Report Airline Passenger Satisfaction

Kilian Sennrich

26 3 2021

This is the report of Kilian Sennrich for the first assignment of the class "10616 Machine Learning" held by Prof. Dr. Dietmar Maringer (University of Basel). The task is to perform a complete analysis for one of the 10 datasets provided in the lecture. I chose the "Airline Passenger Satisfaction" dataset from kaggle: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

From the Website:

What factors lead to customer satisfaction for an Airline?

This dataset contains an airline passenger satisfaction survey. What factors are highly correlated to a satisfied (or dissatisfied) passenger? Can you predict passenger satisfaction?

Note that this dataset was modified from another dataset by John D on Kaggle. It has been cleaned up for the purposes of classification.

1. Gender: Gender of the passengers (Female, Male)
2. Customer Type: The customer type (Loyal customer, disloyal customer)
3. Age: The actual age of the passengers
4. Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
5. Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
6. Flight distance: The flight distance of this journey
7. Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
8. Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
9. Ease of Online booking: Satisfaction level of online booking
10. Gate location: Satisfaction level of Gate location
11. Food and drink: Satisfaction level of Food and drink
12. Online boarding: Satisfaction level of online boarding
13. Seat comfort: Satisfaction level of Seat comfort
14. Inflight entertainment: Satisfaction level of inflight entertainment
15. On-board service: Satisfaction level of On-board service
16. Leg room service: Satisfaction level of Leg room service
17. Baggage handling: Satisfaction level of baggage handling
18. Check-in service: Satisfaction level of Check-in service
19. Inflight service: Satisfaction level of inflight service
20. Cleanliness: Satisfaction level of Cleanliness
21. Departure Delay in Minutes: Minutes delayed when departure
22. Arrival Delay in Minutes: Minutes delayed when Arrival
23. Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Introduction

This Report explains the analysis of the dataset. It is organized according to the main.R file, that contains the analysis. This report explains the thought process behind the analysis. Assumptions in the analysis are clearly documented here!!

Table of contents

Inspection, Balance, Missing values

A first look into the data revealed, that the data is pretty clean. There were 2 columns to be deleted (V1, id) that contain unique identifiers to single observations and that don't add value to the analysis, since we do not try to make statements about single customers. The data is generally well balanced, with approx. 50 percent of the criterion (satisfaction) in both groups. Therefore, in this regard no further steps are necessary. Missing values (NAs and sometimes NULL objects) can only be upserved in the "Arrival Delay in Minutes" column. There are 310 missing values in total, that distributed approximately equally between the two attributes of the criterion. I had the hypothesis, that the 310 values might all originate from the same plane, since an average airplane could transport about that amount of passengers. To check this i made the following assumption **If the NA values originated from passengers of the same flight, the flight distance for all these passengers would be equivalent** . The hypothesis is wrong, since the max number of passengers that had the same flight distance was 5. In order to get good data for the classifier, I wanted to get rid of the NAs. For imputation I had three ideas in mind: replacement by mean, delete rows or MICE. Mice seemed an over-the-top approach for only 310 missing values. Both of the other approaches would work, but i chose to omit the NA rows, since the proportion of data that falls away is very small (0.2 %) and having to many identical values (mean imputation) in the data, sometimes affects the accuracy of the model.

