Short-term stock market price prediction

DS 5220: Supervised Machine Learning and Learning Theory

Team Number: 4

TA: Ewen

Team Members:

- Farhanur Rahim Ansari (001376195)
- Vidhey Oza (001059237)

Overview

- **Goal:**
  - Predicting prices of stocks listed on NYSE and NASDAQ on a short-term basis using technical indicators.
  - Time Series Analysis
  - Regression Problem
- **Dataset:**
  - Stocks: Apple (AAPL), Amazon (AMZN), Microsoft (MSFT), Tesla (TSLA), and Walmart (WMT)
  - Minute-wise records from 1-1-2018 to 11-11-2019
  - Total 181578 record per stock
  - 9 features – Ticker, Per, Date, Time, open, high, low, close, and volume

<TICKER>	<PER>	<DATE>	<TIME>	<OPEN>	<HIGH>	<LOW>	<CLOSE>	<VOL>
US1.WMT	1	20180102	173100	99.3	99.75	99.3	99.75	3587
US1.WMT	1	20180102	173200	99.64	99.71	99.52	99.52	3306
US1.WMT	1	20180102	173300	99.56	99.68	99.56	99.67	4376
US1.WMT	1	20180102	173400	99.64	99.67	99.57	99.65	1594
US1.WMT	1	20180102	173500	99.62	99.64	99.57	99.57	1400
US1.WMT	1	20180102	173600	99.59	99.59	99.42	99.43	1400
US1.WMT	1	20180102	173700	99.46	99.47	99.39	99.39	4028
US1.WMT	1	20180102	173800	99.38	99.41	99.35	99.4	1991
US1.WMT	1	20180102	173900	99.41	99.43	99.32	99.36	2463
US1.WMT	1	20180102	174000	99.38	99.47	99.35	99.4	2000
US1.WMT	1	20180102	174100	99.42	99.45	99.42	99.42	2500
US1.WMT	1	20180102	174200	99.36	99.36	99.36	99.36	200
US1.WMT	1	20180102	174300	99.34	99.34	99.29	99.33	1000
US1.WMT	1	20180102	174400	99.3	99.31	99.29	99.31	950
US1.WMT	1	20180102	174500	99.33	99.35	99.32	99.35	400
US1.WMT	1	20180102	174600	99.34	99.34	99.21	99.22	1796
US1.WMT	1	20180102	174700	99.24	99.27	99.24	99.24	600
US1.WMT	1	20180102	174800	99.26	99.28	99.26	99.28	200
US1.WMT	1	20180102	174900	99.27	99.29	99.26	99.28	1058
US1.WMT	1	20180102	175000	99.27	99.27	99.24	99.24	200
US1.WMT	1	20180102	175100	99.27	99.27	99.14	99.22	3885
US1.WMT	1	20180102	175200	99.21	99.27	99.21	99.27	2000

