Forecasting Short-Term Stock Prices Using Machine Learning Models

Steven Whang
swhang00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Emilio Sagre
esagre00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Niha Sachin
nsachin00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Gus Dutra
gdutra01@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Langara.

Intro

The Stock Market can be **unpredictable** and making informed decisions has always been a **challenge**...

with **the rise of machine learning** researches and other professionals are incorporating new models to improve stock prices forecasting

Langara.

objective

Our project aims to **create accurate predictions** in the Stock Market...

by analyzing **historical data** of stocks, bonds, stock indexes, and economic commodities.

We explore the use of machine learning by comparing **four different algorithms** regarding prediction accuracy

4 models

01 XGBOOST

02 RANDOM FOREST

03 SUPPORT VECTOR REGRESSION

04 MULTILAYER PERCEPTRON

Langara.

method

We focused on **refining the models** by adding the following indicators

Stocks	Bonds	Stock Indexes	Commodities

Stocks

TSLA , NVDA , AAPL

Why?

Past performance can provide trends and **indicate future performance**, and **how the market has reacted to a variety of different variables**, from regular economic cycles to sudden, exogenous world events.

Langara.

We focused on **refining the models** by adding the following indicators

method

Stocks | Bonds | Stock Indexes | Commodities

Bonds

2-year treasury bond (TWOVX)

5-year treasury bond (FVX)

10-year treasury bond (TVX)

Why?

related to Interests rates; which can affect **the borrowing power of investors**.

method

We focused on **refining the models** by adding the following indicators

Stocks | Bonds | Stock Indexes | Commodities

Stock Indexes

Dow Jones (DOW)

Nasdaq Composite (NASX)

S&P 500

Why?

Dictates how the stock market moves on a daily basis as **they compose the largest stocks in the market**.

Langara.

We focused on **refining the models** by adding the following indicators

Stocks	Bonds	Stock Indexes	Commodities

Commodities

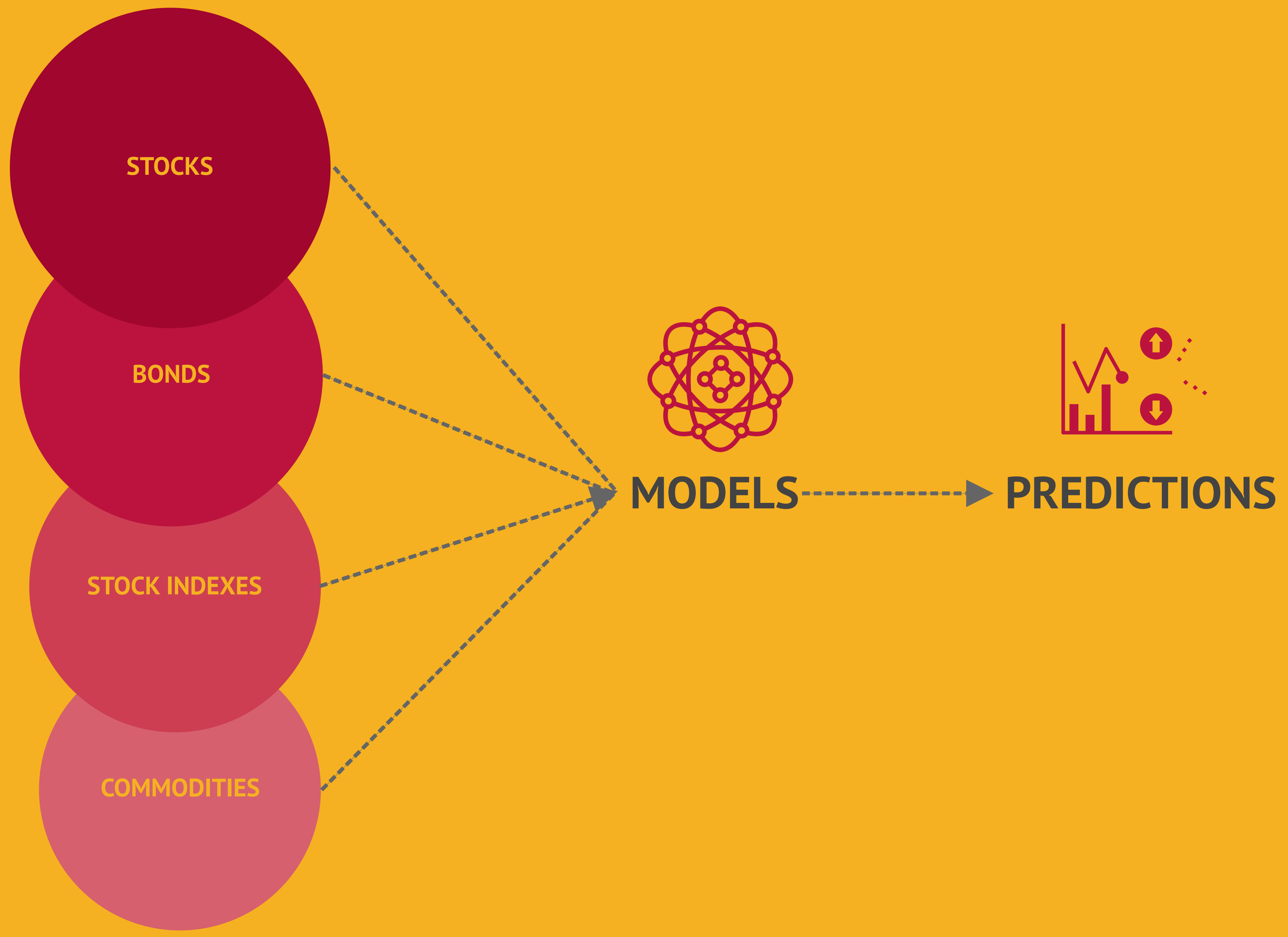
Gold, Oil

Why?

Contributes to the world's economic outlook and **heavily influences inflation.**

