Stock Price Forecast with Time Series Models: Comparisons of Their Accuracy across Sectors

Richard Lu

New York University

MATH-UA 251 Introduction to Math Modeling

Abstract. This essay explores the application of time series models, specifically ARIMA and SARIMA, in forecasting stock prices across different sectors (technology, food manufacture, and energy). The study uses daily closing prices of 30 publicly traded US companies over three years as sample data. The author conducts a practical example using Apple Inc. stock data, determining optimal parameters for both ARIMA and SARIMA. Evaluation metrics, such as R-square and MAPE, are employed for model comparison. Results suggest varying performance across sectors, with SARIMA being better in food manufacture, ARIMA in energy, and mixed outcomes in the technology sector. The study acknowledges the simplicity of the models, emphasizing the need for refinement and consideration of additional factors for more accurate predictions. The essay concludes with recommendations for potential improvements, including the use of daily return rates, introducing extra explanatory variables, exploring different seasonal lengths, and considering machine learning models for future research.

Key Words: time series, ARIMA, SARIMA, stock forecast

1.Introduction

Investors have always been interested in predicting the future performance of stocks to help decide whether to buy or sell them and the best timing for such actions. These prediction attempts can be generally classified into two categories: fundamental analysis and technical analysis. Fundamental analysis considers information like a company's financial statements (revenue, profits, assets and liabilities, cash flows, etc.), macroeconomic indicators and related policies (GDP, inflation rate, interest rate, etc.), plus the analyst's own interpretation of the whole industry's outlook and the company's strengths and weaknesses against its competitors. In comparison, technical analysis focuses on using a stock's past prices and trading volumes to predict its future prices via tools like statistical models and computer programming. [1] Technical forecasting is especially trending in quantitative finance nowadays thanks to advancements in machine learning technologies. There are a lot of debates about the validity of technical analysis within academia, due to a lack of evidence proving its profitability in the long run and it being a direct contradiction of the efficient market hypothesis (EMH), which states that current stock prices fully reflect all information contained in past prices so it's impossible to predict the future using past data. [1] Based on EMH, stock prices movements should be stochastic with equal chances of going up or down with the same amount of profits or losses, thus making this a martingale process: at any time, the expected future value of the stock price is equal to the current price. In this article, it is assumed that EMH does not apply and two basic deterministic time series models ARIMA and SARIMA are trained using stock data of different US companies across three sectors to forecast their future prices, followed by a comparison of the performances of the two models using statistical metrics. There are three sections: introduction, methods and results, discussion and conclusions.

2. Methods and Results

2.1 Autoregressive Model (AR)

The Autoregressive (AR) model of order *p*, often denoted as AR(p), is a time series model where the current value of the series is modeled as a linear combination of its *p* previous values ("lags") plus a random error term. The mathematical equation is:

$$X_t = c+\phi_1 X_{t-1}+\phi_2 X_{t-2}+\ldots+ \phi_p X_{t-p}+\varepsilon_t$$

where X_t is the value of the stock price at time t, c is a constant term, ϕ_1, $\phi_2,\ldots,\phi_p$ are the autoregressive coefficients, and ε_t is a white noise error term at time t, the random and unpredictable component of the observed data that is not explained by the model. If it is believed that the stock price today is correlated with yesterday's price only, then the order of lags is 1 and an AR(1) model could be built with the equation $X_t = c+\phi_1 X_{t-1}+\varepsilon_t$.

2.2 Moving Average Model (MA)

The Moving Average (MA) model of order *q*, often denoted as MA(q), is a time series model where the current value of the series is modeled as a linear combination of the white noise error terms from the *q* previous time points ("lags"). The mathematical equation for an MA(q) process is:

$$X_t = \mu+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}$$

where X_t is the stock price at time t, μ is the mean of the stock prices, ε_t is the white noise error term at time t, θ_1, θ_2,…, θ_q are the moving average coefficients, and q is the order of the lags. If it is believed that the stock price today is correlated with yesterday's error term only, then the equation for MA(1) would be $X_t = \mu+\varepsilon_t+\theta_1\varepsilon_{t-1}$.

