Improving Price Prediction Accuracy for Magnificent Seven Stocks Utilizing Cross-Stock Historical Data

Kunasekaran Nirmalkanna

Independent Study Project

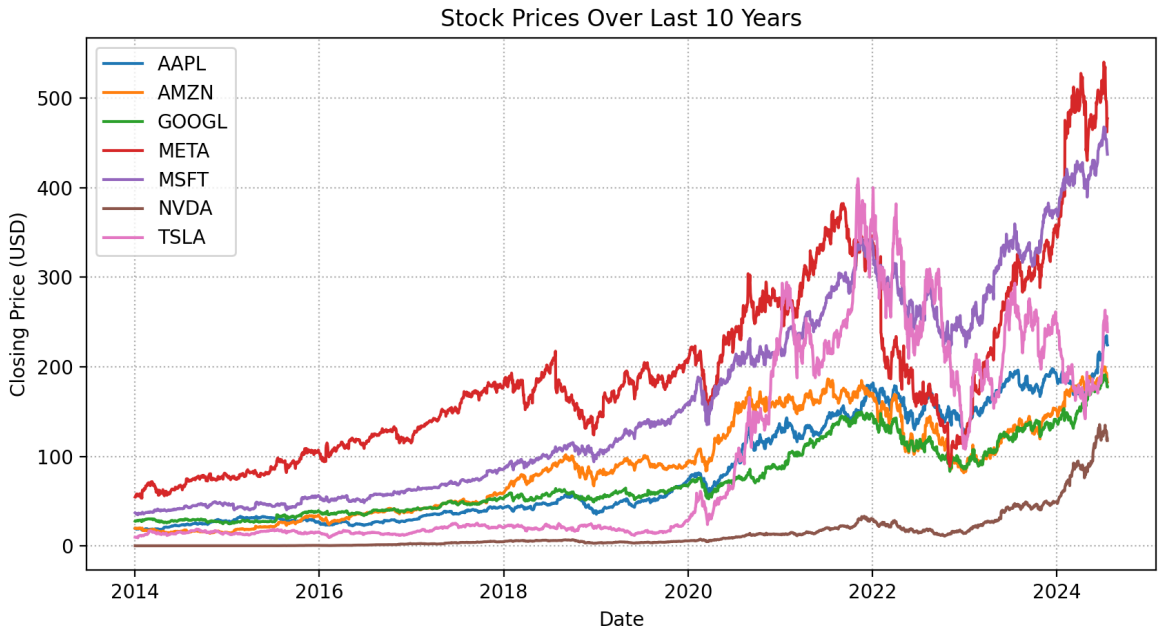
ACENET Microcredential in Advanced Computing