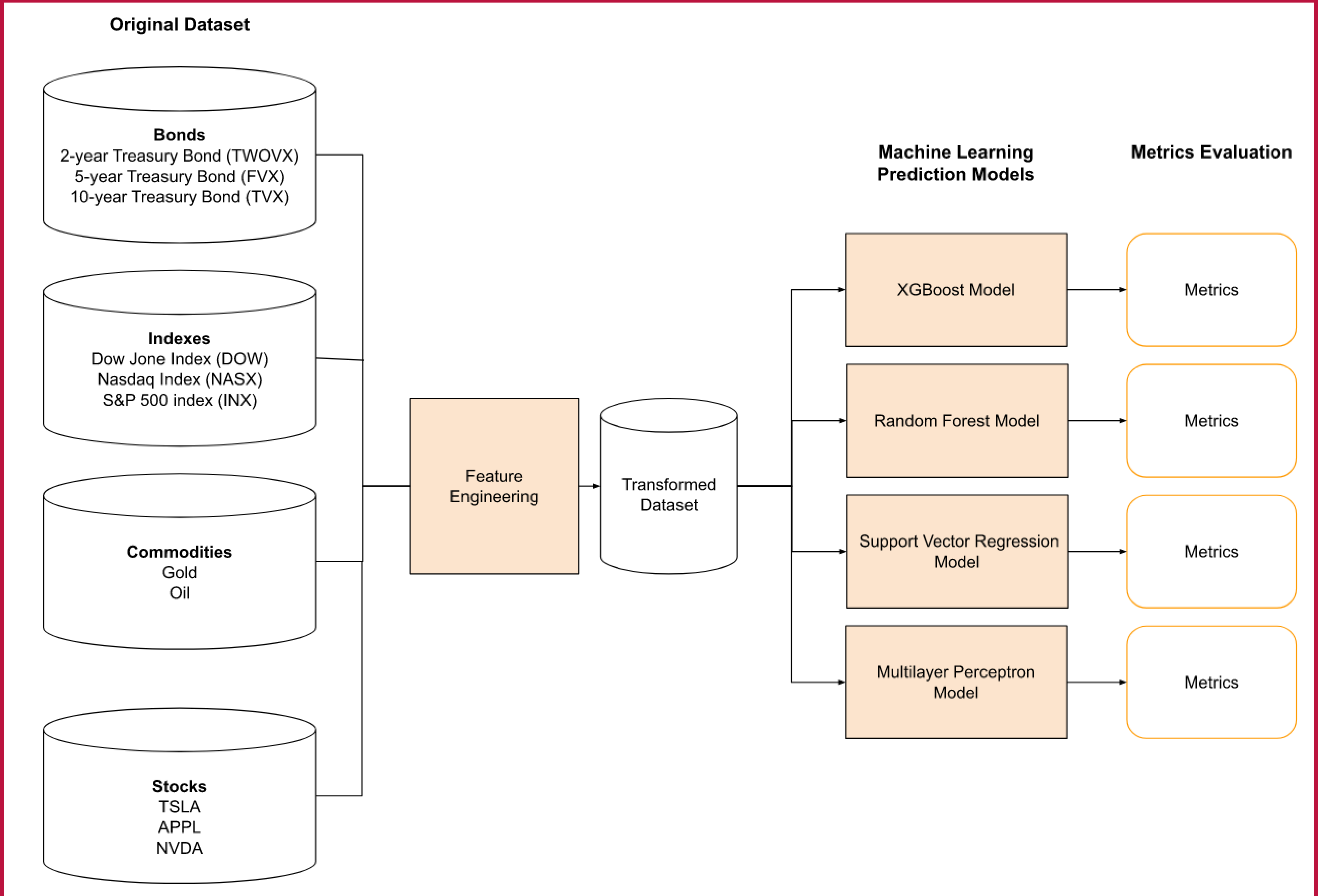
method

Our goal is to make **short-term predictions**, specifically forecasting 1 day ahead and 5 days ahead for **Tesla** (TSLA), **Apple** (AAPL), and **Nvidia** (NVDA)



Langara.

model building process



Langara.

feature engineering

Timeframe: March 2020 to May 2022.

Normalization details to follow

Numerical Var	Time Var	Target Var

Numerical Variables

Price of 2-year treasury bond (TWOVX);

5-year treasury bond (FVX);

10-year treasury bond (TVX);

Value of Dow Jones Index;

Value of Nasdaq Index;

Value of S&P 500 Index;

Price of Gold;

Price of Oil.

feature engineering

Timeframe: March 2020 to May 2022.

Normalization details to follow

Numerical Var	Time Var	Target Var

Time variables

Months of the year (12 variables);

Day of the month (31 variables);

Week day (5 variables for Monday to Friday);

Hours of the day (6 variables for hours 9 to 16);

Minute Segment of the hour (4 for minute segment 0, 15, 30, and 45);

Whether the time period is in Monday morning (1 variable);

Whether the time period is in Friday afternoon (1 variable);

Whether the time period is in a ”Pre-holiday” afternoon (1 variable);

Whether the time period is in a ”post-holiday” morning (1 variable).

Langara.

feature engineering

Timeframe: March 2020 to May 2022.

Normalization details to follow

Numerical Var	Time Var	Target Var

Target Variables

Price of Tesla Stock - TSLA; Target Variable 1

Price of Apple Stock - AAPL; Target Variable 2

Price of Nvidia Stock - NVDA; Target Variable 3

normalization and performance evaluation

Min-max normalization process applied across **all numerical variables** to lessen the effects of outliers;

4 accuracy measures to evaluate the performance of the machine learning models;

1 **MAPE**

Mean Absolute Percentage Error: It emphasizes on the percentage rather than the raw value, as it disregards different scales of the data resulting in easier interpretations.

2 **MPE**

Mean Positive Error: MPE is a business metric where we are trying to check if the forecasted value of the stock price is bigger than the actual value of the stock price.

3 **MTT**

Mean Train Time: Measures the amount of time it takes the model to train the dataset.

4 **RMSE**

Root Mean Squared Error: tells how far the predicted value is from the actual value.

Langara.

model iterations

specific parameters

XGBoost | Random Forest | Multilayer Perceptron | Support Vector Regression

</> XGBoost

XGBoost 1.0: n_estimators = 100, max_depth = 100

XGBoost 2.0: n_estimators = 300, max_depth = 100

model iterations

specific parameters

XGBoost	Random Forest	Multilayer Perceptron	Support Vector Regression

Random Forest

RF 1.0: n_estimators = 100, max_depth = 100

RF 2.0: n_estimators = 300, max_depth = 100

model iterations

specific parameters

XGBoost | Random Forest | Multilayer Perceptron | Support Vector Regression

Multilayer Perceptron

MLP 1.0: neurons = 100, activation = relu, dropout = 0.25, opt = Adam (amsgrad=True, lr =0.001,beta_1=0.79, beta_2 = 0.999), loss = mse

MLP 2.0: neurons = 100, activation = relu, dropout = 0.25, opt = Adam (amsgrad=True, lr =0.001,beta_1=0.79, beta_2 = 0.999), loss = mse, epochs=8, batch_size=256

MLP 3.0: neurons = 100, activation = relu, dropout = 0.25, opt = Adam (amsgrad=True, lr =0.001,beta_1=0.79, beta_2 = 0.999), loss = mse, epochs=20, batch_size=256

model iterations

specific parameters

XGBoost | Random Forest | Multilayer Perceptron | Support Vector Regression

</> Support Vector Regression

SVR 1.0: kernel = 'rbf', C=1.0, gamma = "scale"

SVR 2.0: kernel = 'rbf', C=5.0, gamma = "scale"

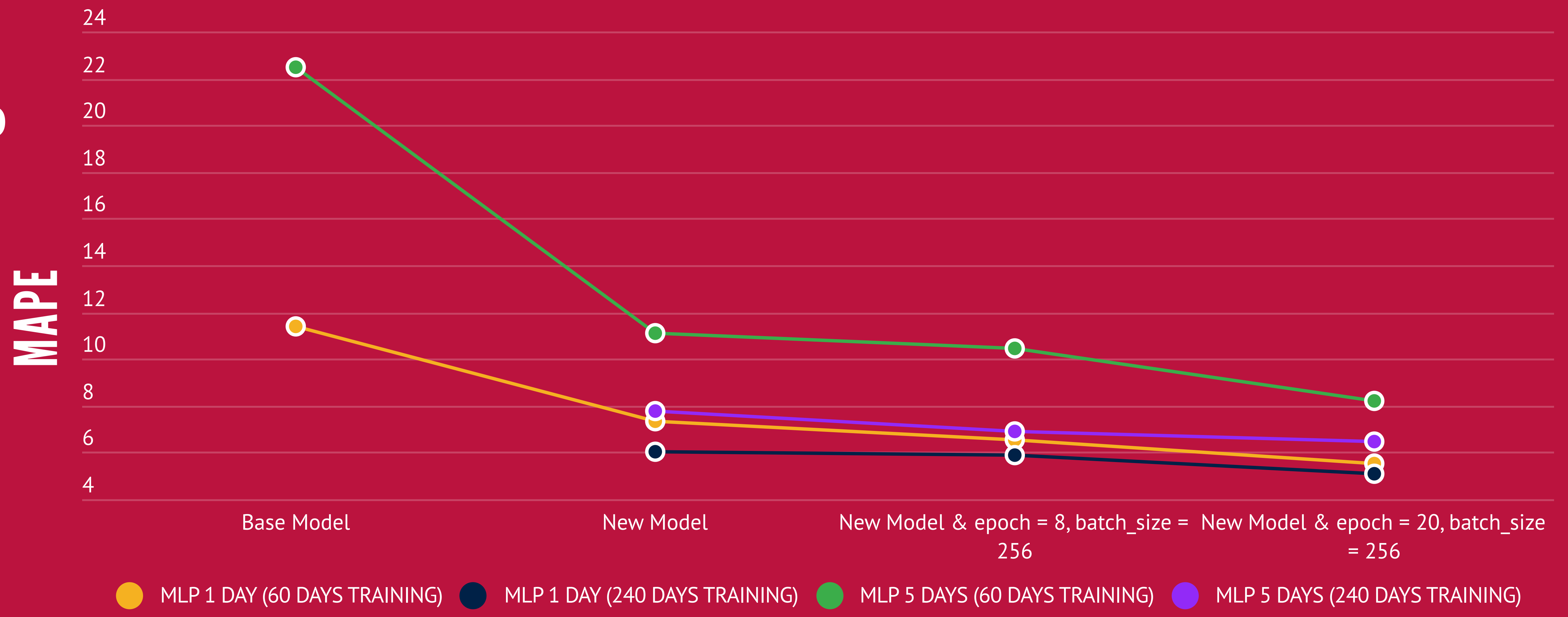
SVR 3.0: kernel = 'rbf', C=10.0, gamma = "scale"

