Capstone 3: Stock Prediction Model

By: Mohsin Syed
Aug. 1/2024

Problem Statement and Context

- Predicting stock price can be challenging.
- Myriad of factors influence stock price.
  - Economic indicators.
  - Market trends
  - Company specific events.
- Investors and financial analyst seek advanced methods to gain market insight and predict stock price.
- Can we leverage Machine Learning and Deep Learning algorithms to enhance decision making in stock market?

Purpose and Objective

- Explore ML and DL techniques
- Understand the architecture to build a model
- Evaluate model and understand the accuracy
- Highlight complexity and limitations
- Propose potential improvements

Data Collection & Wrangling

- Yahoo Finance API - (Free)
  - Tesla
  - AMD
  - NVDA
- Features:
  - Date, open, high, low, close, and volume
- Data for last 10 years.
- Data Wrangling
  - No Null Values
  - Date -> datetime index (Time Series Dataset)

EDA - Visualizing Historical stock price

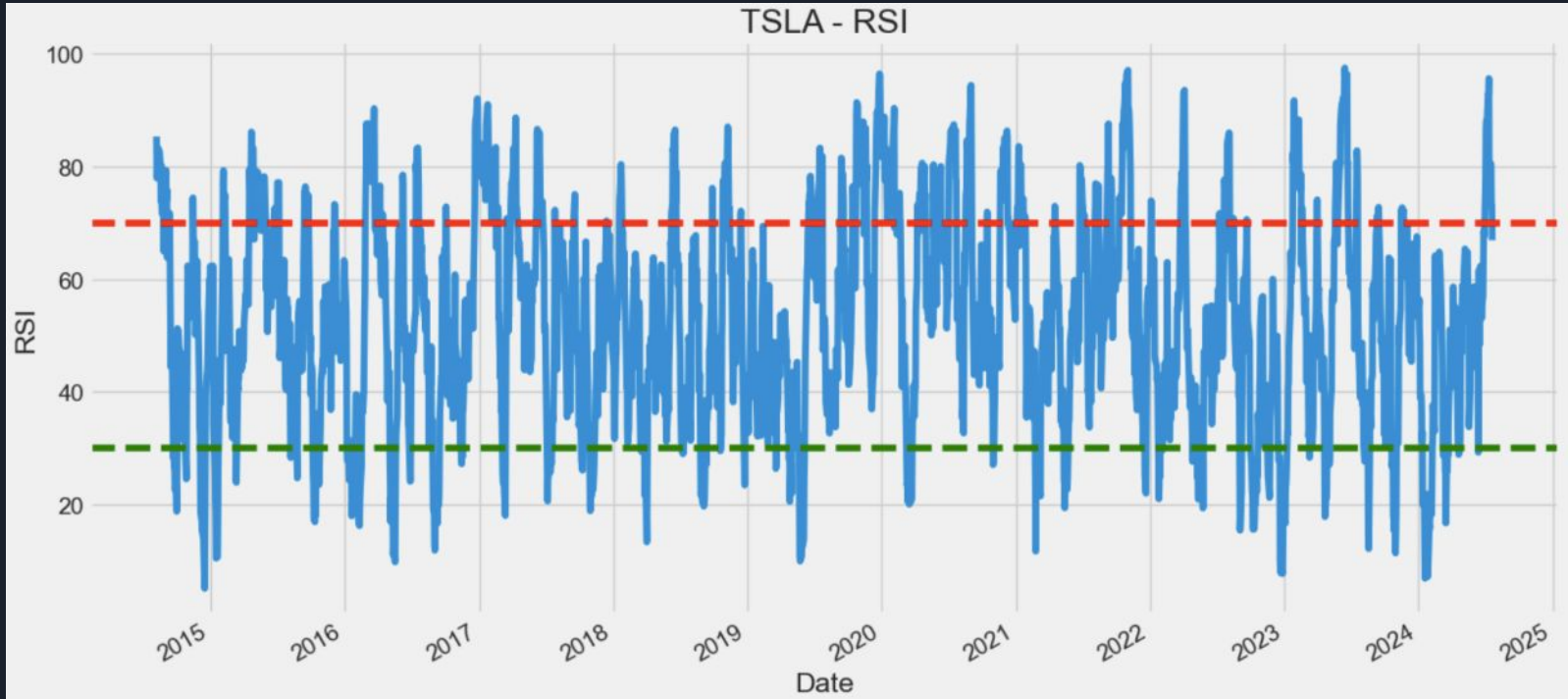Last 10 year historical stock price trend