Approach

Data Extraction & Preliminary Analysis

Feature Engineering

Data Preprocessing

Feature selection

Model Learning

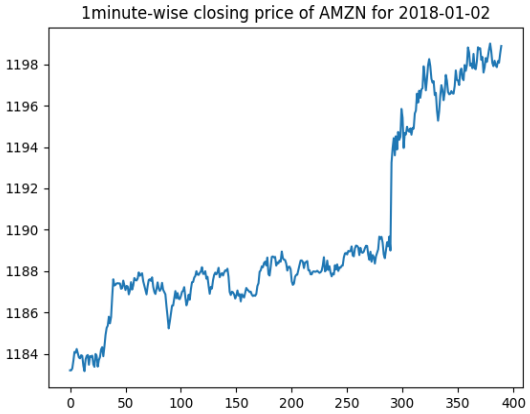
Model Evaluation

AAPL

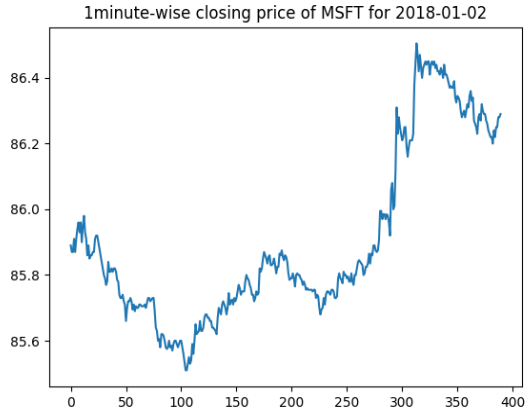
TSLA

WMT

AMZN

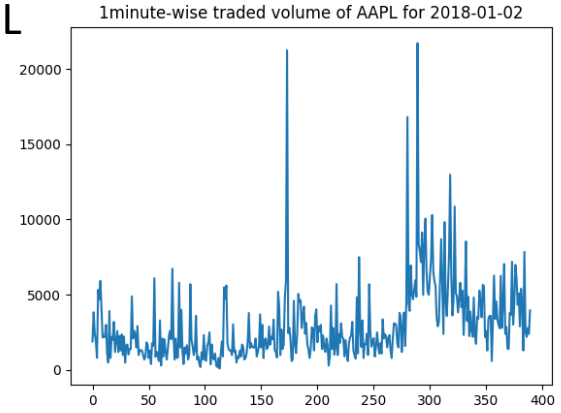


MSFT

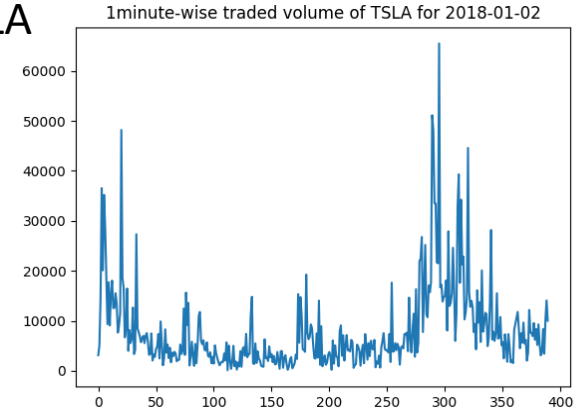


Preliminary Data Analysis: Closing Price

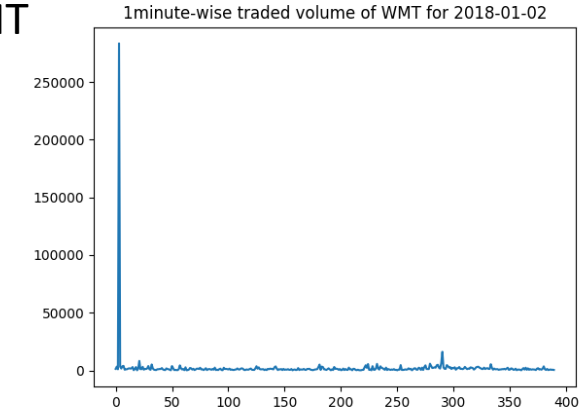
AAPL



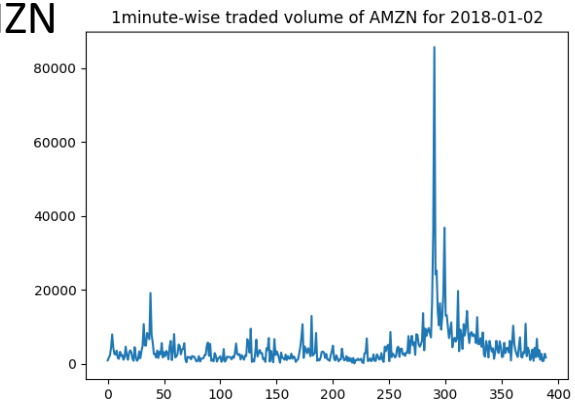
TSLA



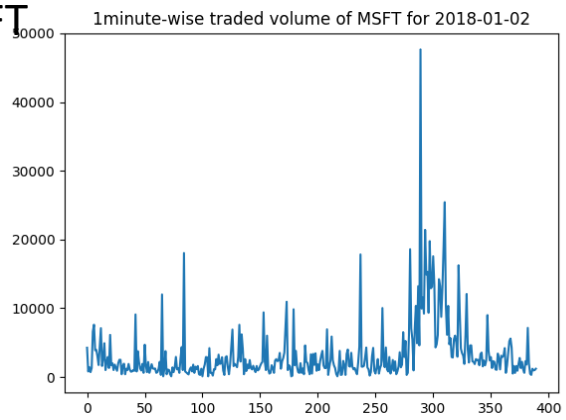
WMT



AMZN



MSFT

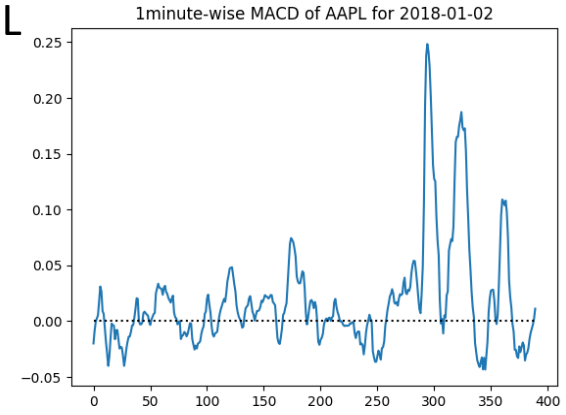


Preliminary Data Analysis: Volume

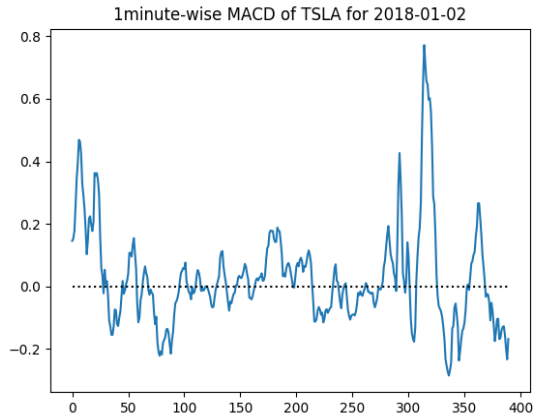
Feature Engineering

- 5 main features - open, high, low, close, volume
- Engineered features are technical indicators from 6 main categories:
  - Smoothing (moving averages)
  - Momentum (shows the continuous 'momentum' of stock price)
  - Volume (analysis from volume of stock traded)
  - Overbought/oversold signals (generate signals when stock is overbought or oversold)
  - Volatility (analysis from variability of stock in a given period)
  - Trends (analysis of general trend of price movement)
