STOCK PRICE ANALYSIS WITH ESM AND ARIMA

Mitali Bharali
Yihan Gao
Arpit Shrivastava
Anirudha Tatavarthy
Luo Yang

Executive Summary

The financial services industry has adopted big data and predictive analytics in a wide manner, and it has helped online traders to make great investment decisions that would generate consistent returns. With rapid changes in the stock market, investors have access to a lot of data. In this paper, we analyze the stock market's ups and downs. For this analysis, we used data provided by the course and used Yahoo Finance API to extend the data from 2005 through 2020. We mainly used predictive analytical techniques such as ESM and ARIMA for our analysis, we will adopt these techniques for each five-chosen pharmaceutical stocks and compare the outcome in terms of suitability and accuracy.

Data Description and Preprocessing

There are 5 datasets that contain 3769 records each totaling to 18845 records between the years 2005 and 2020. The data is presented in CSV format including variables Date, Open, High, Low, Close, Volume and OpenInt, in our analysis, we will use Close, the stock price at close time, as our variable.

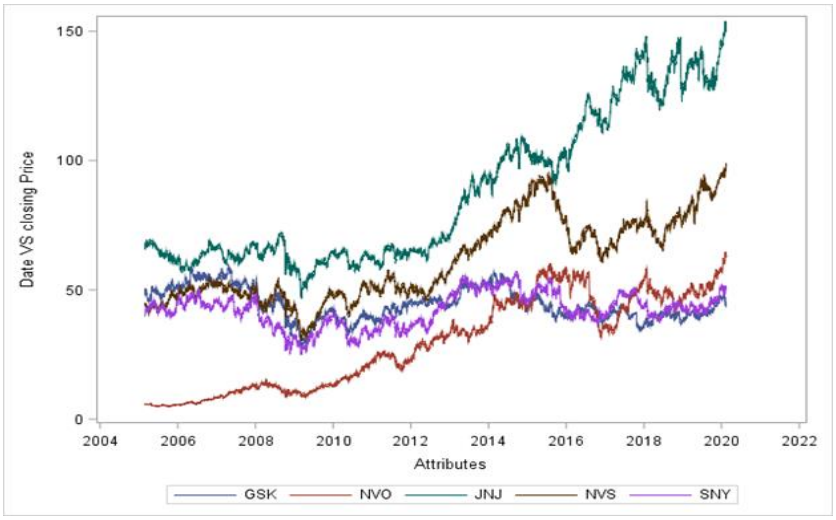
No Preprocessing of the data was required since the data was not unstructured. Considering the characteristics of stock price data, data on weekends is missing. In our analysis, we will omit these missing values and regard the whole dataset as consecutive. Furthermore, in order to keep consistency and timeliness of our time series analysis, we have requested a more completed dataset from Yahoo Finance API and extended our end date, hence our time series analysis started from 25th February 2005 to 14th February 2020.

Data Exploration

Trend

Graph 1.1 shows the movement closing price of each stock with respect to date. Here we can observe that NVO, NVY, JNJ are having almost the same pattern of graphs which represents that they are potentially correlated as one goes up the other two goes up at the same time.

