Using a time series model, can we effectively forecast the company’s revenue?

By:

Kevin Sandoval

**A1.**

I attempted to re-clone the D603 Machine Learning repository, but since there isn’t a separate one from Task 1,2 and 3, it did not work. Thus, I am using the same repo from Task 1 but creating additional branches. The branches in Task 3 will be in the format “Task3\_Part#\_Number\_#\_Update#” and Task 1 will be in the format of “Task1\_Part#\_Number\_#\_Update#”.

**A2.**

Each part of D has their own committed branch on the repo.

**A3.**

A link to the GitLab repository is provided in the “Comments to Evaluator” section.

**A4.**

A .txt file containing the history of the activity from the repo is provided in the submission.

**B1.**

The research questions I will be examining is: “Using a time series model, can we effectively forecast the company’s revenue?” This is relevant to a real-world organization as having some idea on how the company is doing and will do in the future can be used when presenting to shareholders. It can provide useful insight to if the company is growing and if not, could indicate that some additional analysis might help turn profits around.

**B2.**

The goal(s) of the analysis will be to create time series model that can help predict future revenue for the company. This model will account for stationarity, autocorrelated data, and seasonality. It will then be plotted, and the future revenue projections will be added on to the plot and clearly defined.

**C1.**

“Stationarity is an important concept in time series analysis. Loosely speaking, it means that the statistical properties of the process remain constant over time. In practice, this implies that all values of the process are comparable, no matter at what time they were observed. In turn, comparability of the observations allows us to draw statistical conclusions about the whole process.” Palma, W. (2016). *Time Series Analysis.* Wiley Global Research (STMS). “However, ordinary regression models do not account for dependence between values in different periods, which in cross‐sectional data is assumed to be absent. Yet, in the time series context, values in neighboring periods tend to be correlated. Such correlation, called autocorrelation, is informative and can help in improving forecasts. If we know that a high value tends to be followed by high values (positive autocorrelation), then we can use that to adjust forecasts.” Shmueli, G., Bruce, P. C., Gedeck, P., Yahav, I., & Patel, N. R. (2023). *Machine Learning for Business Analytics (2nd ed.)*. Wiley Global Research (STMS). This gives some insight into the assumption of autocorrelation. This assumption is that a positive autocorrelation is likely to remain positive (commodities and currencies) while a negative autocorrelation is likely to remain negative (individual stocks). Negative autocorrelation of a stock is “mean reverting” while a positive autocorrelation of a stock is “trend following”

**D1.**

The line graph visualizing the realization of the time series is shown below.

A graph on a computer screen

AI-generated content may be incorrect.

**D2.**

The format of the realization is in days. Specifically starting at Day 1, and counting to Day 731, which is the length of the sequence as well. There are no actual dates in the data, just a number representing the number of days. This would imply that observations are recorded once per day. I did not find any inconsistencies in the data. I checked for missing values, duplicates, and the number of unique values in the Day column, which should equal the length of the sequence, which it does. The total time span covered is 731 days, which is approximately 2 years. The code and output for these calculations is shown in a screenshot below.

**A screenshot of a computer

AI-generated content may be incorrect.**

**D3.**

Looking at the initial line graph for the time series, we can see that the overall mean is increasing, which means that the time series is non-stationary. From checking the autocorrelation plot of revenue, we can see that the autocorrelation is positive but decreasing slowly over time. This is another indication that the time series is not stationary. The last check for stationarity is to perform the Augmented Dickey-Fuller test. From the output we can see the p-value to be 0.32 which is not less than 0.05 thus we fail to reject the null and can assume that the series is non-stationary. To edit this, we can create a new data frame taking the difference of the previous revenues and recheck the plots and autocorrelation. After doing so, the line graph looks stationary. The line graph looks to have no trend or seasonality, and the ACF plot looks normal.

**D4.**

The first step when I import a dataset is to check for missing values and any duplicates. This is partly shown in Part D2, but I observed no missing values, and no duplicates. Another part of the data cleaning process, I chose to use the ‘Day’ column to create a new index for the data frame that starts on January 1st, 2023. Thus, the data’s dates will include January 1st, 2022, until December 31st, 2023. The training and test splits were created using the original data frame. I checked to see what 80% of the length of the data frame was, which was approximately 512. I then set the first 512 values in the data frame to be the training data and the remaining values to be the test data. The files are included in the submission.

**D5.**

Prepared CSV is included in the submission.

**E1.**

1. A graph showing the growth of a company

   AI-generated content may be incorrect.  
   From the initial line graph shown above, we can see that there does not appear to be any seasonality. However, there does appear to be an upwards trend for the time series which would be cause for non-stationary.
2. A graph with blue dots

   AI-generated content may be incorrect.  
   From this plot of the autocorrelation function above, we can see a steady decline at each lag. This is another clue into the time series being non-stationary.
3. A graph showing a line

   AI-generated content may be incorrect.  
   From the spectral density graph, there does not appear to be any cycles, or cyclical behavior. Thus, I do not believe there is seasonality in the time series, however the graph suggests that the series is not stationary. The differenced data time series spectral density plot is shown below and appears to be stationary.  
   A graph showing a number of blue lines

   AI-generated content may be incorrect.
4. A screenshot of a graph

   AI-generated content may be incorrect.  
   Using the decomposition function for time series and the “additive” method, we can see here that the entire series was taken as the trend component and there was no seasonality found in the time series.
5. From the graph above, we can also see that there is no trend in the residuals of the decomposed series.