2.3 Autoregressive Moving Average model (ARMA) and with Differencing (ARIMA)

If it is decided that both autoregressive and moving average models are suitable for a group of time series data, then they could be combined to form an autoregressive moving average model (ARMA). ARMA (p, q) is a widely used model in time series analysis for capturing both autoregressive dependencies and moving average noise in a series. The mathematical equation for an ARMA(p, q) process is given by:

$$X_t = c+\phi_1 X_{t-1}+\phi_2 X_{t-2}+\ldots+\phi_p X_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}$$

where all components have the same definitions as in AR and MA models. One constraint for all time series models is that they only work for stationary data, which have constant statistical properties over time, such as a constant mean and variance. If the original data is not stationary (which is usually the case for stock data), then differencing could make it stationary: the ARIMA (p,d,q) model can be fitted on the difference between adjacent terms in the original data. For example, if differencing once then here X_t would equal to stock price changes between day t and day (t-1), which is P_t-P_{t-1}, with ε_t being the error of (P_t-P_{t-1}) and d=1, and if the data are still non-stationary (fairly rare) then difference-of-difference could be calculated as X_t=(P_t-P_{t-1})-(P_{t-1}-P_{t-2}) with d=2 (not P_t-P_{t-2}). The one-step math equation for ARIMA is much more complicated than the previous ones, so it is omitted here, and a two-step equation will be provided as an example when d=1:

$$X_t = P_t - P_{t-1}$$

$$X_t = c+\phi_1 X_{t-1}+\phi_2 X_{t-2}+\ldots+\phi_p X_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}$$

2.4 Seasonal Autoregressive Moving Average Model (SARIMA)

Sometimes time series data entries show seasonality, which is patterns or fluctuations that occur at regular intervals of time (annually, quarterly, monthly, etc.) The data may be not only correlated to its values in the past few days but also its value a month ago. A regular ARMA/ARIMA model may not be able to capture these trends because the number of lags (p, q) involved is far smaller than that required for a month ago (lag=30) or a quarter (lag=90). We can specifically add the relatively distant terms we want into the ARIMA model without including all the irrelevant terms between them and the basic p/q lags to form a Seasonal Autoregressive Moving Average Model, SARIMA(p,d,q)(P,D,Q)s. The (P,D,Q)s part is an ARIMA process by itself intended for the autocorrelation and moving average noise of the seasonal terms X_{t-1s}, X_{t-2s}, X_{t-3s}… only, s being the number of periods in one season while D is the order of seasonal differences (d is daily differences). For example, for daily stock prices the period is equal to one day and if assuming that there is a monthly trend then there are 30 days/periods in a season and s=30. The (P,D,Q) part then focuses on seasonal terms: stock prices P_{t-30}, P_{t-60}, P_{t-90}… The math equation for SARIMA(p,d,q)(P,D,Q)s when d=D=1, s=30 (both differencing once, monthly trend) is:

$$Y_t = P_t - P_{t-1} \text{ (for d=1 in non-seasonal part)}$$

$$X_t = Y_t - Y_{t-30} \text{ (for D=1 and s=30 in the seasonal part)}$$

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} +\ldots+ \phi_p X_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} +\ldots+ \theta_q\varepsilon_{t-q} + \phi_{p+1} X_{t-30} + \phi_{p+2} X_{t-60}+\ldots+\phi_{p+P} X_{t-30P} + \theta_{q+1}\varepsilon_{t-30}+\theta_{q+2}\varepsilon_{t-60}+\ldots+\theta_{q+Q}\varepsilon_{t-30Q}$$

The double differencing process suppresses the constant term. Same as the ARIMA equation, practical analytics work only requires a general understanding of this equation.

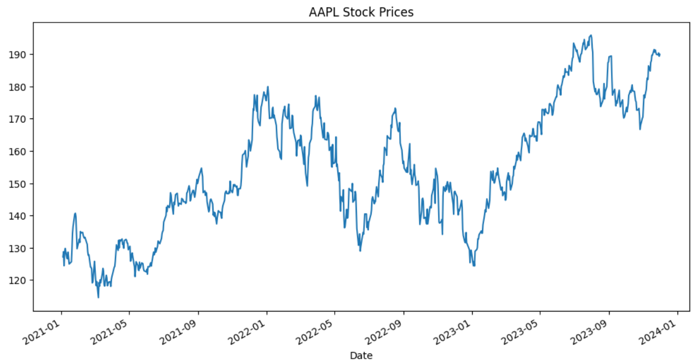
2.5 Data Source and Software

As the topic of this article is a cross-sectional comparison of the effectiveness of the ARIMA and SARIMA models, three sectors are chosen (technology, food manufacture, and energy) and in each sector 10 leading publicly traded companies are selected. The daily closing stock prices of each company in the last 3 years are downloaded from Yahoo Finance and processed with Python and its data analytics libraries: numpy, pandas, scikit-learn, and matplotlib.

