Stock Price Forecasting Using Time-Series and Regression Models

University of Toronto: Faculty of Applied Science and Engineering

Jonah Ernest

1.0 Historical Price Analysis

The team focused on the 2022 stock prices for Exxon Mobile Corp (XOM) [1]. The following sections describe the team's historical price analysis of the chosen stock prices including a discussion of the relevant patterns, trends, and seasonality with supporting statistical methods and external research. From this analysis, two different forecasting methods are chosen.

1.1 Discussion of Patterns, Trends, and Seasonality

In order to assess any patterns, trends, and seasonality in the data, the team began by observing the stock prices graphed over time between January 2022 and December 2022, shown in Figure 1. As seen, there is a spike in June, which is thought to be due to the Russia-Ukraine war. Although the war began in late February, The New York Times reported that Russia began selling less oil due to sanctions imposed by the European Union and United States around the time of this spike [2]. Since Russia is typically a major natural gas producer, the reduced supply led to a jump in prices which was reflected in the stock prices in June. The team found that similar spikes were observed in the stock prices for other oil and natural gas companies such as Shell PLC (SHEL) [3] and TotalEnergies SE (TTE) [4].

Figure 1. XOM Closing Stock Price Throughout 2022

The team did not observe any seasonal patterns from the data, and so ANOVA was used to assess both weekly and monthly seasonality. The following hypotheses were explored:

Weekly seasonality	H_0: $\mu_{day\,1} = \mu_{day\,2} = \ldots = \mu_{day\,5}$ H_1: $at\ least\ one\ treatment\ mean\ differs\ from\ another$ where $\mu_{day\,n}$ is the average price for the nth weekday over all weeks
Monthly seasonality	H_0: $\mu_{day\,1} = \mu_{day\,2} = \ldots = \mu_{day\,31}$ H_1: $at\ least\ one\ treatment\ mean\ differs\ from\ another$ where $\mu_{day\,n}$ is the average price for the nth day of the month over all months

The results of the ANOVA test for both weekly and monthly seasonality failed to reject the null hypothesis at a 95% significance level, meaning there is not significant evidence to suggest that the treatment means are different (see Exhibit 1). Thus assuming there is no seasonality in the data.

From empirical observation, there is an upwards trend in the stock price over the year. It was found that the data had a linear trend defined by the equation $y = 0.0989x - 4332.2$ with an R^2 value of 0.7957. To determine whether the trend was statistically significant, the team carried out an ANOVA test on the null hypothesis that there is no correlation between the date and the closing price. The p-value for the parameters of the regression line was found to be less than $\alpha = 0.05$, leading to a rejection of the null hypothesis (see Exhibit 1). Therefore, moving forward, it is assumed that there is a positive trend in the data.

1.2 Identifying Appropriate Forecasting Methods

Based on the team's findings about trends, seasonality, and patterns it was determined that Holt's Method (List A) and linear regression (List B) models would be most suitable to forecast the first ten days of 2023 for XOM. Holt's Method, also known as double exponential smoothing, was chosen as it takes into account linear trends and does not have a seasonality

factor making it ideal for the data as it has a trend but no seasonality. This method also allowed the team to define the importance of more recent data through parameterization; since the natural gas industry is highly sensitive to external factors, it is important that the model values more recent data. The linear regression method was chosen for similar reasons. Linear regression can be applied to data that does not display seasonality, like that of Exxon. It is also used to predict future values in data that experience trends, making it one of the favored forecasting methods. Linear regression also allowed the team to change the features in the model to identify which set of available features would result in the most accurate prediction. Furthermore, this method would allow the team to incorporate other prices and causal features besides the closing price, allowing for a more representative forecast and in-depth sensitivity analysis.

2.0 Future Price Forecasting

The team forecasted the closing price for XOM for the first ten days of 2023 using two different methods: Double exponential smoothing (using Holt's method), and a linear regression model. The performance of each method will then be analyzed while comparing it to the actual closing prices for these dates, and future recommendations and improvements will be suggested.

