Performance Assessment

WGU | D213

D213 Task 1

Narcisse, Laurie

2024

Contents

[**Part I: Research Question** 2](#_Toc158380491)

[A.  Describe the purpose of this data analysis by doing the following: 2](#_Toc158380492)

[1.  Summarize one research question that is relevant to a real-world organizational situation captured in the selected data set and that you will answer using time series modeling techniques. 2](#_Toc158380493)

[2.  Define the objectives or goals of the data analysis. Ensure your objectives or goals are reasonable within the scope of the scenario and are represented in the available data. 2](#_Toc158380494)

[**Part II: Method Justification** 2](#_Toc158380495)

[B.  Summarize the assumptions of a time series model including stationarity and autocorrelated data. 2](#_Toc158380496)

[**Part III: Data Preparation** 3](#_Toc158380497)

[C.  Summarize the data cleaning process by doing the following: 3](#_Toc158380498)

[1.  Provide a line graph visualizing the realization of the time series. 3](#_Toc158380499)

[2.  Describe the time step formatting of the realization, including any gaps in measurement and the length of the sequence. 3](#_Toc158380500)

[3.  Evaluate the stationarity of the time series. 4](#_Toc158380501)

[4.  Explain the steps you used to prepare the data for analysis, including the training and test set split. 5](#_Toc158380502)

[5.  Provide a copy of the cleaned data set. 6](#_Toc158380503)

[**Part IV: Model Identification and Analysis** 6](#_Toc158380504)

[D.  Analyze the time series data set by doing the following: 6](#_Toc158380505)

[1.  Report the annotated findings with visualizations of your data analysis, including the following elements: 6](#_Toc158380506)

[2.  Identify an autoregressive integrated moving average (ARIMA) model that accounts for the observed trend and seasonality of the time series data. 12](#_Toc158380507)

[3.  Perform a forecast using the derived ARIMA model identified in part D2. 13](#_Toc158380508)

[4.  Provide the output and calculations of the analysis you performed. 16](#_Toc158380509)

[5.  Provide the code used to support the implementation of the time series model. 16](#_Toc158380510)

[**Part V: Data Summary and Implications** 17](#_Toc158380511)

[E.  Summarize your findings and assumptions by doing the following: 17](#_Toc158380512)

[1. Discuss the results of your data analysis, including the following points: 17](#_Toc158380513)

[2.  Provide an annotated visualization of the forecast of the final model compared to the test set. 19](#_Toc158380514)

[3.  Recommend a course of action based on your results. 19](#_Toc158380515)

[**Part VI: Reporting** 19](#_Toc158380516)

[F.  With the information from part E, create your report using an industry-relevant interactive development environment (e.g., an R Markdown document, a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation. 19](#_Toc158380517)

[G.  Cite the web sources you used to acquire third-party code to support the application. 20](#_Toc158380518)

# **Part I: Research Question**

## A.  Describe the purpose of this data analysis by doing the following:

### 1.  Summarize **one** research question that is relevant to a real-world organizational situation captured in the selected data set and that you will answer using time series modeling techniques.

Is it possible to identify patterns within the revenue time series data for the next 90 days?

### 2.  Define the objectives or goals of the data analysis. Ensure your objectives or goals are reasonable within the scope of the scenario and are represented in the available data.

My primary objective in this analysis is to delve into the daily revenue data to unearth any hidden trends, seasonality, or patterns. Additionally, I aim to construct an ARIMA model capable of accurately predicting future daily revenues. This analysis entails applying time series modeling techniques to the dataset, which comprises information on days and revenue, with the overarching goal of identifying potential patterns. The insights gleaned from this exploration could prove instrumental in forecasting future outcomes.

# **Part II: Method Justification**

## B.  Summarize the assumptions of a time series model including stationarity and autocorrelated data.

Time series models rely on key assumptions, with stationarity being a fundamental requirement. This implies that the statistical properties of the data, such as mean, variance, and autocorrelation structure, remain consistent over time. The absence of trends or seasonal patterns in the time series data is essential for meeting the stationarity criterion. This assumption simplifies the modeling process and underpins many time series methods. Additionally, autocorrelation plays a crucial role in time series analysis, measuring the correlation of the data with its past values. This examination of how the time series is correlated at different time points is a critical step in forecasting and has broad implications across various industries.

