Final Project Report

Tommy Tseng

11/19/2020

1. Summary

1.1 Inferential Questions

1. What are 3 factors that affect passenger's satisfaction the most? 3 factors that affect passengers' satisfaction the least?
2. From the proposed model, are there any predictors that are often overlooked by airline companies but end up quite significantly impacting passenger's satisfaction? Any commonly emphasized predictors that end up not positively impacting passengers' satisfaction?
3. What are your overall recommendations to the airline companies?

1.2 Method & Final Model

As the response variable for this analysis (passenger satisfaction) was recorded in two levels, satisfied or neutral/dissatisfied, logistic regression with multiple predictors was chosen as the modeling method. Below is the final model:

$$\begin{gathered} Logit(\pi_i) = \beta_0 + \beta_1\ OnlineBoarding_i + \beta_2\ Travel\ Type_i + \beta_3\ Inflight\ Service_i + \beta_4\ Class_i + \\ \beta_5\ Ease\ of\ OnlineBooking_i + \beta_6\ Seat\ Comfort_i + \beta_7\ Cleanliness_i + \beta_8\ Flight\ Distance\ Cent_i + \\ \beta_9\ Age_i + \beta_{10}\ Food\ and\ Drink_i + \beta_{11}\ Gender_i + \beta_{12}\ log\ (TotalDelay)_i + \beta_{13}\ Travel\ Type : AgeCent_i \end{gathered}$$

Note: Except log(Total_Delay), FlightDistanceCent, and AgeCent from TravelType:AgeCent are continuous predictors, other predictors are categorical.

1.3 Most Important Results to Inferential Questions

1. The airline company should focus on promoting to business travelers and building good online boarding experience.
2. At the same time, don't spend too much effort improving seat comfortableness and ease of online booking.

2. Introduction

As many industries are becoming more and more competitive, many companies are actually making use of big data to relocate resources and to maximize profit by providing goods and services that are valued most by most customers. This is not an exception for the airline industry. In this project, a passenger satisfaction data set is provided by an airline company. I will investigate the associations between different rated/recorded factors and passenger's satisfaction, and will come up with a model that recommends the top 3 and least 3 factors this airline company should focus on when relocating their resources.

3. Data

3.1 Data Description & Pre-Processing

Data Source: The data was obtained from https://www.kaggle.com/johndddddd/customer-satisfaction, and the number observations is 129881.

Variables Included: In the original data set, there were 21 predictors initially. To prevent multi-collinearity, predictors that were similar in nature were removed. For example, wifi service, food and drink, and inflight entertainment can be categorized into the entertainment and food category of the predictors, so only food and drink was kept as the representative predictor from the category as most flights provide food and drink regardless of the flight distance. In addition, departure delay in minutes and arrival delay in minutes were summed up as total delay and was later transformed to log total delay for EDA purpose.

Thirteen predictors were chosen for this analysis. These predictors were Gender (male vs female), Age, Customer Type (loyal vs disloyal), Travel Type (personal vs business), Class (economy, economy+, business), Flight Distance, Food and Drink (rated from level 0-5), Ease of Online Booking (rated from level 0-5), Online Boarding (rated from level 0-5), Seat Comfort (rated from level 0-5), Inflight Service (rated from level 0-5), Cleanliness (rated from level 0-5), and log total delay (minutes).

In order to prevent multi-collinearity, continuous variables were mean-centered. Besides, to ensure minimum amount of data at each level, categorical level that had less than 150 observations was dropped. In this case, level 0 from Food and Drink, Seat Comfort, Cleanliness, and Inflight Service were dropped.

3.2 Exploratory Data Analysis

Due to the iterative nature of modeling process, two rounds of EDA were performed before and after the main predictors modeling. For the first round, each predictor was plotted against the response variable to find interesting patterns. After the main predictors modeling, the focus shifted to finding interesting patterns from potential interaction terms. First, the relationship between the response variable and each continuous predictor swapped by other ten categorical predictors were examined. Next, the relationship between the response variable and each categorical predictor swapped by other categorical predictors were investigated. Due to large number of categorical predictors, the categorical predictors were categorized into three groups based on their distinct nature, and representative categorical predictor from each group was selected for investigating the potential interaction between categorical predictors.

