# Aztec Quant Consulting


| | | |
|---|---|---|
| **Team Members:** | Sammer Khalaf | Red ID: 827169857 |
| | Bita Etaati | Red ID: 827176552 |
| | Vincent O'Dea | Red ID: 825895090 |
| | Mohammed Al Ani | Red ID: 827161576 |

Table of contents

## Executive Summary

The airline industry has suffered greatly in the last few years due to turbulent market conditions brought about by the Covid-19 pandemic. It is estimated that North American airlines lost upwards of $40 billion in revenue in the year 2020 as passenger volume dropped by sixty percent (Troy, 2021)[1]. Aztec Quant Consulting specializes in the detailed analysis of "big data" and provides clients with uniquely catered solutions to their everyday business problems. This paper will explore the steps and procedures used to investigate customer satisfaction in the airline industry. Our analysis will be based on the results of 129,880 satisfaction survey results, which will enable us to draw statistically valid conclusions that our clients can use and adopt into their daily operations. It is well established that customer satisfaction leads to customer loyalty , and similarly customer loyalty drives profitability and organizational growth (Heskett, 2008)[2]. With this in mind our objective is to clearly identify what are the key determinants that lead to a satisfied customer.

## Introduction

Satisfaction is defined by the Mirriam-Webster dictionary as the "fulfillment of a need or want". Airline passengers come in all shapes and sizes, as do their needs. Some may simply need to get from A to B for business, others may be embarking on a trip of a lifetime so it is important that our solutions acknowledge this range of expectations, whilst at the same time provide concise, practical, and implementable steps that our clients can follow and realize the benefits.

Our analysis was primarily conducted through R Studio using the R programming language and is based on the dataset which was acquired from Kaggle. The original dataset had been pre split into two files, a training and test set. The two files were downloaded and merged using Microsoft Excel, which allowed our team a better control of the preprocessing steps we wished to

implement. The data has just under 130,000 observations, with 25 columns, which we will refer to as the project variables. Please refer to Figure 1 below for a glimpse of the variables that were assessed in our study. Variables were provisionally assessed as seen below but modified which we will outline and rationalize in later passages of this report.

**Figure 1**. *Project Variables and structure.*

```
> str(airline_satisfaction)
spc_tbl_ [129,880 × 25] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
$ ...1                             : num [1:129880] 0 1 2 3 4 5 6 7 8 9 ...
$ id                               : num [1:129880] 19556 90035 12360 77959 36875 ...
$ Gender                           : chr [1:129880] "Female" "Female" "Male" "Male" ...
$ Customer Type                    : chr [1:129880] "Loyal Customer" "Loyal Customer" "disloyal Customer" "Loyal Custom
er" ...
$ Age                              : num [1:129880] 52 36 20 44 49 16 77 43 47 46 ...
$ Type of Travel                   : chr [1:129880] "Business travel" "Business travel" "Business travel" "Business tra
el" ...
$ Class                            : chr [1:129880] "Eco" "Business" "Eco" "Business" ...
$ Flight Distance                  : num [1:129880] 160 2863 192 3377 1182 ...
$ Inflight wifi service            : num [1:129880] 5 1 2 0 2 3 5 2 5 2 ...
$ Departure/Arrival time convenient: num [1:129880] 4 1 0 0 3 3 5 2 2 2 ...
$ Ease of Online booking           : num [1:129880] 3 3 2 0 4 3 5 2 2 2 ...
$ Gate location                    : num [1:129880] 4 1 4 2 3 3 5 2 2 2 ...
$ Food and drink                   : num [1:129880] 3 5 2 3 4 5 3 4 5 3 ...
$ Online boarding                  : num [1:129880] 4 4 2 4 1 5 5 4 5 4 ...
$ Seat comfort                     : num [1:129880] 3 5 2 4 2 3 5 5 5 4 ...
$ Inflight entertainment           : num [1:129880] 5 4 2 1 2 5 5 4 5 4 ...
$ On-board service                 : num [1:129880] 5 4 4 1 2 4 5 4 2 4 ...
$ Leg room service                 : num [1:129880] 5 4 1 1 2 3 5 4 2 4 ...
$ Baggage handling                 : num [1:129880] 5 4 3 1 2 1 5 4 5 4 ...
$ Checkin service                  : num [1:129880] 2 3 2 3 4 1 4 5 3 5 ...
$ Inflight service                 : num [1:129880] 5 4 2 1 2 2 5 4 3 4 ...
$ Cleanliness                      : num [1:129880] 5 5 2 4 4 5 3 3 5 4 ...
$ Departure Delay in Minutes       : num [1:129880] 50 0 0 0 0 0 0 77 1 28 ...
$ Arrival Delay in Minutes         : num [1:129880] 44 0 0 6 20 0 0 65 0 14 ...
$ satisfaction                     : chr [1:129880] "satisfied" "satisfied" "neutral or dissatisfied" "satisfied" ...
```

The structure of the project was based on answering research questions that we at Aztec Quant Consulting established with our client through a series of interviews. The questions are based on their needs and questions they wanted investigated and answered. The research questions we established can be found in Table 1 below.

**Table 1**. *Research Questions.*

| | |
|---|---|
| 1 | What are the most important features of a good flight from the customer's perspective? |

| | |
|---|---|
| 2 | Is there a difference in satisfaction rate between travelers traveling for business and those traveling for leisure? |
| 3 | What is the association between age and gender and flight satisfaction? |
| 4 | Does customer loyalty have a significant effect on satisfaction? |
| 5. | What effect does the boarding process have on overall customer satisfaction? |

This paper will explore a few analytical models that were builts at the behest of the client.

