Predicting Airline Customer Satisfaction

Elif KARTAL

2024-05-31

Introduction

Predicting airline customer satisfaction enhances customer loyalty by increasing the likelihood of repeat business and aids in acquiring new customers through word-of-mouth marketing. High customer satisfaction provides a competitive advantage, increasing market share and supporting differentiation. It improves financial performance, generates additional revenue, and reduces operational costs. Moreover, it identifies operational inefficiencies and enhances service quality. It personalizes the customer experience and allows for proactive improvements by predicting potential dissatisfaction. It contributes to data-driven decision-making processes, facilitates the setting of strategic goals, and ensures compliance with regulatory requirements. Ultimately, it strengthens brand reputation and promotes sustainable growth.

Features

- **id :** Unique id number to each passenger. (numerical)
- **Gender:** Gender of the passengers (Female, Male) (categorical)
- **Customer Type:** The customer type (Loyal customer, disloyal customer) (categorical)
- **Age:** The actual age of the passengers (numerical)
- **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)

Features

- **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)
- **Departure/Arrival time convenient:** Satisfaction level of Departure/Arrival time convenient (numerical)
- **Ease of Online booking:** Satisfaction level of online booking (numerical)
- **Gate location:** Satisfaction level of Gate location (numerical)
- **Food and drink:** Satisfaction level of Food and drink (numerical)

Features

- **Online boarding:** Satisfaction level of online boarding (numerical)
- **Seat comfort:** Satisfaction level of Seat comfort (numerical)
- **Inflight entertainment:** Satisfaction level of inflight entertainment (numerical)
- **On-board service:** Satisfaction level of On-board service (numerical)
- **Leg room service:** Satisfaction level of Leg room service (numerical)
- **Baggage handling:** Satisfaction level of baggage handling (numerical)

Features

- **Check-in service:** Satisfaction level of Check-in service (numerical)
- **Inflight service:** Satisfaction level of inflight service (numerical)
- **Cleanliness:** Satisfaction level of Cleanliness (numerical)
- **Departure Delay in Minutes:** Minutes delayed when departure (numerical)
- **Arrival Delay in Minutes**: Minutes delayed when Arrival (numerical)
- **Satisfaction:** Airline satisfaction level(Satisfaction, neutral or dissatisfaction)(categorical)

Steps that I followed

- Importing Dataset.
- Loading needed packages.
- Checking and omitting missing values.
- Checking structure and summary statistics of the data.
- Splitting dataset into two parts as Train and Test.

Steps that I followed

- Checking the balance on the trainset.Steps that I followed
- Fixing the imbalance by using oversampling method.
- Training logistic regression model.
- Detecting non-significant variables by using glm.
- Removing non-significant variables in order to increase performance.

Steps that I followed

- Training RandomForests.
- Training Decision Tree.
- Comparing performances and interpretablity among models.
- Interpreting Decision Tree.
- Conclusion.

