Bus Terminal Arrival Forecast

The goal of this project is to forecast the number of passengers arriving at the city's bus terminal to help your transportation company make informed decisions about increasing the terminal size and adding buses. This report will begin by introducing the data and strategies used to evaluate forecasting models. The report will then explore several forecasting methods that were implemented; finally, the most suitable method will be applied to predict the number of people expected at the terminal from March 22nd at 6:30 am to March 24th at 10:00 pm.

Introduction

Data on passengers at the terminal extends from March 1st (Tuesday morning) to March 21st (Monday night) in 15-minute intervals. Data collection begins at 0630 and ends at 2200 each day.

The data will be converted into a time series and divided into training and validation sets; models will be run on the training set of data and used to forecast the validation period; results of the prediction will be compared to the actual values to determine the model's accuracy. The project calls for an 80% training-validation split; however, to keep the seasons intact and not divide the time series mid-season, I have chosen to use the first 882 observations (66%) as the training set. The first two seasons will be used as training, with the final season as validation; this will allow me to easily run forecasting models.

The metrics, mean absolute error (MAE) and root mean squared error (RMSE), will be used to measure model performance and will be reported for training and validation periods. Mean absolute scaled error (MASE) will not be reported, as all models returned a value of infinity. This is potentially because 108 points in the time series record zero people at the terminal, which could have interrupted MASE calculations.

An initial visualization of the time series suggests that multiple seasonality likely exists. By observing the dataset, it seems that there are 63 observations in a day and 441 observations in a week. This seasonality is confirmed by observing the plotted autocorrelation function, where there is strong autocorrelation at 63 and 441 lags, respectively. Therefore, it will be essential for the model to capture daily seasonality (spikes in usage during commute hours) and weekly seasonality (less usage on weekends) to make accurate predictions.

Model Selection and Application

Data Driven Methods:

Initially, simple data-driven forecasting methods were implemented. First, the average forecasting method is run as the baseline, not accounting for seasonality or trend. An MAE of 19.4 and RMSE of 25 are recorded for the validation set. The training set had an MAE of 19 and RMSE of 23; subsequent models will be compared to this value. Considering that the season is consistent with almost no trend, I suspect that a seasonal naive method may be appropriate. This model performs very well with an MAE of 5.9 and RMSE of 10.8 for the validation set, and an MAE of 5.9 and RMSE of 7.5 for the training set. Despite receiving excellent results, other models were explored to improve the validation set forecast; however, I was not able to get a lower MAE from subsequent models.

Model Based Forecasting Methods:

The average and seasonal naive methods are data-driven forecasting methods. To select which model-based method to use, the Partial AutoCorrelation Function (PACF) and AutoCorrelation Function (ACF) are observed. The ACF displays the correlation between the most current observation and the observed value at each lag, while PACF shows the correlation of each lag independent from other lags. By observing the ACF and PACF plots, it is noted that there is significant autocorrelation when intermediate lags are considered and when they are not. Furthermore, the distribution of residuals is not normal, being right-skewed. To address these issues, several models will be applied.

ARIMA:

ARIMA models are designed specifically to deal with autocorrelation. I try to predict using ARIMA(2,0,1)(2,0,0), the equation for this model is:

$$Y_t = c + \phi 1\, Y_{t-1} + \phi 2\, Y_{t-2} + \theta 1\, \varepsilon_{t-1} + 0.1808\, Y_{t-63} + 0.0768\, Y_{t-126} + \varepsilon t$$

