632 - Group Project R script

Van Phan, Hou U Kuong, Sabrina Liu

2024-04-03

```
pacman::p_load(ggplot2, tidyverse, gtsummary, dplyr, naniar, pROC, rpart, rpart.plot,
               randomForest, yardstick, tidymodels, xgboost, vip, openxlsx, partykit)
```

Part 1: Data and Data Description

```
# import datasets
test <- read.csv("air_test.csv", stringsAsFactors=TRUE)
train <- read.csv("air_train.csv", stringsAsFactors=TRUE)
```

```
# remove columns X and id for the data set since it is not related to our finding
dat <- train[,-1:-2]

# check all the variables structure
str(train)
```

```
## 'data.frame':    103904 obs. of  25 variables:
##  $ X                                : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ id                               : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65
##  $ Gender                           : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
##  $ Customer.Type                    : Factor w/ 2 levels "disloyal Customer",..: 2 1 2 2 2 2 2 2 2 1
##  $ Age                              : int  13 25 26 25 61 26 47 52 41 20 ...
##  $ Type.of.Travel                   : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 .
##  $ Class                            : Factor w/ 3 levels "Business","Eco",..: 3 1 1 1 1 2 2 1 1 2 ..
##  $ Flight.Distance                  : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
##  $ Inflight.wifi.service            : int  3 3 2 2 3 3 2 4 1 3 ...
##  $ Departure.Arrival.time.convenient: int  4 2 2 5 3 4 4 3 2 3 ...
##  $ Ease.of.Online.booking           : int  3 3 2 5 3 2 2 4 2 3 ...
##  $ Gate.location                    : int  1 3 2 5 3 1 3 4 2 4 ...
##  $ Food.and.drink                   : int  5 1 5 2 4 1 2 5 4 2 ...
##  $ Online.boarding                  : int  3 3 5 2 5 2 2 5 3 3 ...
##  $ Seat.comfort                     : int  5 1 5 2 5 1 2 5 3 3 ...
##  $ Inflight.entertainment           : int  5 1 5 2 3 1 2 5 1 2 ...
##  $ On.board.service                 : int  4 1 4 2 3 3 3 5 1 2 ...
##  $ Leg.room.service                 : int  3 5 3 5 4 4 3 5 2 3 ...
##  $ Baggage.handling                 : int  4 3 4 3 4 4 4 5 1 4 ...
##  $ Checkin.service                  : int  4 1 4 1 3 4 3 4 4 4 ...
##  $ Inflight.service                 : int  5 4 4 4 3 4 5 5 1 3 ...
##  $ Cleanliness                      : int  5 1 5 2 3 1 2 4 2 2 ...
##  $ Departure.Delay.in.Minutes       : int  25 1 0 11 0 0 9 4 0 0 ...
```

```
##  $ Arrival.Delay.in.Minutes          : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                      : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1 2 1 2 1 1
```

```r
dim(train)
```

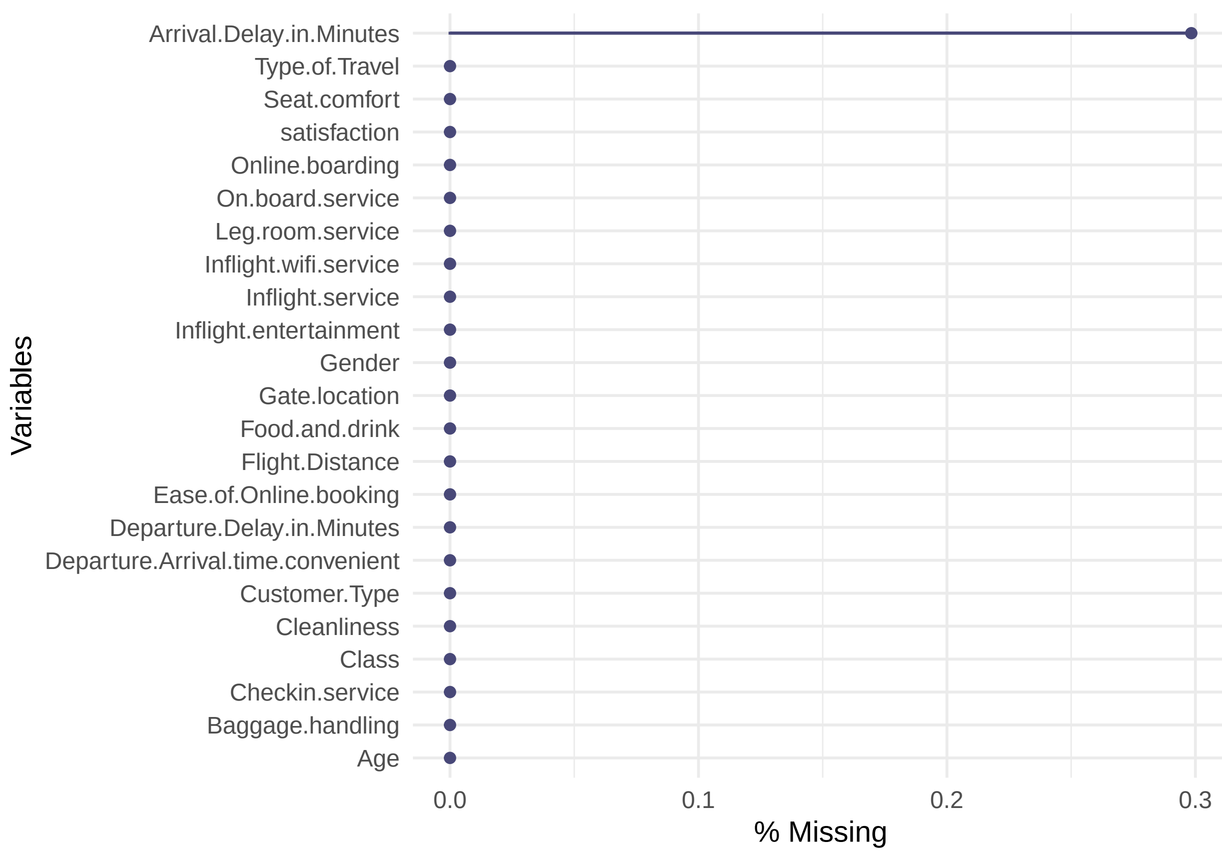
```
## [1] 103904     25
```

```r
# change binary variable satisfaction to 0 and 1, 1 is satisfied
dat$satisfaction <- as.factor(ifelse(dat$satisfaction == "satisfied", 1, 0))

# coerce from chr to factor variables
dat$Gender= as.factor(dat$Gender)
dat$Customer.Type= as.factor(dat$Customer.Type)
dat$Type.of.Travel= as.factor(dat$Type.of.Travel)
dat$Class= as.factor(dat$Class)
summary(dat)
```

```
##     Gender                Customer.Type        Age
##  Female:52727   disloyal Customer:18981   Min.   : 7.00
##  Male  :51177   Loyal Customer   :84923   1st Qu.:27.00
##                                           Median :40.00
##                                           Mean   :39.38
##                                           3rd Qu.:51.00
##                                           Max.   :85.00
##
##          Type.of.Travel       Class       Flight.Distance Inflight.wifi.service
##  Business travel:71655   Business:49665   Min.   :  31    Min.   :0.00
##  Personal Travel:32249   Eco     :46745   1st Qu.: 414    1st Qu.:2.00
##                          Eco Plus: 7494   Median : 843    Median :3.00
##                                           Mean   :1189    Mean   :2.73
##                                           3rd Qu.:1743    3rd Qu.:4.00
##                                           Max.   :4983    Max.   :5.00
##
##  Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location
##  Min.   :0.00                      Min.   :0.000          Min.   :0.000
##  1st Qu.:2.00                      1st Qu.:2.000          1st Qu.:2.000
##  Median :3.00                      Median :3.000          Median :3.000
##  Mean   :3.06                      Mean   :2.757          Mean   :2.977
##  3rd Qu.:4.00                      3rd Qu.:4.000          3rd Qu.:4.000
##  Max.   :5.00                      Max.   :5.000          Max.   :5.000
##
##  Food.and.drink  Online.boarding  Seat.comfort   Inflight.entertainment
##  Min.   :0.000   Min.   :0.00    Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :3.00    Median :4.000   Median :4.000
##  Mean   :3.202   Mean   :3.25    Mean   :3.439   Mean   :3.358
##  3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.00    Max.   :5.000   Max.   :5.000
##
##  On.board.service Leg.room.service Baggage.handling Checkin.service
##  Min.   :0.000    Min.   :0.000    Min.   :1.000    Min.   :0.000
##  1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000
##  Median :4.000    Median :4.000    Median :4.000    Median :3.000
```

```
##  Mean   :3.382    Mean   :3.351    Mean   :3.632    Mean   :3.304
##  3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.000
##  Max.   :5.000    Max.   :5.000    Max.   :5.000    Max.   :5.000
##
##  Inflight.service  Cleanliness    Departure.Delay.in.Minutes
##  Min.   :0.00     Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.00     1st Qu.:2.000   1st Qu.:   0.00
##  Median :4.00     Median :3.000   Median :   0.00
##  Mean   :3.64     Mean   :3.286   Mean   :  14.82
##  3rd Qu.:5.00     3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.00     Max.   :5.000   Max.   :1592.00
##
##  Arrival.Delay.in.Minutes satisfaction
##  Min.   :   0.00          0:58879
##  1st Qu.:   0.00          1:45025
##  Median :   0.00
##  Mean   :  15.18
##  3rd Qu.:  13.00
##  Max.   :1584.00
##  NA's   :310
```

```
# Missing information and visualize
gg_miss_var(dat, show_pct = TRUE)
```



```r
# Remove N/A of the Arrival.Delay.in.Minutes
dat = dat[!is.na(dat$Arrival.Delay.in.Minutes) ,]
```

```r
# Summary Table for Age, Departure.Delay, Arrival.Delay, Flight.Distance
dat %>%
  select(Age, Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, Flight.Distance) %>%
  tbl_summary(statistic = all_continuous() ~ "{mean} ({sd})",
              digits = all_continuous() ~ c(2,2))
```

Characteristic	**N = 103,594**
Age	39.38 (15.11)
Departure.Delay.in.Minutes	14.75 (38.12)
Arrival.Delay.in.Minutes	15.18 (38.70)
Flight.Distance	1,189.33 (997.30)

```r
# Summary Table for 14 variables related to airline services
dat %>%
  select(Inflight.wifi.service, Departure.Arrival.time.convenient, Ease.of.Online.booking,
         Gate.location,Food.and.drink, Online.boarding,Seat.comfort, Inflight.entertainment,
         On.board.service, Leg.room.service, Baggage.handling, Checkin.service,
         Inflight.service, Cleanliness) %>%
  tbl_summary(statistic = all_continuous() ~ "{mean} ({sd})",
              digits = all_continuous() ~ c(2,2))
```

