BUDT 758T: Data Mining and Predictive Analytics

Group Project

Team 15

Project Title:

Prediction of Airline Satisfaction Rates With Diverse Classification Algorithms in R

Team Members:

Gauri Goel, Kai-Wen Chen, Rohin Bhagvatula, Tanya Singh, Xin Lan

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	Signature
Contact Author	GAURI GOEL	Gauri Goel
	KAI-WEN CHEN	Kai-Wen Chen
	ROHIN BHAGAVATULA	Rohin Bhagavatula
	TANYA SINGH	Tanya Singh
	XIN LAN	Xin Lan

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY 3

II. DATA DESCRIPTION 3

III. RESEARCH QUESTIONS 5

IV. METHODOLOGY 10

A. LOGISTIC REGRESSION 10

B. KNN 12

C. NAIVE BAYES 12

D. CLASSIFICATION TREE 13

E. BOOSTING 15

F. XGBOOST 16

V. RESULTS AND FINDINGS 16

VI. CONCLUSION 18

I. Executive Summary

Various airlines conduct flight satisfaction surveys with a list of service items to derive which content passengers would appreciate in their traveling experience. However, the raw numerical outcome could be thoroughly sophisticated or ambiguous to extract the authentic essence of the feedback. Therefore, in this project, we aim to diagnose which service factors are among the most that affect clients' reflection by diverse predictive models and provide practical suggestions on methods to improve the quality of service.

II. Data Description

The data is available at Kaggle: https://www.kaggle.com/datasets/johndddddd/customer-satisfaction

The original airline satisfaction survey dataset contains 129,880 flight records with 24 variables with the following information of each:

Dependent Variable:

Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction) → Categorical

Independent variables:

Age: The actual age of the passengers → Numerical

Gender: Gender of the passengers (Female, Male) → Categorical

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel) → Categorical

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus) → Categorical

Customer Type: The customer type (Loyal customer, disloyal customer) → Categorical

Flight distance: The flight distance of this journey → Numerical

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5) → Numerical

Ease of Online booking: Satisfaction level of online booking (0:Not Applicable;1-5) → Numerical

Inflight service: Satisfaction level of inflight service (0:Not Applicable;1-5) → Numerical

Online boarding: Satisfaction level of online boarding (0:Not Applicable;1-5) → Numerical

Inflight entertainment: Satisfaction level of inflight entertainment (0:Not Applicable;1-5) → Numerical

Food and drink: Satisfaction level of Food and drink (0:Not Applicable;1-5) → Numerical

Seat comfort: Satisfaction level of Seat comfort (0:Not Applicable;1-5) → Numerical

On-board service: Satisfaction level of On-board service (0:Not Applicable;1-5) → Numerical

Leg room service: Satisfaction level of Leg room service (0:Not Applicable;1-5) → Numerical

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient (0:Not Applicable;1-5) → Numerical

Baggage handling: Satisfaction level of baggage handling (0:Not Applicable;1-5) → Numerical

Gate location: Satisfaction level of Gate location (0:Not Applicable;1-5) → Numerical

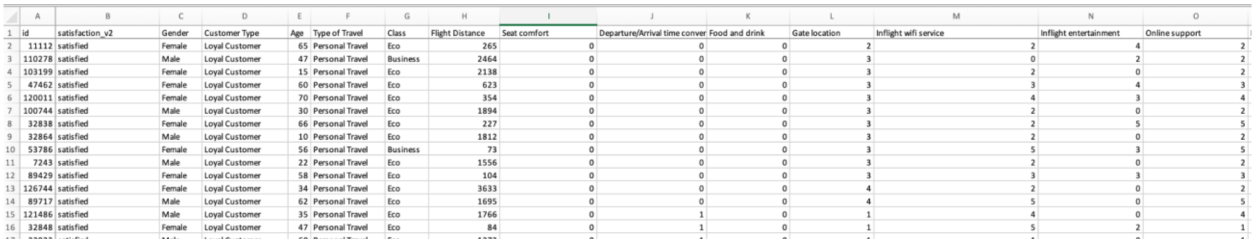
Cleanliness: Satisfaction level of Cleanliness (0:Not Applicable;1-5) → Numerical

Check-in service: Satisfaction level of Check-in service (0:Not Applicable;1-5) → Numerical

Departure Delay in Minutes: Minutes delayed when departure (0:Not Applicable;1-5) → Numerical

Arrival Delay in Minutes: Minutes delayed when Arrival (0:Not Applicable;1-5) → Numerical

Sample of Observations



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	id	satisfaction_v2	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time conver	Food and drink	Gate location	Inflight wifi service	Inflight entertainment	Online support
2	11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco	265	0	0	0	2	2	4	2
3	110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business	2464	0	0	0	3	0	2	2
4	103199	satisfied	Female	Loyal Customer	15	Personal Travel	Eco	2138	0	0	0	3	2	0	2
5	47462	satisfied	Female	Loyal Customer	60	Personal Travel	Eco	623	0	0	0	3	3	4	3
6	120011	satisfied	Female	Loyal Customer	70	Personal Travel	Eco	354	0	0	0	3	4	3	4
7	100744	satisfied	Male	Loyal Customer	30	Personal Travel	Eco	1894	0	0	0	3	2	0	2
8	32838	satisfied	Female	Loyal Customer	66	Personal Travel	Eco	227	0	0	0	3	2	5	5
9	32864	satisfied	Male	Loyal Customer	10	Personal Travel	Eco	1812	0	0	0	3	2	0	2
10	53786	satisfied	Female	Loyal Customer	56	Personal Travel	Business	73	0	0	0	3	5	3	5
11	7243	satisfied	Male	Loyal Customer	22	Personal Travel	Eco	1556	0	0	0	3	2	0	2
12	89429	satisfied	Female	Loyal Customer	58	Personal Travel	Eco	104	0	0	0	3	3	3	3
13	126744	satisfied	Female	Loyal Customer	34	Personal Travel	Eco	3633	0	0	0	4	2	0	2
14	89717	satisfied	Male	Loyal Customer	62	Personal Travel	Eco	1695	0	0	0	4	5	0	5
15	121486	satisfied	Male	Loyal Customer	35	Personal Travel	Eco	1766	0	1	0	1	4	0	4
16	32848	satisfied	Female	Loyal Customer	47	Personal Travel	Eco	84	0	1	0	1	5	2	1

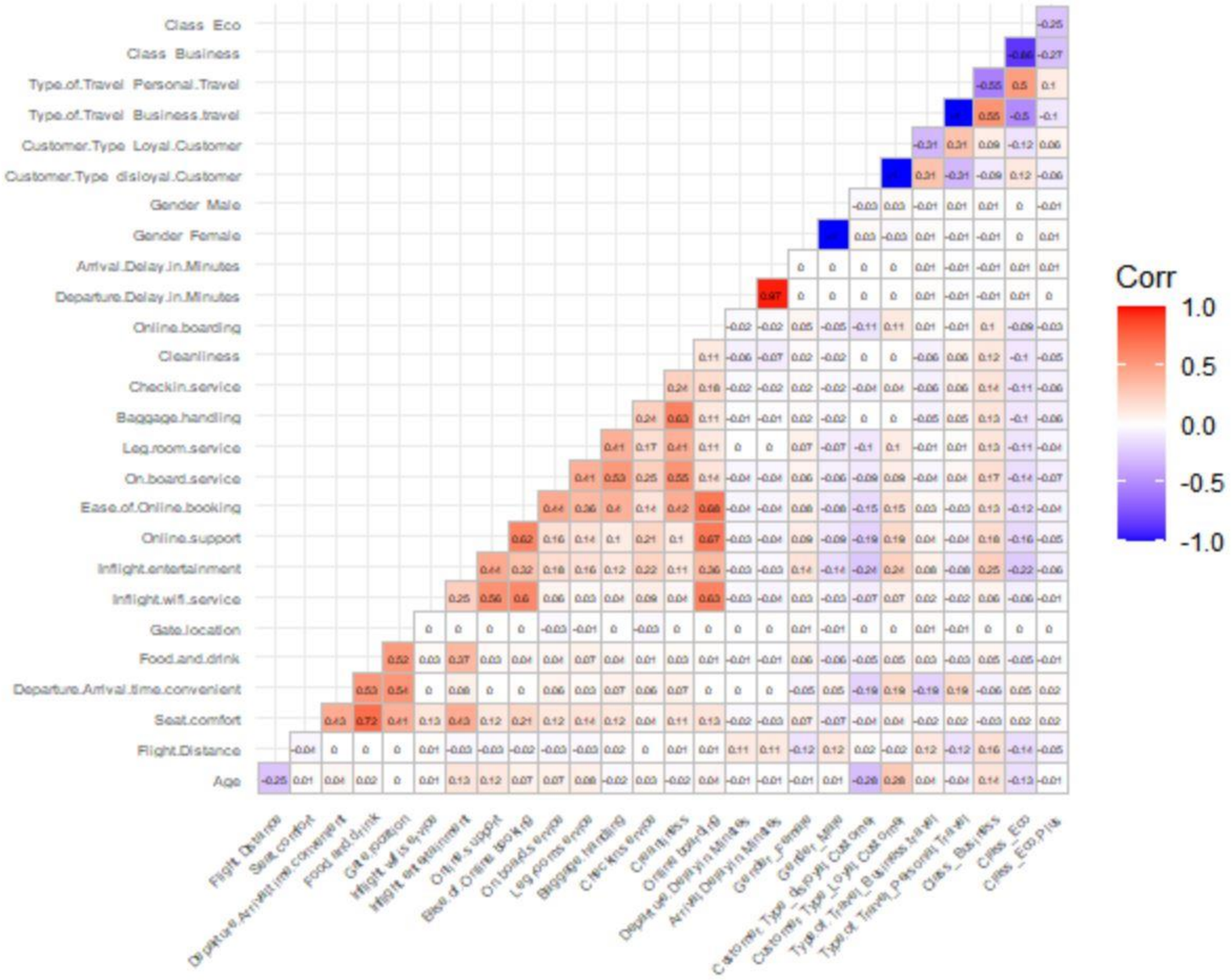
III. Research Questions

1. Which variables are more correlated with the customer satisfaction rating and impact it significantly?

   Using the abs(correlation coefficient)>= 0.2, the following variables are most correlated with RATE variable:-

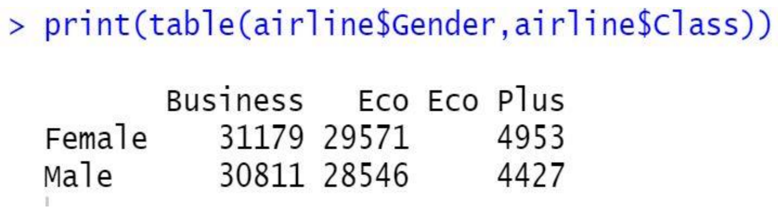
   Onlineboarding,Cleanliness,Checkin service,Baggage handling,Leg room service,On board service,Ease of Online booking,Online support,Inflight entertainment,Inflight wifi service,Seat comfort,Class_Eco,Class_Business,Customer.Type_Loyal.Customer,Gender_Male

   The heatmap is:-



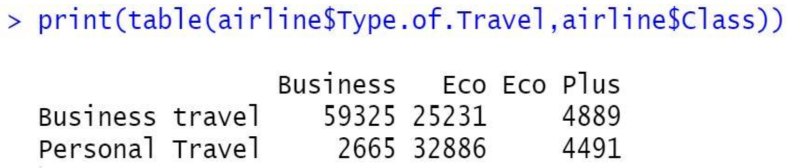
2. What is the preference for travel class between different genders and the difference between business travel and personal travel in choosing class?

It seems that females and males are equally likely to fly Business/Eco/Eco-Plus class :-



```
> print(table(airline$Gender,airline$Class))

         Business   Eco Eco Plus
  Female    31179 29571     4953
  Male      30811 28546     4427
```

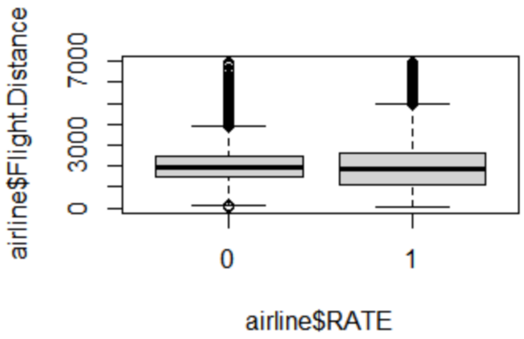
The travel purpose plays a major role in choosing the class of airplane seats. Trips taken for Business purpose are more likely to be of Business Class whereas trips taken for Personal purpose are more likely to be of Eco Class :-



```
> print(table(airline$Type.of.Travel,airline$Class))

                   Business   Eco Eco Plus
  Business travel     59325 25231     4889
  Personal Travel      2665 32886     4491
```

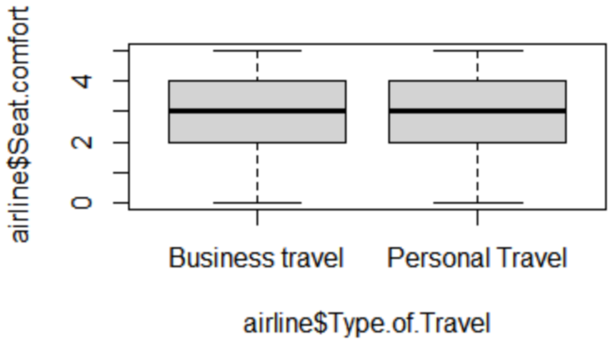
3. With the increase in the flight distance, is an impact seen in the ratings of an airline? Or do the ratings remain the same for an airline no matter the distance they fly?

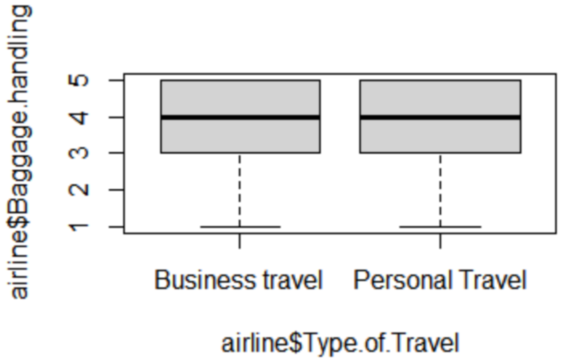
From the correlation heat-map, we have found out that the correlation between RATE and Flight Distance is -0.04, which is close to 0. A correlation close to 0 means there is hardly any linear correlation between the two variables. Also, from the boxplot below, we can hardly find any difference between the satisfaction levels of satisfied passengers vs unsatisfied passengers.

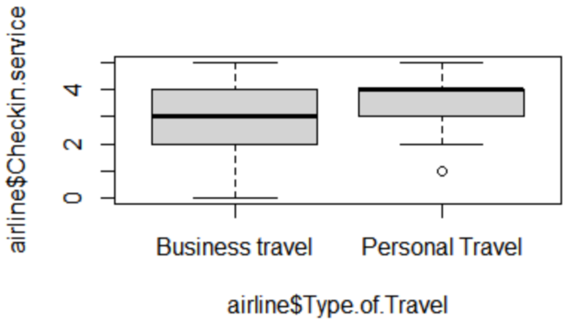


4. Given the type of travel, how do passengers' satisfaction levels change? Do personal travelers rate baggage handling services or check-in services better than the people traveling for business purposes?

   From the three box-plot above, Business-purpose and personal-purpose travelers are equally likely to be satisfied with both seat comfort level and baggage handling services. However, in terms of check-in services, personal travelers are more likely to be satisfied. This makes sense since business-purpose travelers are more concerned with time, and they always want a quick check-in. There is no wonder why they would rate this aspect higher than personal travelers.