- 22 engineered features - MACD, RSI, William's %R, Stochastic %K and %D, MFI, ROC, SMA, WMA, EMA, HMA, CCI, ADL, CMF, OBV, EMV, ATR, Mass Index, Ichimoku clouds (made from 2 indicators), Aroon Index, ADX

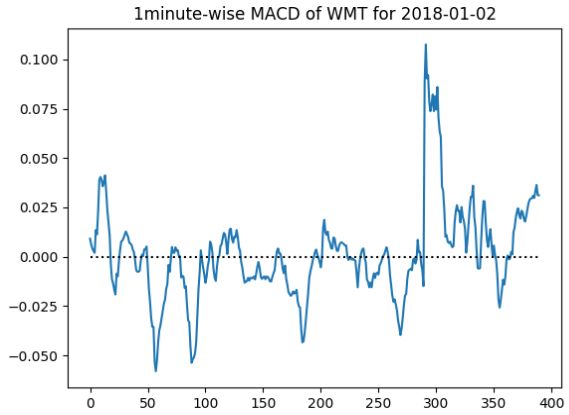
AAPL



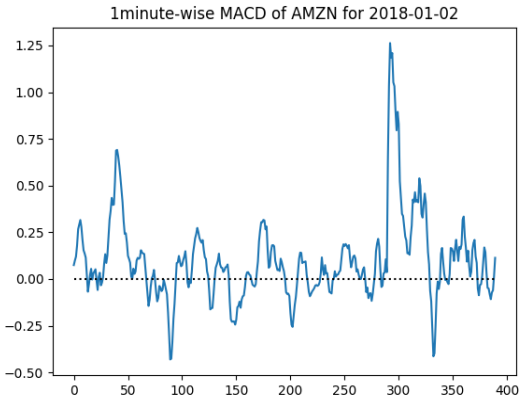
TSLA



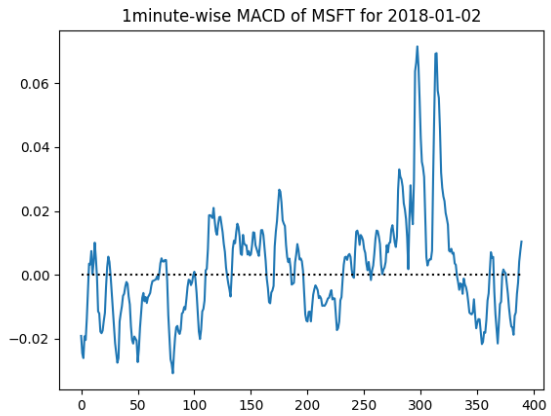
WMT



AMZN



MSFT



Technical Indicator Example: MACD

Data Preprocessing

- **Number of Records:**
  - 1 minute: 50000 records
  - 5 minute: 20000 records
  - 10 minute: 10000 records
- **Data Cleaning:** Unwanted features like ticker, per, date, and timestamp were removed.
- **Data Scaling:** Data is normalized using Z-score normalization.
- **Handling Null Values:** 100 records containing NaN values from both the top and bottom of the dataset were removed.
- **Feature Selection:**
  - Selected 16 features out of the 22 engineered features.
  - Feature selection was done by understanding the meaning behind the features and removing the noisy ones.
  - This was also verified by analyzing the correlation matrix.
- **Train-Test Split:** Training set (75%) and Testing set (25%)