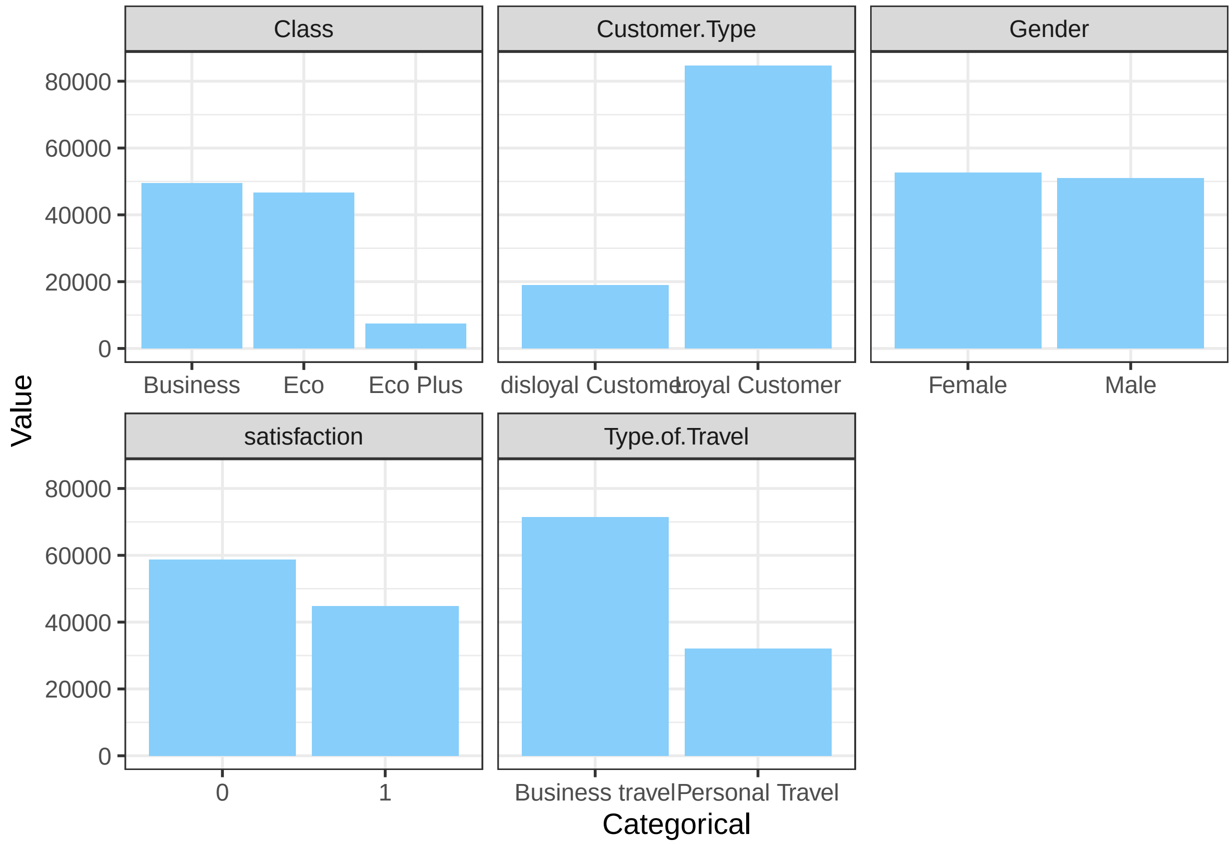
Characteristic	**N = 103,594**
Inflight.wifi.service	
0	3,096 (3.0%)
1	17,781 (17%)
2	25,755 (25%)
3	25,789 (25%)
4	19,737 (19%)
5	11,436 (11%)
Departure.Arrival.time.convenient	
0	5,290 (5.1%)
1	15,452 (15%)
2	17,142 (17%)
3	17,903 (17%)
4	25,474 (25%)
5	22,333 (22%)
Ease.of.Online.booking	
0	4,473 (4.3%)
1	17,466 (17%)
2	23,962 (23%)
3	24,370 (24%)
4	19,508 (19%)
5	13,815 (13%)
Gate.location	
0	1 (<0.1%)
1	17,511 (17%)

Characteristic	**N = 103,594**
2	19,396 (19%)
3	28,489 (28%)
4	24,353 (24%)
5	13,844 (13%)
Food.and.drink	
0	105 (0.1%)
1	12,800 (12%)
2	21,918 (21%)
3	22,238 (21%)
4	24,294 (23%)
5	22,239 (21%)
Online.boarding	
0	2,420 (2.3%)
1	10,658 (10%)
2	17,449 (17%)
3	21,744 (21%)
4	30,671 (30%)
5	20,652 (20%)
Seat.comfort	
0	1 (<0.1%)
1	12,031 (12%)
2	14,846 (14%)
3	18,641 (18%)
4	31,682 (31%)
5	26,393 (25%)
Inflight.entertainment	
0	14 (<0.1%)
1	12,441 (12%)
2	17,579 (17%)
3	19,080 (18%)
4	29,335 (28%)
5	25,145 (24%)
On.board.service	
0	3 (<0.1%)
1	11,832 (11%)
2	14,632 (14%)
3	22,770 (22%)
4	30,773 (30%)
5	23,584 (23%)
Leg.room.service	
0	470 (0.5%)
1	10,310 (10.0%)
2	19,469 (19%)
3	20,042 (19%)
4	28,704 (28%)
5	24,599 (24%)
Baggage.handling	
1	7,223 (7.0%)
2	11,483 (11%)
3	20,567 (20%)
4	37,274 (36%)
5	27,047 (26%)

Characteristic	**N = 103,594**
Checkin.service	
0	1 (<0.1%)
1	12,852 (12%)
2	12,854 (12%)
3	28,356 (27%)
4	28,975 (28%)
5	20,556 (20%)
Inflight.service	
0	3 (<0.1%)
1	7,063 (6.8%)
2	11,414 (11%)
3	20,227 (20%)
4	37,846 (37%)
5	27,041 (26%)
Cleanliness	
0	12 (<0.1%)
1	13,276 (13%)
2	16,081 (16%)
3	24,506 (24%)
4	27,100 (26%)
5	22,619 (22%)

```r
# Visualize for quantitative variables
ggplot(gather(dat %>% select_if(is.numeric)), aes(value)) +
  geom_histogram(fill = "4E84C4") +
  facet_wrap(~key, scales = 'free_x') +
  guides(x= guide_axis(angle=20)) +
  theme(text = element_text(size = 10),
        axis.text.x = element_text(lineheight=0.75)) +
  theme_bw()
```


```r
# Visualize for categorical variables
ggplot(gather(dat %>% select_if(is.factor)), aes(value)) +
  geom_bar(bins = 10, fill = "lightskyblue") +
  facet_wrap(~key, scales = "free_x") + labs(x = "Categorical", y = "Value") + theme_bw()
```




```
# Summary Table for categorical variables
dat %>%
  select(Class, Customer.Type, Gender, satisfaction, Type.of.Travel) %>%
  tbl_summary(statistic = all_continuous() ~ "{mean} ({sd})",
              digits = all_continuous() ~ c(0,0))
```

Characteristic	**N = 103,594**
Class	
Business	49,533 (48%)
Eco	46,593 (45%)
Eco Plus	7,468 (7.2%)
Customer.Type	
disloyal Customer	18,932 (18%)
Loyal Customer	84,662 (82%)
Gender	
Female	52,576 (51%)
Male	51,018 (49%)
satisfaction	
0	58,697 (57%)
1	44,897 (43%)
Type.of.Travel	
Business travel	71,465 (69%)
Personal Travel	32,129 (31%)

Part 2: Data Modeling

```r
# fit the multiple logistic model
mod <- glm(satisfaction ~ ., data = dat, family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = dat)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.860e+00  7.876e-02 -99.793  < 2e-16 ***
## GenderMale                         4.255e-02  1.949e-02   2.183  0.02905 *
## Customer.TypeLoyal Customer        2.035e+00  2.994e-02  67.970  < 2e-16 ***
## Age                               -8.308e-03  7.110e-04 -11.684  < 2e-16 ***
## Type.of.TravelPersonal Travel     -2.722e+00  3.147e-02 -86.494  < 2e-16 ***
## ClassEco                          -7.389e-01  2.566e-02 -28.794  < 2e-16 ***
## ClassEco Plus                     -8.554e-01  4.155e-02 -20.588  < 2e-16 ***
## Flight.Distance                   -1.789e-05  1.132e-05  -1.581  0.11392
## Inflight.wifi.service              3.949e-01  1.148e-02  34.405  < 2e-16 ***
## Departure.Arrival.time.convenient -1.244e-01  8.218e-03 -15.132  < 2e-16 ***
## Ease.of.Online.booking            -1.440e-01  1.135e-02 -12.691  < 2e-16 ***
## Gate.location                      2.914e-02  9.174e-03   3.176  0.00149 **
## Food.and.drink                    -2.860e-02  1.068e-02  -2.677  0.00743 **
## Online.boarding                    6.126e-01  1.025e-02  59.773  < 2e-16 ***
## Seat.comfort                       6.555e-02  1.118e-02   5.862 4.58e-09 ***
## Inflight.entertainment             6.555e-02  1.427e-02   4.594 4.34e-06 ***
## On.board.service                   3.014e-01  1.019e-02  29.582  < 2e-16 ***
## Leg.room.service                   2.532e-01  8.540e-03  29.652  < 2e-16 ***
## Baggage.handling                   1.331e-01  1.144e-02  11.633  < 2e-16 ***
## Checkin.service                    3.234e-01  8.566e-03  37.757  < 2e-16 ***
## Inflight.service                   1.207e-01  1.205e-02  10.018  < 2e-16 ***
## Cleanliness                        2.236e-01  1.210e-02  18.471  < 2e-16 ***
## Departure.Delay.in.Minutes         4.759e-03  9.882e-04   4.815 1.47e-06 ***
## Arrival.Delay.in.Minutes          -9.412e-03  9.745e-04  -9.659  < 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: 141768  on 103593  degrees of freedom
## Residual deviance:  69169  on 103570  degrees of freedom
## AIC: 69217
##
## Number of Fisher Scoring iterations: 6
```

```r
# removed Fligh.Distance from the model
model1 <- glm(satisfaction ~ . -Flight.Distance, data = dat, family = binomial)
summary(model1)
```

```
##
```

```
## Call:
## glm(formula = satisfaction ~ . - Flight.Distance, family = binomial,
##     data = dat)
##
## Coefficients:
##                                    Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                      -7.8805974  0.0776711 -101.461  < 2e-16 ***
## GenderMale                        0.0426306  0.0194941    2.187  0.02875 *
## Customer.TypeLoyal Customer       2.0228375  0.0288773   70.049  < 2e-16 ***
## Age                              -0.0082344  0.0007095  -11.606  < 2e-16 ***
## Type.of.TravelPersonal Travel    -2.7162862  0.0312523  -86.915  < 2e-16 ***
## ClassEco                         -0.7262649  0.0243771  -29.793  < 2e-16 ***
## ClassEco Plus                    -0.8401071  0.0403878  -20.801  < 2e-16 ***
## Inflight.wifi.service             0.3958541  0.0114621   34.536  < 2e-16 ***
## Departure.Arrival.time.convenient -0.1245630  0.0082158  -15.161  < 2e-16 ***
## Ease.of.Online.booking           -0.1443757  0.0113484  -12.722  < 2e-16 ***
## Gate.location                     0.0292781  0.0091723    3.192  0.00141 **
## Food.and.drink                   -0.0283801  0.0106844   -2.656  0.00790 **
## Online.boarding                   0.6121496  0.0102449   59.752  < 2e-16 ***
## Seat.comfort                      0.0652383  0.0111807    5.835 5.38e-09 ***
## Inflight.entertainment            0.0654989  0.0142682    4.591 4.42e-06 ***
## On.board.service                  0.3012244  0.0101866   29.571  < 2e-16 ***
## Leg.room.service                  0.2527880  0.0085340   29.621  < 2e-16 ***
## Baggage.handling                  0.1333193  0.0114348   11.659  < 2e-16 ***
## Checkin.service                   0.3233399  0.0085655   37.749  < 2e-16 ***
## Inflight.service                  0.1210841  0.0120457   10.052  < 2e-16 ***
## Cleanliness                       0.2235309  0.0121055   18.465  < 2e-16 ***
## Departure.Delay.in.Minutes        0.0047425  0.0009879    4.800 1.58e-06 ***
## Arrival.Delay.in.Minutes         -0.0093973  0.0009742   -9.646  < 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: 141768  on 103593  degrees of freedom
## Residual deviance:  69172  on 103571  degrees of freedom
## AIC: 69218
##
## Number of Fisher Scoring iterations: 6
```

```r
# Cross Valid --- create test data set
# Using 50%
probs_test = predict(model1, newdata = test, type = "response")
length1 = length(probs_test)
preds_test = rep(0,length1)
preds_test[probs_test > 0.5] = 1
head(probs_test)
```

```
##          1          2          3          4          5          6
## 0.93520342 0.87319022 0.02970501 0.30729370 0.06400260 0.73443376
```

```r
# make confusion matrix
tb = table(prediction = preds_test,
```

```r
            acutal = test$satisfaction)
addmargins(tb)
```

```
##           acutal
## prediction neutral or dissatisfied satisfied   Sum
##        0                     13146      1940 15086
##        1                      1427      9463 10890
##        Sum                   14573     11403 25976
```

last line is the actual data

```r
# Accuracy percent correctly classified
(tb[1,1] +tb[2,2])/25976
```

```
## [1] 0.8703804
```