2.1 List A: Double Exponential Smoothing (Holt's Method)

Double exponential smoothing using Holt's method was first used to predict XOM's closing stock prices for the first ten days of 2023. It should also be noted that there is no need for any additional add-on methods to handle potential seasonality since it was proven to be statistically insignificant. See the Exhibit 2 Excel file for the full model that was used to perform this method. To start, all 251 of the 2022 closing prices were collected from Yahoo Finance [1] and put into the "Double_Exp_Smoothing_XOM" sheet in Exhibit 2 under the D_t column which represents the demand (closing price) at the end of day t. The other columns are created, where F_t

is the forecasted closing price for the time period t, S_t is the "true" series at time t, and G_t is the gradient at time t. To forecast D_t for the first ten days of 2023, the team started forecasting starting from the first closing price of 2022 (on 2022/01/03) until the tenth day of 2023.

To begin this process, the value of G_1 was initialized to zero, and S_1 and D_1 were initialized to the closing price on the first day of 2022. After initializing the first day's forecast, true series, and gradient, the team utilized the following formulae to solve for F_t, S_t, and G_t for t = 1,...,251 (for each of the 251 days in 2022).

$$S_t = \alpha D_t + (1 - \alpha) F_t$$

$$G_t = \beta(S_t - S_{t-1}) + (1 - \beta) G_{t-1}$$

$$F_{t+1} = S_t + G_t$$

After completing the forecasting for the year of 2022, the team then performed multi-step ahead forecasts to predict the closing prices of XOM for the first ten days of 2023 using the following formula, $F_{251+\tau} = S_{251} + \tau \cdot G_{251}$, where $\tau \in [0,10]$ and is the number of days into 2023. So for example, to forecast the closing price for the third day of 2023, $\tau = 3$.

In addition to the forecasting, the error was calculated for each date by subtracting the forecast and demand for that day. The mean absolute deviation (MAD) was also found using the average of the absolute value of these individual errors.

Sensitivity analysis was performed through parameterization of α and β to determine their optimal values which would minimize error in the model. Excel solver was used to minimize the MAD obtained from the 2022 closing prices with the constraints that both parameters had to be between 0 and 1 (inclusive). This resulted in α and β values of approximately 0.988 and 0.02. From these optimal values, it can be seen that since α is almost one, more recent data greatly contributes much more to the current closing price of the stock when compared to older data.

This means that decreasing this parameter would result in more error due to an overemphasis being placed on older data. Additionally, since β is quite small, the model is placing less emphasis on the recent trend, but instead is more concerned with the cumulative trend over time.

Exhibit 2 shows the results of the forecasting along with the errors, sensitivity analysis, and chart comparing the demand, forecasts, and trend for all of 2022 and 2023 respectively.

2.2 List B: Linear Regression

Linear regression was used to predict the closing prices of the first 10 days in January 2023. Features in the data included the date, opening price, closing price, adjusted closing price, volume of shares, and the highest and lowest prices of the stock on that day. By incorporating the volume of shares as an additional feature, the team ensured the inclusion of a variable that was neither a stock price nor a unit of time. Using Python's sklearn library, the team constructed three different linear regression models to identify which set of features would allow for the most accurate forecast (see Exhibit 3).

The first linear regression model was constructed with a training set composed of all data from 2022 and a testing set composed of the data of the same features for the first 10 days of 2023. The training set was used to train a linear regression model and included the highest price, lowest price, volume, and open price. The column for adjusted closing price was removed due to its strong correlation with the actual closing price. The testing set contained only the close price feature which is what the team wanted to forecast. The outputs of this model include the forecasted closing prices, as well as the training and testing score of the models. The forecasted closing prices allowed the team to compare the accuracy of the model by comparing the number to the true closing prices, and to find the magnitude in deviation from the given 2023 data using the mean absolute error. The second model was formulated in the same way with the only

difference being the features incorporated in the training model. The team removed the low and high price features leaving the training set with only the volume and the opening price since these values would be unknown before the end of the day. The MAD of this forecasting model was then calculated for future comparison with the first model.

In order to further investigate the linear regression discussed above the team conducted a sensitivity analysis that resulted in a third model where the closing price of CNOOC Ltd. [5] (in HKD), a natural gas company based in Hong Kong, was added as a feature to the training set. CNOOC Ltd. was chosen because the HKSE market closes before the NYSE market opens (due to the time difference), meaning that the CNOOC closing price will be known when forecasting the XOM closing price. This change was made in order to observe if there is a relationship between global natural gas prices, and whether having the data of another company's closing price would increase the accuracy of our model.

