Nevada Airline Modeling

By: Corey Dearing, Nick Young, and Sana

4/29/2024

Abstract

Forecasting monthly passengers for airlines is of importance to many sectors of the aviation and tourism industry. Our project is to build a model for a single airline, that predicts total monthly traffic through their busiest hubs in a chosen state. For Nevada, we find Southwest Airline to be the largest continuously operating airline in Nevada for the two-decade duration of our data, January 2000 to December 2019. We chose to build forecasting models based on Southwest Airlines data to find the best model to forecast from January to June 2020.

Introduction

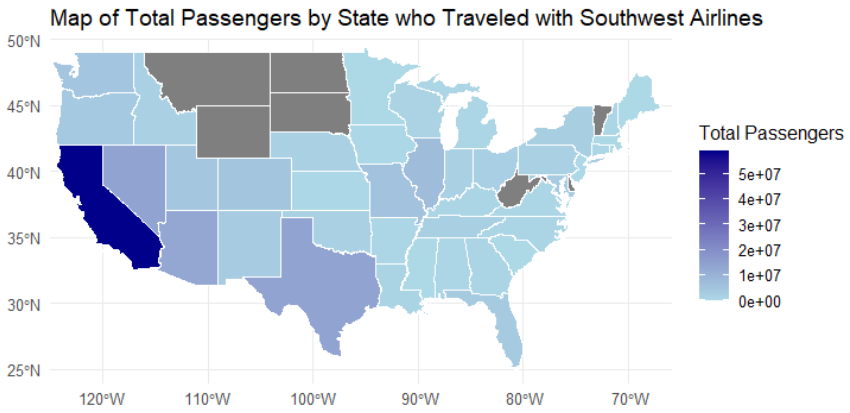
Data Cleaning and Review

To develop a predictive model for total monthly passenger traffic through Southwest Airlines in Nevada, we utilized various data analysis and machine learning techniques in R programming language. The methodology can be divided into three main steps: data manipulation, model creation, and model selection.

To start our data, "US_Monthly_Air_Passengers00_19.csv", contains over 6 million rows in 17 columns, containing air traffic data from across the continental United States from January 2000 through December 2019. The 17 feature columns are as follows:

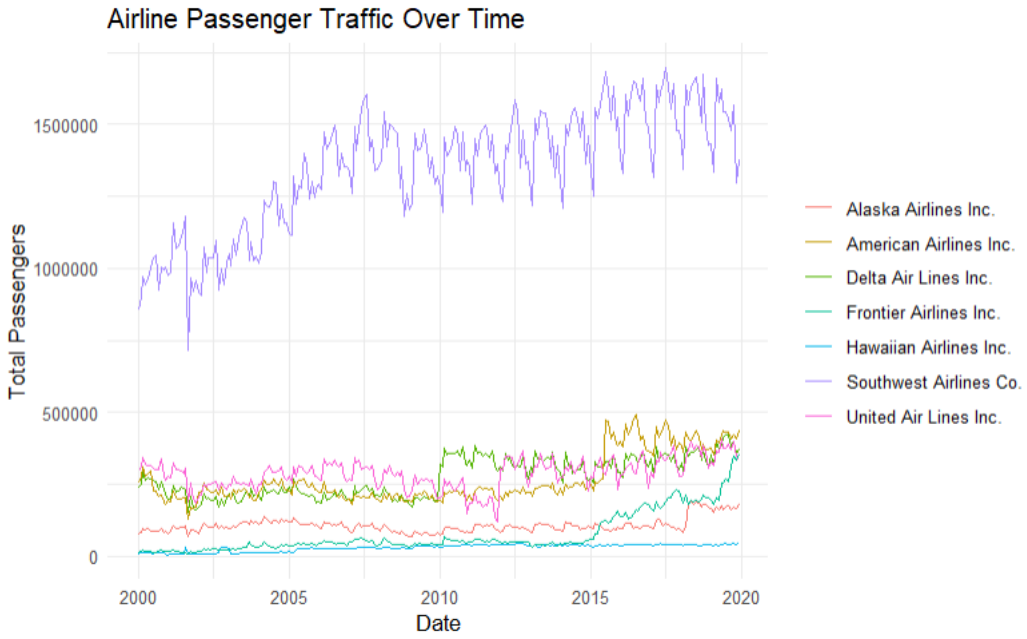
Variable	Description
Sum_PASSENGERS	Total number of passengers
AIRLINE_ID	Unique identifier for the airline
CARRIER_NAME	Name of the airline carrier
ORIGIN	Airport code of the origin airport
ORIGIN_CITY_NAME	Name of the origin city
ORIGIN_STATE_ABR	Abbreviation of the state where the origin is located
ORIGIN_STATE_NM	Full name of the state where the origin is located
ORIGIN_COUNTRY	Country code of the origin country
ORIGIN_COUNTRY_NAME	Name of the origin country
DEST	Airport code of the destination airport
DEST_CITY_NAME	Name of the destination city
DEST_STATE_ABR	Abbreviation of the state where the destination is located
DEST_STATE_NM	Full name of the state where the destination is located
DEST_COUNTRY	Country code of the destination country
DEST_COUNTRY_NAME	Name of the destination country
YEAR	Year of the flight record
MONTH	Month of the flight record

