ARIMA Models for Forecasting Stock Prices

Athiwat Pathomtajeancharoen, Angelina May, Trevor Lee

Department of Applied Mathematics, University of Washington

December 9, 2022

Abstract

This paper evaluates the autoregressive integrated moving average (ARIMA) model for forecasting stock prices of three companies: Walmart, Disney, and Etsy. Specific ARIMA models that best predict the companies' stock prices are found using Python code and deconstructed into solvable second-order linear homogeneous difference equations. We found that the ARIMA models are unable to accurately predict the exact value or general trend of any of the three companies' stock prices over a two year time period.

Introduction

Forecasting, in the most general context, has always been a point of fascination for society. The ability to essentially tell the future is a superpower that time series analysis gives us. Medical experts, policymakers, engineers, and business analysts are few of the many people that rely on time series analysis for optimal decision making and advancement. Therefore, it is important to identify and understand models for forecasting time series. In 1970, George Box and Gwilym Jenkins proposed an iterative modeling approach to time series methods in forecasting which included identification, estimation, and model checking [Tsay paper]. This development gave rise to interest in the autoregressive integrated moving average model, better written as ARIMA [7]. The model has since been explored extensively in literature for time series prediction and has furthered interest in time series analysis.

To investigate further into ARIMA, we have to first understand the concept of stationarity and differencing. The model assumes that past observations affect new observations and that data is stationary such that the time when data is captured does not matter. A time series is stationary if its statistical properties are constant over time [10]. By computing the differences between successive observations within a non-stationary time series, we can change the time series to become stationary [7, 10]. These time series can be simplified as 'signal' and 'noise' where the signal could be the trend, cycle, or pattern of the graph and the noises are like statically independent errors [8, 10]. An ARIMA model is viewed as a 'filter' that helps us identify 'signal' while ignoring the 'noises' in order to formulate better future predictions [10].

The model consists of three parts, autoregressive terms (AR), moving average terms (MA), and differencing terms (I) that make the times series stationary. The term "autoregressive" refers to the use of a linear combination of past values of a variable of interest to predict a new value of that variable [7]. Similarly, the "moving average" correlates to a linear combination of past error values [7]. The notation ARIMA(p, d, q) is typically used where constant p is the order of the autoregressive part, constant d is the degree of first differencing involved, and constant q is the order of the removing average part. Different values of p, d, and q give special cases of ARIMA models such as random-walk models, autoregressive models, exponential smoothing models, and more [10].

In the text by Hyndman et al., [7] they used ARIMA to forecast quarterly percentage changes in US consumption expenditure. They explained the importance of the value of d on the effect of the long-term

forecasts of the model. Particularly, the greater the value of d, the faster the prediction interval increases in size, meaning there is a wider range of possible values for a given time period. Predicting values on a long-term basis will introduce more error since the prediction interval is wider, suggesting that the performance of the model deteriorates. It is also mentioned that when d=0, the standard deviation of the forecasted data will approach the standard deviation of historical data, resulting in prediction intervals of the same size. In a study done by Mohamed [9], an ARIMA model is identified based on the annual series of inflation rates in Sudan. They emphasized the use of the Box-Jenkins methodology and followed a mixed ARMA process, focusing on the autoregressive and moving average component. Aariyo et al. [1] built a stock price predictive model using ARIMA. They found the best ARIMA model for particular stock data and evaluated performance of the model through comparison of predicted price versus actual stock price. Based on its accuracy for a given month, they emphasized how ARIMA works as an excellent predictor for stock prices on a short-term basis. This is interesting considering what Hyndman et al. describes about the degree of first differencing on long-term behavior. The specific models found in [1] were ARIMA(2, 1, 0) and ARIMA(1, 0, 1) with ARIMA(1, 0, 1) demonstrating the greatest accuracy. This aligns with what was explained in [7] which alternatively put, was that smaller values of d result in slower growth of prediction intervals. This along with short-term forecasting, increases the ability of the model to give accurate predictions.