results

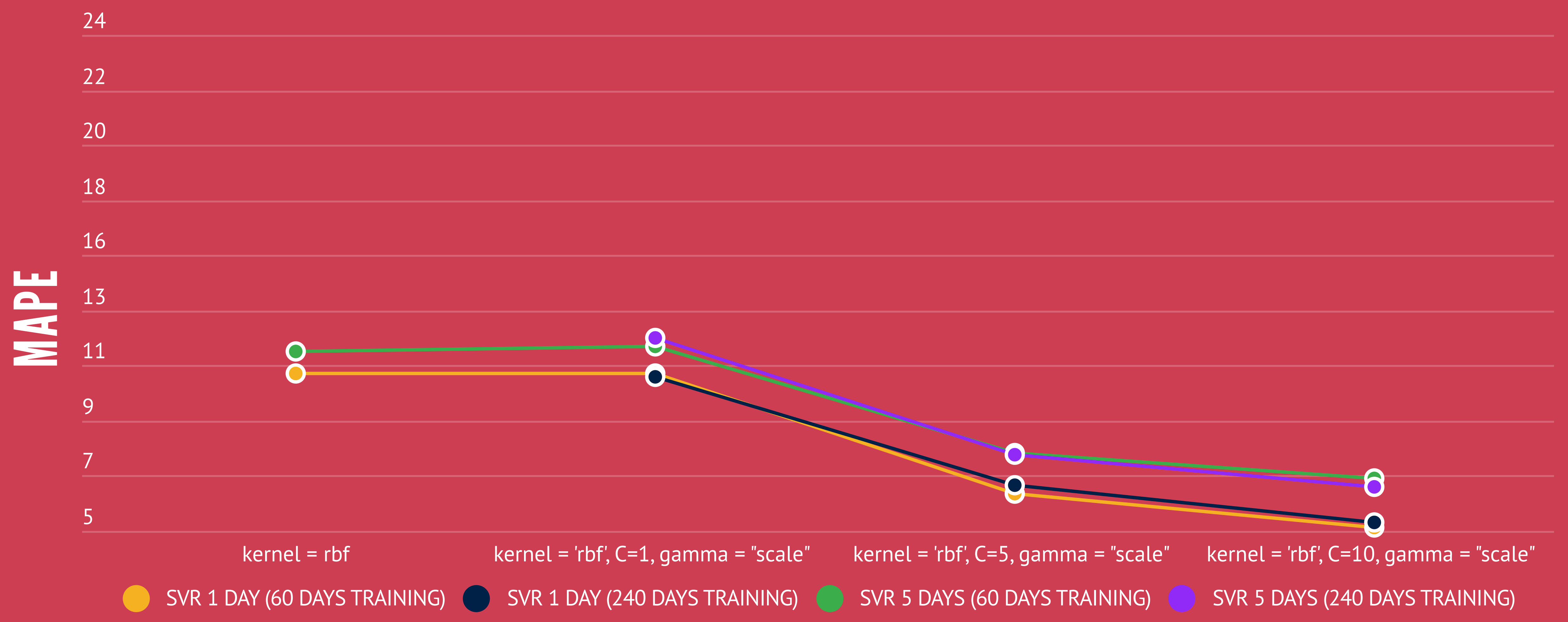
- We **compare the performance** of the models based on the evaluation metrics mentioned above.
- A **lower value** for all evaluation metrics **is favourable** as implies that the **prediction is close to the actual value**.
- For simplicity, results are split into **2 groups for each stock**: forecasts for 1-day ahead and 5-days ahead.
- Although the errors increase, it is advisable to **use as much historical data as possible** for forecasts and predictions.

MODEL COMPARISON

MLP



SVR



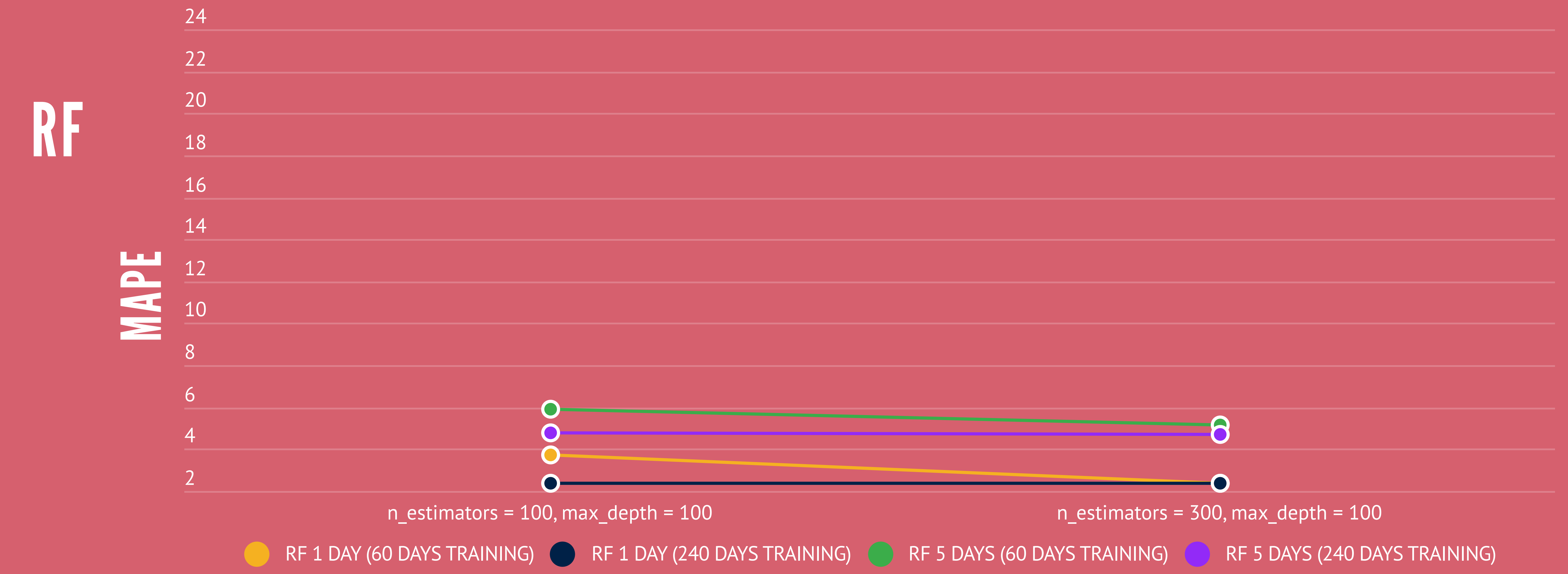
INTERPRETATION

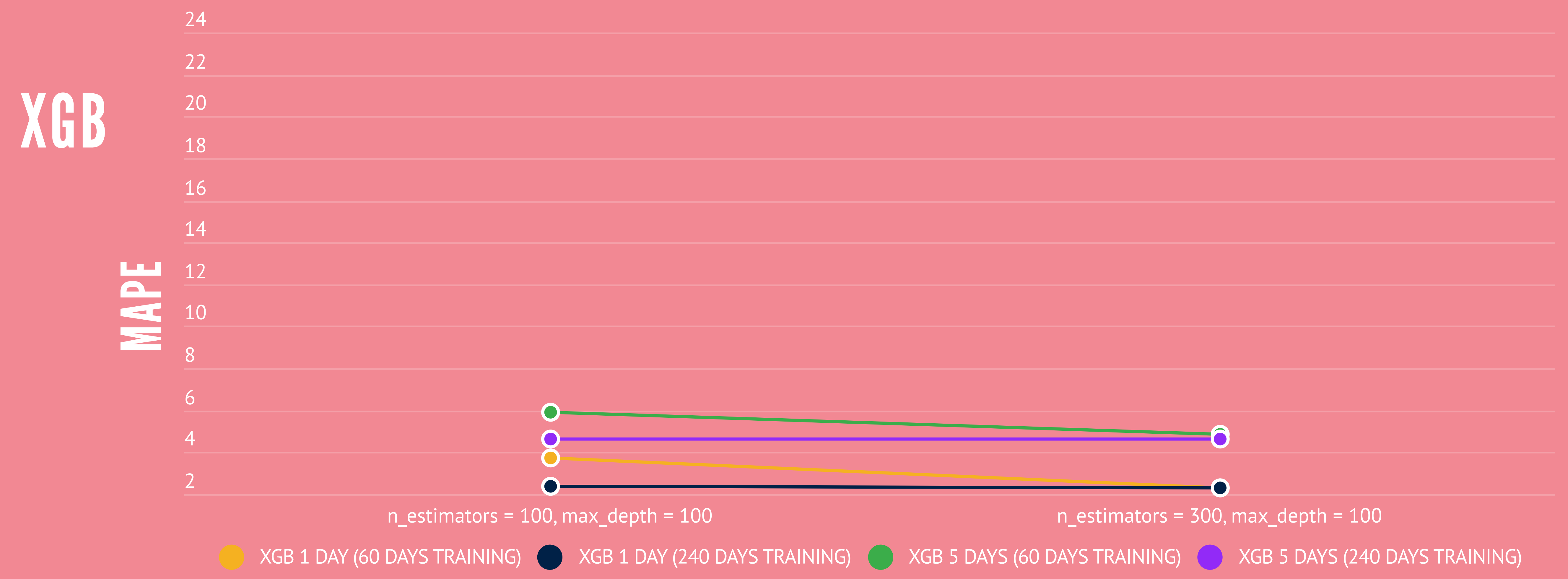
Both MLP and SVR show notable improvements for errors. The significant iteration for MLP is the increase in epochs, which shows a steady 1% improvement for Tesla when increasing epochs from 8 to 20.

