GRADTDA 5620 Random Forest Assignment

Brian M Wisniewski

2023-06-17

Introduction

An airline wants to identify some opportunities to improve customer satisfaction and to identify the most dissatisfied customers to target for coupons. This report intends to address those wants by examining the airlines data set, which contains variables related to customer satisfaction and flights. Plots, context, and the use of random forest modeling is used and take-aways compiled in this report.

1) Identify Some Opportunities to Improve Customer Satisfaction

There are opportunities to improve customer satisfaction related to the variables in Table 1. In Table 1, the variables are listed in descending order of their relative importance. To assess specific actions that can be taken, would require collaboration between financial analysts and airline specialists to determine ROI and feasibility of action. Some example actions, for the highest importance variables, are as follows:

Age: If there are opportunities to upgrade passengers to a seat with more leg room, then it may be good to select older passengers. There are potentially concerns with discrimination, but this is hopefully mitigated by the defense that older people generally have worse leg/hip joints. That defense hopefully avoids accusations of ageism.

Ease of Online Booking, Online Support, online boarding: Cost benefit analysis could be done to see if the upfront cost of hiring designers and engineers to improve the online processes is worth the improvement in customer satisfaction.

Class: Ensure that class upgrades are given, if they are available and/or use class upgrades as incentives every "X" number of miles flown with the airline.

Cleanliness: Cost benefit analysis could be done to see if the upfront cost of more thoroughly cleaning planes between flights, thus potentially delaying the flight, is worth it. A similar analysis could be done to see if more cleaning staff should be employed.

Baggage Handling: If this is due to poor interactions with staff, then additional training could be provided on customer relations.

2) Identify the Most Dissatisfied Customers to Target for Coupons

Because the airline wants to target the most dissatisfied customers, I would recommend prioritizing only people associated with variables with the highest mean decrease gini in Table 1. Variables associated with highest mean decrease gini, are more powerful predictors of customer satisfaction.The computation at the end of this report utilizes the predict() function and these variables. It can be used here.

Though, it's not clear that all variables should be considered. Age should not be considered due to risk of discrimination accusations. In 1), an alternative approach is provided to appease older flyers. Additionally, certain variables are unlikely to be influenced by a coupon. For example, individuals who are dissatisfied due to uncleanliness probably do not care about a coupon- they simply want a cleaner flight. Despite this, my recommendation is to trial offering coupons to all people associated with the variables included in Table 1- excluding age. When collecting data during these trials, it will be important to control for confounding variables that may affect customer satisfaction. For example, flying around a holiday probably decreases customer satisfaction. To minimize confounding variables, I would recommend randomly sampling flights across a full year. On those flights, I would randomly sample passengers that belong to cohorts associated with the Table 1 variables -excluding age- and offer them coupons. I would also randomly sample passengers to not give coupons, to act as a control. After a year, I would perform additional statistical analysis and re-evaluate the coupon policy.

Exploratory Analysis

Exploratory analysis was aimed at reducing the number of variables used in further analysis. Variables were excluded either on the basis of contextual real-world justifications or lack of apparent meaningful association with customer satisfaction. The former was assessed using subjective judgement by the analyst, regarding the data set description. The latter was assessed via box plots for continuous variables and bar charts for categorical variables. All variables in the data set were plotted, but excluded variables are omitted from this report. Figures 1 and 2 show aforementioned plots. Below are informal notes taken by the analyst, including justifications for his judgement.

Variables to Model:

- Class: The airline could upgrade the most unsatisfied customers. - Cleanliness: More time could be spent on more thorough cleaning between flights. The impact on satisfaction looks somewhat minimal, but could be worth investigating. - Online Boarding: More time/resources could be spent on improving the online experience. It could be advertised more and potentially even incentivized with cheaper flights. - Check-in service: The airline could query customers to see what could be improved here. It may be something simple, such as untrained staff. Perhaps more staff training could be implemented. - Checked Baggage handling: The airline could query customers to see what could be improved here. I doubt this is something the airline can fix, but maybe it's something fixable like untrained staff involved in the process. - Ease of Online Booking: More time/resources could be spent on improving the online experience. It could be advertised more and potentially even incentivized with cheaper flights. - Online Support: More time/resources could be spent on improving the online experience. It could be advertised more and potentially even incentivized with cheaper flights.