The success of ARIMA using historical, real-life data underscores the relevance of the model, especially in economics. The goal of this paper is to further explore the ARIMA model for forecasting stock prices. We will use Python to computationally obtain specific cases of ARIMA and transform the models into solvable linear difference equations to understand general behavior and characteristics. By analyzing the properties of the ARIMA model and its ability to accurately predict stock prices, we hope to gain insight into the usefulness of this modeling technique.

Data and Methods

In this study, we investigated the use of ARIMA models in order to forecast stock prices. To do so, we first sourced data from the Nasdaq Stock Market [6]. We obtained 10 years of stock data for Walmart, Disney, and 7 years for Etsy. We then used 80% of the data for training the model to make sure that we are able to learn the patterns in the data without overfiting the data. After retrieving the needed information, we then used Python programming to process the data and fit ARIMA models into different individual stocks.

We utilized the 'auto_arima' function from the pmdarima library, allowing us to automatically select the optimal ARIMA model based on Akaike's Information Criterion (AIC) values. The AIC is a measure of the relative quality of the ARIMA model based on comparing the actual against the predicted distribution of the data, the lower the score the better the model fits the data [2, 9].

Once the ARIMA models were fitted, we used them to make predictions about future stock prices. We then evaluated the accuracy of these predictions and discussed the potential applications and limitations of using ARIMA models in this context.

In this study, we employed numerical analysis to investigate the properties of the fitted ARIMA model. Specifically, we transformed the model into a solvable second-degree homogeneous linear difference equation, enabling us to analyze its behavior and characteristics. This approach provided us with valuable insight into the properties of the ARIMA model and its potential applications in forecasting stock prices [5, 11].

In order to make the analysis of the math model less complicated, we omitted the constant term from all ARIMA models discussed in this study. Although the presence of a constant term would not affect the fundamental arguments of our analysis, its inclusion would add unnecessary complexity to the mathematical details of our study. By excluding it, we are able to focus on the key elements of the model and provide a clear and concise analysis [11].

The general nth order linear difference equation with constant coefficients is a mathematical equation that describes the relationship between the current and past values of a time series [5]. The equation takes the form shown below:

$$C_0Y_t + C_1Y_{t-1} + C_2Y_{t-2} + ... + C_nY_{t-n} = \varepsilon_t$$

Where $C_i, i = 0, 1, 2, ..., n$ are constants. The equation is said to be non-homogeneous if the error term ε_t is non-zero. This means that the relationship between the time series and the error term is not constant, and the equation cannot be solved using methods taught in class. Therefore we need to set ε_t to zero in order for us to solve [5].

Before going any further, we need to introduce the backshift notation, which is a mathematical notation that is used to represent lagged values of a time series. The backshift operator is represented by the letter

B, and it is used to shift the time series one period back in time [7]. The terminology is often defined as:

$$B^k y_t = y_{t-k}$$

With the use of the backshift notation we are able to combine and create

$$C(B) = 1 + C_1 B + C_2 B^2 + ... + C_n B^n$$

$$C(B) Y_t = e_t$$

Time series data can be modeled using a variety of different statistical methods, including autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models. Autoregressive (AR) models are based on the idea that the current value of a time series is related to its past values [7]. An AR(p) model takes the form:

$$Y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where y_t is the current value of the time series at time t, ϕ_1 is the autoregressive coefficient, and ε_t is the error term at time t. Moving average (MA) models are based on the idea that the current value of a time series is related to its past errors [7]. An MA(q) model takes the form:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where y_t is the current value of the time series at time t, θ_1 is the moving average coefficient, and ε_t is the error term at time t [7].