For SVR, by increasing C from 1 to 5, the MAPE decreased by 5% (TSLA). However, once C is increased from 5 to 10, MAPE only decreased by less than 2%.

Langara.

MODEL COMPARISON





INTERPRETATION

Across all experiments, the XGBoost model produces the lowest errors compared to the other machine learning models.

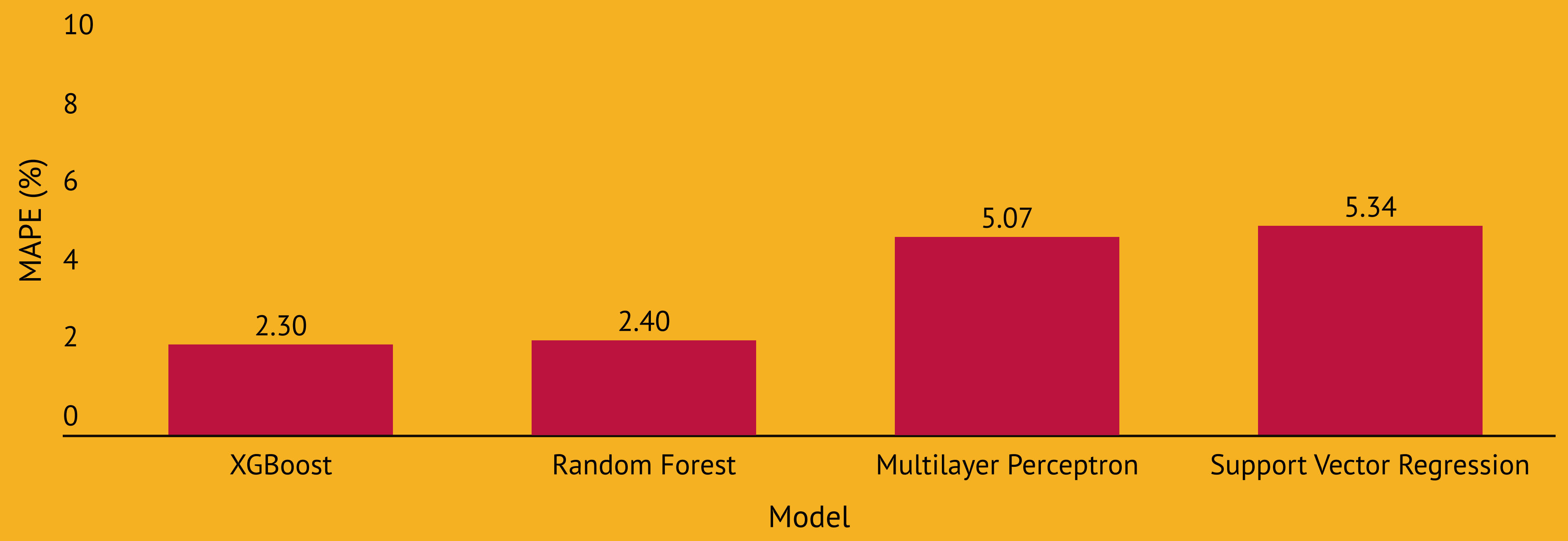
Interestingly, increasing N-estimators from 100 to 300 for both XGBoost and Random Forest with 60 and 240 training days showed little to no signs of improvement for the MAPE measure.

Langara.

TESLA

5-days ahead forecasts

240 days training dataset



1-day ahead forecasts

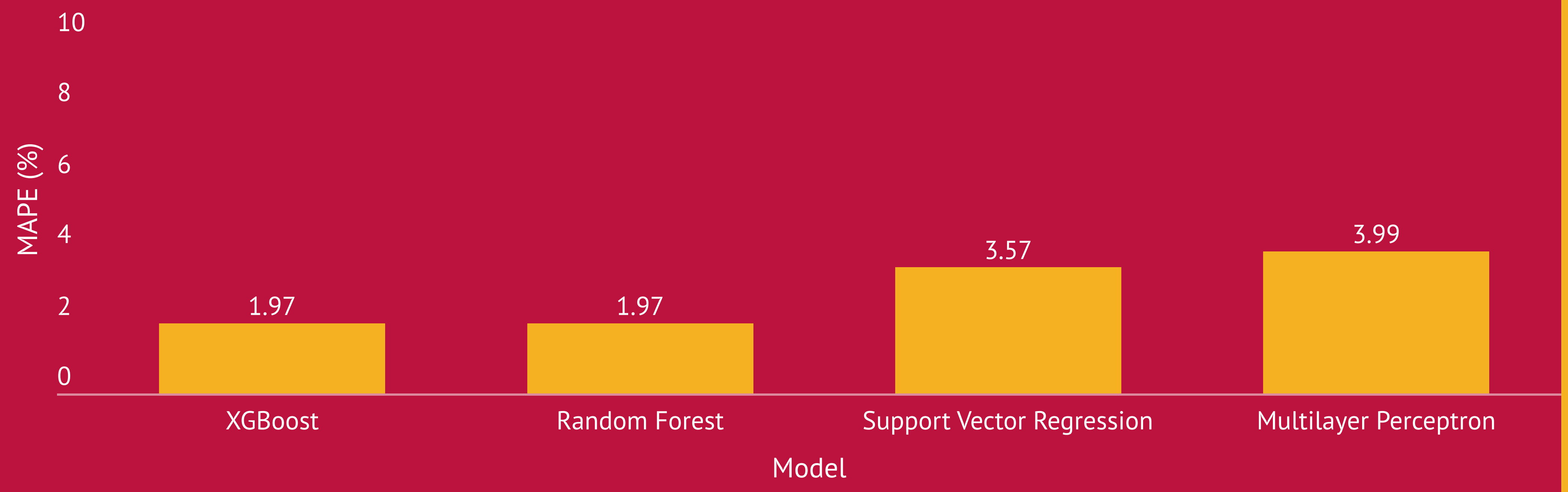
240 days training dataset

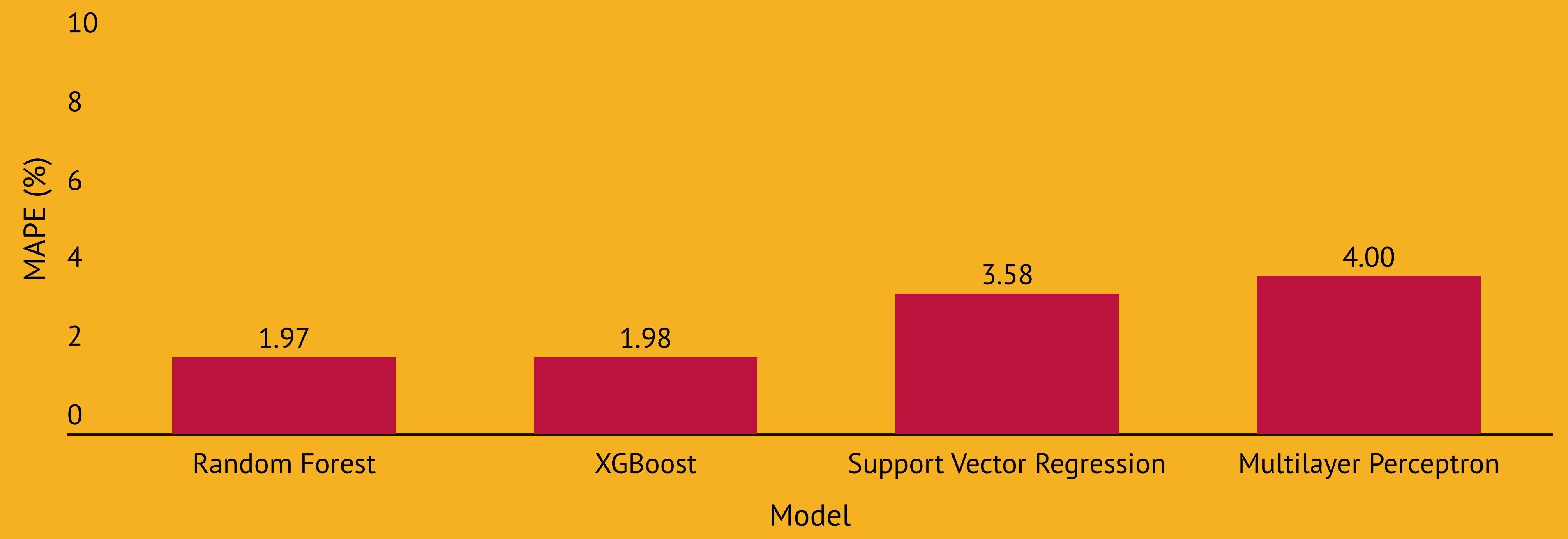
Model	Model Number	RMSE 60	RMSE 240	MAPE 60	MAPE 240	MPE 60	MPE 240	MTT 60	MTT 240
XGBoost	XGB 1.0	53.3438	59.7556	4.8314	4.6295	36.3369	40.6613	0.4993	3.2973
	XGB 2.0	53.2843	59.7160	4.8203	4.62	36.2566	40.5980	1.2594	1.5360
Random Forest	RF 1.0	58.7769	58.8869	5.249	4.779	39.7791	41.6486	0.7560	4.1401
	RF 2.0	58.0970	58.2680	5.18	4.7	0.4540	-0.0039	2.0410	1.2770
Multilayer Perceptron	MLP 1.0	103.4490	93.7580	11.1	7.8	2.3330	4.0870	0.1540	0.3300
	MLP 2.0	98.6997	83.0880	10.4375	6.89	67.9532	4.0870	0.2464	0.9525
	MLP 3.0	78.9532	75.0133	8.1672	6.488	59.4233	57.0666	0.3637	0.8539
Support Vector Regression	SVR 1.0	123.0820	164.5650	12.1	12.4	4.9670	9.5480	0.1120	1.4940
	SVR 2.0	85.9262	112.9910	7.9949	7.95	73.5371	4.1840	1.9469	1.4710
	SVR 3.0	75.3176	98.0147	7.0147	6.6904	52.9710	61.1254	0.1276	1.7576

Model	Model Number	RMSE 60	RMSE 240	MAPE 60	MAPE 240	MPE 60	MPE 240	MTT 60	MTT 240
XGBoost	XGB 1.0	28.6994	33.4075	2.30	2.39	17.7353	21.3779	0.9799	3.3622
	XGB 2.0	28.6149	33.3070	2.29	2.30	17.6374	21.2520	2.5916	7.9480
Random Forest	RF 1.0	31.3919	38.5990	2.41	2.41	18.5857	21.6248	0.9742	4.5863
	RF 2.0	31.3130	38.6290	2.40	2.40	-0.1870	-0.2133	2.0050	10.7240
Multilayer Perceptron	MLP 1.0	72.7950	72.0080	7.30	6.00	1.4913	1.7090	0.1063	0.4010
	MLP 2.0	63.9281	71.8420	6.53	5.90	47.3876	1.7090	0.2112	1.3205
	MLP 3.0	55.7913	60.6597	5.52	5.07	40.6648	44.3550	0.4640	0.8865
Support Vector Regression	SVR 1.0	115.335	114.0490	11.00	10.90	4.5650	113.8240	0.0990	1.4480
	SVR 2.0	73.8169	101.5630	6.43	6.78	49.3822	3.5400	0.1171	1.5210
	SVR 3.0	60.4214	84.7801	5.12	5.34	39.1658	49.3290	0.1448	2.4556

Langara.

NVIDIA





5-days ahead forecasts

240 days training dataset

1-day ahead forecasts

240 days training dataset

Model	Model Number	RMSE 60	RMSE 240	MAPE 60	MAPE 240	MPE 60	MPE 240	MTT 60	MTT 240
XGBoost	XGB 1.0	6.3038	8.3243	1.6995	2.0104	3.4850	4.8472	0.8520	2.9969
	XGB 2.0	6.2513	8.2578	1.668	1.9722	3.4281	4.7653	1.4872	8.7589
Random Forest	RF 1.0	6.9317	9.3045	1.6975	1.9881	3.5297	4.8831	0.7613	3.3682
	RF 2.0	6.8652	9.2339	1.6884	1.9771	3.5129	4.8588	2.5251	13.9600
Multilayer Perceptron	MLP 1.0	13.6289	13.9370	4.4211	4.3813	8.5569	10.2119	0.2191	0.4154
	MLP 2.0	12.4509	13.0708	4.1447	4.0746	8.0167	9.4167	0.3469	0.6758
	MLP 3.0	9.9559	12.4459	3.3086	3.9981	6.4021	9.3130	0.6034	1.4299
Support Vector Regression	SVR 1.0	16.8461	30.0103	4.9412	8.7727	10.6080	21.2391	0.1682	2.4363
	SVR 2.0	9.8268	15.9903	2.729	4.3074	5.7977	10.2527	0.2758	4.6295
	SVR 3.0	8.6046	13.6308	2.2921	3.577	4.8715	8.5171	0.3870	6.4798

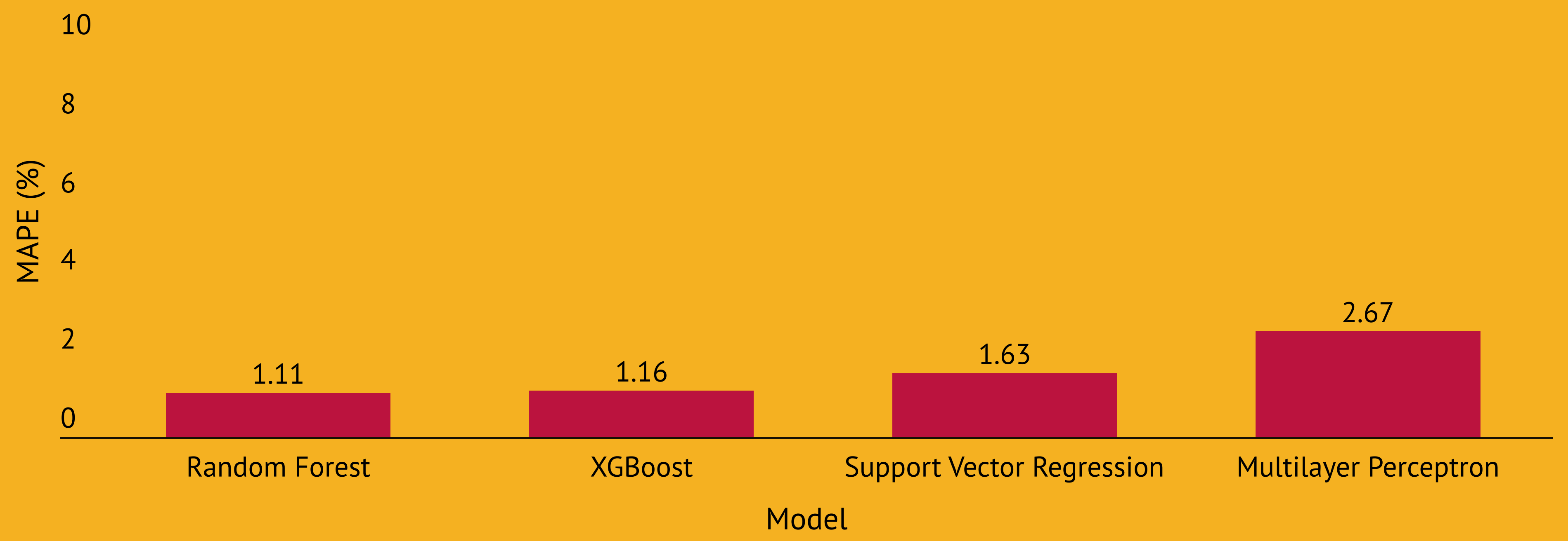
Model	Model Number	RMSE 60	RMSE 240	MAPE 60	MAPE 240	MPE 60	MPE 240	MTT 60	MTT 240
XGBoost	XGB 1.0	6.3038	8.3243	1.70	2.01	3.4850	4.8472	0.8520	2.9969
	XGB 2.0	6.2513	8.2578	1.67	1.97	3.4281	4.7653	1.4872	8.7589
Random Forest	RF 1.0	6.9317	9.3045	1.70	1.99	3.5297	4.8831	0.7613	3.3682
	RF 2.0	6.8652	9.2339	1.69	1.98	3.5129	4.8588	2.5251	13.9600
Multilayer Perceptron	MLP 1.0	13.6289	13.9370	4.42	4.38	8.5569	10.2119	0.2191	0.4154
	MLP 2.0	12.4509	13.0708	4.14	4.07	8.0167	9.4167	0.3469	0.6758
	MLP 3.0	9.9559	12.4459	3.31	4.00	6.4021	9.3130	0.6034	1.4299
Support Vector Regression	SVR 1.0	16.8461	30.0103	4.94	8.77	10.6080	21.2391	0.1682	2.4363
	SVR 2.0	9.8268	15.9903	2.73	4.31	5.7977	10.2527	0.2758	4.6295
	SVR 3.0	8.6046	13.6308	2.29	3.58	4.8715	8.5171	0.3870	6.4798

APPLE

Langara.