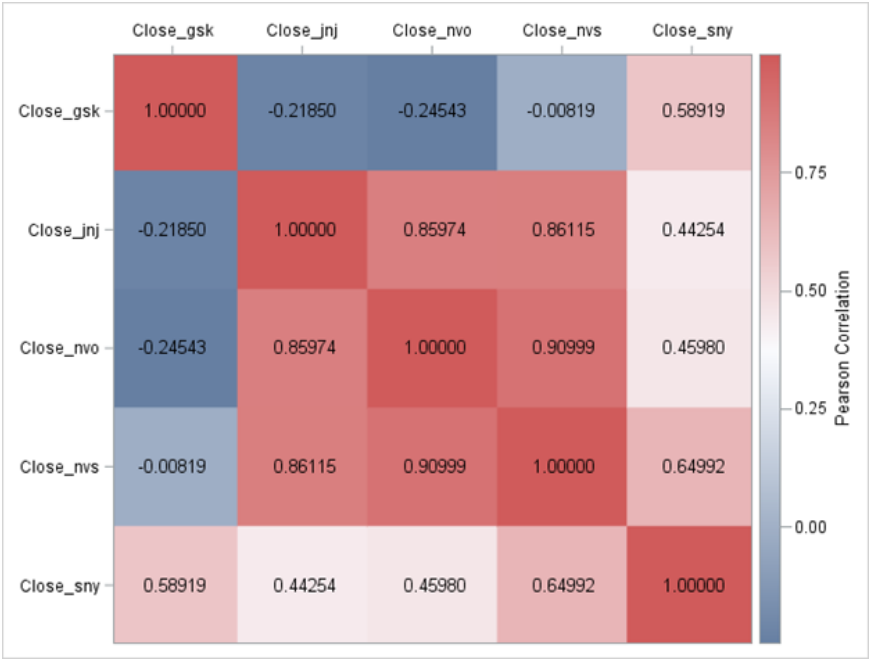
Graph 1.1



Among those five stocks, three of them obviously have an upward trend, while the other two, GSK and SNY, are much steadier across the whole time.

Correlation

From Graph 1.2, we tried to understand correlation between the closing price of five stocks of Pharma companies. It shows that the stocks NVO and NVS are highly positively correlated with a value of .9099, stock JNJ and NVS have a high positive



Graph 1.2

correlation of .86115 and stock NVO and JNJ have a high positive correlation of .85974. These stocks are highly correlated with each other as they belong to the Pharma Sector. This shows that the closing price for the pharma sector is correlated for a particular day as one increase other also increases and as one decreases the other also decreases. Generally, in stock analysis, having the combination of highly correlated stocks will bring volatility into this portfolio, but in our case, we will not regard all the stock as a whole.

Mode Application

Exponential Smoothing

Model Selection

Exponential smoothing forecasting methods are similar in that a prediction is a weighted sum of past observations, but the model explicitly uses an exponentially decreasing weight for past observations. There is essentially 3 broad categories of Exponential Smoothing:

Simple Exponential Smoothing: With no trend and seasonality;

Hyperparamters: Alpha;

SYNTAX: model=SIMPLE;

Double Exponential Smoothing: With Trend and no seasonality;

Hyperparameters: Alpha and Beta;

SYNTAX: model=DOUBLE;

Triple Exponential Smoothing: With Trend and Seasonality;

Hyperparameters: Alpha, Beta and Gamma;

SYNTAX: model=WINTERS;

In order to select the best ESM out of the three, we will run Winters Method Parameter Estimates to examine the weight of trend and seasonality (if p value of the weight is less than alpha, that parameter weight is not significant). Table 1.3 shows the outcome of each dataset based on 0.05 confidence level. According to the result, Simple Exponential Smooth will be applied to two of the datasets and Triple will be applied to the rest. Details are provided in Table 1.4.

Parameters	GSK	JNJ	SNY	NVO	NVS
Level	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**
Trend	0.9434	0.3105	0.3608	0.9514	0.2632
Seasonal	0.0724	**0.0018**	**0.0008**	0.4880	**0.0376**

