STOCK PRICE PREDICTION

Agenda

Introduction

Problem Statement

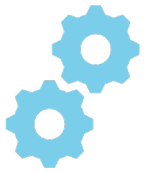
Value

Datasets & Exploration

Data collection

Predictors and target variable

Exploratory Data Analysis



Methodology & Results

Model Application

Model Selection

Implementation of Select Model

Model Validation

Conclusion & Future Work

Findings

Proposed Future Extensions

Introduction

- Stock market is a public market where people can invest or withdraw their funds from any publicly listed company
- It allows us to buy/sell unit or fractional shares of any of the listed companies
- Stock price fluctuations depend on the company’s profits and other financial factors
- If there are more sellers than buyers, the stock price drops and vice versa

Problem Statement

To predict next day's stock price using historic stock prices and financial factors and simultaneously analyze the tweet's sentiment related to that stock.

Value of the Project

Accurate stock price predictions can help people make profit and save them from losing money in the stock market.

Project Purpose

To make accurate predictions of stock prices in order to make informed buying/selling decisions related to the selected companies.

Objectives

- To train different data modeling algorithms
- Select the best performing model using cross validation
- Test the model results on test dataset
- Use time series modeling to compare results

Stock Price Chart - Features

Attributes		Adj Close	Close	High	Low	Open	Volume
Date	Symbols						
2015-07-24	AAPL	114.533791	124.500000	125.739998	123.900002	125.320000	42162300.0
2015-07-27	AAPL	112.942253	122.769997	123.610001	122.120003	123.089996	44455500.0
2015-07-28	AAPL	113.503433	123.379997	123.910004	122.550003	123.379997	33618100.0

Dataset Sources

- Top 5 Movers in the stock Market:
  - Apple
  - Microsoft
  - Netflix
  - Amazon
  - NVIDIA
- Historic data:
  - 2017
  - 2018
  - 2019

- Tickers used for Twitter Data:
  - $AAPL, #AAPL, 'apple stock’
  - $MSFT, #MSFT, 'microsoft stock’
  - $NFLX , #NFLX, 'netflix stock'
  - $AMZN, #AMZN,'amazon stock'
  - $NVDA, #NVDA, 'nvidia stock'
- Historic data:
  - 2017
  - 2018
  - 2019

Data Collection

Attributes	Date	Symbols	Adj Close	Close	High	Low	Open	Volume
0	2016-01-04	AAPL	97.772148	105.349998	105.370003	102.000000	102.610001	67649400.0
1	2016-01-04	NFLX	109.959999	109.959999	110.000000	105.209999	109.000000	20794800.0
2	2016-01-04	AMZN	636.989990	636.989990	657.719971	627.510010	656.289978	9314500.0
3	2016-01-04	MSFT	50.258858	54.799999	54.799999	53.389999	54.320000	53778000.0
4	2016-01-04	NVDA	31.691172	32.369999	32.580002	32.040001	32.290001	8951900.0
...	...	...	...	...	...	...	...	...
6031	2019-12-31	NFLX	323.570007	323.570007	324.920013	321.089996	322.000000	3713300.0
6032	2019-12-31	AMZN	1847.839966	1847.839966	1853.260010	1832.229980	1842.000000	2506500.0
6033	2019-12-31	MSFT	156.833633	157.699997	157.770004	156.449997	156.770004	18369400.0
6034	2019-12-31	NVDA	235.052078	235.300003	235.679993	230.130005	230.899994	5775100.0
6035	2019-12-31	SPY	318.576508	321.859985	322.130005	320.149994	320.529999	57077300.0

6036 rows × 8 columns

	date	text	ticker
0	1/1/17	$ NFLX maturity 01/06/2017 Vol PutCallRatio of...	NFLX
1	1/1/17	2017 Top Picks: Tesla (TSLA), Netflix(NFLX), a...	NFLX
2	1/2/17	Just finished # TheOA Highly recommend it. Bra...	NFLX
3	1/2/17	Breaks the 122 level and we c\u2026 $ NFLX htt...	NFLX
4	1/2/17	Wall Street's Top Picks For 2017: All In One P...	NFLX
...	...	...	...
995	5/16/17	Netflix or Twitter; Which One Should You Buy S...	NFLX
996	5/16/17	# Forex Netflix fails to hold gains - Analysis...	NFLX
997	5/16/17	Dobbs: My greatest regret is being paid in cas...	NFLX
998	5/16/17	music Netflix Flashed Warnings Before The Stoc...	NFLX
999	5/16/17	Yeah they always add things that are garbage, ...	NFLX

Data Collection

Select type of Interest Rate Data

Daily Treasury Yield Curve Rates Go

Select Time Period

2019 Go

Date	1 Mo	2 Mo	3 Mo	6 Mo	1 Yr	2 Yr	3 Yr	5 Yr	7 Yr	10 Yr	20 Yr	30 Yr
01/02/19	2.40	2.40	2.42	2.51	2.60	2.50	2.47	2.49	2.56	2.66	2.83	2.97
01/03/19	2.42	2.42	2.41	2.47	2.50	2.39	2.35	2.37	2.44	2.56	2.75	2.92
01/04/19	2.40	2.42	2.42	2.51	2.57	2.50	2.47	2.49	2.56	2.67	2.83	2.98
01/07/19	2.42	2.42	2.45	2.54	2.58	2.53	2.51	2.53	2.60	2.70	2.86	2.99
01/08/19	2.40	2.42	2.46	2.54	2.60	2.58	2.57	2.58	2.63	2.73	2.88	3.00
01/09/19	2.40	2.42	2.45	2.52	2.59	2.56	2.54	2.57	2.64	2.74	2.90	3.03
01/10/19	2.42	2.42	2.43	2.51	2.59	2.56	2.54	2.56	2.63	2.74	2.92	3.06
01/11/19	2.41	2.43	2.43	2.50	2.58	2.55	2.51	2.52	2.60	2.71	2.90	3.04
01/14/19	2.42	2.43	2.45	2.52	2.57	2.53	2.51	2.53	2.60	2.71	2.91	3.06
01/15/19	2.41	2.43	2.45	2.52	2.57	2.53	2.51	2.53	2.61	2.72	2.92	3.08
01/16/19	2.41	2.40	2.43	2.49	2.57	2.55	2.53	2.54	2.62	2.73	2.92	3.07
01/17/19	2.41	2.41	2.42	2.50	2.57	2.56	2.55	2.58	2.66	2.75	2.93	3.07
01/18/19	2.40	2.40	2.41	2.50	2.60	2.62	2.60	2.62	2.70	2.79	2.95	3.09
01/22/19	2.38	2.40	2.43	2.51	2.59	2.58	2.55	2.57	2.65	2.74	2.91	3.06
01/23/19	2.37	2.38	2.41	2.51	2.59	2.58	2.57	2.59	2.66	2.76	2.93	3.07
01/24/19	2.38	2.41	2.37	2.50	2.58	2.56	2.54	2.55	2.62	2.72	2.89	3.04
01/25/19	2.36	2.41	2.39	2.51	2.60	2.60	2.58	2.59	2.66	2.76	2.92	3.06
01/28/19	2.39	2.41	2.42	2.51	2.60	2.60	2.58	2.58	2.65	2.75	2.92	3.06
01/29/19	2.39	2.41	2.42	2.51	2.60	2.56	2.54	2.55	2.61	2.72	2.90	3.04
01/30/19	2.40	2.39	2.42	2.50	2.57	2.52	2.49	2.49	2.58	2.70	2.90	3.06

Source: LINK

	currentRatio	quickRatio	cashRatio	daysOfSalesOutstanding	daysOfInventoryOutstanding	operatingCycle	daysOfPayablesOutstanding
2019-12	0.901222	0.798290	0.732010	30.335187	0.0	30.335187	535.797489
2019-09	0.734063	0.675290	0.611061	32.441259	0.0	32.441259	625.001788
2019-06	0.847587	0.721699	0.721699	0.000000	0.0	0.000000	642.453272
2019-03	0.607907	0.488284	0.488284	0.000000	0.0	0.000000	674.259637
2018-12	1.494320	1.378947	0.584908	449.069571	0.0	449.069571	623.689714
2018-09	1.386154	1.279136	0.487139	455.223283	0.0	455.223283	764.710028
2018-06	1.537336	1.432566	0.642434	448.765132	0.0	448.765132	795.224682

```
NFLXratio_4Y.columns
```

```
Index(['currentRatio', 'quickRatio', 'cashRatio', 'daysOfSalesOutstanding',
       'daysOfInventoryOutstanding', 'operatingCycle',
       'daysOfPayablesOutstanding', 'cashConversionCycle', 'grossProfitMargin',
       'operatingProfitMargin', 'pretaxProfitMargin', 'netProfitMargin',
       'effectiveTaxRate', 'returnOnAssets', 'returnOnEquity',
       'returnOnCapitalEmployed', 'netIncomePerEBT', 'ebtPerEbit',
       'ebitPerRevenue', 'debtRatio', 'debtEquityRatio',
       'longTermDebtToCapitalization', 'totalDebtToCapitalization',
       'interestCoverage', 'cashFlowToDebtRatio', 'companyEquityMultiplier',
       'receivablesTurnover', 'payablesTurnover', 'inventoryTurnover',
       'fixedAssetTurnover', 'assetTurnover', 'operatingCashFlowPerShare',
       'freeCashFlowPerShare', 'cashPerShare', 'payoutRatio',
       'operatingCashFlowSalesRatio', 'freeCashFlowOperatingCashFlowRatio',
       'cashFlowCoverageRatios', 'shortTermCoverageRatios',
       'capitalExpenditureCoverageRatio', 'dividendPaidAndCapexCoverageRatio',
       'dividendPayoutRatio', 'priceBookValueRatio', 'priceToBookRatio',
       'priceToSalesRatio', 'priceEarningsRatio', 'priceToFreeCashFlowsRatio',
       'priceToOperatingCashFlowsRatio', 'priceCashFlowRatio',
       'priceEarningsToGrowthRatio', 'priceSalesRatio', 'dividendYield',
       'enterpriseValueMultiple', 'priceFairValue'],
      dtype='object')
```

Target Variable

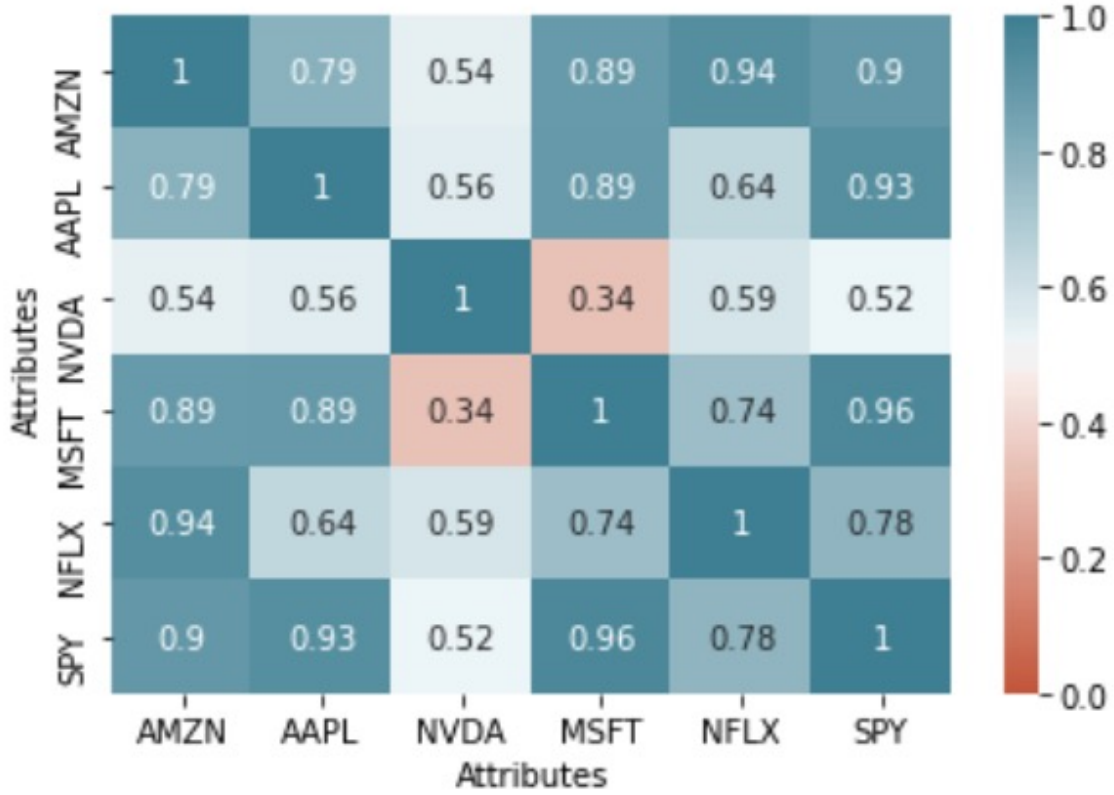
	Date	next_day_price	price	High	Low	Open	Volume	currentRatio	quickRatio	debtEquityRatio	priceEarningsRatio	price_l
0	2017-01-03	103.183281	100.830811	106.370003	99.379997	104.400002	37549900.0	4.070064	3.676629	0.805409	112.347325	6
1	2017-01-04	100.563942	103.183281	105.500000	101.529999	103.400002	29980500.0	4.070064	3.676629	0.805409	112.347325	6
2	2017-01-05	101.908211	100.563942	105.820000	101.050003	104.529999	24607400.0	4.070064	3.676629	0.805409	112.347325	6
3	2017-01-06	106.039902	101.908211	104.250000	101.199997	102.849998	20571400.0	4.070064	3.676629	0.805409	112.347325	6
4	2017-01-09	105.239250	106.039902	108.000000	103.500000	103.500000	22906200.0	4.070064	3.676629	0.805409	112.347325	6
...	...	...	...	...	...	...	...	...	...	...	...	
3755	2019-12-20	282.054138	277.525391	282.649994	278.559998	282.230011	68994500.0	1.540126	1.168041	2.741004	82.259199	24
3756	2019-12-23	282.322266	282.054138	284.250000	280.369995	280.529999	24643000.0	1.540126	1.168041	2.741004	82.259199	24
3757	2019-12-24	287.923645	282.322266	284.890015	282.920013	284.690002	12119700.0	1.540126	1.168041	2.741004	82.259199	24
3758	2019-12-26	287.814392	287.923645	289.980011	284.700012	284.820007	23280300.0	1.540126	1.168041	2.741004	82.259199	24
3759	2019-12-27	289.522614	287.814392	293.970001	288.119995	291.119995	36566500.0	1.540126	1.168041	2.741004	82.259199	24

