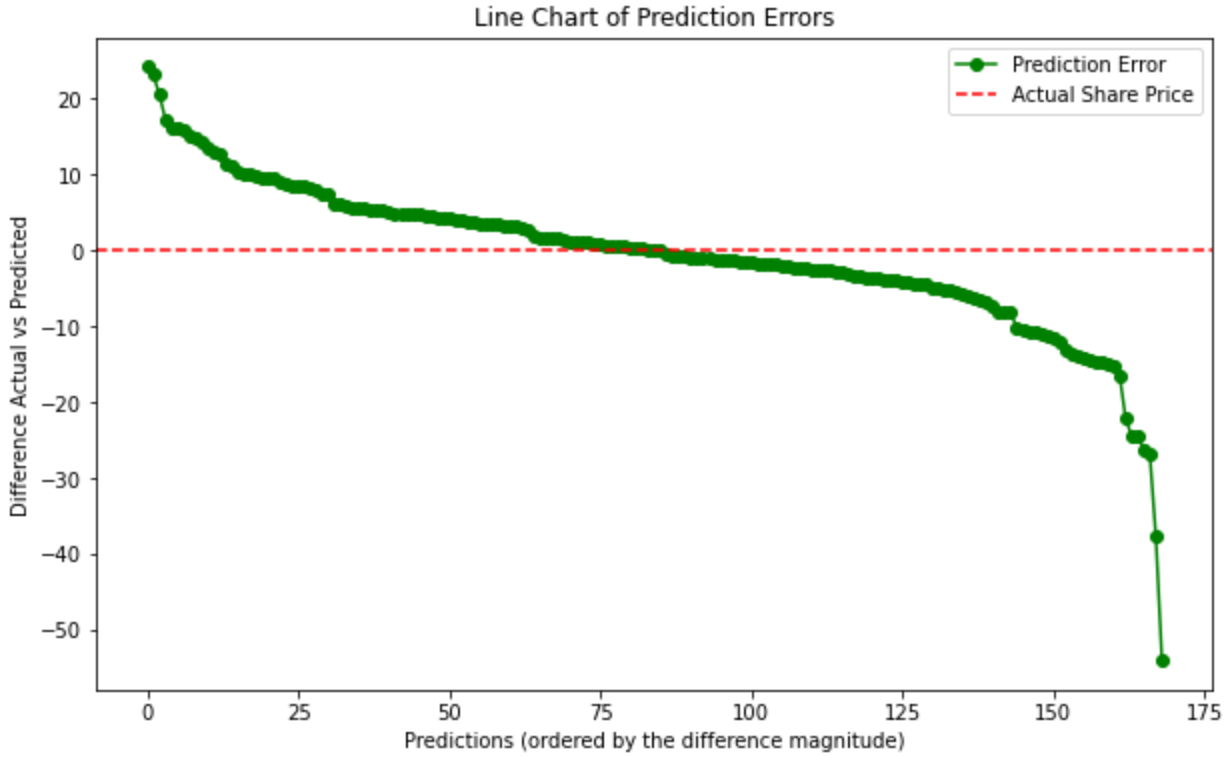
R^2: 0.95

RANDOM FOREST – EVALUATION

MAE - Mean Absolute Error

- a metric used to measure the accuracy of a model's predictions by calculating the average of the absolute differences between predicted and actual values.