Autoregressive moving average (ARMA) models are a combination of AR and MA models. By combining this with another factor (I) differencing, we obtain a non-seasonal ARIMA model [7]. The model is written below:

$$y'_t = \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Using back shift notation, we can then translate the above equation as

$$\begin{array}{ccccc} (1 - \phi_1 B - \cdots - \phi_p B^p) & (1 - B)^d y_t & = & (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t \\ \uparrow & \uparrow & & \uparrow \\ \text{AR}(p) & d \text{ differences} & & \text{MA}(q) \end{array}$$

By rearranging the equation as follow, we can obtain the 2nd order homogeneous linear difference equation by expanding the $(MA)^{-1}$ term

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)(1 + \theta_1 B + \cdots + \theta_q B^q)^{-1} y_t = \varepsilon_t = 0$$

The resulting solvable second-order homogeneneous linear difference equation:

$$Y_t + (\theta_1 - \psi_1) Y_{t-1} + (\theta_1^2 + \theta_p - \psi_p \theta_1 - \theta_1 + \psi_p) Y_{t-2} = \varepsilon_t = 0$$

By putting in the coefficient we generated from the python library, we then will be able to solve this

equation. For example, let $a = (\theta_1 - \psi_1)$ and $b = (\theta_1^2 + \theta_p - \psi_p\theta_1 - \theta_1 + \psi_p)$

$$Y_t + aY_{t-1} + bY_{t-2} = 0$$

$$Y_{t+2} + aY_{t+1} + bY_t = 0$$

By transforming $Y_t = \lambda^t$, we then are able to isolate the λ without the t value by:

$$\lambda^{t+2} + a\lambda^{t+1} + b\lambda^t = 0$$

$$\lambda^2 + a\lambda + b = 0$$

Once we have transformed the ARIMA model, we can solve for the eigenvalues which are a key characteristic of the model, as they determine the behavior of the time series prediction [5]. Then we would be able to obtain and solve the initial value problem for the linear difference equation. Therefore we can observe the behavior and characteristics of the model.

Results

We used time series data on Walmart stock (WMT) to fit an ARIMA(1,1,0) model. This model was selected as the optimal model based on its AIC values, and it allowed us to make predictions about the future evolution of Walmart's stock price. Our results suggest that this ARIMA(1,1,0) model is well-suited for forecasting WMT stock prices, but further analysis is needed to confirm its effectiveness in practice [2].

The mathematical formulation of the ARIMA(1,1,0) model is as follows:

AR(1) : $y_t = \psi y_{t-1} + e_t$

y_t = 1st diff of $y = y_t - y_{t-1}$

$$y_t = \psi B y_t + \varepsilon_t$$

$$(1 - \psi B)y_t = \varepsilon_t$$

By substituting in $y_t = (1 - B)Y_t$

$$(1 - \psi B)(1 - B)Y_t = \varepsilon_t$$

$$(1 - B - \psi B + \psi B^2)Y_t = \varepsilon_t$$

$$Y_t - (1 - \psi)Y_{t-1} + \psi Y_{t-2} = \varepsilon_t = 0$$

$$Y_{t+2} - (1 - \psi)Y_{t+1} + \psi Y_t = 0$$

Applying the transformation of form: $Y_n = \lambda^t$

$$\lambda^{t+2} - (1-\psi)\lambda^{t+1} + \psi\lambda^t = 0$$

$$\lambda^2 - (1-\psi)\lambda^1 + \psi = 0$$

$$(\lambda - 1)(\lambda - \psi) = 0$$

$$\lambda_1 = 1,\ \lambda_2 = \psi$$

The coefficient of parameters of this model was found as $\psi = -0.1188$

$$\lambda_1 = 1$$

$$\lambda_2 = -0.1188$$

Therefore we can conclude that

$$Y_t = C_1\lambda_1^t + C_2\lambda_2^t$$

$$Y_t = C_1 + C_2(-0.1188)^t$$

When $t_0 = 147$ and $t_1 = 145.95$

$$147 = C_1 + C_2(-0.1188)^0$$

$$145.95 = C_1 + C_2(-0.1188)^1$$

Solving for $C_1 = 149.196$ and $C_2 = 2.404$ gets us:

$$Y_t = 149.196 + 2.404(-0.9422)^t$$

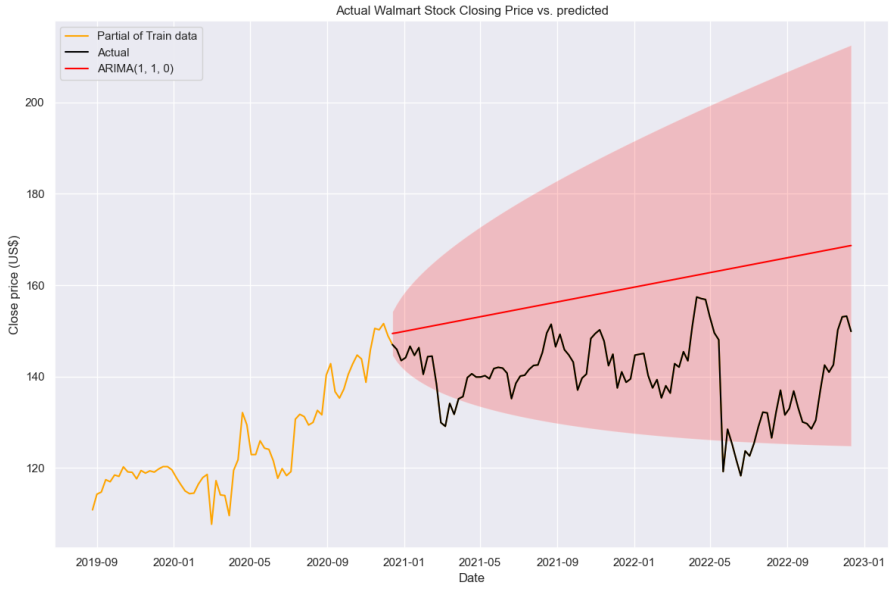

Figure 1: Graph of Walmart actual stock price vs prediction

The graph above shows a comparison between the actual stock price of Walmart and the predicted stock price based on an ARIMA model along with a 95% confidence interval. The x-axis of the figure represents

time, while the y-axis shows the stock price. The black line in the figure represents the actual stock price of Walmart, while the red line represents the predicted stock price based on the ARIMA model. The red area surrounding the red line represents the 95% confidence interval for the predicted stock price. The equation that we obtained earlier by solving the second-degree order linear difference equation is represented by the red line in the figure. As you can see for the case of Walmart stock, the majority of the predicted values for Walmart stock are above the actual prices. This indicates that the model is overestimating the prices of Walmart stock.

To show another example, we applied the ARIMA model to the Disney stock prices and obtained the ARIMA(0,1,2) model. Below is the mathematics calculation of the equation to predict Disney stock prices.

MA(2) : $y_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + e_t$

$$y_t = (1-B)Y_t$$

$$y_t = (1-\theta_1 B - \theta_2 B^2)\varepsilon_t$$

$$(1-B)Y_t = (1-\theta_1 B - \theta_2 B^2)\varepsilon_t$$

$$(1-B)(1-\theta_1 B - \theta_2 B^2)^{-1} Y_t = \varepsilon_t$$

We can expand $(1-\theta_1 B - \theta_2 B^2)^{-1}$ to $(1+\theta_1 B + (\theta_1^2+\theta_2)B^2 + ...)$

$$(1-B)(1+\theta_1 B + (\theta_1^2+\theta_2)B^2 + ...)Y_t = \varepsilon_t$$

$$Y_t(1+\theta_1 B + (\theta_1^2+\theta_2)B^2 - B - \theta_1 B^2 - (\theta_1^2+\theta_2)B^3 + ...) = \varepsilon_t$$

$$Y_t(1+(\theta_1-1)B + ((\theta_1^2+\theta_2)-\theta_1)B^2 - (\theta_1^2+\theta_2)B^3 + ...) = \varepsilon_t$$

$$Y_t + (\theta_1-1)Y_{t-1} + (\theta_1^2+\theta_2-\theta_1)Y_{t-2} + ...) = \varepsilon_t$$

