Predicting Customer Satisfaction

DSC 478: Programming Machine Learning Applications

DePaul University

Table of Contents

[**Executive Summary** 3](#_Toc66109033)

[**Predicting Customer Satisfaction** 4](#_Toc66109034)

[Data 4](#_Toc66109035)

[Attribute Descriptions and Distributions 4](#_Toc66109036)

[Distribution Comparisons 5](#_Toc66109037)

[Tables 6](#_Toc66109038)

[Summary Statistics 6](#_Toc66109039)

[Categorical Value Counts 7](#_Toc66109040)

[Figures 8](#_Toc66109041)

[Attribute Histograms 8](#_Toc66109042)

[Correlations 10](#_Toc66109043)

\

**Executive Summary**

Need to add executive summary

**Predicting Customer Satisfaction**

Despite the current pandemic the US Airline industry is a booming 686 billion US dollar industry that is projected to grow by an additional 110 billion dollars by 2021 (Mazareanu, 2021). With projections increasing, it is important to understand the fierce competition between airlines and their fight for customer retention and satisfaction. In this report, we aim to discover features or latent factors that may be indicative of customer satisfaction by attempting a variety of classification algorithms and tuning parameters. We first describe the data by looking at the different distributions for each of the variables as well looking at potential correlations. Then perform data clean-up and convert all variables into numerical attributes. Finally, we attempt four different classification algorithms and compare the results.

# Data

The raw dataset consists of 129,880 observations and 23 different attributes. The sample is an anonymized dataset from a US Airline describing passenger characteristics, ratings on various airline features and airline flight delays. It is a rather complete dataset, with very few null observations. Only 393 instances under one attribute were found to have null values. Due to the large dataset, it would be best to drop any instances where null values exist. After removing rows that contain null values, the final usable dataset consists of 23 attributes and 129,487 observations; 5 of which, are categorical attributes and 17 are numerical attributes.

## Attribute Descriptions and Distributions.

Many of the attribute had similar distributions, especially for attributes that dealt with passenger ratings. The majority of the rankings had moderate left-skewed distributions that leaned toward ratings of 3-5. The distributions can be found in Figure 1.

**Gender.** Is the passenger’s gender. The distribution is nearly equal with 50.75% female and 49.25% male.

**Customer Type.** Refers to two unique values a loyal customer and a disloyal customer. The attribute distribution leans towards loyal customers with over 81% and 19% disloyal customers.

**Age**. Refers to the actual age of the passengers. The passengers’ age ranges between 7 and 85 with an average age of 39.42 and a standard deviation of 15.12. The distribution is unimodal and nearly normal.

**Type of Travel**. Purpose of the flight of the passengers has two unique values. Passengers were either traveling for business or for personal reasons. In this dataset, 69% of passengers traveled for business, and 31% traveled for personal reasons.

**Class.** The travel class in the had three categories: Business, Eco and Eco Plus. Most passengers were either in Business (48%) or Eco (45%) and in small percentage in Eco Plus (7%).

**Flight distance**. The flight distance of a passenger’s journey in miles. Flights ranged between 31 and 4,983 miles with an average distance of 1,190 miles and a standard deviation 997. It is a strong right-skewed distribution favoring flights that are less than 1,000 miles.

**Inflight Wi-Fi Service**. Refers to a passengers’ satisfaction rating of the inflight Wi-Fi service ranging between 0 to 5. On average, passengers rated the feature at 2.72 with a standard deviation of 1.32. The unimodal distribution is nearly normal.

**Departure/Arrival Time Convenience**. Is the Satisfaction rating of Departure/Arrival time convenience between 0 and 5. On average, passengers rated this feature 3.05 with a standard deviation of 1.52. It is a moderate left-skewed unimodal distribution.

**Ease of Online booking.** Is the satisfaction rating for online booking. The ratings range between 0 and 5 with an average rating of 2.75 and a standard deviation of 1.40. It is a unimodal, nearly normal distribution.

**Gate location.** Is the satisfaction rating of a passengers’ gate location. The ratings range between 0 and 5 with an average of 2.97 and a standard deviation of 1.27. It is a unimodal weak left-skewed distribution.

