Project 1: Prediction of Airline Prices

Aspen Fryberger

April 23, 2019

Websites like Kayak pride themselves on providing a wide variety of flight options from a variety of different airlines. They offer a selection of flights from a variety of different airlines and prices including flight options which require multiple airlines for a single trip. While Kayak definitely has an emphasis on finding the cheapest flight, the price is not the only important characteristic. They also allow individuals to select between different flight times, airline carrier, the number of stops, and the total travel time. I am looking at how these characteristics can be used to predict the prices airline tickets.

For this project I scraped data on one way routes flying from Atlanta Georgia to 220 different destinations both domestic and international for the week of May 19^{th}. These data included the price for the ticket if it was listed, the time of takeoff, the time in travel, the destination, and the number of stops required. Additionally, I also included the day of the week to capture differences in the travel tickets within a week. I then used the data to create a variable on the total number of airlines that operate on a single route and the number of flights in the week on an individual route to capture the effects that competition may have on a specific route's prices.

The data collected from Kayak have two types: flights with prices listed and flights without prices listed. Flights with prices lists the most economical price available and the flights without prices require additional search on the airline website. While flights with listed price are definitely the majority, this project will use the listed prices information as the training data set to best predict the unknown prices. I split the data into two groups. The first group is the group with listed prices which contains 16,184 flights, the majority of the fights collected. The descriptive statistics for these data are in table 1. Table 2 contains the 45 flights without prices.

While it is difficult to say from the descriptive statistics in tables 1 and 2 it appears that flights without prices tend to occur on routes with fewer flights, and shorter routes with fewer stops. Flights without prices don't appear to have much of a difference in the number of other carriers on the given route. While it is possible that flights without prices could come from a different data generating process than flights with prices, it is unlikely and is probably a results of different price strategies for individual airlines. Overall, there were 37 different carriers. Of the total 37 airlines, only Delta, Southwest, and United listed flights without immediately listing the price as well and these 45 flights fly to only 8 destinations.

Table 1: Flights With Prices

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Price	16,184	486	672	30	179	403	6,393
Flights per route	16,184	87.132	25.988	1	62	107	143
Carriers per route	16,184	4.502	1.711	1	3	6	10
Hour of take off	16,184	12.528	5.925	5	7	17	24
Stops	16,184	0.904	0.716	0	0	1	3
Flight time in minutes	16,184	435.809	361.336	35	146	571	3,898

Table 2: Flights Without Prices

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Flights per route	45	67.444	16.922	21	60	60	107
Carriers per route	45	4.711	1.392	4	4	5	8
Hour of take off	45	12.978	5.137	6	9	19	21
Stops	45	0.489	0.506	0	0	1	1
Flight time in minutes	45	214.089	94.243	90	135	255	462

To predict the price of the unknown airline tickets I use 6 models and compare the R^2 and the MSE. Using linear supervised learning I begin in Model 1 by simply including the destination of the route to predict the prices. The R^2 for this initial model is .707 with a MSE of 132,316. It is little surprise that the destination accounts for so much variation in the price because increased distance increases costs substantially. To improve the prediction, in Model 2, the carrier of the route was added to the destination to predict the price. This led to an increase in the R^2 to .919 and the MSE fell to 36,406. Again it is not really surprising that accounting for the airline has increase the predictive power of the model. Often airlines try to differentiate their product by offering rewards, nice seats, wifi, etc. instead of competing exclusively on price and this captures that. While the R^2 indicates that this is a pretty good model it is interesting to see how the other characteristics influence the prediction.

In Model 3 I add the day of the week of the flight to capture increased prices for popular days to fly. This addition led to a small increase in the R^2 to .922 and the MSE again fell from 36,406 to

Table 3: Results

Model	Characteristics	R^2	MSE
Model 1	Destination	0.707	132,316
Model 2	Destination, Carrier	0.919	36,406
Model 3	Destination, Carrier, Day of Week	0.922	35,196
Model 4	Destination, Carrier, Day of Week, Flight Hour	0.922	34,990
Model 5	Destination, Carrier, Day of Week, Stops	0.922	35,190
Model 6	Destination, Carrier, Day of Week, Num Carriers	0.922	35,193

35,196. While a somewhat smaller effect, it appears this does improve the prediction of the price. The 4^{th} Model adds the hour of the day that the flight takes off. However, this doesn't appear to improve the metrics much. The R^2 remains around .922, but the MSE falls some to 34,990. This may not be a good predictor of price because there are many different people who prefer flights at different times of the day meaning there may not be any strictly preferred hour for a plane to depart.

Alternatively, in Model 5 instead of using the hour to depart the number of stops is added to the model. While people basically prefer few stops, this again has little effect on the metrics of the model. The R^2 remained at .922 and saw little change. Last, in Model 6 the number of carriers on the route was added. This also led to minimal change in the metrics of the model. While this is not supported by the broader literature on airlines, in this case it may be that this effect is already being captured by including the destination and therefore doesn't improve the predictive power.

Ultimately, it appears Model 3 is the best Model in terms of getting a good prediction and not over fitting the data. As a result, Model 3 was used to predict the value of the missing prices. The results from this are included in the "results.csv" file. Overall the model has an R^2 of .92 which indicates it is a good model for predicting airline ticket prices. It makes sense that the destination, carrier, and the day of the week influence the price, but it was interesting how little impact the number of stops and the time of day had on predicting the price. Particularly the number of stops may be a characteristic people selecting flights may care a lot about and as a result could have been a better explanatory variable.