1. Logistic Regression.
2. Naive Bayes.
3. Decision Tree.
4. Random Forest.

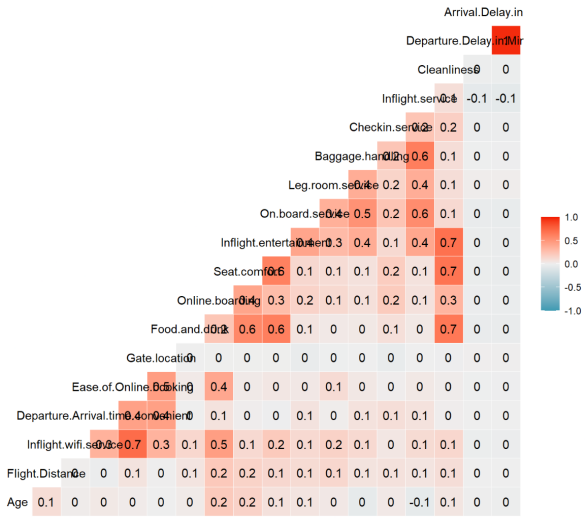
The primary objective of this project is to provide answers to the research questions established with our client. This will be achieved by detailing our exploratory analysis of some of the features that proved interesting and will give a holistic view of the data under investigation. We will then visit our model planning phase where a brief theoretical background of the models in use, as well as the rationale behind their selection. Following this a detailed account of how our models were built will be provided as well a series of figures and visualizations that will validate our findings. To conclude our results will be summarized and discussed. Limitations of the analysis will be explored as well as areas where we believe further research can be carried out to provide further insight into the topic matter.

## Exploratory Data Analysis

As mentioned previously this project is based upon the findings of a survey taken from 129,880 individuals and has 23 features or variables. Below are some some useful insights on the dataset:

- The dataset originally had 393 rows with missing values. These empty fields we omitted which we believe is justified due to shear size of the dataset, which possess 129,880 data points.
- The predictor variables that describe quality of service had ordinal scales that ranged from 1-5 .
- Correlation matrix: only assessed quantitative variables.

**Figure 2**. *Quantitative Variables of Dataset.*




For the visualization aspect of the project and exploratory analysis, we have divided these features into 3 major categories of passengers' demographics, flight characteristics and survey questions. To get a better insight on our dataset, we will dig into each of these categories.

1. Passengers' Demographics:

In terms of gender, we can say the female and male population is almost the same (49% are men and 51% are women). Most of the passengers participating in the survey are loyal customers (82%). While we have passengers as young as 7 years old, the maximum age seen in our dataset is 85. In the exploratory part, based on a guideline from Involve Communities [3], we defined seven age ranges as per the below table:

**Table 2.** *Passenger Demographics.*

| Age Group | Observation Percentage |
|---|---|
| Under 18 | 7% |
| 18 to 24 | 11% |
| 25 to 34 | 19% |
| 35 to 44 | 23% |
| 45 to 54 | 21% |
| 55 to 64 | 14% |
| Above 65 | 5% |

2. Flight Characteristics:

Flight distance is one of the pieces of information that we think might be correlated with customer satisfaction, hence it is important for our project purpose to review this feature more precisely. Based on Wikipedia [4] Flight Length category is defined as short haul (less than 600 miles), medium haul (between 600 to 2100 miles) and long haul (more than 2100 miles) flights. As we can see, most flights were short to medium haul. 48% of these flights were business class and about 52% were economy, where 70% were for business reasons.

**Table 3.** *Flight Characteristics.*

| Flight Length | Observation Percentage |
|---|---|
| Short Haul | 38% |
| Medium Haul | 42% |
| Long Haul | 20% |

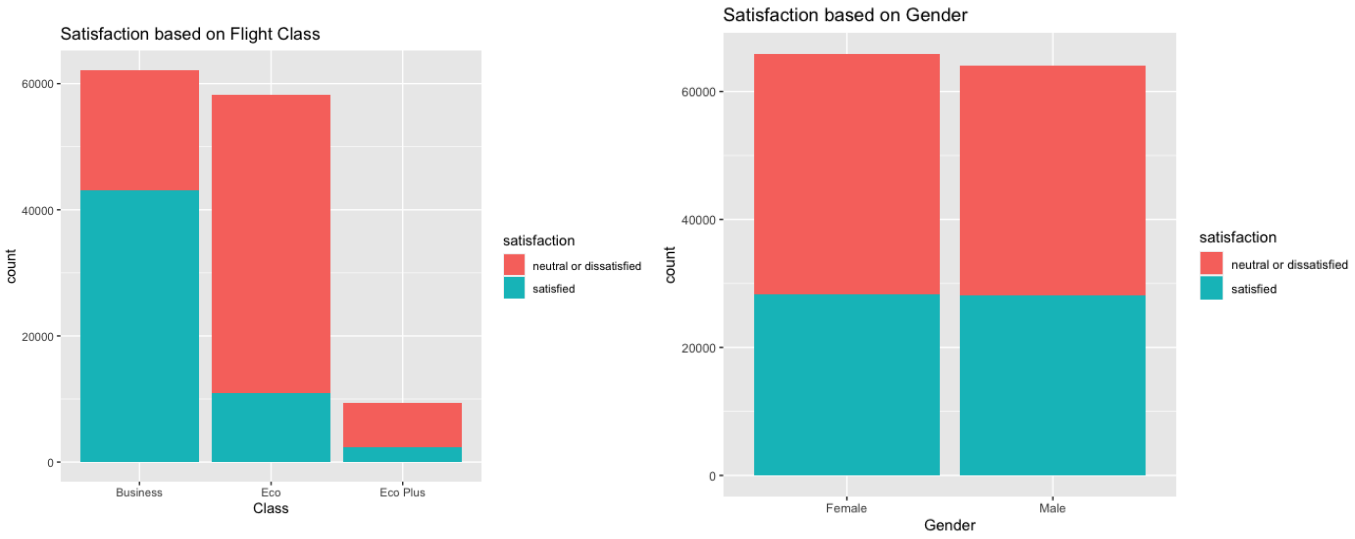
Since our dataset provides arrival and departure delay (in minutes), below is a statistical summary on these two variables. These delays range from 0 to almost 25 hours, with a mean of approximately 15 minutes.

**Figure 3.** *Summary of Delays in Departures (RStudio)*

```
summary(data$`Departure Delay in Minutes`)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.00    0.00    0.00   14.71   12.00 1592.00 
summary(data$`Arrival Delay in Minutes`)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.00    0.00    0.00   15.09   13.00 1584.00 
```

As seen in the chart below (Figure 4) there is no visible difference between satisfaction in men and women. The number of satisfied customers is slightly lower compared to the other group. However, as we were expecting, the satisfaction rate is higher in business class flights.

**Figure 4.** *Satisfaction Based on Flight Class, Satisfaction Based on Gender.*



Regarding age groups, we can see the age ranges with lower discomfort tolerance (senior citizens and children under 18) were mostly dissatisfied with their flights.

**Figure 5**. *Satisfaction Based on Travel Type.*

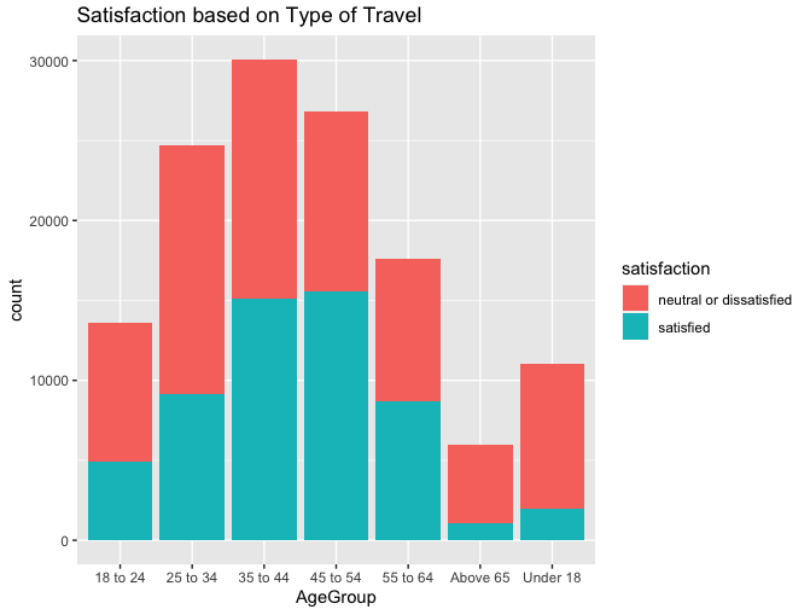
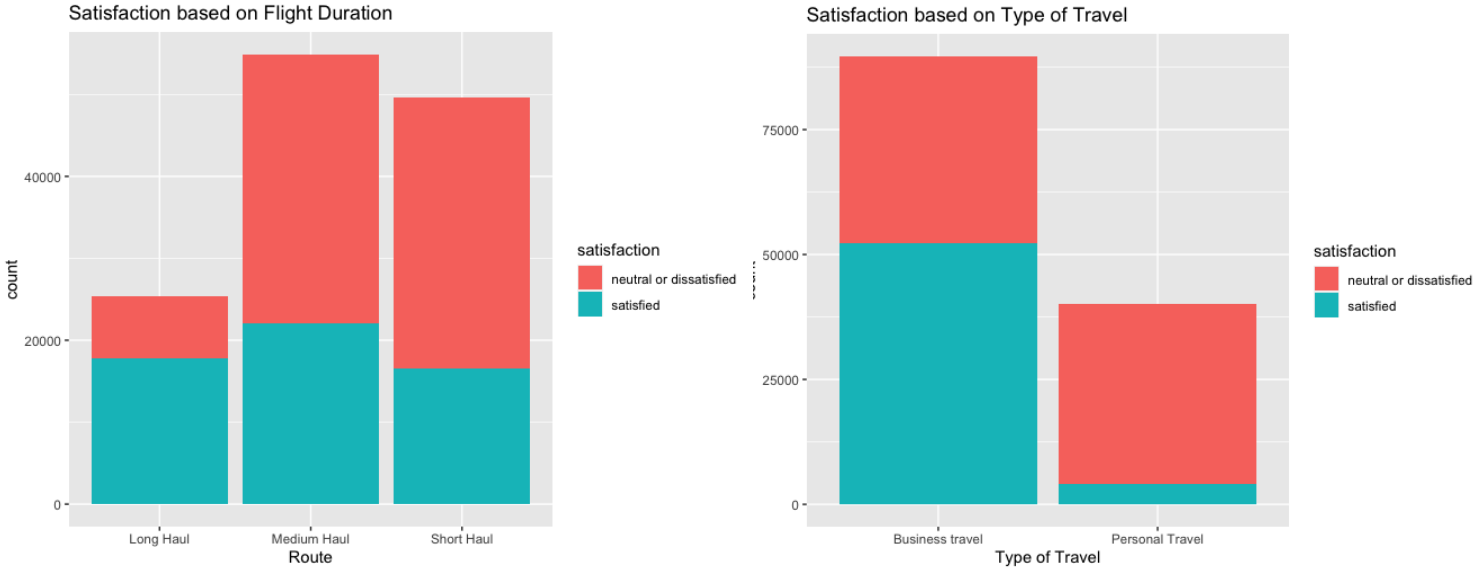



**Figure 6.** *Satisfaction based on Flight Duration & Travel Type.*




We were also curious to discover potential correlation between customer satisfaction and flight duration. Even though we cannot see any visible pattern, the satisfaction ratio seems to be higher

in long haul flights. Besides flight duration, type of travel was another factor that might be correlated with passenger satisfaction. Surprisingly, there were only a few passengers satisfied with their flights when they were having personal trips.