Variables to Exclude from Model:

- Type of Travel: The airline can't control why people travel and probably doesn't know this information reliably. - Gender: I don't think the risk of discrimination is justified, so I'm excluding this. - Customer loyalty: I don't think the airline can control this. Satisfaction probably dictates this in large part, so using this is somewhat circular in my opinion. - Arrival Delay: There isn't much that can be done about this. - Departure Delay: I don't think much can be done here. - Leg Room Service: Can't do much here. You either pay for more or don't - a free cookie or other incentive probably won't do much to make your legs less uncomfortable. - In-flight Wifi service: I'm not sure if this can be changed - I doubt it. We'd need to consult relevant specialists here. I bet the wifi is only as good as our current engineering allows - probably not worth investigating. - Gate Location: Doesn't matter much according to my boxplot, and airlines can't change this. - Food Drink: Barely had an influence, according to the boxplot. Also, you get what you pay for. This probably ties into the class of seat. - Departure/Arrival time convenient: Can't really improve this feasibly.

Boxplots of Continuos Variables for Modeling

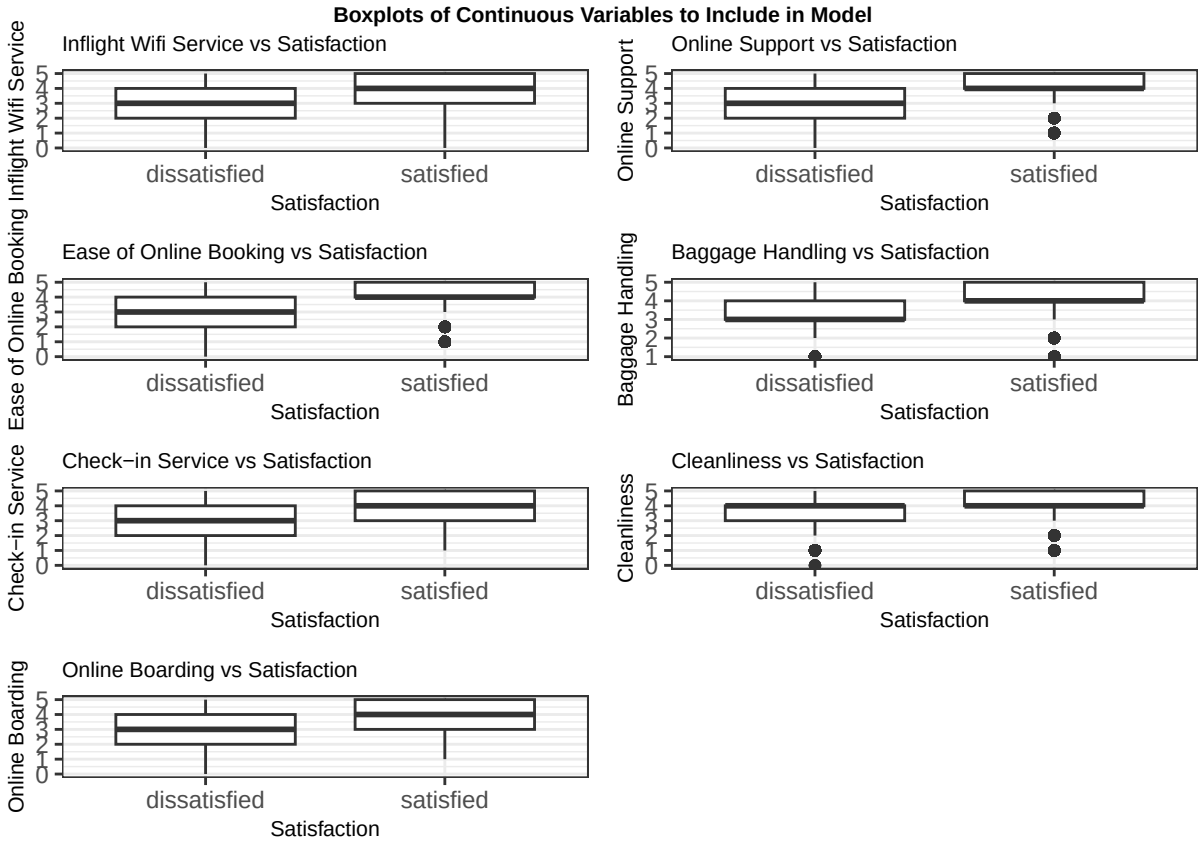

Figure 1: Figure 1

Bar Chart of Categorical Variable for Modeling

Random Forest Modeling and Prediction

Random Forest and Variable Importance

```
# sample size
sample_size <- 1000
# generates random sample
random_sample <- airline[sample(nrow(airline), sample_size), ]