The first category was based on customer demographics, and included categorical predictor was Gender (female vs male). The second category was based on customer's choices, and included categorical predictors were Customer Type (loyal vs disloyal), Travel Type (personal vs business), Class (economy, economy+, business). The third category was based on the rated airline services, and included categorical predictors were Food and Drink (rated from level 1-5), Ease of Online Booking (rated from level 0-5), Online Boarding (rated from level 0-5), Seat Comfort (rated from level 1-5), Inflight Service (rated from level 1-5), and Cleanliness (rated from level 1-5). Three selected categorical predictors were Gender, Travel Type, and Seat Comfort.

One interesting pattern found was that the probability for business travel passengers to report satisfied was 48% higher than that of personal travel passengers'. (See the table below)

```
##                          Travel_Type_Fact
## Satisfaction_Fact         Personal Travel Business travel
##   neutral or dissatisfied        0.899022       0.4162489
##   satisfied                      0.100978       0.5837511
```

4. Model

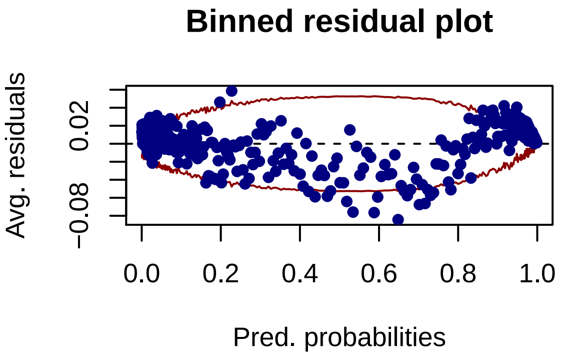
4.1 Model Selection & Model Assessment

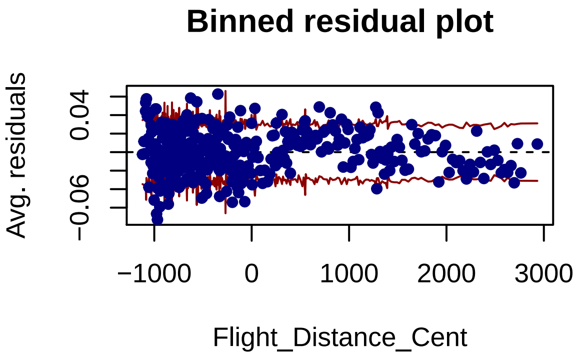
To fit a model that satisfies with assumptions and cover the most predictors, three rounds of model selection and assessment were conducted. To start with, all main predictors were fitted as a baseline model. This baseline model was used to see if there were any major assumption violations before adding any potential interaction terms. If assumption violations exist, predictor transformation, adding higher order term, or adding potential interaction terms will be considered as the result. From the result of assumption check, both the fitted values vs raw residual and mean-centered age vs raw residual plots had quite a lot of points outside of the red bound area.

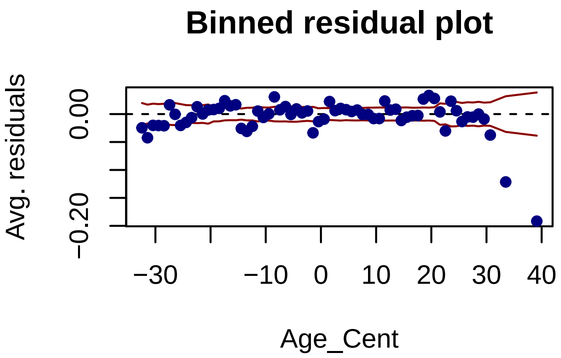
In response to this, Age was categorized into 5 levels based on the distribution of the data set. Also, EDA on the potential interaction terms was performed. Interaction is suspected when there is difference in trend across each level/group of comparison. In this case, eight potential interaction terms were identified. They were Travel Type:Flight Distance, Travel Type:Age, Ease of Online Booking:Age, Seat Comfort:Age, Inflight Service:Flight Distance, Class:Flight Distance, Customer Type:Flight Distance, and Customer Type:Age.