We then get 2nd order homogeneous linear difference equation:

$$Y_t + (\theta_1-1)Y_{t-1} + (\theta_1^2+\theta_2-\theta_1)Y_{t-2} = 0$$

$$Y_{t+2} + (\theta_1-1)Y_{t-1} + (\theta_1^2+\theta_2-\theta_1)Y_t = 0$$

Applying the transformation of form $Y_n = \lambda^t$

$$\lambda^{t+2} + (\theta_1-1)\lambda^{t+1} + (\theta_1^2+\theta_2-\theta_1)\lambda^t = 0$$

$$\lambda^2 + (\theta_1-1)\lambda^1 + (\theta_1^2+\theta_2-\theta_1) = 0$$

$$\lambda_{1,2} = \frac{-(\theta_1-1) \pm \sqrt{(\theta_1-1)^2 - 4(\theta_1^2+\theta_2-\theta_1)}}{2}$$

The coefficient of parameters of this model was found as $\theta_1 = -0.0019$ and $\theta_2 = 0.1108$

$$\lambda_{1,2} = \frac{-(-0.0019 - 1) \pm \sqrt{(-0.0019 - 1)^2 - 4((-0.0019)^2 + 0.1108 + 0.0019)}}{2}$$

$$\lambda_{1,2} = 0.50095 \pm 0.93035i$$

Therefore we can conclude that for complex roots:

$$Y_t = r^t(C_1 \cos(\theta t) + C_2 \sin(\theta t))$$

$$Y_t = 1.0566^t(C_1 \cos(1.0769t) + C_2 \sin(1.0769t))$$

When $t_0 = 147.13$ and $t_1 = 154.14$

$$147.13 = 1.0566^0(C_1 cos(0) + C_2 sin(0))$$

$$154.14 = 1.0566(C_1 cos(1.0769) + C_2 sin(1.0769))$$

Solving for $C_1 = 147.13$ and $C_2 = 86.468$ gets us:

$$Y_t = 1.0566^t(147.13 \cos(1.0769t) + 86.468 \sin(1.0769t))$$

Figure 2: Graph of Disney actual stock price vs prediction

As shown in the figure, the ARIMA model was able to predict the initial upward trend of Disney stock prices. However it was not able to capture the subsequent downward trend in the prices. This shows the limitation of ARIMA model and how it doesn't do well as time goes on. There are moments where the actual stock prices falls outside the confidence interval indicating that either the stock is outperforming the prediction or the prediction is simply wasn't able to predict accurately.

Finally, we applied the ARIMA model to the Etsy stock prices and obtained the ARIMA(1,1,2) model.

Below is the mathematical calculation of the equation to predict the price of ETSY stock.

$$y_t = y_{t-1} + \psi y_{t-1} - \psi y_{t-2} + \varepsilon_t - \theta_1 \varepsilon_{t-2} - \theta_2 \varepsilon_{t-2}$$

$$y_t - y_{t-1} - \psi y_{t-1} + \psi y_{t-2} = \varepsilon_t - \theta_1 \varepsilon_{t-2} - \theta_2 \varepsilon_{t-2}$$

translating the first part into:

$$y_t - By_t + \psi By_t - \psi B^2 y_t$$

$$y_t(1-B)(1-\psi B) = e^t(1 - \theta_1 B - \theta_2 B^2)$$

$$y_t(1-B)(1-\psi B)(1 - \theta_1 B - \theta_2 B^2)^{-1} = \varepsilon^t$$

$$Y_t(1-B)(1-\psi B)(1 + \theta_1 B + (\theta_1^2 + \theta_2)B^2 + ...) = \varepsilon_t$$