# create the random forest model using the random sample
rf1 <- randomForest(satisfaction ~ Age + Cleanliness + Class + Online.boarding +
                        Checkin.service + Baggage.handling + Ease.of.Online.booking  +
                        Online.support, data = random_sample)
rf1
```

```
##
## Call:
```

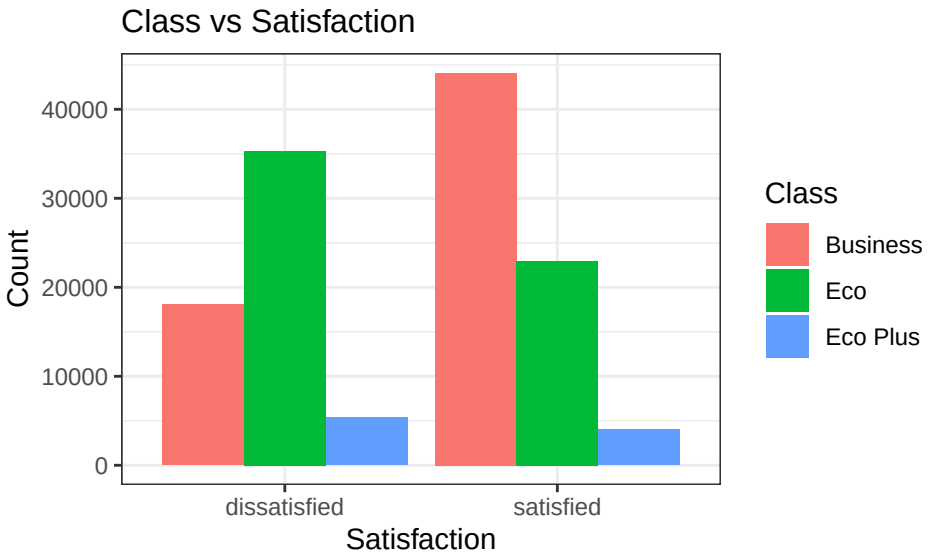

Figure 2: Figure 2

Table 1: Table 1: Importance of Variables for Predicting Customer Satisfaction

	Variable	MeanDecreaseGini
1	Age	81.83807
7	Ease.of.Online.booking	67.69125
8	Online.support	63.96252
3	Class	51.88526
5	Checkin.service	43.60981
4	Online.boarding	43.26304
2	Cleanliness	39.13797
6	Baggage.handling	35.98598

```
##  randomForest(formula = satisfaction ~ Age + Cleanliness + Class +      Online.boarding + Checkin.se
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 21.8%
## Confusion matrix:
##              dissatisfied satisfied class.error
## dissatisfied          336       109   0.2449438
## satisfied             109       446   0.1963964
```

```
#calc var importance
var_importance <- importance(rf1)
```

Prediction Based on Random Forest

```r
#create an empty data frame to contain properly
#new flyers
newflyer <- airline[0,c(4,20,6,21,19,18,15,14)]

#specify the values of my new flyers
newflyer[1,] <- c(60,4,"Eco",3,4,4,4,3)

#make the prediction
predict(rf1,newflyer,type="prob")
```

```
##   dissatisfied satisfied
## 1        0.692     0.308
## attr(,"class")
## [1] "matrix" "array"  "votes"
```