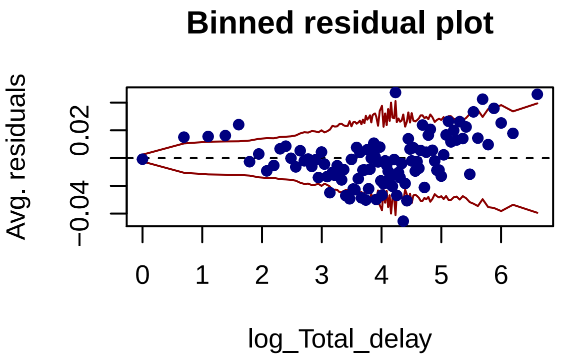
Next, on top of the baseline model, additional transformed term/higher order term/potential interaction term was added one at a time, and assumption check was performed correspondingly. This step was used to figure out the model that satisfied the assumptions the most while covering as many predictors as possible. After several rounds of trial, the tuned model was used as the input for step-wise model selection to find out a model that had the minimized BIC. Below are the final model and corresponding binned residual plots:

$$
\begin{aligned}
Logit(\pi_i) = \beta_0 + \beta_1\ OnlineBoarding_i + \beta_2\ Travel\ Type_i + \beta_3\ Inflight\ Service_i + \beta_4\ Class_i + \\
\beta_5\ Ease\ of\ OnlineBooking_i + \beta_6\ Seat\ Comfort_i + \beta_7\ Cleanliness_i + \beta_8\ Flight\ Distance\ Cent_i + \\
\beta_9\ Age_i + \beta_{10}\ Food\ and\ Drink_i + \beta_{11}\ Gender_i + \beta_{12}\ log\ (TotalDelay)_i + \beta_{13}\ Travel\ Type : AgeCent_i
\end{aligned}
$$


Binned residual plot


Binned residual plot


Binned residual plot


Binned residual plot

4.2 Model Validation & Result Interpretation To validate the model, ROC curve was drawn to find out the classification performance at all classification thresholds. The computed AUC for the best threshold (0.472) is 0.944, along with a sensitivity of 0.89 and a specificity of 0.85. Also, vif was used to check whether multi-collinearity was an issue in the continuous predictors, and all the continuous predictors had a vif below 5.

To answer the inferential questions, the summary table below was used. Only predictors having p-value smaller than 0.05 were interpreted.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.05	0.08332	-24.61	1.065e-133
Online__boarding__Fact1	-1.311	0.08913	-14.71	5.276e-49
Online__boarding__Fact2	-1.411	0.08815	-16.01	1.156e-57
Online__boarding__Fact3	-1.378	0.08695	-15.84	1.548e-56
Online__boarding__Fact4	0.6656	0.08602	7.737	1.018e-14
Online__boarding__Fact5	2.378	0.08887	26.76	9.299e-158
Travel__Type__FactBusiness travel	2.732	0.02906	94.01	0
Inflight__service__Fact2	-0.09579	0.04231	-2.264	0.02356
Inflight__service__Fact3	-0.2172	0.03953	-5.494	3.925e-08
Inflight__service__Fact4	0.9488	0.03689	25.72	7.046e-146
Inflight__service__Fact5	1.654	0.03907	42.33	0
Class__FactEco Plus	0.2102	0.03716	5.656	1.546e-08
Class__FactBusiness	0.803	0.02379	33.75	9.337e-250
Ease__of__online__booking__Fact1	-2.389	0.07565	-31.58	8.099e-219
Ease__of__online__booking__Fact2	-2.625	0.07521	-34.91	5.792e-267
Ease__of__online__booking__Fact3	-2.548	0.07494	-34	2.557e-253
Ease__of__online__booking__Fact4	-1.984	0.07468	-26.56	1.873e-155
Ease__of__online__booking__Fact5	-1.328	0.07558	-17.57	3.861e-69
Seat__comfort__Fact2	-0.207	0.0474	-4.368	1.252e-05
Seat__comfort__Fact3	-1.069	0.0442	-24.18	3.678e-129
Seat__comfort__Fact4	-0.434	0.04324	-10.04	1.054e-23
Seat__comfort__Fact5	0.02651	0.04575	0.5795	0.5622
Cleanliness__Fact2	0.06616	0.04673	1.416	0.1569
Cleanliness__Fact3	0.7538	0.04288	17.58	3.589e-69
Cleanliness__Fact4	0.6745	0.04288	15.73	9.288e-56
Cleanliness__Fact5	0.9044	0.04691	19.28	7.738e-83
Flight__Distance__Cent	0.0002605	1.051e-05	24.78	1.397e-135
Age__Fact2	-0.002707	0.03986	-0.06791	0.9459
Age__Fact3	0.5353	0.06032	8.874	7.038e-19
Age__Fact4	0.5631	0.08238	6.835	8.225e-12
Age__Fact5	0.1668	0.1113	1.499	0.134
Food__and__drink__Fact2	0.4493	0.04471	10.05	9.38e-24
Food__and__drink__Fact3	0.4959	0.0442	11.22	3.312e-29
Food__and__drink__Fact4	0.6872	0.0439	15.65	3.122e-55
Food__and__drink__Fact5	0.2393	0.04516	5.298	1.169e-07
Gender__FactFemale	-0.08562	0.01861	-4.6	4.222e-06
log__Total__delay	-0.1415	0.005044	-28.05	4.165e-173
Travel__Type__FactPersonal Travel:Age__Cent	-0.02385	0.002456	-9.711	2.703e-22
Travel__Type__FactBusiness travel:Age__Cent	0.01473	0.002515	5.858	4.678e-09

