Final Project

Ryan Yu

Introduction and data

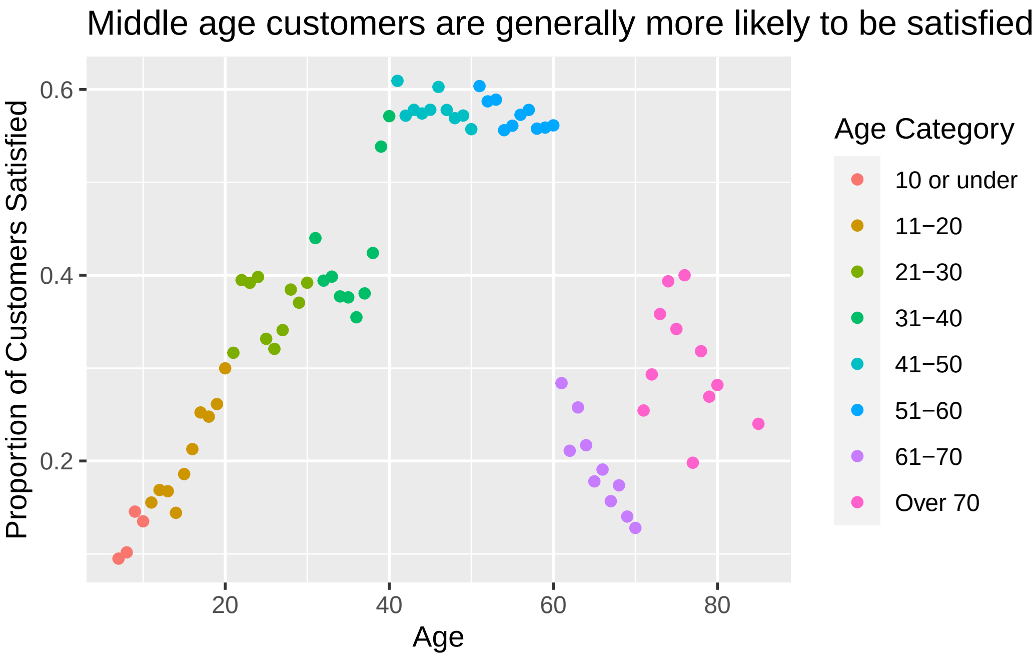
Air travel is an extremely popular form of transportation in the United States with over a million people flying every day. Everyone's experience is unique, whether they are travelling for work, vacation, school, etc. My research is motivated by the many different factors of air travel that contribute to a passenger's satisfaction. Understanding the factors that contribute to passenger satisfaction is crucial for airlines to improve their services. My research aims to answer the question: What are the most important factors that drive overall passenger satisfaction, and how effective are these factors at predicting overall customer satisfaction? My data set was collected through a US passenger satisfaction survey and compiled on Kaggle. The data set was originally split into "training" and "testing" data, which were random mutually exclusive parts of the same survey. In my research, the two data sets are recombined and an additional binary variable for satisfaction was added. Fourteen factors were included in the survey, with participants rating their satisfaction for each factor from 1 to 5, with 1 being least satisfied and 5 being most satisfied. Additionally, a rating of 0 corresponded to "Not Applicable", however I have changed the satisfaction levels of 0 to NA as to not skew the analysis. These factors include many different aspects of a passengers experience during a flight, from booking services to online check-in to the food offered during the flight, etc. Other variables include gender, customer type (loyal/disloyal), age, type of travel (business/person), class of travel, flight distance, departure delay, arrival delay, and overall satisfaction.

The table below shows the fourteen factors in the survey with the average of all responses (grouped by overall satisfaction as well as combined):

	Satisfied	Not Satisfied	Combined
Inflight wifi	3.393511	2.398750	2.813526
Departure/Arrival time	3.14203	3.28529	3.223411
Ease of online booking	3.244406	2.617090	2.883001
Gate location	2.972903	2.980055	2.976948
Food and drink	3.528888	2.961526	3.208034
Online checkin	4.153870	2.708061	3.33164
Seat comfort	3.966417	3.038039	3.441388