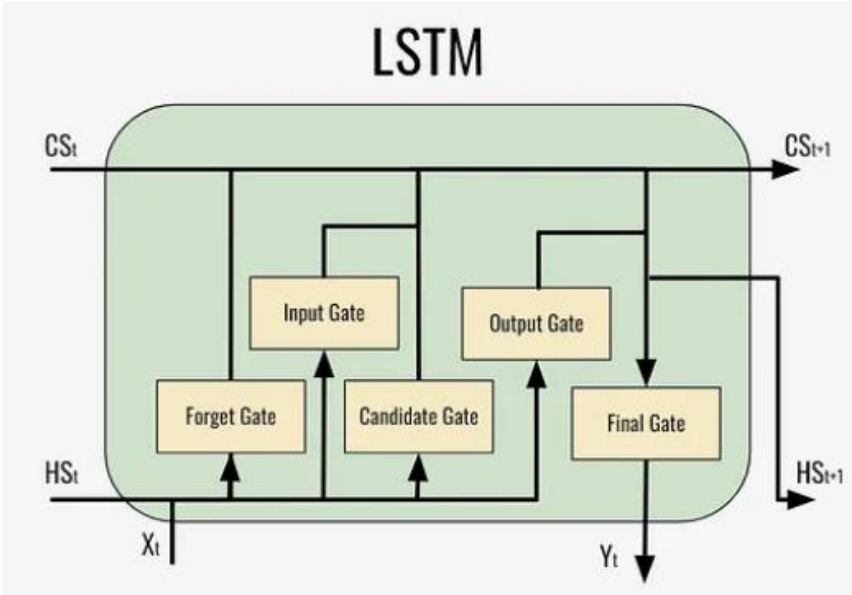
MAE: 7.05



TSLA STOCK PRICE PREDICTION – LONG SHORT-TERM MEMORY (LSTM) MODEL

LONG SHORT-TERM MEMORY MODEL

- **LSTM Overview**: Long Short-Term Memory (LSTM) models are a type of recurrent neural network specifically designed for learning from sequences of data, making them well-suited for time series predictions like TSLA share prices.
- **Memory Mechanism**: LSTMs use memory cells and gate mechanisms to retain relevant information over long periods, enabling them to capture how past investor sentiment influences future price movements.
- **Input Processing**: By processing sentiment scores as input sequences, LSTMs can effectively identify the relationship between investor sentiment and fluctuations in stock prices.



https://medium.com/@CallMeTwitch/building-a-neural-network-zoo-from-scratch-the-long-short-term-memory-network-1cec5cf31b7

LONG SHORT-TERM MEMORY MODEL

Goal: Predict TSLA Closing price on T+1

Data Preprocessing

- Aggregate by day, target is the Closing price on T+1
- Scale all the features to be in range 0 – 1
- Create sequences, for selected number of days (5 days)

Features selection:

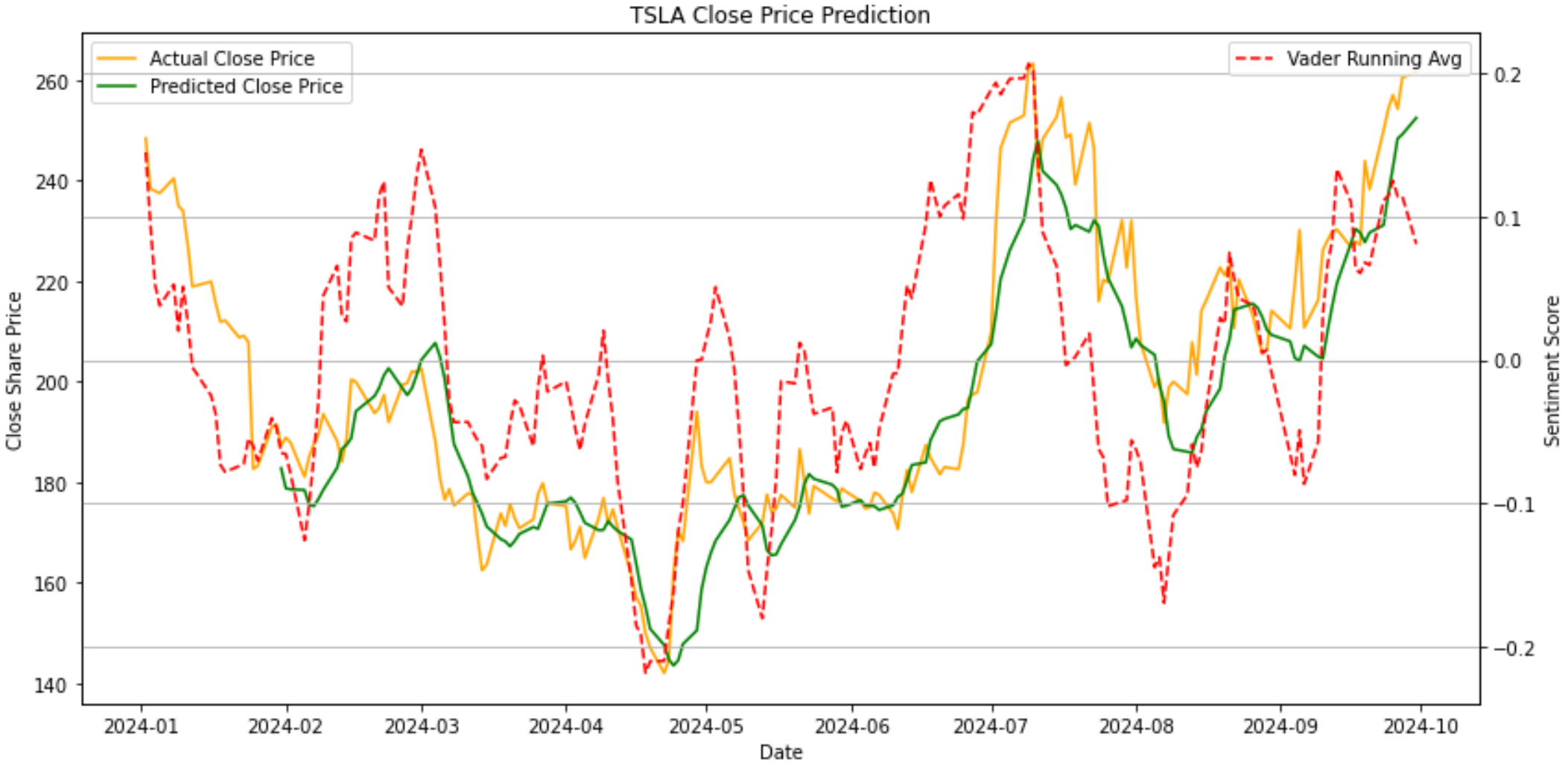
1) 5 days running average of the sentiment score
2) 5 days running average of the TSLA closing price
3) 5 days running average of TSLA daily volume
4) 5 days running average of TSLA daily volatility
5) Closing price

Model Training

- Train/test split = 80%/20%, difference to Random Forest – continous
- Sequential – layers are stacked one after the other
- Initiate a LSTM layer with 20 units (neurons) and ReLU activation.
- The model is compiled with the Adam optimizer and Mean Squared Error (MSE) as the loss function.
- # of epochs = 10 --> 10x throught the entire dataset

LONG SHORT-TERM MEMORY MODEL - RESULTS

MSE: 139 MAE: 9.02 R^2: 0.82



FROM PREDICTION TO PROFIT: TRADING SIMULATION

TRADING STRATEGY

Invested amount: 10 000 USD

Investment period: 31.01.2024 - 30.09.2024

Prediction Model: LSTM

Strategy:

Trade decision every day at market closing

If predicted price on T+1 > Close price T+0 --> buy or hold position

If predicted price on T+1 < Close price T+0 --> sell or keep cash position

Benchmark:

Close price 31.01.2024: 187.29 USD

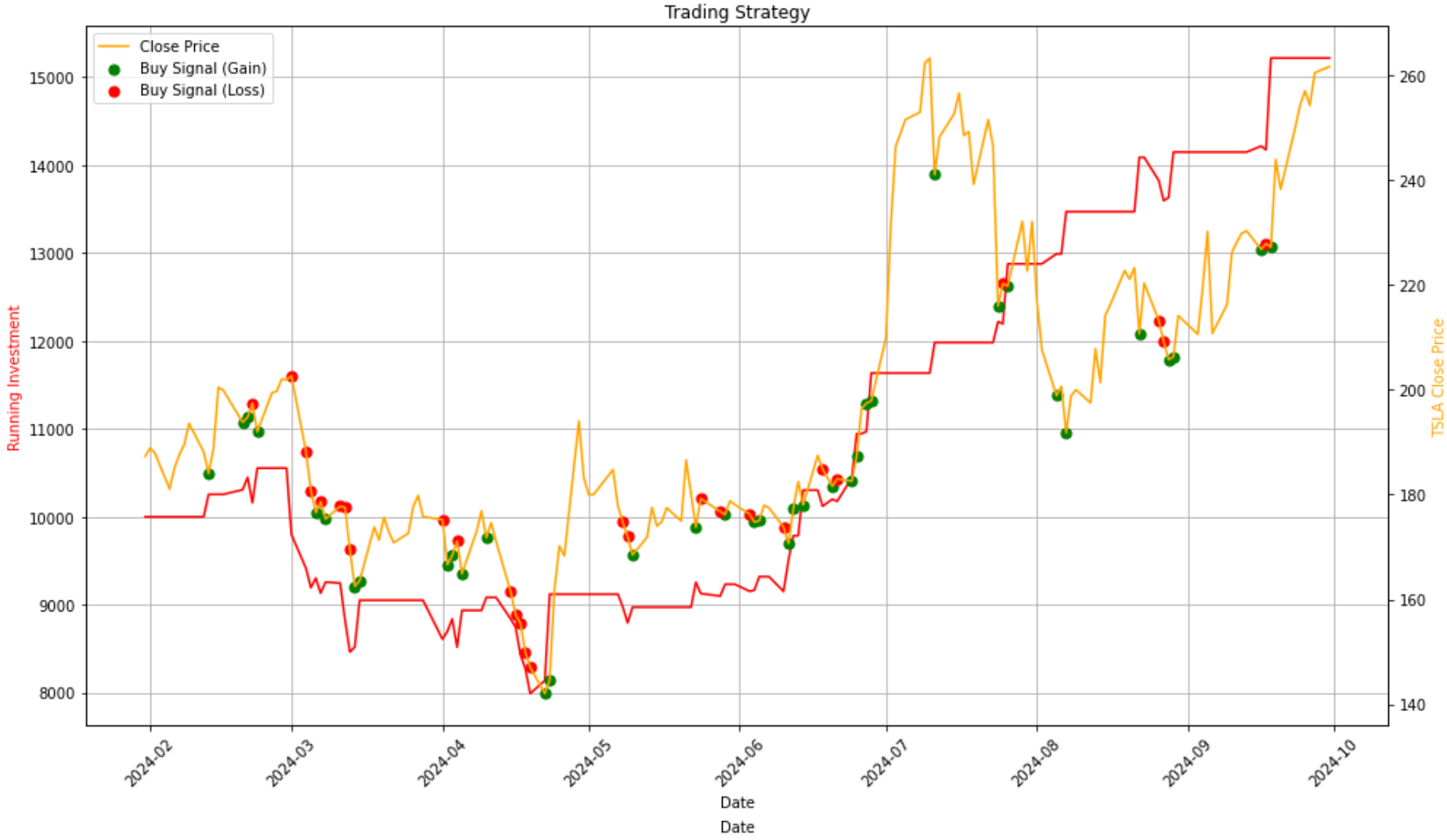
Close price 30.09.2024: 261.69 USD

Performance: +39.72%

ROI: 13 972 USD

TRADING STRATEGY - RESULTS

Trading days: **168** Days with buy signal: **64 (38%),** w/o gain **37** days and loss **27** days