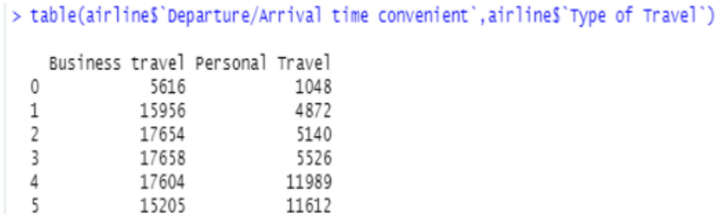


5. Do In-Flight meal and entertainment services impact customer satisfaction? If yes then to what extent?

   From the correlation coefficient heat map, in-flight entertainment is correlated with RATE with a score of 0.52, and Food & drink is correlated with RATE with a score of 0.12. Also, In-flight entertainment and food & drinks have a correlation coefficient of 0.37. We also ran a logistic regression and the coefficients for in-flight entertainment service and food & drink are 0.68 and -0.21 respectively. This means that the in-flight entertainment satisfaction levels are positively correlated , while Food & Drinks' are negatively correlated with RATE, which as per the common sense, is unlikely. This might be due to the fact that these two variables are weakly correlated as per general standards of correlation coefficient (>=0.5), but are relatively significant given our cut-off>=0.2.

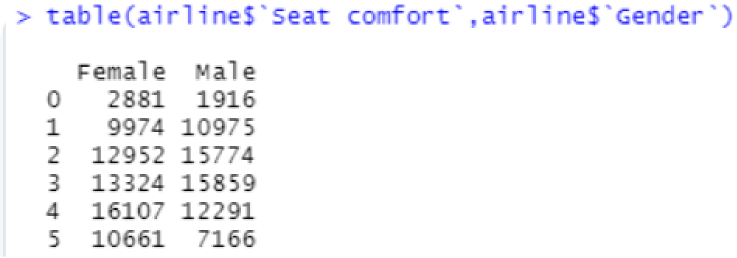
6. How much effort should we put into different customer types based on gender and customer type to optimize the satisfaction rate?

   Firstly, we need to take a look at the way different customers rate different aspects of a flight. Business-purpose travelers care more about whether the arrival and departure time is accurate than personal-purpose travelers.

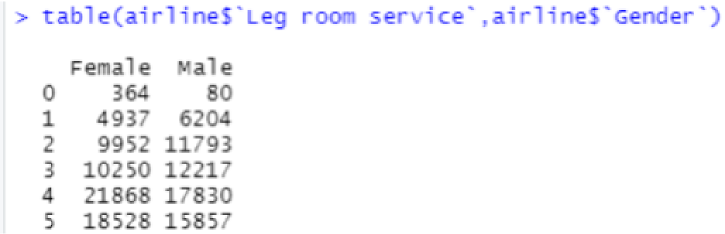


```
> table(airline$`Departure/Arrival time convenient`,airline$`Type of Travel`)

    Business travel Personal Travel
  0            5616            1048
  1           15956            4872
  2           17654            5140
  3           17658            5526
  4           17604           11989
  5           15205           11612
```

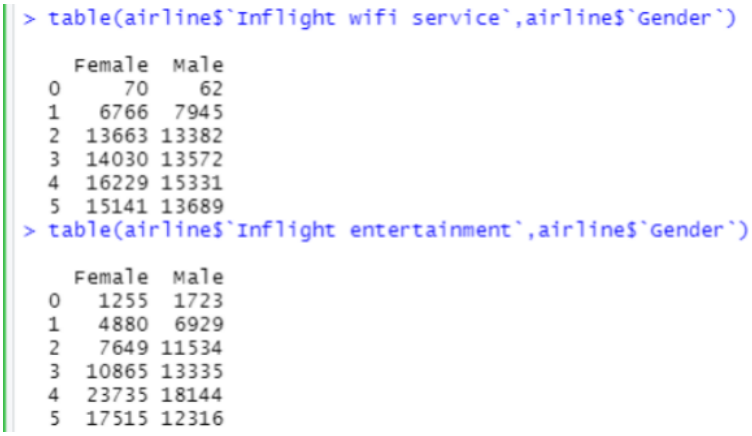
As for different genders, it looks like males are more likely to rate the seat comfort and leg room service lower than how females do. Also, males are less likely to be satisfied with in-flight Wi-Fi/entertainment and food&drinks than females. We can probably tailor our seat assignment to give better leg room to male passengers . However, it can become a gender-bias practice, airlines might probably avoid doing so. Maybe, the airlines can consider passengers' height to solve the issue of leg-room service satisfaction.



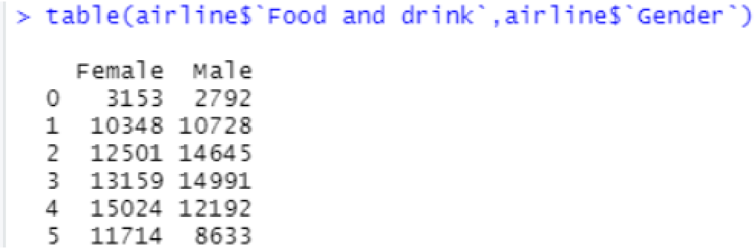
```
> table(airline$`Seat comfort`,airline$`Gender`)

    Female  Male
  0   2881  1916
  1   9974 10975
  2  12952 15774
  3  13324 15859
  4  16107 12291
  5  10661  7166
```



```
> table(airline$`Leg room service`,airline$`Gender`)

    Female  Male
  0    364    80
  1   4937  6204
  2   9952 11793
  3  10250 12217
  4  21868 17830
  5  18528 15857
```



```
> table(airline$`Inflight wifi service`,airline$`Gender`)

    Female  Male
  0     70    62
  1   6766  7945
  2  13663 13382
  3  14030 13572
  4  16229 15331
  5  15141 13689
> table(airline$`Inflight entertainment`,airline$`Gender`)

    Female  Male
  0   1255  1723
  1   4880  6929
  2   7649 11534
  3  10865 13335
  4  23735 18144
  5  17515 12316
```



```
> table(airline$`Food and drink`,airline$`Gender`)

    Female  Male
  0   3153  2792
  1  10348 10728
  2  12501 14645
  3  13159 14991
  4  15024 12192
  5  11714  8633
```