5-days ahead forecasts

240 days training dataset



1-day ahead forecasts

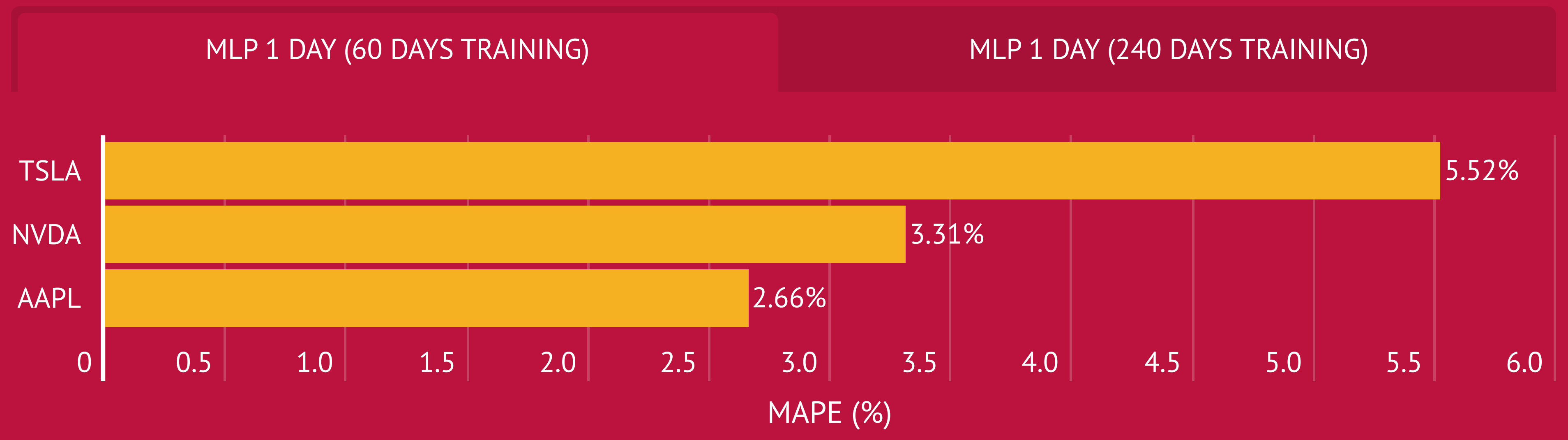
240 days training dataset

Model	Model Number	RMSE 60	RMSE 240	MAPE 60	MAPE 240	MPE 60	MPE 240	MTT 60	MTT 240
XGBoost	XGB 1.0	3.9461	5.0499	1.9389	2.1851	2.8054	3.4577	1.6343	4.7057
	XGB 2.0	3.9231	5.0138	1.9238	2.1601	2.7832	3.4185	3.4009	11.7320
Random Forest	RF 1.0	4.0242	5.1931	1.9783	2.2358	2.8255	3.4728	0.7133	3.6406
	RF 2.0	4.0338	5.1680	1.9849	2.2209	2.8369	3.4482	2.4670	9.6862
Multilayer Perceptron	MLP 1.0	16.5924	9.8825	6.6557	4.7899	9.2570	7.3740	0.1941	0.2556
	MLP 2.0	13.7633	9.0901	5.7839	4.5451	8.0849	6.9782	0.2177	0.5660
	MLP 3.0	9.6381	9.0027	4.3967	4.2568	6.2725	6.6269	0.5641	1.0026
Support Vector Regression	SVR 1.0	5.4410	8.1033	2.6612	3.8982	3.8446	6.1221	0.1947	2.8284
	SVR 2.0	4.7693	7.0648	2.2817	3.1366	3.2996	4.9238	0.3360	5.3325
	SVR 3.0	4.5780	6.9293	2.20%	3.05%	3.1816	4.7912	0.3978	8.3726

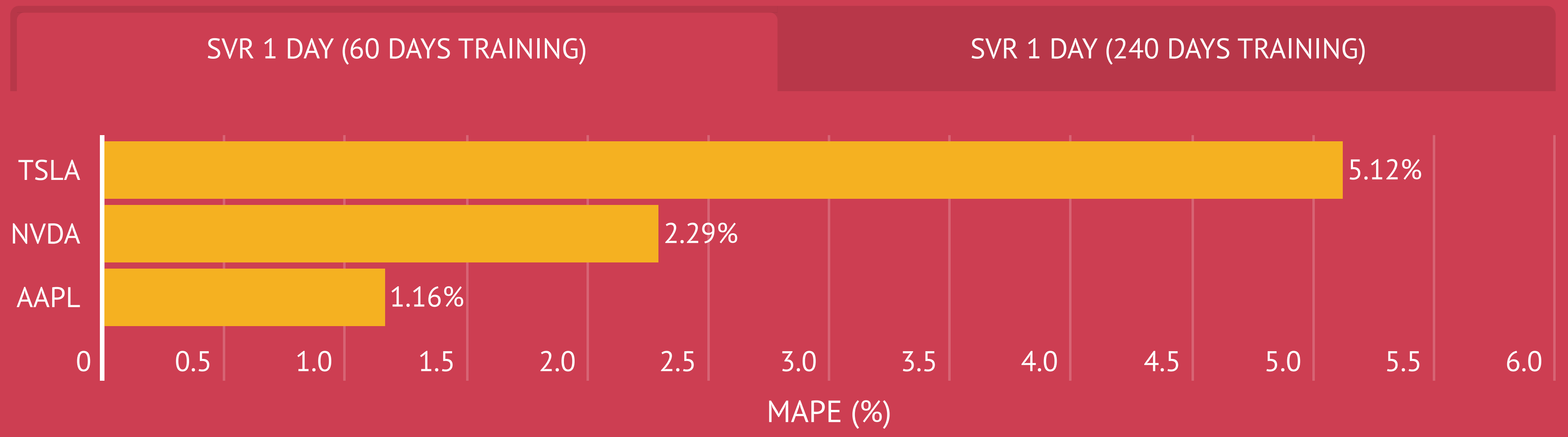
Model	Model Number	RMSE 60	RMSE 240	MAPE 60	MAPE 240	MPE 60	MPE 240	MTT 60	MTT 240
XGBoost	XGB 1.0	2.4639	3.1319	1.10	1.19	1.6028	1.8954	1.3888	4.2025
	XGB 2.0	0.4283	3.0889	1.08	1.16	1.5692	1.8453	1.8335	8.2777
Random Forest	RF 1.0	0.5190	3.2387	1.03	1.11	1.4877	1.7521	1.5357	5.7628
	RF 2.0	2.5115	3.2482	1.03	1.11	1.4906	1.7513	3.9954	11.3276
Multilayer Perceptron	MLP 1.0	8.7370	7.0270	3.59	3.31	5.0992	5.0955	0.1463	0.2555
	MLP 2.0	7.2373	6.8876	3.13	3.12	4.4675	4.8209	0.2159	0.4325
	MLP 3.0	5.3217	5.7895	2.66	2.67	3.7911	4.1880	0.4098	0.9680
Support Vector Regression	SVR 1.0	4.2251	6.4117	1.91	2.94	2.7771	4.6225	0.1632	2.5221
	SVR 2.0	3.0282	4.4575	1.26	1.82	1.8419	2.8683	0.2936	6.1959
	SVR 3.0	2.8071	4.1972	1.16	1.63	1.7029	2.5930	0.1948	10.0036

STOCK COMPARISON

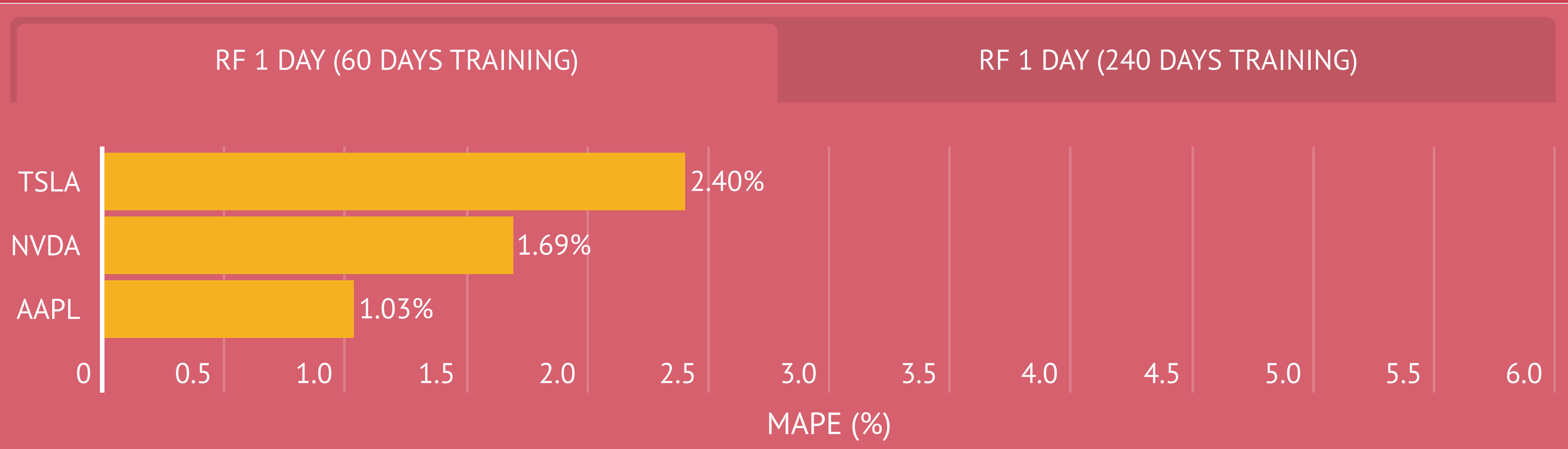
MLP



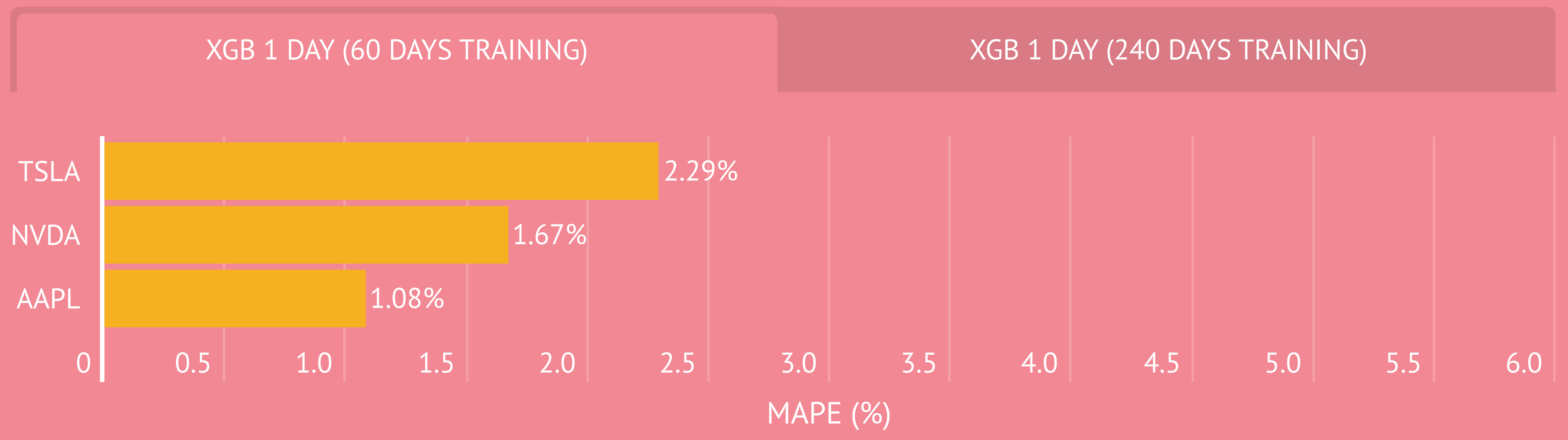
SVR



RF



XGB



OBSERVATIONS

By increasing the training days to 240 days, MAPE values across all 3 stocks increased.

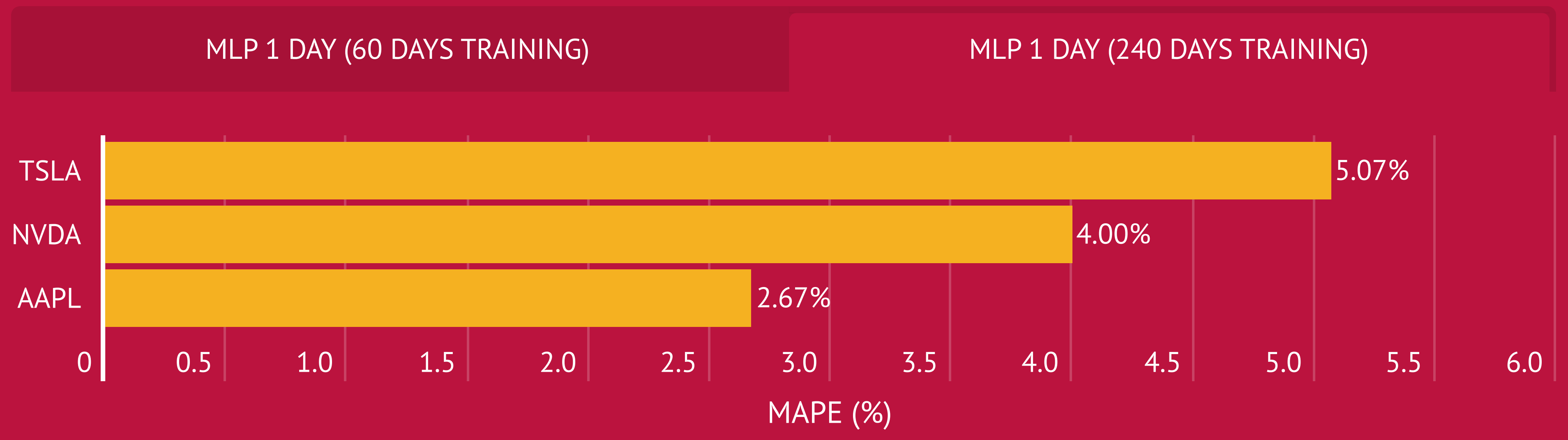
Among the 3 stocks, Apple has the lowest MAPE values, followed by Nvidia then Tesla. This can be attributed to Apple's stability.

Both RF and XGB have significantly lower MAPE values compared to MLP and SVR.

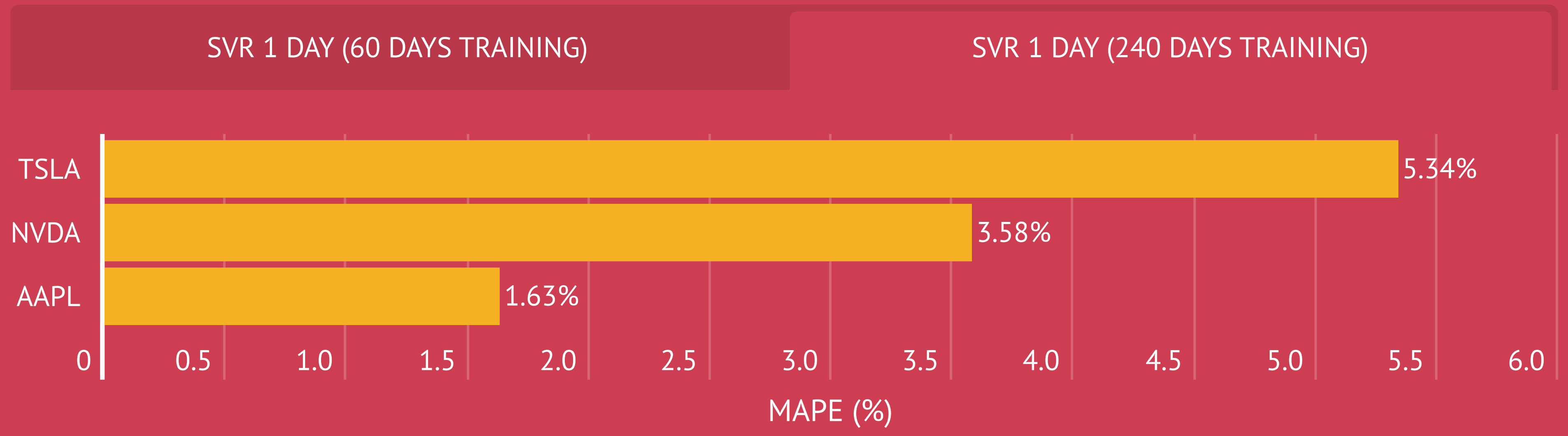
Langara.