In chronical order, the first 80 percent of the data for each company is used for building the ARIMA and SARIMA models while the rest 20 percent is for back testing the goodness of fit of the models.

2.6 A Practical Example

The ARIMA(p,d,q) and the SARIMA(p,d,q)(P,D,Q)s models were developed long ago and are available in Python libraries, so the original work in this article is mainly about choosing the right number of lags and differencing, or in other words most suitable values for p, d, q, P, D, Q and s to achieve the best prediction results. In this article Apple Inc. is used as an example to demonstrate the whole process. Here is a picture of its price changes in the previous three years.



Firstly, decide if the original data is stationary or not. If it is stationary then there is no need for differencing and d=D=0. The Augmented Dickey-Fuller (ADF) test is a statistical test commonly used to determine whether a time series has a unit root. The null hypothesis of the ADF test is that a unit root is present in the time series, indicating non-stationarity. Here is the ADF test result for Apple.

ADF Statistic: -1.6362262728984873

p-value: 0.4642333476157549

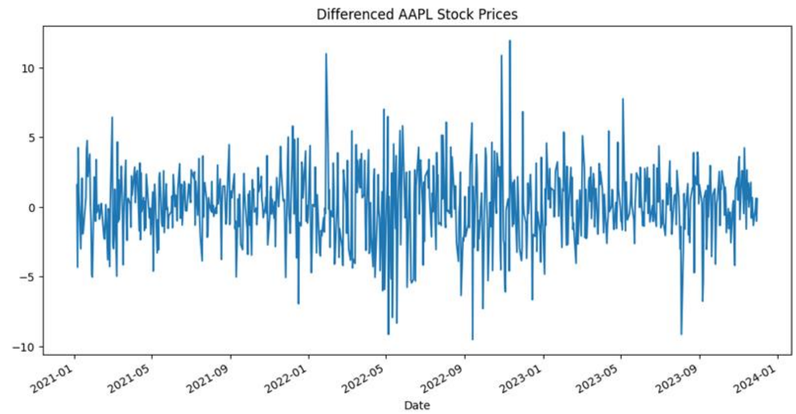
Critical Values:

1%: -3.439314999916068

5%: -2.8654965012008677

10%: -2.5688768817372867

Like most ordinary hypothesis testing scenarios, at 95% confidence level the null hypothesis can be rejected if p-value < 0.05 or ADF Statistic is smaller than 5% critical value. It is clear that in this case the null hypothesis cannot be rejected, thus the original data is non-stationary. Differentiate it once so d=1. Plot the differenced stock prices below. It seems to be stationary.



Next conduct the ADF test again. By rejecting the null hypothesis at a 95% confidence level, the result proves that differencing once is enough to make the data stationary, so d=1.

ADF Statistic: -20.15595859223645

p-value: 0.0

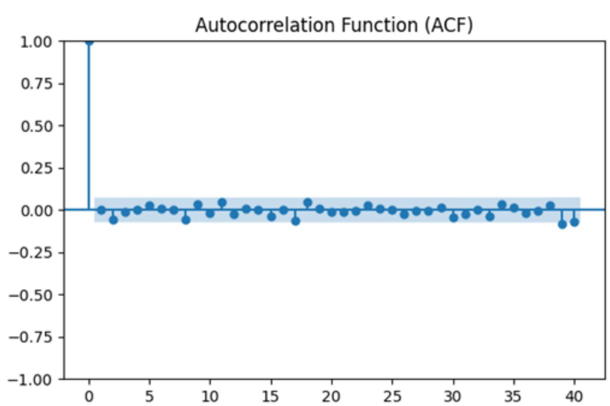
Critical Values:

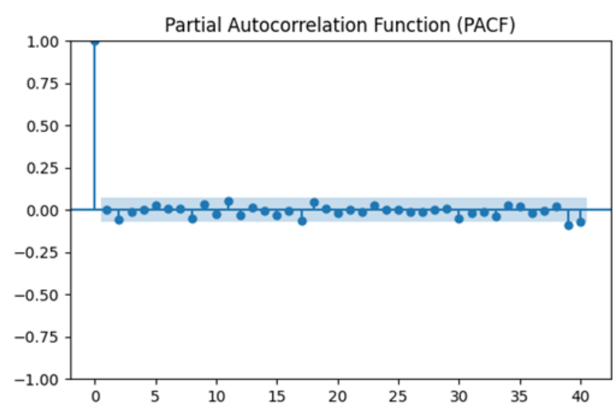
1%: -3.4393396487377155

5%: -2.865507363200066

10%: -2.5688826684180897

After fixing d the following step is to find out the value for p and q in ARIMA(p,d,q). Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) are used for determining p for AR model and q for MA model respectively. They compute the correlation between the value at time t and the previous lags (measured by days in this case).





These two graphs may look very similar but are not completely the same upon closer observation. If the “lollipop” corresponding to the lag is out of the boundary of the blue box then it is said that the correlation is significant. At lag=0 the correlation is always 1 because the data is of course perfectly correlated to itself, so this can be ignored. Starting from lag=1, find the first lollipop that is significant and check if the several lags after it have greatly reduced significance (very short lollipop), if true then the lag of the first lollipop is the p or q number needed, if false then keep going until a lag fulfills the requirement and record its value. In the figure above it can be seen that lag=2 is significant for both ACF and PACF, while lag 3 to 7 are all insignificant, so the conclusion is that p=q=2. An ARIMA(2,1,2) model is completed. Conventionally it is believed that the lag number should be as small as possible since too many lags could cause overfitting for the model, but some researchers found that large lag numbers may be useful in certain situations too, so this article does not impose a limit on the maximum acceptable number of lags. [2]

After finishing the ARIMA part next comes the seasonal (P, D, Q)s part. Similarly, the first step is to determine the order of differencing D. As d=1 in ARIMA, try D=1 first. Also considering the abundance of research discoveries proving that the US stock market tended to perform better in some months than others in some historical periods, [3] it is reasonable to assume that each month is a season so s=30. With D and s fixed an ADF test can be performed on the seasonally differenced data:

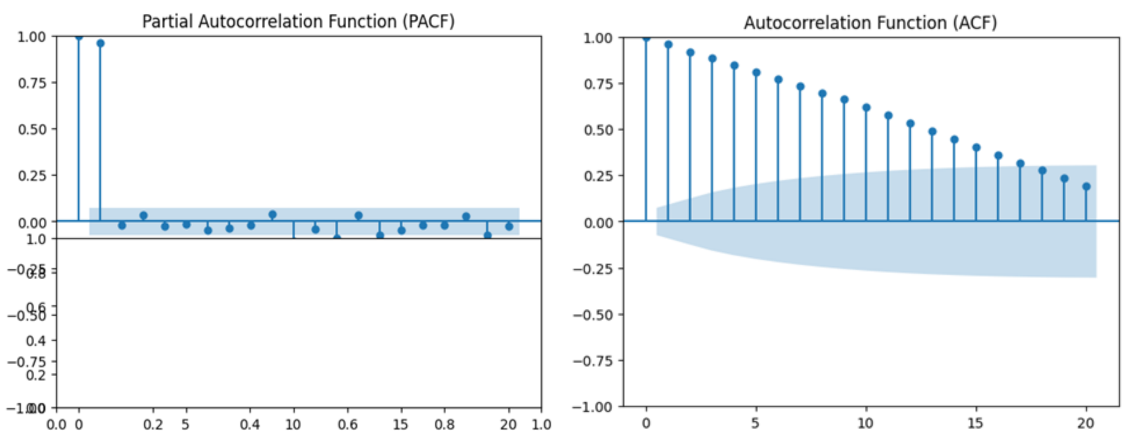
SARIMA:

ADF Statistic: -3.7199444488831817

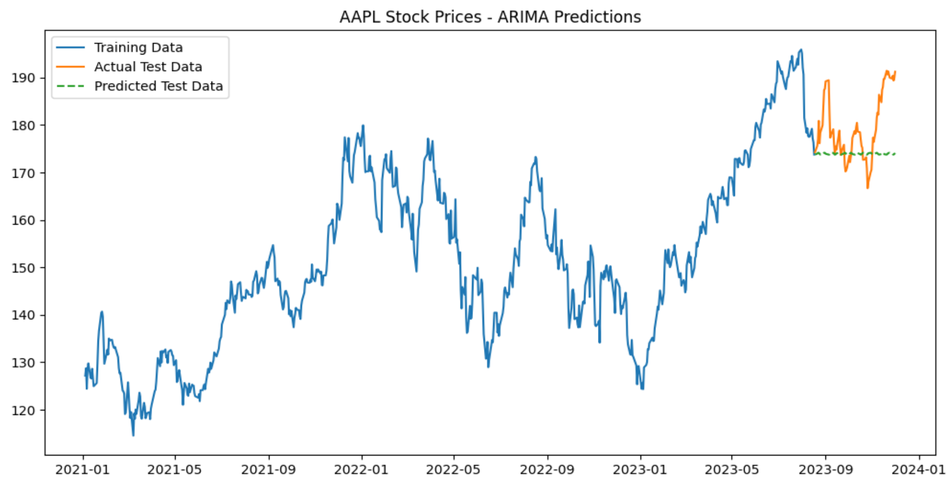
p-value: 0.0038412049403596556

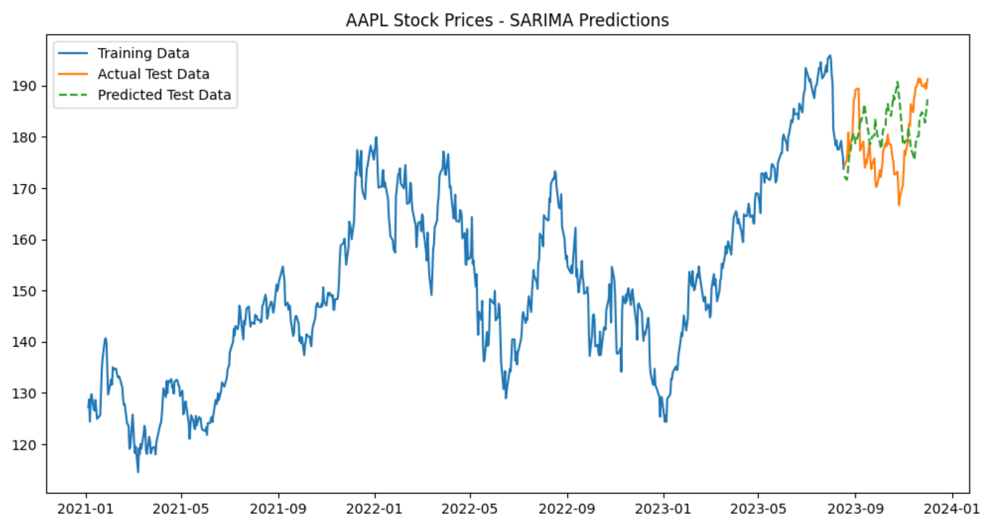
Critical Values: {'1%': -3.4396995339981444, '5%': -2.8656659438580796, '10%': -2.5689671530263554}

The ADF test rejects the null hypothesis and so D=1, s=30 does make Apple stock prices data stationary. Then apply PACF and ACF on the processed data:



The requirements are a bit different here from those of the ARIMA model: According to Robert Nau, usually for the SARIMA seasonal part only one of SAR(p) and SMA(q) will be chosen and the other will be given a lag of 0 and have no effect in the following calculations. The optimal number of lags is 1, so it's either SAR(1) or SMA(1). Nau states that "a pure SAR(1) process has spikes in the ACF at lags s, 2s, 3s, etc., while the PACF cuts off after lag s". [4] This is exactly what the above two graphs depict: the PACF is largely insignificant after lag 1, while the ACF is continuously significant all the way beyond lag 15. Thus a pure SAR(1) process is appropriate here, and combined with ARIMA(2,1,2), the final product is a SARIMA(2,1,2)(1,1,0)$_{30}$ model for Apple Inc. Given the ARIMA and SARIMA models Python can "predict" what the stock prices should be after the period of training data (80% of all data) ends and the back testing period starts. Plot out all predictions on the graphs in contrast to actual data:





It seems at first sight that SARIMA makes a better prediction as it has a zig-zag trend which is similar to the actual test data. However, there are periods when SARIMA goes in the completely opposite direction as the actual data, which might make it worse than the relatively flat line prediction by the ARIMA model. It is best to use objective statistical metrics like R-square and MAPE (mean absolute percentage error) to judge their effectiveness.

2.7 Results Panel

The 3 following graphs demonstrate the models chosen and goodness of fit for the 30 US companies' stocks divided into 3 sectors:

Technology	ARIMA(p,d,q)			Goodness of Fit		SARIMA(P,D,Q) s=30			Goodness of Fit	
Company	AR(p)	d	MA(q)	R-square	MAPE	SAR(P)	D	SMA(Q)	R-square	MAPE
Apple	2	1	2	-0.63	3.62%	1	1	0	-0.62	4.01%
Alphabet	3	1	3	-1.54	4.55%	1	1	0	-1.53	3.93%
Amazon	13	1	13	-0.1	4.56%	1	1	0	-2.74	8.72%
Microsoft	3	1	3	-0.78	5.84%	1	1	0	-0.02	4.72%
Meta	4	1	4	-1.08	5.69%	1	1	0	-3.41	9.64%
Intel	3	1	3	-1.03	9.75%	1	1	0	0.25	5.55%
Cisco	2	1	2	-0.06	3.80%	1	1	0	-1.6	6.16%
Oracle	5	1	5	-0.23	4.97%	1	1	0	-2.89	9.50%
IBM	1	1	1	-0.92	4.12%	1	1	0	-1.04	4.07%
NVIDA	5	1	5	-0.17	5.37%	1	1	0	-7.97	15.48%
Average					5.23%					7.18%

Food Manufacture	ARIMA(p,d,q)			Goodness of Fit		SARIMA(P,D,Q) s=30			Goodness of Fit	
Company	AR(p)	d	MA(q)	R-square	MAPE	SAR(P)	D	SMA(Q)	R-square	MAPE
Nestle	1	1	1	-0.96	3.85%	1	1	0	0.35	2.26%
Mondelez	3	1	3	-1.07	4.37%	1	1	0	-0.26	3.54%
Tyson	3	1	3	-1.42	7.80%	1	1	0	-0.62	6.20%
General Mills	2	1	2	-11.65	9.08%	1	1	0	-9.36	6.19%
Heinz	6	1	6	-0.007	2.41%	1	1	0	-9.95	7.98%
Conagra	2	1	2	-10.04	10.84%	1	1	0	-4.59	5.76%
Campbell	3	1	3	-3.31	4.42%	1	1	0	-21.26	9.61%
Hershey	4	1	4	-3.09	9.32%	1	1	0	-1.08	5.64%
Pepsi	2	1	2	-2.5	6.18%	1	1	0	-0.24	3.06%
Coca-cola	3	1	3	-2.24	5.42%	1	1	0	-0.82	3.87%
Average					6.37%					5.41%

Energy	ARIMA(p,d,q)			Goodness of Fit		SARIMA(P,D,Q) s=30			Goodness of Fit	
Company	AR(p)	d	MA(q)	R-square	MAPE	SAR(P)	D	SMA(Q)	R-square	MAPE
Exxon	2	1	2	-0.56	4.52%	1	1	0	-1.06	5.10%
Chevron	4	1	4	-0.01	5.53%	1	1	0	-0.35	5.82%
ConocoPhillips	11	1	11	-3.88	5.56%	1	1	0	-7.4	6.21%
Schlumberger	5	1	5	-0.02	4.20%	1	1	0	-21.24	19.41%
EOG	12	1	12	0.01	2.62%	1	1	0	-11.6	9.26%
NextEra	1	1	1	-1.33	12.20%	1	1	0	0.15	6.52%
Duke	6	1	6	-0.18	2.34%	1	1	0	-0.36	2.48%
Southern Company	3	1	3	-0.09	2.45%	1	1	0	-2.36	4.18%
Kinder Morgan	5	1	5	-0.16	2.01%	1	1	0	-3.05	4.16%
Phillips 66	2	1	2	-0.85	3.85%	1	1	0	-9.49	9.64%
Average					4.53%					7.28%

