D213: Performance Assessment Task 1

Part A: Research Question

1. One question that could be asked by someone of this dataset is if we can detect meaningful patterns in the hospital’s profits using the revenue data over the last 2 years.
2. The goal of this analysis will be to determine how the hospital is making revenue and what could correlate between revenue growth and loss during operation.

Part B: Method Justification

1. For this time series analysis, we will be looking at the model using a stationarity and autocorrelated lens to help justify our analytical approach. We will see if the time series is time dependent and if the present value has any similarity or relationship with past values.

Part C: Data Preparation

1. See attached for line graph
2. For this step we checked for general information from the dataset such as null values and missing values. We also can see the general summary statistics to prove there are no gaps within the dataset. See the attached code
3. The stationarity was evaluated in order to check the time series and was found to have a p-value of 0.2 which demonstrates that our dataset is non-stationary since it’s greater than 0.05.
4. The main thing that’s going to be done to this dataset is to split it into a proper testing and training set. I’ll be splitting the data with the training dataset containing 80% of the data and the test set containing 20% of the data. The code for this will be attached.
5. See attached for the cleaned testing and training dataset.

Part D: Model Identification and Analysis

1. For the seasonality component of this time series, we can see based on the attached chart that there is an apparent seasonal component. For the trend component, we can see it goes up for the first year, and then the second year seems to plateau. The autocorrelation shows that the trend is trending downwards over time when checked for the 30 day time lag to represent each month. Also attached is our spectral density function. The decomposed time series is also attached along with multiple charts to demonstrate the trends in the decomposed series.
2. The attached shows our ARIMA model and shows the expected value of 16.06 with a forecast of 19.31 with a standard error of 0.51.
3. See attached for auto forecast model
4. The main values calculated were the mean squared error which resulted in 228.18 and a root mean square error of 15.11. See attached for calculations and code.
5. See attached for code

Part E: Data Summary and Implications

1. For my model I ended up using SARIMAX with an accompanying ARIMA model in order to produce a forecast of results. My prediction forecast was trained using the 2 years of revenue data that was provided and the prediction interval was set for another year in advance. The 1-year forecast makes sense because an organization like a hospital would prefer to use a yearly estimate for if revenue is likely to go up or down. The errors that were calculated for my analysis were a standard error of 0.51 which is pretty good. We also calculated the mean square error of 228.18 and root mean square error of 15.11 which are both comically terrible. The attached code also shows the confidence intervals of the revenue in the 75, 90, 95, and 99 percentiles for what revenue ranges can be expected in the upper and lower bounds. This is interesting as our standard error is small when computing our MSE and RMSE. This is most likely since the training set was steadily going up so when the SARIMAX model ran a time series prediction, it thought it was still going up. After training the model with the test data, it immediately changed to going downward. This is likely due to the function believing in the sample acting as a wave function since time series forecasting uses sine and cosine for its calculations.
2. The interesting wrinkle in my analysis is that it appears that using a SARIMAX forecast model, the graph depicts revenue decreasing over the next year, which it might be thinking in the form of a wave function, which believes that the plateau is more of a peak instead of an indication of it increasing. However, when I run my confidence interval over the information provided, the standard error is small, but the MSE and RMSE are high, which seems to show that it’s confident that it doesn’t know. While the MSE and RMSE aren’t very good, the values they reveal are quite telling. The MSE being 228.18 shows that the values might be close to 228 million dollars in revenue(which is rather extreme) while the RMSE makes a bit more sense, which predicts that the revenue should be around 15 million which seems far more reasonable given the graph and prediction our SARIMAX model created. Attached is the code with the graph that shows this.
3. If I had to make a recommendation to the stakeholders at this hospital, I would tell them to interpret the data as more or less staying steady in revenue with a chance of decline. The high confidence in revenue staying rather steady should give reassurance that revenue will stay steady, however because the model does predict a revenue decrease, it’s important to also expect revenue to decrease and to prepare for it under a worst-case scenario.

Part F: Video

Panopto video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8ebb3be7-6b46-4160-a2b5-ae290183ea37>

Sources:

[1] <https://towardsdatascience.com/time-series-forecast-in-python-using-sarimax-and-prophet-c970e6056b5b> Time Series Forecast in Python using SARIMAX and PROPHET, Michele De Filippo

[2]<https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAXResults.predict.html> statsmodels documentation

[3] <https://machinelearningmastery.com/time-series-data-stationary-python/> How to Check if Time Series Data is Stationary with Python, Jason Brownlee

[4] <https://builtin.com/data-science/time-series-forecasting-python> A Guide to Time Series Forecasting in Python, Sadrach Pierre

[5] <https://www.greenbook.org/marketing-research/how-to-interpret-standard-deviation-and-standard-error-in-survey-research-03377> How to Interpret Standard Deviation and Standard Error in Survey Research, DataStar Inc.

[6] <https://towardsdatascience.com/cyclical-features-encoding-its-about-time-ce23581845ca> Cyclical features encoding, it’s about time! Pierre-Louise Bescond