Las Vegas 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 the state of 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. Hence, we chose to build forecasting models based on the Southwest Airlines data to find a best model to forecast from January to June, 2020.

Introduction

Data Cleaning

To develop a predictive model for traffic in 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:

“Sum_PASSENGERS", "AIRLINE_ID", "CARRIER_NAME", "ORIGIN" "ORIGIN_CITY_NAME", "ORIGIN_STATE_ABR", "ORIGIN_STATE_NM" "ORIGIN_COUNTRY", "ORIGIN_COUNTRY_NAME", "DEST, "DEST_CITY_NAME", "DEST_STATE_ABR", "DEST_STATE_NM”, "DEST_COUNTRY", "DEST_COUNTRY_NAME", "YEAR", “MONTH”.

After reviewing our data, we filtered our dataset to only include records where ORIGIN_STATE_ABR or DEST_STATE_ABR 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. Additionally, a majority of these passengers were going with Southwest Airlines, which is why we chose them for our analysis.

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. This would mean creating a new dataset known as, “nv_complete_airlines”. To start 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 started our data in the year 2000 moving at a frequency of 12.

Model Creation

Model Selection

Analysis