**Time Series Analysis of Chicago Transit Authority (CTA) Boarding Rates**

RathiPriyanka Srinivasan

Master of Science in Data Analytics, Western Governors University

D214: Data Analytics Graduate Capstone

Dr. Daniel Smith

April 25, 2023

**Research Question**

The Chicago Transit Board has approved nine renovation projects with the goal of modernizing and improving the city’s buses and trains, stations, tracks, and implemented technology so that transit is more accessible to Chicago’s residents (“System improvement projects,” 2023). The significant costs of renovations, including the $19 million needed for “repairs and improvements for the Western Brown Line Station” (“Chicago Transit Board approves plan*,”* 2023), may have room for improvement when planning is paired with Time Series Modeling (TSM). This analysis seeks to answer whether CTA boarding rates can be reliably forecasted at 90 percent accuracy via Time Series Analysis:

* Hypothesis – CTA boarding rates can be forecasted with 90 percent accuracy.
* Null Hypothesis – CTA boarding rates cannot be forecasted with at least 90 percent accuracy.

By forecasting future boarding rates, the Chicago Transit Board will be able to reduce the cost of renovations by strategically scheduling them during lower boarding rate times.

**Data Collection**

The CTA boarding rate data was provided by the CTA within Chicago’s Open Data Portal (“CTA – Ridership,” 2023). Monthly boarding rates for each station were collected through an automated process that involves recording station facecard usage (“Ridership Readme,” 2011). The advantage of having an automated data collection process is crucial as it prevents human error and provides consistent, and reliable data. However, one thing to note is that cross-platform transfers were not counted as ‘entries’, this exclusion could lead to some bias within the dataset as many Chicago residents need to transfer between lines to reach their destination (“Ridership Readme,” 2011). Despite the challenge of bias within the dataset, the Chicago Open Data Portal is updated several times a day which allows the ridership dataset and the TSM analysis to have a real-world context (“Chicago Data Portal,” 2023). The Chicago Open Data Portal also specifies that the hosted datasets are open source which means that they can be downloaded and analyzed without undergoing a formal IRB approval process (“Chicago Data Portal,” 2023).

**Data Extraction and Preparation**

Graphical user interface, application, table, Excel

Description automatically generatedSince TSM requires a univariate dataset (Pandian, 2021), the original CTA ridership dataset was processed in Excel using the PivotTable function to sum total monthly boarding rates for each L station by month and year. This way, the dataset is not segmented by each station and provides continuous and univariate fields ‘month’ and ‘monthlytotal’ for analysis. A screenshot of the process is provided below:

Please note that ‘monthlytotal’ was then divided to represent ridership by the billions so that visualizations within Python are easy-to-read. The PivotTable was then saved as a separate .csv file as ‘cta\_univariate.csv’.

Next, the following dataset preparation will then be performed using Python and the following packages:

* Pandas – for data manipulation
* Numpy – to create and manipulate arrays
* Seaborn – for plotting and visualization
* MatPlotLib – to format x-axis ticks and line plots
* Statsmodels – for ARIMA and Augmented Dickey Fuller Test (ADF)

Text

Description automatically generatedThe ‘cta\_univariate.csv’ file was stored as a dataframe called ‘cta\_data’ using the pd.DataFrame() function and then plotted to visualize the original dataset and determine any obvious trends or lack of stationarity. A screenshot of the code and output is provided below:

Chart

Description automatically generated

CTA ridership ‘monthlytotal’ is incremented by month for years ranging between 2001 and 2022. By plotting ‘cta\_data’, it is easy to see that there is an obvious drop between 2019 and 2020 - which could be attributed to the drastic drop in train and bus ridership due to COVID-19 (“CTA ridership jumps,” 2023). This drop could impact the accuracy of the TSM forecast. At the beginning, the revenue jumps from .000793 to Next, the Augmented Dickey Fuller Test (ADF) was conducted to measure the stationarity of the dataset. A screenshot of the code output is given below: Text

Description automatically generated

The P-Value does not seem to be significant at .4815, and the absolute value comparison of the Critical Values versus the ADF statistic shows that the Critical Value is greater than the ADF statistic. These two findings convey that the dataset is non-stationary. The following steps were taken to prepare the data for TSM:

1. Apply differencing to make the dataset stationary.

Text

Description automatically generatedChart

Description automatically generated with low confidence

1. Text

   Description automatically generatedSplit the dataset into train and test sets and save them as ‘train.csv’ and ‘test.csv’.

**Analysis**

**A picture containing background pattern

Description automatically generated**A screenshot of the decomposed time series is included below:

**Chart, line chart, histogram

Description automatically generated**Although there is a trend and seasonality present within the dataset, both components do not seem to be significant after differencing. The same conclusion can be observed by plotting the autocorrelation of the differenced dataset, screenshot included below:

**\**

**Graphical user interface, chart, line chart

Description automatically generated**In addition, the spectral density plot (screenshot below) displays a relatively flat line that also supports the non-significance of seasonality within the differenced dataset.