Model Learning

- Response variable: Closing Price (close)
- Regression algorithms used:
    1. Support Vector Regressor (SVR)
    2. Random Forest Regressor (100 estimators)
    3. AdaBoost Regressor (100 estimators)
    4. XGBoost Regressor
    5. Artificial Neural Network (ANN)
        - 1 Input layer, 2 Fully Connected Hidden Layers, 1 output layer
        - Activation functions:- Hidden layer: Relu, Output Layer: linear
        - Hidden nodes:- Layer 1: 100 units, Layer 2: 50 units
        - Total Parameters: 6,801
        - Epochs: 50

Hyperparameter Tuning

- **Dataset:** indicator window
  - Denotes no. of data points (mins in this case) algorithm looks back to generate indicator values
  - We checked for 5-min, 10-min, 15-min, 20-min on one stock; best results for 10-min window
- **Regression Models:**
  - SVR: C = {1, 5, 10, 100} best at C = 5; γ = auto supported by sklearn package
  - RandomForest: n_estimators = {25, 50, 100, 150} best at 100
  - XGBoost: n_estimators = {25, 50, 100, 150} best at 100
  - Adaboost: n_estimators = {25, 50, 100, 150} best at 100
  - ANN:
    - Varied the number of hidden layers, no advantage for > 2 hidden layers
    - Varied the activation function {sigmoid, tanh, ReLU}, ReLU worked the best for hidden layers
    - Linear activation at output since other functions have an upper or lower bound, not suitable for price prediction

Results

Time Interval = 1 min

Stocks	SVR			Random Forest			AdaBoost			XGBoost			ANN		
	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE
AAPL	**0.40**	**0.16**	**0.17**	1.17	1.39	0.31	2.27	5.18	0.75	1.25	1.57	0.34	3.81	0.22	1.62
AMZN	**10.39**	**108.1**	**0.52**	78.61	6180	3.89	101.4	10294	5.30	79.35	6296	3.93	63.95	1585	3.26
MSFT	**0.30**	**0.09**	**0.22**	2.00	4.01	1.54	2.49	6.22	2.05	2.09	4.37	1.63	1.88	0.91	1.5
TSLA	1.24	1.55	0.31	**1.70**	**2.90**	**0.14**	3.49	12.21	0.66	1.83	3.35	0.17	42.02	1.26	10.73
WMT	0.37	0.14	0.35	**0.04**	**0.001**	**0.02**	0.70	0.50	0.70	0.06	0.004	0.05	1.67	0.008	1.57

Results

Time Interval = 5 min

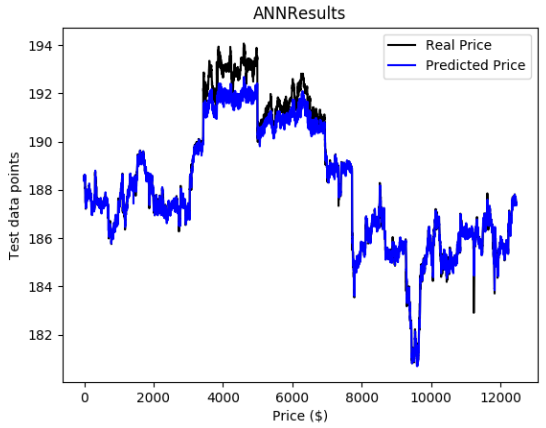
Stocks	SVR			Random Forest			AdaBoost			XGBoost			ANN		
	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE
AAPL	1.32	1.74	0.52	1.75	3.06	0.46	2.20	4.84	0.74	**1.72**	**2.96**	**0.45**	36.4	0.75	16.05
AMZN	10.93	119	0.5	**6.74**	**45.5**	**0.25**	16.38	268	0.78	6.98	48.7	0.27	164.3	32.9	8.12
MSFT	0.58	0.34	0.41	0.27	0.07	0.15	0.58	0.34	0.43	**0.18**	**0.03**	**0.12**	5.4	0.07	4.07
TSLA	1.42	2.01	0.33	1.01	1.02	0.17	2.79	7.82	0.68	**0.79**	**0.63**	**0.16**	50.06	0.89	12.47
WMT	0.29	0.08	0.23	0.17	0.02	0.09	0.45	0.21	0.37	**0.12**	**0.01**	**0.09**	5.98	0.01	4.94