July 23, 2024

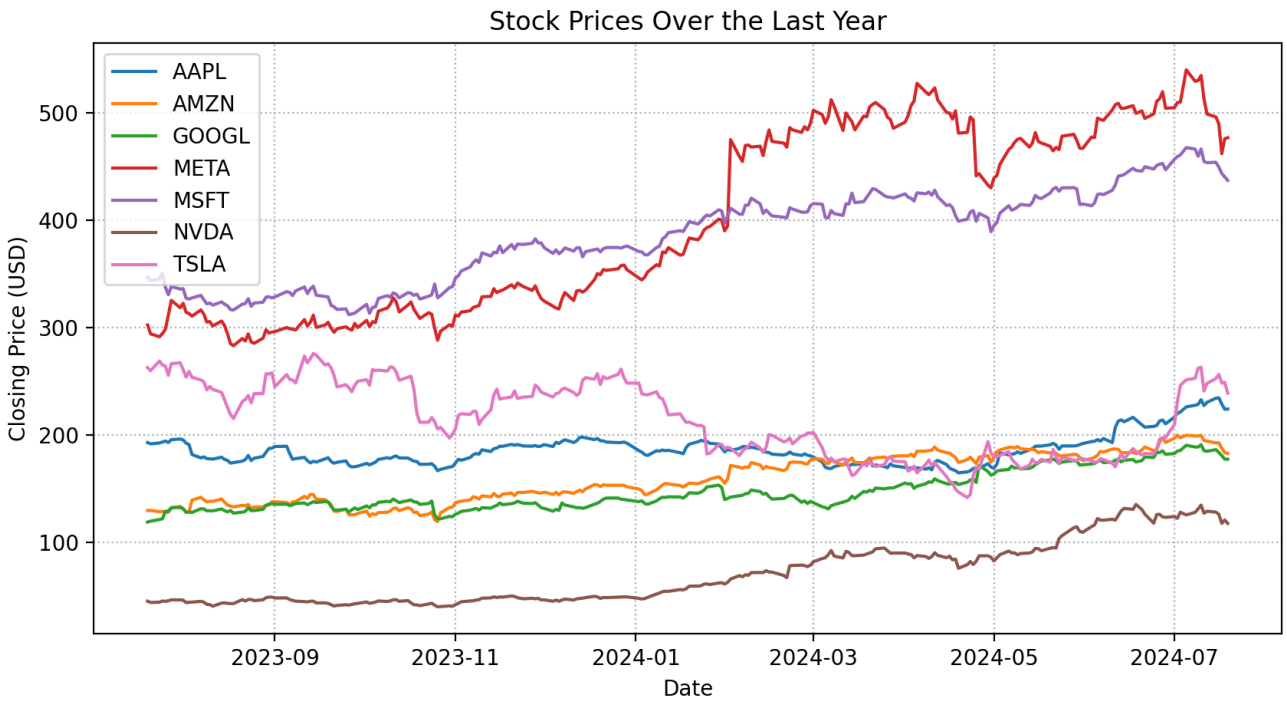
Outline

Introduction

- "Magnificent Seven Stocks."
  - **AAPL**: Apple Inc.
  - **AMZN**: Amazon.com Inc.
  - **GOOGL**: Alphabet Inc. (Google)
  - **META**: Meta Platforms Inc. (Facebook)
  - **MSFT**: Microsoft Corporation
  - **NVDA**: NVIDIA Corporation
  - **TSLA**: Tesla Inc.
- Data
  - **`yfinance`** package in python (Aroussi, 2019) offers a threaded and Pythonic way to download market data from Yahoo! finance.



Introduction



- AutoRegressive Integrated Moving Average (ARIMA) is a popular method (Ariyo et al., 2014) to predict stock price.
- Can we improve prediction using other stocks?

Granger Causality

- Granger causality (Granger, 1969) is a statistical hypothesis test used to determine if one time series can predict another time series.

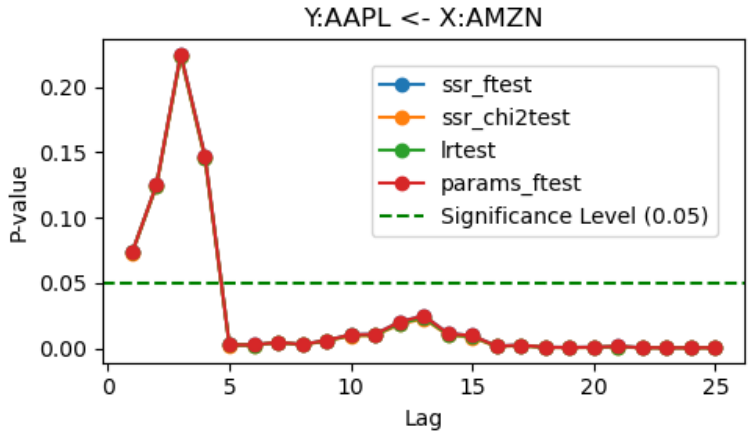
$$
\begin{aligned}
M1: \quad & Y_t = \phi_0 + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \epsilon_t \\
M2: \quad & Y_t = \phi_0 + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \phi_1^* X_{t-1} + \ldots + \phi_p^* X_{t-p} + \epsilon_t
\end{aligned} \tag{1}
$$