STOCK COMPARISON

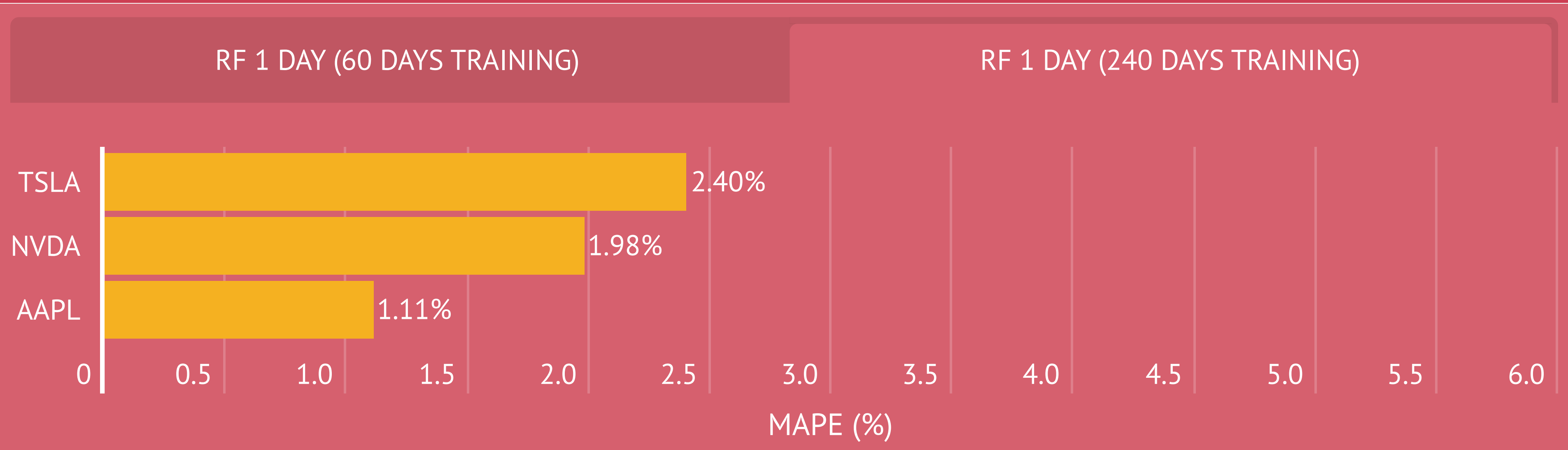
MLP



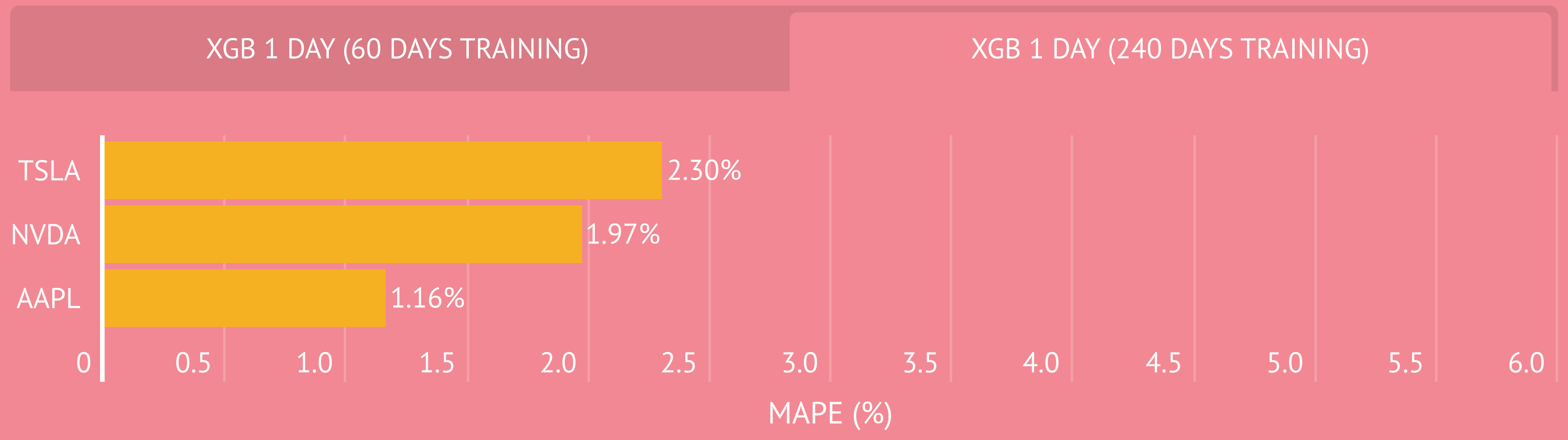
SVR



RF



XGB



OBSERVATIONS

By increasing the training days to 240 days, MAPE values across all 3 stocks increased.

Among the 3 stocks, Apple has the lowest MAPE values, followed by Nvidia then Tesla. This can be attributed to Apple's stability.

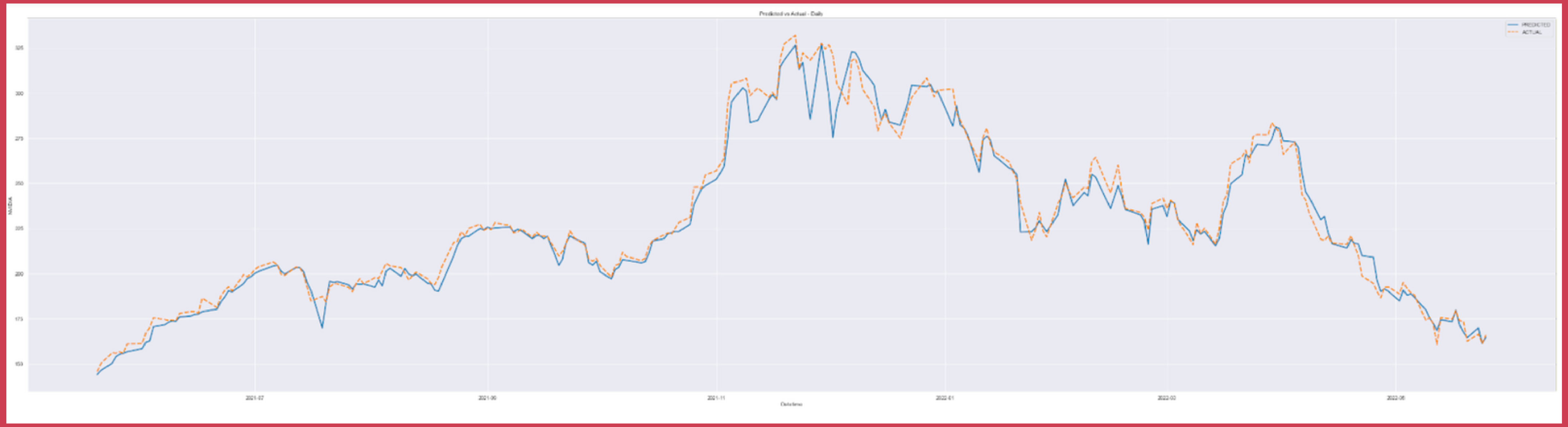
Both RF and XGB have significantly lower MAPE values compared to MLP and SVR.

Langara.

MODEL COMPARISON

XGB - TSLA

XGB - NVDA



XGB - AAPL

INTERPRETATION

Prediction accuracy is higher during periods with low volatility.

Errors occur when the observed price of the stocks fluctuate.

Among the three stocks Apple has the lowest evaluation metrics followed by Nvidia then Tesla - Apple is more mature, and less volatile than the other two stocks.

1-day ahead forecast
240 days training dataset

Langara.

conclusion

XGBoost has the **highest accuracy**. It can also be concluded that greater accuracy occurs **during low-volatility periods**.

A disadvantage is that XGboost has **the highest training time**.

Thank you!

Q&A

Steven Whang
swhang00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Emilio Sagre
esagre00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Niha Sachin
nsachin00@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Gus Dutra
gdutra01@mylangara.ca
Data Analytics
Langara College
Vancouver, BC

Langara.