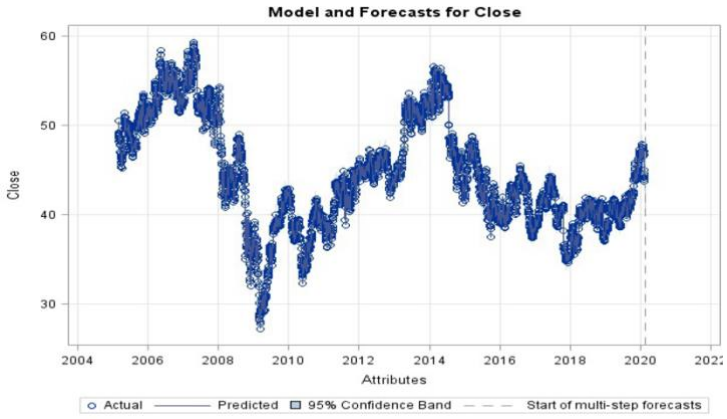
Table 1.3

Model	Datasets
Simple	GSK, NVO
Triple	JNJ, SNY, NVS

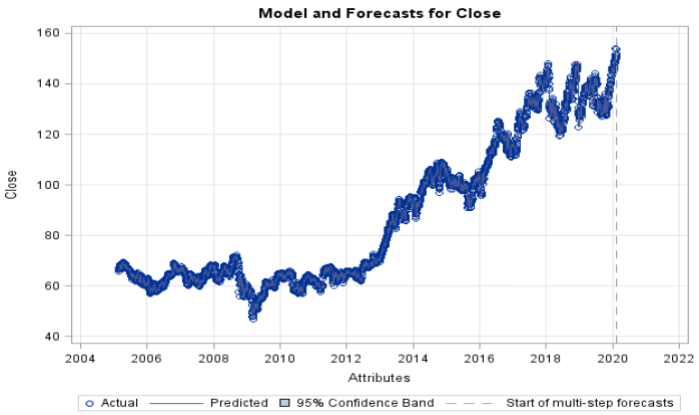
Table 1.4.

Prediction and Forecast

The best model for the dataset is then applied to plot the entire lifetime of the data. Based on Graph 1.5 and 1.6, we can conclude that both simple and triple models fit the

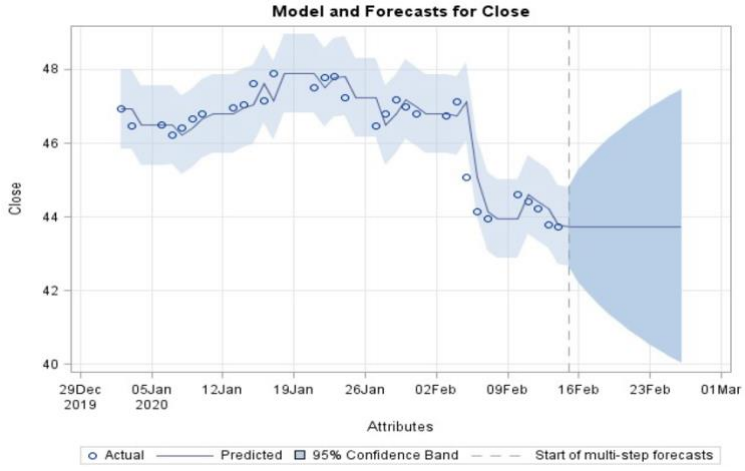


Graph 1.5 How simple model fits GSK

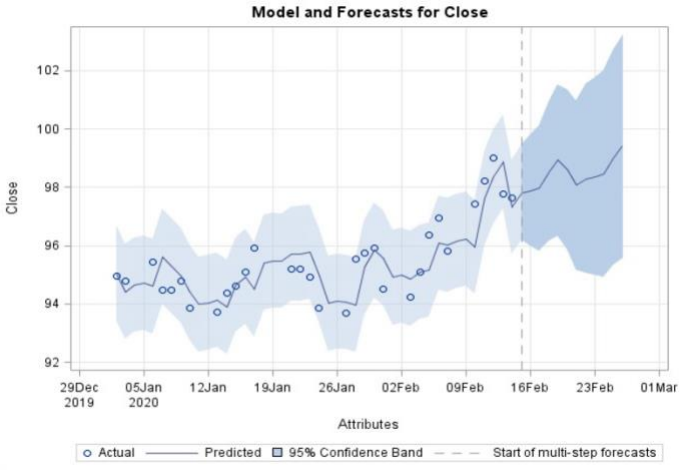


Graph 1.6 How triple/Winter model fits NVO

data really well, capturing most of the fluctuations. Graph 1.7 and 1.8 demonstrate the difference of predicted (straight line)and actual values (circles) alongside the range of 95% confidence interval around the predictions. This is then applied to forecast the data point for the year 2020 with an additional lead of 12 days, all of them labelled in the similar fashion. The Winters model captures the trend (upward, downward) and



Graph 1.7 Residual and forecast plot for GSK



Graph 1.8 Residual and forecast plot for NVO

seasonality of the data which is clearly visible from its forecast of 12 additional days, while the forecast of the Simple ESM model clearly indicates constant value with no trend or seasonality in the data.