7. Does inflight wifi availability compensate for an absence of poor inflight entertainment or vice-e-versa?

The correlation between inflight wifi service and inflight entertainment service is 0.25, meaning they are positively correlated. However, this cannot answer the question above. In order to find whether inflight wifi availability compensates for an absence of poor inflight entertainment or vice-e-versa, we need to run a controlled random experiment, which is out of the scope of this project.

IV. Methodology

- Removed the first column, named #id, as it was irrelevant
- Dropped the 393 missing values
- Renamed variables in order to remove whitespaces in between
- Created dummies for categorical variables- 'Gender','Customer Type','Type of Travel' and 'Class' using fastDummies library
- Renamed variables to remove whitespaces in between the newly created dummy variables
- Created RATE variable and coded it as 0 and 1 on the basis of the independent categorical variable “satisfaction_v2” using the following code:-

```
airline$RATE <- as.numeric(ifelse(airline$satisfaction_v2=='satisfied',1,0))
```

- We then converted RATE variable using factor() with labels 0 and 1
- We have used the seed of 1 to maintain the consistency across the results of our models
- Splitted the data into training and testing dataset in 70-30 ratio
- Applied R built-in functions and methods specific to each model in order to calculate respective confusion matrix, accuracy, error rate, sensitivity, and specificity. We built the following classification models to predict an airline’s customer satisfaction rate:-

1. Logistic Regression:

Firstly, we conducted feature selection by calculating accuracies for 3 different sets of variables on the basis of the correlation heatmap.

Secondly, we applied the glm() function with the family attribute set to binomial to calculate 3 logistic regression equations for the RATE variable using the below 3 set of variables.

The 3 models were built on the criteria of:-

- abs(correlation coefficient) >= 0.30

  ```
  model_0.3<- glm(RATE ~ Online.boarding+Leg.room.service+On.board.service+Ease.of.Online.booking+Online.support+Inflight.entertainment+Class_Business, data=traindata, family="binomial")
  ```

- abs(correlation coefficient) >= 0.25

  ```
  model_0.25<- glm(RATE ~ Online.boarding+Cleanliness+Checkin.service+Baggage.handling+Leg.room.service+On.board.service+Ease.of.Online.booking+Online.support+Inflight.entertainment+Class_Eco+Class_Business+Customer.Type_Loyal.Customer, data=traindata, family="binomial")
  ```

- abs(correlation coefficient) >= 0.20

  ```
  model_0.2<- glm(RATE ~ Online.boarding+Cleanliness+Checkin.service+Baggage.handling+Leg.room.service+On.board.service+Ease.of.Online.booking+Online.support+Inflight.entertainment+Inflight.wifi.service+Seat.comfort+Class_Eco+Class_Business+Customer.Type_Loyal.Customer+Gender_Male, data=traindata, family="binomial")
  ```

Thirdly, we predicted the values using the testing dataset and calculated accuracies for each of the 3 equations, using:-

```
predicted.probability <- predict(model_0.XX, type = "response", newdata=testdata[,-1])
```

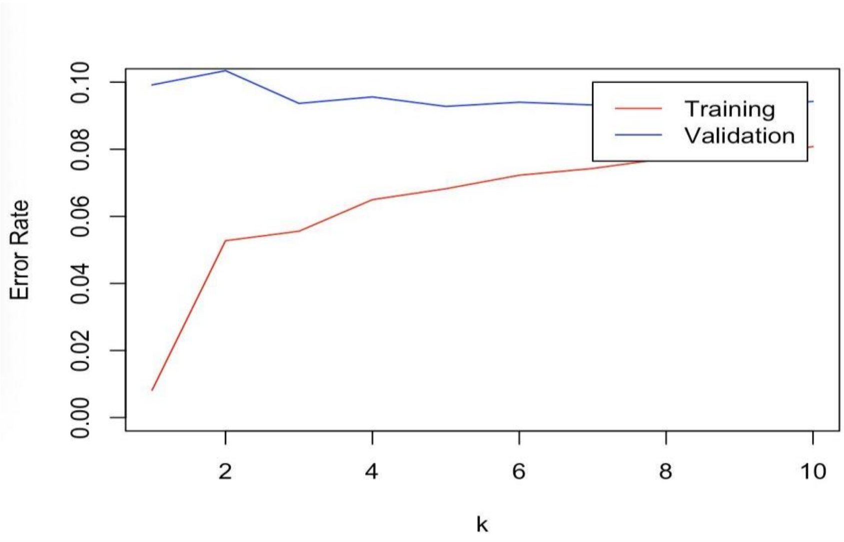
We went ahead with the model with abs(correlation coefficient)>=0.2 as the accuracy for that model was at the highest with 83%, on the test data.