After reviewing our data, we filtered our dataset to only include records where both origin and destination state name were equal to Nevada. This narrows our data to only include records involving our state of interest. After this, we counted the number of flights inbound and outbound flights from each airport in Nevada and found Las Vegas and Reno accounted for over 90% of flight traffic. Also, most of these passengers flew with Southwest Airlines, which is why we chose them for our analysis. We likewise see that most of the passenger travel during the period of investigation occurs on the west coast of the U.S.



Now that we had an idea of which airline we would focus our attention on we needed to manipulate our data and conduct some feature creation before building our model. First, it would consist of the total sum of passengers for each airline in Nevada sorted by month. The airlines were referred to by their id, with Southwest being 19393. We then created a “MONTH” and

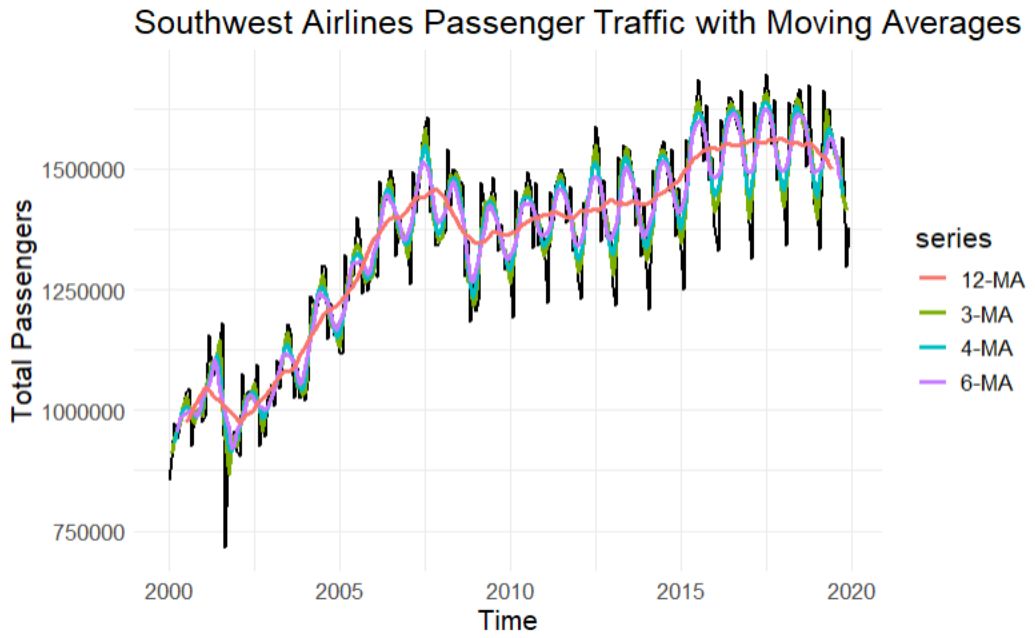
"YEAR" column and extracted their respective values from "datetime." This time-series pivot table started our data in the year 2000 moving at a frequency of 12.



From this image we see that Southwest Airlines, designated 19393, exhibits a strong seasonal trend across time. Additionally, they have the largest share of passengers in Nevada, far exceeding other airlines in the state.