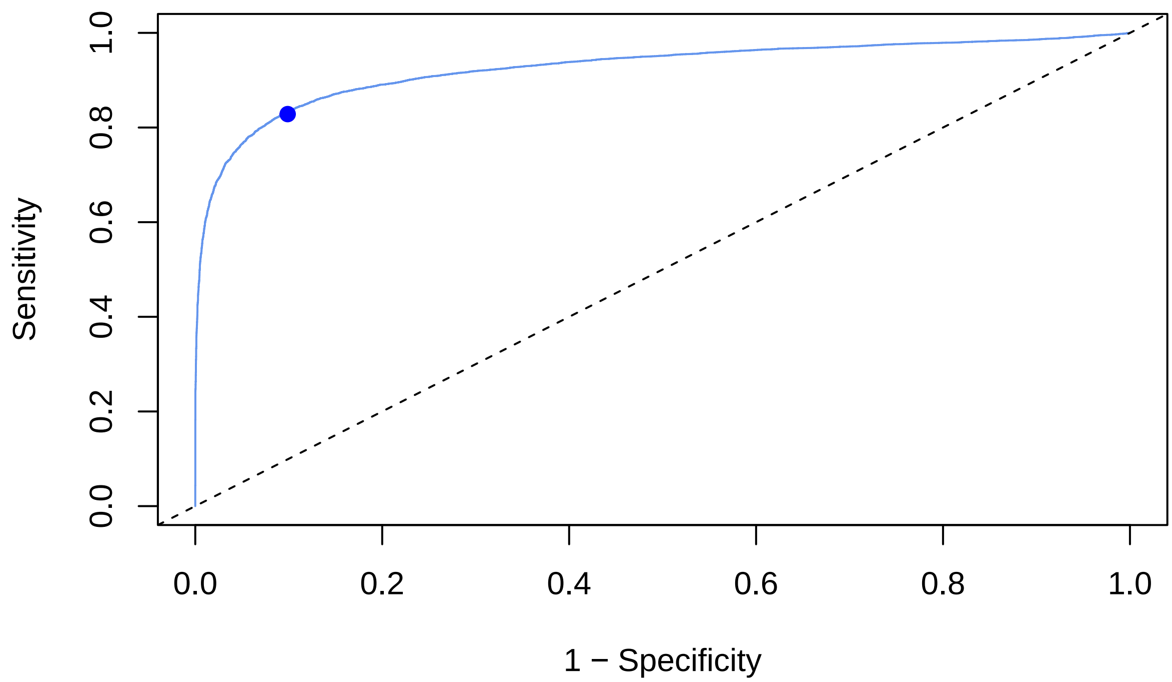
```r
# Sensitivity percent of customer satisfied correctly classified
sensitivity = tb[2,2]/11403
sensitivity
```

```
## [1] 0.8298693
```

```r
# Specificity percent of customers are NOT satisfied correctly classified
specificity = tb[1,1]/14573
specificity
```

```
## [1] 0.9020792
```

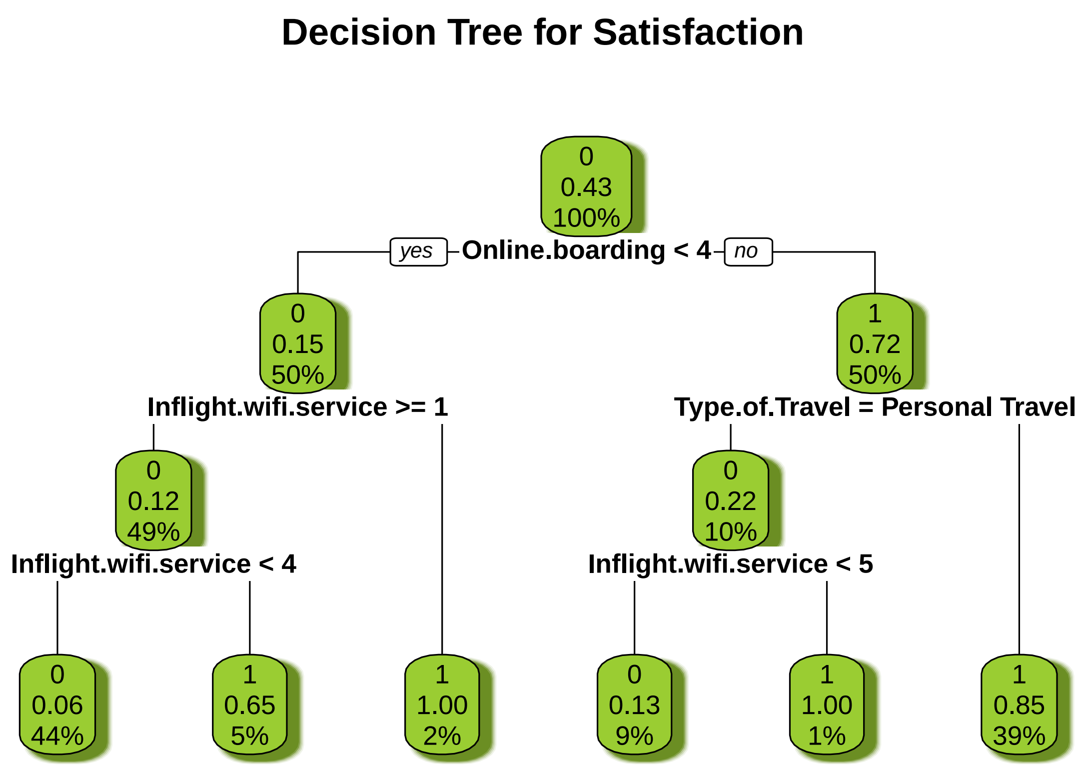
```r
# ROC Curve
roc_obj <- roc(test$satisfaction, probs_test)
plot(1 - roc_obj$specificities, roc_obj$sensitivities, type="l", col = "cornflowerblue",
xlab = "1 - Specificity", ylab = "Sensitivity")

# plot red point corresponding to 0.5 threshold:
points(x = 423/4278, y = 2891/3490, col="blue", pch=19)
abline(0, 1, lty=2) # 1-1 line
```




```
auc(roc_obj)
```

```
## Area under the curve: 0.9255
```

```
# Decision Tree
t1 = rpart(satisfaction ~. -Flight.Distance, data = dat)

#plot the tree with rpart.plot for more customization options
rpart.plot(t1, main = "Decision Tree for Satisfaction",
           box.palette = "yellowgreen",
           shadow.col = "olivedrab", cex = 0.8)
```



