Laycee Glass

NLM3 Task 1

Advanced Data Analytics

4/1/25

Western Governors University

## NLM3 Performance Assessment, Task 1

### Student Information

Student Name: Laycee Glass

ID Number: 001896026

Date: 4/1/25

### Part I: Research Question

### A1. Research Question

The question chosen for this task was the following: Can we forecast the revenue of the hospital system for the next quarter (3 months/90 days)?

### A2. Objectives/Goals of Analysis

The goal of this analysis was to forecast revenue of the hospital for the next quarter using time series modeling techniques. The forecast would allow for the hospital system to make appropriate business decisions based on projected revenue.

### Part II: Method Justification

### B1. Assumptions

Time series modeling with ARIMA relies on several key assumptions (Wormuth, 2024).

1. Stationarity: The data should be stationary, meaning its statistical properties remain stable over time. If trends or seasonality are present, differencing may be needed to make it stationary.
2. Autocorrelation: ARIMA assumes that past values influence future values. For us, that means that past revenue values influence future revenue values. There is a consistent pattern in the data that can be leveraged to forecast values.
3. Univariate analysis: ARIMA models a single variable over time. In this case, we are focusing solely on daily revenue without incorporating external factors.
4. Minimal outlier influence: Extreme outliers in the revenue variable may disrupt the model’s ability to detect patterns. Handling of outliers, whether it be removal or transformation, may be necessary.

### Part III: Data Preparation

### C1. Line Graph Visualization

The code to generate a line graph of the time series, and its subsequent output, is shown below:

A graph with lines and numbers

Description automatically generated

In general, there is an initial upward trend. After that, there appears to be several fluctuations with many peaks and valleys, which may be attributed to seasonality. Overall, however, the daily revenue increased during the observed two-year period.

### C2. Time Step Formatting

The dataset had no missing days, meaning there were no gaps in the time series. The time step is daily, with a total of 731 recorded days. This covers two full years; here, I chose to begin this on January 1, 2016 and end on December 31, 2017. Since 2016 was a leap year (366 days) and 2017 was a regular year (365 days), the total number of recorded days aligns with expectations.

The code for choosing and implementing the start date and timedate format is shown below:

A screenshot of a computer

Description automatically generated

### C3. Stationarity Evaluation

A screenshot of a computer

Description automatically generated As discussed earlier, stationarity is a key assumption for time series modeling. Therefore, we needed to check our data before we started the modeling process. To do this, I used the Augmented Dickey-Fuller (ADF) test. This test uses the criteria that the null hypothesis (H0) is the data is non-stationary, and the alternative (H1) is the data is stationary. The code for the execution of the ADF test, and its results, is shown here:

Our results showed a p-value of 0.1997. This is greater than the standard significance threshold of 0.05. As such, we fail to reject the null hypothesis, indicating the data is likely non-stationary.

### C4. Data Preparation Steps

To prepare the data for analysis, I had to do a few key things. The first thing I did was check for missing or duplicate values. As there were none, I then moved on to check for outliers in the Revenue column. Based on the zscore, there were no outliers that needed to be dealt with. The created zscore column was then dropped.

A screenshot of a computer code

Description automatically generated

A close-up of a white background

Description automatically generated

As mentioned previously, the Day column needed to be converted to proper datetime format. This column then needed to become the index of the time series. This process can be seen in the screenshot of code found in section C2.

The data, as we found, was non-stationary. Therefore, we needed to make it stationary by way of differencing. Once this was done, the ADF test was run again and the data was found to be stationary (p-value of 0.00). Once the data had been differenced, I exported this to its own CSV file for submission.

A screenshot of a phone

Description automatically generated

A screenshot of a computer

Description automatically generated

The final part of the data preparation was splitting the data into training and testing sets. I used an 80/20 split by way of positional indexing. I then made sure the data was split appropriately by checking the shape of each set. Once the sets were verified to be correct, they were then exported to their own csv files for submission. This code is shown here:

A screenshot of a computer code

Description automatically generated

### A close-up of a number Description automatically generated

### C5. Copy of Cleaned Data

A copy of the cleaned, differenced dataset has been submitted alongside this report. A screenshot of the code used to complete this can be seen in the above section.

### Part IV: Model Identification and Analysis

### D1. Time Series Findings

For the annotated findings of the data analysis, the following elements were evaluated:

* A screen shot of a graph

  Description automatically generatedPresence or lack of a seasonal component

This plot indicates there is still a degree of seasonality present within the data. Peaks and valleys occur at regular intervals across the data, suggesting recurring fluctuations likely influenced by time-related factors. We can see the repeating shapes that emerge. This helps to justify the inclusion of a seasonal component in the model (e.g. SARIMAX).

* A graph of a graph

  Description automatically generated with medium confidenceTrends

Visual inspection of the trend plot shows no consistent upward or downward movement over the two-year period. This suggests no strong trend is present in the data.

* Autocorrelation function

A screenshot of a computer screen

Description automatically generated

The ACF plot reveals significant spikes at lag 1, indicating autocorrelation exists in the revenue data. This means that past revenue values (especially the preceding value) influence current values. This supports the use of an autoregressive component in our model.

* A screen shot of a graph

  Description automatically generatedSpectral density

The spectral density plot displays multiple valleys rather than distinct peaks, suggesting a lack of dominant periodic patterns in the data. This indicates that no single frequency overwhelmingly contributes to the series’ variance.

* A screenshot of a computer screen

  Description automatically generatedDecomposed time series

The decomposition of the time series clearly separates the data into trend, seasonal, and residual components. As noted earlier, the seasonal component shows repeating fluctuations. The trend component appears to have no distinct pattern. And the residual component captures the random variation that is not explained by the other components.