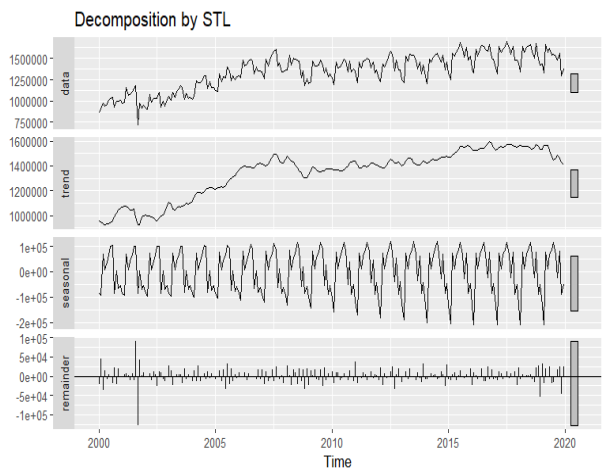
Model Creation

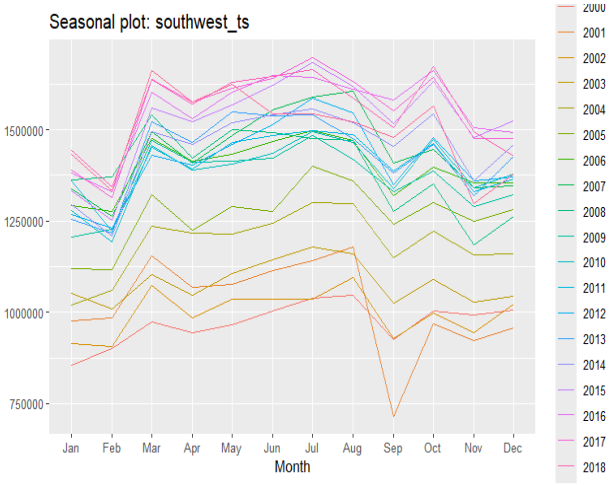
With this data we approached creating our first model. Starting simple, we wanted to model Southwest Airlines total monthly passenger transport, including inbound and outbound passengers. This would give us an idea of what time of year Southwest Airlines should prepare for higher volumes of passengers. For this, we created a time series object for Southwest, starting in 2000 with a frequency of 12, named "southwest_ts". We then plotted a moving average series at intervals of 3, 4, 6, and 12 months.



This graph reveals that there is a strong seasonal component to this data. Months 3, 4, and 6 exhibit a strong trend like the movement of the data, while the 12-month average shows a much smoother curve. This makes sense as the 12-month average would conceal seasonal movement.

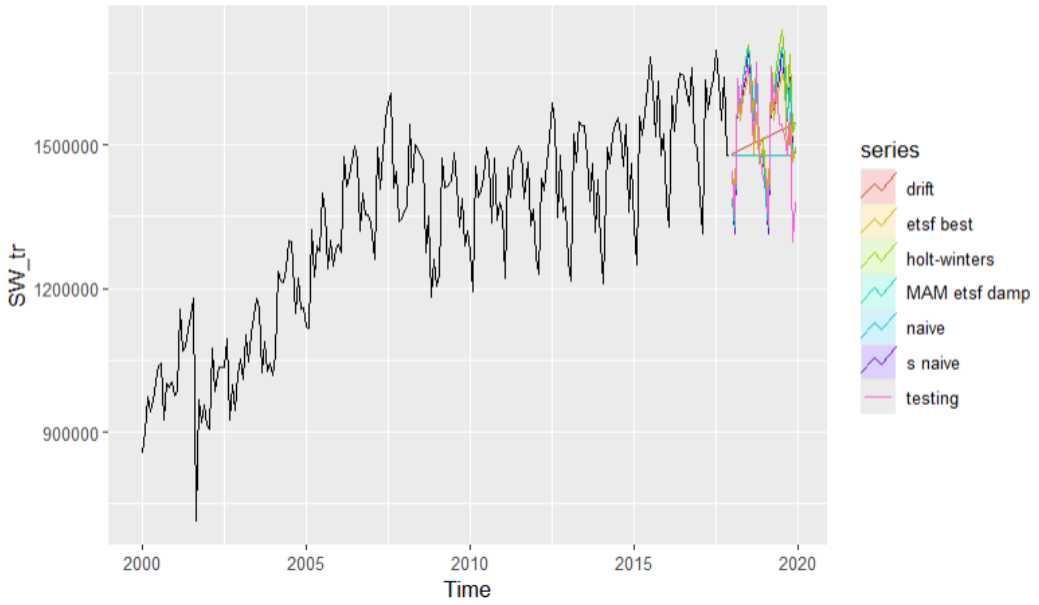
Following this we conducted decomposition of the Southwest Airlines time-series using X11, SEATS, and STL methods. Overall, there was slight variation between the three.