As for metrics of model accuracy, we have decided to choose Sum of Square Errors to measure the errors the model made for each dataset, SSE score detail for each respective model can be referred from table 1.9. In the last chapter of the report, we will use SSE as a criterion to compare ESM models amd ARIMA models.

Dataset	Model	SSE Score
GSK	Simple ESM	1160.30
JNJ	Triple/Winters	1083.82
SNY	Triple/Winters	3255.27
NVO	Simple ESM	1937.83
NVS	Triple/Winters	1561.15

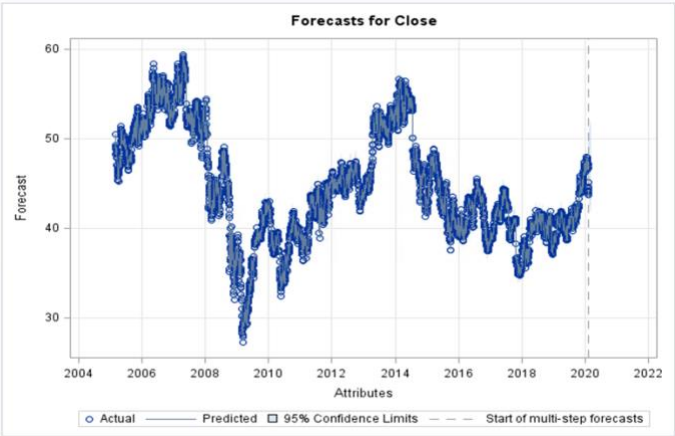
Table 1.9 SSE Score of each model

Auto Regressive Integrated Moving Average

Model Estimate and Selection

Auto Regressive Integrated Moving Average (ARIMA) is a powerful model for forecasting time series data and most people have used it to predict stock trend. We will apply ARIMA model to our five pharmaceutical stocks to find their best fit ARIMA model and to check how well the best model fit the data. The model we will apply is non-seasonal ARIMA, as we accept the fact that there rarely exists seasonality in pharmaceutical stocks. Any ARIMA model involves three parameters. That is, p, d, and q where p stands for lags of auto regression, d stands for level of differencing and q stands for lags of moving average. We will use AIC score to identify the optimal parameters for our final model.

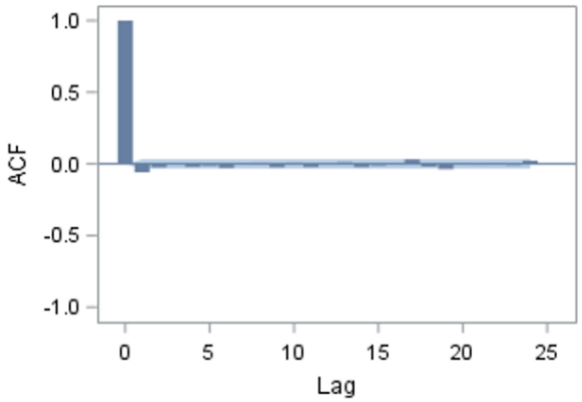
The requisite prior to estimate any ARIMA model is to have a dataset with stationary series, which is usually not the case of stock price data. By differencing our data with lag one as it is enough for majority of data, we managed to bring all five datasets stationary based on Dickey-Fuller test results and to identify two white noise series, which in this case the best ARIMA model will be (0, 1, 0).That is, a random walk model where $y(t)=y(t-1)+\varepsilon(t)$. This has been proven after we ran a several model and found out ARIMA (0, 1, 0) yields the lowest AIC score. Graph 2.0 shows the fitness of ARIMA models on GSV dataset. As for another three stocks, GSK, SNY and JNJ, we need to