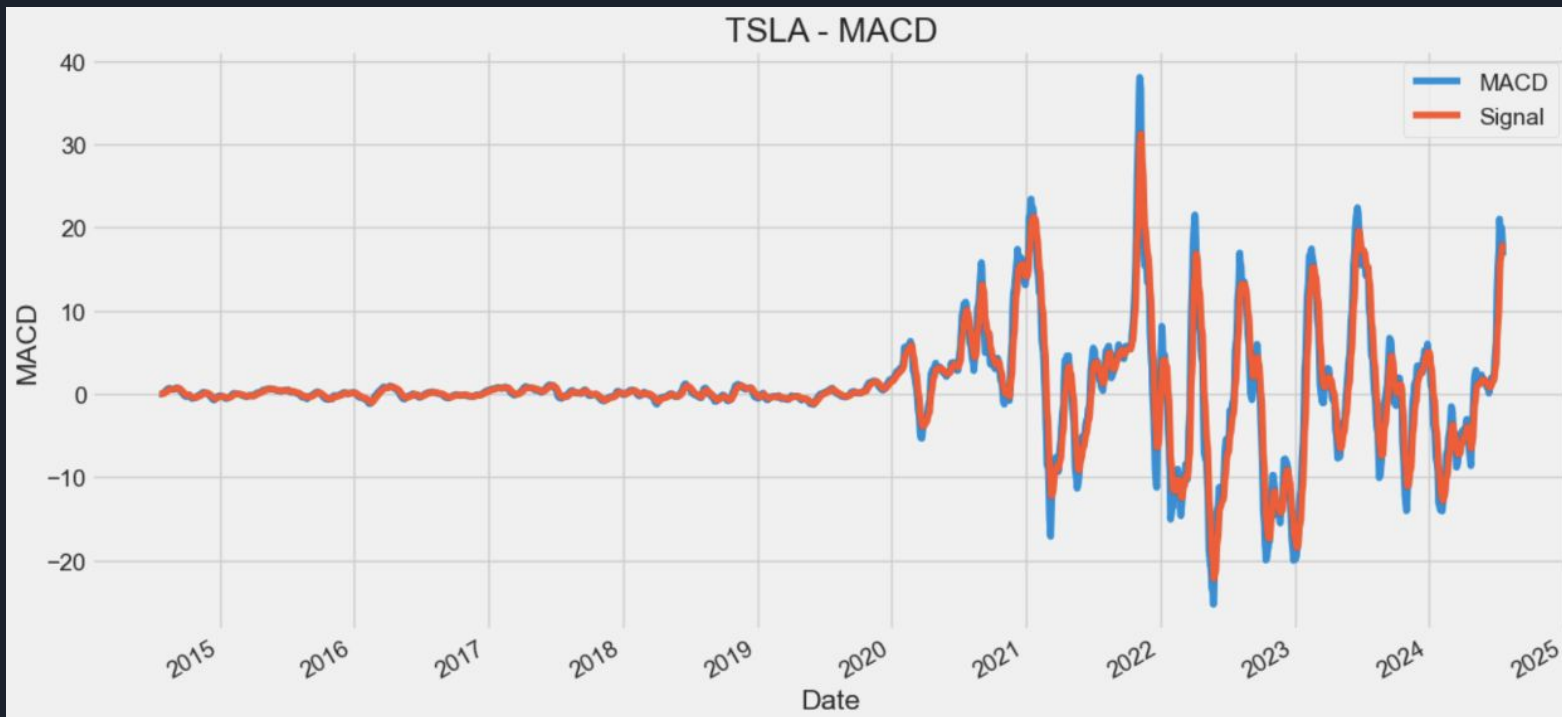# EDA - Feature Engineering 1

1. Moving Average (30 & 90 day MA)

EDA - Feature Engineering 2

2. Relative Strength Index (RSI)

EDA - Feature Engineering 3
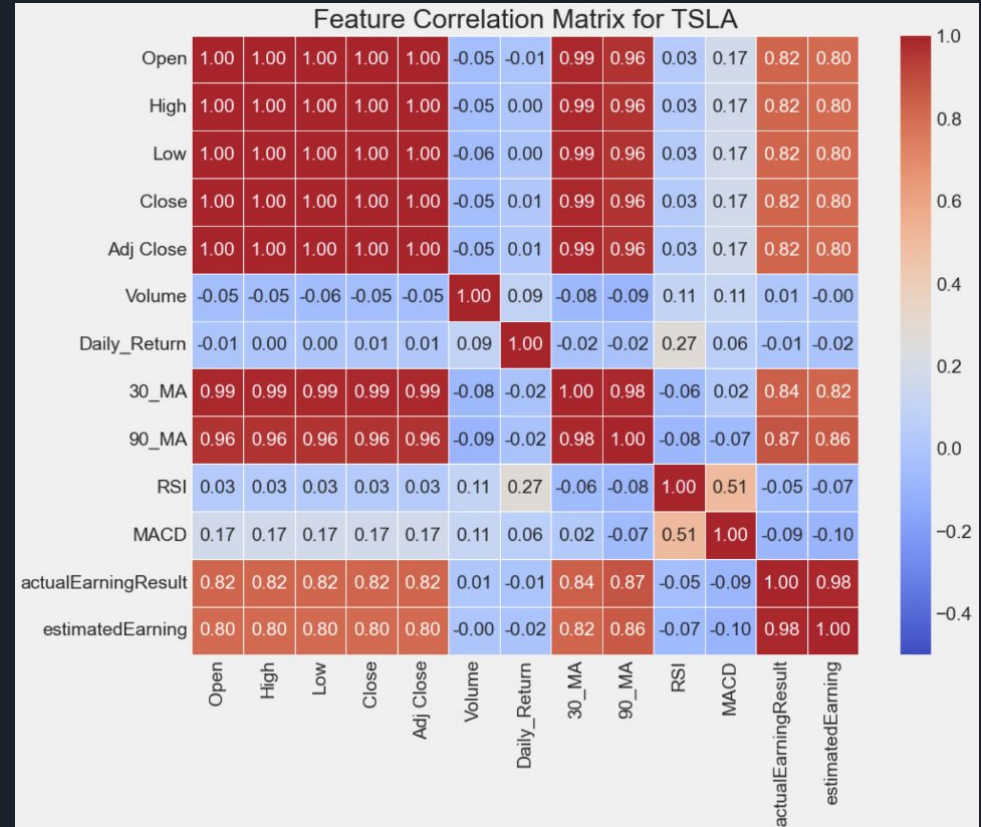
3. Moving Average Convergence Divergence (MACD)