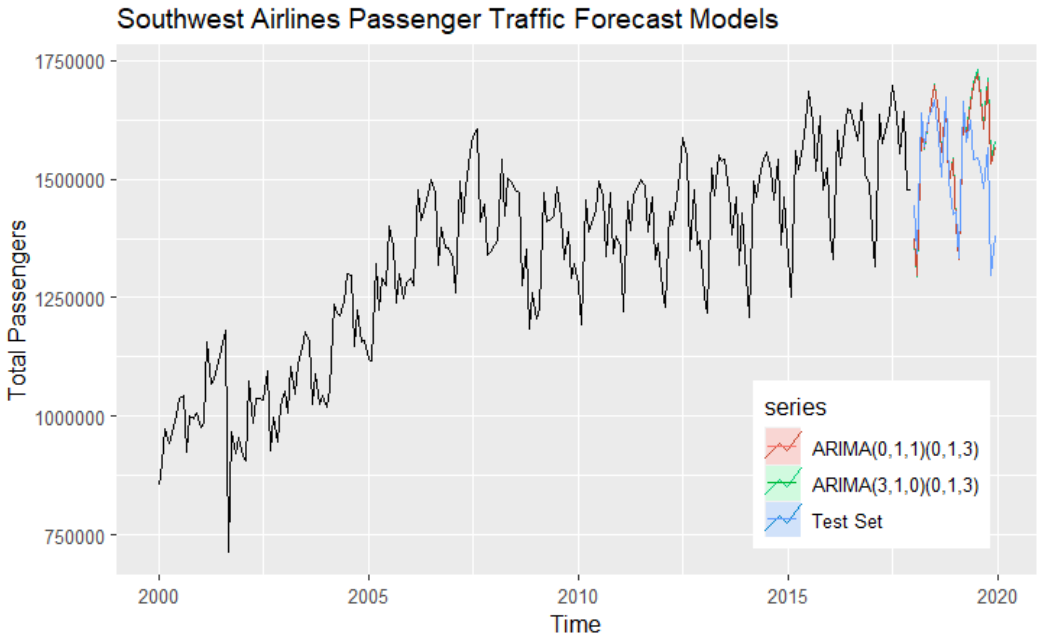
From this STL decomposition we see that the seasonal trend dominates the data throughout the 20 years. Visually, we see that there is some change beginning around 2010 where the trough of the seasonal cycle gradually becomes deeper through the end of 2019. We also note the peak months for air travel appear to be March, October, and the summer months climbing from May to peak in July. The slowest months for air travel are September, January, and February. The remaining residuals do not exhibit a white noise series. Now we built our first

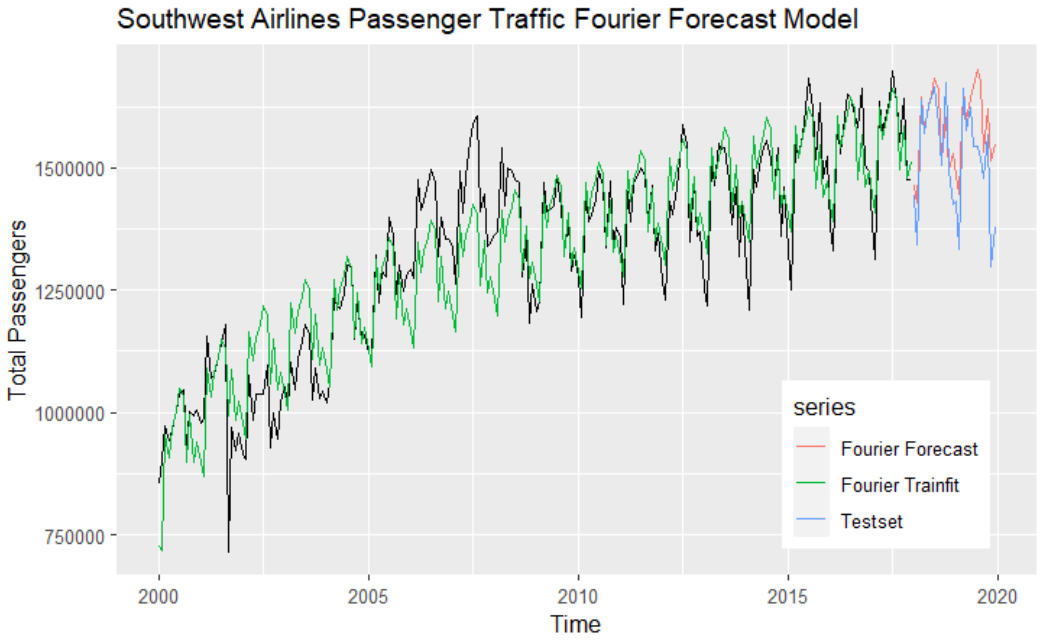
basic forecasting models, being Naïve, Seasonal Naïve, Drift, Holt-Winters, dampened ETS, and best ETS.