Graph 2.0 How ARIMA (0,1,0) fits GSV

Graph 2.1 How ARIMA (1,1,1) fits NVO

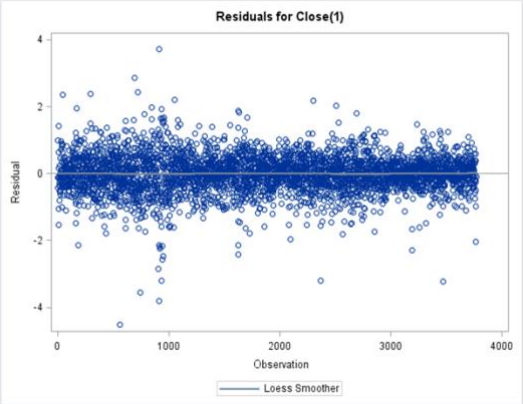
move forward to determine p and q by observing spikes on Autocorrelated Function (ACF) plot and Partial Autocorrelated Function (PACF). For GSK and JNJ, the parameters that yield the lowest AIC are found by referencing directly from the plot, while SNY encounters an exception in which plots suggests p=1 due to a sudden decrease after spike lag 1, as shown on Graph 2.2, however, the model with p = 2 generates a lower AIC score.



Graph 2.2 Spike at lag = 1 on SNY

Model Estimate and Selection

Based on model selection, we ran ARIMA with lowest AIC on each of the dataset. The residual of all the models are white noises according to Ljung-Box test, Graph 2.3 shows randomness of residual from GSK ARIMA (1, 1, 1), and in general, the distributions of residual tend to be normal, hence we can conclude that ARIMA models ran for these datasets are good models.



Graph 2.3 Residual plot of ARIMA (1,1,1) on GSK

In order to compare ARIMA with ESM, we have recorded SSE as a metrics for accuracy, more detailed can be found in Table 2.4 below.

Stock	Best ARIMA	AIC	SSE
GSK	(1, 1, 1)	6352.1	1158.21
SNY	(2, 1, 1)	7357.9	1551.6
JNJ	(2, 1, 2)	10117.7	3225.7
NVS	(0, 1, 0)	8181.9	1083.11
NVO	(0, 1, 0)	5997.5	1933.9

Table 2.4 ARIMA models and their AIC and SSE

ESM and ARIMA Comparison

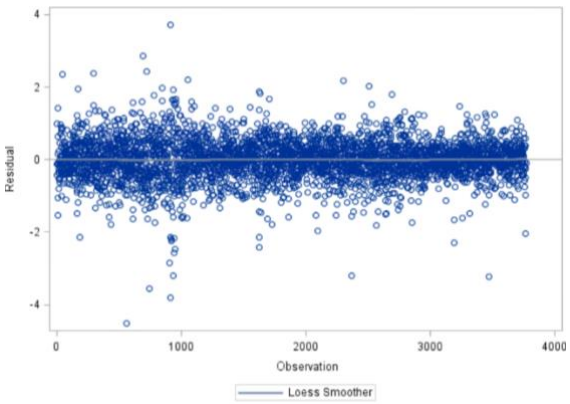
After going through parameter selection and model comparison within different versions of ARIMA and ESM, here we are going to compare the performance metrics of the best ARIMA and ESM models in each data set and decide which one performs better.

Though ARIMA and ESM are both very widely used time series models, there are times one is preferred over another. ARIMA has many advantages over ESM, as the AR and MA parts capture both autoregressive components and moving average components. Also, ARIMA models are generally more flexible than ESM. In fact, some ESM, such as simple exponential smoothing model is a special case of ARIMA (0,0,1). However, when a time series demonstrates non-stationary traits, ARIMA is generally not preferred unless procedures like differencing can be done before the modeling to ensure a stationery model. In our analysis, as we know that all our 5 models demonstrate different patterns of trends and seasonality, we applied both models on all the datasets to get a more wholesome comparison of the two models. As for the performance metric, we choose Sum of Squared Error (SSE). Details can be found in Table 2.5.

