Taylor Cina

D213 – Advanced Data Analytics

***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.*

What is the expected daily revenue over 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.*

The objective of this analysis is to develop an autoregressive model that can predict company revenue over the next 90 days with acceptable accuracy.

***Part II: Method Justification***

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

Time series autoregression models assume that the dataset is stationary and has some level of autocorrelation that would allow new elements to be predicted by previous ones. Without stationarity, the model would have too many variables to control for to accurately forecast data. Without autocorrelation, the model would have no patterns strong enough to make meaningful predictions.

***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 the growth of the company's revenue

Description automatically generated

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

The series uses daily revenue measurements over the span of 731 consecutive days. There are no missing measurements in the series.

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

An augmented Dickey-Fuller test on the unmodified revenue time series gives a p value of ~0.393, which is significantly higher than the critical value of 0.05. This means the series is a random walk and therefore non-stationary.

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

To prepare the data, an arbitrary range of dates starting from January 1st, 2020 was generated to act as an index for the dataset. A new column is added to the dataset for the numeric difference between revenue entries. Using percent change requires dropping more columns at the head of the dataset due to the first two entries being 0 and near 0. Augmented Dickey-Fuller tests are run on both the revenue variable and the new revenue difference column to determine the stationarity of the revenue difference column. The data is split into train and test sets using the first 70% of the data for the training set, and the remaining 30% for the test set.

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

Cleaned dataset attached as “teleco\_time\_series\_clean.csv”

***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*

The non-stationary time series appears to have a seasonal component with a 90-day period.

A graph of blue lines

Description automatically generated

*•   trends*

The trend line of the non-stationary time series is shown below.

A line graph with numbers and a line

Description automatically generated with medium confidence

*•   the autocorrelation function*

The autocorrelation function of the stationary time series is shown below:

A graph with blue dots and numbers

Description automatically generated

The dataset has significant mean reverting autocorrelation on lag 1 and 2.

*•   the spectral density*

The spectral density of the stationary time series is shown below:

A graph of a power spectrum

Description automatically generated

*•   the decomposed time series*

The full decomposition of the non-stationary time series is shown below:

A graph of growth and revenue

Description automatically generated with medium confidence

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

The graph of the residuals of the decomposed time series is shown below:

A graph of blue lines

Description automatically generated

A Ljung-Box statistic of 0.12 in the model generated in the next section indicates that the residuals of the model do not have significant autocorrelation.

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

Using auto-ARIMA, a model with the order (1, 1, 0) and seasonal order (1, 1, 0, 90) was determined to have the lowest AIC. The Ljung-Box statistic of 0.18 indicates that the model does not have significant autocorrelation in its residuals.

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

Below is a forecast generated by a SARIMAX model compared to the testing set of data excluded from model training:

A graph of a graph showing a number of data

Description automatically generated with medium confidence

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

Output and calculation code is attached as the file “TCina D213 T1.ipynb”

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

Code is attached as the file “TCina D213 T1.ipynb”

***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 appropriate ARIMA model was selected using auto-ARIMA with the observed 90 day seasonal cycle.

*•   the prediction interval of the forecast*

The prediction interval of the forecast is daily, given the original dataset consists of a daily time series and no resampling was performed.

*•   a justification of the forecast length*

A forecast length of 3 months, or 90 days, is an acceptable amount of time for a company to prepare for the next quarter. Predictions longer than this period may be less accurate. Additionally, the test set used to evaluate model performance is longer than this period.

*•   the model evaluation procedure and error metric*

The model was evaluated using the AIC and mean absolute error. The model generated by auto-ARIMA has an AIC of 858 and a mean absolute error of 0.49, indicating that the model is a good fit for the data.

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

Below is the final model and 90-day forecast after the end of the test set compared to the full dataset:

A graph with numbers and lines

Description automatically generated

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

The model can be used to prepare for fluctuations in revenue predicted in the next 90 days based on historical revenue data.

***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.*

PDF attached as the file “TCina D213 T1.pdf”

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

No web sources used.

*H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.*

No third party sources used.