Visualisation

Variable by Criterion

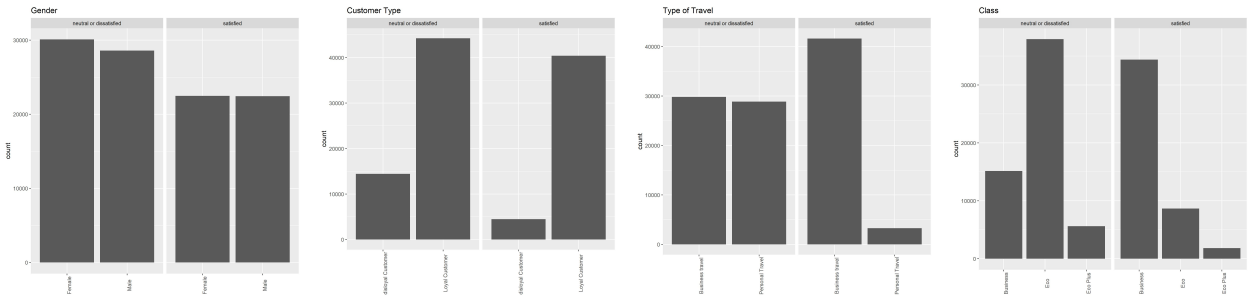

Figure 1: Vizualisation of all nominal variables by criterion

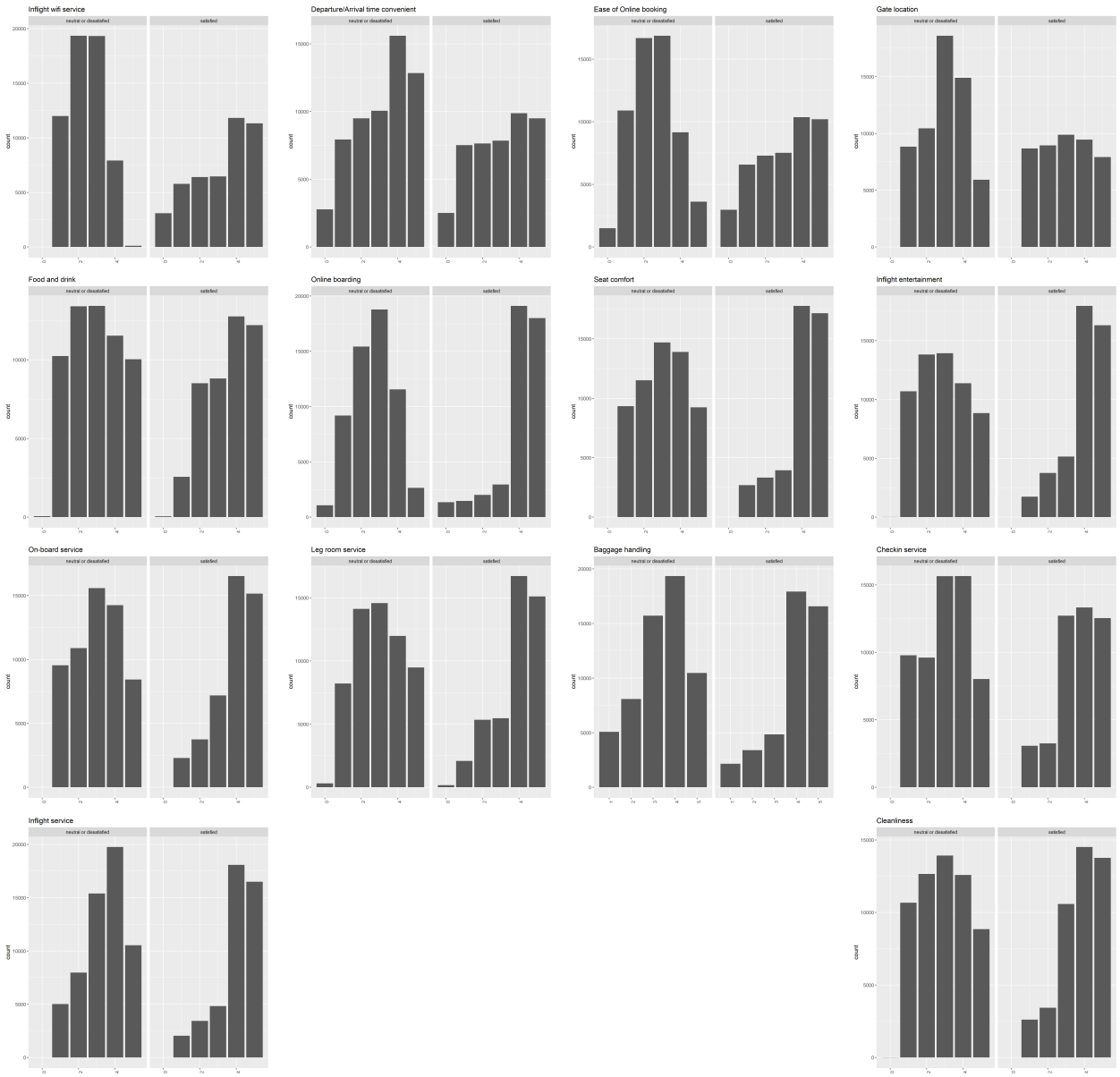

Figure 2: Vizualisation of all ordinal variables by criterion

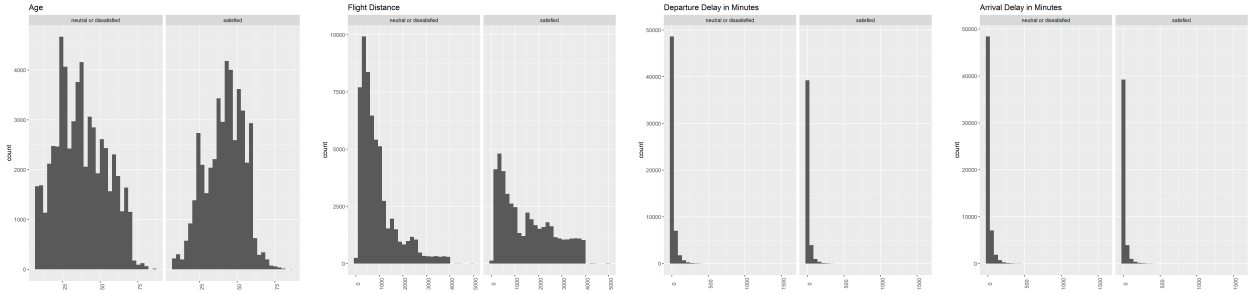

Figure 3: Vizualisation of all intervall variables by criterion

This is the text for the visualisation part,

Correlation Analysis

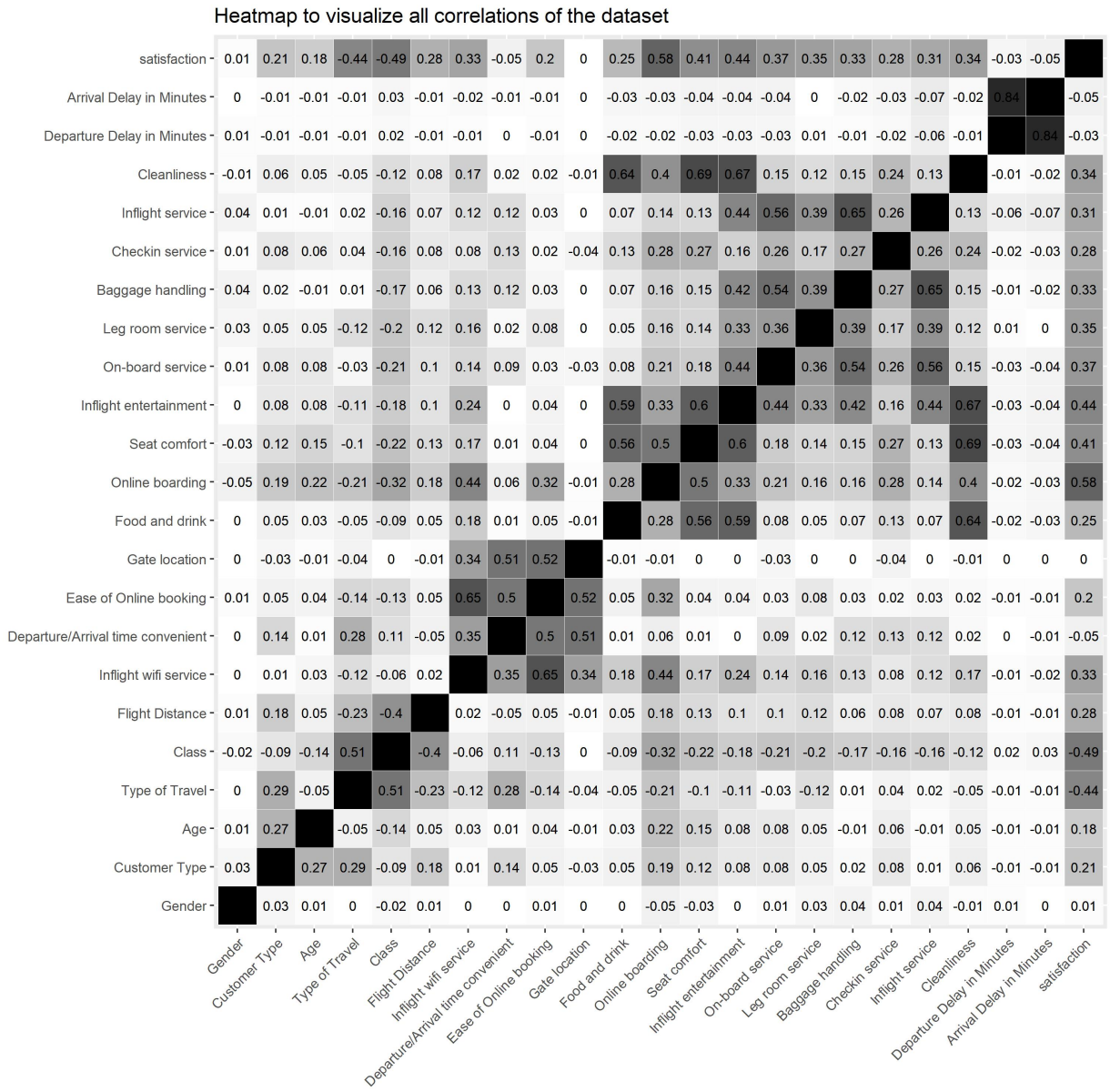

Figure 4: The darker the color, the more significant the correlation. The correlation-coefficient is spearmans rho. Be careful with interpretation since there were nominal variables in there!

Table 1: correlations higher than +-0.5

y	x	cor