Part 3: Using together test and train dataset for model comparison purpose

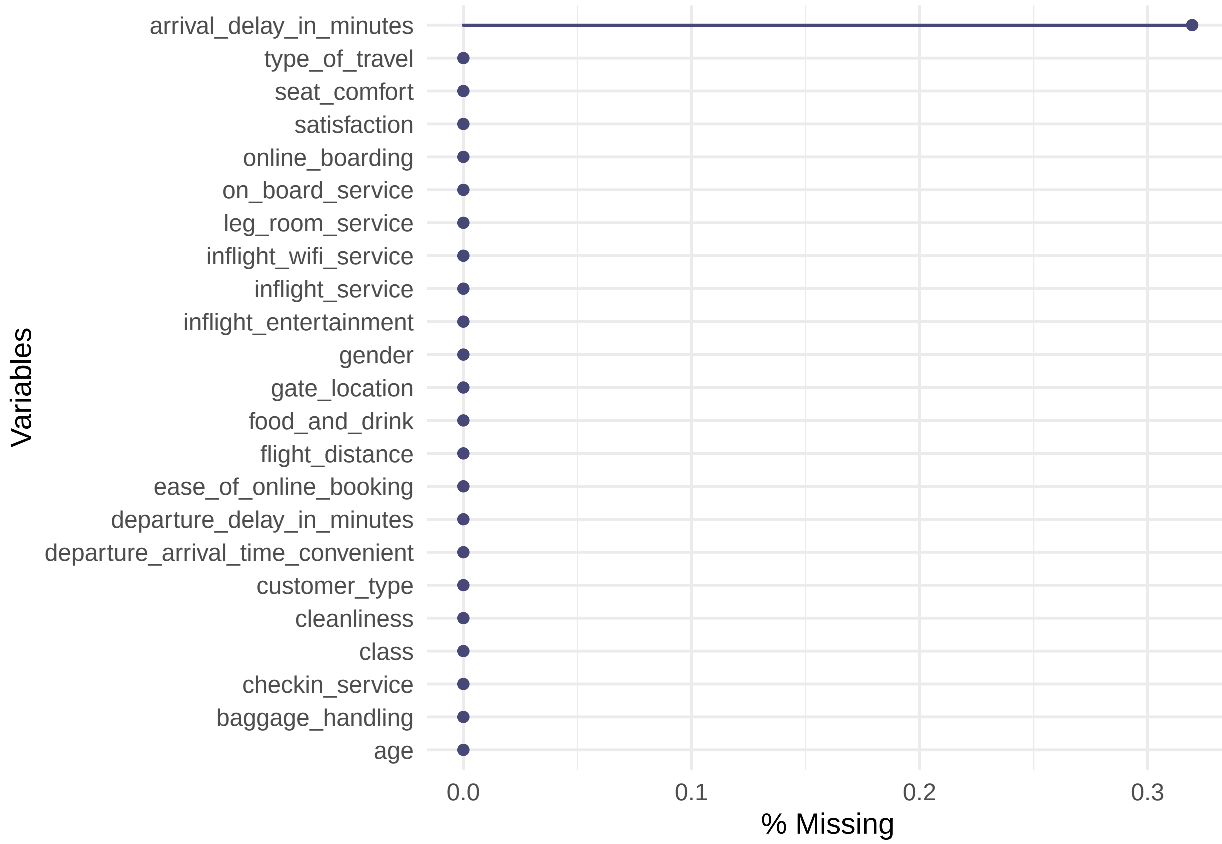
```r
# remove columns X and id for the test data
airtest <- test[,-1:-2]
```

```r
# change binary variable satisfaction to 0 and 1, 1 is satisfied
airtest$satisfaction <- as.factor(ifelse(airtest$satisfaction == "satisfied", 1, 0))

# coerce from chr to factor variables
airtest$Gender= as.factor(airtest$Gender)
airtest$Customer.Type= as.factor(airtest$Customer.Type)
airtest$Type.of.Travel= as.factor(airtest$Type.of.Travel)
airtest$Class= as.factor(airtest$Class)

airtest <- airtest |>
  janitor::clean_names()
summary(airtest)
```

```
##     gender                  customer_type        age
##  Female:13172   disloyal Customer: 4799   Min.   : 7.00
##  Male  :12804   Loyal Customer   :21177   1st Qu.:27.00
##                                           Median :40.00
##                                           Mean   :39.62
```

```
##                                         3rd Qu.:51.00
##                                         Max.   :85.00
##
##          type_of_travel       class       flight_distance inflight_wifi_service
##  Business travel:18038   Business:12495   Min.   :  31    Min.   :0.000
##  Personal Travel: 7938   Eco     :11564   1st Qu.: 414    1st Qu.:2.000
##                          Eco Plus: 1917   Median : 849    Median :3.000
##                                           Mean   :1194    Mean   :2.725
##                                           3rd Qu.:1744    3rd Qu.:4.000
##                                           Max.   :4983    Max.   :5.000
##
##  departure_arrival_time_convenient ease_of_online_booking gate_location
##  Min.   :0.000                     Min.   :0.000          Min.   :1.000
##  1st Qu.:2.000                     1st Qu.:2.000          1st Qu.:2.000
##  Median :3.000                     Median :3.000          Median :3.000
##  Mean   :3.047                     Mean   :2.757          Mean   :2.977
##  3rd Qu.:4.000                     3rd Qu.:4.000          3rd Qu.:4.000
##  Max.   :5.000                     Max.   :5.000          Max.   :5.000
##
##  food_and_drink  online_boarding  seat_comfort   inflight_entertainment
##  Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :4.000   Median :4.000   Median :4.000
##  Mean   :3.215   Mean   :3.262   Mean   :3.449   Mean   :3.358
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##
##  on_board_service leg_room_service baggage_handling checkin_service
##  Min.   :0.000    Min.   :0.00     Min.   :1.000    Min.   :1.000
##  1st Qu.:2.000    1st Qu.:2.00     1st Qu.:3.000    1st Qu.:3.000
##  Median :4.000    Median :4.00     Median :4.000    Median :3.000
##  Mean   :3.386    Mean   :3.35     Mean   :3.633    Mean   :3.314
##  3rd Qu.:4.000    3rd Qu.:4.00     3rd Qu.:5.000    3rd Qu.:4.000
##  Max.   :5.000    Max.   :5.00     Max.   :5.000    Max.   :5.000
##
##  inflight_service  cleanliness    departure_delay_in_minutes
##  Min.   :0.000    Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000    1st Qu.:2.000   1st Qu.:   0.00
##  Median :4.000    Median :3.000   Median :   0.00
##  Mean   :3.649    Mean   :3.286   Mean   :  14.31
##  3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000    Max.   :5.000   Max.   :1128.00
##
##  arrival_delay_in_minutes satisfaction
##  Min.   :   0.00          0:14573
##  1st Qu.:   0.00          1:11403
##  Median :   0.00
##  Mean   :  14.74
##  3rd Qu.:  13.00
##  Max.   :1115.00
##  NA's   :83
```

```
# Missing information and visualize
gg_miss_var(airtest, show_pct = TRUE)
```




```r
# Remove N/A of the Arrival.Delay.in.Minutes
#  = airtest[!is.na(airtest$Arrival.Delay.in.Minutes) ,]
airtest <- airtest %>% drop_na()
```

```r
# remove columns X and id for the test data
airtrain <- train[,-1:-2]
```

```r
airtrain$satisfaction <- as.factor(ifelse(airtrain$satisfaction == "satisfied", 1, 0))

# coerce from chr to factor variables
airtrain$Gender= as.factor(airtrain$Gender)
airtrain$Customer.Type= as.factor(airtrain$Customer.Type)
airtrain$Type.of.Travel= as.factor(airtrain$Type.of.Travel)
airtrain$Class= as.factor(airtrain$Class)

airtrain <- airtrain |>
  janitor::clean_names()
summary(airtrain)
```

```
##     gender                   customer_type         age
##  Female:52727   disloyal Customer:18981   Min.   : 7.00
##  Male  :51177   Loyal Customer   :84923   1st Qu.:27.00
##                                           Median :40.00
##                                           Mean   :39.38
##                                           3rd Qu.:51.00
```

```
##                                         Max.   :85.00
##
##          type_of_travel       class       flight_distance inflight_wifi_service
##  Business travel:71655   Business:49665   Min.   :  31    Min.   :0.00
##  Personal Travel:32249   Eco     :46745   1st Qu.: 414    1st Qu.:2.00
##                          Eco Plus: 7494   Median : 843    Median :3.00
##                                           Mean   :1189    Mean   :2.73
##                                           3rd Qu.:1743    3rd Qu.:4.00
##                                           Max.   :4983    Max.   :5.00
##
##  departure_arrival_time_convenient ease_of_online_booking gate_location
##  Min.   :0.00                      Min.   :0.000          Min.   :0.000
##  1st Qu.:2.00                      1st Qu.:2.000          1st Qu.:2.000
##  Median :3.00                      Median :3.000          Median :3.000
##  Mean   :3.06                      Mean   :2.757          Mean   :2.977
##  3rd Qu.:4.00                      3rd Qu.:4.000          3rd Qu.:4.000
##  Max.   :5.00                      Max.   :5.000          Max.   :5.000
##
##  food_and_drink  online_boarding  seat_comfort   inflight_entertainment
##  Min.   :0.000   Min.   :0.00    Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :3.00    Median :4.000   Median :4.000
##  Mean   :3.202   Mean   :3.25    Mean   :3.439   Mean   :3.358
##  3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.00    Max.   :5.000   Max.   :5.000
##
##  on_board_service leg_room_service baggage_handling checkin_service
##  Min.   :0.000    Min.   :0.000    Min.   :1.000    Min.   :0.000
##  1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000
##  Median :4.000    Median :4.000    Median :4.000    Median :3.000
##  Mean   :3.382    Mean   :3.351    Mean   :3.632    Mean   :3.304
##  3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.000
##  Max.   :5.000    Max.   :5.000    Max.   :5.000    Max.   :5.000
##
##  inflight_service  cleanliness    departure_delay_in_minutes
##  Min.   :0.00     Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.00     1st Qu.:2.000   1st Qu.:   0.00
##  Median :4.00     Median :3.000   Median :   0.00
##  Mean   :3.64     Mean   :3.286   Mean   :  14.82
##  3rd Qu.:5.00     3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.00     Max.   :5.000   Max.   :1592.00
##
##  arrival_delay_in_minutes satisfaction
##  Min.   :   0.00          0:58879
##  1st Qu.:   0.00          1:45025
##  Median :   0.00
##  Mean   :  15.18
##  3rd Qu.:  13.00
##  Max.   :1584.00
##  NA's   :310
```

```r
# Remove N/A of the Arrival.Delay.in.Minutes
airtrain<- airtrain %>% drop_na()
```

```r
str(airtrain)
```

```
## 'data.frame':    103594 obs. of  23 variables:
##  $ gender                           : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
##  $ customer_type                    : Factor w/ 2 levels "disloyal Customer",..: 2 1 2 2 2 2 2 2 2 1
##  $ age                              : int  13 25 26 25 61 26 47 52 41 20 ...
##  $ type_of_travel                   : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 .
##  $ class                            : Factor w/ 3 levels "Business","Eco",..: 3 1 1 1 1 2 2 1 1 2 ..
##  $ flight_distance                  : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
##  $ inflight_wifi_service            : int  3 3 2 2 3 3 2 4 1 3 ...
##  $ departure_arrival_time_convenient: int  4 2 2 5 3 4 4 3 2 3 ...
##  $ ease_of_online_booking           : int  3 3 2 5 3 2 2 4 2 3 ...
##  $ gate_location                    : int  1 3 2 5 3 1 3 4 2 4 ...
##  $ food_and_drink                   : int  5 1 5 2 4 1 2 5 4 2 ...
##  $ online_boarding                  : int  3 3 5 2 5 2 2 5 3 3 ...
##  $ seat_comfort                     : int  5 1 5 2 5 1 2 5 3 3 ...
##  $ inflight_entertainment           : int  5 1 5 2 3 1 2 5 1 2 ...
##  $ on_board_service                 : int  4 1 4 2 3 3 3 5 1 2 ...
##  $ leg_room_service                 : int  3 5 3 5 4 4 3 5 2 3 ...
##  $ baggage_handling                 : int  4 3 4 3 4 4 4 5 1 4 ...
##  $ checkin_service                  : int  4 1 4 1 3 4 3 4 4 4 ...
##  $ inflight_service                 : int  5 4 4 4 3 4 5 5 1 3 ...
##  $ cleanliness                      : int  5 1 5 2 3 1 2 4 2 2 ...
##  $ departure_delay_in_minutes       : int  25 1 0 11 0 0 9 4 0 0 ...
##  $ arrival_delay_in_minutes         : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                     : Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 2 1 1 ...
```

```r
# Performance on train data
# Logisitcs
log_model <- glm(satisfaction ~ ., data = airtrain, family = binomial)

summary(log_model)
```

```
##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = airtrain)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.860e+00  7.876e-02 -99.793  < 2e-16 ***
## genderMale                         4.255e-02  1.949e-02   2.183  0.02905 *
## customer_typeLoyal Customer        2.035e+00  2.994e-02  67.970  < 2e-16 ***
## age                               -8.308e-03  7.110e-04 -11.684  < 2e-16 ***
## type_of_travelPersonal Travel     -2.722e+00  3.147e-02 -86.494  < 2e-16 ***
## classEco                          -7.389e-01  2.566e-02 -28.794  < 2e-16 ***
## classEco Plus                     -8.554e-01  4.155e-02 -20.588  < 2e-16 ***
## flight_distance                   -1.789e-05  1.132e-05  -1.581  0.11392
## inflight_wifi_service              3.949e-01  1.148e-02  34.405  < 2e-16 ***
## departure_arrival_time_convenient -1.244e-01  8.218e-03 -15.132  < 2e-16 ***
## ease_of_online_booking            -1.440e-01  1.135e-02 -12.691  < 2e-16 ***
## gate_location                      2.914e-02  9.174e-03   3.176  0.00149 **
## food_and_drink                    -2.860e-02  1.068e-02  -2.677  0.00743 **
## online_boarding                    6.126e-01  1.025e-02  59.773  < 2e-16 ***
```

```
## seat_comfort                      6.555e-02  1.118e-02   5.862 4.58e-09 ***
## inflight_entertainment            6.555e-02  1.427e-02   4.594 4.34e-06 ***
## on_board_service                  3.014e-01  1.019e-02  29.582  < 2e-16 ***
## leg_room_service                  2.532e-01  8.540e-03  29.652  < 2e-16 ***
## baggage_handling                  1.331e-01  1.144e-02  11.633  < 2e-16 ***
## checkin_service                   3.234e-01  8.566e-03  37.757  < 2e-16 ***
## inflight_service                  1.207e-01  1.205e-02  10.018  < 2e-16 ***
## cleanliness                       2.236e-01  1.210e-02  18.471  < 2e-16 ***
## departure_delay_in_minutes        4.759e-03  9.882e-04   4.815 1.47e-06 ***
## arrival_delay_in_minutes         -9.412e-03  9.745e-04  -9.659  < 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: 141768  on 103593  degrees of freedom
## Residual deviance:  69169  on 103570  degrees of freedom
## AIC: 69217
##
## Number of Fisher Scoring iterations: 6
```

```r
log_step <-stats::step(log_model)
```

```
## Start:  AIC=69217.21
## satisfaction ~ gender + customer_type + age + type_of_travel +
##     class + flight_distance + inflight_wifi_service + departure_arrival_time_convenient +
##     ease_of_online_booking + gate_location + food_and_drink +
##     online_boarding + seat_comfort + inflight_entertainment +
##     on_board_service + leg_room_service + baggage_handling +
##     checkin_service + inflight_service + cleanliness + departure_delay_in_minutes +
##     arrival_delay_in_minutes
##
##                                     Df Deviance   AIC
## <none>                                    69169 69217
## - flight_distance                    1    69172 69218
## - gender                             1    69174 69220
## - food_and_drink                     1    69176 69222
## - gate_location                      1    69179 69225
## - inflight_entertainment             1    69190 69236
## - departure_delay_in_minutes         1    69193 69239
## - seat_comfort                       1    69204 69250
## - arrival_delay_in_minutes           1    69264 69310
## - inflight_service                   1    69270 69316
## - baggage_handling                   1    69305 69351
## - age                                1    69306 69352
## - ease_of_online_booking             1    69331 69377
## - departure_arrival_time_convenient  1    69397 69443
## - cleanliness                        1    69512 69558
## - leg_room_service                   1    70054 70100
## - on_board_service                   1    70061 70107
## - class                              2    70113 70157
## - inflight_wifi_service              1    70392 70438
## - checkin_service                    1    70642 70688
## - online_boarding                    1    72969 73015
```

```
## - customer_type                      1    74247 74293
## - type_of_travel                     1    77964 78010
```

```r
summary(log_step)
```

```
##
## Call:
## glm(formula = satisfaction ~ gender + customer_type + age + type_of_travel +
##     class + flight_distance + inflight_wifi_service + departure_arrival_time_convenient +
##     ease_of_online_booking + gate_location + food_and_drink +
##     online_boarding + seat_comfort + inflight_entertainment +
##     on_board_service + leg_room_service + baggage_handling +
##     checkin_service + inflight_service + cleanliness + departure_delay_in_minutes +
##     arrival_delay_in_minutes, family = binomial, data = airtrain)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.860e+00  7.876e-02 -99.793  < 2e-16 ***
## genderMale                         4.255e-02  1.949e-02   2.183  0.02905 *
## customer_typeLoyal Customer        2.035e+00  2.994e-02  67.970  < 2e-16 ***
## age                               -8.308e-03  7.110e-04 -11.684  < 2e-16 ***
## type_of_travelPersonal Travel     -2.722e+00  3.147e-02 -86.494  < 2e-16 ***
## classEco                          -7.389e-01  2.566e-02 -28.794  < 2e-16 ***
## classEco Plus                     -8.554e-01  4.155e-02 -20.588  < 2e-16 ***
## flight_distance                   -1.789e-05  1.132e-05  -1.581  0.11392
## inflight_wifi_service              3.949e-01  1.148e-02  34.405  < 2e-16 ***
## departure_arrival_time_convenient -1.244e-01  8.218e-03 -15.132  < 2e-16 ***
## ease_of_online_booking            -1.440e-01  1.135e-02 -12.691  < 2e-16 ***
## gate_location                      2.914e-02  9.174e-03   3.176  0.00149 **
## food_and_drink                    -2.860e-02  1.068e-02  -2.677  0.00743 **
## online_boarding                    6.126e-01  1.025e-02  59.773  < 2e-16 ***
## seat_comfort                       6.555e-02  1.118e-02   5.862 4.58e-09 ***
## inflight_entertainment             6.555e-02  1.427e-02   4.594 4.34e-06 ***
## on_board_service                   3.014e-01  1.019e-02  29.582  < 2e-16 ***
## leg_room_service                   2.532e-01  8.540e-03  29.652  < 2e-16 ***
## baggage_handling                   1.331e-01  1.144e-02  11.633  < 2e-16 ***
## checkin_service                    3.234e-01  8.566e-03  37.757  < 2e-16 ***
## inflight_service                   1.207e-01  1.205e-02  10.018  < 2e-16 ***
## cleanliness                        2.236e-01  1.210e-02  18.471  < 2e-16 ***
## departure_delay_in_minutes         4.759e-03  9.882e-04   4.815 1.47e-06 ***
## arrival_delay_in_minutes          -9.412e-03  9.745e-04  -9.659  < 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: 141768  on 103593  degrees of freedom
## Residual deviance:  69169  on 103570  degrees of freedom
## AIC: 69217
##
## Number of Fisher Scoring iterations: 6
```