2. KNN:

Before we ran the KNN model we standardized the variables with abs(correlation coefficient)>=0.2, to make sure they are all on the same scale. Then we ran a loop for 10 different k (k=1:10). We found out at k=5, the error rate is the lowest with the value of 0.0928 and the accuracy is 0.907.

```
for (i in 1:kmax){

  prediction <- knn(train_input, train_input,train_output, k=i)

  prediction2 <- knn(train_input, test_input,train_output, k=i)
```

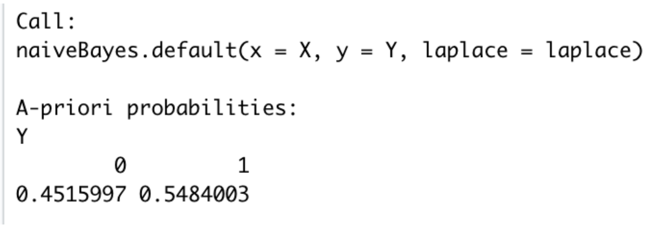
The error rate for the best model is: -



3. Naive Bayes:

We built the Naive Bayes model with the e1071 package.

```
model <- naiveBayes(RATE~., data=traindata)
```



```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.4515997 0.5484003 
```

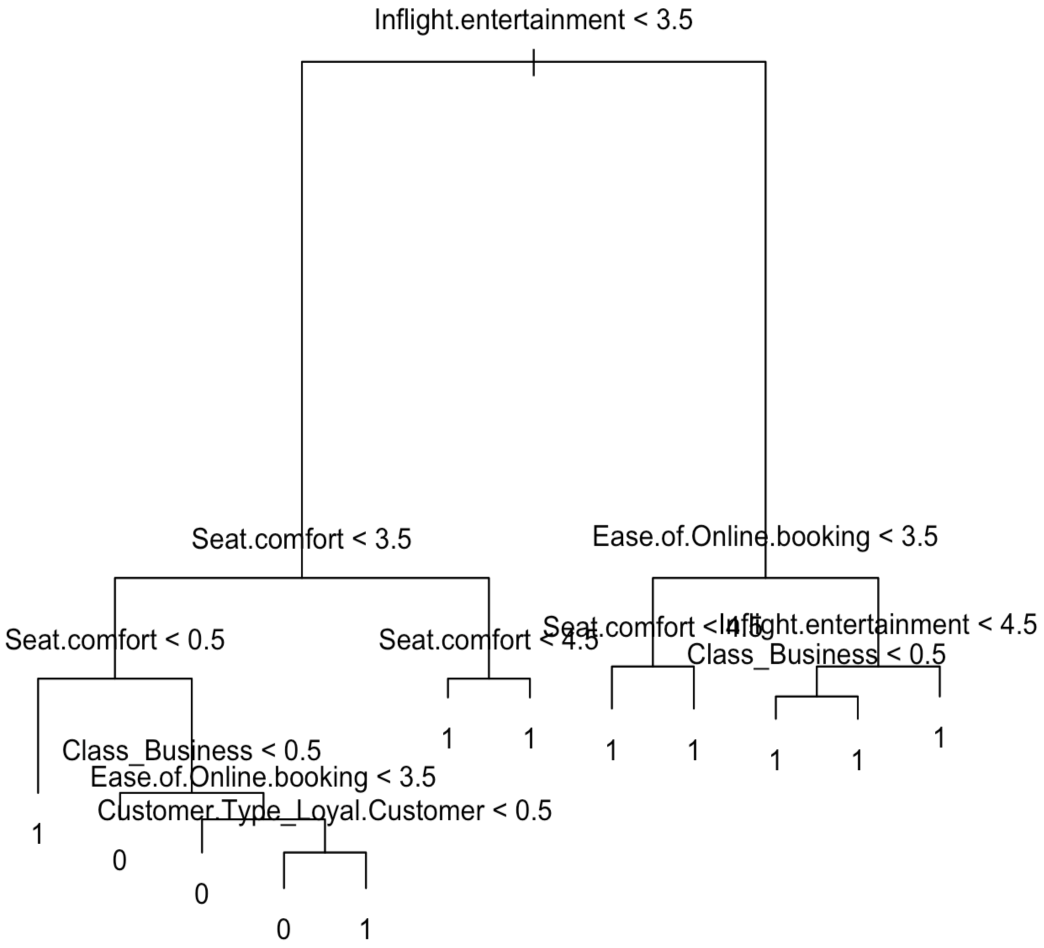
The apriori probabilities output demonstrates that there is a chance of 0.45 that the flight reflection would be dissatisfied and a chance of 0.55 that it would be rated as satisfied. The prediction offers an 0.82 accuracy rate on the test data.

```
prediction <- predict(model, newdata = testdata[,-1])
```

4. Classification Tree:

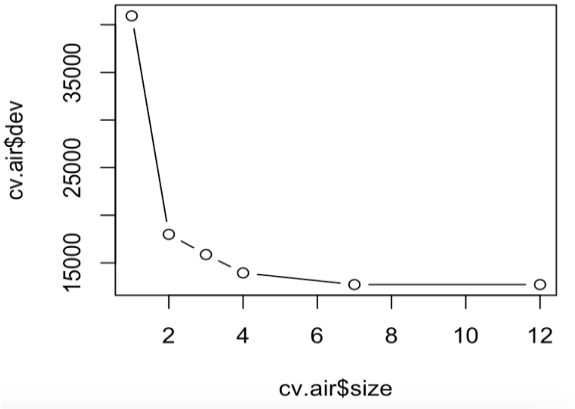
We applied the tree() library to build our initial classification model and plotted the unpruned tree.

```
tree.air = tree(RATE~., data = traindata2)
```



Inflight.entertainment < 3.5
Seat.comfort < 3.5
Ease.of.Online.booking < 3.5
Seat.comfort < 0.5
Seat.comfort < 4.5
Seat.comfort < 4.5
Inflight.entertainment < 4.5
Class_Business < 0.5
Class_Business < 0.5
Ease.of.Online.booking < 3.5
Customer.Type_Loyal.Customer < 0.5
1
0
0
0
1
1
1
1
1
1
1
1

Next, we used cross-validation to determine the best node size and received the minimum standard deviation at node size=7.