  - Y_t: Time series data at time t
  - X_t: Predictor time series data at time t
  - $\phi_0, \phi_1, \ldots, \phi_p, \phi_1^*, \ldots, \phi_p^*$: Regression parameters

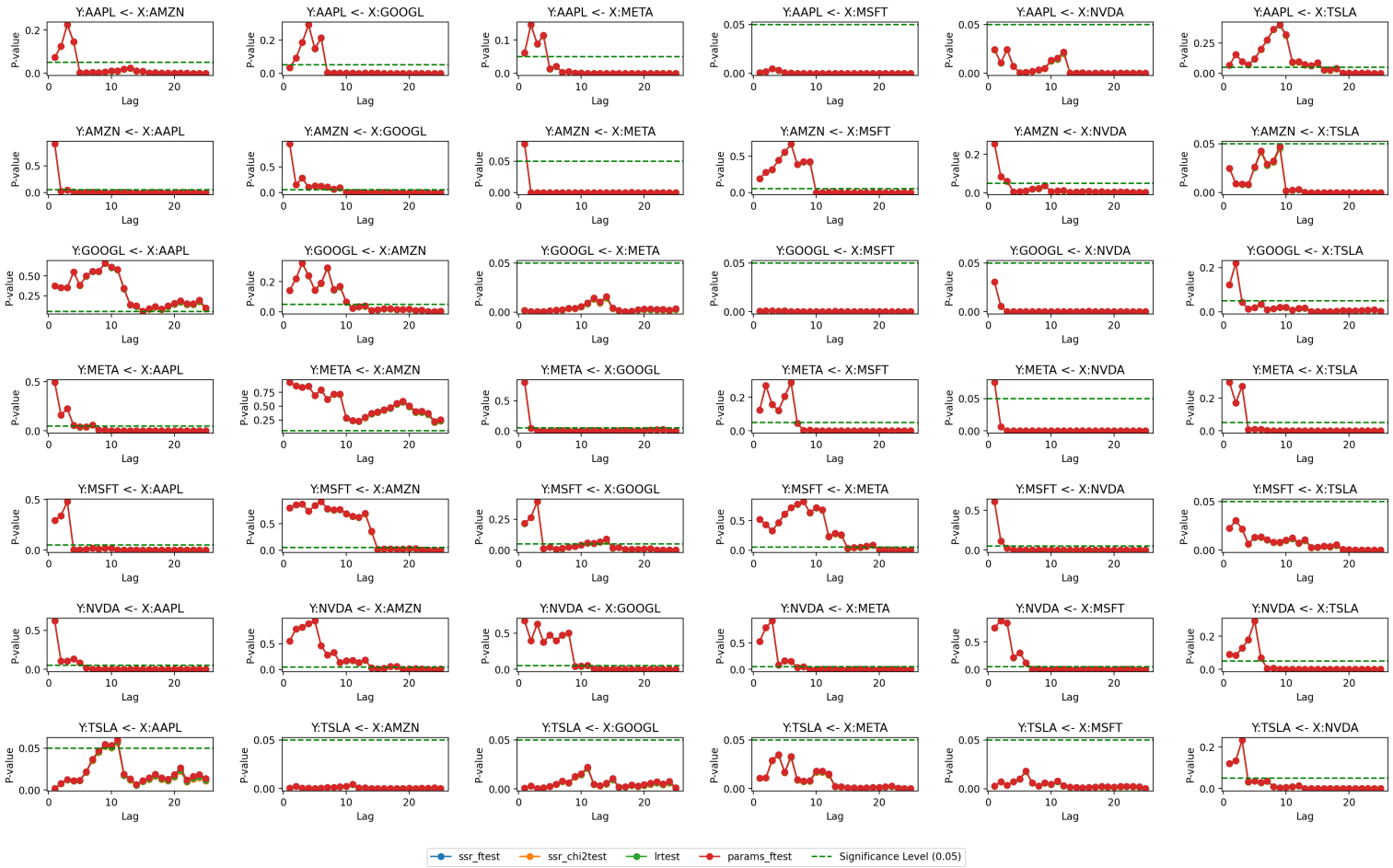
- **`grangercausalitytests`** function in **`statsmodels`** (Seabold and Perktold, 2010) package in python is used to conduct the test.