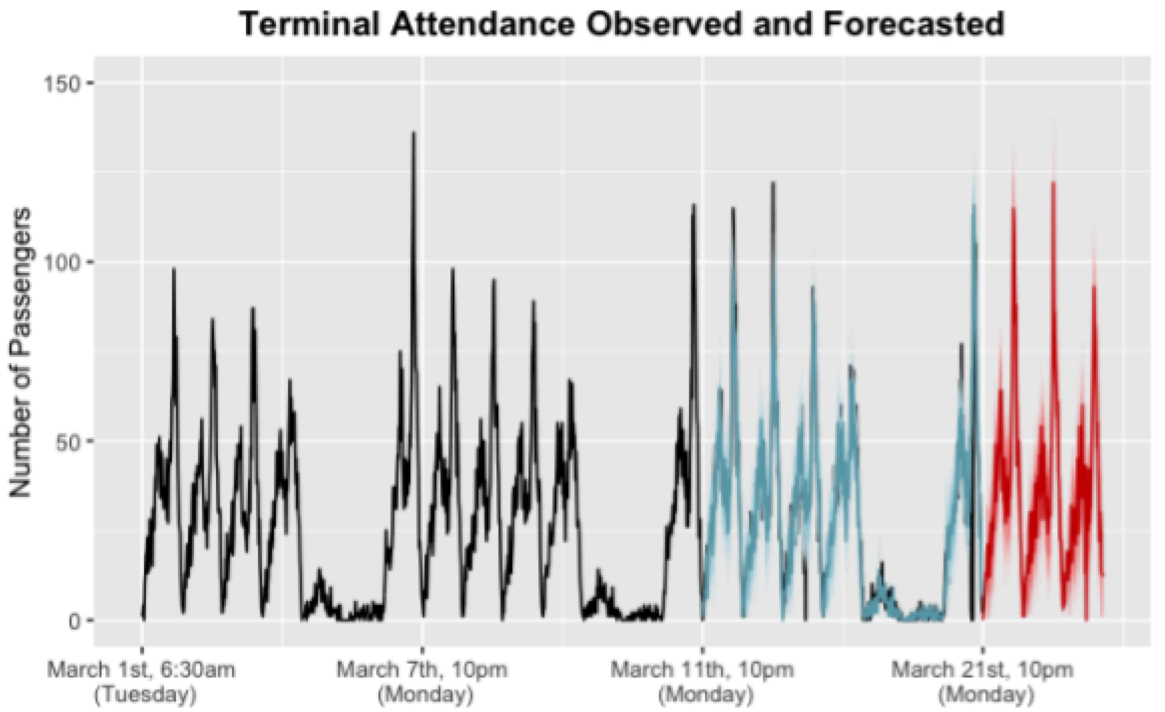
This equation considers the past two lags as well as the error from the previous lag. It also incorporates lags from the two previous days or "seasons." Note that the model does not account for weekly seasons. My model performs poorly compared to the seasonal naive method, with an MAE score of 18.718078 and an RMSE of 23 for the validation set. The training set had an MAE of 7.7 and an RMSE of 4. Auto ARIMA suggested using ARIMA(0,1,4) using the “cox-box” method to select lambda; however, as expected the model performed extremely poorly. Other ARIMA models that account for multiple seasonality could perform better; however, these models were not further explored due to time constraints.

Regression:

To improve predictions, it was crucial to consider both daily and weekly seasonality. Regression with trend and seasonality as predictors was run, with the lambda value determined using the "cox-box" method. Based on visualizations, this model adequately accounts for both weekly and daily trends, yielding a low MAE of 5.60 and RMSE of 10.4 for the validation set, and an MAE of 5.5 and RMSE of 3.7 for the training set. While this model represents a significant improvement over the ARIMA model, the MAE is no better than the seasonal naive method. Moreover, there is a significant risk of overfitting.

Final Forecast and Considerations

The seasonal naive method was employed for the final forecast as it outperformed other model-driven forecasting techniques, achieving the lowest MAE score of 5.9 for the validation set and an RMSE of 10.8, which was only slightly higher than that of the regression model. The model was applied to the entire dataset from March 1st to March 21st, providing forecasts for the requested dates of March 22nd to March 24th. The figure below illustrates the validation set prediction overlaid on the actual data as well as the forecast for March 22nd to 24th.



Observed Data (Black), Validation Prediction (Blue), Forecast of March 22nd -24th (Red)

While the model effectively estimates the training set, its predictions rely solely on time series analysis and overlook other factors such as weather or disruptions in alternative travel methods, which may influence ridership. It's essential to acknowledge this limitation, as unforeseen external influences could lead to surges or declines in terminal attendance. Like the regression approach, there is a high risk that the model is overfit, especially considering that the naive model is a data-driven method simply using evaluations from the previous seasons to make predictions. Since the model solely accounts for seasonality while disregarding any existing trend, it would not be appropriate to use this model for long-term or medium-term forecasting. However, this model suits the need of the limited three-day forecast which was requested.