Results

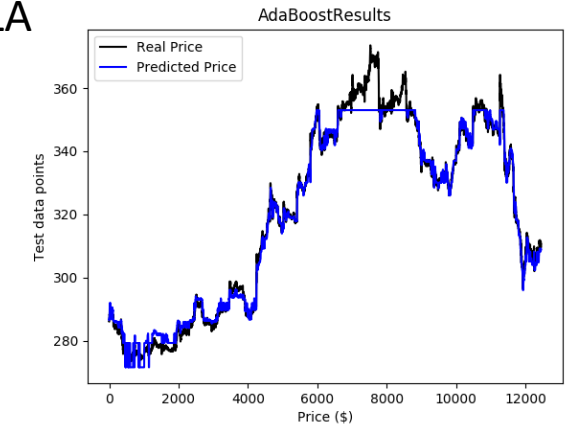
Time Interval = 10 min

Stocks	SVR			Random Forest			AdaBoost			XGBoost			ANN		
	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE
AAPL	1.67	2.81	0.67	2.28	5.20	0.66	2.42	5.90	0.82	**2.12**	**4.51**	**0.59**	36.18	1.88	15.89
AMZN	13.47	181	0.64	9.69	93.9	0.38	16.5	274	0.75	**8.22**	**67.5**	**0.32**	168	60.8	8.3
MSFT	0.73	0.54	0.55	0.36	0.13	0.21	0.64	0.41	0.45	**0.29**	**0.08**	**0.18**	5.43	0.12	4.07
TSLA	1.65	2.74	0.37	1.32	1.76	0.25	3.74	14.0	0.91	**0.94**	**0.89**	**0.20**	51.37	6.46	12.9
WMT	0.38	0.15	0.29	0.23	0.05	0.15	0.54	0.29	0.45	**0.17**	**0.03**	**0.13**	6.00	0.03	4.96