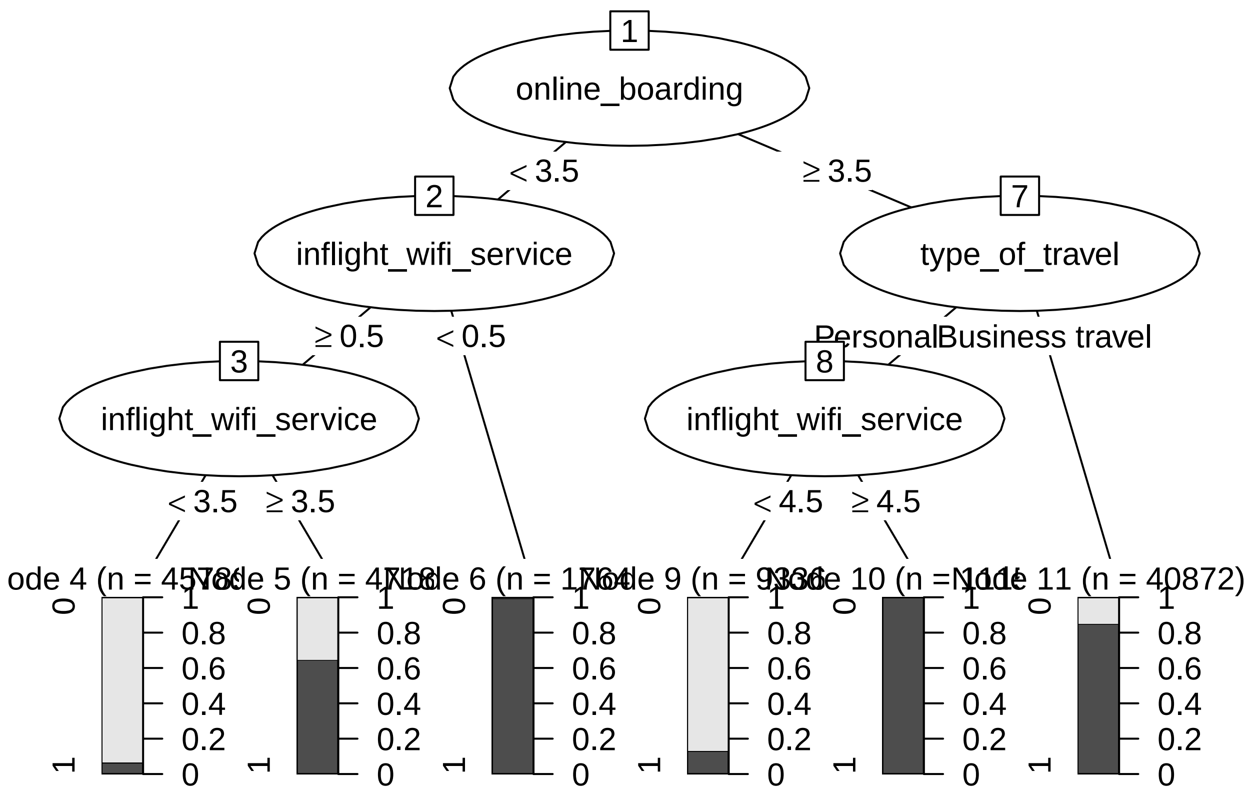
```r
# Performance on train data
pred <- airtrain %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    list(.pred_class = as.factor(as.integer(predict(log_step, newdata = airtrain, type = "response") >0
  ) %>%
  rename(sat_log = .pred_class)

confusion_log_1 <- pred %>%
  conf_mat(truth = 1, estimate = sat_log)

log_train_acc<-accuracy(pred, satisfaction, sat_log)


# Performance on test data
pred <- airtest %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    list(.pred_class2 = as.factor(as.integer(predict(log_step, newdata = airtest, type = "response") >0
  ) %>%
  rename(sat_log = .pred_class2)


confusion_log_2 <- pred %>%
  conf_mat(truth = 1, estimate = sat_log)

confusion_log_2
```

```
##           Truth
## Prediction     0     1
##          0 13104  1898
##          1  1424  9467
```

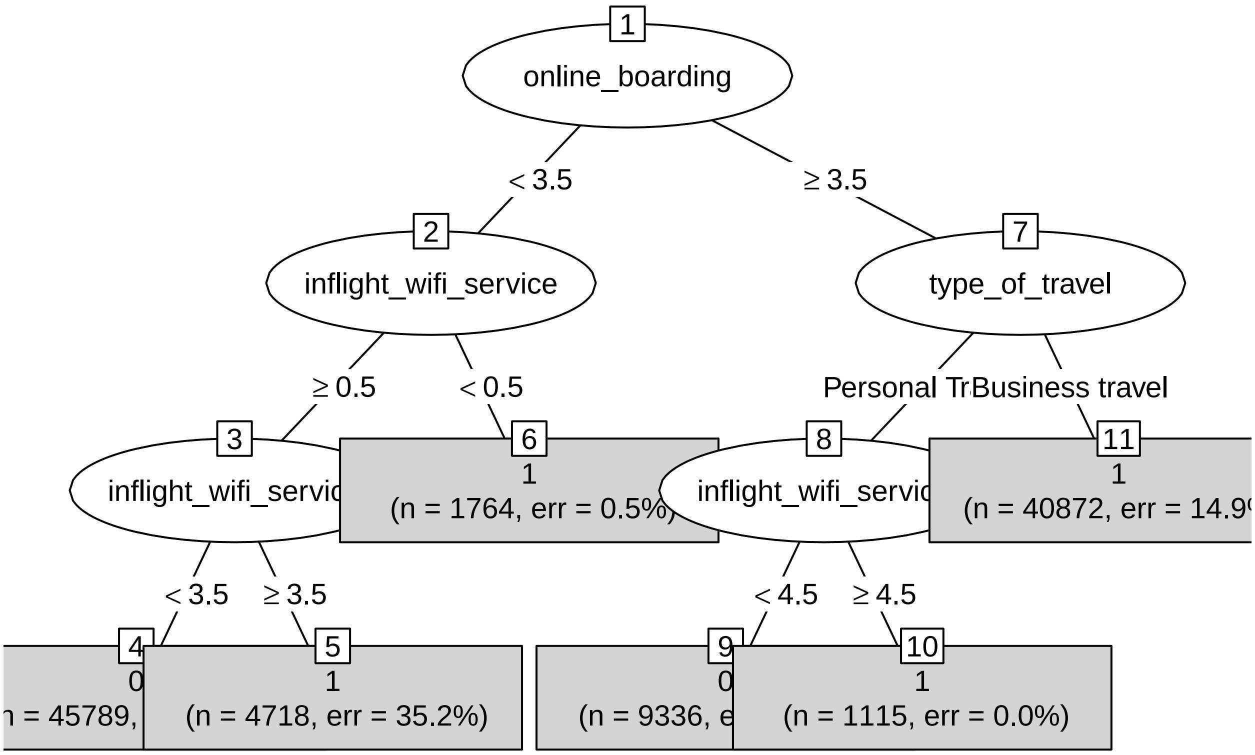
```r
log_test_acc<-accuracy(pred, satisfaction, sat_log)

# Predict probabilities
predicted_probs <- predict(log_step, type = "response",newdata = airtrain)

# Calculate AUC
roc_obj <- roc(airtrain$satisfaction, predicted_probs)
log_train_auc<- auc(roc_obj)


# Predict probabilities
predicted_probs <- predict(log_step, type = "response",newdata = airtest)

# Calculate AUC
roc_obj <- roc(airtest$satisfaction, predicted_probs)
log_test_auc<- auc(roc_obj)
```

Decision tree model

```r
mod_dtree <- decision_tree(mode = "classification") %>%
  set_engine("rpart") %>%
  fit(satisfaction ~., data = airtrain)

split_val <- mod_dtree$fit$splits %>%
  as_tibble() %>%
  pull(index)

plot(as.party(mod_dtree$fit))
```




```r
plot(as.party(mod_dtree$fit), type = "simple",gp=gpar(cex=0.9))
```

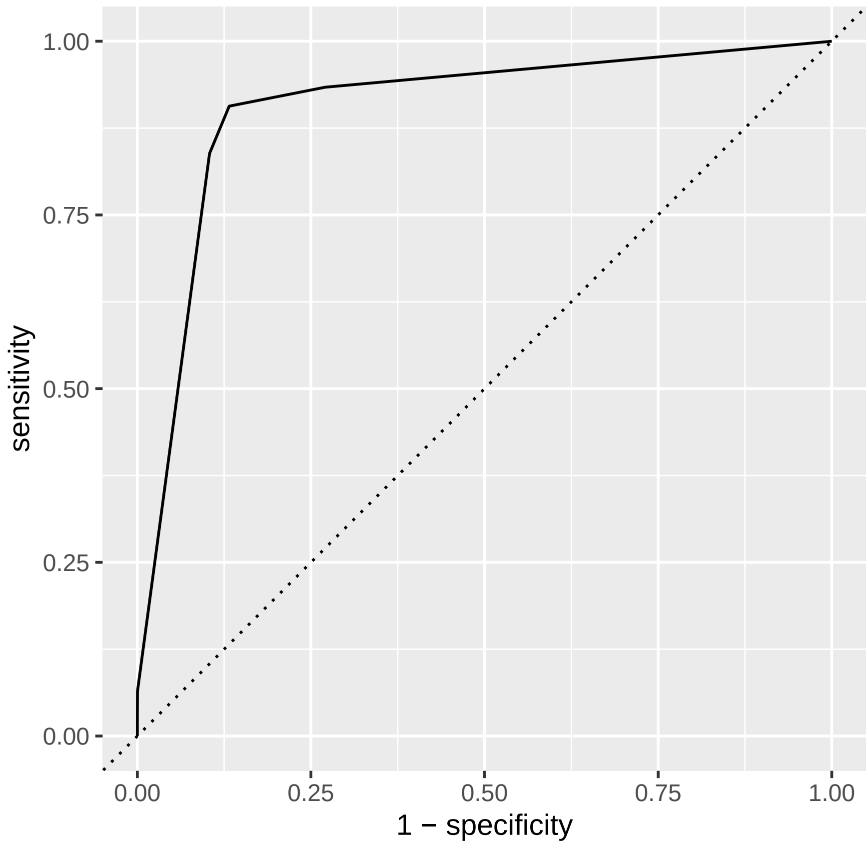



```r
##train###
pred <- airtrain %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_dtree, new_data = airtrain, type = "class")
  ) %>%
  rename(sat_log = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = sat_log)
confusion
```

```
##            Truth
## Prediction     0     1
##          0 50931  4194
##          1  7766 40703
```

```r
dtree_train_acc<-accuracy(pred, satisfaction, sat_log)

mod_dtree %>%
  predict(airtrain, type = "prob") %>%
  bind_cols(airtrain) %>%
  roc_curve(satisfaction, .pred_1,event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_path() +
```

```r
  geom_abline(lty = 3) +
  coord_equal()
```



