Data Analysis Assignment 4

Due: 11:55pm, Friday, Nov 18

Instructions

This assignment involves generalized linear models. The data can be found on Sakai: go to Resources → Data Analysis Assignment Datasets → Assignment 4. You are required to type your solutions using R Markdown. You will submit 1) a PDF produced from R Markdown with answers to the questions, and 2) your Rmd file. Submissions should be made on gradescope: go to Assignments → Data Analysis Assignment 4 and Assignments → Data Analysis Assignment 4 RMD CODE.

DO NOT INCLUDE R CODE OR OUTPUT IN YOUR SOLUTIONS/REPORTS *All R code must be submitted separately, and R outputs should be converted to nicely formatted tables. Feel free to use R packages such as* `kable`*,* `xtable`*,* `stargazer`*, etc. Please consult the example report and corresponding Rmd file for guidance.*

Also, you should round up ALL numbers/estimates to 2 decimal places (4 decimal places at the most to avoid exact zeros when possible).

Reminder: You are allowed and even encouraged to talk to each other about general concepts, or to the instructor/TAs. However, the write-ups, solutions, and code MUST be entirely your own work.

The exercise is based on the airline customer satisfaction dataset found here: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.

The data for this question can be found in the file "airline_survey" on Sakai. Note that you are analyzing a modified version of the dataset found on kaggle.

Code book

Variable	Description
Age	Passenger age
Gender	Passenger gender
Type.of.Travel	Purpose of the passenger's flight (personal, business)
Class	Travel class in the plane (business, eco, eco plus)
Customer.Type	Loyal or disloyal customer
Flight.Distance	Flight distance
Inflight.wifi.service	Satisfaction level of inflight wifi service (0: Not applicable; 1-5 where 5 is completely satisfied)
Ease.of.Online.booking	Satisfaction level of online booking
Inflight.service	Satisfaction level of inflight service
Online.boarding	Satisfaction level of online boarding
Inflight.entertainment	Satisfaction level of inflight entertainment
Food.and.drink	Satisfaction level of food and drink
Seat.comfort	Satisfaction level of seat comfort
On.board.service	Satisfaction level of On-board service
Leg.room.service	Satisfaction level of leg room
Departure.Arrival.time.convenient	Satisfaction level of departure/arrival time convenience
Baggage.handling	Satisfaction level of baggage handling
Gate.location	Satisfaction level of gate location
Cleanliness	Satisfaction level of cleanliness
Checkin.service	Satisfaction level of check-in service
Departure.Delay.in.Minutes	Departure delay in minutes
Arrival.Delay.in.Minutes	Arrival delay in minutes
Satisfaction	Airline satisfaction level (satisfied, neutral, dissatisfied)

An airline called LaneAir is seeking a data scientist consultant to better understand drivers of customer satisfaction. The airline distributed a survey to customers who have flown with LaneAir in the last six months. Customers rated their overall satisfaction as dissatisfied, neutral, or satisfied. Then they rated their satisfaction with various aspects of the flight. The airline also has information on the passengers' flight details, including flight distance and departure delay. LaneAir would like to know which services are worth investing in to improve customer satisfaction. However, they would like the data science consultant to keep in mind that some services are more difficult to improve than others. LaneAir is considering the following investments, ranked from **most difficult** to **least difficult** to implement:

- Newer, larger seats to improve seat comfort and leg room. This will reduce the number of seats per plane, which is LaneAir's last choice
- Newer plane models that improve reliability to minimize delays
- Hire more flight attendants or other staff to improve services including inflight service, cleanliness, or onboarding.
- Technology investment to improve wifi and entertainment service

- Marketing or promotion initiatives to improve customer loyalty or appeal to different customer types (e.g., different age demographic, business/personal travelers)

You will complete this assignment in two parts. First, you will create a 2-page report **for the client.** This report should be understood by LaneAir executives with very little understanding of statistics. This report should include the following:

- Introduction: Provide an overview of the dataset and the goals of the analysis. Keep in mind that the LaneAir team is familiar with the questions on the survey, but not the results. So you should include, for example, basic summary statistics to show the team the distribution of customers in the different satisfaction categories.
- Methods: Explain the model you used to analyze the data without getting into technical details. Why did you decide to use that model for this dataset and how does it answer the airline's question?
- Results: What are the key results of the analysis as they relate to the airline's question? Present at least one figure that effectively communicates a key takeaway of the analysis.
- Conclusion: Keeping in mind LaneAir's cost considerations outlined above, what are your recommendations for how they can balance impact with cost? Do you have any recommendations that the airline has not considered? Finally, are there any limitations that the client should be aware of? For example, could certain customers be more likely to respond to the survey than others?

The second part of the assignment will be a 3-4 page report that is suitable for **other data scientists.** Here, you will present details of your model to justify the conclusions you presented to the client. This section should present technical details that someone with a data science background can understand. This report must include the following, though you may wish to provide additional details relevant to the analysis:

- Data overview and analysis plan: You should present details of the data that were not included in part 1, for example an appropriate distributional assumption for the outcome variable. Then, present the type of model you used for the analysis. Which type of generalized linear model is best suited for this problem? What is the link function?
- Model results: Present a table of model results including odds ratios, confidence intervals, and p-values. Interpret the results that you think are most compelling.
- Model assessment: Present the confusion matrix and explain your conclusion for the model's predictive accuracy. Additionally, the model you should use for this analysis relies on a key assumption that is unique to this model. Using a different, more "precise" model, compare predictions using the predictors `Gender` and `Customer type`. Show the confusion matrix for the more precise model and compare the accuracy.
- Conclusion: What do you conclude about the validity of this analysis?

Grading

40 points