**Food and drink.** Is the passengers’ satisfaction rating of food and drinks for the airline. The ratings ranged between 0 and 5 with an average of 3.20 and a standard deviation of 1.32. It is a unimodal left-skewed distribution.

**Online boarding.** Is the passengers’ satisfaction rating of online boarding. Most passengers offered a high rating. They ranged between 0 and 5 with an average of 3.25 and a standard deviation of 1.35. It is a unimodal strong left-skewed distribution.

**Seat comfort.** Refers to the satisfaction level of seat comfort within the airplane. The ratings range between 0 and 5 with an average of \_\_ and a standard deviation of \_\_. It is a moderately strong left skew distribution.

**Inflight entertainment.** Refers to the satisfaction level of inflight entertainment. The ratings range between 0 and 5 with an average of \_\_ and a standard deviation of \_\_. It is a moderately strong left skewed distribution.

**On-board service.** Refers to the satisfaction level of On-board service. The ratings range between 0 and 5 with an average of \_\_\_ and a standard deviation of \_\_. It is a moderately strong left skewed distribution.

**Leg room service.** Refers to the satisfaction level of Leg room service. The ratings range between 0 and 5 with an average of \_\_ and a standard deviation of \_\_. It is a moderately strong left skewed distribution.

**Baggage handling**: Refers to the satisfaction level of baggage handling. The ratings range between 0 and 5 with an average of \_\_ and a standard deviation of \_\_. It is a moderate left skewed distribution.

**Check-in service**: Refers to the satisfaction level of Check-in service. The ratings range between 0 and 5 with an average of \_ and standard deviation of \_\_. It is a strong left skewed distribution.

**Inflight service**: Refers to the satisfaction level of inflight service. The ratings ranged between 0 and 5 with an average of \_ and a standard deviation of \_\_. It is a strong left skewed distribution.

**Cleanliness**. Refers to the satisfaction level of cleanliness on the flight. The ratings range between 0 and 5 and have an average of \_ with a standard deviation of \_. It is a strong left skewed distribution.

**Departure Delay in Minutes**. Refers the to the number of minutes delayed for departure. The delays range between \_ and \_ with an average of \_ and a standard deviation of \_. It is a strong right skewed distribution.

**Arrival Delay in Minutes**. Refers to the number of minutes delayed on arrival. Like departure delay, it ranges between \_ and \_ with an average of \_ and a standard

deviation of \_. It is a strong right skewed distribution.

**Satisfaction**. Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

### 

### Distribution Comparisons

As shown in Figure 2, there are numerous differences in the distributions between satisfied and unsatisfied customers. Some of the most obvious differences lie in the ratings. For customers that were satisfied, they tended to rate air line features higher than unsatisfied customers; most notably, their online booking experience. Satisfied customers tended to be slightly older, traveled farther, and mainly traveled for business rather than pleasure.

Tables

## Summary Statistics

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | **Age** | **Flight Distance** | **Inflight Wi-Fi** | **Departure/Arrival time convenient** | **Online booking** | **Gate location** | **Food & Drink** | **Online boarding** | **Seat comfort** |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **mean** |  | 39.43 | 1190.21 | 2.73 | 3.06 | 2.76 | 2.98 | 3.20 | 3.25 | 3.44 |  | | **std** |  | 15.12 | 997.56 | 1.33 | 1.53 | 1.40 | 1.28 | 1.33 | 1.35 | 1.32 |  | | **min** |  | 7.00 | 31.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |  |  | | **25%** |  | 27.00 | 414.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 |  |  | | **50%** |  | 40.00 | 844.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 | 4.00 |  |  | | **75%** |  | 51.00 | 1744.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 4.00 | 5.00 |  |  | | **max** |  | 85.00 | 4983.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 |  |  | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Entertainment** | **On-board service** | **Leg room service** | **Baggage handling** | **Check-in service** | **Inflight service** | **Cleanliness** | **Departure Delay in Minutes** |
| **mean** |  | 3.36 | 3.38 | 3.35 | 3.63 | 3.31 | 3.64 | 3.29 | 14.64 |
| **std** |  | 1.33 | 1.29 | 1.32 | 1.18 | 1.27 | 1.18 | 1.31 | 37.93 |
| **min** |  | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **25%** |  | 2.00 | 2.00 | 2.00 | 3.00 | 3.00 | 3.00 | 2.00 | 0.00 |
| **50%** |  | 4.00 | 4.00 | 4.00 | 4.00 | 3.00 | 4.00 | 3.00 | 0.00 |
| **75%** |  | 4.00 | 4.00 | 4.00 | 5.00 | 4.00 | 5.00 | 4.00 | 12.00 |
| **max** |  | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 1592.00 |

## Categorical Value Counts

|  |  |  |
| --- | --- | --- |
|  | **Gender** | **%** |
|  |  |  |
| Female | 65,703 | 50.75% |
| Male | 63,784 | 49.25% |
|  |  |  |
|  | **Travel Type** |  |
| Business | 89,445 | 69.07% |
| Personal | 40,042 | 30.93% |
|  |  |  |
|  | **Class** |  |
| Business | 61,990 | 47.87% |
| Eco | 58,117 | 44.88% |
| Eco Plus | 9,380 | 7.24% |
|  |  |  |
|  | **Satisfaction** |  |
| Neutral or Dissatisfied | 73,225 | 56.55% |
| Satisfied | 56,262 | 43.45% |
|  |  |  |
|  | **Customer Type** |  |
| Loyal | 105,773 | 81.69% |
| Disloyal | 23,714 | 18.31% |

Figures

## Attribute Histograms

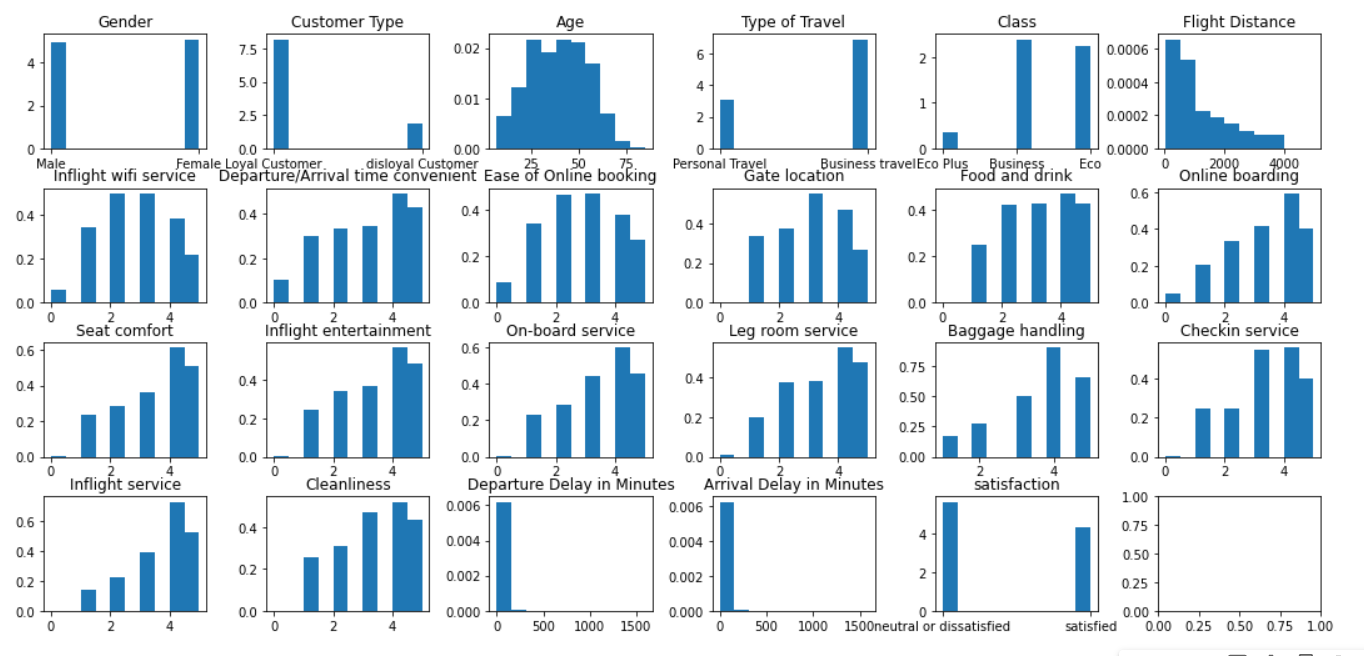


Figure 1 – Univariate Distributions

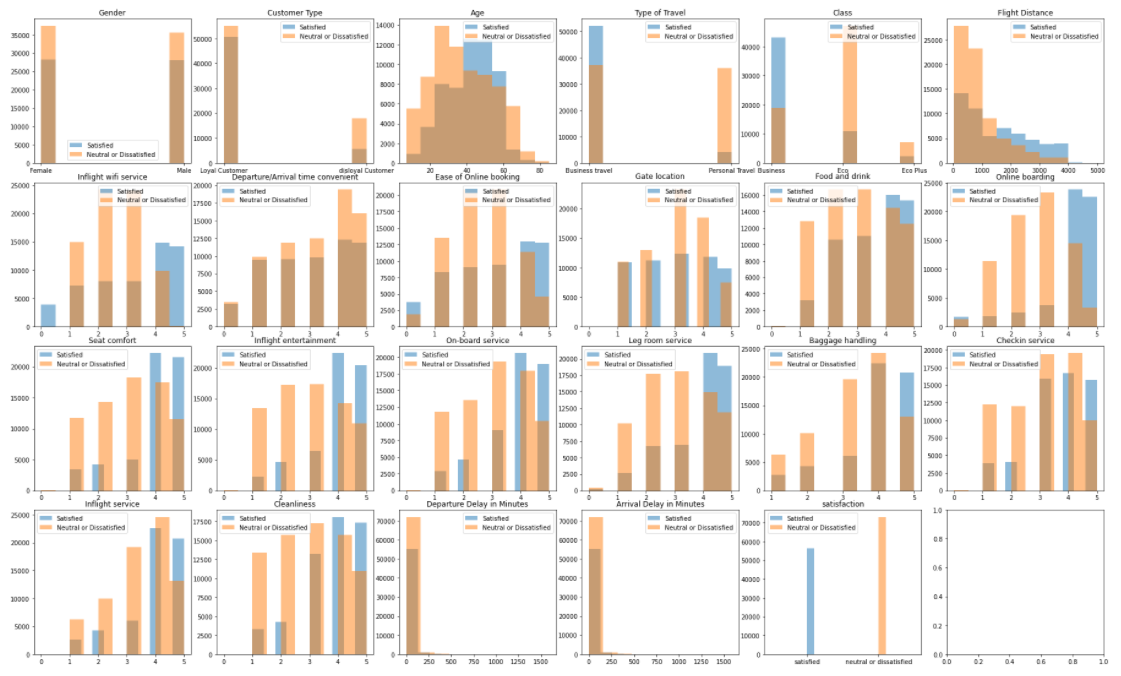


Figure 2 - Distributions between satisfied and unsatisfied customers

## CorrelationsChart Description automatically generated

Figure 3 - Significant Correlations

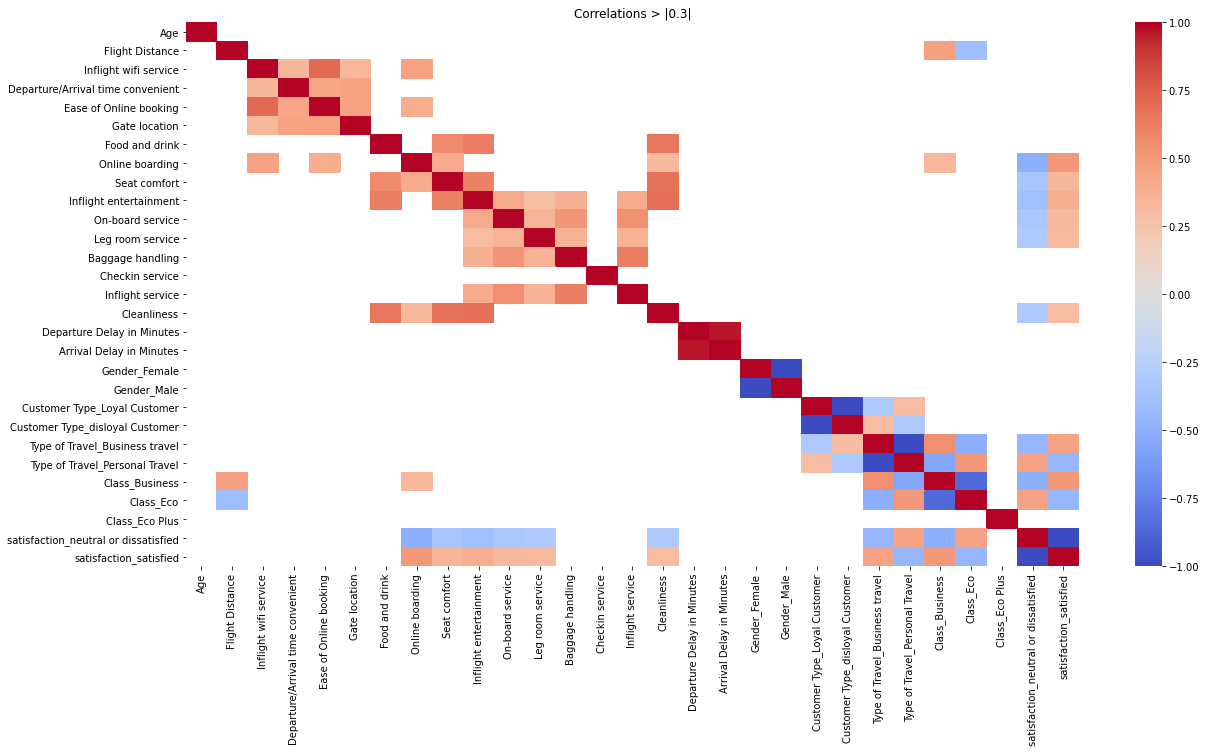