Inflight entertainment	3.964202	2.893142	3.358542
On-board service	3.856171	3.019742	3.383153
Leg room	3.834051	3.006488	3.366377
Baggage handling	3.966914	3.374912	3.632114
Check-in service	3.649004	3.043008	3.306293
Inflight service	3.970990	3.389832	3.642333
Cleanliness	3.746509	2.933359	3.28668

The average satisfaction score varies between 2.81 for in-flight WiFi service to 3.64 for in-flight service. For the most part, being overall satisfied corresponds to a higher average score for the factors (with departure/arrival time and gate location being the only two exceptions). Additionally, the factors have varying spreads between average scores of overall satisfied passengers and not satisfied passengers. For example, online checkin has a much higher average score when looking at overall satisfied passengers compared to not satisfied passengers while gate location has a very small difference in average scores when looking at overall satisfied compared to not satisfied.


Middle age customers are generally more likely to be satisfied

The probability of being satisfied seems highest for middle age (groups aged 41-50 and 51-60) people while younger and older passengers have a smaller probability of being satisfied. There is a notable difference in the proportion of satisfied passengers between age groups with a sharp drop off once the passenger age reaches 60 years old.

Methodology

For my analysis, I chose to use a logistic regression model. Since overall satisfaction is a binary variable (satisfied or not satisfied), a logistic regression model provides valuable comparisons between different values of the predictor variables in terms of odds ratios for satisfaction. The model will show which variables affect the probability of being satisfied the greatest. We can then test the model on our data to determine the effectiveness of our model at predicting overall satisfaction. For predictor variables, I used all fourteen factors in the survey, as well as age (categorized), flight distance, customer type (loyal/not loyal), arrival delay, gender, type of travel, travel class, and an interaction term between type of travel and travel class. The main effects were chosen as they are all factors that could affect a passenger's satisfaction in their flying experience. The interaction term is included because I believe that the relationship between travel class and satisfaction depends on the type of travel due to business travels having a stronger preference and desire to fly business class. Departure delay is not included in the model because the departure delay does not matter as much as the arrival could still be on time, therefore I only included the arrival delay in the model since the two variables are highly co-linear and arrival delay matters much more to passengers than departure delay.

The original data set had missing data points that have been addressed through Multiple Imputation via Chained Equations (MICE). The missing data seems random throughout and could probably be explained by some flights not offering some of the factors (for example, some flights don't offer in flight entertainment).

Assess conditions and diagnostics

Afterwards, in order to determine the model's effectiveness, I calculated the sens/spec/etc... fill in later.

Results

$$\widehat{Satisfaction}\ odds = e^{-10.72 + 0.0353*male + 2.34*loyal\ customer + 0.0000043*flight\ distance + (-3.67)*personal\ travel}$$

$$+ (-0.98)*economy + (-1.10)*economy\ plus + 0.80*inflight\ wifi\ service + (-0.29)*time\ convenience + 0.27*ease\ of\ online\ booking$$

$$+ (-0.22)*gate\ location + (-0.05)*food\ and\ drink + 0.85*online\ checkin + 0.015*seat\ comfort + 0.053*inflight\ entertainment$$

$$+ 0.32*onboard\ service + 0.26*legroom + 0.13*baggage\ handling + 0.34*checkin\ service + 0.13*inflight\ service + 0.23*cleanliness$$

$$+ (-0.0047)*arrival\ delay + 0.36*11\ to\ 20\ years\ old + 0.42*21\ to\ 30\ years\ old + 0.13*31\ to\ 40\ years\ old + 0.44*41\ to\ 50\ years\ old$$

$$+ 0.36*51\ to\ 60\ years\ old + (-0.092)*61\ to\ 70\ years\ old + (-1.08)*over\ 70\ years\ old$$

$$+ 0.86*personal\ travel*\ economy + 0.76*personal\ travel*economy\ plus$$

Out of the fourteen factors, in-flight WiFi service, online check-in, boarding service, leg room, check-in service have relatively high magnitude slope coefficients while controlling for all other variables, meaning that changes in the scores of these factors have the greatest impact on the predicted probability of overall satisfaction. Meanwhile, seat comfort and food and drink had relatively low magnitude slope coefficients while controlling for all other variables, meaning that changes in the scores of those factors have a relatively smaller impact on the predicted probability of overall satisfaction. Being a loyal customer = more satisfy. do math on these later