## Model Planning

### Naive Bayes

Naive Bayes assumes that features are independent of each other and there is no correlation between features. The outcome of the naive bayes model depends on a set of independent variables that have no correlation with each other. Naive Bayes performs quite well when the training data doesn't contain all possibilities so it's dynamic with missing values of data. Naive bayes is a supervised learning algorithm for classification so the underlying functionality is to find the class of observation (data point) given input. Since we are using multiple models in this project, our team members believed it was necessary to use Naive Bayes as the threshold when comparing accuracy. Naive Bayes is predominantly notorious for being a classification model that is optimal and reliable with big data sets. The Naive Bayes model is easy to build and particularly useful for very large data sets. When dealing with a large dataset our first thought was about Naive classification.

Ultimately, predictor variables aren’t always independent of each other, there are always some correlations between them. Since Naive Bayes considers each predictor variable to be independent of any other variable in the model, it is using a ‘Naive’ approach. The principle behind Naive Bayes is the Bayes theorem also known as the Bayes Rule. The Bayes theorem is used to calculate the conditional probability, which means the probability of an event occurring based on information about past events. Therefore, utilizing the satisfaction rates in the past we are able to predict future probabilities using Naive Bayes.

**Logistic Regression**

Since the response variable is qualitative our team believed that Logistic Regression should be the first model considered in our project. Logistic regression is more concerned with the probability of our dependent variable (satisfaction) belonging to a certain class of “satisfied” or “neutral or dissatisfied”. Logistic regression uses the underlying linear regression model to predict the probability of the dependent variable being classified in a particular category. The methods used for classification first predict the probability that the observation belongs to each of the categories of a qualitative variable, as the basis for making the classification..Since the outcome of the binary logistic regression will fall between 0 and 1, it is essential to model by using the logistic function to ensure that our predicted result will not go over range of 0 and 1.

If we use linear regression, some of our estimates might be outside the [0,1] interval, making them hard to interpret as probabilities. Logistic regression introduces the concept of log odds in the model, where it calculates the probability of an event occurring divided by the probability of an event not occurring. For estimating the regression coefficients, the logistic regression model uses the technique of maximum likelihood where it uses estimates of $\beta_0$ and $\beta_1$ to align the predicted probability of risk for each consumer against the observed value of risk for each consumer. To evaluate the effectiveness of our binary logistic regression model, we can compute the standard error of our coefficients and analyze the *z-statistic* results. Ultimately, logistic regression can help make predictions about binary responses by using many independent variables.

**Decision Trees:**

A decision tree machine learning algorithm was performed to describe customer satisfaction in terms of other variables in the dataset. Decision trees work with classification problems. Decision trees are highly interpretable by humans because they resemble human decision

making. On the other hand, decision trees are not expected to provide the highest accuracy compared to random forests. Decision trees provide a good baseline accuracy measurement for random forests especially as random forests become more computationally expensive to run.

**Random Forests:**

A random forest is a more sophisticated machine learning algorithm than decision trees. Random forests average out many trees by taking advantage of bagging and bootstrapping. Bagging is a method of aggregating different bootstrapped samples while bootstrapping is random cross validated sampling. While random forests are not the perfect solution to all problems, they tend to provide more accuracy over decision trees.

## Model Building

### Naive Bayes

Naive Bayes is implemented in R using the naiveBayes() function, which is part of the e1071 library. By default, this implementation of the Naive Bayes classifier models each quantitative feature using the preprocessed data distribution. The first step in our model building is to set up a train and test split. We have dedicated 70% of our data for training and the remaining 30% for testing our predictions to measure the accuracy of our model. We used the coefficient of each independent variable relating to the dependent variable to understand which are significant enough to keep in our model. Some of the dropped variables were ***Gender and ID Number.*** Their p value was above our threshold and gave us a good indication that we should keep the variables out of our model. After all, using predictors that have no relationship with the response tends to cause a deterioration in the test error rate since such predictors cause an increase in variance without a corresponding decrease in bias.

### Logistic Regression

Once again, we have followed the same methodology that we implemented in the naive bayes model. In this model we really emphasized the importances of dropping null values from the data set to ensure the classification is performing at the most optimal level. We used na.omit(data) to conclude this situation .We found which variables were significantly correlated with the predictors to keep in our model. We had a little change of variables being dropped and included ***Gender, ID number, and Type of Travel***. After removing the variables from the model we also had to ensure that our models were leveled in the same factor setting consistently across the entire dataset and categorized the input variables in an ordinal measurement scale. Our response

indicated that anything below a 0.5 was considered dissatisfied or neutral and anything above a 0.5 would be considered satisfied.