Exploratory Data Analysis

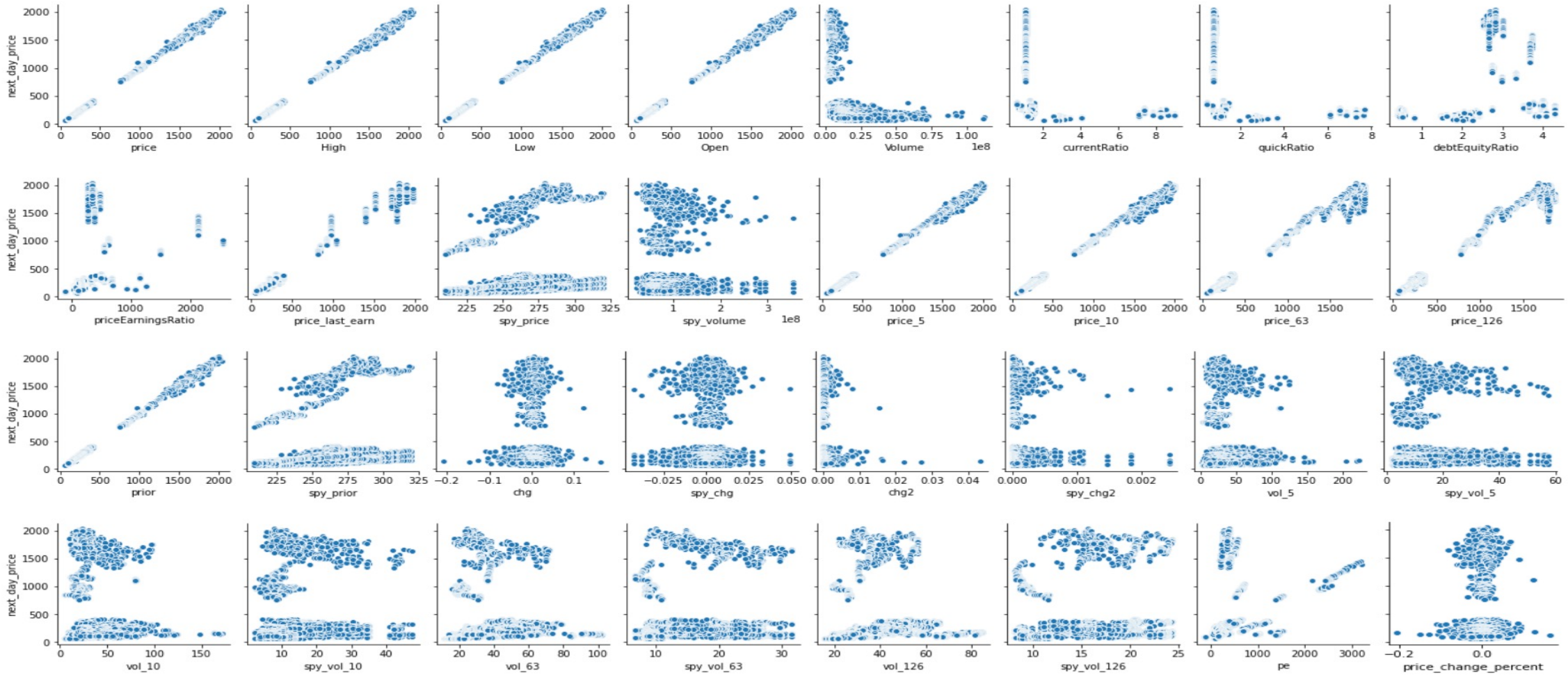
Attributes	AMZN	AAPL	NVDA	MSFT	NFLX	SPY
count	754.000000	754.000000	754.000000	754.000000	754.000000	754.000000
mean	1466.788952	177.857955	184.436273	98.375531	271.255504	259.672692
std	384.674691	35.569313	48.038133	26.776836	83.418412	25.696077
min	753.669983	110.070328	94.505310	58.563644	127.489998	210.625656
25%	1003.184998	150.924412	150.985268	71.940742	185.692493	237.280960
50%	1605.509949	170.951103	179.384689	98.194199	291.565002	261.110184
75%	1787.742462	200.732815	223.900753	117.171322	346.647491	279.063171
max	2039.510010	291.638000	287.642822	157.705505	418.970001	319.645508

EDA – Correlation Matrix

Attributes	AMZN	AAPL	NVDA	MSFT	NFLX	SPY
Attributes						
AMZN	1.000000	0.794179	0.544210	0.888374	0.936881	0.898121
AAPL	0.794179	1.000000	0.556869	0.894986	0.642273	0.930942
NVDA	0.544210	0.556869	1.000000	0.337134	0.585352	0.522064
MSFT	0.888374	0.894986	0.337134	1.000000	0.744531	0.964495
NFLX	0.936881	0.642273	0.585352	0.744531	1.000000	0.776067
SPY	0.898121	0.930942	0.522064	0.964495	0.776067	1.000000



EDA – Pairplot



Tweet Sentiment Analysis

Sentiment Classification

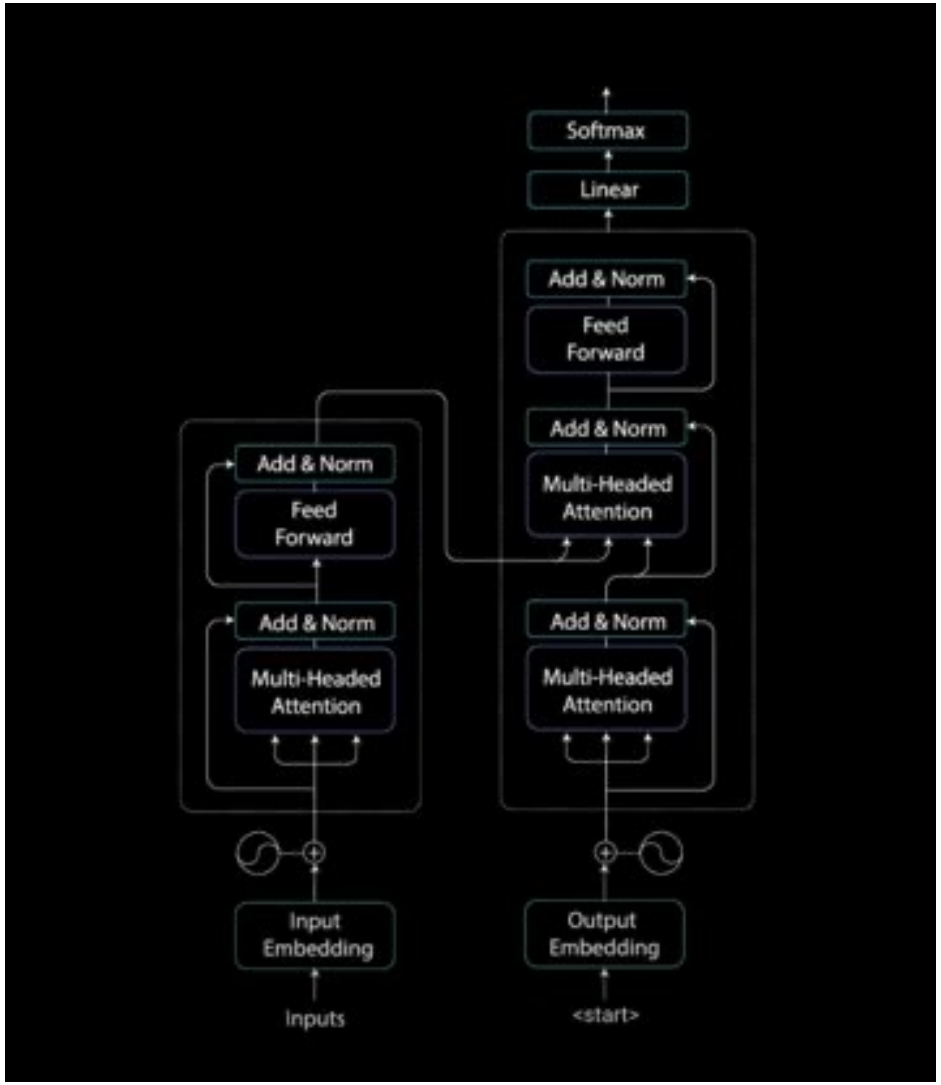
- We are using BERT(Bi-Directional Encoder Representation Transformer) to tackle the sentiment classification problem.

- We need to understand Transformers to understand BERT.

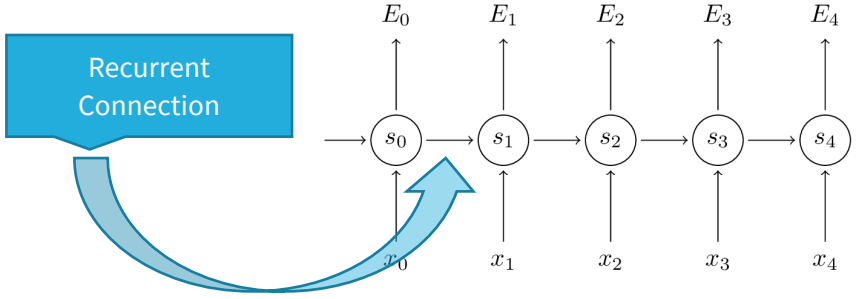
Sentiment Classification

- First Transformer model was proposed in the paper “Attention is all you need”, Ashish(Google Brain)

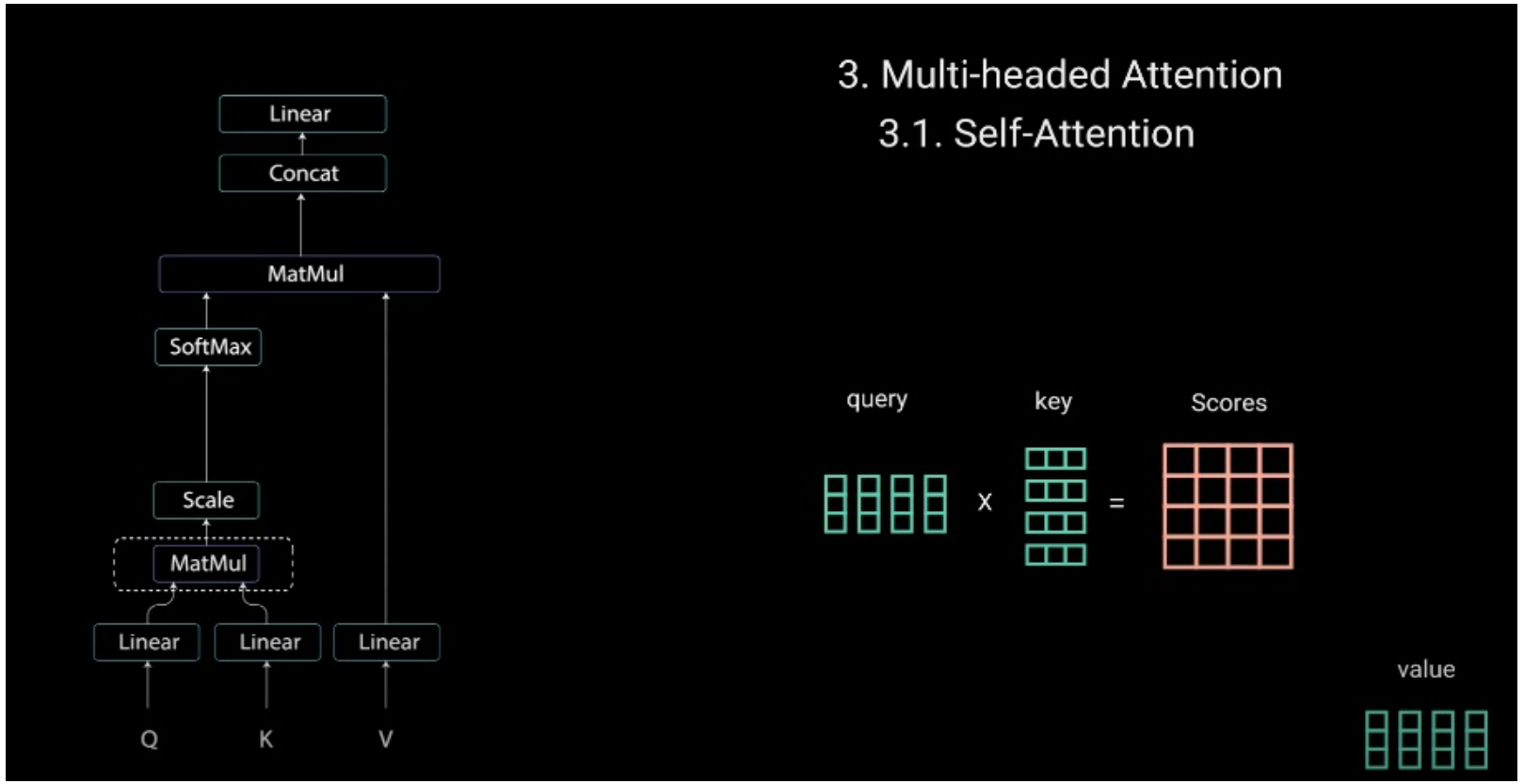
- The key features of a transformer model are:
  - It doesn’t use recurrent connections in solving language tasks.
  - It uses the concept of positional embeddings to capture the temporal information.
  - It uses Multi-Headed Attention.



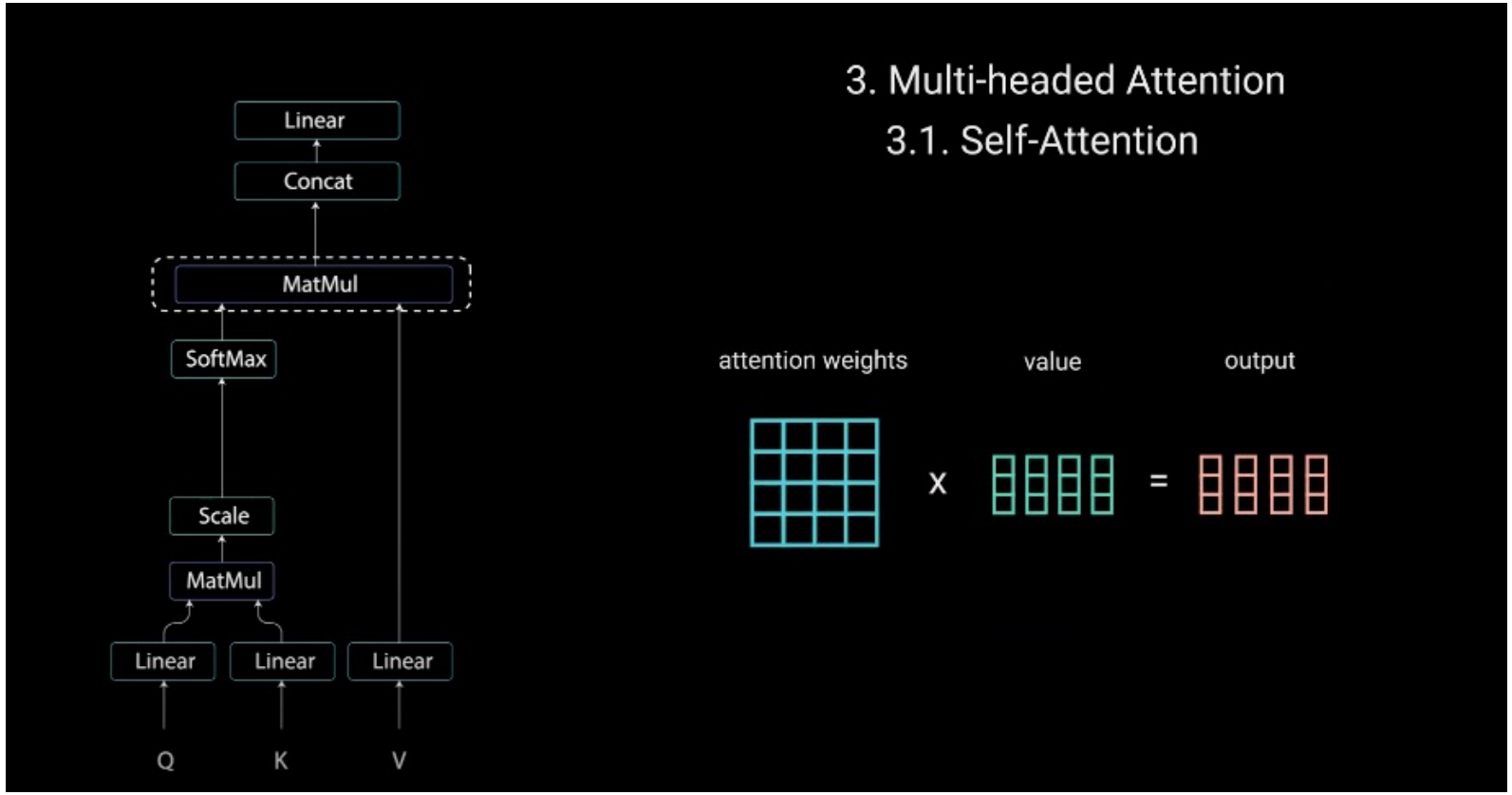
Recurrent Neural Network and Vanishing Gradient Problem:



Sentiment Classification



Sentiment Classification



Sentiment Classification

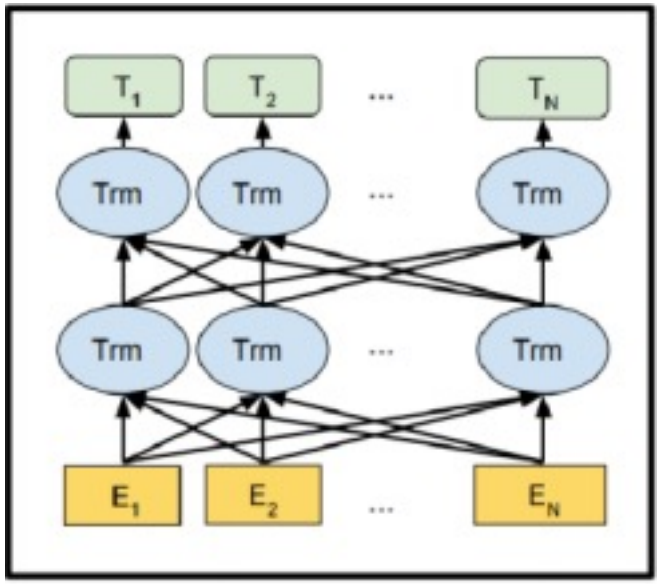


Figure 1: BERT architecture

Sentiment Classification

- We used pre-trained transformer(BERT) to generate our embeddings. Then we fed the embeddings into a Gated Recurrent Unit and a Fully Connected Network to predict the sentiment score of the text.

