**Part I:  Research Question**

Research Question: Can we predict our organization’s revenue for the next quarter (90 days)? We can answer this question using time series modeling techniques.

The goal of this data analysis is to learn the financial wellbeing of our organization. Ultimately (subsequent to and outside the scope of this analysis), we want to learn the impact readmissions will have on that financial wellbeing. Here, however, we simply want to give the executives an idea of our operations using revenue as the key metric.

**Part II:  Method Justification**

Time series models operate on several assumptions, one of which being that the data are stationary. Stationarity implies that “the mean, variance, and autocorrelation structure do not change over time” (*Stationarity*). In other words, there should be no upwards or downwards trend in the data, the variance should be approximately constant, and the relationship between a value and its neighbors should be approximately the same.

**Part III:  Data Preparation**

A line graph visualizing the time series is shown below. This graph shows the revenue, in millions, for each day in the first two years of operation, totaling 731 separate measurements of revenue per day.

Graphical user interface, chart, line chart

Description automatically generated

We can see, from this visualization alone, that this data is not stationary. This is evident from the clear upwards trend in the data (stationary data does not have a trend over time). Furthermore, running an Augmented Dickey-Fuller (ADF) test can test the null hypothesis that the time series is non-stationary. An ADF statistic of -2.218 and p-value of 0.199 tells us to not reject this null hypothesis. However, we can make the data stationary in this scenario by using the difference of the data, by subtracting the previous value from each value in the time series. The resulting line graph appears below and appears to be stationary based on an ADF statistic of -17.37 and p-value of close to zero.

Chart, bar chart

Description automatically generated

To prepare this data for analysis, I first imported the data frame into a Jupyter notebook, which was the environment I used for the analysis. Upon initial inspection, I found there were 731 rows and one column showing daily revenue over the course of two years (731 days). There were no missing values. I split the data into two sets, a training and a test set. These datasets are attached separately.

**Part IV:  Model Identification and Analysis**

A visual inspection of the stationary data above, as well as the ACF and PACF plots below, tell us that there is no seasonality in this data.

Timeline

Description automatically generated

The lack of peaks in these plots indicate there are no lags in which the data tends to have repeating patterns, and therefore no seasonality. When we decompose the data, we get the visualizations below:

A picture containing background pattern

Description automatically generated

The second plot titled “Trend” tell us the trends are slight, if present at all. It appears in the first year, there is a slight upward trend, and perhaps a slight downward trend in the second year. Furthermore, the fourth plot titled “Resid” showing the residuals over time confirms the residuals are not growing larger nor smaller, and thus, suggest there is no trend.

A plot showing the spectral density is below:

Chart

Description automatically generated

Since our initial dataset is non-stationary, we know we need to use an ARIMA(p, d, q) model for the analysis. Choosing an ARIMA model comes down to choosing values for p, d, and q. Since we learned earlier that we are one degree of differencing from making our data stationary, we already know that d=1. Our ACF and PACF plots indicated that our data can be explained by 1 lagged value, so p can equal 1. We also discovered that there is no seasonality in our data, so 0 is an appropriate value for q. Using the auto.ARIMA function in Python’s pmdarima.arima library helps us confirm, as do AIC and BIC tests. An ARIMA (1, 1, 0) model is appropriate for this time series analysis.

I used the training dataset, which contained the daily revenues of the first 641 days, to create the initial model. The test set contained the final 90 days, and I used the forecast that was created by the initial model to evaluate the model’s precision. One measurement I calculated was the mean absolute percentage error, (MAPE) which suggests the accuracy of a forecast as a percentage. Our training forecast’s MAPE was about 0.09. This means that the average difference between our model’s prediction and the actual values (in the test set) was about 9%. In other words, this model is about 91% accurate. The visualization below shows our forecast (shown on the green line) compared with the actual values (orange line).

Chart, line chart

Description automatically generated

After performing a forecast of revenue for the next 90 days, the first and last 5 values are returned and shown below. The code used to support the implementation of the time series model is attached separately.

731 16.171557

732 16.213858

733 16.231380

734 16.238637

735 16.241643

...

816 16.243769

817 16.243769

818 16.243769

819 16.243769

820 16.243769

We can evaluate our model using the Box-Jenkins method, which is a process that involves three steps: identification of model, estimation of parameters, and diagnostic checking (*Brownlee, 2017*). These steps have been followed in a separate Jupyter notebook, attached and titled “Task 1\_Part V\_Data Summary and Implications.pdf”.

Creating diagnostic plots for this model gave the following output:

Graphical user interface

Description automatically generated

The uncorrelated and normally distributed residuals in the top two plots tell us this is a good model, as the residuals have no obvious patterns. The third and fourth plots also tell us this is a good working model.

I chose to forecast for the next 90 days because quarters are commonly used when discussing revenue. Furthermore, it is not too lengthy of a period, in which case our forecast could be inaccurate. The forecasts were made with a 95% confidence interval. The first and last five values of our forecast have the following prediction intervals, meaning we can say with 95% confidence the actual revenue will fall between the following values:

lower Revenue upper Revenue

731 15.307542 17.035571

732 14.717347 17.710370

733 14.202390 18.260370

734 13.755479 18.721796

735 13.362614 19.120673

.. ... ...

816 2.662033 29.825505

817 2.582180 29.905358

818 2.502791 29.984747

819 2.423859 30.063679

820 2.345374 30.142164

Our final model gives the following forecast on the red line, and the confidence intervals in the light red shading:

Chart, line chart

Description automatically generated

Based on this forecast, we can predict with a high level of certainty the range what our revenue will be for the next quarter. However, this is a very wide range, so forecasts must be taken with a grain of salt, particularly the further in the future they are. With that said, revenue is expected to remain relatively constant next quarter. My recommendation is to put our focus on limiting readmissions, and we can reevaluate our model at the end of the next quarter. This will provide us with a comparison, and we can implement that exogenous factor into our next forecast to see the difference it makes.

References

Brownlee, J. (2017, January 13). *A gentle introduction to the box-jenkins method for time series forecasting*. Machine Learning Mastery. Retrieved December 17, 2021, from https://machinelearningmastery.com/gentle-introduction-box-jenkins-method-time-series-forecasting/

National Institute of Standards and Technology. (n.d.). *Stationarity*. Engineering Statistics Handbook. Retrieved December 17, 2021, from https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm