Important factors Affect Airline Customers Preference

Group 13
Group Members:
Chow Hin Shing 58089000
Li Wang Cheong 57149051
Chau Yu Shing 57910971
Keung Wan Hei 57137691
Ho Wing Tat Tommy 56400912

01

Problem statement

Problem Statement

Problem Motivation - Why it is important?

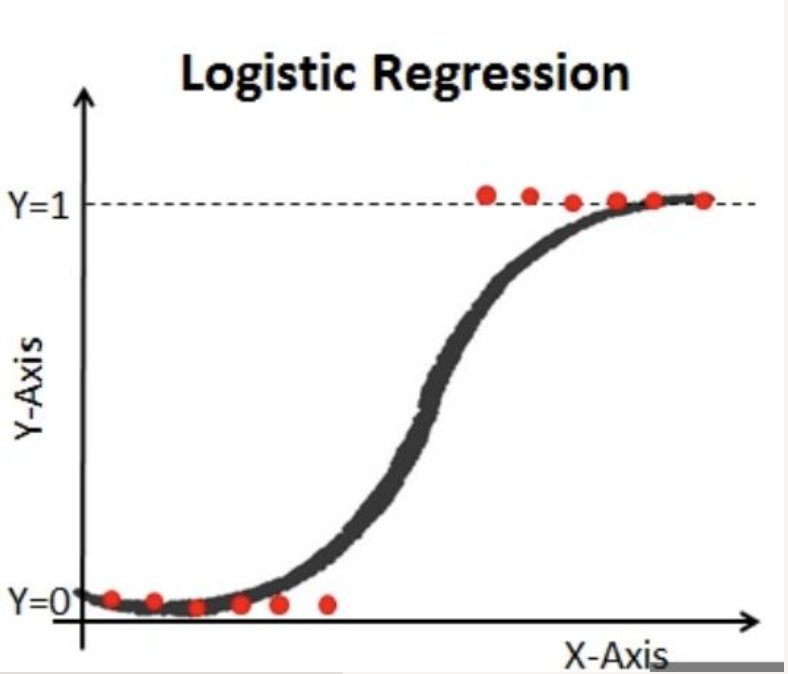
Customer satisfaction :

- Increases brand loyalty
- Attracts new customers
- Reduce customers complaints

Therefore, we hope to answer the following questions:

1. What variables affect customer satisfaction the most?
2. Based on the above analysis, what suggestions can we make to the airlines?

2. Problem Statement



2.2 Problem Type and Model

- We will use logistic regression modeling to forecast the association between a certain factor (X) and customer satisfaction (Y) since customer satisfaction (Y) is a nominal data. This will help us identify the specific factor that requires to be improved.

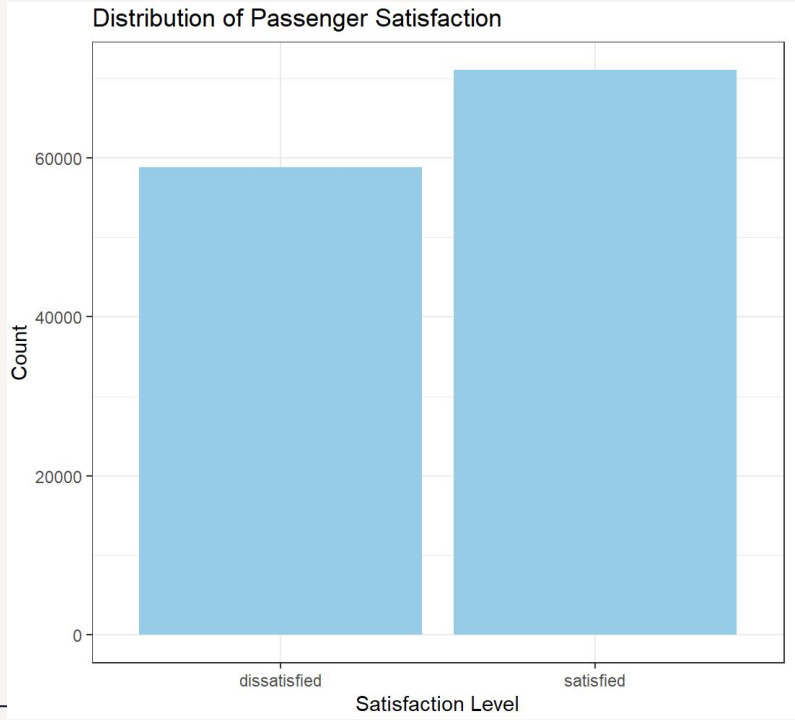
02

Data Description, Exploration

2. Data Description & Exploration

- Dataset: Airline Customer's Preferences & Satisfaction
- 129880 Observations
- 23 columns
- Y variables:
  - Satisfaction
- X variables:
  - Departure/ Arrival time convenient
  - Inflight entertainment
  - On board service

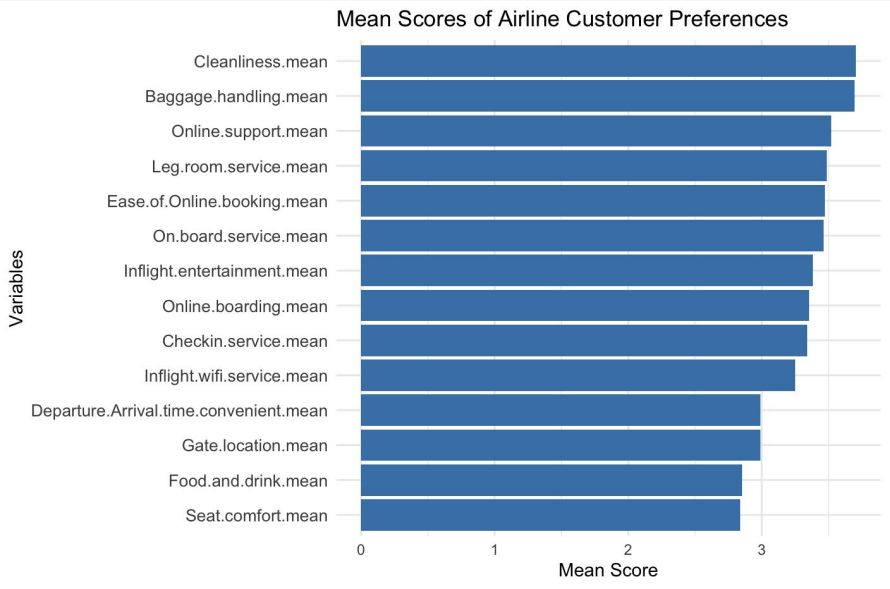
Distribution of Y : (satisfaction)



Satisfaction Level	Count	Proportion
Satisfied	71087	0.55
Dissatisfied	58793	0.45

- Proportion of Satisfied customer is not high

Data Description & Exploration



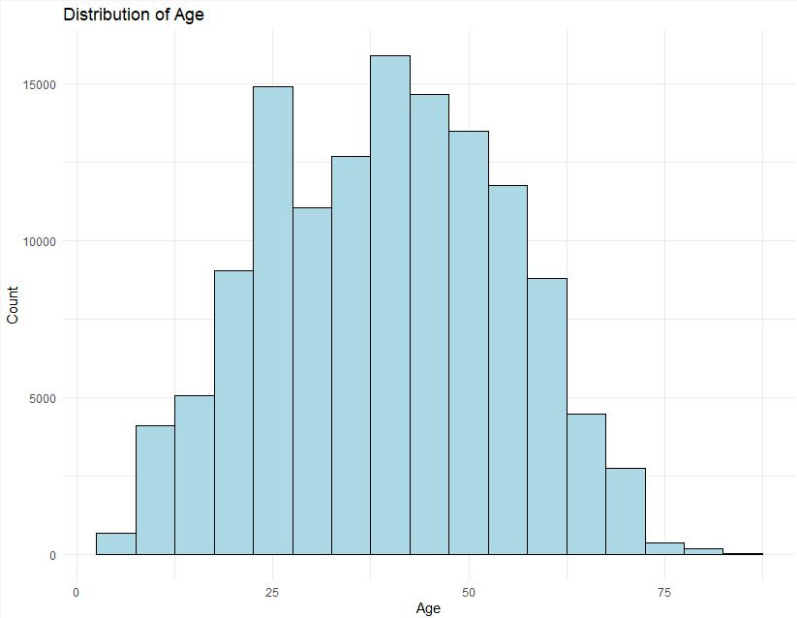
Key preferences identified:

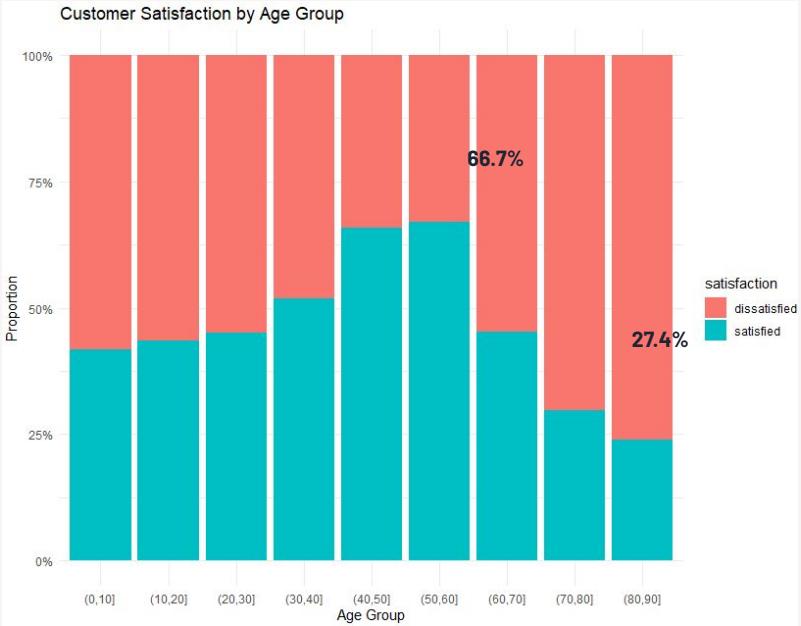
Cleanliness has the highest mean score. Seat comfort has the lowest mean score.

Lower scores may indicate possible sources of discontent that can result in client attrition.

Companies in this industry may improve quality of food and drinks and seat comfort to gain better customer loyalty.

Data Description & Exploration

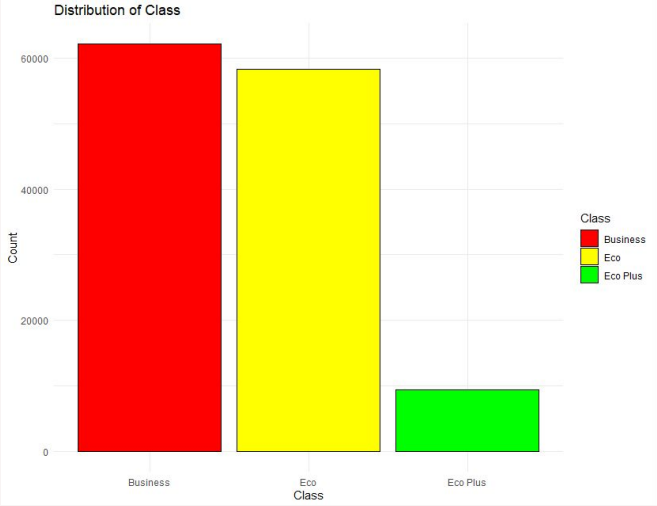


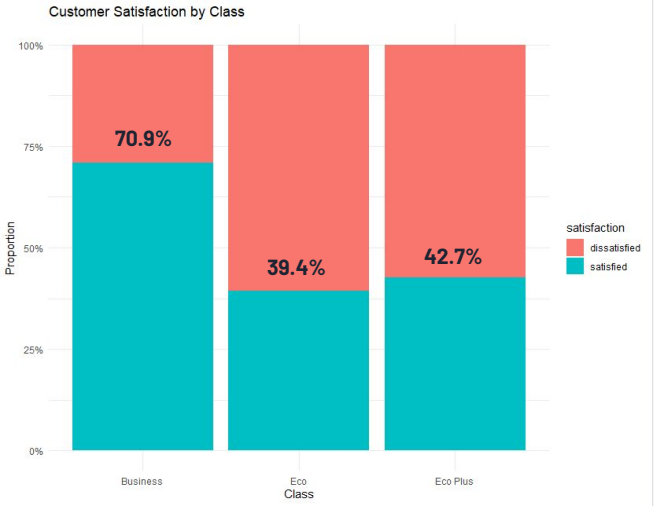


For satisfaction rate among different **Age**:

[40-60] >> Other age group

Data Description & Exploration



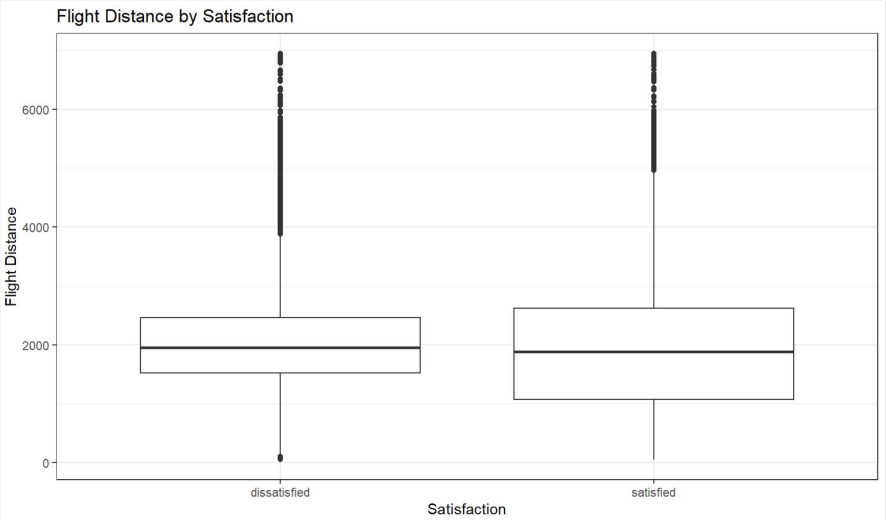


For satisfaction rate among different **Class**:

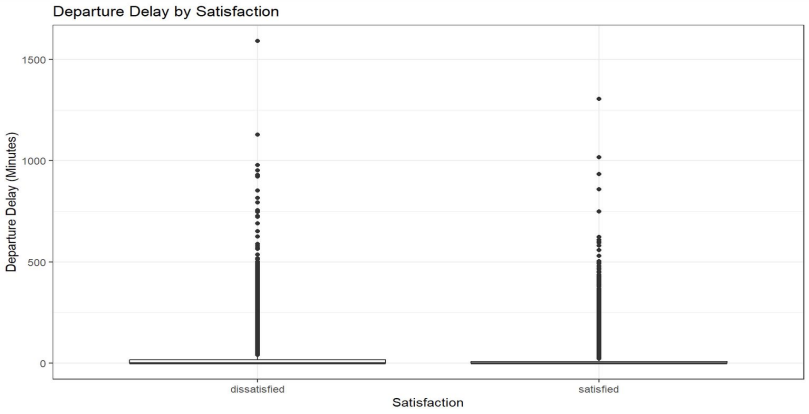
Business class passenger >> Economy, Economy Plus

Data Description & Exploration

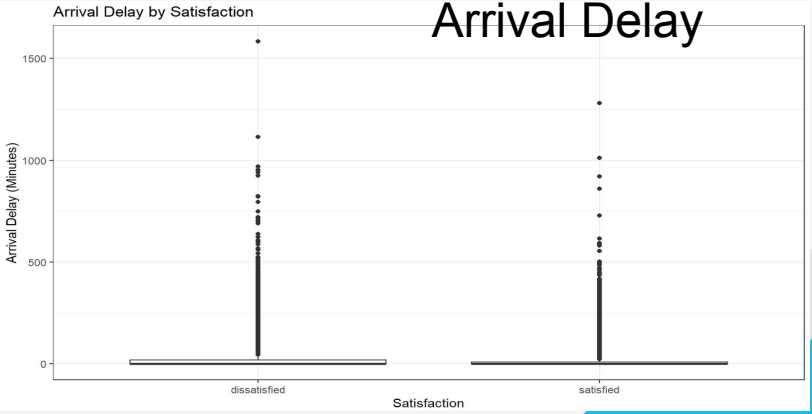
Flight Distance



Departure Delay

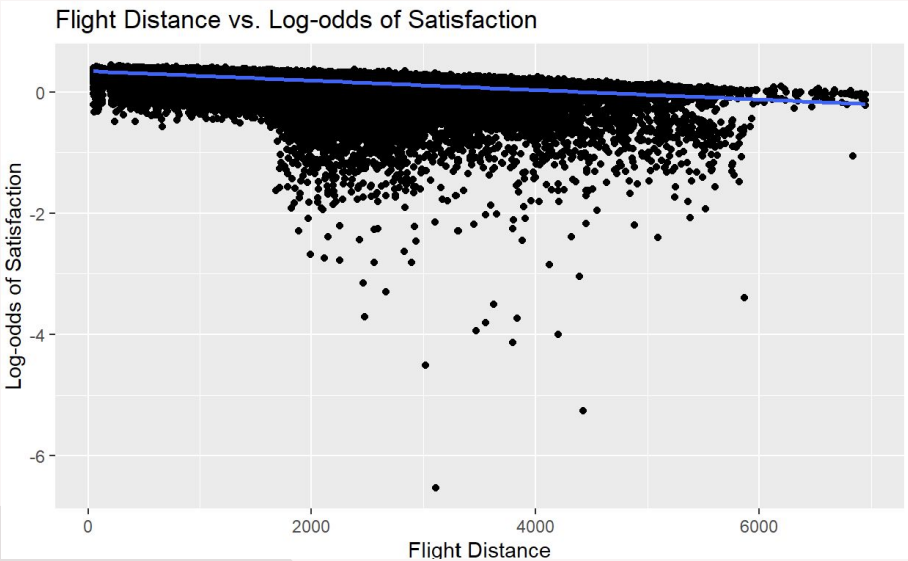


Arrival Delay

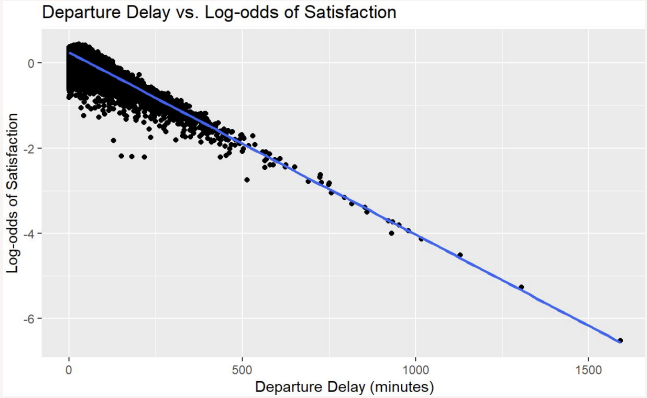


Data Description & Exploration

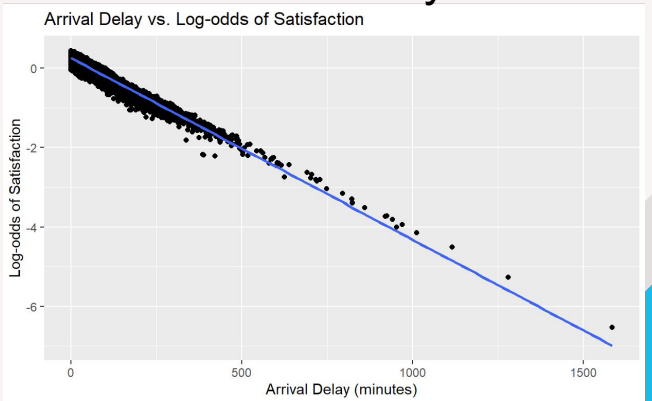
Flight Distance



Departure Delay



Arrival Delay



03

Data cleaning & processing

Data cleaning & processing

1. Delete rows that contains missing values

```
#read csv
airline <- read.csv('Airline.csv')

#Clear Missing data
air <- na.omit(airline)
```

Data	
air	129487 obs. of 23 variables
airline	129880 obs. of 23 variables

2. Modify three variables that contains outliers

Log Transformation (for Delays)

```
air$log_DepartureDelay <- log(air$Departure.Delay.in.Minutes + 1)#For Departure.Delay.in.Minutes
air$log_ArrivalDelay <- log(air$Arrival.Delay.in.Minutes + 1)#For Arrival.Delay.in.Minutes
air$log_FlightDistance <- log(air$Flight.Distance + 1)#For Flight.Distance
```

Data cleaning & processing

3. Transform Y variable and other variables (Which have character) as factor

```
air <- air %>% mutate(satisfaction = ifelse(satisfaction == "dissatisfied",0,1)) #change Y variable as factor
air <- air %>% mutate_if(is.character,as.factor) #change variable that are characters as factor
```

4. Split into Training and Testing Data

```
#Split dataset into training and testing
smp_size <- floor(0.8 * nrow(air))
set.seed(123)
train_ind <- sample(seq_len(nrow(air)), size = smp_size)
train <- air[train_ind, ]
test <- air[-train_ind, ]
```

04

Model & Evaluation and Results

A. Logistic Regression Model

```
Call:
glm(formula = satisfaction ~ ., family = "binomial", data = train)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.3863734  0.1262155 -42.676  < 2e-16 ***
GenderMale                      -0.9664452  0.0185703 -52.042  < 2e-16 ***
Customer.TypeLoyal Customer      1.9611116  0.0278844  70.330  < 2e-16 ***
Age                             -0.0087951  0.0006442 -13.653  < 2e-16 ***
Type.of.TravelPersonal Travel   -0.7866657  0.0261777 -30.051  < 2e-16 ***
ClassEco                        -0.7060701  0.0235694 -29.957  < 2e-16 ***
ClassEco Plus                   -0.7686493  0.0363547 -21.143  < 2e-16 ***
Seat.comfort                     0.2860538  0.0103272  27.699  < 2e-16 ***
Departure.Arrival.time.convenient -0.1956350  0.0075739 -25.830  < 2e-16 ***
Food.and.drink                  -0.2216542  0.0105095 -21.091  < 2e-16 ***
Gate.location                    0.1152660  0.0085319  13.510  < 2e-16 ***
Inflight.wifi.service           -0.0727751  0.0099085  -7.345 2.06e-13 ***
Inflight.entertainment           0.6739900  0.0092741  72.674  < 2e-16 ***
Online.support                   0.0931419  0.0100955   9.226  < 2e-16 ***
Ease.of.Online.booking           0.2202168  0.0129994  16.940  < 2e-16 ***
On.board.service                 0.3057681  0.0092268  33.139  < 2e-16 ***
Leg.room.service                 0.2243625  0.0078597  28.546  < 2e-16 ***
Baggage.handling                 0.0980526  0.0103869   9.440  < 2e-16 ***
Checkin.service                  0.2954457  0.0077737  38.006  < 2e-16 ***
Cleanliness                      0.0898061  0.0108154   8.304  < 2e-16 ***
Online.boarding                  0.1739493  0.0111492  15.602  < 2e-16 ***
log_DepartureDelay               0.0204325  0.0093849   2.177   0.0295 *
log_ArrivalDelay                -0.1790026  0.0093297 -19.186  < 2e-16 ***
log_FlightDistance              -0.1900500  0.0135909 -13.984  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 142659  on 103588  degrees of freedom
Residual deviance:  79639  on 103565  degrees of freedom
AIC: 79687
```

At 95% confidence level,
Significant variables: all variables are significant

A. Logistic Regression Model

Coefficient	Estimate	Pr(>|t|)
Loyal Customer	1.9612	<2e-16
In-flight Entertainment	0.674	2.06e-13
On-Board Service	0.3057	<2e-16

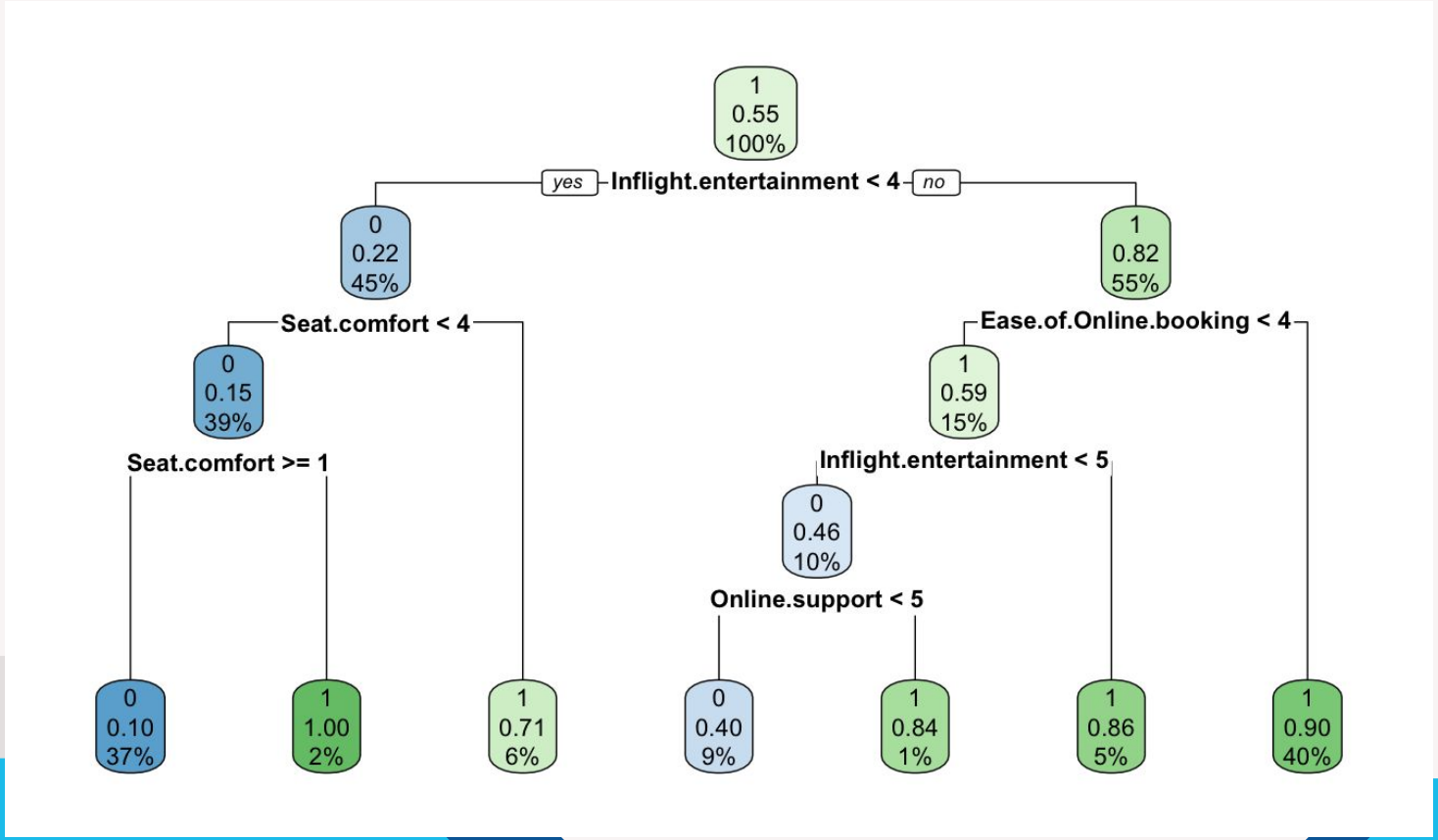
- Being a loyal customer, the **log odds** of being satisfied **increase** by **1.9612**, while holding other variables constant.
- A unit increase in In-flight Entertainment **increases** the **log odds** of of being satisfied by **0.674**, while holding other variables constant.
- A unit increase in On-Board Service **increases** the **log odds** of of being satisfied by **0.3057**, while holding other variables constant.