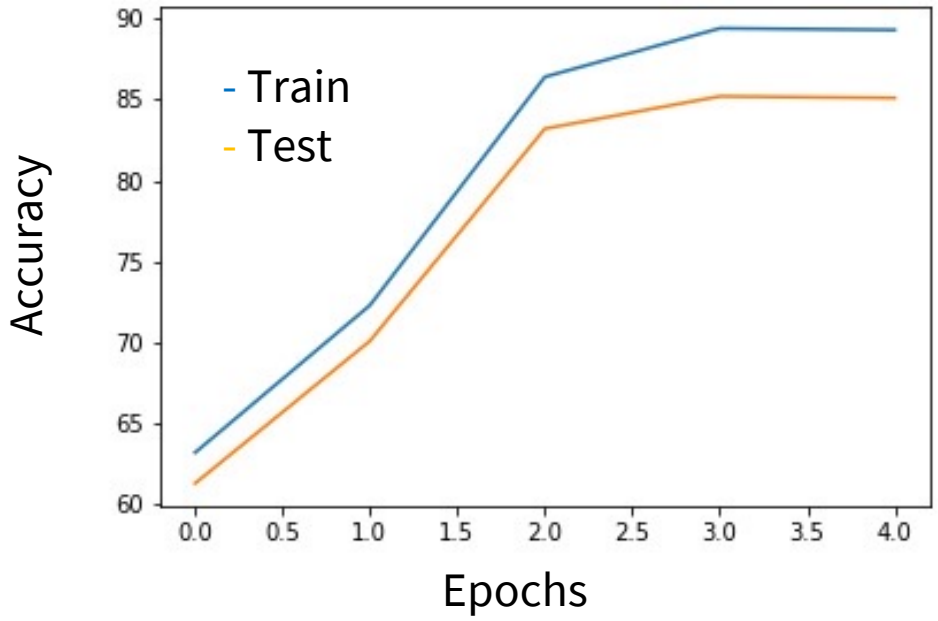
- First, we trained our model on 1.6 million arbitrary tweets with sentiments.

- Then, we fine-tuned the model on 5,000 manually labelled stock related tweets to align the model for our case.

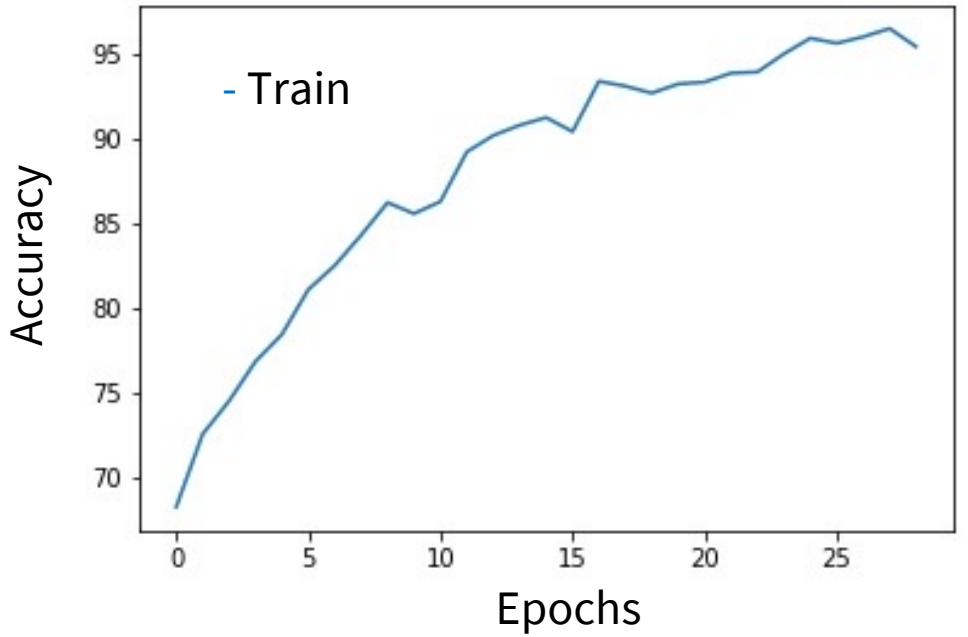
- Finally, used the resulting model to perform predictions on rest of the tweets.

