DS740

Final Project

Jeff Watson

In 2023, international airlines are expected to post a profit, in aggregate, of $4.7B. This is the first profit posted by airline since 2019][1]. This is also just a 0.6% net profit margin. It is imperative that airlines retain existing customers and entice as many new ones as possible.

The data for my final project is from surveys filled out by airline passengers describing their experience on a recent flight. They were ultimately asked about their overall satisfaction. The objective of this project is to accurately predict the satisfaction of the passengers. This data could be used by airlines to make financial decisions about where to reinvest earnings to maximize customer satisfaction and retention.

The data set contains 129,880 completed customer satisfaction surveys. Of those, 129,487 were completely complete. 393 of the surveys had missing information. I will remove those from this analysis as they make up such a fraction of a percent of the data. Below are the items from the survey. Note that a 0 for any of the scored questions indicates an NA for that flight.

* Airline satisfaction level (Satisfaction, neutral or dissatisfaction)
* The actual age of the passengers
* Gender of the passengers (Female, Male)
* Purpose of the flight of the passengers (Personal Travel, Business Travel)
* Travel class in the plane of the passengers (Business, Eco, Eco Plus)
* The customer type (Loyal customer, disloyal customer)
* The flight distance of this journey
* Satisfaction level of the inflight Wi-Fi service   
  (0-5)
* Satisfaction level of online booking (0-5)
* Satisfaction level of inflight service (0-5)
* Satisfaction level of online boarding (0-5)
* Satisfaction level of inflight entertainment (0-5)
* Satisfaction level of Food and drink (0-5)
* Satisfaction level of Seat comfort (0-5)
* Satisfaction level of On-board service (0-5)
* Satisfaction level of Leg room service (0-5)
* Satisfaction level of Departure/Arrival time convenient (0-5)
* Satisfaction level of baggage handling (0-5)
* Satisfaction level of Gate location (0-5)
* Satisfaction level of Cleanliness (0-5)
* Satisfaction level of Check-in service (0-5)
* Departure Delay in Minutes
* Arrival Delay in Minutes
* Flight cancelled (Yes, No)
* Flight time in minutes

I will quickly check the correlations of the variables in the data set to make sure we don’t have any issues with that. There are a couple of small pockets of correlation. Seat.comfort, Food.and.drink, and Inflight.entertainment make up the first cluster and Onboard.service, Baggage.handling, and Inflight.service make up the second. Both of those clusters make sense intuitively. Someone concerned about seat comfort might expect more from other material comforts on the flight.

Chart, scatter chart

Description automatically generated

I will employ two machine learning methods for predicting customer satisfaction- random forests and artificial neural networks. Random forests are large collections of randomly selected decision trees created with the data. For my cross-validation/parameter tuning, I will test multiple variables to be considered at each point for the random forest and will test multiple sizes and decays for the artificial neural network.

The three most important predictors of customer satisfaction according to the random forest are Inflight.wifi.service, Online.boarding, and Checkin.service. An interesting phenomenon is revealed when you look at the partial dependence plot for Inflight.wifi.service. As mentioned earlier, a value of zero for the survey items indicates an NA. What this allows is flights without wifi (Inflight.wifi.service == 0) to have higher satisfaction scores than those with poor wifi service (Inflight.wifi.service %in% c(1,2,3)). I looked through the other survey items to check for this U-shaped partial dependence plot. Ease.ofOnline.booking, Food.and.drink, Online.boarding, Seat.comfort, and Leg.room.service all showed some degree of NA scoring higher satisfaction than low scores (1, 2, 3).

Graphical user interface

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

Chart

Description automatically generated

As there are a few categorical variables in this data set, there is a lot of opportunity for subset analysis. And with the size of the data set being so large, there is a reasonable likelihood that the finding will have enough observations to support the findings. I went through to check the most important predictors for the subgroups and found that Inflight.wifi.service, Online.boarding, and Checkin.service were unsurprisingly near the top for all groups.

Chart, bar chart

Description automatically generated

1. https://www.iata.org/en/pressroom/2022-releases/2022-12-06-01/