A. Logistic Regression Model

Coefficient	Estimate	e^b-1	%	Pr(>|t|)
Loyal Customer	1.9612	6.108	+610.8%	<2e-16
In-flight Entertainment	0.674	0.962	+96.2%	2.06e-13
On-Board Service	0.3057	0.358	+35.8%	<2e-16

- Being a loyal customer, the **odds** of being satisfied **increase** by 610.8%, while holding other variables constant.
- A unit increase in In-flight Entertainment **increases** the **odds** of of being satisfied by **96.2%**, while holding other variables constant.
- A unit increase in On-Board Service **increases** the **odds** of of being satisfied by **35.8%**, while holding other variables constant.

B. Pruned Decision Tree Model



B. Pruned Decision Tree Model

2 Different Node path:

If Inflight.entertainment not <4

AND Ease.of.Online.booking not < 4

➔ Number of Observations =40% of total

Predicted: dissatisfied = 0.10

Predicted: satisfied = 0.90

The customer will be satisfied with their flight experience..

If Inflight.entertainment <4

AND Seat.comfort < 4

AND Seat.comfort >= 1

➔ Number of Observations =37% of total

Predicted: dissatisfied = 0.90

Predicted: satisfied = 0.10

The customer will be dissatisfied with their flight experience..

C. 4 Model Comparison by Accuracy

Model	Accuracy
Logistic Regression	0.8388
Pruned Decision Tree	0.8651
Bagging	0.9546
Random Forest	0.9577

D. Model's Feature Importance (Top 3)

Logistic regression(0.8388)

```
> print(top3_lr)
                                                Variable Odds_Ratio       P_Value
Customer.TypeLoyal Customer Customer.TypeLoyal Customer   7.107223  0.000000e+00
Inflight.entertainment           Inflight.entertainment   1.962050  0.000000e+00
On.board.service                       On.board.service   1.357667 8.104239e-241
```

Pruned Decision Tree(0.8651)

```
> print(top3_dt)
Inflight.entertainment          Seat.comfort        Online.support
             20738.298              12626.088               9315.325
```

Random forest(0.9577)

```
> print(top_rf_variables)
                                          Variable Importance
Inflight.entertainment Inflight.entertainment  10244.283
Seat.comfort                     Seat.comfort   6730.287
Ease.of.Online.booking Ease.of.Online.booking   3949.564
```

05

Conclusion and Implications

Conclusion

ALL THREE MODEL INCLUDES Inflight Entertainment

TWO MODELS INCLUDES Seat Comfort

Random Forest Includes Ease of Online Booking
(Best Accuracy Model)

- Therefore, we will provide recommendation for Airline based on these three variables that are influential to customer satisfaction

Recommendation 1 (Enhance Inflight Entertainment)

The model shows that passenger pleasure is greatly impacted by the **quality** and **accessibility** of in-flight entertainment. Improving entertainment alternatives can result in better experiences for passengers. For example:

- Movies
- TV-series
- Games
- Wi-Fi connections (Social media access)
- Live TV

Recommendation 2 (Improve Seat Comfort)

Another important variable affecting satisfaction is **seat comfort**. Comfortable seating is important to passengers since it might affect how they travel overall. For example:

- More ergonomic seating
- Lumbar support
- Adopt high quality fabric seats
- Provide blankets and pillows

Recommendation 3 (Efficient Online Booking)

The third important variable affecting satisfaction is how **easy** it is for passengers to book flights online. For example:

- Support mobile and web based system
- Clean and clear navigation
- Real-time updates
- Support multiple payment methods
- Fast response time

Thank you