* A screen shot of a graph

  Description automatically generatedConfirmation of the lack of trends in the residuals of the decomposed series

Upon reviewing the residual component of the decomposed series and the enlarged graph here, we can see that there is no clear trend present. The residuals fluctuate around zero with magnitude of approximately 1.0 in either direction. This indicates the residuals are random and uncorrelated, which confirms the lack of significant trends in the data.

### D2. ARIMA Model

Getting back to the main objective of forecasting next quarter’s revenue, an ARIMA model was built. ARIMA stands for autoregressive integrated moving average, and it’s a common technique for forecasting future values. There are 3 parameters of ARIMA that require focus: p, d, q (Ibm, 2024).

* p: order of autoregression
* d: degree of differencing
* q: order of moving average

These are typically written as ARIMA(p,d,q). In addition to the standard ARIMA model, there is also a model called SARIMAX. This is a variation of ARIMA that accounts for seasonality components and exogenous variables. I opted to use SARIMAX for this data because there remained a degree of seasonality in our data per the above graph. I also chose this model type because I used auto\_arima, and it gave SARIMAX(1,1,0) as the best model. The code, and output, for auto\_arima can be seen here:

A screenshot of a data sheet

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated The model was then run using the identified parameters on the training set. The model set up and fitting can be seen below:

A screenshot of a computer

Description automatically generated Using this model, predictions were made on the test set. The code for this, and the visualized comparison, can be seen here:

A graph with orange lines

Description automatically generated

### D3. Forecast

The SARIMAX model was then run on the entire dataset to forecast values for next quarter. The code for this process can be seen here:

A screenshot of a computer code

Description automatically generated

### D4. Output and Calculations

All of the necessary output and calculations can be seen in the above sections.

### D5. Code

All of the code used to implement the model can be seen in the above sections.

### Part V: Data Summary and Implications

### E1. Data Analysis Results

The specific model I chose for this task was a SARIMAX(1,1,0) model. This was chosen through the auto\_arima process as I mentioned previously. This process indicated SARIMAX(1,1,0) would minimize the AIC value. Our actual model ended up with an AIC of 861.243.

The data provided to us was daily over the course of two full years. Because of this frequency, the forecast was also daily – 90 days, or the next quarter. The model used previous values and correlations

Given our data covered two full years, forecasting up to a year would be considered accurate. Beyond that, it would be unreliable due to data limitations. My choice of 90 days fell within the year mark, and as such, could be considered accurate for our purposes.

A screenshot of a computer code

Description automatically generated The model was evaluated using both the mean squared error (MSE) and root mean squared error (RMSE). It produced an MSE of 0.2997 and an RMSE of 0.5474. While lower values closer to zero generally indicate "perfect" predictions, these results are relatively small compared to the overall scale of the revenue (millions of dollars). This suggests that the SARIMAX model is reasonably accurate in forecasting future revenue, although some variability still exists. These figures can be seen here:

### E2. Annotated Visualization

Using the model created earlier, the following graph compares the predicted values with the actual test set values.

A graph with a line going up

Description automatically generatedA screen shot of a computer program

Description automatically generated As shown, the forecasted values remain steady at approximately $16.2 million per day over the next 90 days. While this reflects the model’s ability to capture an overall trend, it does not respond well to the daily fluctuations seen in the actual data. The increasing forecast confidence intervals further into the future also highlight the growing uncertainty over time. This pattern suggests that the model may be too simplistic to capture any complex dynamics within the dataset. Future models could improve accuracy by incorporating additional seasonal components or adjusting model parameters to better account for variability.

### E3. Course of Action

Returning to our original question, the model predicts next quarter's revenue to be approximately $16.2 million per day, or approximately $1.458 billion for the whole quarter. however, the practical significance of this is limited. While the SARIMAX model does fit the data reasonably well, the forecasted values appear to be a flat line with considerable variability still. These factors suggest the model may not generalize well to predictions further out.

Given these limitations, I would not recommend the hospital system rely solely on this model to predict next quarter's revenue. The flat forecast trend, combined with the wide confidence intervals, raises concerns about the reliability and usefulness of the prediction for decision-making. The model may be more effective as a general trend indicator as opposed to a precise forecasting tool.

Instead, I would recommend the hospital system take steps to improve future modeling efforts. This could mean creating tailored forecasts at a more granular level, whether it be by individual hospital or even further, by specific departments. this could account for operational differences. additionally, including exogenous variables such as patient volume, insurance reimbursements, and economic indicators could improve forecast accuracy. These variables might help explain some of the revenue fluctuations the current model failed to capture. finally, I would suggest removing the first few months of data for revenue, as this period appears to be a "ramp-up" phase. This data does not reflect the true patterns in the data and excluding it could help strengthen correlations for future predictions.

In summary, while the SARIMAX model offered a reasonable baseline forecast, its limitations highlight the need for more complex and flexible modeling approaches. Forecasts derived from richer, more localized data will better support hospitals in making informed, strategic decisions.

### Part VI: Reporting

### F. Report

### G. Code Sources

D213 Webinars

### H. Text Sources

Ibm. (2024, December 19). *What are Arima Models?*. IBM. https://www.ibm.com/think/topics/arima-model

Wormuth, B. (2024, December 26). *The Stationary Data Assumption in time series analysis*. Statistics Solutions. https://www.statisticssolutions.com/stationary-data-assumption-in-time-series-analysis/#:~:text=In%20time%20series%20analysis%2C%20the%20assumption%20of,in%20the%20data%20are%20stable%20over%20time.