- Statistical test:
  - `params_ftest`, `ssr_ftest` are based on F distribution
  - `ssr_chi2test`, `lrtest` are based on chi-square distribution
  - Lower P-value suggests that X Granger-causes Y, meaning X has predictive power for Y in lag p.

- E.g. AMZN Granger causes AAPL.



Granger Causality



- *Objective of this study:* How much improvement in prediction by considering the history of other stocks in the model.

Model 1

- Let's define **M1** as

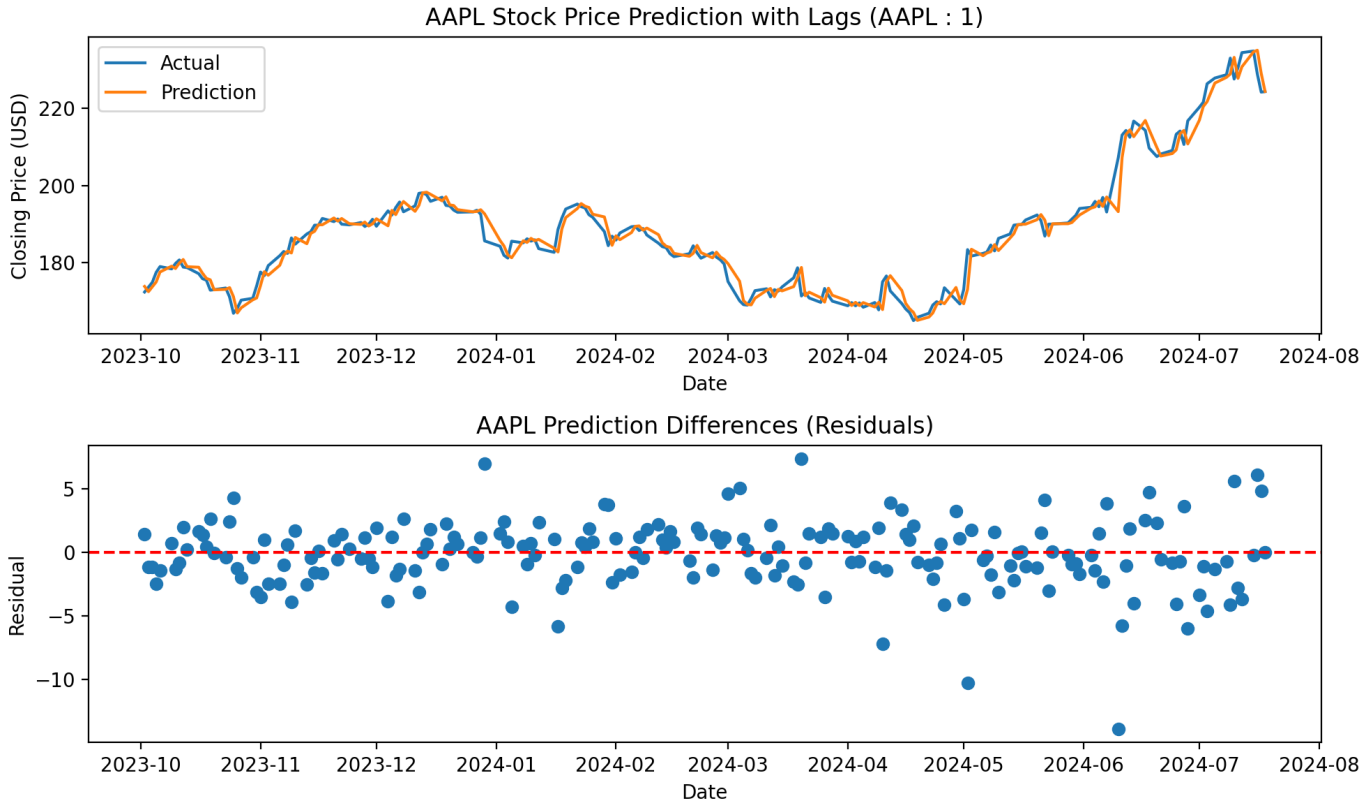
$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \epsilon_t, \qquad t = 1, \cdots, n, \tag{2}$$