Sentiment Classification

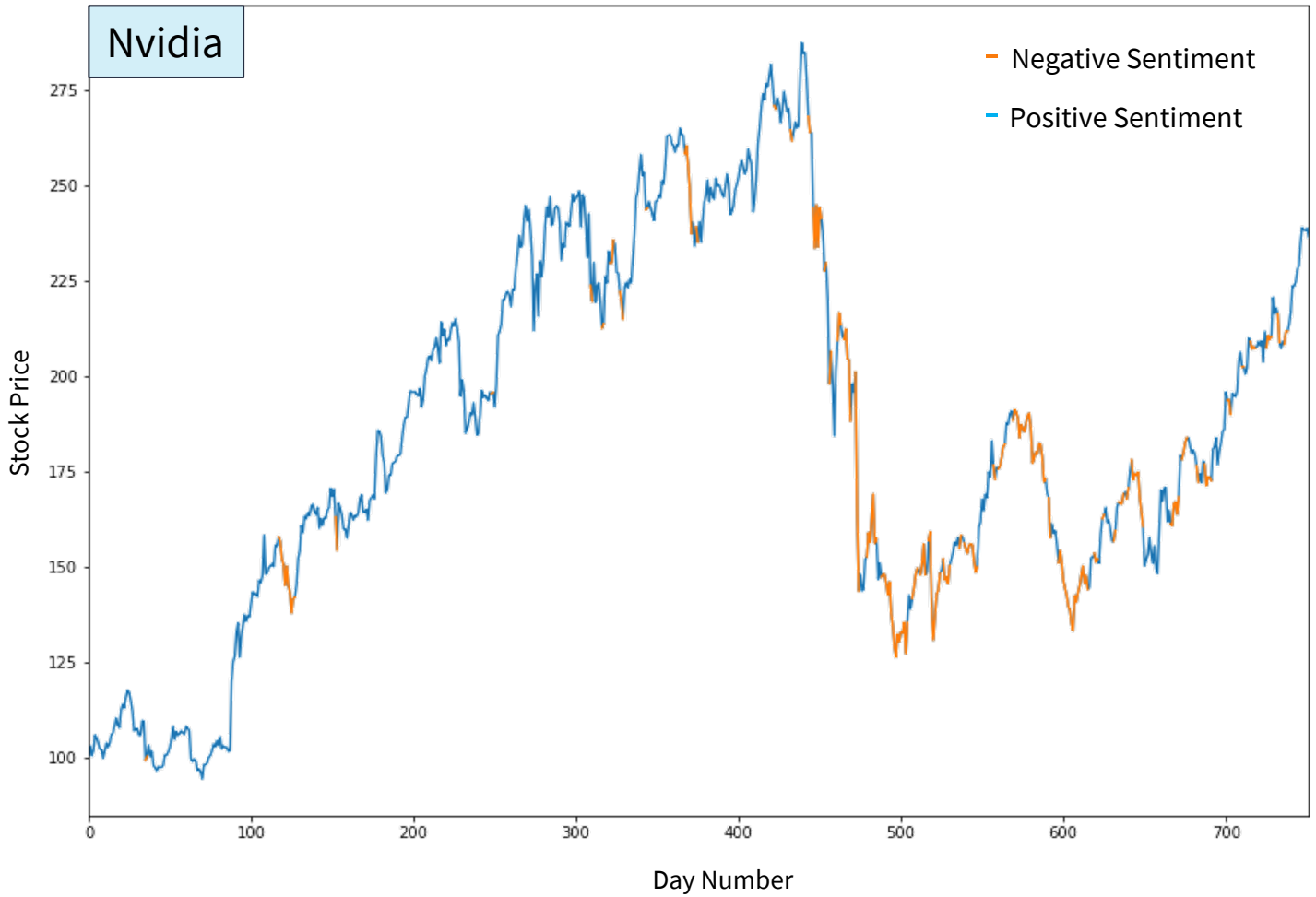
Pre-Training on Arbitrary Tweets



Pre-Training on Stock Related Tweets

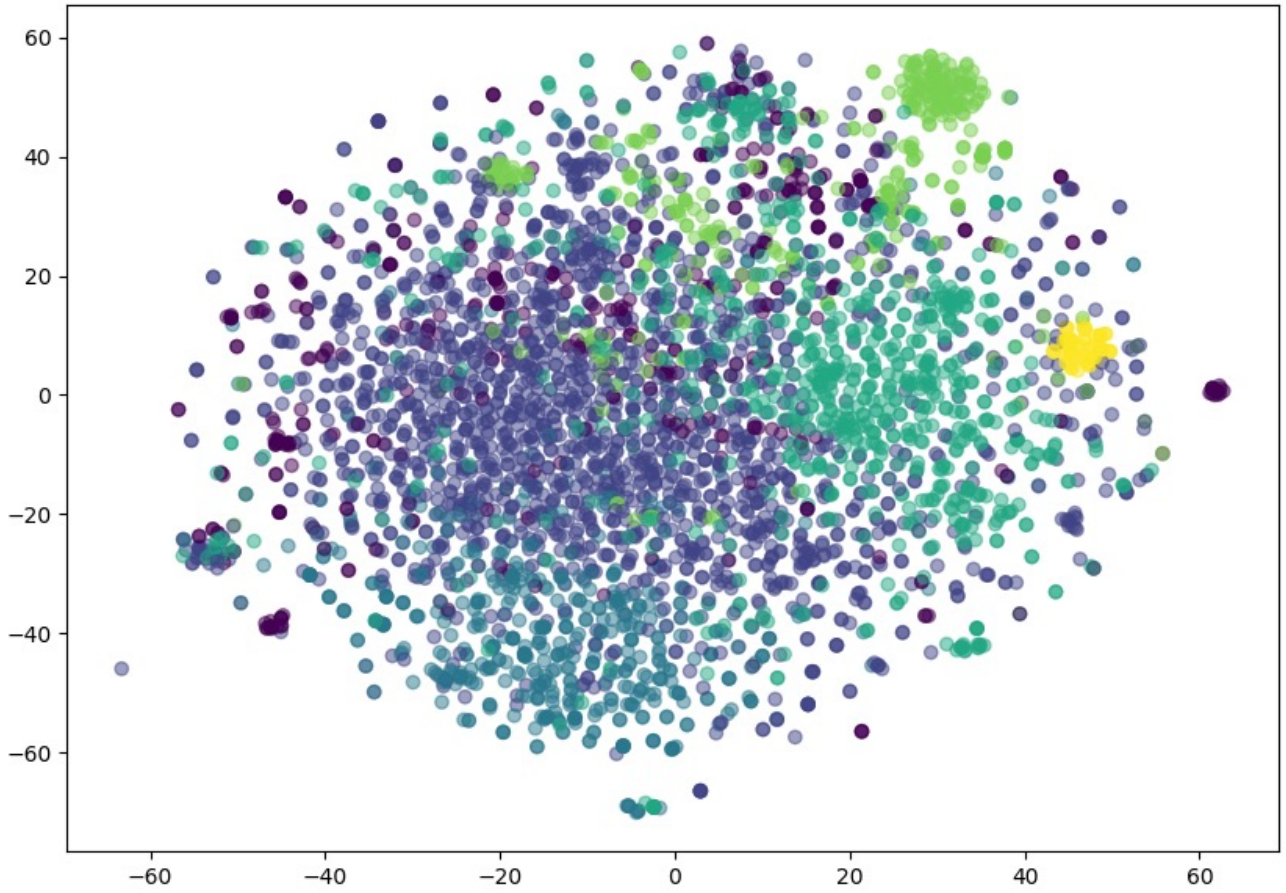


Sentiment Analysis

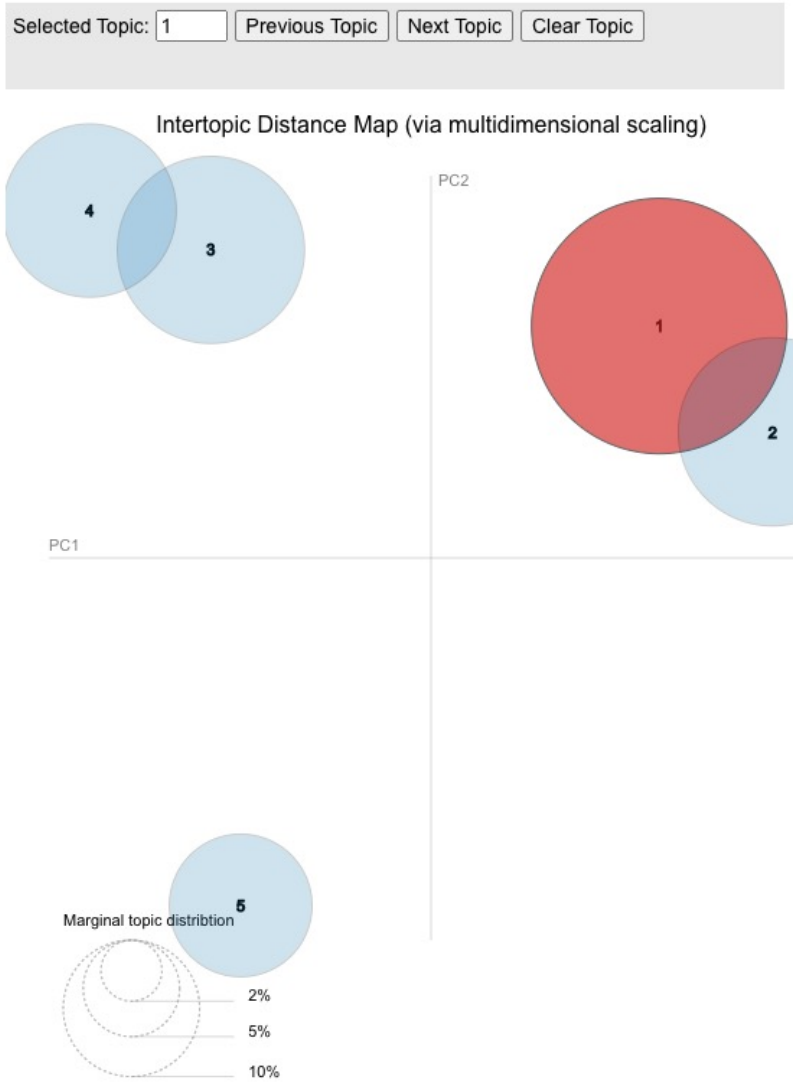


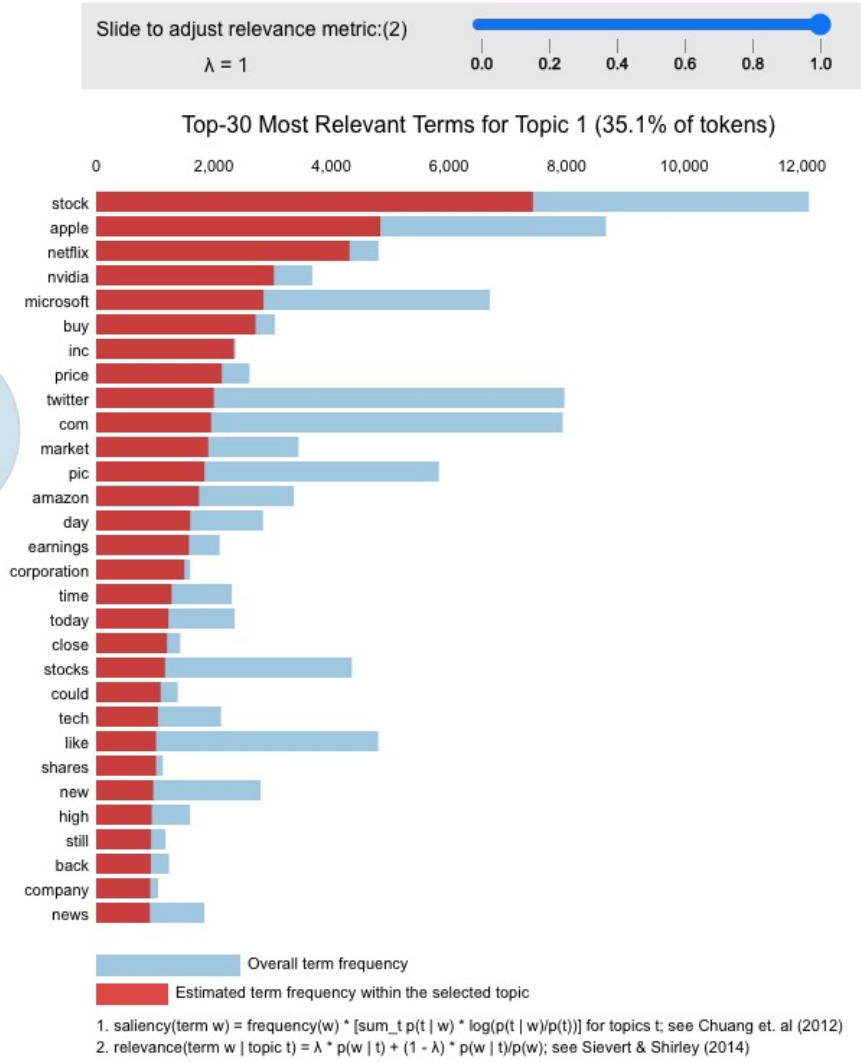
Sentiment Analysis

TSNE plot on the KMeans clusters



Sentiment Analysis





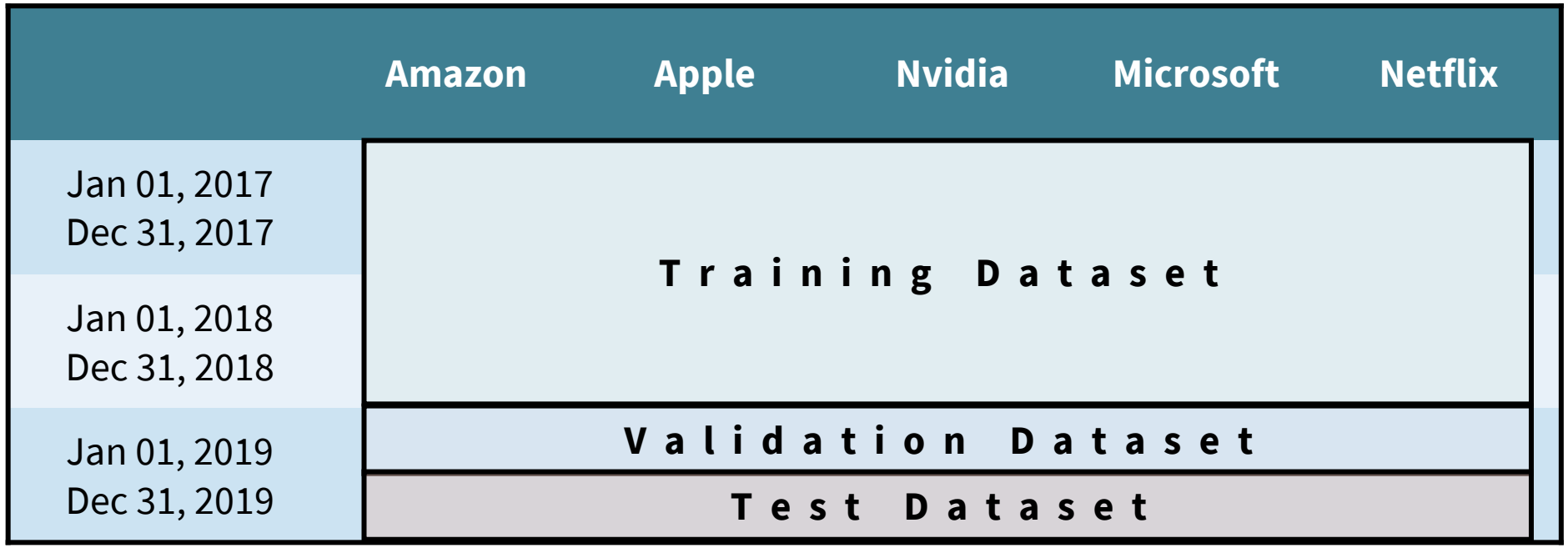
Tweet Analysis

Data Modeling

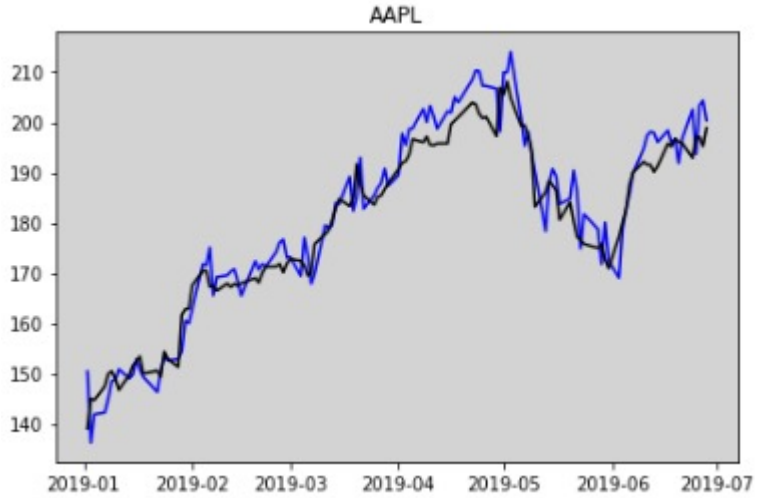
Model Training

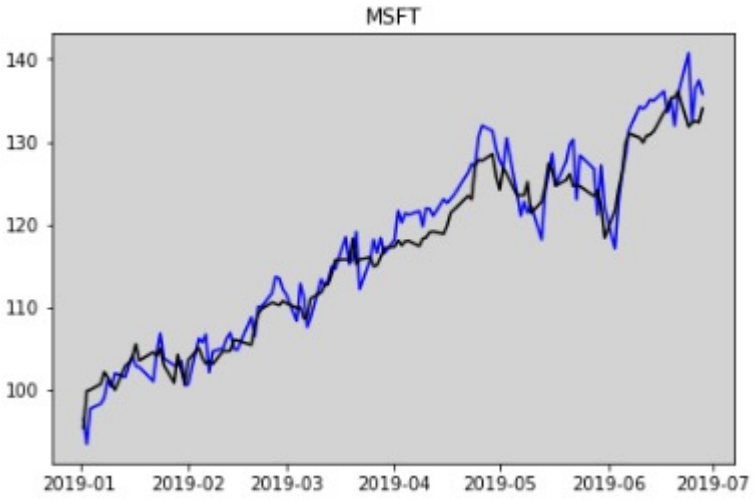
- We had two different perspectives towards the problem statement:
  - Immediate day stock prediction using Regression
    - Decision Tree
    - Random Forest
    - Gradient Boosting
    - Support Vector Regressors
  - Stock price prediction using Time Series
    - ARIMA

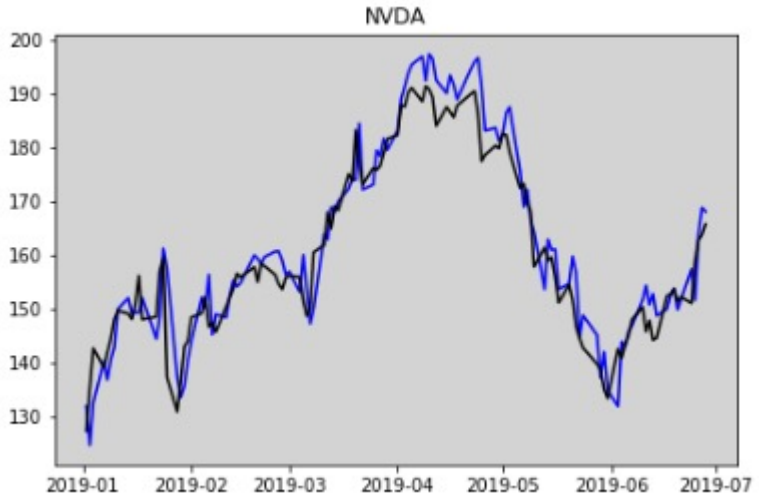
Train Test Split

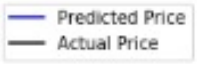


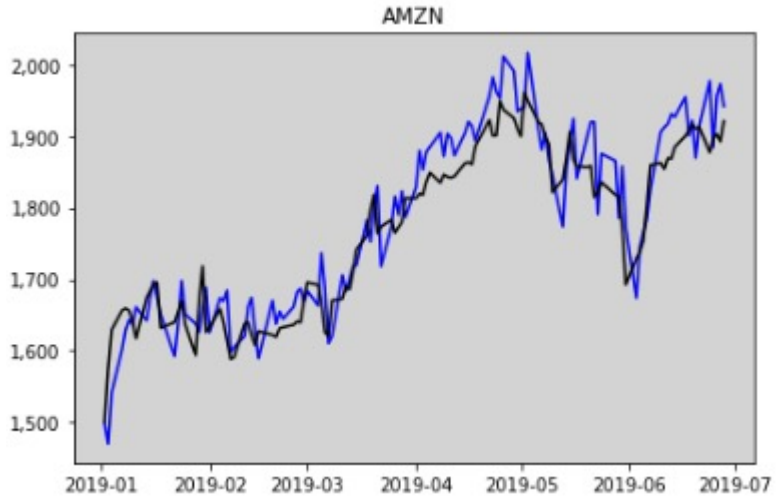
SVR

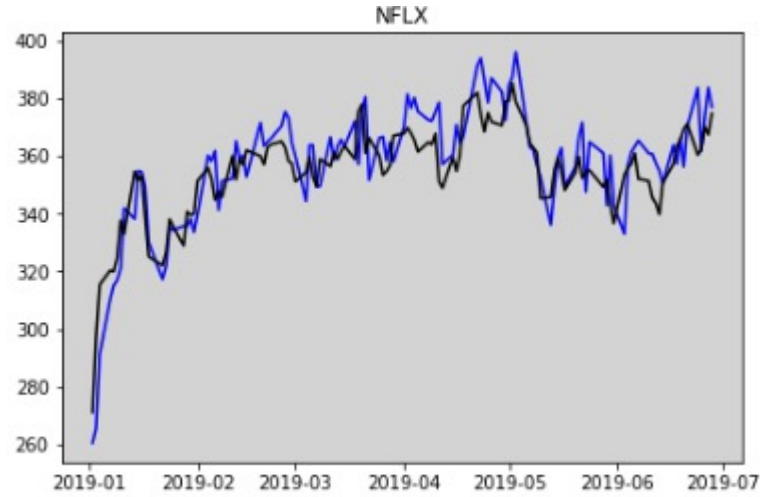

AAPL
210
200
190
180
170
160
150
140
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07


MSFT
140
130
120
110
100
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07

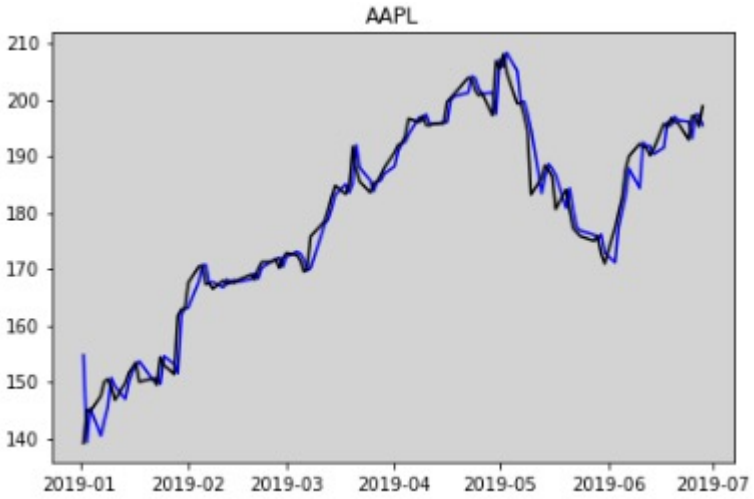

NVDA
200
190
180
170
160
150
140
130
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07

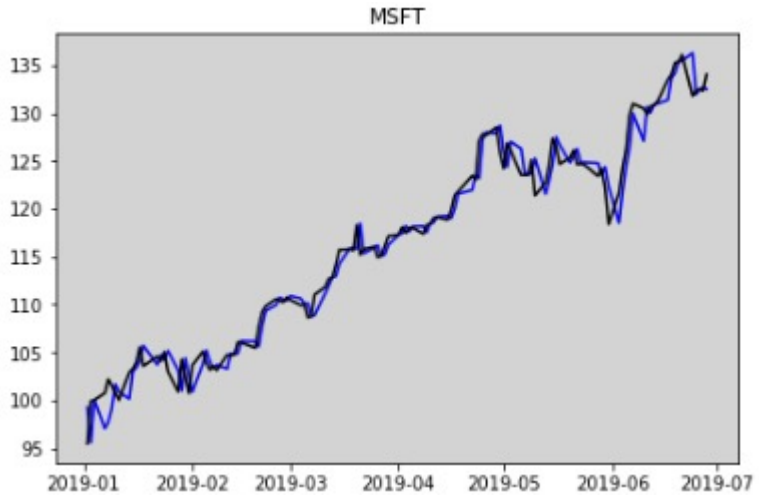

Predicted Price
Actual Price

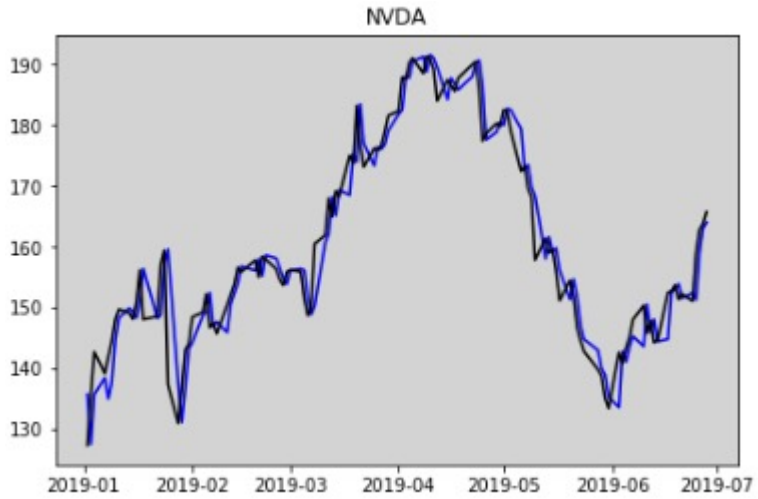

AMZN
2,000
1,900
1,800
1,700
1,600
1,500
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07