satisfaction	Online boarding	0.5899783

Table 2: correlations between +-0.3 and +-0.5

y	x	cor
satisfaction	Type of Travel	-0.4489949
satisfaction	Class	-0.4801065
satisfaction	Inflight wifi service	0.3722062
satisfaction	Seat comfort	0.3618708
satisfaction	Inflight entertainment	0.3995540
satisfaction	On-board service	0.3283335
satisfaction	Leg room service	0.3200487
satisfaction	Cleanliness	0.3031766

Table 3: correlations between +-0.1 and +-0.3

y	x	cor
satisfaction	Customer Type	0.1875584
satisfaction	Age	0.1470610
satisfaction	Flight Distance	0.2573588
satisfaction	Ease of Online booking	0.2335138
satisfaction	Food and drink	0.2075063
satisfaction	Baggage handling	0.2693956
satisfaction	Checkin service	0.2323059
satisfaction	Inflight service	0.2654399

Table 4: items not associated with criterion (-0.1 and 0.1)

y	x	cor
satisfaction	Gender	0.0123560
satisfaction	Departure/Arrival time convenient	-0.0468449
satisfaction	Gate location	-0.0003662
satisfaction	Departure Delay in Minutes	-0.0737981
satisfaction	Arrival Delay in Minutes	-0.0735977

Bad idea to use cor. test since we had to correct with bonferroni and high samplesize would make everything significant

Why not leave uncorrelated items out? Bc we use random forrest classifier which does not care about uncorrelated items

why not leave out 0s in the satisfaction ratings? bc it can still give some insight to the data to know which people do not have an opinion