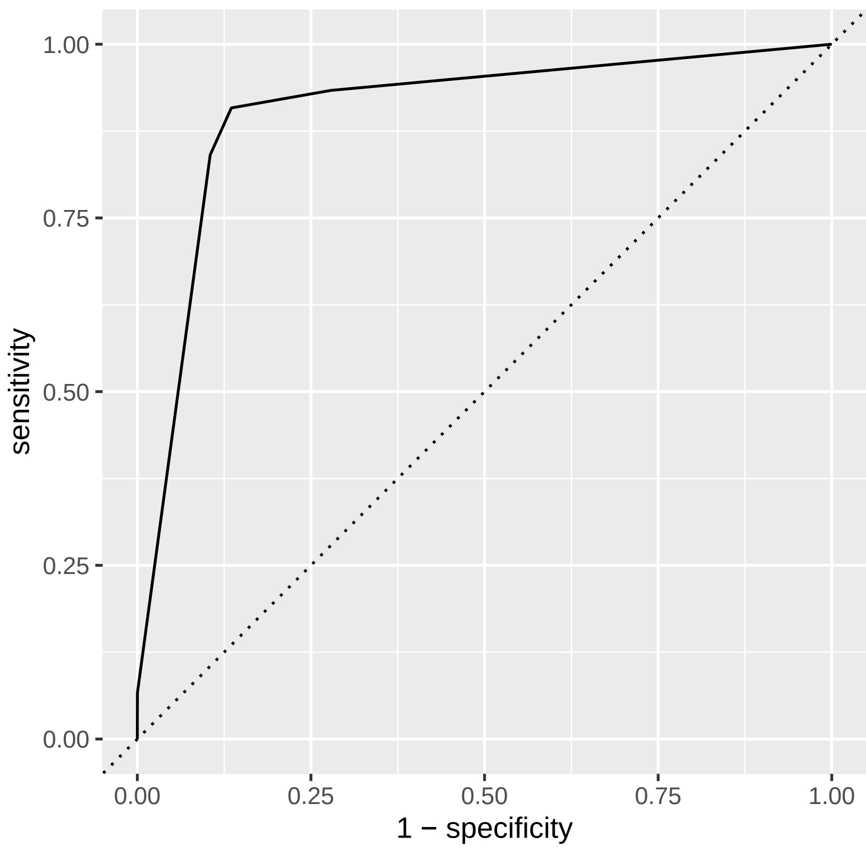

```r
mod_dtree %>%
  predict(airtrain, type = "prob") %>%
   bind_cols(airtrain) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.904
```

```r
##test###

pred <- airtest %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_dtree, new_data = airtest, type = "class")
  ) %>%
  rename(sat_log = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = sat_log)
confusion
```

```
##           Truth
## Prediction     0     1
##          0 12561  1042
##          1  1967 10323
```

```
dtree_test_acc<-accuracy(pred, satisfaction, sat_log)

mod_dtree %>%
  predict(airtest, type = "prob") %>%
  bind_cols(airtest) %>%
  roc_curve(satisfaction, .pred_1,event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_path() +
  geom_abline(lty = 3) +
  coord_equal()
```




```
mod_dtree %>%
  predict(airtest, type = "prob") %>%
   bind_cols(airtest) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.904
```

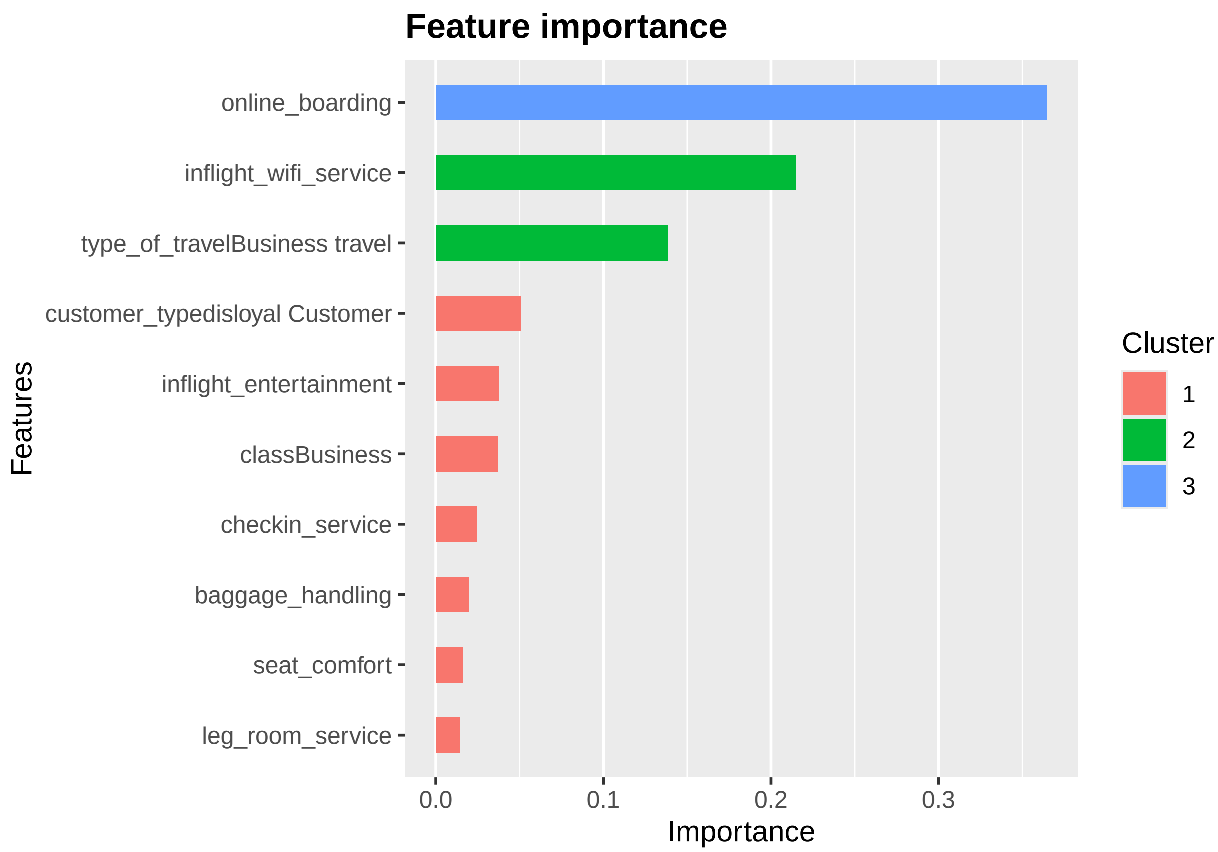
```r
###
# Predict probabilities
predicted_probs <- predict(mod_dtree, type = "prob",new_data = airtrain) %>% dplyr::select(.pred_1) %>%

# Calculate AUC
roc_obj <- roc(airtrain$satisfaction, predicted_probs)
dtree_train_auc<- auc(roc_obj)


# Predict probabilities
predicted_probs <- predict(mod_dtree, type = "prob",new_data = airtest) %>% dplyr::select(.pred_1) %>% p

# Calculate AUC
roc_obj <- roc(airtest$satisfaction, predicted_probs)
dtree_test_auc<- auc(roc_obj)
```

xgb Boosting tree

```r
mod_xgb <- boost_tree(trees = 50) %>%
  set_engine("xgboost") %>%
  set_mode("classification") %>%
  fit(satisfaction ~., data = airtrain)

xgb.importance(model=mod_xgb$fit)
```

```
##                               Feature         Gain        Cover   Frequency
##  1:                   online_boarding 3.648373e-01 0.1207968382 0.048015679
##  2:              inflight_wifi_service 2.147169e-01 0.2133385339 0.129348359
##  3:       type_of_travelBusiness travel 1.387206e-01 0.0861316807 0.057324841
##  4:      customer_typedisloyal Customer 5.060494e-02 0.0620980243 0.049975502
##  5:             inflight_entertainment 3.739522e-02 0.0344723230 0.044096031
##  6:                      classBusiness 3.724004e-02 0.0645776076 0.044585987
##  7:                    checkin_service 2.444558e-02 0.0379445514 0.030867222
##  8:                   baggage_handling 1.986446e-02 0.0424393109 0.040666340
##  9:                       seat_comfort 1.596779e-02 0.0273989616 0.040666340
## 10:                   leg_room_service 1.434980e-02 0.0174940673 0.033806957
## 11:                   on_board_service 1.406622e-02 0.0277691851 0.027927487
## 12:                   inflight_service 1.312813e-02 0.0352214229 0.043606075
## 13:                      gate_location 1.204409e-02 0.0147422608 0.041646252
## 14:                                age 1.172356e-02 0.0550266832 0.086232239
## 15:                        cleanliness 1.140180e-02 0.0254860227 0.024497795
## 16: departure_arrival_time_convenient 4.755717e-03 0.0168062568 0.031847134
## 17:           arrival_delay_in_minutes 4.713017e-03 0.0298409752 0.042626164
## 18:                    flight_distance 3.829179e-03 0.0389811319 0.078882901
## 19:             ease_of_online_booking 3.345284e-03 0.0335242816 0.035766781
## 20:                     food_and_drink 1.182115e-03 0.0043309922 0.019108280
## 21:         departure_delay_in_minutes 1.059773e-03 0.0105928950 0.030377266
## 22:                       genderFemale 3.606839e-04 0.0004302178 0.010289074
## 23:                           classEco 1.817609e-04 0.0002327061 0.004409603
## 24:                      classEco Plus 6.612234e-05 0.0003230698 0.003429691
##                               Feature         Gain        Cover   Frequency
```

```r
xgb.importance(model=mod_xgb$fit) %>% xgb.ggplot.importance(
top_n=10, measure=NULL, rel_to_first = F)
```

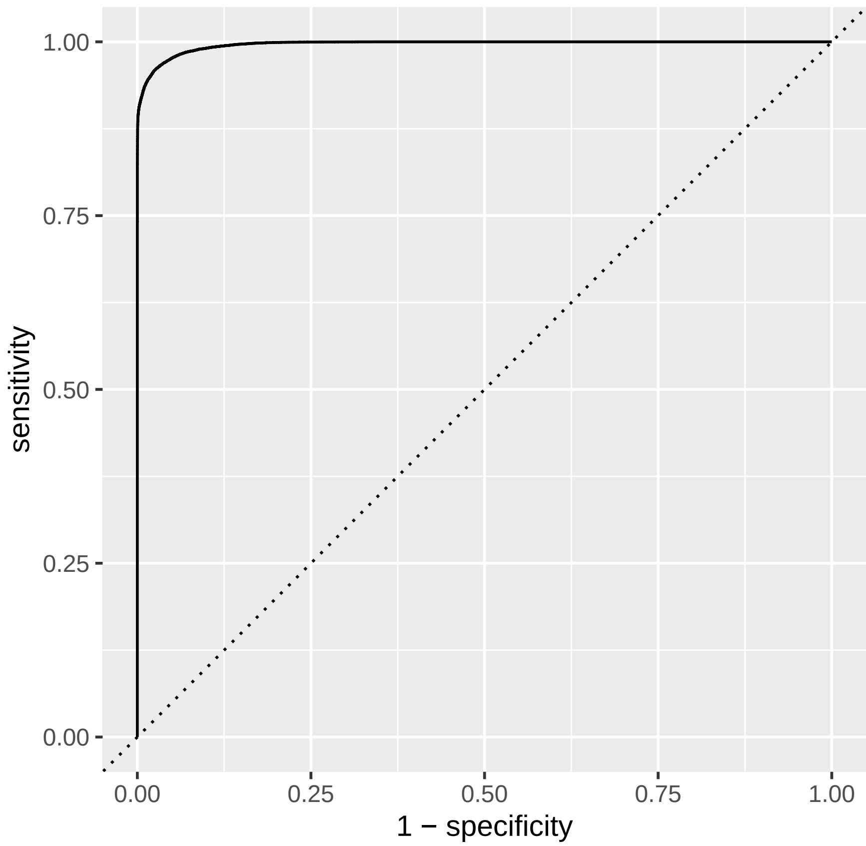



```r
summary(mod_xgb)
```

```
##             Length Class       Mode
## lvl         2      -none-      character
## spec        8      boost_tree  list
## fit         9      xgb.Booster list
## preproc     4      -none-      list
## elapsed     1      -none-      list
## censor_probs 0     -none-      list
```

```r
##train###
pred <- airtrain %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_xgb, new_data = airtrain, type = "class")
  ) %>%
  rename(satisfaction_null = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = satisfaction_null)
```

```r
mod_xgb %>%
  predict(airtrain, type = "prob") %>%
  bind_cols(airtrain) %>%
  roc_curve(satisfaction, .pred_1,event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_path() +
  geom_abline(lty = 3) +
  coord_equal()
```




```r
mod_xgb %>%
  predict(airtrain, type = "prob") %>%
   bind_cols(airtrain) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.996
```