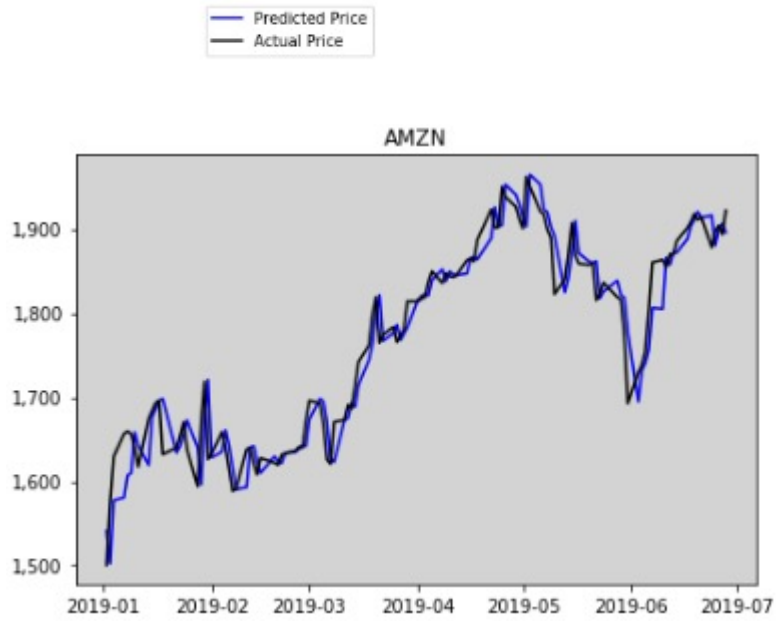

NFLX
400
380
360
340
320
300
280
260
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07

