D213 Times Series Modeling

Telecommunication Revenue Data

Darien Nguyen

Western Governors University

D213 Task One: Times Series Modeling

July 6, 2023

**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.

Can we identify patterns or trends in the daily revenue data over the first two years of operation for the telecommunications company through the use of a time series?

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 objectives or goals of the data analysis:

a) Identify Seasonality: Determine if there are any seasonal patterns in the revenue data, such as recurring patterns related to specific times of the year, months, weeks, or days.

b) Detect Trends: Identify any long-term increasing or decreasing trends in the revenue data, which can help understand the overall growth or decline in the company's revenue over time.

c) Forecasting: Develop a time series forecasting model to predict future revenue based on the historical data. This can assist the telecommunications company in estimating their future revenue and making informed business decisions.

e) Identify Anomalies: Detect any outliers or abnormal observations in the revenue data that may require further investigation. These anomalies could indicate specific events or circumstances that had a significant impact on revenue.

**Part II: Method Justification**

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

Stationarity:

A fundamental assumption in time series modeling is stationarity. Stationarity implies that the statistical properties of a time series remain constant over time. The following assumptions are associated with stationarity:

Constant Mean: The mean of the time series remains constant over the entire time period. This assumption suggests that the average revenue or any other variable under consideration does not exhibit a significant upward or downward trend over time.

Constant Variance: The variance of the time series remains constant across all time points. This assumption implies that the dispersion or variability of the revenue data does not show systematic changes as time progresses.

Constant Autocovariance: The autocovariance between observations at different time points remains constant. Autocovariance represents the covariance between observations at different lags. In a stationary time series, the autocovariance does not depend on the specific time points being considered.

Autocorrelated Data:

Autocorrelation refers to the correlation between observations at different time points in the series. The assumptions related to autocorrelation include:

Serial Correlation: Serial correlation indicates that the correlation between two observations decreases as the time lag between them increases. In other words, the values of the time series at adjacent time points are not independent but exhibit some level of correlation.

No Spurious Relationships: Spurious relationships occur when variables appear to be correlated, but the relationship is actually driven by a common underlying factor or the presence of other variables. It is important to identify and account for any spurious relationships in time series analysis to ensure accurate modeling.

Lack of Autocorrelation in Residuals: After fitting a time series model, the residuals (the differences between observed and predicted values) should not exhibit autocorrelation. Autocorrelated residuals indicate that the model has not captured all the systematic patterns present in the data.

**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 with a red line

Description automatically generated

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

From the data dictionary, no time frame was given to us from the collected data. We know that the data was collected within a two-year span of 731 days with no gaps in between.

3.  Evaluate the stationarity of the time series.

To assess stationarity, an Augmented Dickey-Fuller test was performed. The test statistic was -1.9246, and the p-value was 0.32057282. A more negative test statistic indicated a stronger rejection of the null hypothesis (stationarity), but the value was not convincingly negative. Additionally, the p-value was higher than the commonly used significance level of 0.05, indicating that we could not reject the null hypothesis, and the data was likely non-stationary.

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

1. Data Loading: The data was loaded from a CSV file using Pandas' **pd.read\_csv()** function into a DataFrame named **df**.
2. Data Preprocessing: The DataFrame was visually inspected using **df.info()** and **pd.set\_option("display.max\_columns", None)** to check the data types, number of values, and overall size. The 'Day' column, which presumably represents the date, was converted to a datetime format using **pd.to\_datetime()**. The 'Day' column was then converted to differences in time (time deltas) by subtracting one from each value and adding it to a start date ('2020-01-01') to align the dates properly.
3. Data Exploration: The DataFrame columns were renamed to have Pythonic names using **df.columns = ['date', 'revenue']**. The 'date' column was set as the DataFrame index using **df.set\_index('date', inplace=True)**. The time series data was visually inspected with a plot using Matplotlib to get an overview of the daily revenue pattern over time. A trend line was also plotted on the data to identify any underlying trends.
4. Stationarity Check: The Augmented Dickey-Fuller test was performed on the 'revenue' column using **adfuller()** from the **statsmodels.tsa.stattools** module. The test statistic and p-value were obtained from the test results to assess the stationarity of the time series.
5. Data Transformation: To address non-stationarity, the time series was differenced using the **.diff()** function to calculate the first difference of the revenue data. The differenced DataFrame, **df\_trans**, was created by dropping the first row containing NaN values resulting from the differencing process.
6. Train-Test Split: The differenced data, **df\_trans**, was split into training and test sets using **train\_test\_split()** from **sklearn.model\_selection**. The training set was assigned 80% of the data, and the test set was assigned the remaining 20%. The data was split without shuffling, ensuring the sequential order of the time series was maintained.
7. Data Decomposition: The seasonal decomposition of the differenced data was performed using **seasonal\_decompose()** from **statsmodels.tsa.seasonal** to separate the time series into trend, seasonal, and residual components.
8. ARIMA Model Fitting: An ARIMA model with the order (1, 0, 0) was fitted on the training data using **ARIMA()** from **statsmodels.tsa.arima.model**. The order (1, 0, 0) indicates a first-order autoregressive (AR) component with no differencing (I) and no moving average (MA) component.
9. Forecasting: The fitted ARIMA model was used to forecast future revenue values using **get\_prediction()**. The forecasted values were obtained for the time steps corresponding to the test set.
10. Model Evaluation: The root mean squared error (RMSE) was calculated to evaluate the performance of the ARIMA model on the test set. The RMSE measures the difference between the forecasted and actual revenue values.
11. Confidence Intervals: Confidence intervals were calculated for the forecasted values using **confidence\_intervals** obtained from the ARIMA model. The intervals represent the uncertainty of the forecast.
12. Final Visualization: The final forecasted values, along with confidence intervals, were plotted against the observed data to visualize the accuracy of the forecast and the uncertainty associated with it.

5.  Provide a copy of the cleaned data set.

**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 – no seasonal component found

•   trends – No trend

•   the autocorrelation function

•   the spectral density – downward then levels

•   the decomposed time series

A group of blue waves

Description automatically generated•   confirmation of the lack of trends in the residuals of the decomposed series – lack of trends

A close-up of a graph

Description automatically generatedA graph showing a blue line

Description automatically generated

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

We have identified an ARIMA(1, 0, 0) model that accounts for the observed trend and seasonality in the time series data. The model includes an autoregressive term of order 1 (AR(1)) to capture the relationship between current revenue values and the previous value. Additionally, non-seasonal differencing of order 0 (I(0)) is used to account for the trend, and there is no moving average term (MA) as its order is 0.

The ARIMA(1, 0, 0) model is represented as:

revenue\_t = const + ar.L1 \* revenue\_{t-1} + residual\_t

Where:

* revenue\_t represents the current revenue value.
* const represents the constant term with an approximate value of 0.0234.
* ar.L1 corresponds to the coefficient of the autoregressive term, which is approximately -0.4597.
* revenue\_{t-1} represents the previous revenue value.
* residual\_t represents the model's residuals.

The model suggests that the current revenue value can be predicted using the previous revenue value (AR(1) term) and a constant term. The autoregressive coefficient is statistically significant (p < 0.001), indicating a strong relationship between current and lagged revenue values.

The residuals of the ARIMA(1, 0, 0) model show no evidence of autocorrelation (Ljung-Box test p-value = 0.96), and they are approximately normally distributed (Jarque-Bera test p-value = 0.40), confirming that the model's assumptions are met.

The decision to employ an ARIMA(1, 0, 0) model with an autoregressive term of order 1 (AR(1)) stems from the observed dynamics within the daily revenue time series. The primary motivation behind this choice lies in the potential persistence of revenue values over time. In a business context, revenue figures are often influenced by historical performance, and there may be a substantial correlation between the current day's revenue and the revenue from the previous day.