EDA - Feature Engineering 4

4. Earnings Impact (actual vs estimated EPS)

EDA - Feature Correlation

- Why chose these Features?
- Dependent variable - Close price

Feature Correlation Matrix for TSLA

	Open	High	Low	Close	Adj Close	Volume	Daily_Return	30_MA	90_MA	RSI	MACD	actualEarningResult	estimatedEarning
Open	1.00	1.00	1.00	1.00	1.00	-0.05	-0.01	0.99	0.96	0.03	0.17	0.82	0.80
High	1.00	1.00	1.00	1.00	1.00	-0.05	0.00	0.99	0.96	0.03	0.17	0.82	0.80
Low	1.00	1.00	1.00	1.00	1.00	-0.06	0.00	0.99	0.96	0.03	0.17	0.82	0.80
Close	1.00	1.00	1.00	1.00	1.00	-0.05	0.01	0.99	0.96	0.03	0.17	0.82	0.80
Adj Close	1.00	1.00	1.00	1.00	1.00	-0.05	0.01	0.99	0.96	0.03	0.17	0.82	0.80
Volume	-0.05	-0.05	-0.06	-0.05	-0.05	1.00	0.09	-0.08	-0.09	0.11	0.11	0.01	-0.00
Daily_Return	-0.01	0.00	0.00	0.01	0.01	0.09	1.00	-0.02	-0.02	0.27	0.06	-0.01	-0.02
30_MA	0.99	0.99	0.99	0.99	0.99	-0.08	-0.02	1.00	0.98	-0.06	0.02	0.84	0.82
90_MA	0.96	0.96	0.96	0.96	0.96	-0.09	-0.02	0.98	1.00	-0.08	-0.07	0.87	0.86
RSI	0.03	0.03	0.03	0.03	0.03	0.11	0.27	-0.06	-0.08	1.00	0.51	-0.05	-0.07
MACD	0.17	0.17	0.17	0.17	0.17	0.11	0.06	0.02	-0.07	0.51	1.00	-0.09	-0.10
actualEarningResult	0.82	0.82	0.82	0.82	0.82	0.01	-0.01	0.84	0.87	-0.05	-0.09	1.00	0.98
estimatedEarning	0.80	0.80	0.80	0.80	0.80	-0.00	-0.02	0.82	0.86	-0.07	-0.10	0.98	1.00