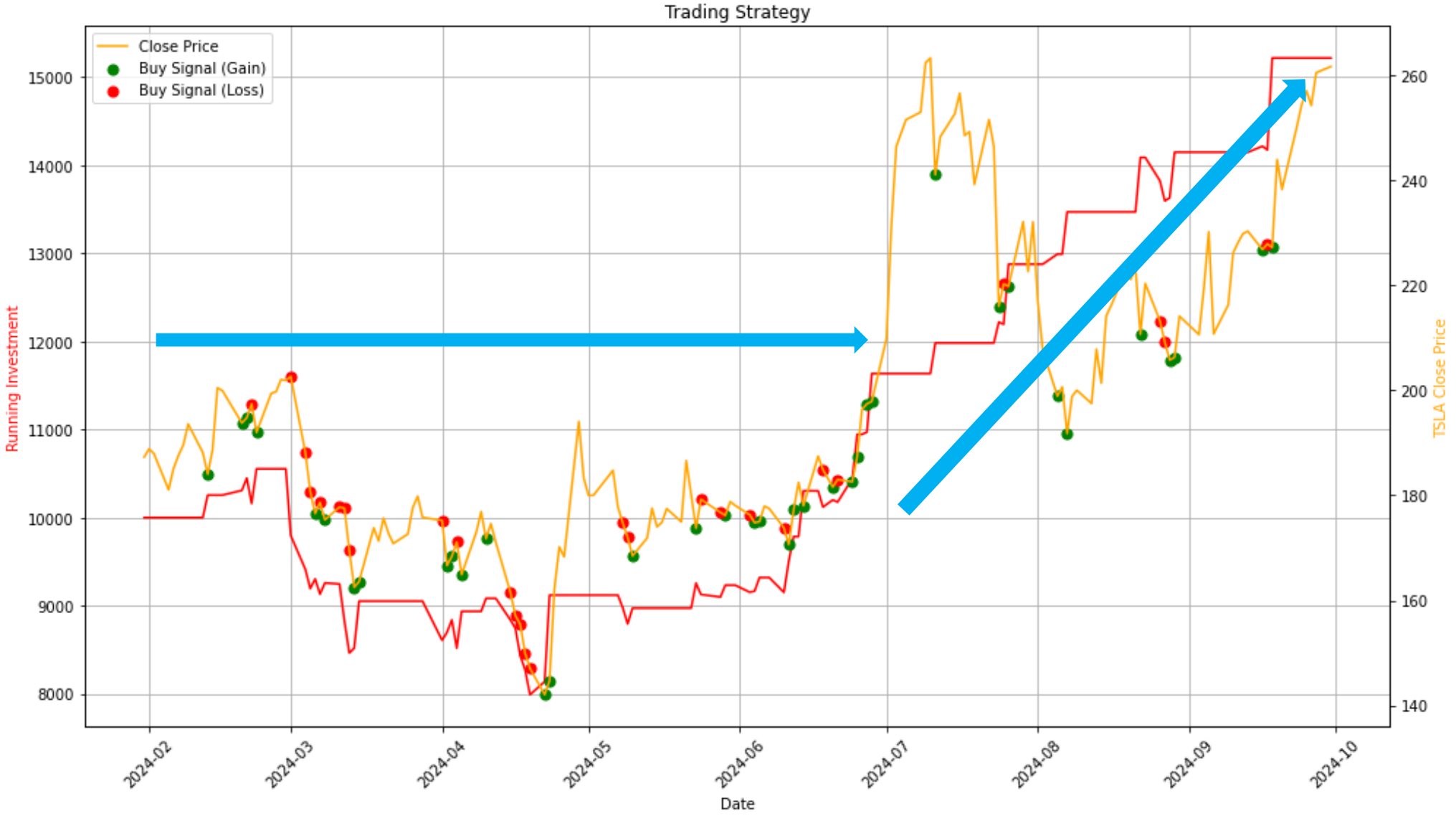


ROI: 15 215 USD

+52,2%

the strategy has exceeded the benchmark by + 1 243 USD

TRADING STRATEGY - RESULTS



REFERENCES

To see the full implementation, you can view the complete code on my GitHub:

https://github.com/lukaszmacias01/CAS_UniBern_Applied_Data_Science/tree/master/MODULE3

THANK YOU