  where $\epsilon_t \sim N(0, \sigma^2)$.

- **`LinearRegression`** function in **`sklearn`** package in python (Pedregosa et al., 2011) is used to fit the linear model.
- It is also know as AutoRegressive AR(p) model.
- Considered $n = 2000$ (≈ 8 *years*) to estimate $\phi_1, \ldots, \phi_p$.
- Considered $p = 1$, assuming that the stock price depends only on previous day's price of itself.
- With the estimated ϕs, calculate the Y_{t+1} and denote the estimated value as $\hat{Y}_{t+1}$.

Model 1

- E.g. The prediction and actual stock prices of AAPL



- **Target**: Reduce the absolute residuals.
  - Same as reducing root mean square error $\left(RMSE = \sqrt{\frac{\sum_{j=1}^{k} \left(Y_j - \hat{Y}_j \right)^2}{k}} \right)$.

Model 2

- Let's define **M2** as

$$\begin{aligned} Y_t =& \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \\ & \phi_1^* X_{t-1} + \ldots + \phi_{p_x}^* X_{t-p_x} + \epsilon_t, \qquad t = 1, \cdots, n, \end{aligned} \tag{3}$$

  where $\epsilon_t \sim N(0, \sigma^2)$.

- Considered $n = 2000$ (≈ 8 *years*) to estimate $\phi_1, \ldots, \phi_p$.
- Considered $p = 1$, assuming that the stock price depends only on previous day's price of itself.
- Considered $p_x = 1, \cdots, 5$.
  - Only single lag where