Modelling - Dummy Regressor

Baseline Reference

Modelling - Traditional ML models

- Train/Test Split - 95%
- Offset by day.

Date	X1	X2	X..	y
Jan 1	1	1	1	1
Jan 2	2	2	2	2
Jan 3	3	3	3	3
:	:	:	:	:
Jan 29	29	29	29	29
Jan 30	30	30	30	30
Jan 31	31	31	31	31

Train

Date	X1	X2	X..	y[1:]
Jan 1	1	1	1	2
Jan 2	2	2	2	3
Jan 3	3	3	3	??

Test

Date	X1	X2	X..	y[1:]
Jan 29	29	29	29	30
Jan 30	30	30	30	31
Jan 31	31	31	31	??

```
Evaluation results for TSLA:
Linear Regression: MAE = 4.761281658580042, RMSE = 6.569041060293892
Decision Tree: MAE = 8.795233356065985, RMSE = 11.056552610346348
Random Forest: MAE = 5.117666692812578, RMSE = 7.1659821955096525
Evaluation results for AMD:
Linear Regression: MAE = 3.7955258079826337, RMSE = 5.225885299599488
Decision Tree: MAE = 9.825703392344073, RMSE = 13.465099004964687
Random Forest: MAE = 7.238814009359062, RMSE = 10.82822204466515
Evaluation results for NVDA:
Linear Regression: MAE = 2.420699357476408, RMSE = 3.2430916311819407
Decision Tree: MAE = 34.997586526161385, RMSE = 40.44316384673505
Random Forest: MAE = 35.609578680873895, RMSE = 41.023065318460425
```

Modelling - Traditional ML models

- TSLA RMSE = $6.56, TSLA MAE = $4.76
- Prediction lag?
- High Accuracy but not real-time prediction.
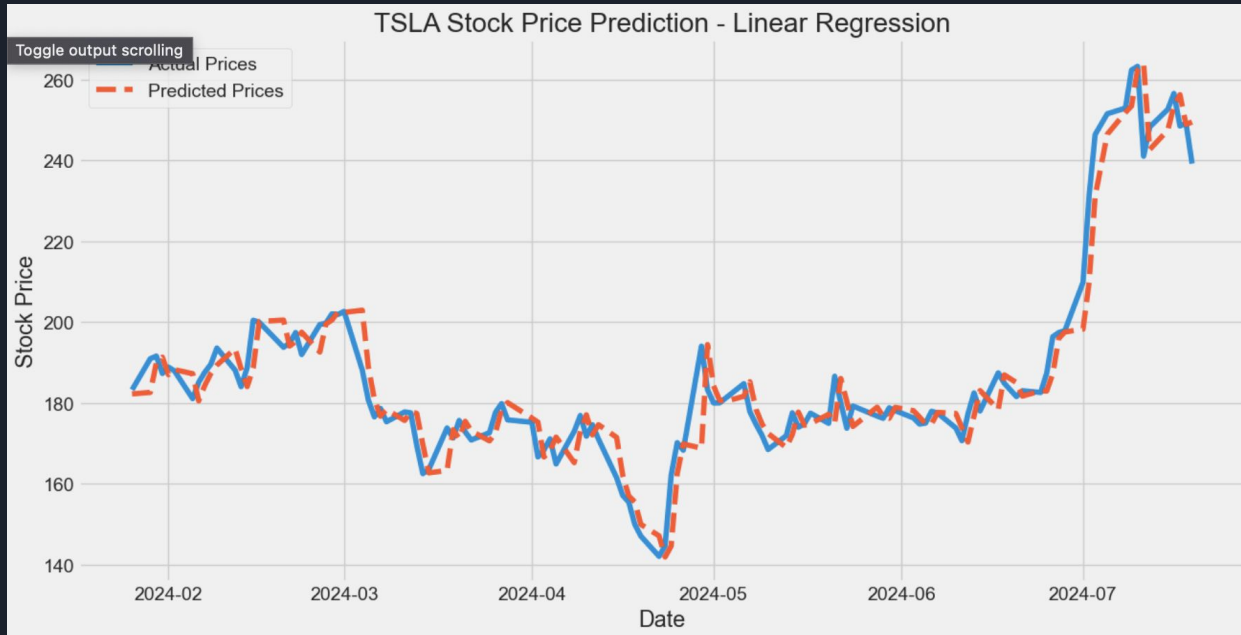
Modelling - LSTM - Data split