For each company, if the ARIMA model performs better both in terms of R-square (a larger R-square) and MAPE (smaller MAPE) than the SARIMA model then a red color block would be added at the end of the row. If the opposite is true then a yellow block is attached. If there is a tie then a blue block is added. It can be seen that SARIMA outperforms ARIMA significantly for the food industry but is far worse than ARIMA in the energy sector. When it comes to the technology industry there is a slight advantage for ARIMA.

3. Discussion and Conclusions

The results from the cross-sector comparisons seem to suggest that the monthly trend is stronger in the food manufacture industry than in the energy industry, while it is unclear whether there is a monthly trend in the tech industry. It can also be observed that the SARIMA model tends to perform very poorly when the number of lags is big, especially for p, q > 4 cases, in which the R-square skyrockets in the negative direction. This might partly explain why large lag numbers are traditionally considered undesirable. In the case of ARIMA models, most firms' stock prices, like that of Apple's, are given a near flat-line prediction, which indicates that there is no clear enough trend in past price changes for ARIMA model to work on and the software can only conclude that the stock is behaving in an almost stochastic way with equal likelihood of up or down, thus $E(P_{t+1}) = P_t$. And for both models the R-square value in most companies' cases are negative, indicating a very poor fit between the model and actual data. This is expected because if stock forecasting is easily achievable with such basic models then everyone can become rich by some simple programming.

It is also worth noting that all the modeling methods and criteria used in this essay are very crude at best and should be treated more as a learning process rather than finished work. For example, when determining the p, d, and q values for ARIMA and SARIMA models, the most rigorous method is to combine results from other statistical tests like AIC and BIC with the ACF and PACF graphs to determine several possible sets of values for (p, q) in each company's case and then use various statistical metrics not limited to R-square and MAPE to find out which set would lead to the best model. This is not feasible given the time constraints of this essay unless using more advanced and complicated black-box algorithms to automate the whole process. And additionally larger R-square and smaller MAPE do not necessarily indicate a better model: the analyst should also use his/her financial knowledge and common sense to decide whether the predictions are "reasonable" or not, which could be viewed as a more qualitative element in the research process.

There are a few changes that are worth trying to improve the prediction results in the future. The first is that instead of feeding the stocks' daily closing prices to the models, replacing these numbers with daily return rate ($\frac{P_t - P_{t-1}}{P_{t-1}}$) might achieve better results and save the time spent in differencing as the return is naturally differenced. Secondly, extra explanatory variables (EEV) could be introduced into the ARIMA and SARIMA models to form ARIMAX or SARIMAX models. These EEVs could be economic indicators like interest rates at time t or dummy variables depending upon the existence of certain conditions (e.g. whether the largest shareholder has a controlling stake in the company, if yes then EEV equals to 1, if not then 0) as an effort of combining fundamental and technical analysis. Thirdly, there are 252 trading days in a year instead of 365 days, so it could be argued that one "month" is equal to 252/12=21 days so s=21 in the SARIMA model. Fourthly, more industries and companies could be included in the research. Last but not least, other seasonal lengths like quarters are also worth exploring because companies usually release their financial statements on a quarterly bases, which contains profit and loss information that could greatly impact their stock prices. Another interesting question which only time can answer is whether such traditional models can still compete with machine learning models.

References

1. Nazário, Rodolfo Toribio Farias, et al. "A literature review of technical analysis on stock markets." *The Quarterly Review of Economics and Finance* 66 (2017): 115-126.

2. Khan, Shakir, and Hela Alghulaiakh. "ARIMA model for accurate time series stocks forecasting." *International Journal of Advanced Computer Science and Applications* 11.7 (2020).

3. Patel, Jayen B. "The January effect anomaly reexamined in stock returns." *Journal of Applied Business Research (JABR)* 32.1 (2016): 317-324.

4. Nau, Robert. Statistical forecasting: notes on regression and time series analysis. https://people.duke.edu/~rnau/411home.htm