$$Y_t = \phi_1 Y_{t-1} + \phi_{p_x}^* X_{t-p_x} + \epsilon_t$$

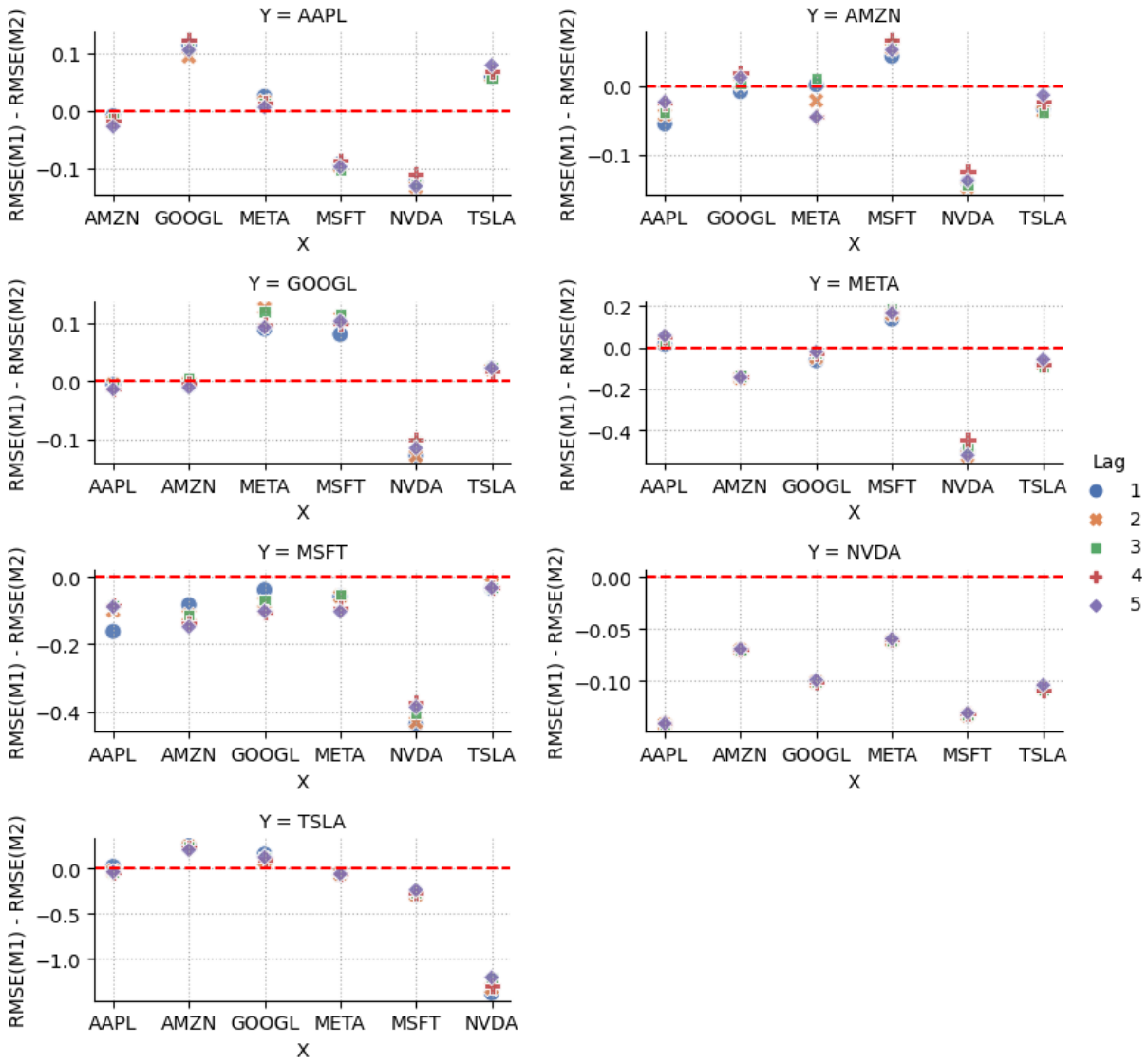
  - Multiple lag

$$Y_t = \phi_1 Y_{t-1} + \phi_1^* X_{t-1} + \ldots + \phi_{p_x}^* X_{t-p_x} + \epsilon_t$$

- With the estimated ϕs, calculate the Y_{t+1} and denote the estimated value as $\hat{Y}_{t+1}$.