In summary, the assumptions of a time series model encompass the need for stationarity, ensuring constancy in statistical properties, and the consideration of autocorrelation, which gauges the relationship of the data with its historical values. These assumptions form the basis for creating reliable models and making accurate predictions in the dynamic field of time series analysis.

# **Part III: Data Preparation**

## C.  Summarize the data cleaning process by doing the following:

### 1.  Provide a line graph visualizing the realization of the time series.

A graph showing a line

Description automatically generated with medium confidence

A graph with a line going up

Description automatically generated

### 2.  Describe the time step formatting of the realization, including any gaps in measurement and the length of the sequence.

The time series mentioned above is formatted in datetime format for each time step. It follows a daily frequency, spanning a total length of 731 days. There appears to be no missing gaps in the measurement of the data.

### 3.  Evaluate the stationarity of the time series.

A graph with a line

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screen shot of a graph

Description automatically generated

Upon visually inspecting the plot, it became evident to me that the data were not stationary, as a noticeable upward trend was observed. Seeking further confirmation, I proceeded to utilize the Augmented Dickey-Fuller test to evaluate the stationarity of the data. To evaluate the stationarity of the time series, I utilized the 'adfuller' module from the Statsmodels library, which enabled me to perform the test. The ADF test returned a p-value of 0.32 and a test statistic of -1.92. Given that the ADF statistic is a small negative number and the p-value is much greater than 0.05, I concluded that we cannot reject the null hypothesis that the time series is not stationary. This conclusion was further supported by the data's clear upward trend, characteristic of non-stationary time series, which lack trends in stationary data. Consequently, I determined that the data would need to be transformed for it to be properly processed by the ARIMA model.

### 4.  Explain the steps you used to prepare the data for analysis, including the training and test set split.

To prepare the data for analysis, I followed several steps. Firstly, I imported necessary libraries including pandas, numpy, matplotlib, seaborn, pmdarima, and statsmodels. Then, I loaded the dataset named 'teleco\_time\_series.csv' into a DataFrame using pandas. I set the index of the DataFrame to a date range starting from '2020-01-01' with a daily frequency. Afterward, I dropped the original 'Day' column from the DataFrame as it was redundant. Next, I visualized the revenue data over time using a line plot, showing the daily revenue in million dollars. Additionally, I plotted the rolling mean of the revenue data using a window size of 30 days to assess its stationarity visually. To further evaluate the stationarity of the revenue time series, I conducted the Augmented Dickey-Fuller test, obtaining the test statistic and p-value. Based on the test results, if the p-value is less than or equal to 0.05, I concluded that the time series is stationary; otherwise, I considered it non-stationary. Moreover, I computed the first difference of the revenue data and performed the Augmented Dickey-Fuller test again to confirm stationarity. Finally, I split the time series into a training set and a test set using a proper 80/20 split, ensuring that the data was shuffled to prevent any potential biases in the splitting process.

### 5.  Provide a copy of the cleaned data set.

A white rectangular object with red text

Description automatically generated

I also combined the files in one as well.

A screenshot of a computer program

Description automatically generated

# **Part IV: Model Identification and Analysis**

## D.  Analyze the time series data set by doing the following:

### 1.  Report the annotated findings with visualizations of your data analysis, including the following elements:

#### •   the presence or lack of a seasonal component

#### •   trends

#### •   the autocorrelation function

#### •   the spectral density

#### •   the decomposed time series

#### •   confirmation of the lack of trends in the residuals of the decomposed series

A screen shot of a graph

Description automatically generated

Seasonality is evident in the data with a repeating pattern observed at regular intervals.

A graph with green lines

Description automatically generated

Plotting the seasonal component of some data, highlighting the Mondays within a specific date range.

A screen shot of a graph

Description automatically generated

The line graph exhibits no discernible upward or downward trends. The data appears to be flat or stationary.

A screenshot of a graph

Description automatically generated

The plot indicates that the values fall within the blue shaded region, suggesting they lack statistical significance. Additionally, it confirms stationarity as the autocorrelation function (ACF) rapidly declines to zero, while the partial autocorrelation function (PACF) displays small values within the blue shaded area, also indicating insignificance.