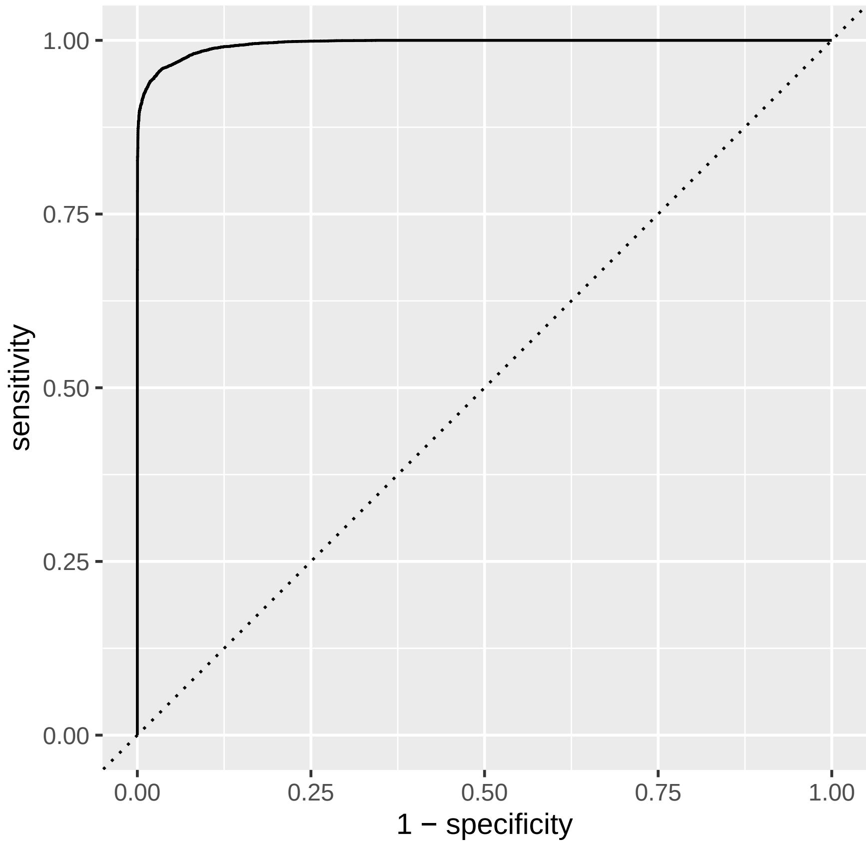
```r
confusion
```

```
##           Truth
## Prediction     0     1
##          0 57729  2375
##          1   968 42522
```

```r
xgb_train_acc<-accuracy(pred, satisfaction, satisfaction_null)

###test###
pred <- airtest %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_xgb, new_data = airtest, type = "class")
  ) %>%
  rename(satisfaction_null = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = satisfaction_null)

confusion
```

```
##           Truth
## Prediction     0     1
##          0 14226   647
##          1   302 10718
```

```r
xgb_test_acc<-accuracy(pred, satisfaction, satisfaction_null)

mod_xgb %>%
  predict(airtest, type = "prob") %>%
  bind_cols(airtest) %>%
  roc_curve(satisfaction, .pred_1,event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_path() +
  geom_abline(lty = 3) +
  coord_equal()
```




```
mod_xgb %>%
  predict(airtest, type = "prob") %>%
   bind_cols(airtest) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.995
```

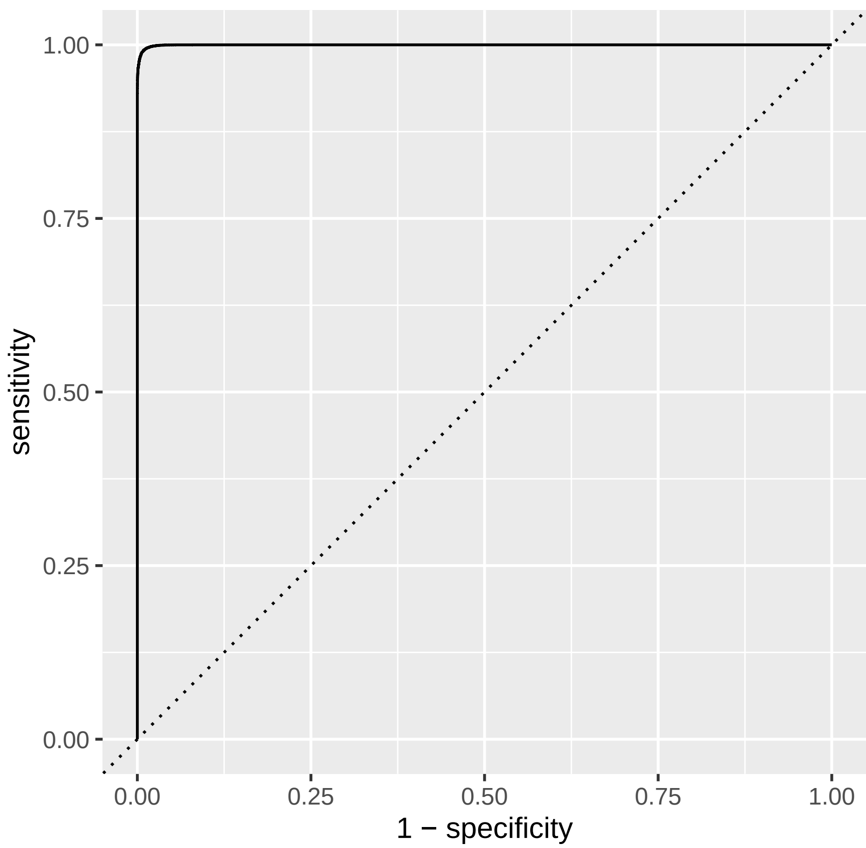
```
predicted_probs <- predict(mod_xgb, type = "prob",new_data = airtrain) %>% dplyr::select(.pred_1) %>% pu

# Calculate AUC
roc_obj <- roc(airtrain$satisfaction, predicted_probs)
xgb_train_auc<- auc(roc_obj)


# Predict probabilities
predicted_probs <- predict(mod_xgb, type = "prob",new_data = airtest) %>% dplyr::select(.pred_1) %>% pul

# Calculate AUC
roc_obj <- roc(airtest$satisfaction, predicted_probs)
xgb_test_auc<- auc(roc_obj)
```

Random Forest

```
##train###
mod_rf_ranger <- rand_forest(trees = 50) %>%
  set_engine("ranger",importance = "impurity") %>%
  set_mode("classification") %>%
  fit(satisfaction ~ ., data = airtrain)

perf_train <-mod_rf_ranger %>%
  predict(airtrain) %>%
  bind_cols(airtrain) %>%
  metrics(truth = satisfaction, estimate = .pred_class)

RF_train_acc<-perf_train[1,3]

mod_rf_ranger %>%
  predict(airtrain) %>%
  bind_cols(airtrain) %>%
  conf_mat(truth = satisfaction, estimate = .pred_class)
```

```
##           Truth
## Prediction     0     1
##          0 58490   875
##          1   207 44022
```

```
mod_rf_ranger %>%
  predict(airtrain, type = "prob") %>%
  bind_cols(airtrain) %>%
  roc_curve(satisfaction, .pred_1,event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_path() +
  geom_abline(lty = 3) +
  coord_equal()
```




```
mod_rf_ranger %>%
  predict(airtrain, type = "prob") %>%
   bind_cols(airtrain) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

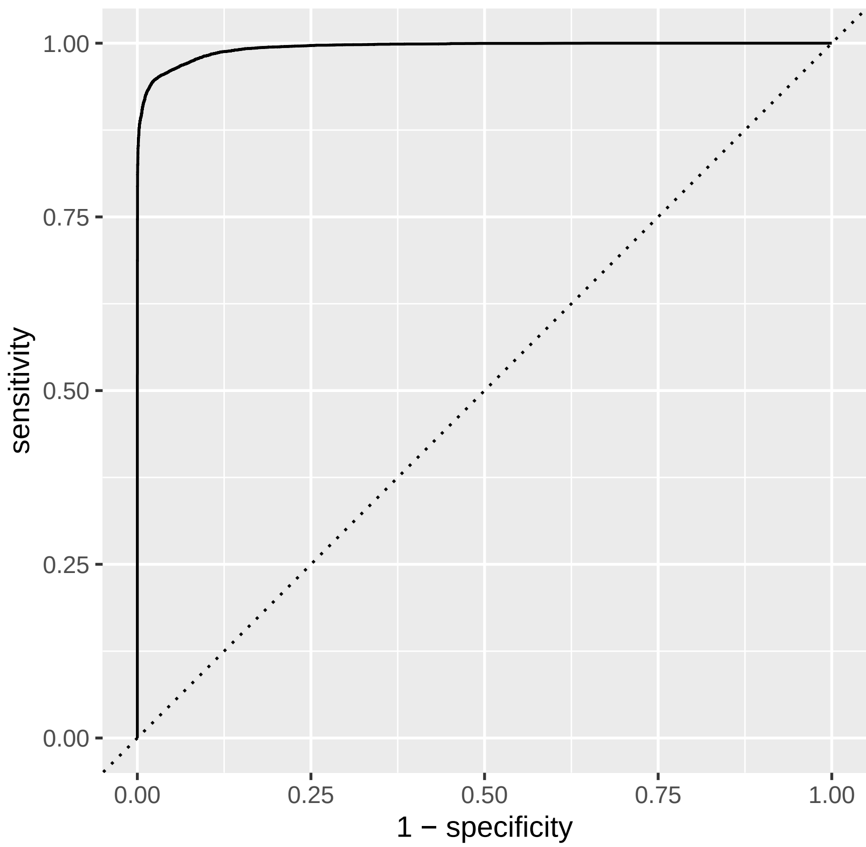
```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary          1.00
```

```
##test###
perf_test <-mod_rf_ranger %>%
  predict(airtest) %>%
  bind_cols(airtest) %>%
  metrics(truth = satisfaction, estimate = .pred_class)

RF_test_acc<-perf_test[1,3]

mod_rf_ranger %>%
  predict(airtest) %>%
  bind_cols(airtest) %>%
  conf_mat(truth = satisfaction, estimate = .pred_class)
```