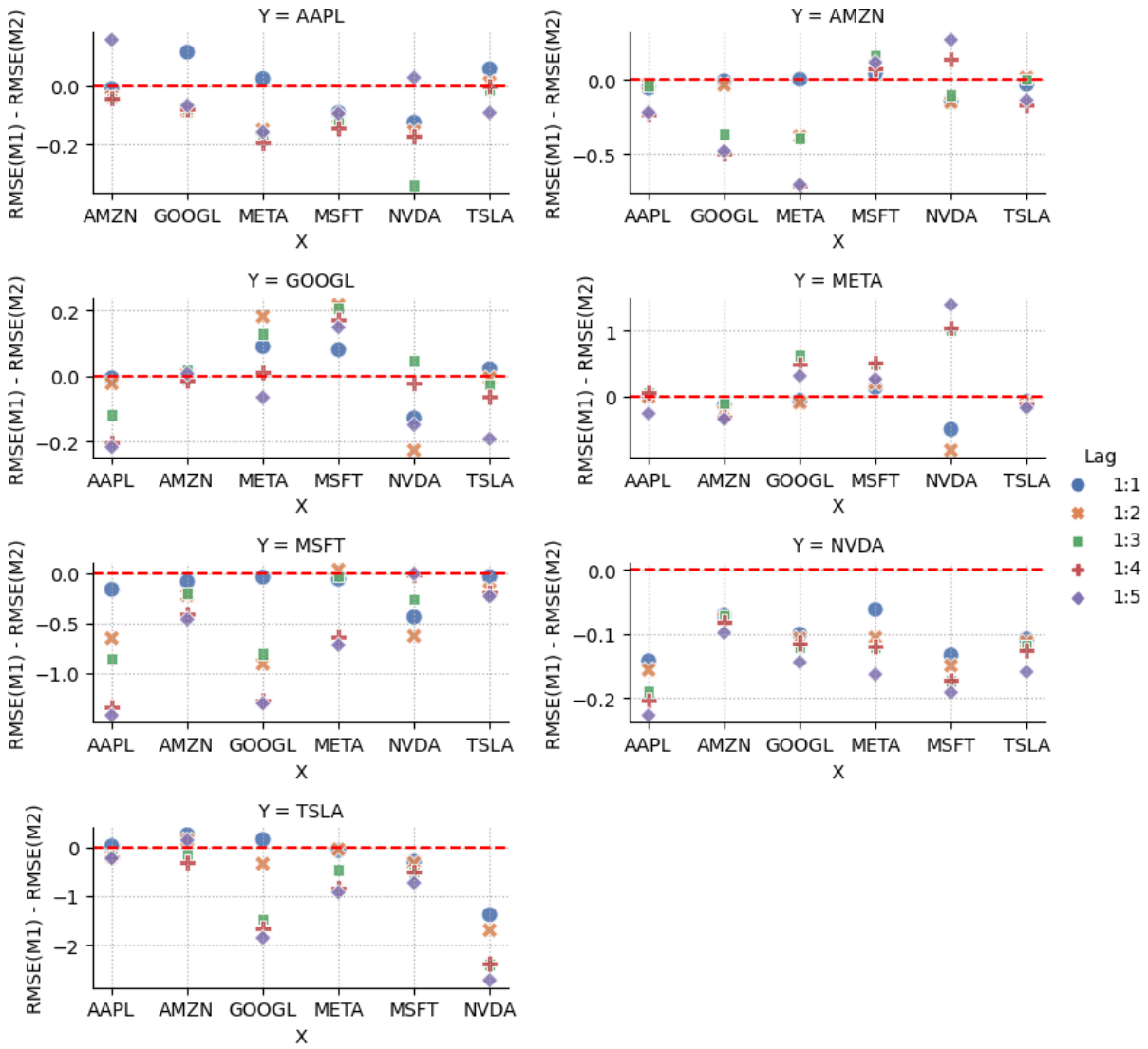
Difference in RMSEs

- $RMSE(M1) - RMSE(M2) > 0 \implies M2$ is better than $M1$ in prediction.
- Only single lag