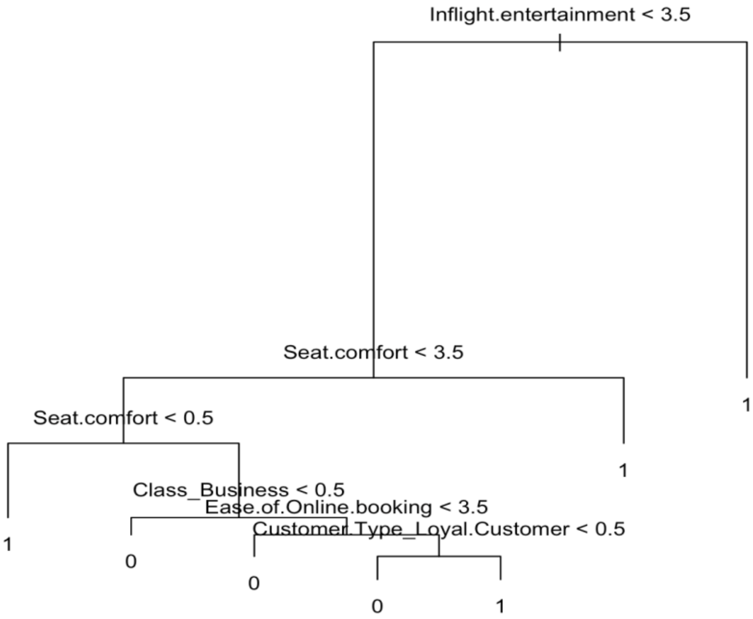


Next, we pruned the tree to the best size of 7 nodes.

```
prune.air=prune.misclass(tree.air,best=7)
```

```
plot(prune.air)
```

```
text(prune.air,pretty=0)
```


Inflight.entertainment < 3.5
Seat.comfort < 3.5
Seat.comfort < 0.5
Class_Business < 0.5
Ease.of.Online.booking < 3.5
Customer Type_Loyal.Customer < 0.5
1
1
1
0
0
0
1

Finally, we predicted with the following command, securing a 0.86 accuracy rate on the test data.

```
prune.pred.air=predict(prune.air,testdata2,type="class")
```

5. Boosting:

For the boosting model, we executed the gbm library, setting n.tree = 1000 and distribution method to bernoulli.

```
boost.air = gbm(RATE~., data = traindata, distribution="bernoulli", n.trees=1000, interaction.depth=4)
```

After predicting with the test data, we took a cutoff of 0.5 to compute the confusion matrix. Eventually, we captured an accuracy rate of 0.93 from the test model.

```
prediction = ifelse(boost.predict>0.5, 1, 0)
```

6. XGBoost:

For our final model, we imported the xgboost package to build the xgboost model with the following order:

```
xgb.air = xgboost(data = xgb.traindata,  label = xgb.label, max.depth = 2, eta = 1, nround = 5, objective = "binary:logistic")
```

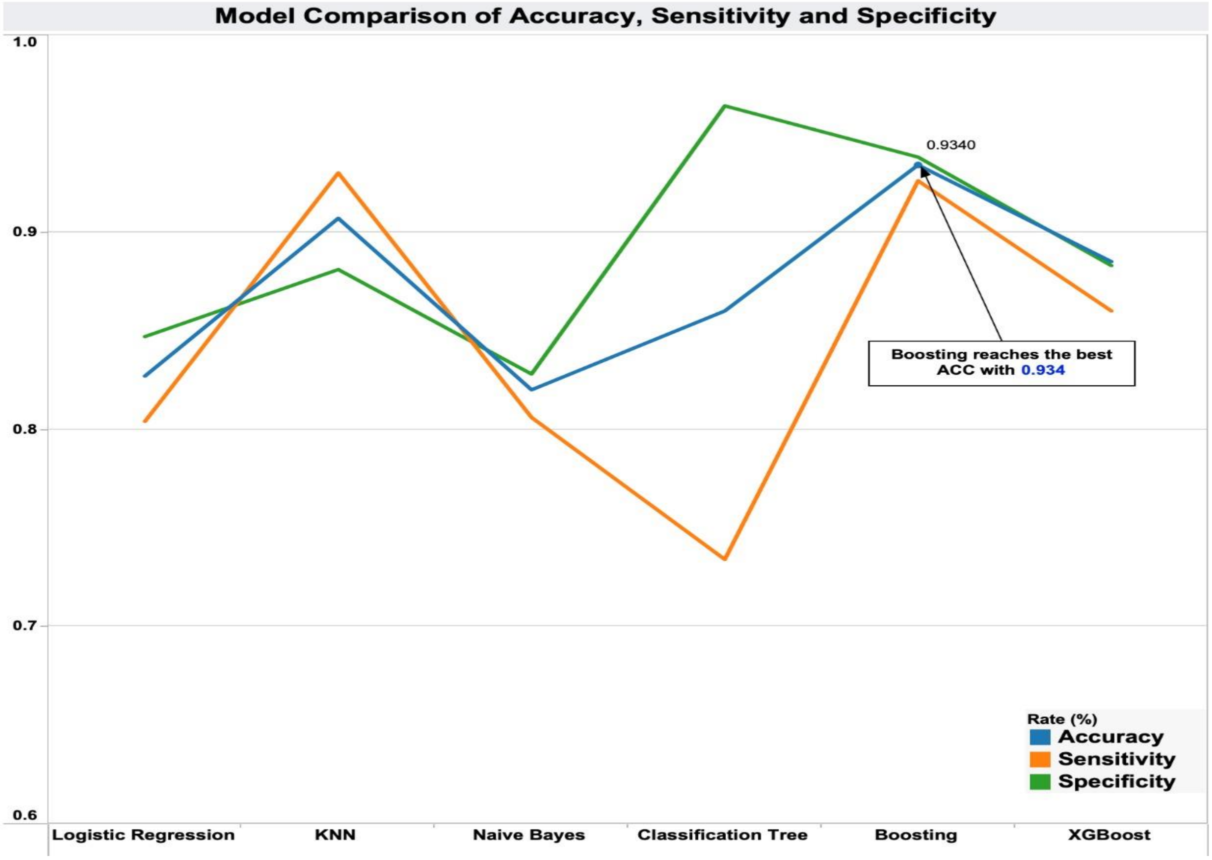
After executing the prediction, we acquired a 0.88 accuracy rate on the test dataset.

```
xgb.pred = predict(xgb.air, xgb.testdata)

prediction = ifelse(xgb.pred > 0.5, 1, 0)
```

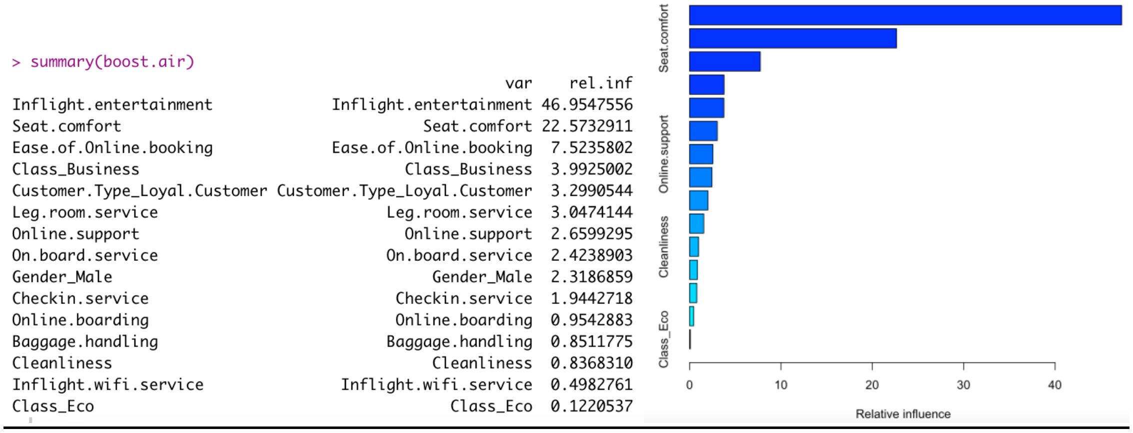
V. Results and Findings

Following is the table and the chart with a comparative analysis of Accuracy, Sensitivity and Specificity of all the classification models we have built in this project:-

Classification Model	Accuracy	Sensitivity	Specificity
Logistic Regression	0.827	0.804	0.847
KNN	0.907	0.930	0.881
Naive Bayes	0.82	0.806	0.828
Classification Tree	0.86	0.734	0.964
Boosting	0.934	0.926	0.938
XGBoost	0.885	0.86	0.883



The boosting model offers the greatest prediction accuracy rate with 0.934 among all the classification models that we performed. It also gives us the best sensitivity with 0.926 and the second best specificity of 0.938, after the classification tree.



```
> summary(boost.air)
                                                      var    rel.inf
Inflight.entertainment             Inflight.entertainment 46.9547556
Seat.comfort                                 Seat.comfort 22.5732911
Ease.of.Online.booking             Ease.of.Online.booking  7.5235802
Class_Business                             Class_Business  3.9925002
Customer.Type_Loyal.Customer Customer.Type_Loyal.Customer  3.2990544
Leg.room.service                         Leg.room.service  3.0474144
Online.support                             Online.support  2.6599295
On.board.service                         On.board.service  2.4238903
Gender_Male                                   Gender_Male  2.3186859
Checkin.service                           Checkin.service  1.9442718
Online.boarding                           Online.boarding  0.9542883
Baggage.handling                         Baggage.handling  0.8511775
Cleanliness                                   Cleanliness  0.8368310
Inflight.wifi.service               Inflight.wifi.service  0.4982761
Class_Eco                                       Class_Eco  0.1220537
```

Additionally, from the summary output of the boosting model, we can discover that Inflight entertainment and Seat comfort have the highest relative influence. We can conclude the two are among the most significant variables that would affect passengers' reflection on their flight.

VI. Conclusion

After running various classification models for predictions of customer satisfaction rates for airlines, we have found out that Boosting has the best performance. From our data analysis, we have found out that females and males do value aspects of flights differently and so do people with different flying purposes. Given our prediction accuracy of 0.93 with the Boosting model with the top 2 highest relative influence features, we can definitely tailor different aspects of various airline services on the basis of the nature of people who board their flights to get a better satisfaction level from their customers along with a major focus on Inflight entertainment and Seat comfort.