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	177095 on 129350 degrees of freedom
Residual deviance:	76492 on 129312 degrees of freedom

Result Interpretation

1. What are 3 factors that affect passenger's satisfaction the most?

Ans: Holding other predictors constant, the top three factors that affect passenger's satfisfaction the most in odds scale are Travel Type Fact Business travel (14.37 times greater compare to personal travel) , Online boarding Fact5 (9.78 times greater compare to passengers who rated online boarding as 0), Inflight service Fact5 (4.23 times greater compare to passengers who rated inflight service as 1) .

2. What are 3 factors that affect passenger's satisfaction the least? In other words, even if a passenger rates a particular predictor with a very high score, the corresponding predictor still doesn't bring much increase, or even decrease to getting a satisfied feedback.

Ans: To answer this question, we will look for three predictors with lowest coefficients in odds scale. If the predictor is categorical and also belongs to service category, we will only look at level 4 and 5 assuming airline company will have to allocate extra resources into a specific service in order to get rating score higher than or equal to 4, which is above average score of 3. Holding other predictors constant, three factors that affect passenger's satisfaction the least are Ease of online booking Fact4 (86% chance lower compare to passengers who rated ease of online boarding as 0), Ease of online booking Fact5 (74% chance lower compare to passengers who rated ease of online boarding as 0), and Seat comfort Fact4 (35% chance lower compare to passengers who rated seat comfort as 1) .

3. From the proposed model, are there any predictors that are often overlooked by airline companies but end up quite significantly impacting passenger's satisfaction?

Ans: Online boarding experience, Business travelers

4. From the proposed model, are there any predictors that are often emphasized by airline companies but end up quite insignificantly impacting passenger's satisfaction?

Ans: Seat Comfort, Ease of Online Booking

5. What are your overall recommendations to the airline companies?

Ans: Focus on promoting to business travelers and building good online boarding experience. At the same time, don't spend to much effort improving seat comfortableness and ease of online booking.

5. Conclusions

5.1 Important Findings

1. The airline company should focus on promoting to business travelers and building good online boarding experience.
2. At the same time, don't spend to much effort improving seat comfortableness and ease of online booking.

5.2 Limitations

1. For the binned residual plots of fitted values vs raw residual, some data points are still outside of the red bound area for the final model.
2. Seasonal impact was not included in the data set which might lead to different data distribution and inference result. (peak season vs off season)