Text

Description automatically generatedA screenshot of the code and the output for the stepwise summary of the auto ARIMA function and the final model summary is included below: Text

Description automatically generated

Text

Description automatically generated

**Graphical user interface, application, table

Description automatically generated**The ARIMA model order (0,0,0) is selected after a stepwise search for the lowest AIC, -2531. There is an advantage to using the stepwise method for discovering the best AIC value and in turn, the most appropriate ARIMA model order. However, an ARIMA model with “all zero components” indicates a “White Noise model” (Zuniga, 2023). This means that the TSM model itself cannot be depended upon for accurate forecasting. For the sake of visualizing the TSM model, a forecast interval of 1 step was chosen. This interval is defined by the sum of the length of ‘train’ and ‘test’ datasets minus 1, so that the standard deviation of both the forecast distribution and the standard deviation of the residuals are similar (Athanasopoulos, 2023). A screenshot of the graph comparing the output of the final model forecast (‘pred’), test dataset (‘test’), train dataset (‘train’), and the confidence interval (95%) is included below:

**Data Summary and Implications**

Next, the model’s performance was evaluated with the ‘plot\_diagnostics()’ function and by calculating the Mean Absolute Error (MAE) of the test and train dataset with their associated predictions. A screenshot of the code and the output of the four plots is given below:

Text

Description automatically generated

Chart

Description automatically generated

Graphical user interface, text

Description automatically generated

The low MAE values for both the ‘train’ dataset (.000609) and ‘test’ dataset (.000807) indicate a high model accuracy. However, the standardized residual plot (top left) shows that there could be a pattern based on the drops located near the beginning and the end of the plot. The histogram depicting the KDE estimate (top right) displays a curve that has a relatively normal distribution. The Normal Q-Q plot (bottom left) shows that although most of the datapoints occur on the red line, there were a couple datapoints that were far from the redline at the start. Finally, the correlogram (bottom right) shows the correlations for lags 3, 8, and 9 may be significant since they fall at the edge of the shaded area. All four of the diagnostic plots imply that there is a lack of statistical soundness within the model. In addition, the fact that all three of the ARIMA model’s PDQ values are zero suggests that the model itself is imprecise and should not be used to forecast CTA ridership.

With these results, CTA Board members can choose to pursue TSM with a more extensive dataset. For example, having a dataset that includes a daily ridership total rather than monthly ridership total will allow for the model training dataset to be larger, which may lead to non-zero PDQ values, and eventually allow for a model that is both statistically sound and provide accurate predictions. Another option is for CTA board members to abandon predictive modeling entirely and pursue a clustering technique like Hierarchal Agglomerative Clustering (HAC) to derive insights into rider behavior and plan renovations based on the resulting clusters.

**References**

Athanasopoulos, G. & Hyndman, R. (2023). Forecasting: Principles and Practice. OTexts. Retrieved from <https://otexts.com/fpp2/prediction-intervals.html>

Chicago Data Portal. (2023). Retrieved from https://data.cityofchicago.org/

*Chicago Transit Board approves plan for rehabilitation of the Western Brown Line Station.* (2023). 2023 Chicago Transit Authority. Retrieved from https://www.transitchicago.com/chicago-transit-board-approves-plan-for-rehabilitation-of-the-western-brown-line-station/

*CTA ridership jumps nearly 25% in 2022, but still trails far behind pre-pandemic levels.* (2023). CBS Chicago. Retrieved from https://www.cbsnews.com/chicago/news/cta-ridership-2022-covid-19-pandemic/

*CTA - Ridership - 'L' Station Entries - Monthly Day-Type Averages & Totals.* (2023). Chicago Transit Authority. Retrieved from https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Monthly-Day-Type-A/t2rn-p8d7

Pandian, S. (2021). *Time series analysis and forecasting: Data driven insights (Updated 2023)*. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/#What\_Are\_the\_limitations\_of\_Time\_Series\_Analysis?

Ridership Readme. (2011). Chicago Transit Authority. Retrieved from https://data.cityofchicago.org/api/assets/7971211C-57DF-4650-8EBD-BE295CCD3151?download=true

System improvement projects. (2023). 2023 Chicago Transit Authority. Retrieved from https://www.transitchicago.com/projects/

Zuniga, J. (2023). Arima Models. Amazon AWS. Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/363796\_2c2581467b02403bbdcc77c6b5660ec7.html#:~:text=The%20ARIMA(0%2C%200%2C,variety%20of%20time%20series%20models.