Figure Correlation Heatmap, threshold = |0.3|

Most of the variables did not display strong correlations. We opted to filter the correlation plot with a threshold of 0.3 to showcase the most relevant relationships. Customer satisfaction was the most heavily correlated with higher ratings for ease of online booking, seat comfort, inflight entertainment, on-board service, leg room, and cleanliness. Business travelers and travelers that upgraded to Business class seats were more likely to report a satisfactory flight experience.

There is some multicollinearity between individual ratings, but not to a degree that caused concern. For instance, flights with high satisfaction scores for online boarding also have favorable scores for ease of online booking and inflight wifi quality.

## Plan

1. Data cleanup / preprocessing
   1. Missing/wrong values/ NA / NAN
   2. Dummify
   3. Train/test split
   4. Scale
   5. Describing data
      1. Datatypes
      2. Nominal/ordinal/continuous
         1. Cant use distance measures on nominal data
2. Var exploration
   1. Predictor counts
   2. ~~Class counts / imbalanced data?~~
      1. Stratified sampling?
   3. 5-num summary
      1. Outliers?
         1. Outside of interquartile range
   4. Distributions
      1. Skewness / normality
   5. Cross tabulations?
   6. Clustering as exploration. Outliers?
   7. Visualization
3. Model building
   1. Try Models
      1. Decision Tree
      2. Random Forest
      3. Logistics Reg
      4. Naïve Bayes
      5. KNN
      6. Ensemble
      7. SVC
   2. Tuning params
   3. Pick the best model
      1. Bias/variance
   4. Final evaluation
      1. Confusion matrix

Works Cited

Mazareanu, E. (2021, January 6). *www.statista.com*. Retrieved from Statista: https://www.statista.com/statistics/1110342/market-size-airline-industry-worldwide/#:~:text=The%20global%20airline%20industry%20was,as%20well%20as%20cargo%20airlines.