We see the naïve and drift models do not follow the seasonal pattern of the data, as expected. Excluding these two the other models may appear to be sufficient. However, only our dampened ETS model exhibits white noise, the other models do not. These results indicate that we needed to do more complex modeling. Moving on from these results we decided to check whether our data was stationary. Conducting and ADF test on our Southwest time series revealed it to be nonstationary, which could result in poor model performance or spurious correlations. To remedy this, we seasonally differenced the data with a lag of 12 to remove any seasonal or residual trend and then applied a first difference.

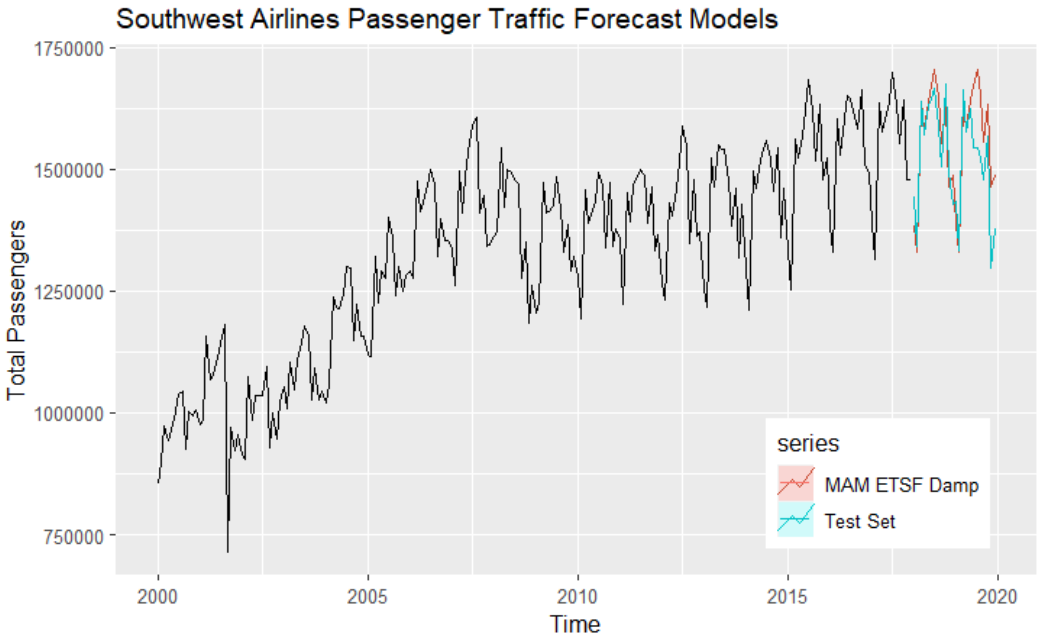
Now we have created an ARIMA model, Auto ARIMA model, and a Fourier model. All of which automatically detects and corrects for nonstationary trends in a dataset. The parameters of our ARIMA model were (3,1,0)(0,1,3) and our Auto ARIMA selected model was (0,1,1)(0,1,3). Both fit our data well, but our Auto ARIMA follows the data trend better. Our last model, the Fourier, used a seasonal pattern of 6. We see it follow the overall trend quite well, close to the shape of a square root function.