- Train/Test split - 95%
- Time-step - 30 days

Modelling - LSTM - Model/Hypertuning

- Multi-Layered model creation.
- Hypertune using Keras - Random Search

```
Trial 5 Complete [00h 20m 41s]
val_loss: 0.0003068597870878875

Best val_loss So Far: 0.00028865167405456305
Total elapsed time: 00h 57m 46s
Optimal hyperparameters for NVDA:
units1: 160
dropout1: 0.0
units2: 256
dropout2: 0.4
```

```
3/3 [==============================] – 0s 31ms/step
Hypertuned Evaluation results for TSLA:
MAE = 4.963618891839849, RMSE = 7.159550316857211
```
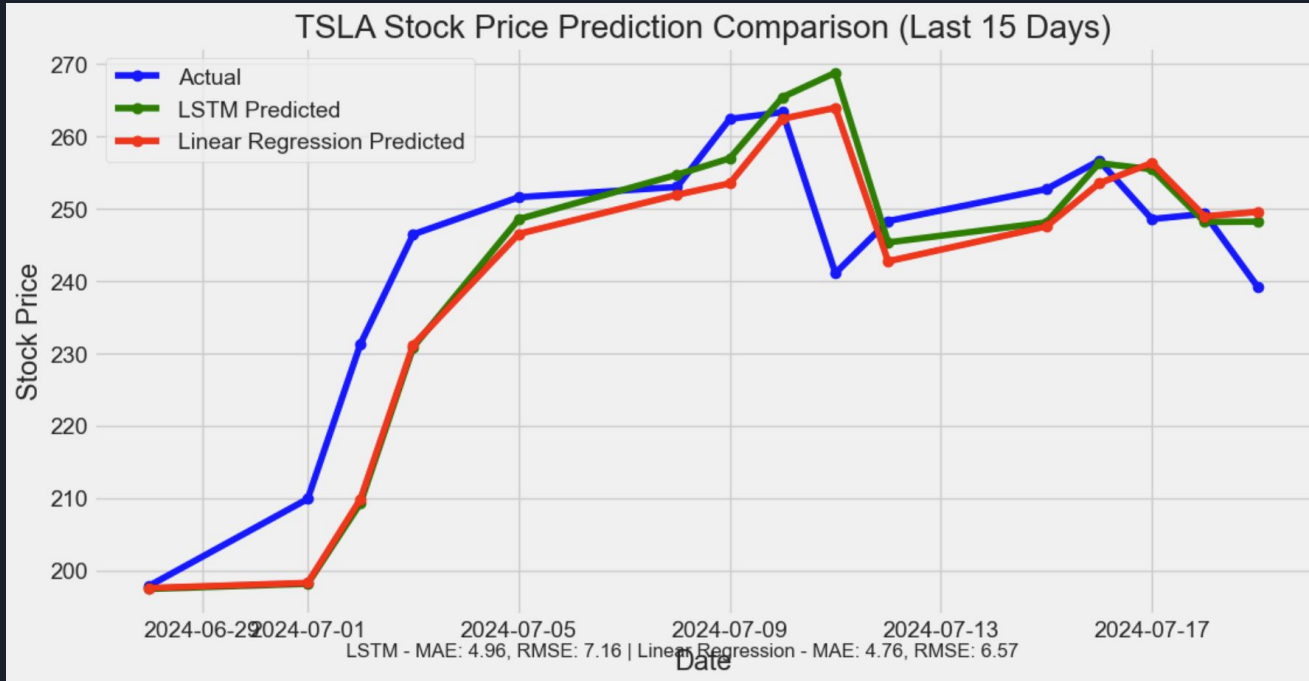
Modelling - LSTM - Prediction Trend