```
##           Truth
## Prediction     0     1
```

```
##           0 14210   635
##           1   318 10730
```

```
mod_rf_ranger %>%
  predict(airtest, type = "prob") %>%
  bind_cols(airtest) %>%
  roc_curve(satisfaction, .pred_1,event_level = "second") %>%
  ggplot(aes(x = 1 - specificity, y = sensitivity)) +
  geom_path() +
  geom_abline(lty = 3) +
  coord_equal()
```

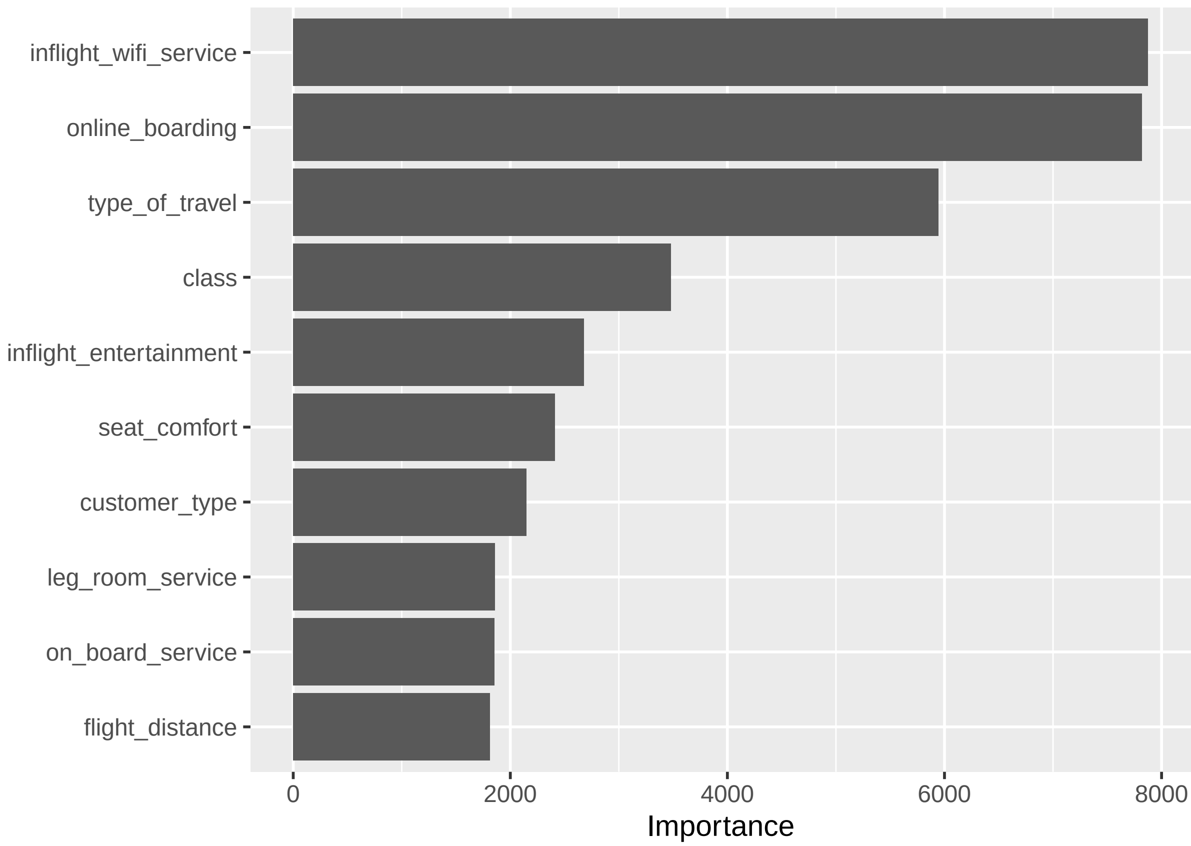



```
mod_rf_ranger %>%
  predict(airtest, type = "prob") %>%
   bind_cols(airtest) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.994
```

```
######using workflow to get variable importance:###
rf_mod<- rand_forest(trees = 50) %>%
```

```r
  set_engine("ranger",importance = "impurity") %>%
  set_mode("classification")

rf_recipe <-
  recipe(satisfaction ~ ., data = airtrain)

rf_workflow <-
  workflow() %>%
  add_model(rf_mod) %>%
  add_recipe(rf_recipe)

rf_workflow %>%
  fit(airtrain) %>%
  extract_fit_parsnip() %>%
  vip(num_features = 10)
```




```r
predicted_probs <- predict(mod_rf_ranger, type = "prob",new_data = airtrain) %>% dplyr::select(.pred_1)

# Calculate AUC
roc_obj <- roc(airtrain$satisfaction, predicted_probs)
rf_train_auc<- auc(roc_obj)


# Predict probabilities
predicted_probs <- predict(mod_rf_ranger, type = "prob",new_data = airtest) %>% dplyr::select(.pred_1)
```

```r
# Calculate AUC
roc_obj <- roc(airtest$satisfaction, predicted_probs)
rf_test_auc<- auc(roc_obj)
```

LASSO

```r
mod_lasso <- logistic_reg(penalty = 0.001, mixture = 1) %>%
  set_engine("glmnet") %>%
  set_mode("classification") %>%
  fit(satisfaction ~ ., data = airtrain)

summary(mod_lasso)
```

```
##              Length Class        Mode
## lvl           2     -none-       character
## spec          8     logistic_reg list
## fit          13     lognet       list
## preproc       4     -none-       list
## elapsed       1     -none-       list
## censor_probs  0     -none-       list
```

```r
broom_lasso<-broom::tidy(mod_lasso)
broom_lasso[order(abs(broom_lasso$estimate),decreasing = TRUE),]
```

```
## # A tibble: 24 x 3
##    term                           estimate penalty
##    <chr>                             <dbl>   <dbl>
##  1 (Intercept)                      -7.74    0.001
##  2 type_of_travelPersonal Travel    -2.67    0.001
##  3 customer_typeLoyal Customer       1.95    0.001
##  4 classEco Plus                    -0.793   0.001
##  5 classEco                         -0.711   0.001
##  6 online_boarding                   0.600   0.001
##  7 inflight_wifi_service             0.368   0.001
##  8 checkin_service                   0.312   0.001
##  9 on_board_service                  0.295   0.001
## 10 leg_room_service                  0.247   0.001
## # i 14 more rows
```

```r
write.xlsx(broom_lasso[order(abs(broom_lasso$estimate),decreasing = TRUE),], "lasso_output.xlsx")
pred <- airtrain %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_lasso, new_data = airtrain, type = "class")
  ) %>%
  rename(satisfaction_null = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = satisfaction_null)
confusion
```

```
##           Truth
## Prediction     0     1
##          0 53131  7339
##          1  5566 37558
```

```
lasso_train_acc <- accuracy(pred, satisfaction, satisfaction_null)

###test####
pred <- airtest %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_lasso, new_data = airtest, type = "class")
  ) %>%
  rename(satisfaction_null = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = satisfaction_null)
confusion
```

```
##           Truth
## Prediction     0     1
##          0 13092  1893
##          1  1436  9472
```

```
lasso_test_acc <-accuracy(pred, satisfaction, satisfaction_null)
lasso_test_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.871
```

```
mod_lasso %>%
  predict(airtest, type = "prob") %>%
   bind_cols(airtest) %>%
   roc_auc(satisfaction, .pred_1,event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.926
```

```
predicted_probs <- predict(mod_lasso, type = "prob",new_data = airtrain) %>% dplyr::select(.pred_1) %>%

# Calculate AUC
roc_obj <- roc(airtrain$satisfaction, predicted_probs)
lasso_train_auc<- auc(roc_obj)


# Predict probabilities
predicted_probs <- predict(mod_lasso, type = "prob",new_data = airtest) %>% dplyr::select(.pred_1) %>% p
```

```r
# Calculate AUC
roc_obj <- roc(airtest$satisfaction, predicted_probs)
lasso_test_auc<- auc(roc_obj)
```

RIDGE

```r
mod_ridge <- logistic_reg(penalty = 0.001, mixture = 0) %>%
  set_engine("glmnet") %>%
  set_mode("classification") %>%
  fit(satisfaction ~ ., data = airtrain)

summary(mod_ridge)
```

```
##              Length Class        Mode
## lvl           2     -none-       character
## spec          8     logistic_reg list
## fit          13     lognet       list
## preproc       4     -none-       list
## elapsed       1     -none-       list
## censor_probs  0     -none-       list
```

```r
broom_ridge <-data.frame(broom::tidy(mod_ridge))
broom_ridge[order(abs(broom_ridge$estimate),decreasing = TRUE),]
```

```
##                                 term      estimate penalty
## 1                        (Intercept) -6.572837e+00   0.001
## 5       type_of_travelPersonal Travel -1.849045e+00   0.001
## 3         customer_typeLoyal Customer  1.302840e+00   0.001
## 6                            classEco -7.852932e-01   0.001
## 7                       classEco Plus -7.058586e-01   0.001
## 14                    online_boarding  4.683931e-01   0.001
## 9                inflight_wifi_service  2.842534e-01   0.001
## 20                    checkin_service  2.325140e-01   0.001
## 17                   on_board_service  2.195893e-01   0.001
## 18                   leg_room_service  2.086892e-01   0.001
## 22                        cleanliness  1.457418e-01   0.001
## 16             inflight_entertainment  1.291706e-01   0.001
## 10 departure_arrival_time_convenient -1.118574e-01   0.001
## 19                   baggage_handling  1.088633e-01   0.001
## 15                       seat_comfort  9.785189e-02   0.001
## 21                   inflight_service  9.604891e-02   0.001
## 2                          genderMale  4.087340e-02   0.001
## 11              ease_of_online_booking -3.362130e-02   0.001
## 12                      gate_location -6.905945e-03   0.001
## 13                     food_and_drink -5.112998e-03   0.001
## 24            arrival_delay_in_minutes -2.590738e-03   0.001
## 4                                 age -1.592116e-03   0.001
## 23          departure_delay_in_minutes -9.372022e-04   0.001
## 8                     flight_distance  8.277506e-05   0.001
```

```
write.xlsx(broom_ridge[order(abs(broom_ridge$estimate),decreasing = TRUE),], "ridge_output.xlsx")


pred <- airtrain %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_ridge, new_data = airtrain, type = "class")
  ) %>%
  rename(satisfaction_null = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = satisfaction_null)
confusion
```

```
##           Truth
## Prediction     0     1
##          0 53218  7685
##          1  5479 37212
```

```
ridge_train_acc <- accuracy(pred, satisfaction, satisfaction_null)

###test####
pred <- airtest %>%
  dplyr::select(satisfaction) %>%
  bind_cols(
    predict(mod_ridge, new_data = airtest, type = "class")
  ) %>%
  rename(satisfaction_null = .pred_class)

confusion <- pred %>%
  conf_mat(truth = 1, estimate = satisfaction_null)
confusion
```

```
##           Truth
## Prediction     0     1
##          0 13153  1971
##          1  1375  9394
```

```
ridge_test_acc <-accuracy(pred, satisfaction, satisfaction_null)
ridge_test_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.871
```

```
predicted_probs <- predict(mod_ridge, type = "prob",new_data = airtrain) %>% dplyr::select(.pred_1) %>%

# Calculate AUC
roc_obj <- roc(airtrain$satisfaction, predicted_probs)
ridge_train_auc<- auc(roc_obj)
```

```r
# Predict probabilities
predicted_probs <- predict(mod_ridge, type = "prob",new_data = airtest) %>% dplyr::select(.pred_1) %>% p

# Calculate AUC
roc_obj <- roc(airtest$satisfaction, predicted_probs)
ridge_test_auc<- auc(roc_obj)
```

Result for model perfomance and comparison

```r
c(
log_train_acc[,3],
lasso_train_acc[,3],
ridge_train_acc[,3],
dtree_train_acc[,3],
RF_train_acc,
xgb_train_acc[,3],
log_test_acc[,3],
lasso_test_acc[,3],
ridge_test_acc[,3],
dtree_test_acc[,3],
RF_test_acc,
xgb_test_acc[,3])
```

```
## $.estimate
## [1] 0.8751086
##
## $.estimate
## [1] 0.8754271
##
## $.estimate
## [1] 0.872927
##
## $.estimate
## [1] 0.8845493
##
## $.estimate
## [1] 0.9895554
##
## $.estimate
## [1] 0.9677298
##
## $.estimate
## [1] 0.8717028
##
## $.estimate
## [1] 0.8714324
##
## $.estimate
## [1] 0.8707759
```

```
##
## $.estimate
## [1] 0.883791
##
## $.estimate
## [1] 0.9631947
##
## $.estimate
## [1] 0.9633492
```

```r
c(
log_train_auc,
lasso_train_auc,
ridge_train_auc,
dtree_train_auc,
rf_train_auc,
xgb_train_auc,
log_test_auc,
lasso_test_auc,
ridge_test_auc,
dtree_test_auc,
rf_test_auc,
xgb_test_auc)
```

```
##  [1] 0.9268080 0.9268305 0.9254647 0.9040932 0.9997435 0.9964044 0.9255069
##  [8] 0.9255009 0.9236595 0.9035144 0.9937804 0.9949842
```