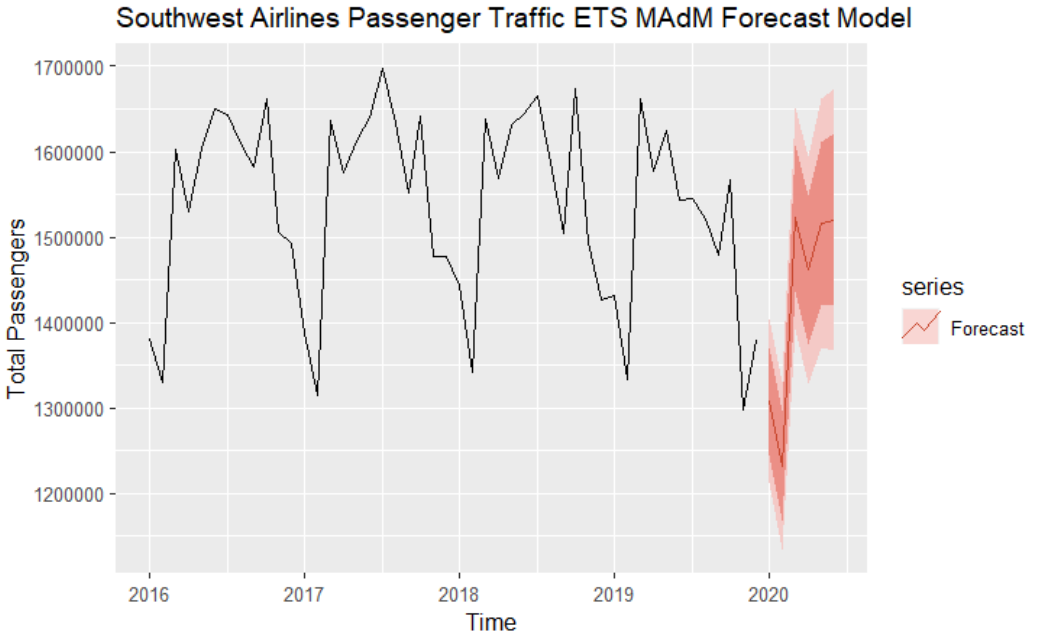




Model Selection

Of our created models, we have many candidates for predicting air traffic with Southwest through Nevada. When choosing our best model, disposing of some imprecise models is simple. As discussed earlier, some models such as our Naïve, Seasonal-Naïve, Drift, Holt-Winters, and Best ETS have residuals that do not exhibit white noise when observing their ACF plots. Additionally, despite our data looking like it would complement a Fourier model, it also does not exhibit white noise in its residuals. This leaves our ETS Dampened and our Auto ARIMA model.





To decide between these two, we used the Root Mean Squared Error metric (RMSE). This is a commonly used metric that measures how much, on average, a given model's prediction is different from the true value. Using the summary() function we were able to see that the ETS Dampened model has a better RMSE score, 73,470.62, compared to our Auto ARIMA RMSE of 98,461.

Model	RMSE Test	White Noise	P-Values
Naive	119,853.10	No	2.2e-16
Drift	117,480.10	No	2.2e-16
Seasonal-Naive	69,067.65	No	2.2e-16
Holt-Winters	98,172.53	No	0.00011
ETS Damp	73,470.62	Yes	0.05024
Best ETS	65,258	No	1.177e-07
Auto ARIMA (0,1,1)(0,1,3)	98,461	Yes	0.5009
Fourier	101,260.01	No	2.2e-16

Conclusion

Our best model for predicting Southwest Airlines month-to-month passenger averages in Nevada is our ETS Dampened model. However, our Auto ARIMA model is a close runner up. These were the only two models that exhibited white noise within their residuals. If we disregard the desire for white noise residuals, then we find the Best ETS has the lowest RMSE of all the models. Also, we include the forecast on the previous page for the ETS Dampened model as trained on the January 2000 to December 2019 Southwest timeseries to forecast from January to June 2020. We expect our model to be fairly accurate up until the end of March 2020 given that our model is trained domestically, and the U.S. had not shut down for SARS-COVID-19 until late March 2020. We should expect large deviations no later than April 2020 from the actual airline data.

Criticisms/Future Work

In our future work, we plan to expand the scope that our models capture in why passenger travel may change independently of seasonal trends. While our models have performed well and given valuable results, there are likely extraneous features not included in our models. Such an example could be the inclusion of an event feature, which could provide information such as whether an event is being held in Las Vegas, or a new casino or hotel has been opened. Such information could help airlines gage when demand for air travel will increase independently of the trends outlined in this report.