A graph with purple lines

Description automatically generated

Spectral Density shown above.

A screen shot of a computer program

Description automatically generated

A graph of different colored lines

Description automatically generated with medium confidence

A screen shot of a computer screen

Description automatically generated

There is a lack of trend shown here in the graph.

### 2.  Identify an autoregressive integrated moving average (ARIMA) model that accounts for the observed trend and seasonality of the time series data.

A screenshot of a computer

Description automatically generated

### 3.  Perform a forecast using the derived ARIMA model identified in part D2.

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

A graph with blue lines and white text

Description automatically generated

A screenshot of a computer error message

Description automatically generated

A screenshot of a graph

Description automatically generated

### 4.  Provide the output and calculations of the analysis you performed.

I have provided all the calculations above.

### 5.  Provide the code used to support the implementation of the time series model.

The code required to implement the model has been provided in its entirety.

# **Part V: Data Summary and Implications**

## E.  Summarize your findings and assumptions by doing the following:

### 1. Discuss the results of your data analysis, including the following points:

#### •   the selection of an ARIMA model

#### •   the prediction interval of the forecast

#### •   a justification of the forecast length

#### •   the model evaluation procedure and error metric

The final model was chosen by combining the outcomes of the auto-ARIMA analysis with individual decomposition and autocorrelation examinations. This selected model was built upon the first-differenced data.

A screenshot of a computer

Description automatically generated



Above displays a series of plots based on the models performance. The standardized residual plot shows no obvious patterns. The normal q-q has most of the data points on the red line which is what should happen. The histogram KDE curve should be as near to the normal distribution. The correlogram should have correlations for lags greater than zero and should all be statistically insignificant, indicated by their placement within the shaded area.

### 2.  Provide an annotated visualization of the forecast of the final model compared to the test set.

A graph with blue lines and white text

Description automatically generated

### 3.  Recommend a course of action based on your results.

The ARIMA time series effectively forecasted the next 90 days with upper and lower error margins. I suggest leveraging this forecast for planning and anticipating increases in the customer base, as well as projecting revenue for the upcoming two quarters. While there's room for improvement, I advise exploring more combinations of model hyperparameters to enhance accuracy. Although incorporating seasonality into the ARIMA model might seem promising, my testing revealed that it resulted in a more complex yet less accurate model.

# **Part VI: Reporting**

## F.  With the information from part E, create your report using an industry-relevant interactive development environment (e.g., an R Markdown document, a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

Please see the attached html and pdf file labeled as Task 1.

## G.  Cite the web sources you used to acquire third-party code to support the application.

Brownlee, J. (2020a, April 23). *How to Model Residual Errors to Correct Time Series Forecasts with Python*. MachineLearningMastery.com. https://machinelearningmastery.com/model-residual-errors-correct-time-series-forecasts-python/

Brownlee, J. (2020b, December 9). *How to Decompose Time Series Data into Trend and Seasonality*. MachineLearningMastery.com. https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

*Converting day count to date time*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/61389654/converting-day-count-to-date-time

Data 360 YP. (2020, December 21). *Time Series Analysis in Python Tutorial - V1* [Video]. YouTube. https://www.youtube.com/watch?v=axjgEgBgIY0

*Detect strings with non English characters in Python*. (2021, March 24). https://discuss.dizzycoding.com/detect-strings-with-non-english-characters-in-python/

GeeksforGeeks. (2020a, February 19). *Python ARIMA model for Time series Forecasting*. https://www.geeksforgeeks.org/python-arima-model-for-time-series-forecasting/

GeeksforGeeks. (2020b, November 26). *AutoCorrelation*. https://www.geeksforgeeks.org/autocorrelation/?ref=lbp

GeeksforGeeks. (2022, July 5). *Plot the power spectral density using Matplotlib  Python*. https://www.geeksforgeeks.org/plot-the-power-spectral-density-using-matplotlib-python/

*Good Resource For Converting ARIMA output in R to equation form?* (n.d.). Cross Validated. https://stats.stackexchange.com/questions/340838/good-resource-for-converting-arima-output-in-r-to-equation-form

*How to export current notebook in HTML on Jupyter | Saturn Cloud Blog*. (2024, January 3). https://saturncloud.io/blog/how-to-export-current-notebook-in-html-on-jupyter/