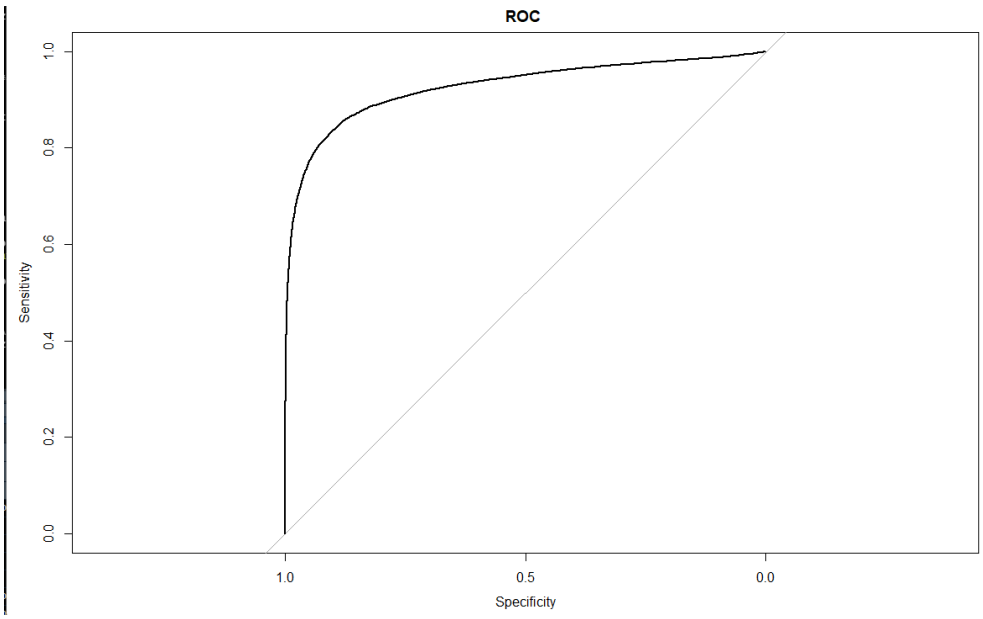
## Results and Performance

### Naive Bayes

**Figure 7.** *Naive Bayes Confusion Matrix.*

```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   19425      2835
  satisfied                                  2690     14014

               Accuracy : 0.8582
                 95% CI : (0.8547, 0.8617)
    No Information Rate : 0.5676
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7108

 Mcnemar's Test P-Value : 0.05271

            Sensitivity : 0.8784
            Specificity : 0.8317
         Pos Pred Value : 0.8726
         Neg Pred Value : 0.8390
             Prevalence : 0.5676
         Detection Rate : 0.4985
   Detection Prevalence : 0.5713
      Balanced Accuracy : 0.8551

       'Positive' Class : neutral or dissatisfied
```

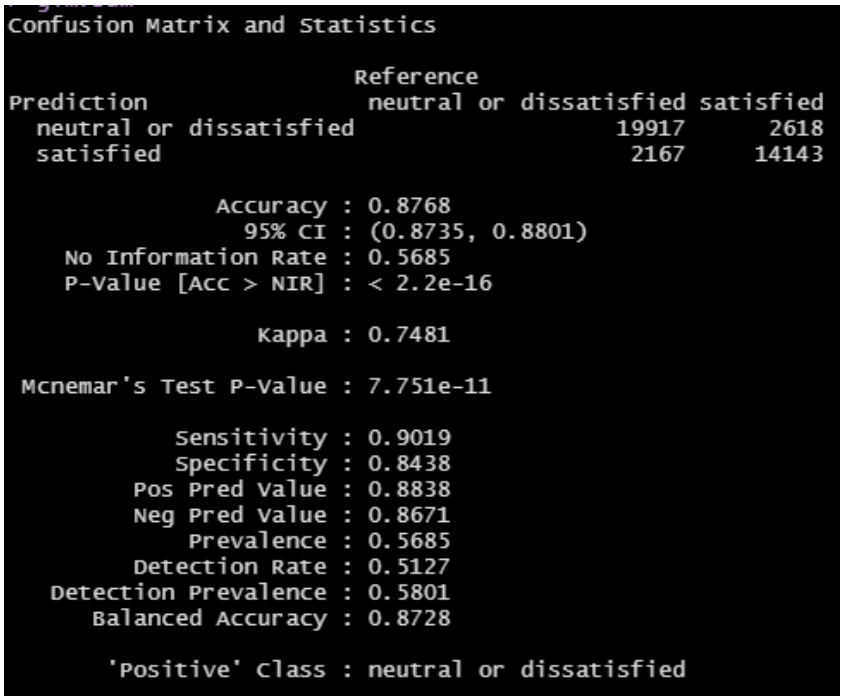
**Figure 8.** *Naive Baye ROC Curve.*



Since our prediction is about satisfaction, we can see where the false positives and false negatives are presented from the preprocessed data. The model produced an accuracy level of 85% derived from the true response in the confusion matrix. Naive Bayes model was the second to lowest accurate model that we have implemented, but it proves that our other models are performing in a reliable and accurate way. Our precision recall is still significant at 87% and we should tune it down to a lower value by adjusting the data used in the model. Overall it does not seem that there is significant bias between our dependent variable prediction results. The AUC for the ROC Curve is at 92% which signifies good accuracy. The ROC plot is also hugging the top left corner which further supports our model performance.