- RMSE = $7.16, MAE = $4.96.
- Worse than LR Model??
- Still not real-time
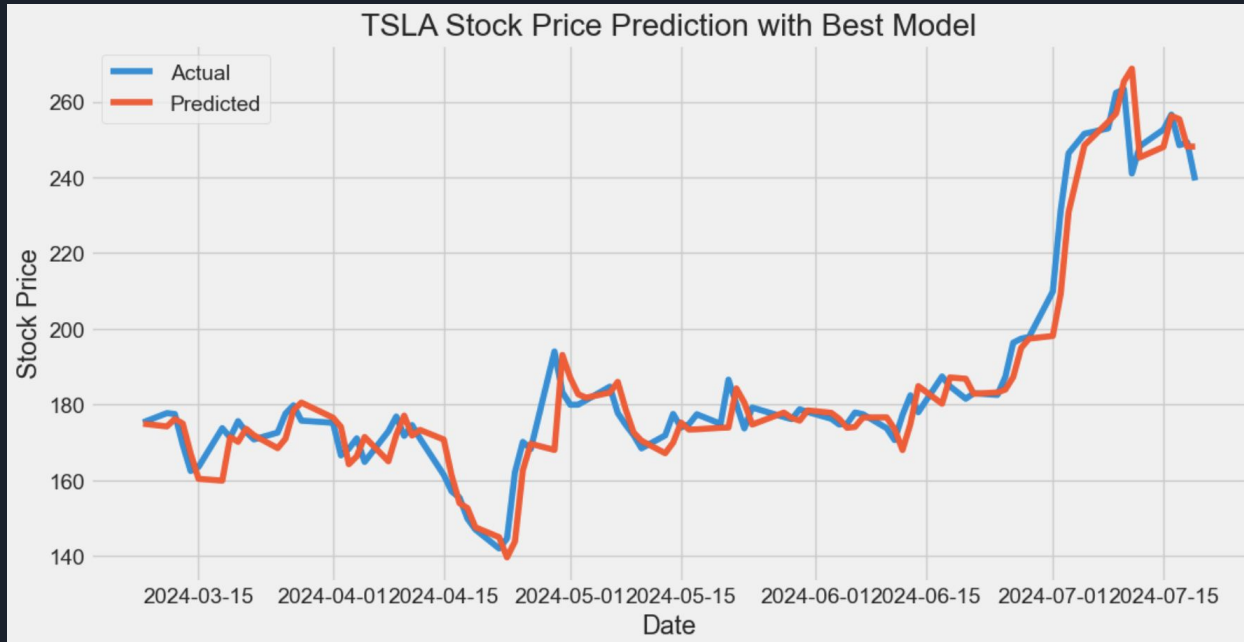
Modelling - LR VS LSTM - Closer look

- LR seems much better.
- Yesterday's prediction + little something.

TSLA Stock Price Prediction Comparison (Last 15 Days)

Improvement Recommendations

1. Data Collection
   a. Real time data - min trend.
2. Feature Engineering
   a. More Technical indicators - US $ index, unemployment rate, interest rates, consumer sentiment index.
   b. Sentiment Analysis - company events, news articles, social media trends, X thread trends.
3. Model Complexity
   a. Bi-directional LSTM multi-layers