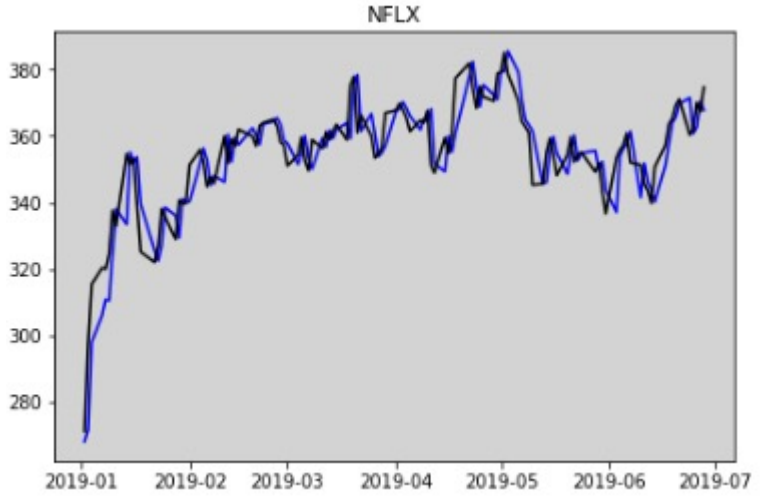
Decision Tree



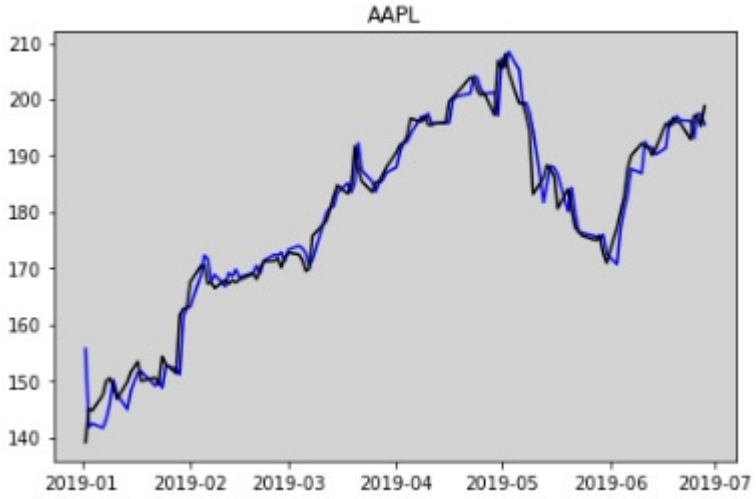


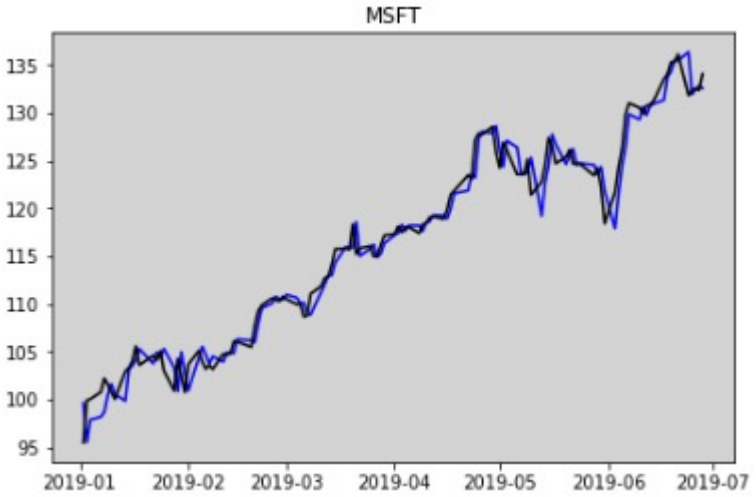


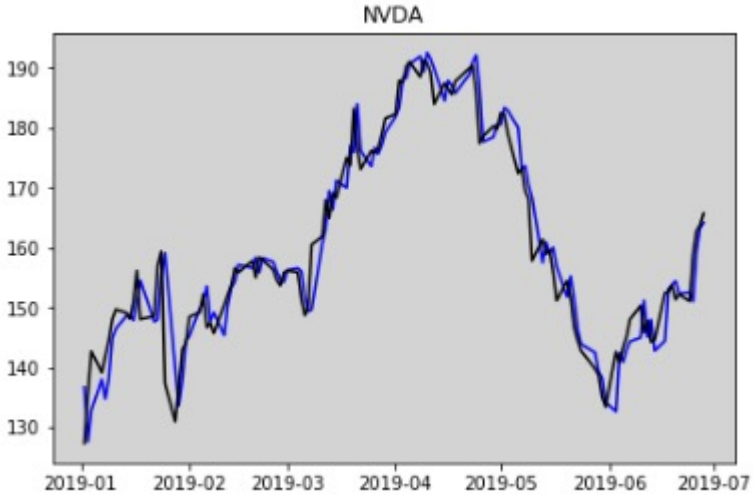


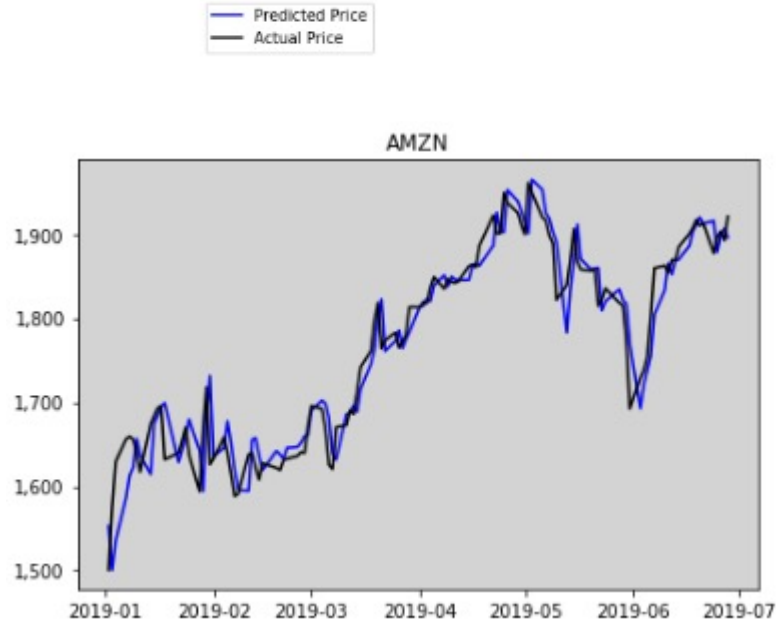


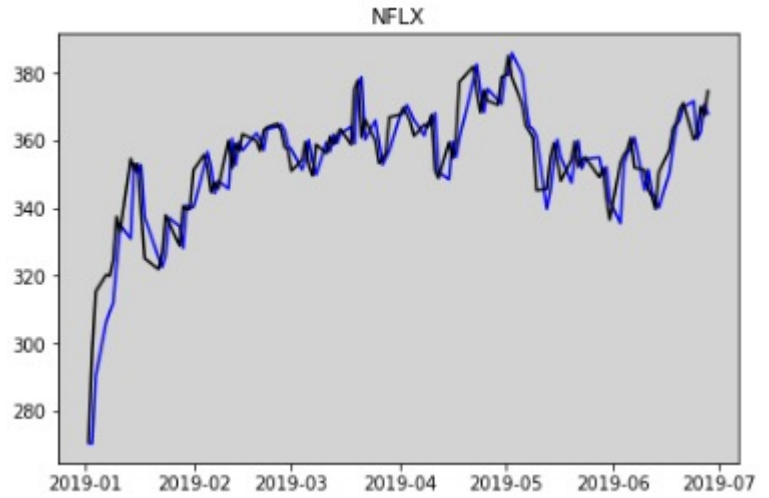
Random Forest



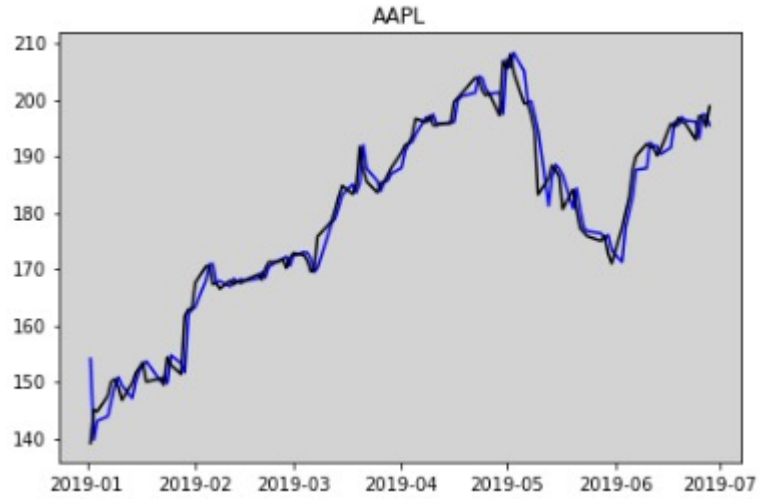


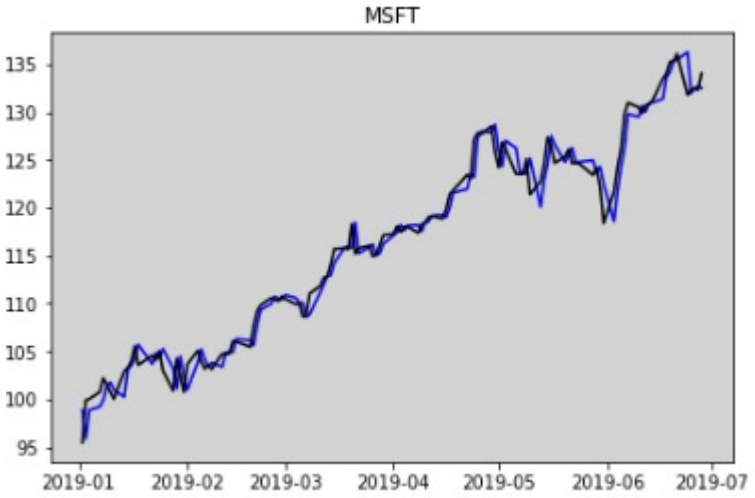


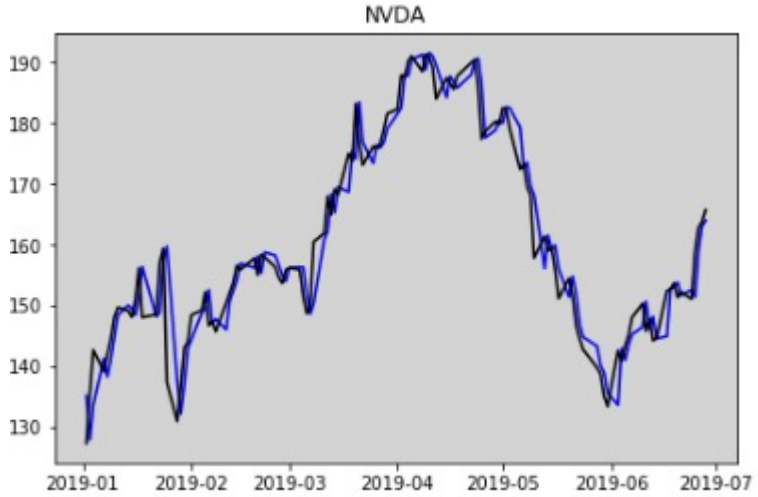


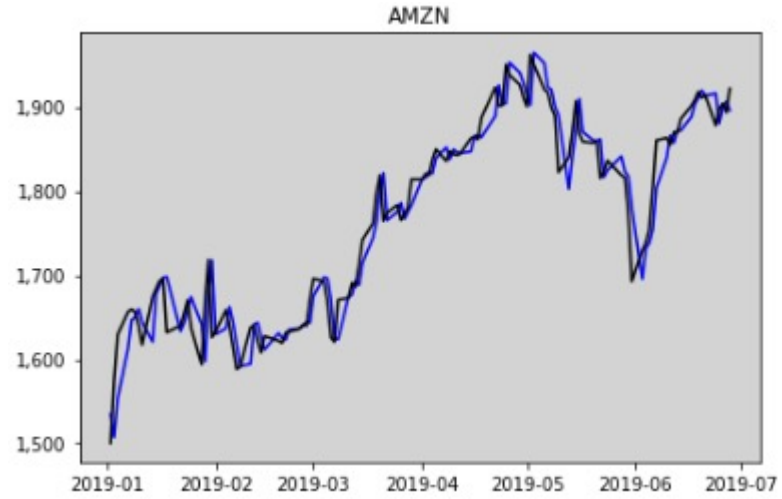


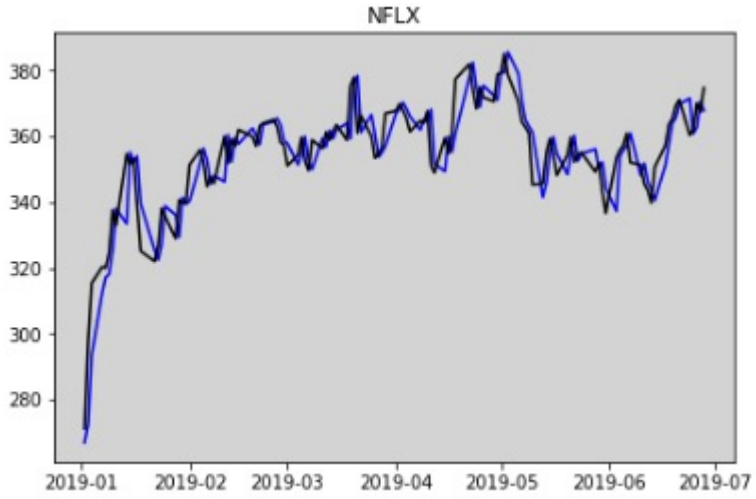
Gradient Boosting



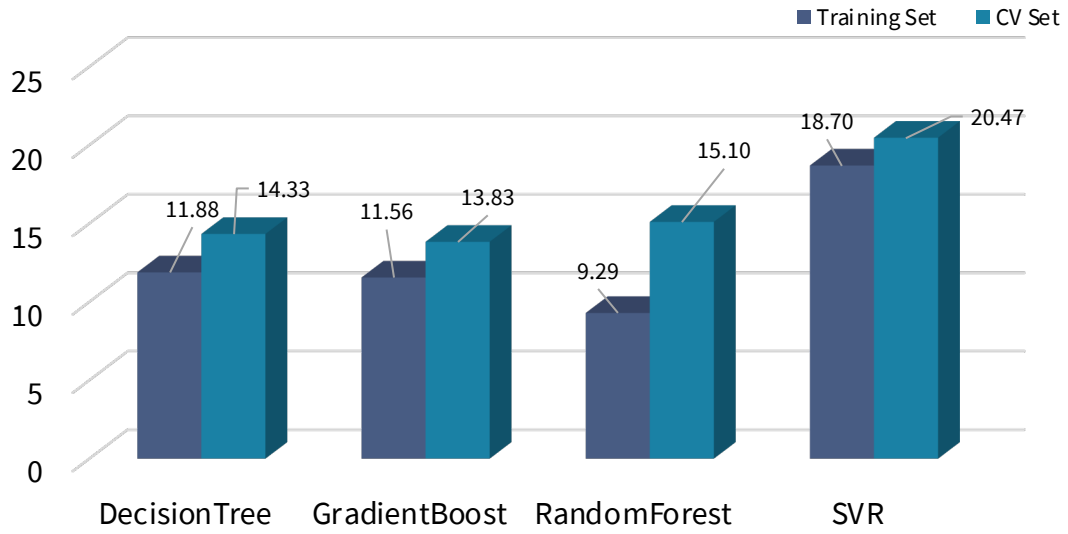


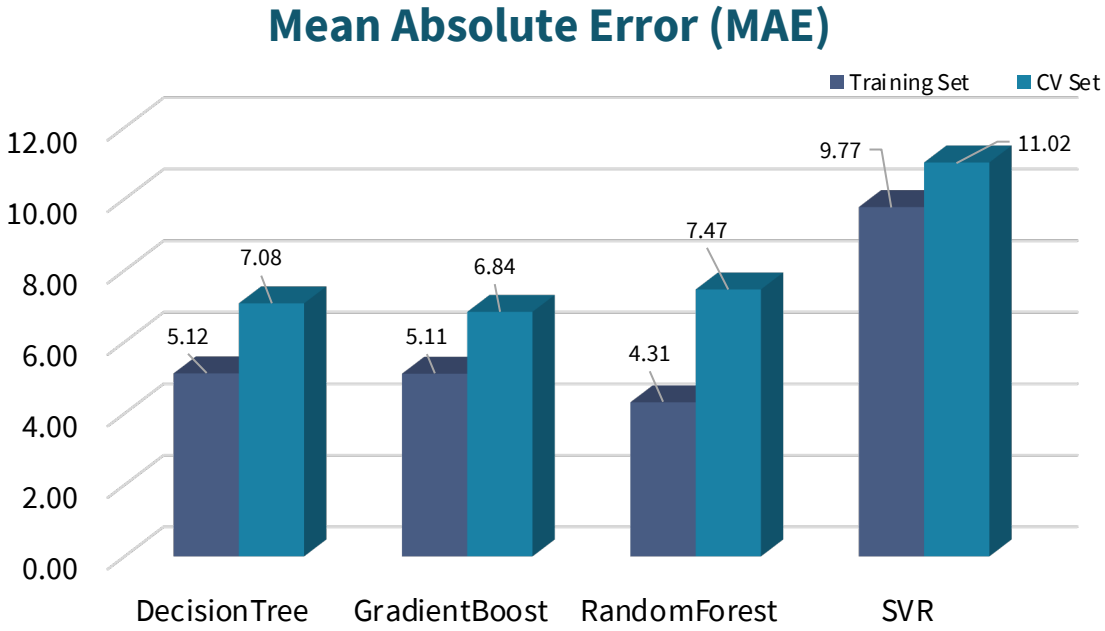


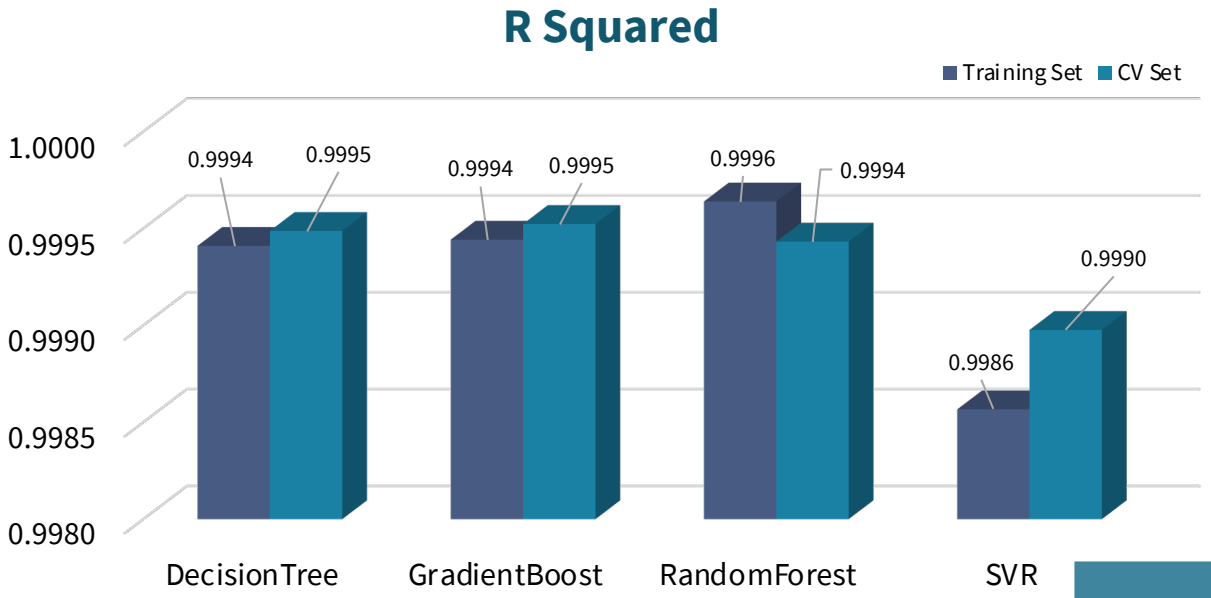




Model Comparison



