# Predicting Customer Satisfaction in the Airline Industry

**Team 6: Joshua Farina (joshuapfarina), Henri Salomon (henri_salomon), Raajitha Middi (Raajitha_Middi), Ryan Chandler (zujin87), Alejandro Martinez (amzeta)**

## 1. Objective / Problem Statement

We intend to use airline passenger satisfaction survey data to determine which factors drive customer satisfaction (or dissatisfaction) so that we can provide actionable recommendations on how money should be invested or re-allocated to keep passengers satisfied.

Customer satisfaction is a key factor in attracting and retaining business.Identifying factors that have the strongest effect on customer satisfaction will provide key insights for airlines to enhance their services, improve the customer experience and optimize business operations.Some estimates indicate that increasing customer retention by as little as 5% increases profits by 25%-95%.[1]

## 2. Methodology / Approach

In our approach, after completing the **Exploratory Data Analysis (EDA)**, we have narrowed down the focus of our modeling to i) Logistic Regression, ii) Decision tree, iii) Random Forest, and iv) Support Vector Machines (SVM):

**i. Logistic Regression.** We will systematically test each variable using a logistic regression model individually. For each variable, we will examine its distribution and regression diagnostics to determine if transformations or binning are necessary. If a log transform appears suitable, we will apply it and iterate through the prior steps. Alternatively, we will evaluate the variable for binning. By comparing the adjusted R squared of the transformations to the original logistic regression, we will decide if there is an improvement. We will repeat this process for each variable. Once we have a suitable set of variables, we will run the full regression and assess the R squared against the untransformed regression to determine overall improvement.After identifying our best logistic regression model, we will cross-validate it to measure accuracy and assess the ROC AUC.

**ii. Decision tree.** We will use a decision tree model to establish a hierarchical structure that splits the dataset based on different features and their importance, enabling effective prediction and interpretation.

**iii. Random Forest.** We will utilize a random forest model to evaluate relative feature importance.

**iv. SVM.** We will repeat the cross-validation process with SVM to determine if it yields better accuracy and AUC.

This comprehensive approach will allow us to assess the performance and suitability of different models and choose the most appropriate one for our project.

## 3. Progress, challenges and next steps

*a) Progress*

Our project has made progress in various aspects:

- **Kaggle primary data set**

Regarding the primary (main) data set from Kaggle, we have been finalizing the exploratory data analysis (EDA) phase, gaining valuable insights into the dataset that are presented later in this document. We also focused our work on finding the most suitable solution for data imputation (mode) given our data set.

[1] https://www.forbes.com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/?sh=3cef46fc1c7d

Additionally, we have made significant progress in data modeling with preliminary results thanks to different logistic regressions, feature engineering (forward selection) and random forest. Some of the preliminary results are presented below. These modeling techniques have allowed us to reach overall very good accuracy and uncover patterns and relationships within the data.

- **ACSI data set**

Upon reviewing the ACSI dataset and related papers, we have concluded that using it for a parallel analysis of customer satisfaction is not feasible. The dataset's broad industry focus and the lack of a direct mapping of factors driving customer satisfaction make it impractical. Additionally, a robust understanding of their techniques is beyond the scope of this course. As an alternative, we are exploring the use of the ACSI data to determine the financial value of a satisfied customer, although challenges arise due to methodological and scale differences between the ACSI dataset and our primary data source. Further details and analysis are presented below.

- **Further research**

We have identified useful insights from recent research papers that can orient our works for the next few weeks. They are presented below.

*b) Challenges*

So far, we have encountered a few challenges:

- *Data imputation for the primary data set:* While we are in the final stages of this task, it required careful consideration and selection of appropriate imputation techniques to ensure data completeness and accuracy.

- *Reconciliation of findings from the primary data set with the American Customer Satisfaction Index (ACSI) dataset:* As mentioned above, using the ACSI dataset to perform a parallel analysis of customer satisfaction does not seem possible, and we are now exploring alternative methods of using the ACSI data. One option is using the ACSI data to determine the financial value of a satisfied customer.

- *Impossibility to merge the primary data set with the flight path data set:* As explained in the previous report, we will not be able to merge flight path data to try to infer airport locations due to insufficient information as the primary data set does not specify number of legs in a trip. In addition, flight distances in the primary dataset are too similar in many instances, which would make it impossible to narrow down with any certainty which airports were involved in the trip. As such, we decided to stop analyzing the flight path data set.

*c) Next steps*

In the upcoming two weeks, we have outlined the following activity to effectively conclude our results:

1. Define a common train/validate/test data sets for comparing models by July 11th
2. Enhance our models by incorporating valuable insights and performing feature engineering by July 14th
3. Conduct a thorough analysis of the data using random forest, decision tree, and SVM techniques by July 16th
4. Select the primary model for the final report by July 16th.
5. Evaluate the results obtained and extract meaningful business implications by July 19th.
6. Complete final report by July 19th.

By following this plan, we aim to complete our project as well as prepare requested deliverables (video, final report, finalized code & documentation).

## 4. Results from the Exploratory Data Analysis of the primary dataset from Kaggle

*a) Data Overview and exploration*

**Data source.** The dataset is taken from surveys from a US airline. It includes user ratings for multiple aspects of air travel such as food and drink and seat comfort, as well as customer information such as age and gender. The data set, available on kaggle, was already split into train (103,904 rows and 24 features) and test sets.

**Likert Scaled and Continuous Variables**

| Statistic | N | Mean | Median | St. Dev. | Min | N | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|---|---|
| Age | 103,904 | 39.380 | 40 | 15.115 | 7 | 103,904 | 27 | 51 | 85 |
| Flight.Distance | 103,904 | 1,189.448 | 843 | 997.147 | 31 | 103,904 | 414 | 1,743 | 4,983 |
| Inflight.wifi.service | 100,801 | 2.814 | 3 | 1.257 | 1 | 100,801 | 2 | 4 | 5 |
| Departure.Arrival.time.convenient | 98,604 | 3.225 | 3 | 1.386 | 1 | 98,604 | 2 | 4 | 5 |
| Ease.of.Online.booking | 99,417 | 2.881 | 3 | 1.299 | 1 | 99,417 | 2 | 4 | 5 |
| Gate.location | 103,903 | 2.977 | 3 | 1.278 | 1 | 103,903 | 2 | 4 | 5 |
| Food.and.drink | 103,797 | 3.205 | 3 | 1.326 | 1 | 103,797 | 2 | 4 | 5 |
| Online.boarding | 101,476 | 3.328 | 4 | 1.267 | 1 | 101,476 | 2 | 4 | 5 |
| Seat.comfort | 103,903 | 3.439 | 4 | 1.319 | 1 | 103,903 | 2 | 5 | 5 |
| Inflight.entertainment | 103,890 | 3.359 | 4 | 1.333 | 1 | 103,890 | 2 | 4 | 5 |
| On.board.service | 103,901 | 3.382 | 4 | 1.288 | 1 | 103,901 | 2 | 4 | 5 |
| Leg.room.service | 103,432 | 3.366 | 4 | 1.299 | 1 | 103,432 | 2 | 4 | 5 |
| Baggage.handling | 103,904 | 3.632 | 4 | 1.181 | 1 | 103,904 | 3 | 5 | 5 |
| Checkin.service | 103,903 | 3.304 | 3 | 1.265 | 1 | 103,903 | 3 | 4 | 5 |
| Inflight.service | 103,901 | 3.641 | 4 | 1.176 | 1 | 103,901 | 3 | 5 | 5 |
| Cleanliness | 103,892 | 3.287 | 3 | 1.312 | 1 | 103,892 | 2 | 4 | 5 |
| Departure.Delay.in.Minutes | 103,904 | 14.816 | 0 | 38.231 | 0 | 103,904 | 0 | 12 | 1,592 |
| Arrival.Delay.in.Minutes | 103,594 | 15.179 | 0 | 38.699 | 0 | 103,594 | 0 | 13 | 1,584 |

Figure 1: Data description of the training set used

**Data types.** Of these, the variables from *Inflight.wifi.service* to *Cleanliness* take values of Likert scale (satisfaction level from 1 to 5 and 0 as “not applicable”). They were considered as categorical variables (ordinal).

**Missing values.** There were two cases of missing values. Firstly, in the ordinal categorical variables, some rows had 0 (“Not applicable”) when passengers haven’t rated the respective variables. As they accounted for less than 5% of the total data for each variable, we imputed those values with mode. Secondly, the variable *Arrival.Delay.in.Minutes* has 310 missing values. We first assumed that arrival delay and departure delay might be highly correlated as the formal depends on the latter because flights rarely make up for the delayed time in air. This hypothesis was demonstrated with the correlation analysis (see below) for which the 310 rows were removed. Based on this, we decided to drop the feature *Arrival.Delay.in.Minutes* and consider *Departure.Delay.in.Minutes* instead.

*b) Correlation Analysis*

As part of our EDA, we performed different correlation analysis. Firstly, we focused on numeric variables, with the below correlation heatmap (Figure 2: Heatmap for numeric variables). This shows that *Arrival.Delay.in.Minutes* and *Departure.Delay.in.Minutes* are highly correlated. For this reason, we decided to exclude *Arrival.Delay.in.Minutes*.

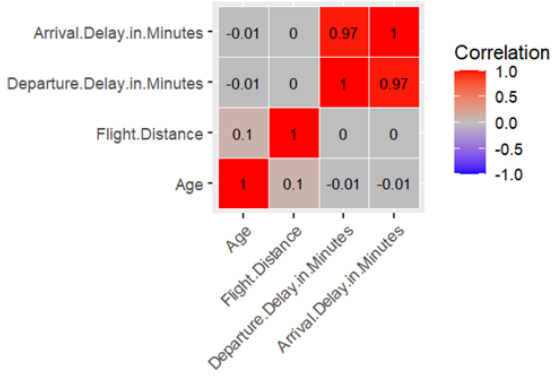



Figure 2: Heatmap for numeric variables

In a similar manner, we generated a heat map for the categorical variables as well (Figure 3: Heatmap for categorical variables). All the values are less than 0.7 which suggests that there are no strong correlations between any of the variables.

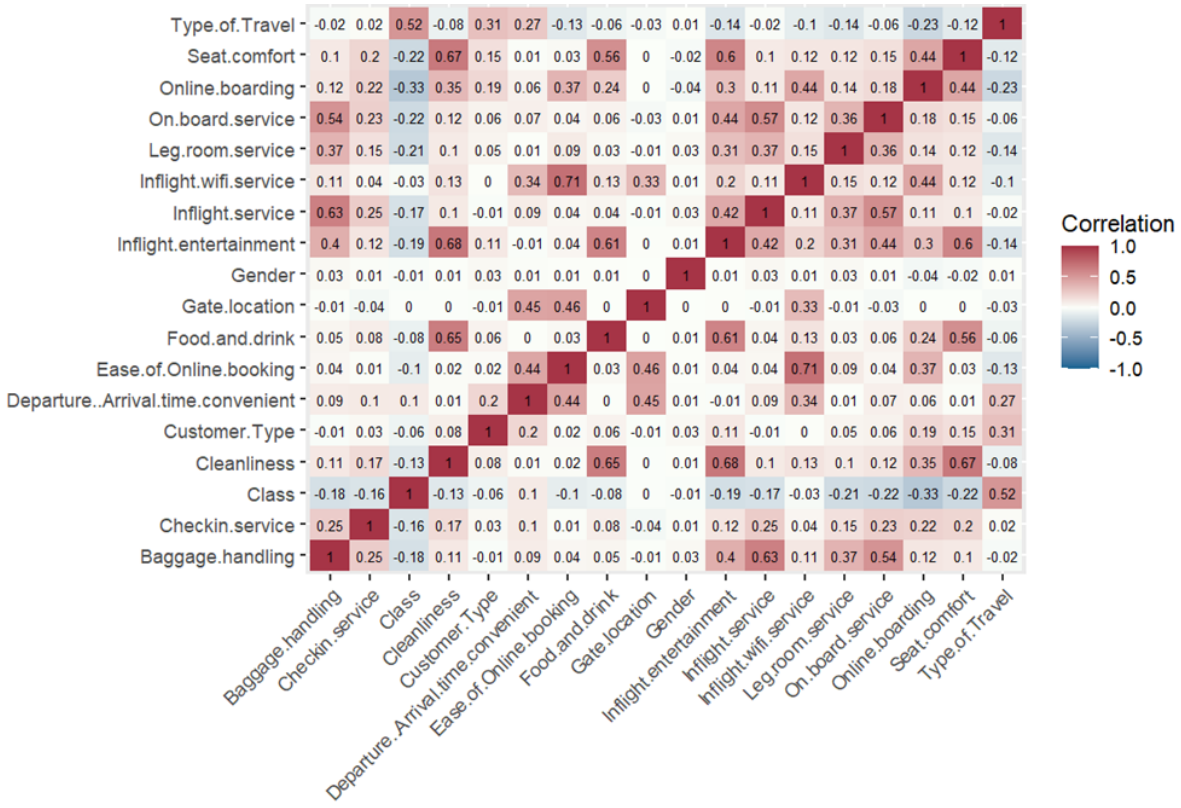



Figure 3: Heatmap for categorical variables

*c) Data Visualization*

Among the numeric variables, *Flight.Distance* has values in the range of 31-4983 and *Departure.Delay.in.Minutes* in the range of 0-1592. We visualized the data using QQ plots, histograms, box plots after applying various transformations like log, square root, box-cox. All the visualizations are listed in the appendix. Data of both variables doesn't assume normal distribution even after transformations and box plot shows lots of data as outliers except for the log transformed data. Hence, based on our group's judgment, we decided to apply log transformation. Removal of outliers is explained in the next section.

We also looked at the visualizations of categorical variables w.r.t the response variable (Figure 4: Categorical Variables (except Age) against response variable). Few key observations from the plots:

- *Gender* doesn't seem to play a major role as both have almost the same satisfaction levels.
- People in the age group of approximately 20-35 seem more dissatisfied than any other age group. And interestingly, people in this age group are more likely to be disloyal.
- Business class passengers have high satisfaction levels as expected. They rated high for *Checkin.service*, *Seat.Comfort* and *Food.and.drink* compared to the Eco plus and Eco class.
- Surprisingly, *Cleanliness* with even 4 and 5 ratings has passengers with satisfaction level as 0.
- Passengers have expressed more dissatisfaction with other services like *Inflight.wifi.service*, *Inflight.entertainment*, *Baggage.handling*, *On.board.service*, *Inflight.service*. Only when the ratings of these services are above 3, there are more passengers who are satisfied but then this number doesn't seem to be significantly large.

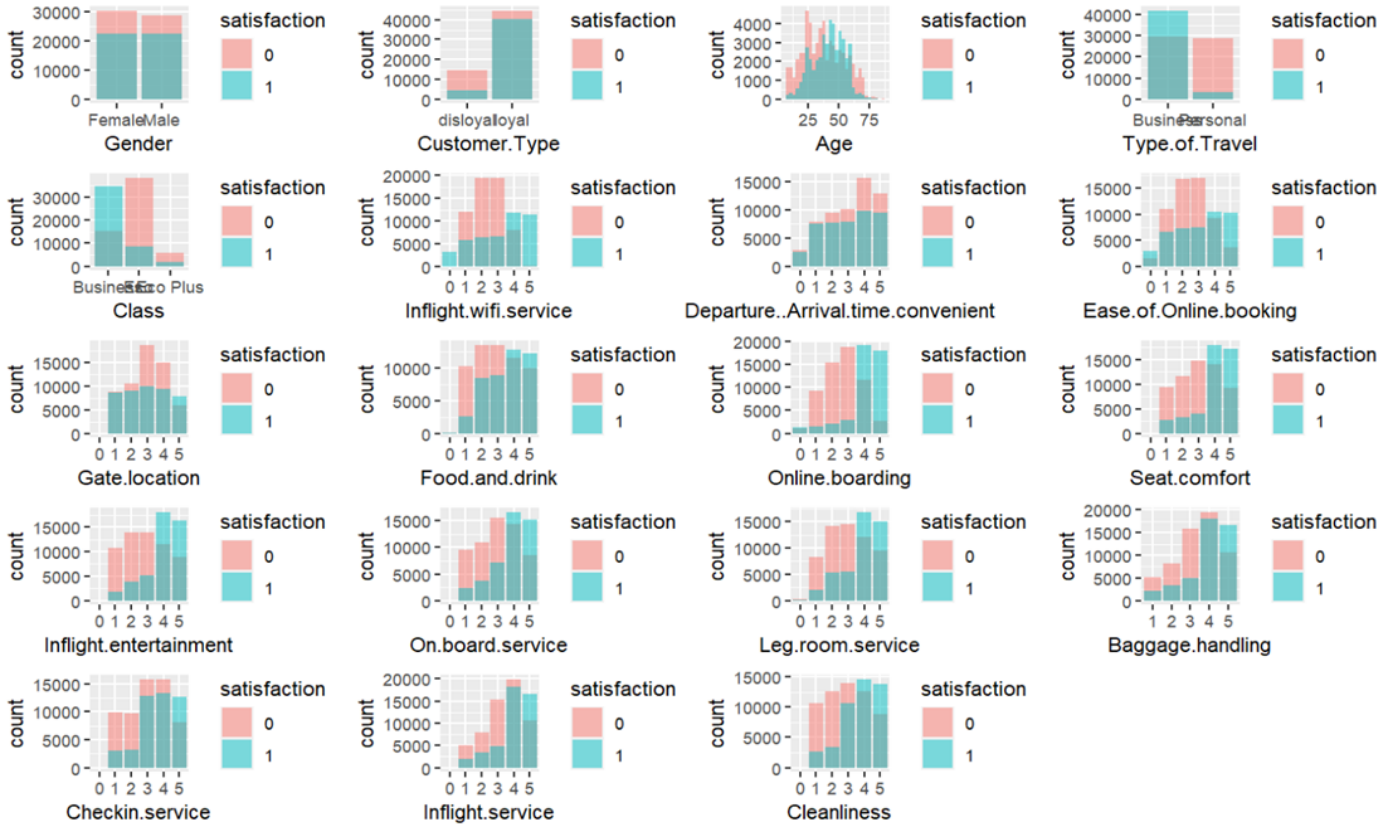



Figure 4: Categorical Variables (except Age) against response variable

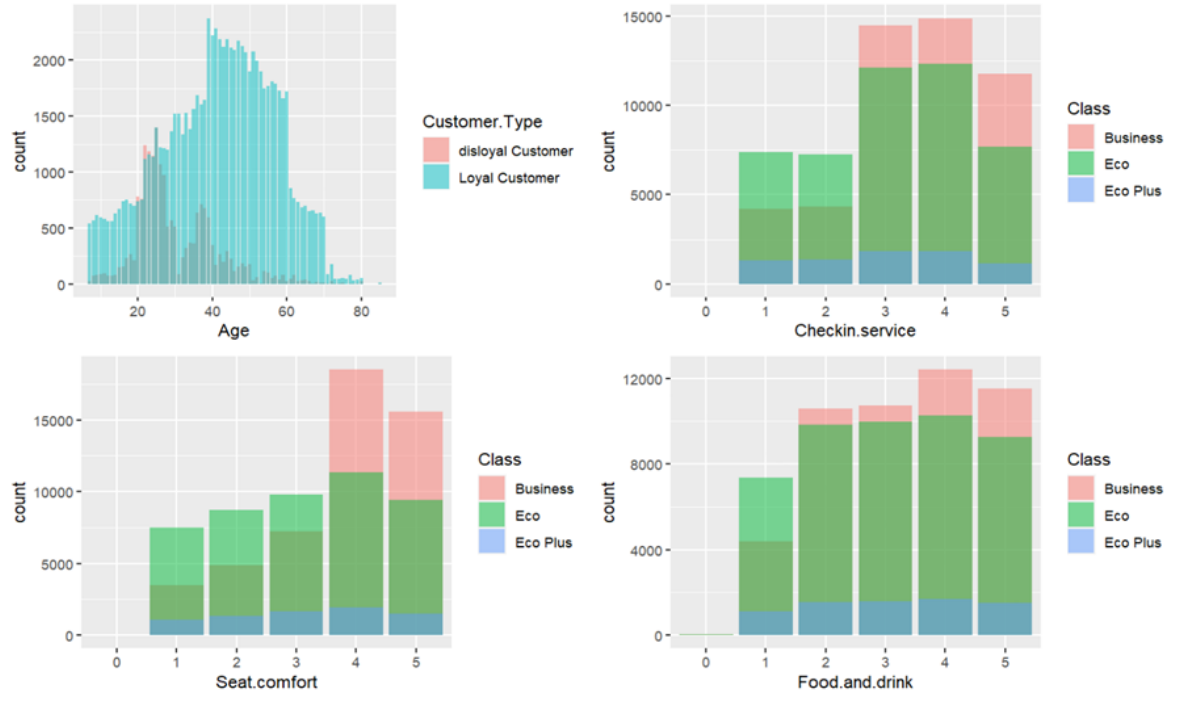



Figure 5: Categorical variables (contd.)

*d) Outliers*

We examined the box plots after applying log transformations to *Flight.Distance* and *Departure.delay.in.Minutes* and there are outliers only in *Departure.Delay.in.Minute*. We explored two methods to remove the outliers – one is by interquartile range (IQR) and the other by z-score (values<3). Z-score is an effective method if the data is normally distributed, which is not the case with our data. Hence, we chose the IQR method to remove outliers. After removing the outliers, the data has reduced to 103,886 rows from 103,904.

## 5. Preliminary results from Data Modeling (primary data set of Kaggle)

*a) Logistic regression and forward selection models*

To understand the important features to be included in the modeling, we used forward selection to choose the factors (Figure 6: Forward Selection to choose features). The results are compelling:

- Our initial hypothesis was that *Flight.Distance* was an important feature but as per forward selection, it isn't.
- We didn't anticipate that *Online.boarding*, *Type.of.travel*, *Inflight.wifi.service*, *Customer.Type* and *Inflight.service* are the top 5 important features.
- We expected *Departure.Delay.in.Minutes* to be among the top important features but as per forward selection, it is the 16$^{th}$ important feature.

Except *Flight.Distance*, forward selection has chosen all the features. Using these results, we would like to build a logistic regression model and check p-values if there are still any insignificant variables.

| | Step | Df | Deviance | Resid. Df | Resid. Dev | AIC |
|---|---|---|---|---|---|---|
| 1 | | NA | NA | 103885 | 142165.13 | 142167.13 |
| 2 | + Online.boarding | -4 | 43771.50806 | 103881 | 98393.62 | 98403.62 |
| 3 | + Type.of.Travel | -1 | 19766.82322 | 103880 | 78626.8 | 78638.8 |
| 4 | + Inflight.wifi.service | -4 | 10164.54812 | 103876 | 68462.25 | 68482.25 |
| 5 | + Customer.Type | -1 | 6546.868551 | 103875 | 61915.38 | 61937.38 |
| 6 | + Inflight.service | -4 | 7182.501528 | 103871 | 54732.88 | 54762.88 |
| 7 | + Leg.room.service | -4 | 2211.843474 | 103867 | 52521.04 | 52559.04 |
| 8 | + Checkin.service | -4 | 1750.071613 | 103863 | 50770.97 | 50816.97 |
| 9 | + Seat.comfort | -4 | 1186.632167 | 103859 | 49584.34 | 49638.34 |
| 10 | + Baggage.handling | -4 | 980.203787 | 103855 | 48604.13 | 48666.13 |
| 11 | + Inflight.entertainment | -4 | 864.653205 | 103851 | 47739.48 | 47809.48 |
| 12 | + Class | -2 | 874.104126 | 103849 | 46865.37 | 46939.37 |
| 13 | + Departure..Arrival.time.convenient | -4 | 652.479314 | 103845 | 46212.9 | 46294.9 |
| 14 | + Ease.of.Online.booking | -4 | 539.367682 | 103841 | 45673.53 | 45763.53 |
| 15 | + On.board.service | -4 | 409.296861 | 103837 | 45264.23 | 45362.23 |
| 16 | + Departure.Delay.in.Minutes | -1 | 341.223488 | 103836 | 44923.01 | 45023.01 |
| 17 | + Cleanliness | -4 | 219.708532 | 103832 | 44703.3 | 44811.3 |
| 18 | + Gate.location | -4 | 135.231073 | 103828 | 44568.07 | 44684.07 |
| 19 | + Age | -1 | 50.884056 | 103827 | 44517.18 | 44635.18 |
| 20 | + Food.and.drink | -4 | 29.758659 | 103823 | 44487.42 | 44613.42 |
| 21 | + Gender | -1 | 4.628407 | 103822 | 44482.8 | 44610.8 |

Figure 6: Forward Selection to choose features

*b) Decision tree and random forest models*

The decision tree and random forest models correlate nicely with the forward regression model. The current tree models exclude any NA values and found that imputing data did not make any significant changes.The random forest model was validated that the errors converge within 200 trees. The decision tree uses only 5 factors which are *Inflight.wifi.service*, *Checkin.service*, *Inflight.entertainment*, *Type.of.Travel*, and *Online.boarding* to achieve a 90% accuracy. Random forest agrees with these factors and achieves a 95% accuracy with only 10 factors.

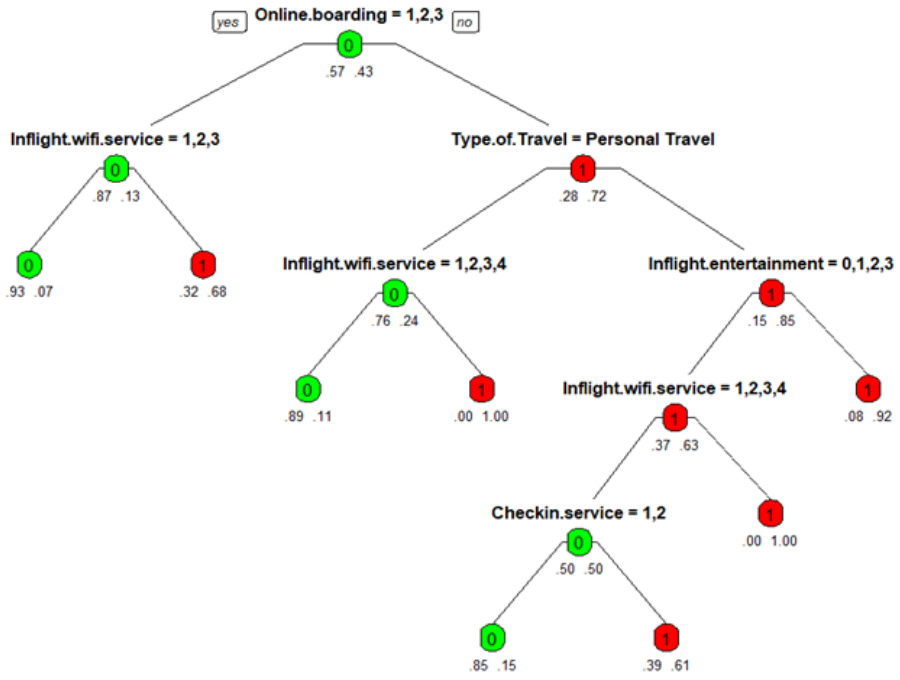



Figure 7: Decision Tree Model

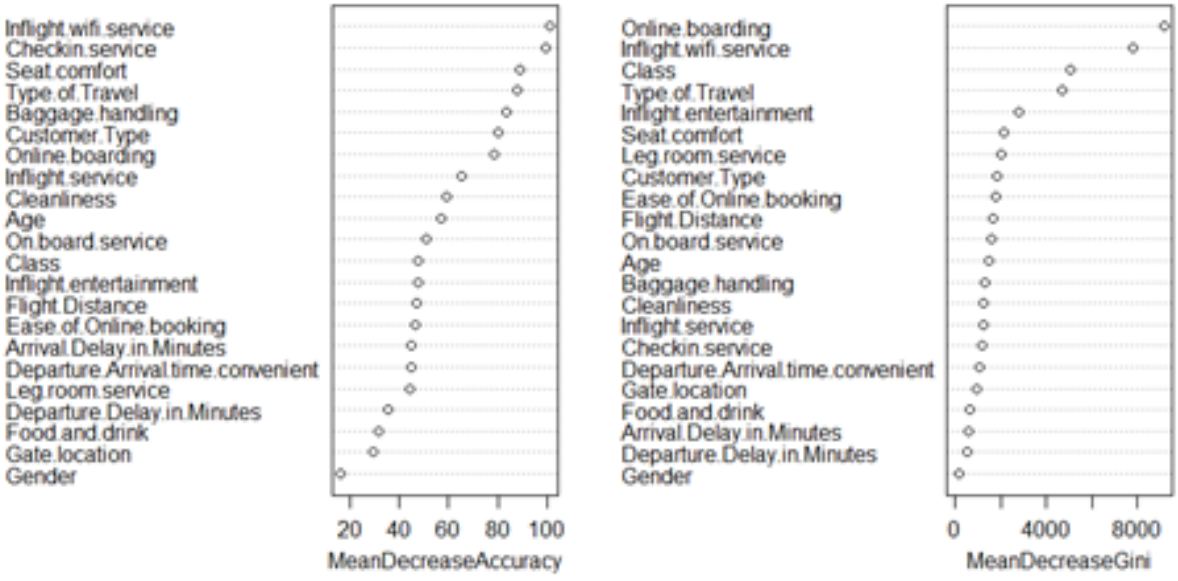



Figure 8: Random Forest Feature Importance

## 6. Results of the analysis of the ACSI data set

In our initial proposal, we described an intent to use the ACSI dataset to perform a parallel analysis of customer satisfaction. After investigating the data set and reviewing some of the related papers, we no longer think this is a practical approach.

In our primary dataset, variables are narrowly targeted and closely related to the airline industry. Variables such as “Baggage Handling” and “Leg Room” are two such examples. The ACSI data crosses multiple industries and therefore the questions serve a more general purpose. For example, it includes a variable assessing whether the customer’s expectations were met by asking, “Considering all of your expectations, to what extent has the company/brand fallen short of or exceeded your expectations?”. Although the question is valid, it is difficult to derive a one-to-one mapping of the factors which drive customer satisfaction.

Additionally, the analysis of the dataset uses partial least squares structural equation modeling (PLS-SEM) for its analysis. The method uses the included variables to derive several latent variables, which it then uses as a tool for

analysis. Each latent variable combines underlying factors constructed using methods comparable to PCA. Although we find their work interesting, a robust understanding of their techniques is beyond the scope of this course.

Given these challenges, we are exploring alternative methods of using the data. One option is using the ACSI data to determine the financial value of a satisfied customer. The ACSI data relates the latent variable of customer satisfaction to the latent variable of loyalty with a coefficient of 0.677. This analysis comes with its own challenges, however. The two different datasets use different methodologies and different scales. Our primary source of information treats customer satisfaction as a binary variable (satisfied/other than satisfied). The ACSI dataset uses a numeric index. In the ACSI data, a 1-point increase in satisfaction leads to a 0.677-point increase in loyalty, but it is not entirely clear how that would translate back to the primary data source.

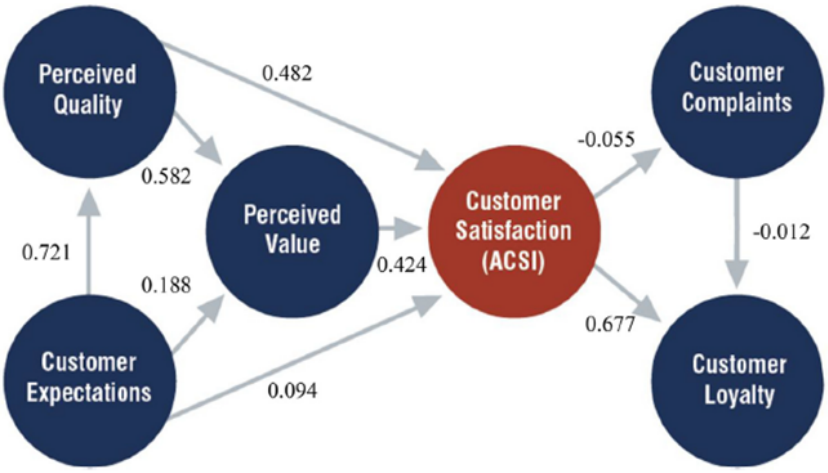



Fig. 2. Commercial airlines.

Figure 9: ACSI model for Commercial airlines

## 7. Insights from further research

*a) Insightful findings from three research papers*

- **The impact of the COVID-19 pandemic on airlines' passenger satisfaction[2]**

Written reviews were analyzed using a text mining tool. The critical features to the flight experience were identified by evaluating the frequency at which each feature was mentioned. The most important attributes according to this paper are staff behavior, employee attitude towards dissatisfied customers, booking and cancellation policies, baggage handling, seat quality, boarding, check in, customer service, and food and drink.

A related study that is mentioned in this paper concluded that the most impacted passenger satisfaction was the queuing time, lounge comfort, cabin crew quality, and seat legroom. Of these features only seat legroom is included as a feature in our dataset. This could be a significant gap in our findings derived from our main dataset. Other related studies have identified cabin crew quality as one of the most critical factors in passenger satisfaction.

This study concluded that the COVID-19 pandemic didn't change which factors influence passenger satisfaction but did give an increased importance to refunds and cabin cleanliness. Our data was taken before the COVID-19 pandemic and thus may not reflect the heightened importance of cleanliness. The data was taken from European airlines so there could be a cultural difference with our dataset since ours is taken from an American airline.

- **Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews[3]**

The authors of this paper utilized Latent Dirichlet Allocation (LDA) to analyze over 55,000 written customer reviews to identify the topics that most influence customer satisfaction. The most important features identified were cabin staff, onboard service, value for money, and seats which is consistent with the findings in the other research papers.

---

[2] https://www.sciencedirect.com/science/article/pii/S0969699723000844#bib20

[3] https://www.sciencedirect.com/science/article/abs/pii/S0969699719302959

The authors also found that certain segmentations can significantly change what customers find valuable. Some segmentations that showed significant differences were nationality, likely due to cultural reasons, and cabin flown. The practical implications to maximize customer satisfaction by cabin flown are: focus on customer service for first class passengers, comfort for premium economy passengers, and checking luggage and waiting time for economy class travelers.

This study used a database with over 55,000 online customer reviews, and covered over 400 airlines and passengers from over 170 countries. The authors claim that by having such a wide-reaching dataset that they have attained more reliable generalizations.

- **An Optimized Deep Learning Approach for Improving Airline Services**[4]

The author of this paper analyzed the same dataset that we are using with the purpose of creating the highest accuracy model possible. While the goals between our research and this author's are different, the findings serve as an interesting comparison to our models.

An interesting approach that the author took for figuring out which features mattered most was to look at the correlation between satisfaction and the 24 features. The author concluded that the most important factors are online boarding, class, type of travel, and in-flight entertainment. This approach jives with our analysis which also identified in-flight entertainment, online boarding, and type of travel as top factors. Likewise the least important factors by correlation are age, gender, departure/arrival time convenience, departure delay, and arrival delay which the author decided to exclude from the model.

The techniques that the author tried were deep neural network with Adam (99.3% accuracy), artificial neural network with Adam (95% accuracy), support vector machine (95.19% accuracy), and random forest (95.90% accuracy).

*b) Research Summary and learnings for our project*

Two of the papers that we investigated took a novel approach to customer satisfaction of analyzing publicly available written reviews and using natural language processing techniques to figure out which features airline customers care about. This seems like a more reliable approach because it doesn't limit the possible factors that may turn out to be important. On the other hand, surveys such as the one from our main dataset assume that the factors that matter are known and merely give customers the option to rate those factors. This is a problem because it implies that our model could be missing a critical factor.

For instance cabin crew quality has been found to be a top contributor to customer satisfaction across multiple research papers on airline passenger satisfaction and it is not considered in our dataset. As a counter-example, seat legroom is also a top contributor to customer satisfaction across multiple research papers and it is included in our dataset.

Another concern with the approach taken by our dataset which is mentioned in multiple research papers is that there is a tendency for customers to not take questionnaires seriously. This means that many customers fill them out randomly which just creates noise in the data.

## 8. References


https://www.forbes.com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/?sh=3cef46fc1c7d

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

https://www.sciencedirect.com/science/article/pii/S0969699723000844#bib20

https://www.sciencedirect.com/science/article/abs/pii/S0969699719302959

https://www.sciencedirect.com/science/article/pii/S2352340923002421

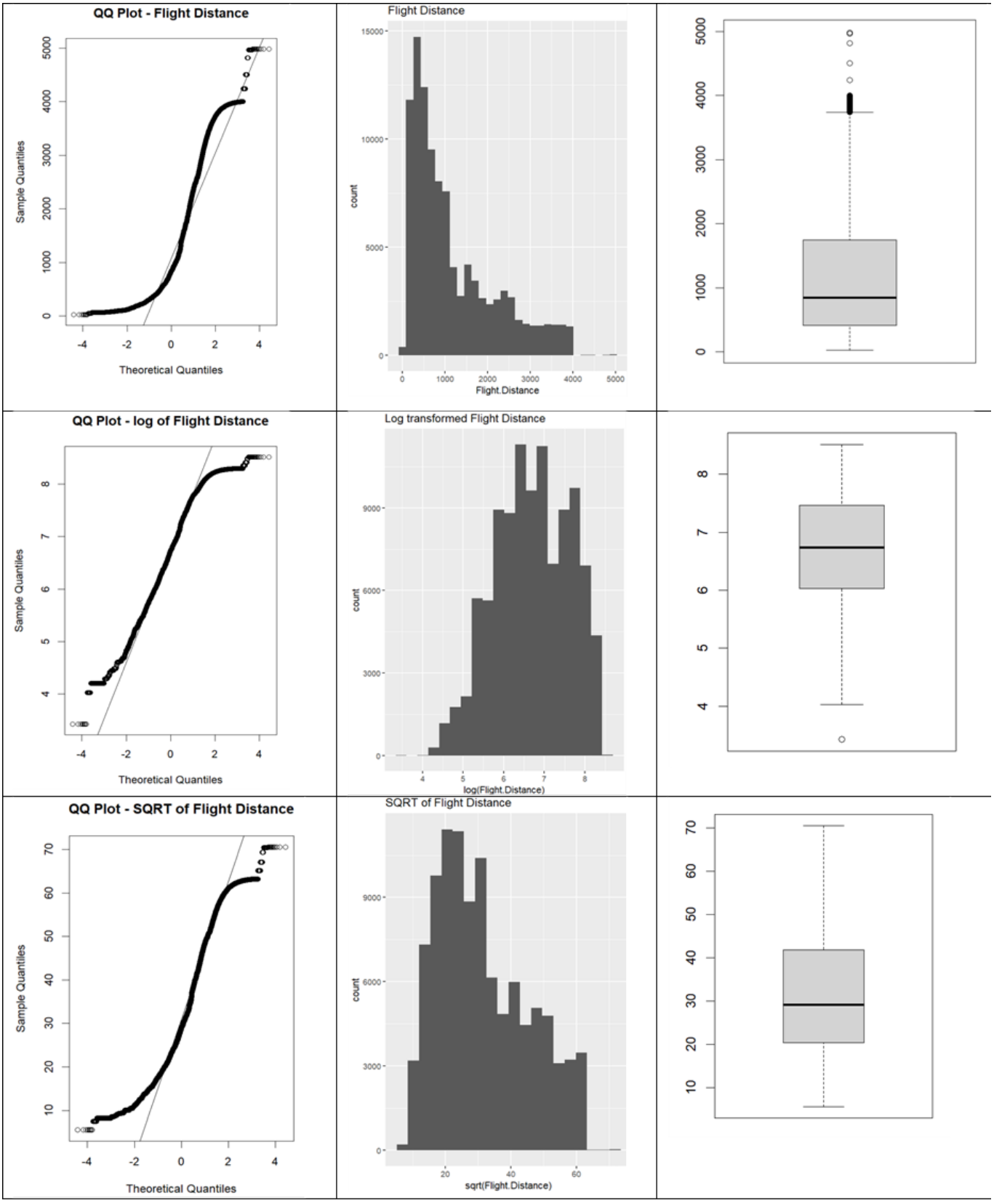
https://psycnet.apa.org/record/1989-10632-001

https://www.techscience.com/cmc/v75n1/51460/pdf


[4] https://www.techscience.com/cmc/v75n1/51460/pdf

## 9. Appendix

Appendix 1: Plots of flight distance



Appendix 2: Plots of Departure delay in minutes