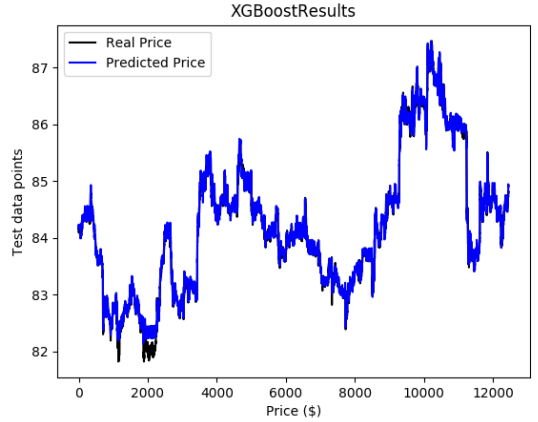
AAPL



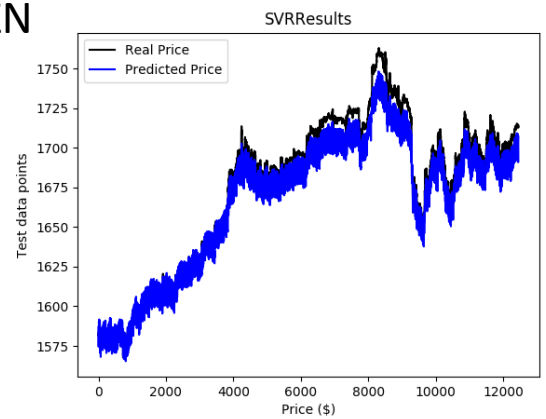
TSLA



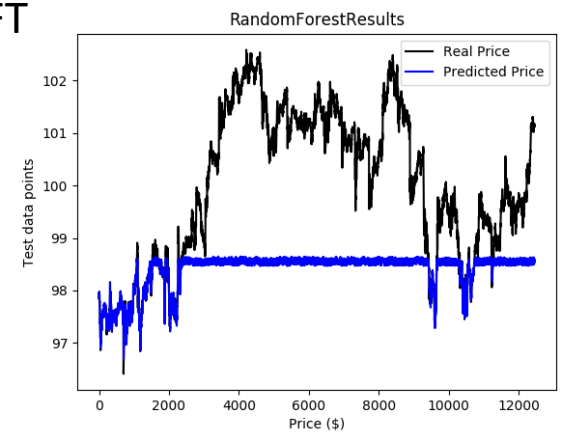
WMT



AMZN

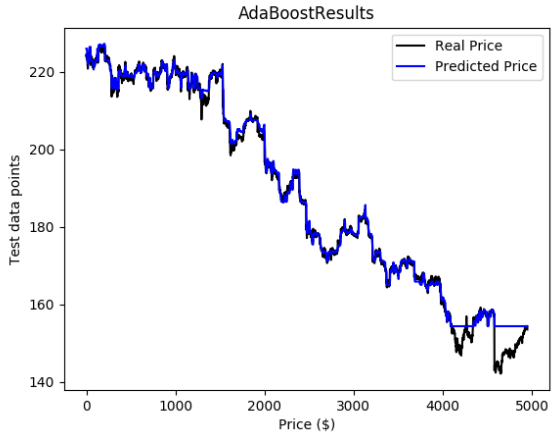


MSFT

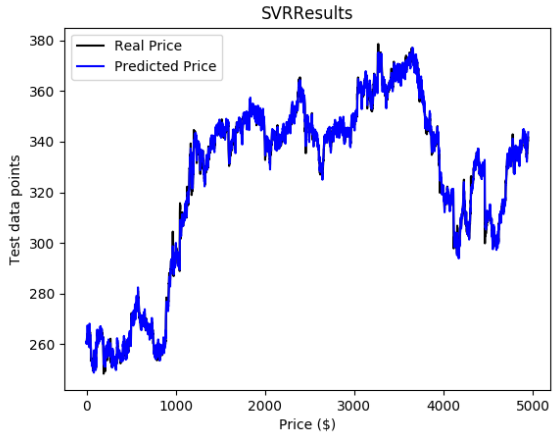


Evaluation of Results: 1-min interval

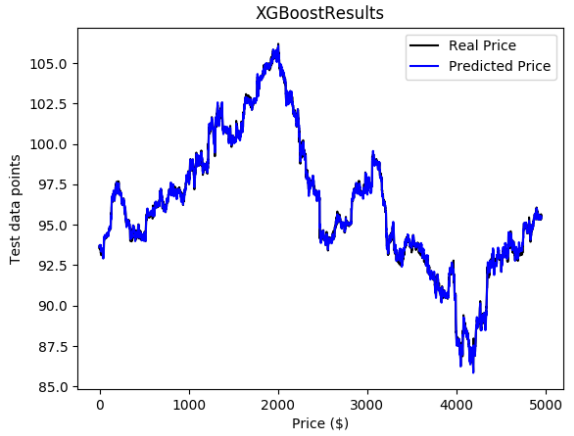
AAPL



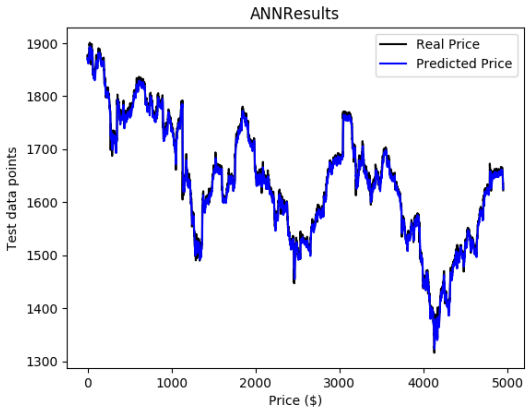
TSLA



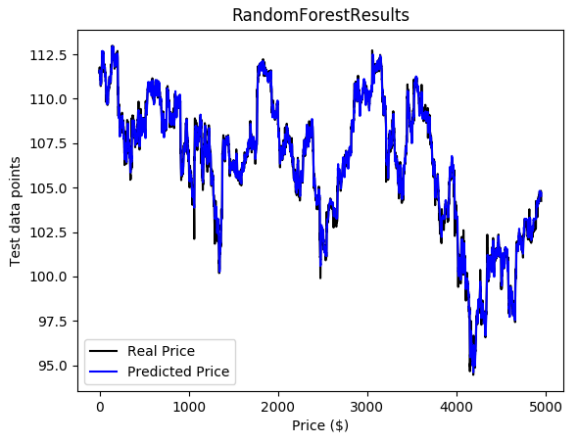
WMT



AMZN

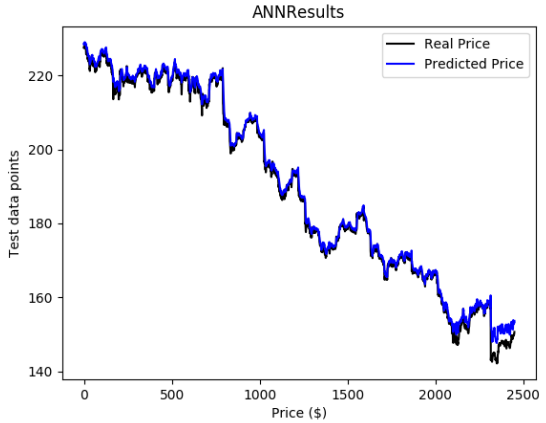


MSFT

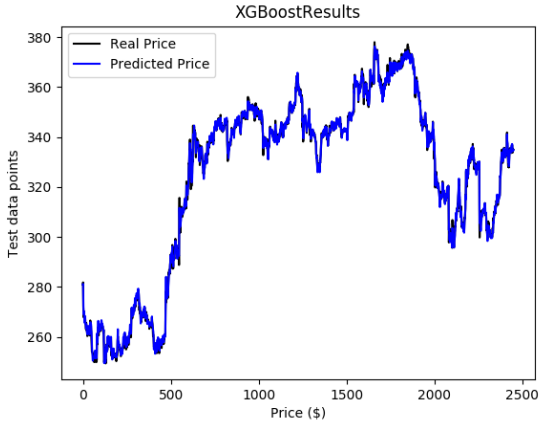


Evaluation of Results: 5-min interval

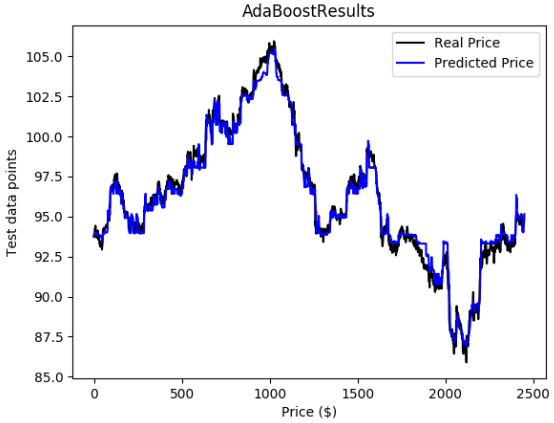
AAPL



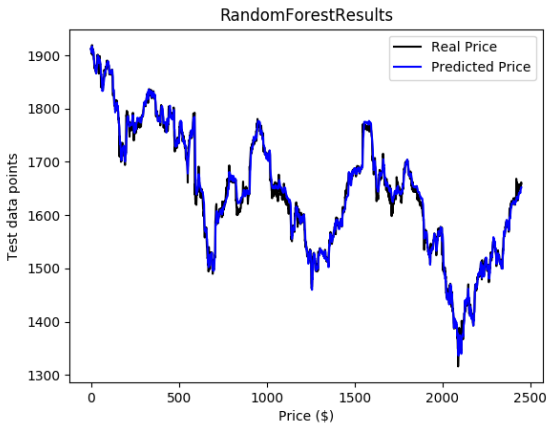
TSLA



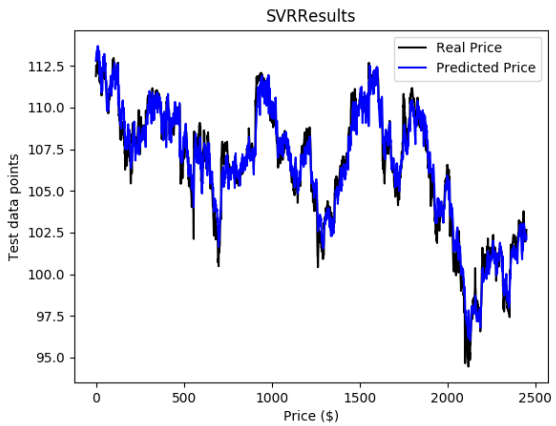
WMT



AMZN



MSFT



Evaluation of Results: 10-min interval

Conclusion

- ANN reports best MSE scores in general, but doesn't perform well on other metrics
  - Sometimes falls in local minima problem, so we don’t get desirable results all the time
- XGBoost performs the best among the three ensemble models, AdaBoost performs the worst.
  - In general, tree-based ensembles don't work well
  - Since stock prices are generally increasing and testing data came chronologically after training data, tree-based predictions quickly lead to highest price possible, but cannot go further than that (due to average as output at leaf)
  - XGBoost tried for linear as well as tree-based models, best results on linear
- SVR performs second best after XGBoost
  - Above other ensemble models
  - Validates the powerful math behind the model
- Interval wise, 1-min interval is the most volatile, so doesn’t give as consistent results as 10-min data

Challenges Faced

Data Extraction:

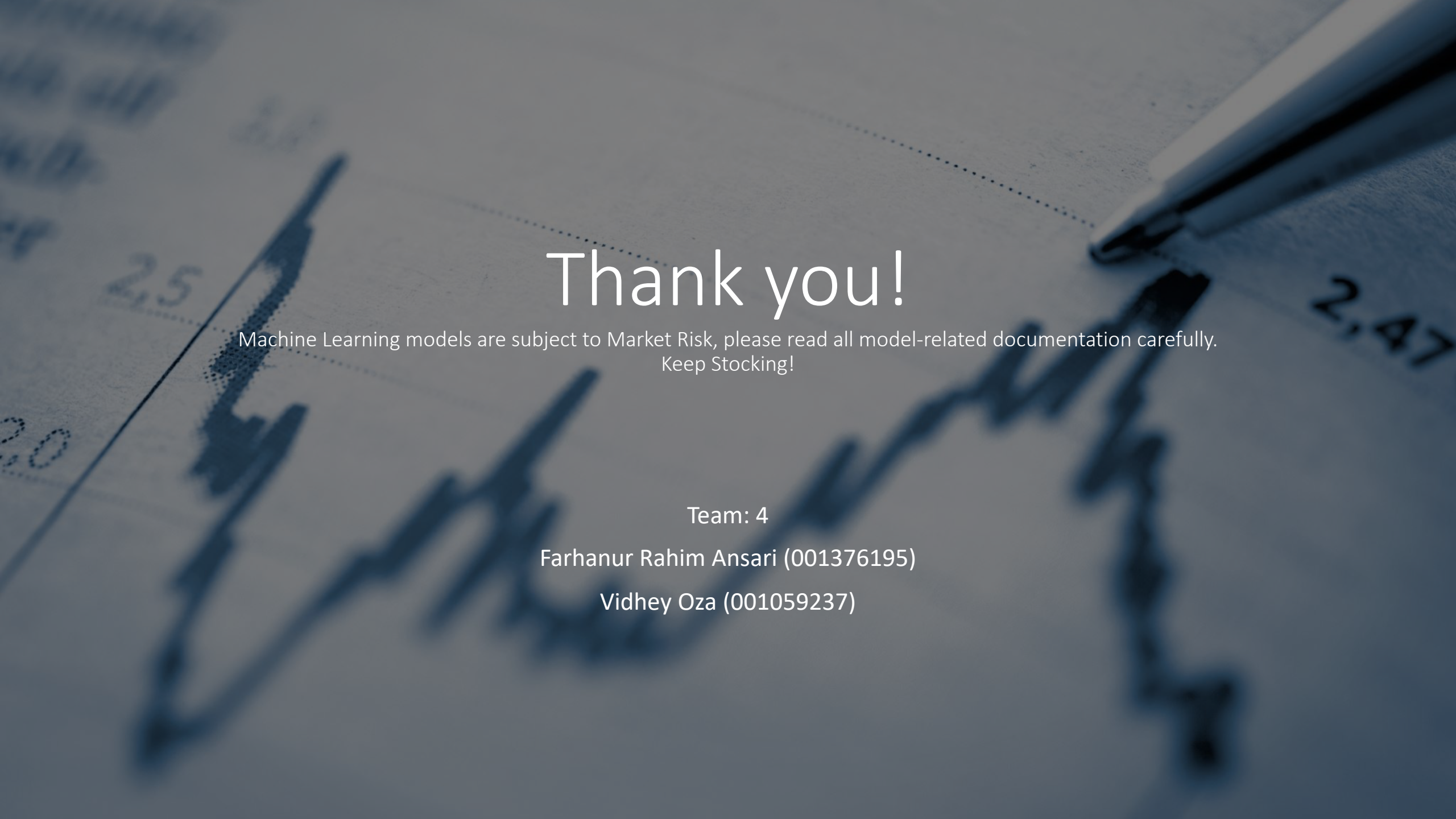
- No API supports direct extraction of intra-day stock prices.
- The dataset was manually downloaded and stored as a CSV file.
- Large volume of data

Train-Test Split:

- Training was done on the first 75% of the dataset, testing on last 25%
- Time series problem, so couldn't apply cross-validation; stock market prediction is done for future based on past values, but CV works on every subset of data, which gives model data that it won't have on test time

Feature Engineering:

- Deriving 22 engineered features from 5 main features was a challenge



Thank you!

Machine Learning models are subject to Market Risk, please read all model-related documentation carefully.
Keep Stocking!

Team: 4

Farhanur Rahim Ansari (001376195)

Vidhey Oza (001059237)