2.3 Performance and Analysis of the Forecasting Methods

When comparing the linear regression forecasts to the double exponential smoothing forecast, it can be seen through their respective MAD values that the first linear regression model is the most accurate. Since all three models will be compared to the same set of true closing prices, the resulting MAD calculations can be compared to identify the most accurate forecast. With a MAD of ~0.48, the linear regression model is just over three times more accurate than the double exponential smoothing model, which has a MAD of 1.57. This error metric allowed the team to identify how far off the forecast of each model was relative to the true closing prices.

As mentioned earlier, the Russia Ukraine war impacted the outcome of the forecasting models. It is believed that the upwards trend in the closing price is due to the decreased natural gas supply from Russia, as a result of the sanctions introduced in the spring of 2022. As shown in

Figure 2, the upwards trend did not exist until 2022. Therefore, if the forecast was based on a dataset beyond 2022, the accuracy may have been negatively impacted.

Figure 2. XOM Closing Prices from 2018-2023 [1]

The team recommends using the first linear regression model over the double exponential smoothing model due to its accuracy, versatility, and incorporation of causal factors into the forecasting. As discussed above, the linear regression model has a lower MAD, meaning that it outperforms double exponential smoothing with its accuracy. Besides its quantitative advantages, this model is much more versatile in terms of its applications and considerations of external factors. Due to its ability to add any causal event (given appropriate data) as a new feature, if a potential investor were to anticipate an external factor impacting the business or industry, they would be able to forecast the new stock prices using an additional feature which would take this new factor into consideration. Overall, the team feels confident in suggesting an investment into XOM using this linear regression model due to the increasing trend in the stock prices (with a slope of 0.0989, as seen in section 1.1), which has been proven to be significantly significant, and also because the forecasted closing prices for early 2023 are set to follow this upwards trend with a very small mean absolute error. With this in mind, it is still important to remember that the XOM's stock prices can rapidly change (as seen in the past year) due to environmental and political external factors which impact the gas and oil industry.

2.4 Improving the Model for Future Stock Predictions

To improve the performance of the linear regression model, other external features could have been explored since it is a causal model, such as consumption and production rates of natural gas, however for the scope of this project, the team was unable to retrieve such data for the 2022-2023 timeframe. If this additional data was obtained and was found to be telling of the true closing price, indicated by a high absolute value of the linear regression coefficients (β), then it would be expected that the forecast would have higher accuracy. Additionally, the linear regression model could be further improved by reducing overfitting which leads to high performance on the training set and poorer performance on the testing set. Overfitting may be caused by having a small dataset or from multi-collinearity which occurs when many of the features are highly correlated with each other. Since the current model is trained on data from 2022 only (with 250 data points), it may be less telling than a model trained on data from multiple years. That being said, since the oil and natural gas market is highly sensitive to external factors, using more recent data may actually lead to a more accurate forecast. Due to the small number of features in the models, removing highly correlated features was not feasible; however, if more features were added, highly correlated features should be removed which would likely improve the testing score, thus improving the forecast's accuracy.

To improve the double exponential smoothing model, the time period of the forecasting in 2022 can be altered to change the impact of older data. For instance, as shown in the sensitivity analysis in Section 2.1, the model minimizes error with a large value of α, therefore further sensitivity analysis could be conducted by exploring if shorter time periods from 2022 can be used in the forecasting model for 2023.

3.0 References

[1] “Exxon Mobil Corporation (XOM) stock historical prices & data,” *Yahoo! Finance*, 18-Feb-2023. [Online]. Available: https://finance.yahoo.com/quote/XOM/history?p=XOM. [Accessed: 17-Feb-2023].

[2] E. Koeze and C. Krauss, “Why gas prices are so high,” *The New York Times*, 14-Jun-2022. [Online]. Available: https://www.nytimes.com/interactive/2022/06/14/business/gas-prices.html. [Accessed: 17-Feb-2023].

[3] “Shell plc (Shel) Stock Historical Prices & Data,” *Yahoo! Finance*, 18-Feb-2023. [Online]. Available: https://finance.yahoo.com/quote/SHEL/history?p=SHEL. [Accessed: 17-Feb-2023].

[4] “Totalenergies SE (TTE) stock historical prices & data,” *Yahoo! Finance*, 18-Feb-2023. [Online]. Available: https://finance.yahoo.com/quote/TTE/history?p=TTE. [Accessed: 17-Feb-2023].

[5] “CNOOC Limited (0883.HK) stock historical prices & data,” *Yahoo! Finance*, 18-Feb-2023. [Online]. Available: https://finance.yahoo.com/quote/0883.HK/history?p=0883.HK. [Accessed: 17-Feb-2023].