**Logistic Regression**

**Figure 9.** *Logistic Regression Confusion Matrix.*

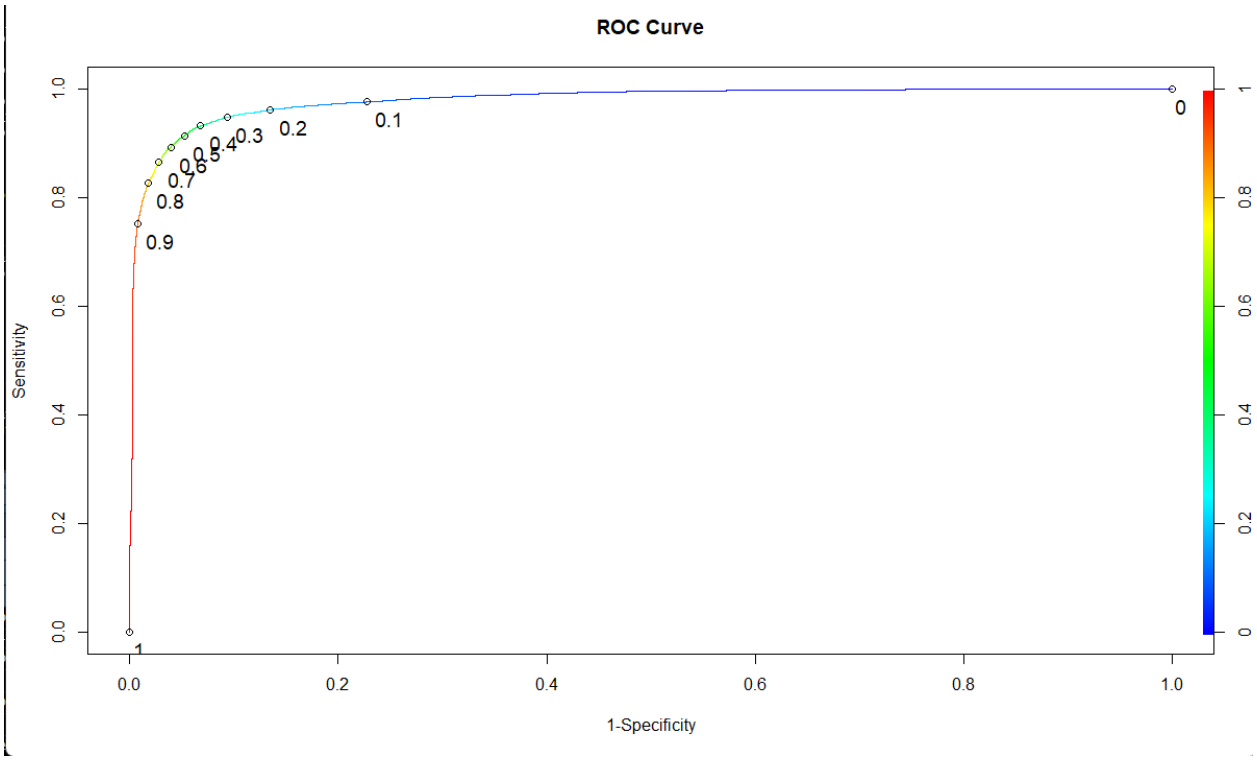


```
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   19917      2618
  satisfied                                  2167     14143

               Accuracy : 0.8768
                 95% CI : (0.8735, 0.8801)
    No Information Rate : 0.5685
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7481

 Mcnemar's Test P-Value : 7.751e-11

            Sensitivity : 0.9019
            Specificity : 0.8438
         Pos Pred Value : 0.8838
         Neg Pred Value : 0.8671
             Prevalence : 0.5685
         Detection Rate : 0.5127
   Detection Prevalence : 0.5801
      Balanced Accuracy : 0.8728

       'Positive' Class : neutral or dissatisfied
```

We fitted a logistic regression model in order to predict Direction using all variables except Gender , ID number , and Type of Travel . Logistic regression estimates the probability of

an event occurring, such as satisfied or not, based on a given dataset of independent variables. The logistic regression model's prediction accuracy is 88% and misclassification is 12%. Both measures indicate reliability of the model.Some area of improvements can be focused on Precision with its value being ranked at 88% . Logistic regression showed that for customer satisfaction, the following factors are positively significant which are loyal customers, convenient departure and arrival time, and ease of booking. This means when these variables increase, the probability of satisfaction increases. We have also discovered that an increase in delay of flight arrival decreases the probability of satisfaction .

**Figure 10.** *Logistic regression ROC Curve.*




The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the (ROC) curve (AUC). An ideal ROC curve will hug the top left corner, so the larger area under the AUC the better the classifier. For this data the AUC is 0.97, which is close to the maximum of one so would be considered very good. We expect a classifier that

performs no better than chance to have an AUC of 0.5. ROC curves are useful for comparing different classifiers, since they take into account all possible thresholds.The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value.

**Decision Tree:**

Our implementation produced a test accuracy of 84% with an AUC of approximately 0.92. It is important to note that it is most likely the case that this decision tree was overfit. When we reported the accuracy and AUC for the tree, it was referring to the unpruned tree.

***Figure 11.*** *Decision Tree*

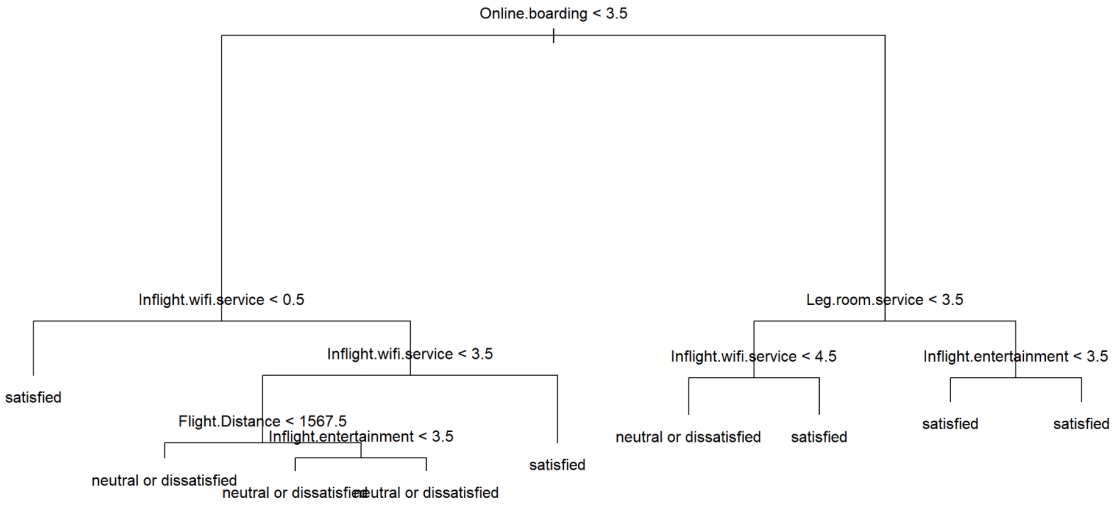
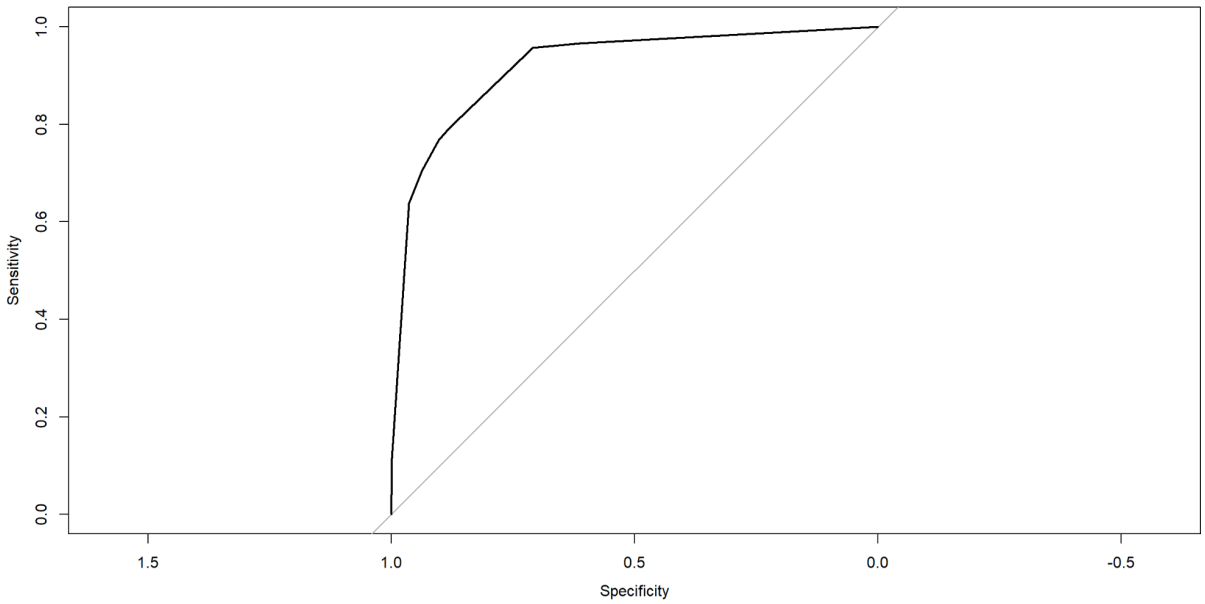


**Figure 12.** *Decision Tree Confusion Matrix.*

```
> cm = confusionMatrix(data= as.factor(tree.pred), reference=as.factor(test.y))
> cm
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   19897      3897
  satisfied                                  2157     12895

               Accuracy : 0.8442
                 95% CI : (0.8405, 0.8477)
    No Information Rate : 0.5677
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6785

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9022
            Specificity : 0.7679
         Pos Pred Value : 0.8362
         Neg Pred Value : 0.8567
             Prevalence : 0.5677
         Detection Rate : 0.5122
   Detection Prevalence : 0.6125
      Balanced Accuracy : 0.8351

       'Positive' Class : neutral or dissatisfied
```

**Figure 13.** *Decision Tree ROC Curve.*



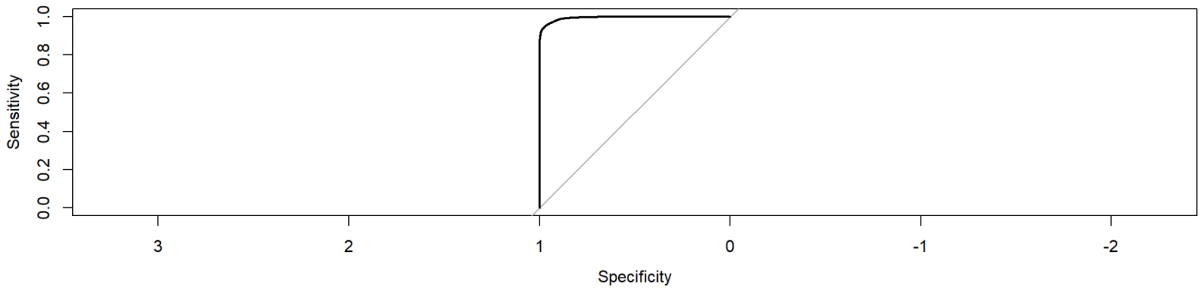
**Random Forest:**

Our implementation produced a test accuracy of 96% with an AUC of approximately 0.99. This is a great result in terms of prediction accuracy and predictive power gained over a random guess. It is also clear that our random forest implementation is more accurate than the naive bayes, logistic regression and decision tree models.

**Figure 14**. *Random Forest Confusion Matrix.*

```
> plot(rf)
> tree$pred <- predict(rf, test, type = "class")
> cm = confusionMatrix(data= as.factor(tree$pred), reference=as.factor(test.y))
> cm
Confusion Matrix and Statistics

                          Reference
Prediction                 neutral or dissatisfied satisfied
  neutral or dissatisfied                   21632       997
  satisfied                                   422     15795

               Accuracy : 0.9635
                 95% CI : (0.9616, 0.9653)
    No Information Rate : 0.5677
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9253

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9809
            Specificity : 0.9406
         Pos Pred Value : 0.9559
         Neg Pred Value : 0.9740
             Prevalence : 0.5677
         Detection Rate : 0.5569
   Detection Prevalence : 0.5825
      Balanced Accuracy : 0.9607

       'Positive' Class : neutral or dissatisfied
```

**Figure 15**. *Random Forest ROC Curve.*




**Insights from the random forest model:**

- Gender is not highly important in predicting satisfaction at least from a relative perspective when compared to other variables in the dataset

- Online boarding seems to be the the most important variable while inflight wifi service is the second most important variable relatively speaking
- Class and type of travel equally important are equally important next
- It is surprising that age, gender and type of the flight are not high on the variable importance list. This is despite a chi square test of independence showing a statistical dependence between gender with satisfaction and customer type with satisfaction. The p-values were extremely small for both tests.

**Figure 16**. *Variable Identified as Important.*

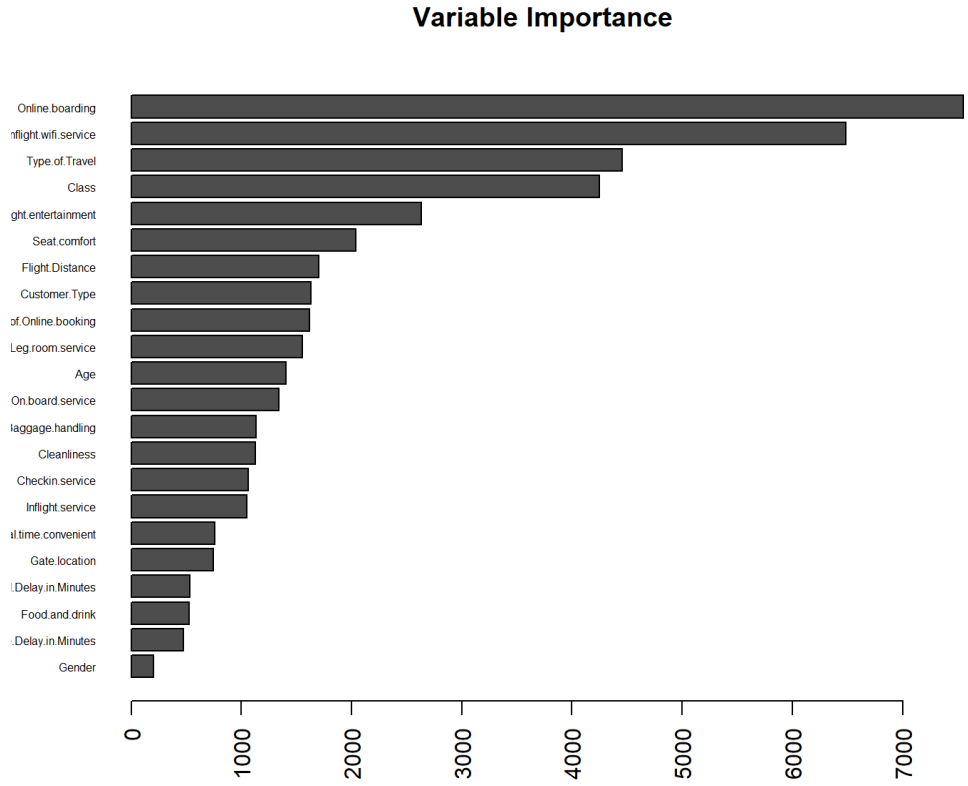


**Table 4.** *Summary of models used and their Accuracy and AUC .*

| Model | Accuracy | AUC |
|---|---|---|
| Logistic Regression | 88% | 0.97 |
| Naive Bayes | 85% | 0.92 |
| Decision Tree | 84% | 0.92 |
| Random Forest | 96% | 0.99 |

## Discussion and Recommendations:

The exploratory analysis of this project uncovered many useful insights of the data analyzed. To add clarity to the age variable, we elected to segment it into seven different age brackets, of which the 35-44 age group was the largest cohort of passengers surveyed at 23%.

The flight distance in miles was similarly split into groups which we labeled short, medium and long haul. Medium haul flights were the most common flights seen in the dataset. Gender and Flight Class was next assed. Unsurprisingly those individuals traveling business class proved far more likely to be satisfied than those traveling economy class. There appears to be no clear difference in satisfaction rate between genders. Regarding age groups, we can see the age ranges with lower discomfort tolerance (senior citizens and children under 18) were mostly dissatisfied with their flights. These age ranges offer a fantastic opportunity for our clients to target and aim to appease them, which will bolster their overall satisfaction ratings.

The model planning and performance areas proved crucial in enabling us to determine what are the key factors driving satisfaction amongst those surveyed. The logistic regression model detailed that loyal customers, convenient departure / arrival time, and ease of booking are areas that airlines can aim to improve and cater for. It additionally reinforced the well known truth that an increase in flight delay in minutes negatively affects an individual's satisfaction. Figure 11 of the previous section outlines the results obtained from our decision tree model which is a fantastic visual aid and really provides a straightforward intuitive assessment of variables affecting our response variable satisfaction. Online boarding proved to be the greatest predictor of satisfaction through this model which is later supported by the random forest model. A summary of models built and their accompanying accuracy as well as AUC score can be seen in

Table 4 above. Random Forest proved to be the most accurate model, who's results can be found in Figure 16.

The dataset proved to be challenging, yet rewarding to work with. We were able to correctly address all of the research questions earmarked by the client however, there are definite areas that we feel further investigations are needed to fully address satisfaction rate. The primary shortcoming identified is the absence of flight cost / baggage cost in the survey questions. This is one area that we would like to incorporate into future investigations. Additionally we feel that knowing if a passenger is on a connecting flight, the destination climate, and flight frequency might prove significant in whether an individual is likely to be satisfied with the service. It is our belief that all of these fields mentioned above warrant research.

## References


[1] Troy, M., (2021, May 19). *Attentive Flight Crews, Flexible Fares and Charges during Pandemic Drive Record High Customer Satisfaction with North America Airlines, J.D. Power Finds.* JD Power. from https://www.jdpower.com/business/press-releases/2021-north-america-airline-satisfaction-study

[2] Heskett, J.L.,et al. (2008, July). *Putting the Service-Profit Chain to Work.* Harvard Business Review, from https://hbr.org/2008/07/putting-the-service-profit-chain-to-work

[3] Lewis Hankinson, 2010. (2008, October 28). *Neighbourhood renewal*. Kirklees Council. Retrieved December 13, 2022, from https://www.kirklees.gov.uk/involve/

*[4] Flight length*. (2022, December 13). Wikipedia, the free encyclopedia. Retrieved December 13, 2022, from https://en.wikipedia.org/wiki/Flight_length

https://www.upgrad.com/blog/random-forest-vs-decision-tree/