$$Y_t(1-B)(1 + \theta_1 B + (\theta_1^2 + \theta_2)B^2 - \psi B - \psi\theta_1 B^2 - \psi(\theta_1^2 + \theta_2)B^3 + ...) = \varepsilon_t$$

$$Y_t(1 + \theta_1 B + (\theta_1^2 + \theta_2)B^2 - \psi B - \psi\theta_1 B^2 -$$

$$\psi(\theta_1^2 + \theta_2)B^3 - B - \theta_1 B^2 - (\theta_1^2 + \theta_2)B^3 +^2 + \psi\theta_1 B^3 + \psi(\theta_1^2 + \theta_2)B^4 + ...) = \varepsilon_t$$

$$Y_t(1 + (\theta_1 - \psi - 1)B + ((\theta_1^2 + \theta_2) - \psi\theta_1 - \theta_1 + \psi)B^2 + ...) = \varepsilon_t$$

We then get 2nd order homogeneous linear diff equation:

$$Y_t + (\theta_1 - \psi 1)Y_{t-1} + (\theta_1^2 + \theta_2 - \psi\theta_1 - \theta_1 + \psi)Y_{t-2} = \varepsilon_t = 0$$

$$Y_{t+2} + (\theta_1 - \psi 1)Y_{t+1} + (\theta_1^2 + \theta_2 - \psi\theta_1 - \theta_1 + \psi)Y_t = \varepsilon_t = 0$$

Applying the transformation of form $Y_n = \lambda^t$

$$\lambda^{t+2} + (\theta_1 - \psi 1)\lambda^{t+1} + (\theta_1^2 + \theta_2 - \psi\theta_1 - \theta_1 + \psi)\lambda^t = \varepsilon_t = 0$$

$$\lambda^2 + (\theta_1 - \psi 1)\lambda^1 + (\theta_1^2 + \theta_2 - \psi\theta_1 - \theta_1 + \psi) = \varepsilon_t = 0$$

The coefficient of parameters of this model was found as $\psi = -0.6718$, $\theta_1 = 0.6968$ and $\theta_2 = -0.1254$

$$\lambda^2 + 0.3686\lambda - 0.5405 = 0$$

$$\lambda_{1,2} = \frac{-0.3686 \pm \sqrt{(0.3686)^2 - 4(-0.5405)}}{2}$$

$$\lambda_1 = 0.5736,\ \lambda_2 = -0.9422$$

The solution will be in a form of $Y_t = C_1\lambda_1^t + C_2\lambda_2^t$:

$$Y_t = C_1(0.5736)^t + C_2(-0.9422)^t$$

When $t_0 = 164.18$ and $t_1 = 165.82$

$$164.18 = C_1(0.5736)^0 + C_2(-0.9422)^0$$

$$165.82 = C_1(0.5736)^1 + C_2(-0.9422)^1$$

Solving for $C_1 = 211.446$ and $C_2 = -47.266$ gets us:

$$Y_t = 211.446(0.5736)^t - 47.266(-0.9422)^t$$

Figure 3: Graph of Etsy actual stock price vs prediction

Similarly to Disney stocks, as shown in the figure, the ARIMA model was able to predict the initial upward trend of ETSY stock prices. However it was not able to capture the subsequent downward trend in the prices. there is a significant increase in the stock prices at the beginning of the time period. However, the ARIMA model wasn't able to capture that showing that the model may have limitation in its ability to handle highly volatile data. This indicates the inflexibility of the model over time.