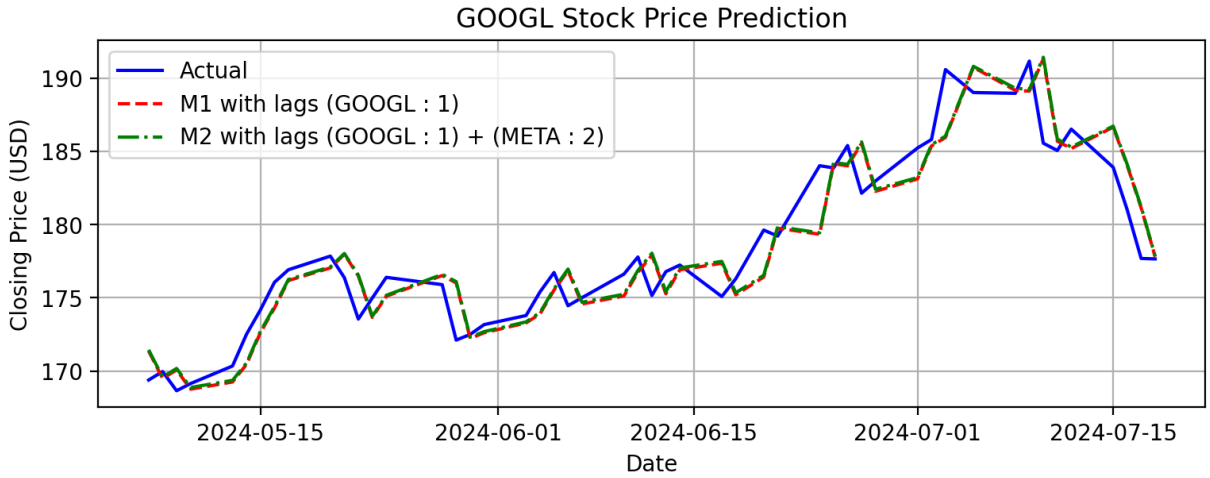
Difference in RMSEs

- $RMSE(M1) - RMSE(M2) > 0 \implies M2$ is better than $M1$ in prediction.
- Multiple lag

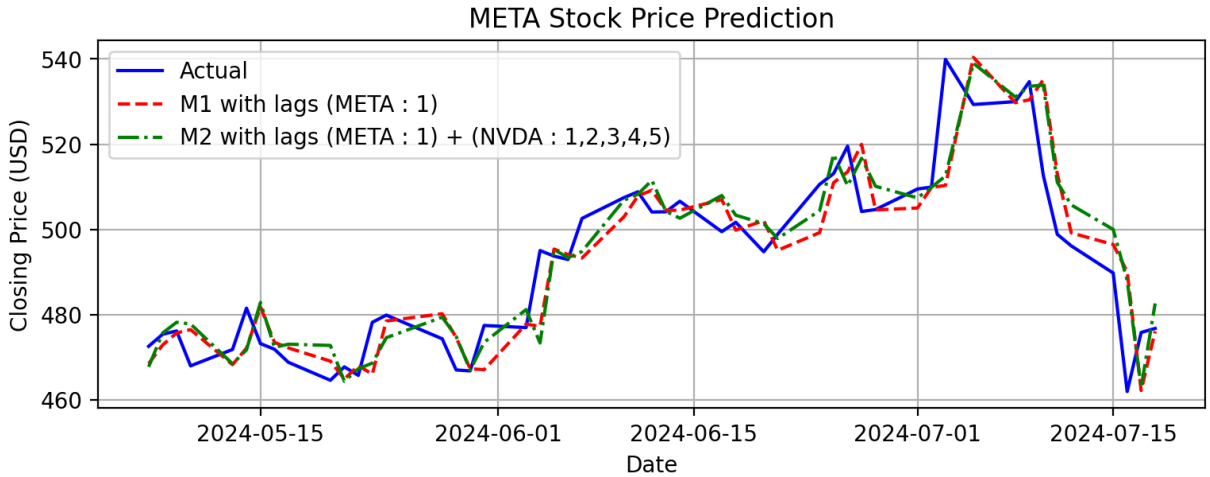


Conclusion

- Including other lagged stock prices in the model improves the prediction.
- E.g.
  - Only single lag



  - Multiple lag



Future Work

- Can consider other values than $n = 2000$ (≈ 8 *years*).
- Can consider $p = 1, \cdots$.
- Can consider $p_x = 1, \cdots, 5, \cdots$.
- **M2** can be

$$\begin{aligned} Y_t =& \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \\ & \phi^*_{1,1} X_{t-1} + \ldots + \phi^*_{1,p_{x_1}} X_{t-p_{x_1}} + \\ & \vdots \\ & \phi^*_{6,1} X_{t-1} + \ldots + \phi^*_{6,p_{x_6}} X_{t-p_{x_6}} + \epsilon_t, \qquad t - 1, \cdots, n, \end{aligned} \tag{4}$$

  where $\epsilon_t \sim N(0, \sigma^2)$.
- Can include other stocks than Magnificent Seven Stocks.

Thank you

References

[1] Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pages 106–112. IEEE.

[2] Aroussi, R. (2019). yfinance: Yahoo! finance market data downloader. Version 0.2.41.

[3] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[5] Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.