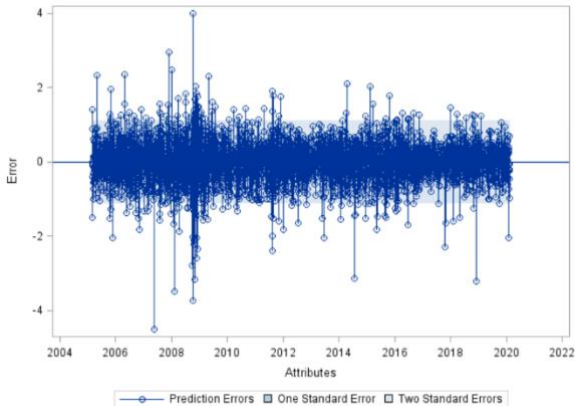
Dataset	ARIMA (SSE)	ESM(SSE)	Better Model
GSK	(1 ,1, 1) (1158.21)	Simple (1160.3)	ARIMA (1, 1, 1)
JNJ	(2, 1, 2) (3225.7)	Winters (3255.27)	ARIMA (2, 1, 2)
SNY	(2, 1, 1) (1551.6)	Winters (1561.15)	ARIMA (2, 1, 1)
NVS	(0, 1, 0) (1083.11)	Winters (1937.83)	ARIMA (0, 1, 0)
NVO	(0, 1, 0) (1933.9)	Simple (1083.82)	Simple

Table 2.5 ARIMA and ESM comparison

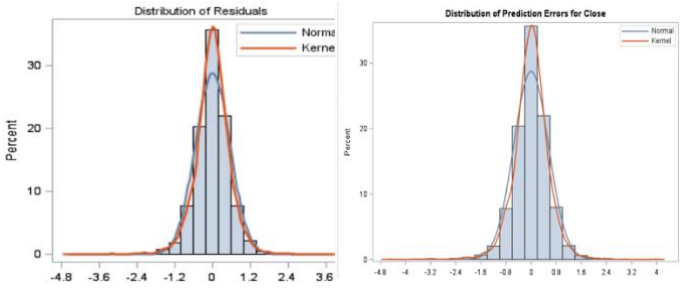
From the table, we can conclude that ARIMA performs slightly better than ESM models yielding lower SSE on GSK, JNJ, SNY and NVS datasets. However. ESM dramatically outperforms ARIMA on the NVO data set. Besides SSE, we also compared residual plots. The residuals in time series models are what is left over after fitting the models. A good forecasting model's residual would appear to be not correlated and have a mean of zero. Its residuals plot would also appear normally distributed and shows constant variance. We took GSK out of our datasets as an example, ARIMA and ESM had a very close SSE. Graph 2.6 and 2.7 demonstrated residual plot of ARIMA and ESM respectively. Both plots show a randomly distributed residual with constant



Graph 2.6 ARIMA (1,1,1) residual plot GSV



Graph 2.7 Simple ESM residual plot GSV



Graph 2.8 GSK Residual normality (ESM on the right)

mean of zero with similarities. In Graph 2.8, we can observe the normality of distribution of residuals. From this graph, both ARIMA and ESM show a very similar normal distribution, indicating that they are both good forecasting models.

Conclusion

In this study, we compared ARIMA and ESM on the same datasets, some of which both models perform similarly, and on other data one model dramatically outperform the other. When selecting a model in real case, there are more criterions to considerate. One might need to involve more metrics such as Mean Absolute Error or Mean Absolute Percentage Error. For suggestions for our further analysis, we can bring these metrics into analysis, also, we can possibly take advantages of data transformation such as logarithm to then compare the performance of the models.

Reference

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.Accessed on 04/23/2020

How is Big Data Analytics Used for Stock Market Trading? (2019, May 18). Retrieved April 21, 2020, from https://blog.imarticus.org/how-is-big-data-analytics-used-for-stock-market-trading-data-analytics-blog/

The NYSE and NASDAQ: How They Work. (2020, March 23). Retrieved April 21, 2020, from https://www.investopedia.com/articles/basics/03/103103.asp/