ARIMA Model Performance Metrics				
Stock	Model	AIC	RMSE	RMAE
WMT	(1,1,0)	1933.895	21.857	4.357
DIS	(0,1,2)	2208.107	46.462	6.147
ETSY	(1,1,2)	2074.956	72.584	8.043

The table above presents the performance metrics of three ARIMA models that were applied to three different stocks: WMT, DIS, and ETSY. The metrics shown in the table include the AIC values of the models, the root mean square error (RMSE), and the root mean absolute error (RMAE). Based on the values shown in the table, we can first see that the AIC values of the models vary greatly: the (1,1,0) model for WMT has the lowest AIC value, and in contrast, the (0,1,2) model for DIS has the highest AIC value. Therefore, this result highlights how the (1,1,0) model is the most effective, as it has the lowest AIC value. This outcome also implies that the (1,1,0) model is best in terms of fitting the data and explaining the underlying processes that generate the stock prices [7].Furthermore, the RMSE and RMAE values of each

model differ significantly. The (1,1,2) model for ETSY has the highest RMSE and RMAE values whereas the (1,1,0) model for WMT has the lowest RMSE and RMAE values. These results indicate that the (1,1,2) model has the largest average error in terms of predicting the stock prices, and therefore may be less reliable than the other models. These results also reinforces how the (1,1,0) mode is the most efficient model out of the three due to its small values for error [7]. In all, these findings strongly suggest that the (1,1,0) model for WMT is the most effective of the three models in terms of fitting the data and making accurate predictions, contrary to the (1,1,2) model for ETSY where it has the highest average error implying how it is the least effective model and has less reliability [7].

There are several sources of error in the results obtained from applying ARIMA models to predict stock prices in our study. The primary ones are measurement error and model misspecification. Measurement error occurs when the data used to fit the ARIMA model is inaccurate. For example, stock prices may be measured with missing data points within the time series. This can lead to errors in the model fitting and prediction process, which can affect the accuracy of the results. Another source error is model misspecification. This error occurs when the ARIMA model does not accurately factor in the hidden processes that generate the stock prices. For instance, the model may not include all the relevant variables, such as seasonality or wrong lag/differencing order. As a result, this can lead to inaccuracies in the predictions. These sources of error and uncertainty can be quantified using statistical measures, such as the mean squared error or the mean absolute error. These measures provide an estimate of the average error in the model predictions. However, it is important to note that these measures do not account for all sources of error and uncertainty, and additional analysis may be needed to fully understand the limitations of the results. It is also important to note that sampling error was not considered in this study due to the large size of our data sources. These sources are sufficiently large enough to accurately represent the entire population of stock prices, and therefore sampling error is not expected to have a significant impact on the results.

The results of this study indicate that the ARIMA model has limitations that need to be researched further. These limitations include the inflexibility of the model to adapt over time and the inability to handle highly volatile data. However it is also important to note that the results show that the ARIMA model is generally able to capture the initial trend of the data accurately. This suggests that the model have some value in forecasting future prices, but it is not sufficient on its own to provide accurate predictions.

Discussion

Figures 1, 2, and 3 show how each ARIMA model predicts both the general trend of a company's stock price (shown by the solid red line) and a 95% confidence interval where it believes the stock price will stay within (shown by the red area surrounding said line). Despite using the stock's price from the past ten years (or seven years for Etsy, as it went public in 2015), the ARIMA models still largely failed in accurately predict the general trend of each company's stock price over the course of two years. This is in line with the conclusion of [1], that ARIMA models ultimately fail in predicting long-term stock prices.

It should be noted that the stock market has seen extreme fluctuation these past years; the start of the pandemic saw many technologies and entertainment companies rise in stock price, and the reopening of the economy lead to large amounts of inflation and a subsequent rise in interest rates, causing those same companies to fall in stock prices. This led to all of the models having an upward trajectory (as the testing data started after the pandemic but before the rise in inflation) and the actual stock price's falling. However, this information does not excuse the poor performance of the ARIMA models as much as it highlights the inherent flaw in using the models to predict stock prices, which is that ARIMA models do not take into account exogenous variables at all, and instead solely use past information of the stock to predict its future.

While this method may work in other contexts, the stock price of a company can fluctuate wildly on many outside factors, from public perception to sales reports to the overall health of the economy. Although these factors come into play in the stock's past price, their influence is too random both in timing and magnitude to be accurately deciphered by the ARIMA model, and thus we are left with an inaccurate prediction line and a wide confidence interval that still does not contain all of the stock's actual price.

Still, despite this fairly poor performance, there may be more value in using the ARIMA model to predict stock prices on a shorter scale. Knowing the general trend of a stock over a month, for example, would still be valuable, and the odds of a large spike or drop in stock in any given month are generally low. For longer time periods, it is worth finding how a Seasonal ARIMA, or SARIMA, performs in forecasting stock prices. While this would not fix the issue of stocks being heavily affected by outside forces, there is a cyclical up-and-down nature to the economy that could be captured in a SARIMA model, but not an ARIMA model.

Conclusions

In this paper, we studied how well ARIMA models could predict a company's stock price over a period of two years. To do this, we used Python code, along with the past ten years (or seven, for Etsy) of Etsy's, Disney's, and Walmart's stock prices to obtain ARIMA models that would then predict those company's stock prices roughly two years into the future. We also converted those models into solvable linear difference equations to get a better understanding of the models, as well as their accuracy in predicting stock prices. Firstly, we found that the optimal ARIMA model for one company was not the same for another; this could be for many reasons, including that the three companies vary greatly in size, location, and structure. We also found that even with many years of data, the best ARIMA models for each company still failed to accurately predict the overall trend of the company's stock over a two year period.

This leads us to the conclusion that a sole ARIMA model would not aid in predicting which stocks will rise or fall over a span of multiple years, most likely due to the fact that ARIMA models do not take into account extraneous variables, while stock prices are affected much more by those variables, as opposed to past stock prices. Our study was affected by the economic turbulence from the Covid-19 Pandemic and subsequent reopening of the economy, which may have been better encapsulated by SARIMA, as economic fluctuations, while certainly not perfectly predictable, do work on a roughly cyclical basis.

Thus, a look into how SARIMA models predict stock prices over several years may be more fruitful. The paper "Stock Price Prediction Using Sarima and Prophet Machine Learning Model" [13] researches just this, along with other machine learning models capable of predicting future values. It should also be noted that ARIMA models may have better results in predicting a company's stock price in a shorter period of time, when there would be less economic change and therefore a stock's price is much more likely to be heavily influenced by its past price.

References

[1] A. Adebiyi et al., "Stock Price Prediction Using the ARIMA Model," IEEE (2014)

[2] D. Argarwal, "Time Series Part 3 - Stock Price prediction using ARIMA model with Python," LinkedIn (2022)

[3] P. Bartlett, "Introduction to Time Series Analysis. Lecture 7," Berkeley

[4] J. Brownlee, "How to Create an ARIMA Model for Time Series Forecasting in Python," Machine Learning Mastery (2017)

[5] S. Celik, "Linear Difference Equation Applications on Time Series Models," Quest Journals (2021)

[6] "Etsy, Inc. Common Stock," Nasdaq (2022)

[7] R. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," OTexts (2018)

[8] R. Kaiser and A. Maravall, "NOTES ON TIME SERIES ANALYSIS ARIMA MODELS AND SIGNAL EXTRACTION," Working Studios (2000)

[9] E. Mohamed, "FORECASTING (ARIMA) MODELS TO INFLATION RATE IN SUDAN," IJDR (2015)

[10] R. Nau, "Statistical Forecasting: notes on regression and time series analysis," Duke University (2020)

[11] R. Nau, "The mathematical structure of ARIMA models," Duke University (2014)

[12] R. Tsay, "Time Series and Forecasting: Brief History and Future Research," Taylor Francis (2000)

[13] A. Vishwakarma, "Stock Price Prediction Using Sarima and Prophet Machine Learning Model," IJARSCT (2020)

Appendix

Python code use to create the ARIMA model: https://www.github.com/earthathiwat/Stock_ARIMA