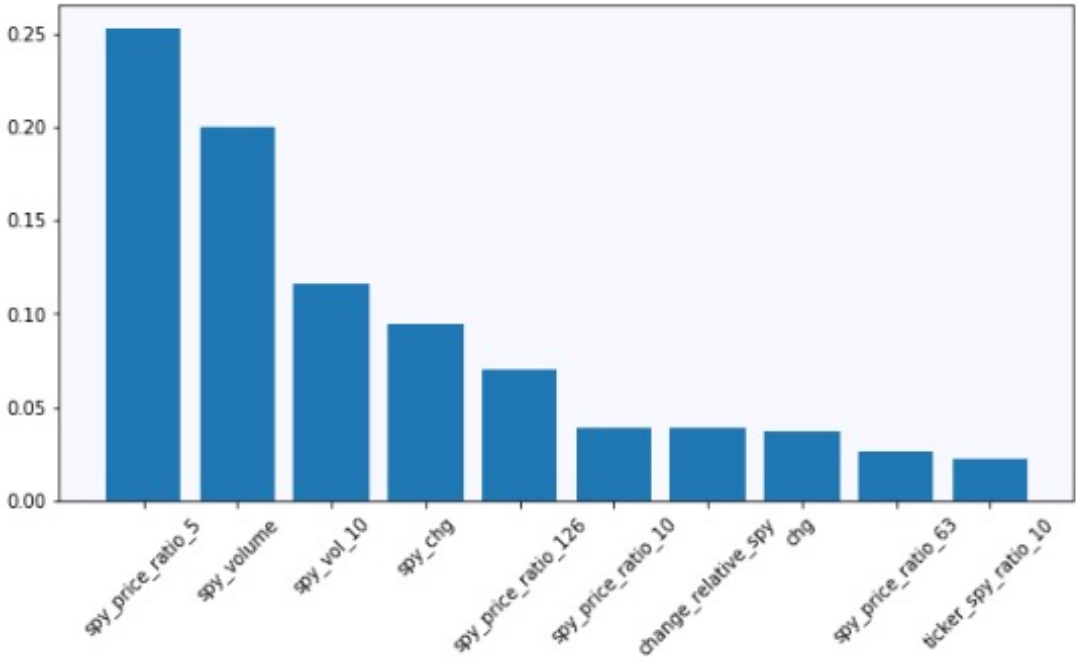


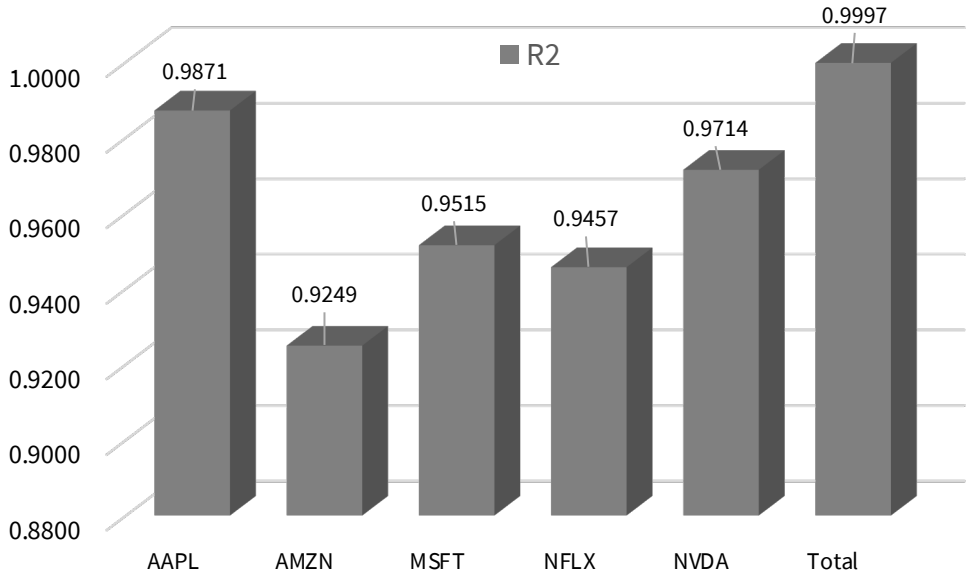


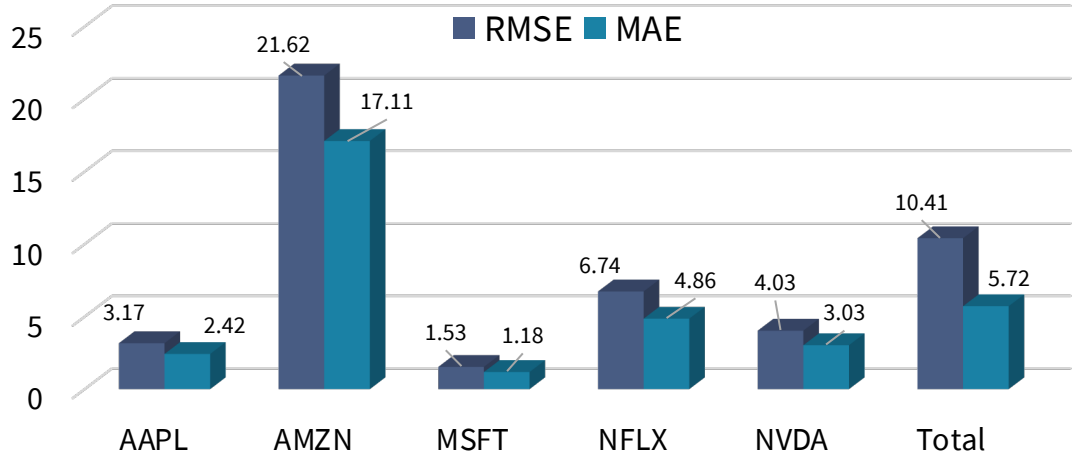
Application of Selected Model

Test Dataset

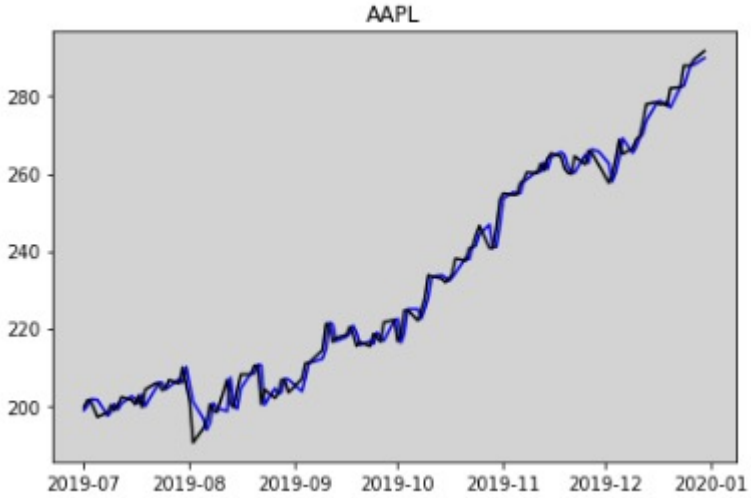
Final Model Results

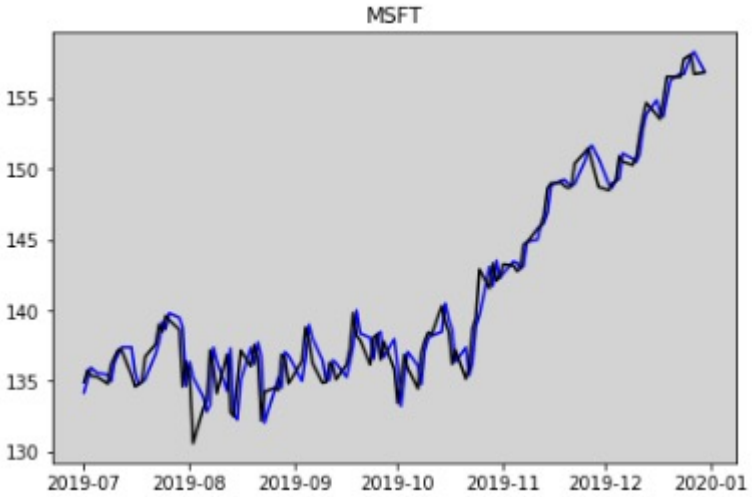


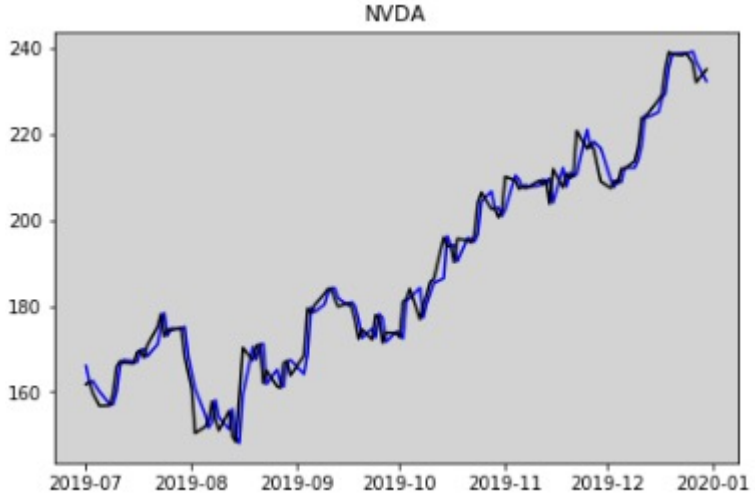


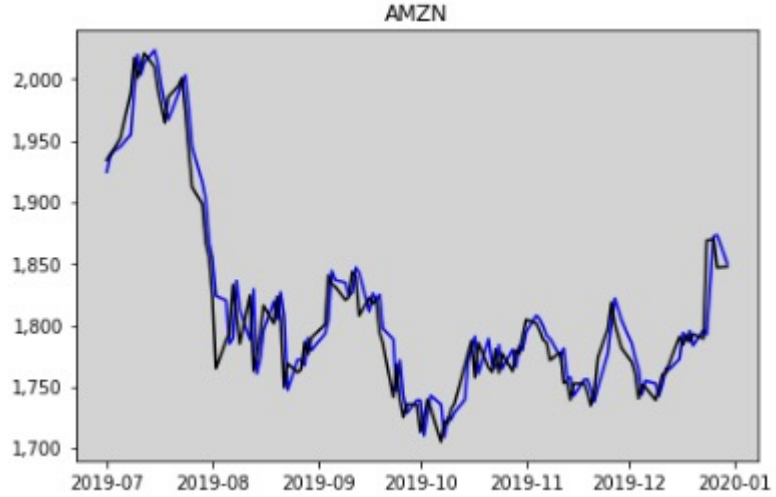


Gradient Boosting – Test Dataset

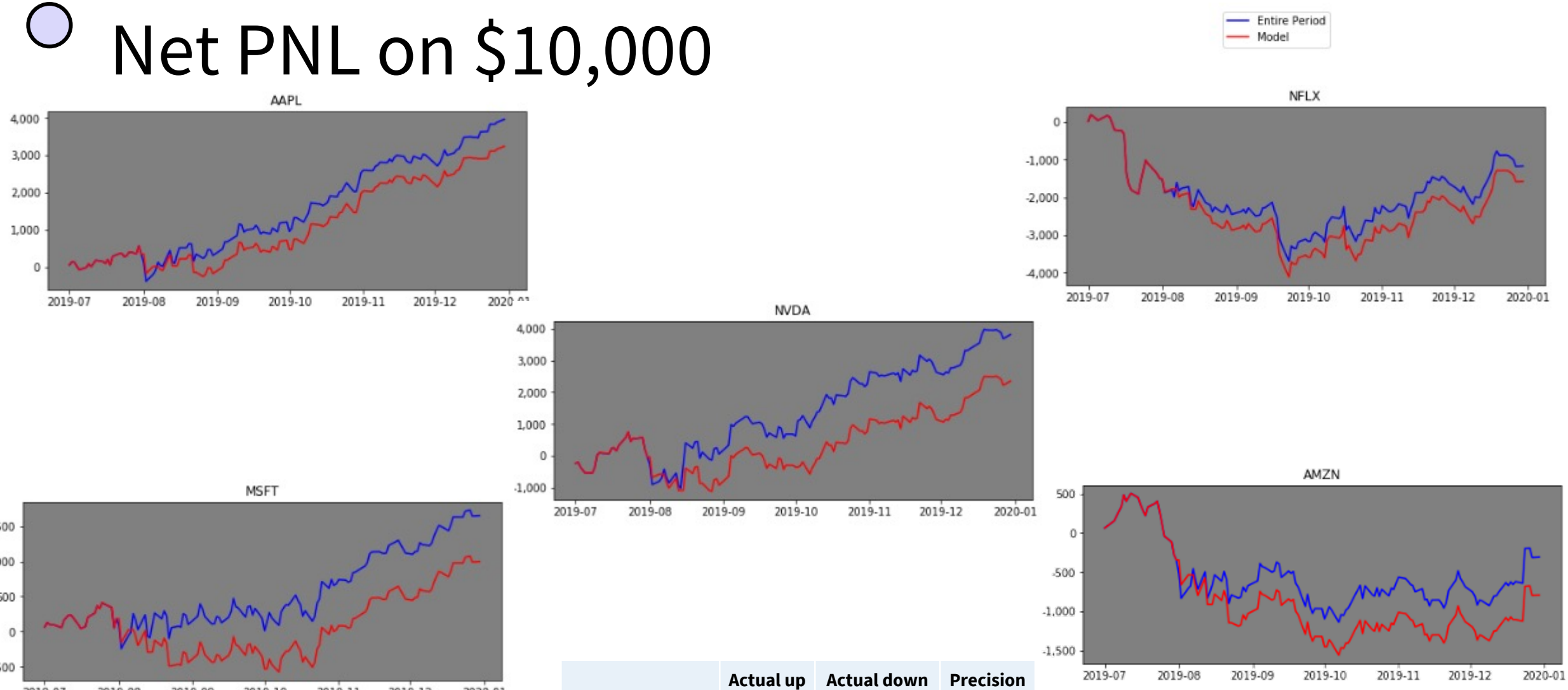






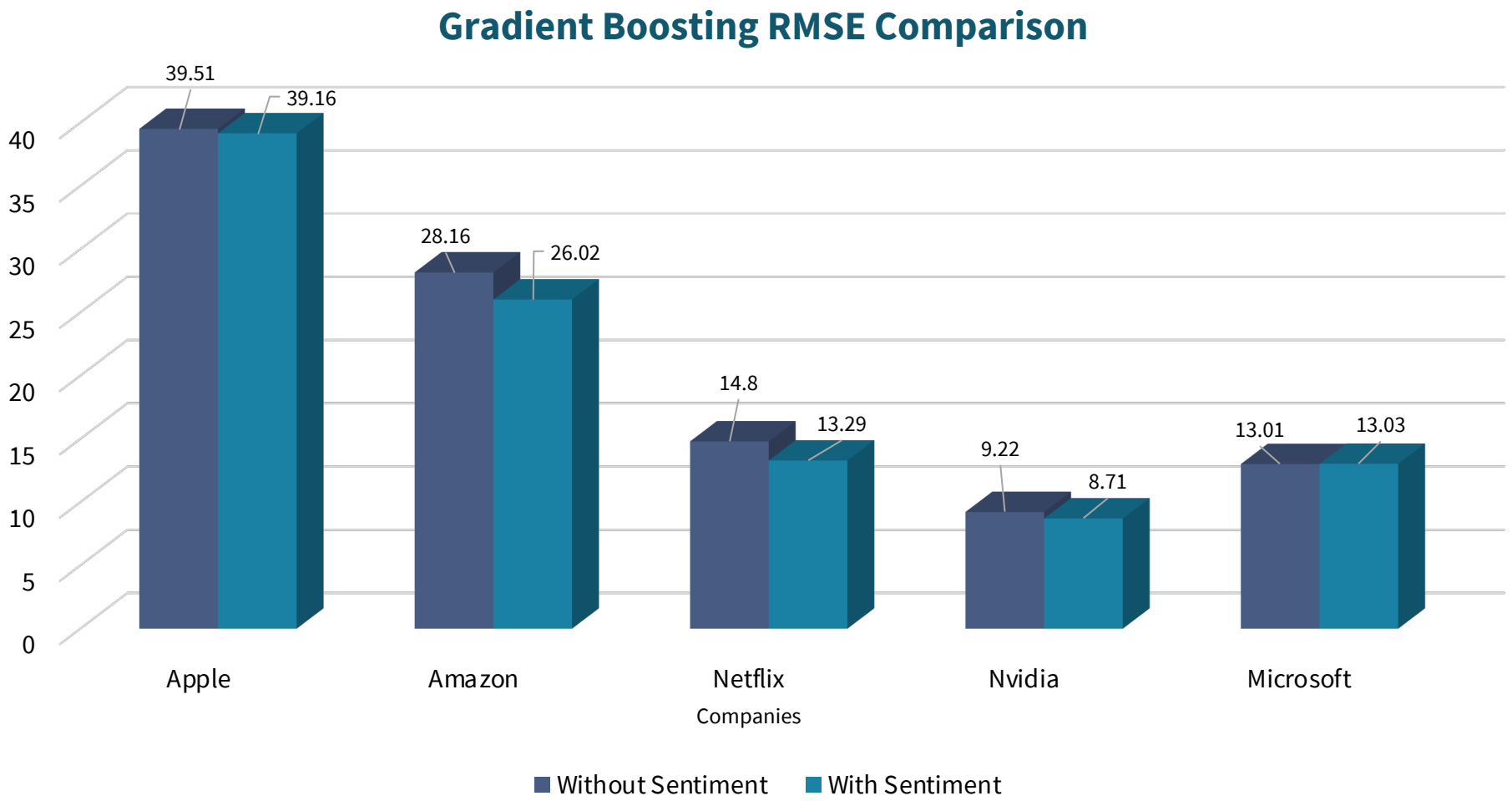


Net PNL on $10,000

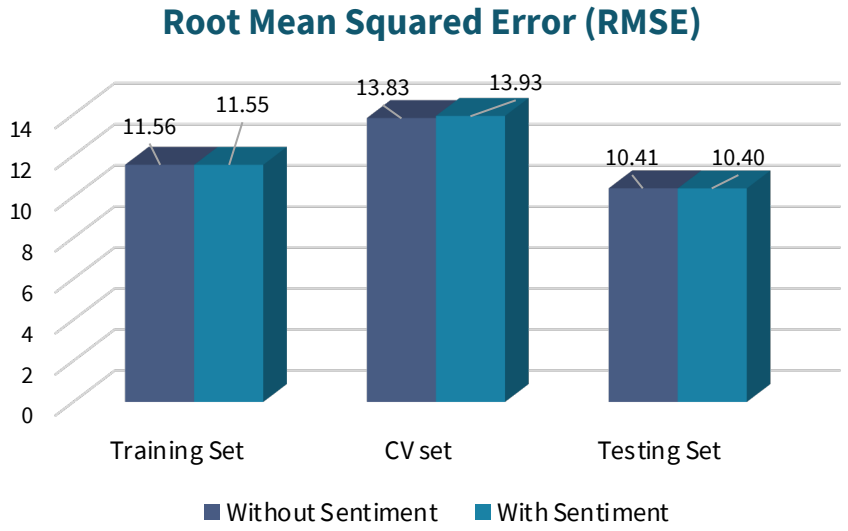


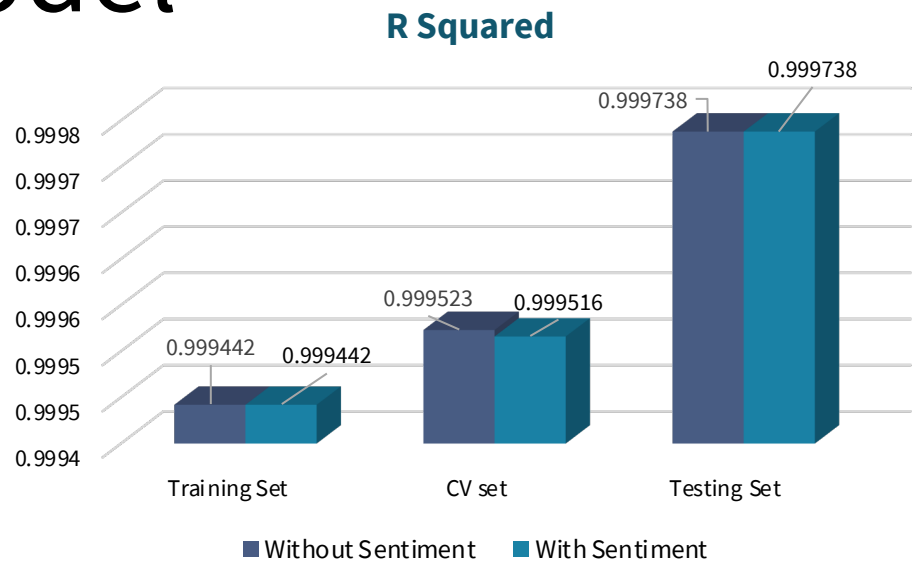
	Actual up	Actual down	Precision
Predicted up	313	250	55.6%
Predicted down	30	127	80.9%