By including an AR(1) term in the model, we aim to capture this inherent temporal relationship. Specifically, the coefficient associated with the AR(1) term represents the impact of the previous day's revenue on the present day's revenue. The statistically significant nature of the AR(1) coefficient (p < 0.001) highlights a robust connection between current revenue and its immediate historical counterpart. This signifies that changes in revenue tend to propagate over time, which aligns with the intuition that business performance doesn't undergo abrupt fluctuations but exhibits a degree of continuity.

Moreover, the Ljung-Box test p-value of 0.96 for the model's residuals suggests that there is no evidence of significant autocorrelation in the residuals, further supporting the validity of the chosen ARIMA structure. Additionally, the normal distribution of the residuals, as indicated by the Jarque-Bera test p-value of 0.40, validates the assumption of normality and independence in the model's errors.

By leveraging an ARIMA(1, 0, 0) model with an AR(1) term, we're effectively capturing the sequential relationship between revenue values while accounting for the observed trend and seasonality. This choice is grounded in the understanding that past revenue performance holds valuable insights for forecasting future revenue trends.

A screenshot of a computer screen

Description automatically generated

A graph showing a line

Description automatically generated3.  Perform a forecast using the derived ARIMA model identified in part D2.

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

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generated

Please zoom in to see calculations.

The root mean squared error of this forecasting model is 2.47394

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

**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 analysis resulted in the selection of an ARIMA(1, 0, 0) model. This model includes an autoregressive term (AR(1)) with a coefficient of approximately -0.4597, indicating a significant negative relationship between the current and lagged revenue values. The non-seasonal differencing term (I(0)) was used to account for the trend in the data, and there is no moving average term (MA) in the model.

•   the prediction interval of the forecast:

The forecasted daily revenue values are presented with a confidence interval, which helps assess the uncertainty around the predictions. The confidence interval provides a range within which the actual values are likely to fall with a certain level of confidence. This interval is visually represented in the annotated visualization of the forecast.

•   a justification of the forecast length

I did not find any specific information in the provided data dictionary or external documentation regarding the date range or duration of the telecom churn data. Therefore, I made the decision to analyze the data for the timeframe of 2020 to 2021 based on my own judgment and assumptions.

Given the available data, my focus was to predict the daily revenue for the remaining months of 2021. This choice was driven by the need to make near-term forecasts and gain insights into revenue patterns and trends for the latter part of the dataset.

It is important to acknowledge that the selection of the date range for time series forecasting can significantly impact the results and the model's performance. As such, while I chose the 2020-2021 timeframe for this analysis, it's essential to document and communicate such assumptions to ensure the reproducibility and transparency of the findings.

•   the model evaluation procedure and error metric:

The ARIMA model's performance was evaluated using the root mean squared error (RMSE) metric, which measured the difference between the forecasted values and the observed values. The calculated RMSE of approximately 2.47394 indicated the average magnitude of the forecast errors, providing an estimate of the model's accuracy in predicting daily revenue.

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

The annotated visualization of the forecast shows the predicted daily revenue (green dashed line) along with the observed revenue (blue line) for the entire period from 2020 to 2021. The confidence intervals (pink shaded region) represent the uncertainty around the forecasted values. This visualization helps compare the forecast with the actual test set, enabling us to assess the accuracy of the model's predictions and how well it captures the underlying patterns.

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

Based on the results, the ARIMA(1, 0, 0) model has provided a reasonable forecast of daily revenue for the remaining months of 2021. However, to ensure the model's reliability and generalizability, further validation and testing against out-of-sample data are warranted. The current analysis is based on a dataset spanning from 2020 to 2021, and while it captures revenue trends within this timeframe, it may not fully account for potential shifts or patterns in more recent data.

To enhance the forecasting accuracy and gain a better understanding of any missed trends or patterns specific to our clients, it would be valuable to compare the telecom churn data with datasets from other companies in the same industry. Analyzing broader industry trends and benchmarks can provide additional insights into the factors influencing daily revenue and uncover patterns that might be unique to specific clients.

**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.

G.  Cite the web sources you used to acquire third-party code to support the application.  
No sources were used.

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.  
No sources were used.

I.  Demonstrate professional communication in the content and presentation of your submission.