**E2.**

To create the ARIMA model, I needed to find the parameters p, d, and q. To do this, I will look at the PACF plot of the differenced data for p, the ACF plot of the differenced data for q, and 1 for d since I only differenced the data once. The plots are shown below.   
A screen shot of a graph

AI-generated content may be incorrect.  
From this PACF graph, we can see that we have 1 lag well outside of the interval, so the value for p will be 1.   
A screen shot of a graph

AI-generated content may be incorrect.  
From this ACF graph, we can see that we have 2 lags that are well outside of the interval. Thus, the value of q in the model will be 2.

The final model will then be ARIMA(1,1,2) and the observed trend is accounted for because I chose those values based on the stationary differenced data. There was no observed seasonality in the data set. The results of the model are shown below.   
A screenshot of a computer

AI-generated content may be incorrect.  
Since these parameters fail the Ljung-Box statistic, thus implying that the residuals are correlated, I will try to edit the model. Instead of using the first lag for p, I will try 0 instead. And instead of using the second lag for q, I will try using the first, or a value of 1. Thus the new model would be ARIMA(0,1,1). The results of this model are shown below.   
A screenshot of a computer

AI-generated content may be incorrect.  
The AIC is slightly increase but the model now fails to reject the null for the Ljung-Box Statistic so we can assume that the residuals are not correlated. This is the model I will choose.

**E3.**For Part E3, the code for creating the forecast is shown in screenshots below. Code snippets partly taken from Dr. Elleh’s PowerPoint lecture “D213 Task 1 Cohort Webinar PPT”

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

**E4.**

The code used for the calculations has been provided throughout the write up. Some extra code that has not been included is below. The output of the forecast created in Part E3 using the results of the ARIMA model created in Part E2 is shown in a screenshot below.

A graph with a line graph and numbers

AI-generated content may be incorrect.

In the screenshot below, I included the code to calculate the predicted mean of a 6 month interval after the conclusion of the test data set.

A screenshot of a computer

AI-generated content may be incorrect.

**F1.**

1. For choosing the ARIMA model, I decided to do a by-hand approach. I used the PACF, number of times differenced, and ACF to get the values of p, d, and q. respectively. I initially found that they were 1,1,2. However, that model failed the Ljung-Box test, so I decided to try 0,1,1 instead based on the plots. This resulted in a much successful Ljung-Box test score.
2. The prediction interval for the forecast is 1 day. Since our time series data is on daily 2-year revenue, the ARIMA model predicts revenue at a by-day interval. (Dr. Elleh, “D213 Task 1 Cohort Webinar PPT”)
3. I decided to choose 180 days, or roughly 6 months as the forecast length. I believed this could potentially provide valuable insight for the company. The tradeoff is that the model starts to struggle the longer the forecast length. Long term predictions require significantly more historical data to maintain accuracy.
4. The AIC score plays an important role in the selection of the best model. The lower the AIC score, the better the model. I also calculated the Mean Absolute Error, as well as the mean squared error. The code and output for those values are shown below.  
   A screenshot of a computer program

   AI-generated content may be incorrect.

**F2.**

The annotated visualization with correct labeling showing the forecast of the final model compared to the test set is shown below. As a note, since I trained my initial ARIMA model with the training set, I had to create a new ARIMA model with the test set to plot. The training set did not span the test set, as the test set represented the last 20% of data. Since I already found the optimal values for p,d, and q, I chose to re-run the model using the test data to create this display. If I created the initial model on the initial data set rather than the cut training set, I would not have had to do this step.

**A graph showing a line graph

AI-generated content may be incorrect.**

**F3.**

From the forecast graph of using the training data to predict the test data, we can see that the model struggles to give an accurate window. This is largely due to the data not having seasonality, which makes future values extremely difficult to predict. As a result, I can’t say that this model effectively forecasts company revenue. Thus, I would recommend to the company to find variables that contribute to revenue, but that might have seasonality components to them. If we can figure out trends of cancellation rates/spikes, bandwidth usage, and other interesting variables, then we can properly optimize the company model to take advantage of these trends. This in turn would increase revenue for the company. That said, there does appear to be a positive trend in revenue that is largely consistent outside of a large dip from around August 2023 to October 2023. The revenue rebounded quickly however and looks to be on a good trend again.

**G:**

A PDF of the Jupyter Notebook file is included in the submission.

**H/I:**

Datacamp. (Jan 7, 2025). *ARIMA for Time Series Forecasting: A Complete Guide.* Retrieved March 9th, 2024,From [*https://www.datacamp.com/tutorial/arima*](https://www.datacamp.com/tutorial/arima)

Dr. Elleh, F. (n.d). *D213 Task 1 Building Arima Model in Python video.* Retrieved March 4th, 2024,From <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1aaf2389-6483-4000-b498-b14f00441d57>

Dr. Elleh, F. (n.d). *D213 Task 1 Cohort Webinar PPT.* Retrieved March 4th, 2024,From D603 Course Search

Palma, W. (2016). *Time Series Analysis.* Wiley Global Research (STMS).

Shmueli, G., Bruce, P. C., Gedeck, P., Yahav, I., & Patel, N. R. (2023). *Machine Learning for Business Analytics (2nd ed.).* Wiley Global Research (STMS).