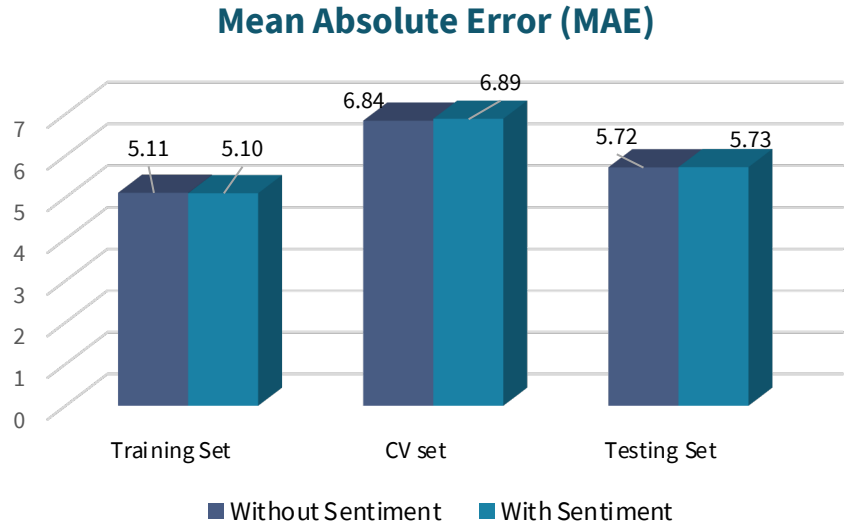
Sentiment Impact – Baseline Model



Sentiment Impact – Final Model





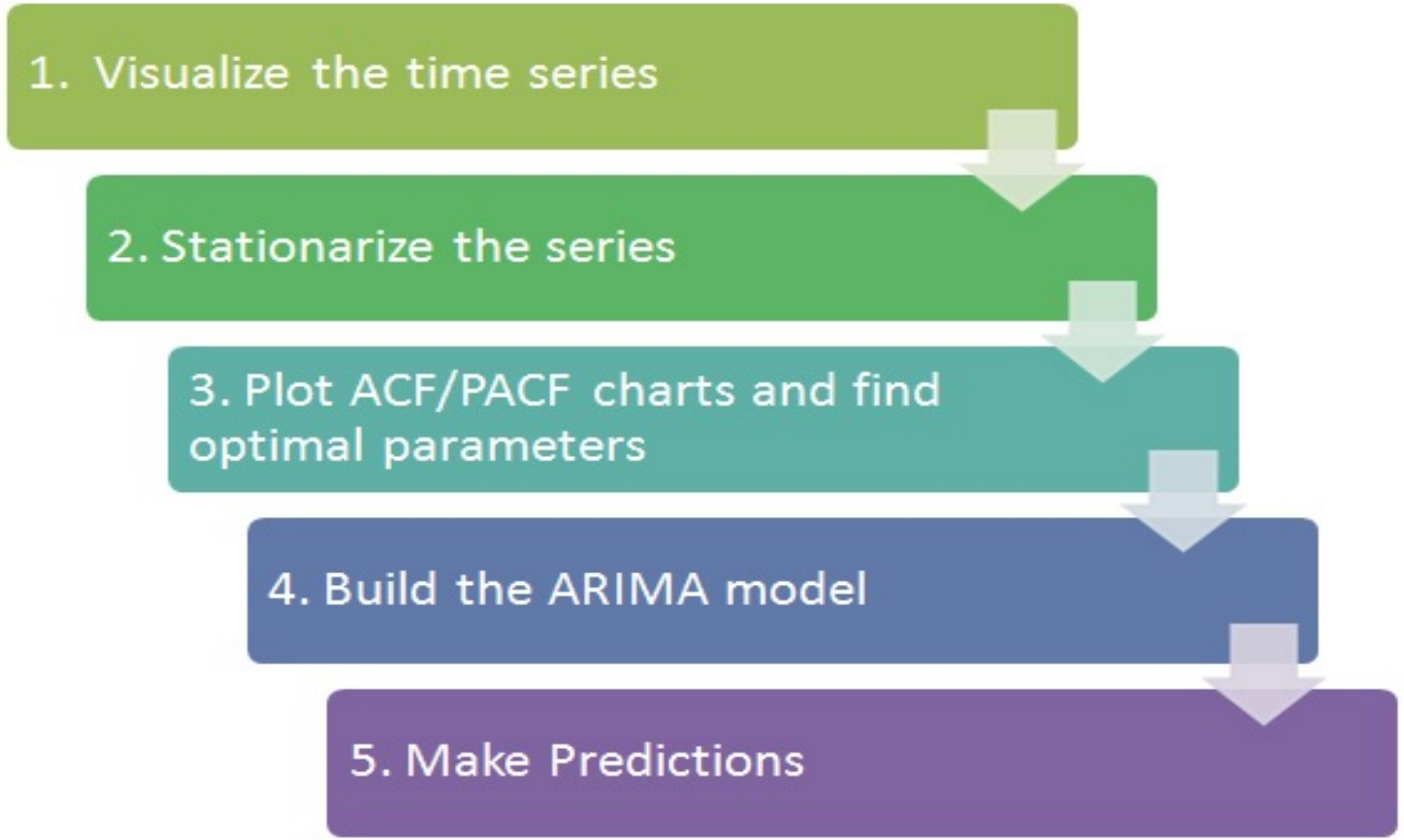


Algorithm used: Gradient Boosting

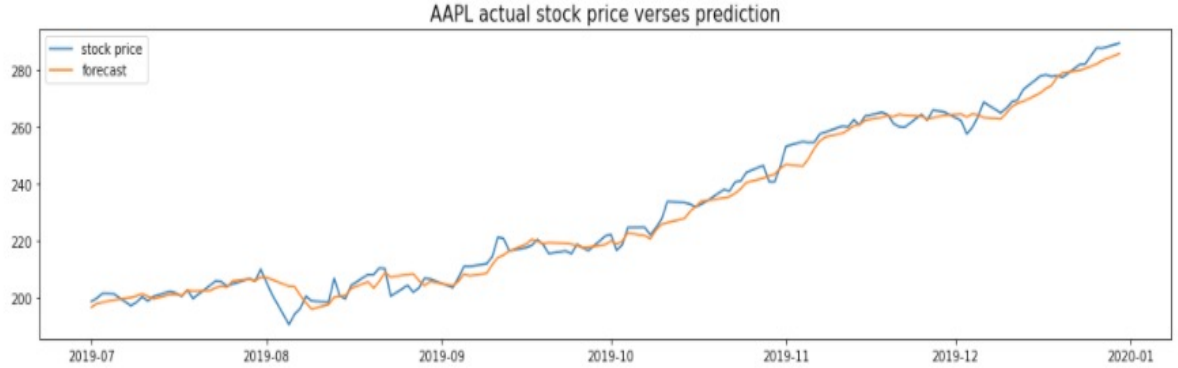
Time Series Data Modeling

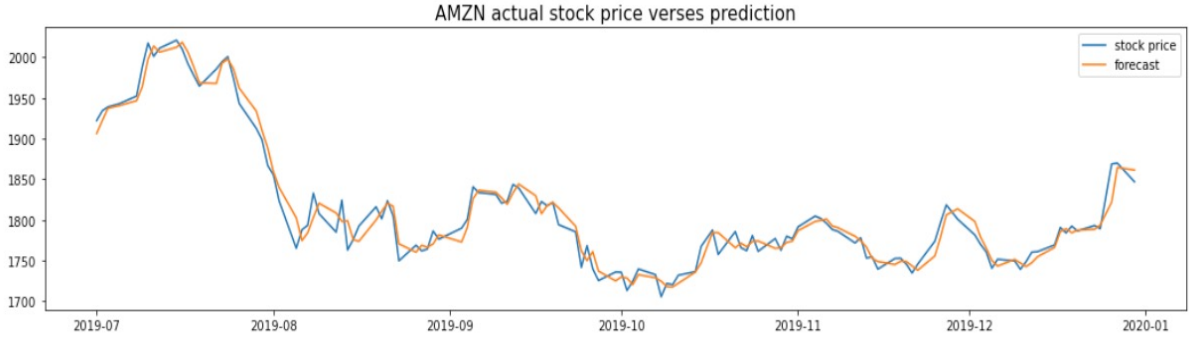
Time Series Exploration

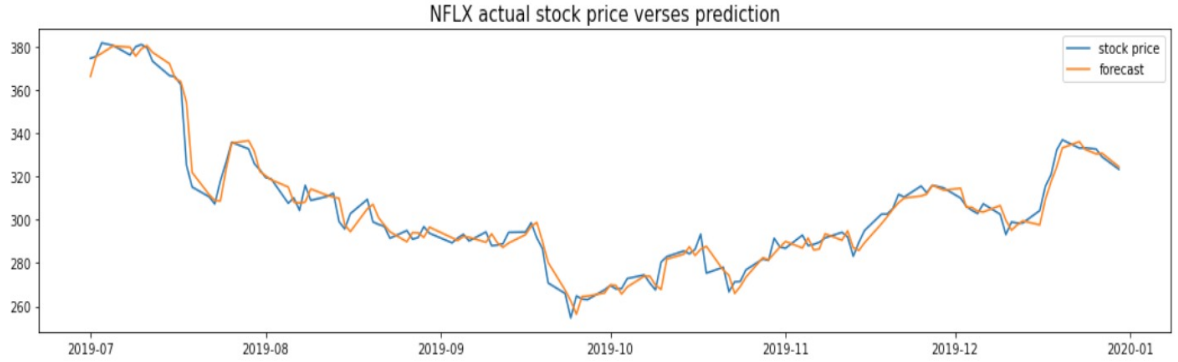
- Method used : ARIMA (AutoRegressive Integrated Moving Average)

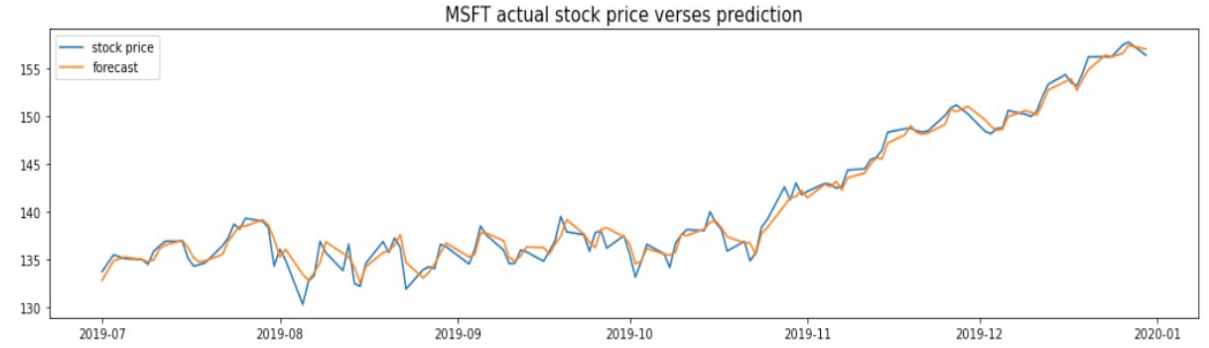


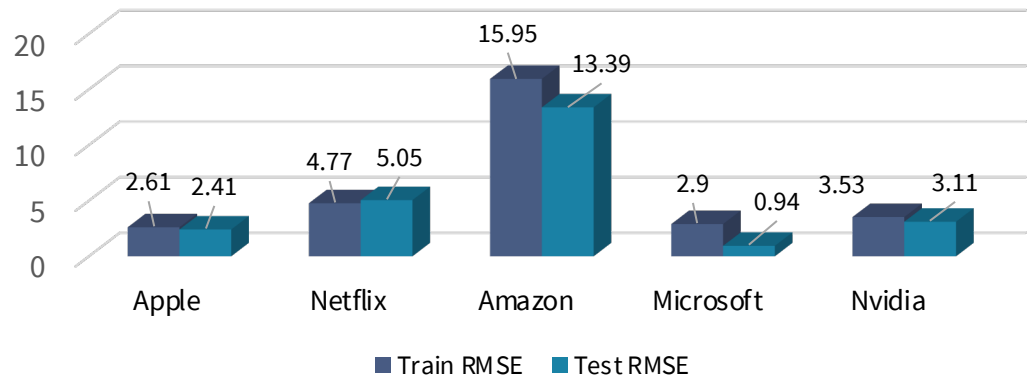
Time Series Prediction Result

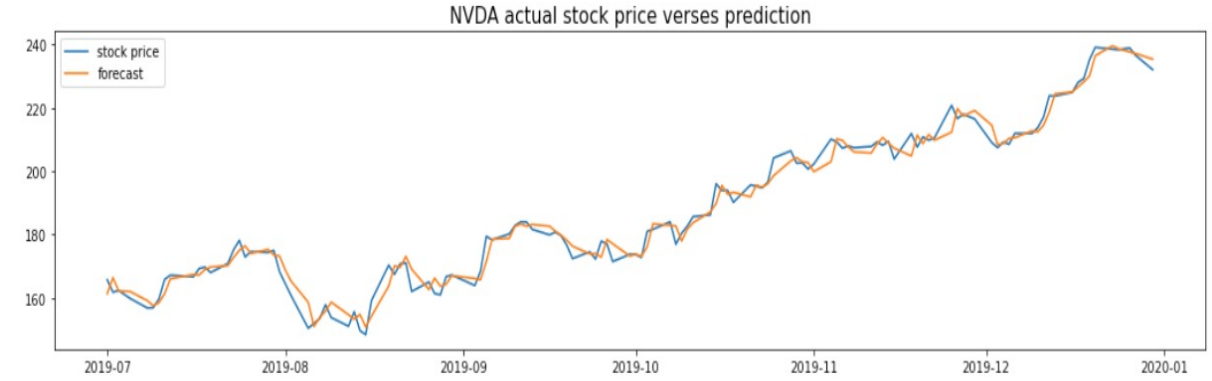




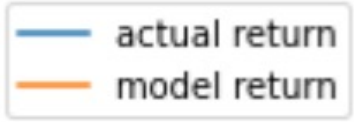


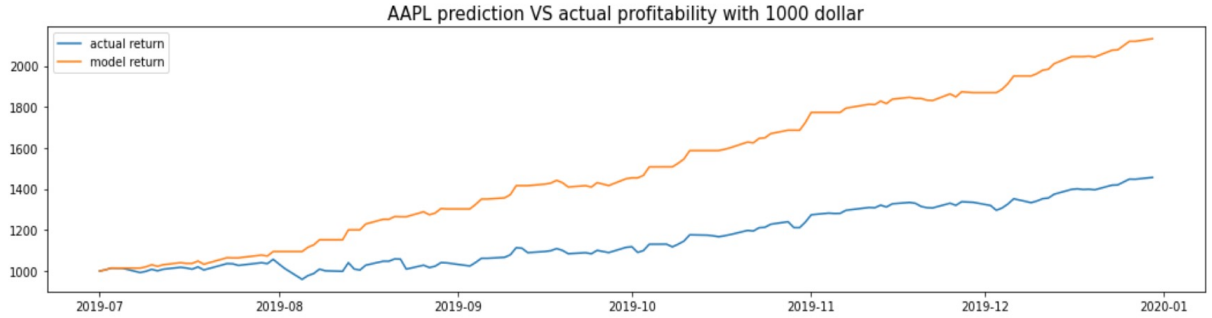


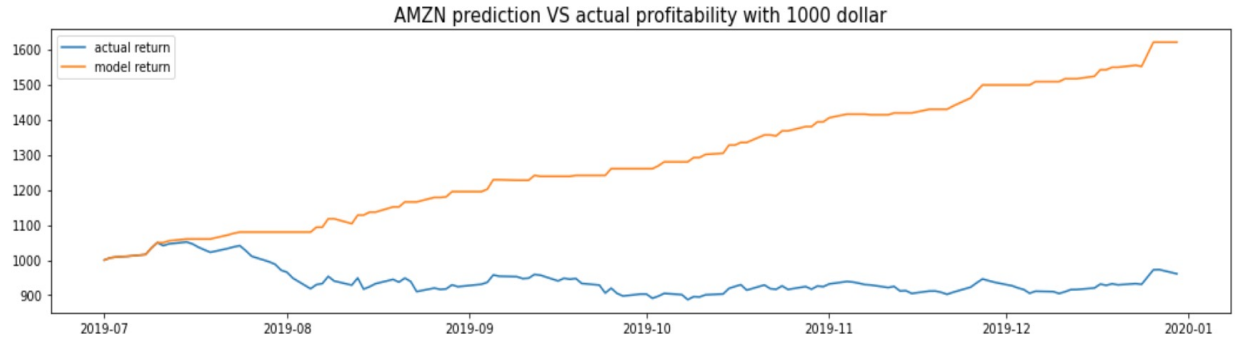


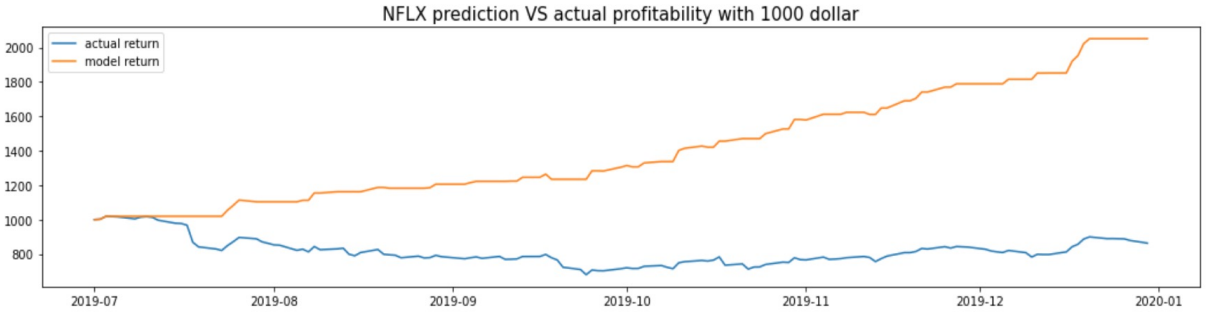


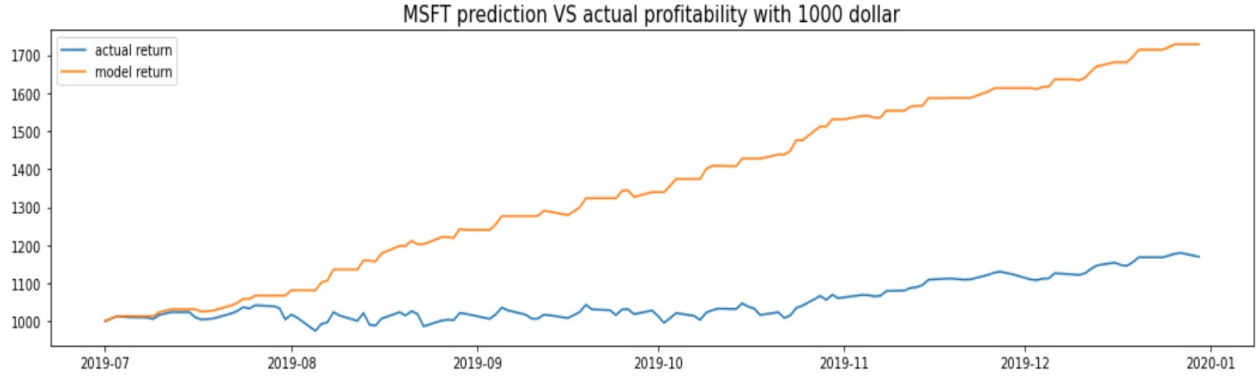
Investment Simulation on $1,000

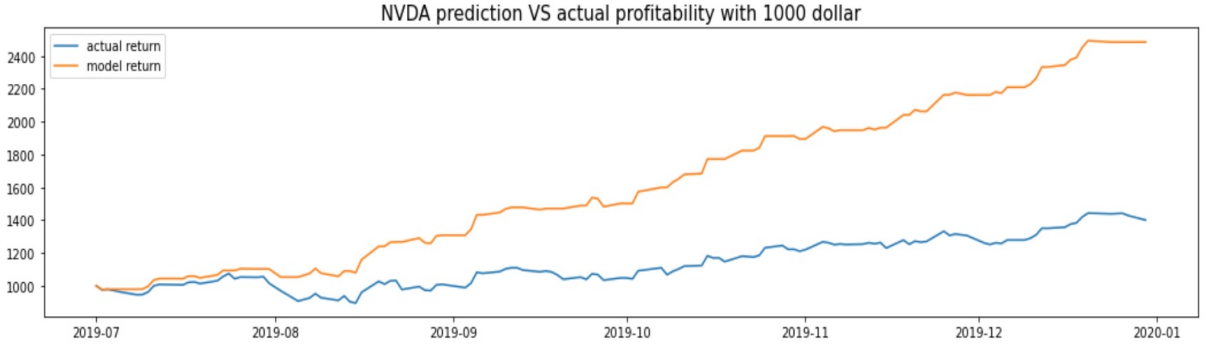












	Actual up	Actual down	Precision
Predicted up	225	126	64.10%
Predicted down	116	163	58.42%

Findings and Conclusion

- Daily stock price changes are largely market driven
- Identifying idiosyncratic stock behavior requires more in-depth data and analysis
- The twitter sentiment of the stock was inconclusive. Filtering to only include “meaningful” tweets could result in a stronger signal
- Machine learning and time series models are useful for stock price prediction

Proposed Future Extensions

Data

- Use more targeted, clean and reliable sentiment data
- Include news data, e.g. breaking news, regulation news
- Obtain the stock data drilled down to hourly time series

Methodology

- To use more complex models like LSTM or Transformers to address the regression problem
- To move towards online machine learning approach

THANKS FOR LISTENING

QUESTIONS

References

- US Treasury Rates
- https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
- https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
- https://www.sciencedirect.com/science/article/pii/S2405918817300247
- https://learndatasci.com/tutorials/python-finance-part-yahoo-finance-api-pandas-matplotlib/
- https://www.youtube.com/watch?v=d4Sn6ny_5LI
- https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/
- https://dataoutpost.wordpress.com/2018/04/03/eda-part-1-full-python-code/amp/
- https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/