Checking the missing value

```
## [1] TRUE
```

```
## [1] FALSE
```

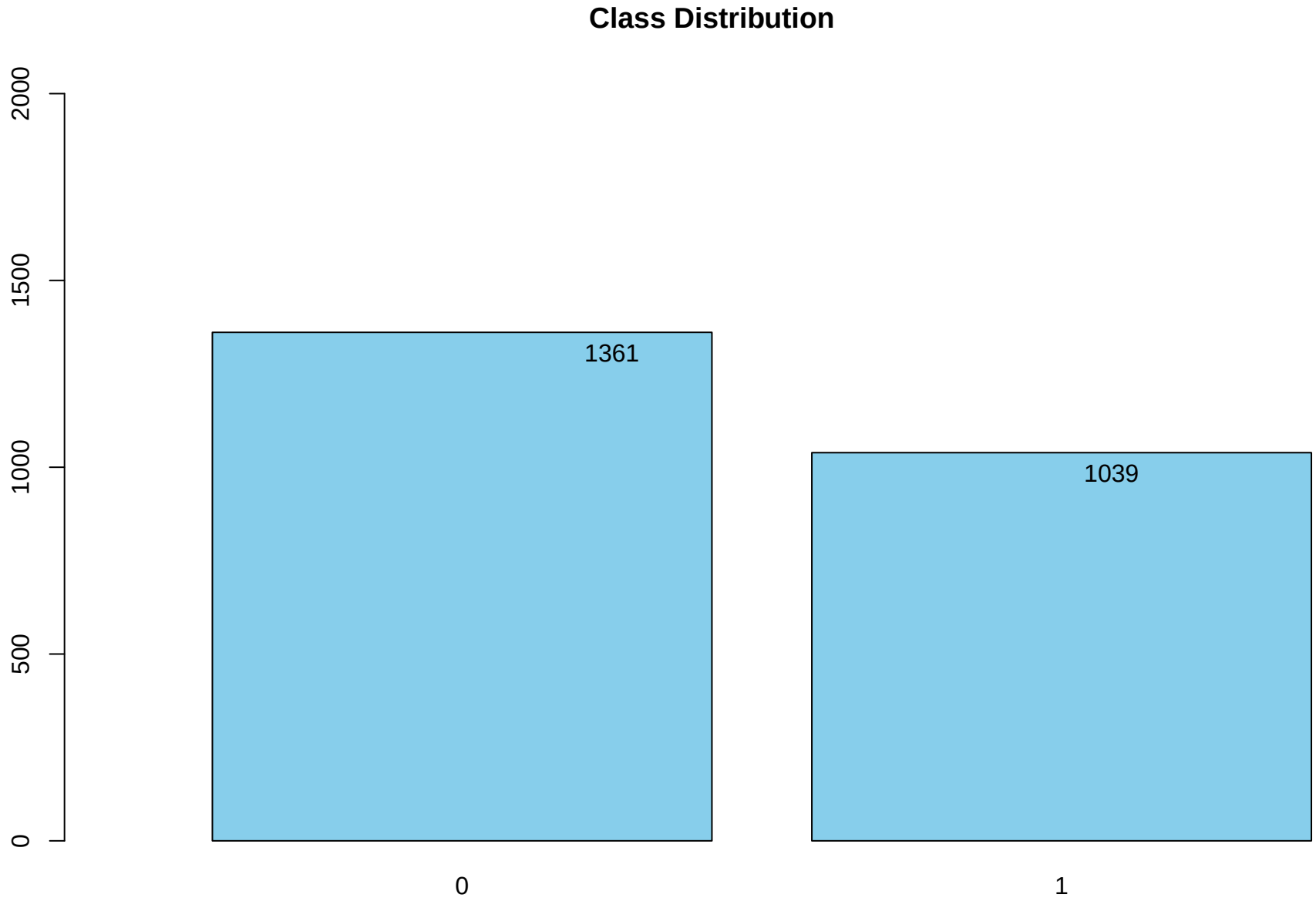
Checking balance of target variable on the traindata

Class distributions and class proportions:

```
##
##    0    1
## 1361 1039
```

```
##
##         0         1
## 0.5670833 0.4329167
```

Visualization of Target Variable



Fixing the imbalance:

In order to fix imbalance problem, oversampling method is used.

Balanced class proportions:

```
##
##   0   1
## 0.5 0.5
```

The number of observations in the classes is equalized and the imbalance problem is solved.

Logistic Regression Model

```
Call:
glm(formula = as.numeric(target) ~ ., family = "binomial", data = balanced_train)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -7.692e+00  5.116e-01 -15.037  < 2e-16 ***
id                              -3.545e-06  1.719e-06  -2.063 0.039143 *
GenderMale                       2.823e-02  1.247e-01   0.226 0.820948
`Customer Type`Loyal Customer    2.348e+00  1.925e-01  12.199  < 2e-16 ***
Age                             -5.309e-03  4.670e-03  -1.137 0.255652
TypeofTravelPersonal Travel     -3.144e+00  2.063e-01 -15.239  < 2e-16 ***
ClassEco                        -6.391e-01  1.669e-01  -3.828 0.000129 ***
ClassEco Plus                   -2.492e-01  2.601e-01  -0.958 0.338089
FlightDistance                  -9.077e-05  7.293e-05  -1.245 0.213269
Inflightwifiservice              5.324e-01  7.219e-02   7.375 1.64e-13 ***
`Departure/Arrivaltimeconvenient` -2.611e-01  5.170e-02  -5.051 4.40e-07 ***
EaseofOnlinebooking             -2.842e-01  7.079e-02  -4.015 5.94e-05 ***
Gatelocation                     6.773e-02  5.744e-02   1.179 0.238345
Foodanddrink                     4.034e-02  6.612e-02   0.610 0.541807
Onlineboarding                   5.498e-01  6.410e-02   8.578  < 2e-16 ***
Seatcomfort                     -1.041e-01  7.502e-02  -1.387 0.165383
Inflightentertainment           -2.970e-03  9.199e-02  -0.032 0.974242
`On-boardservice`                4.399e-01  6.662e-02   6.604 4.01e-11 ***
Legroomservice                   2.456e-01  5.526e-02   4.444 8.82e-06 ***
Baggagehandling                  1.925e-01  7.505e-02   2.565 0.010314 *
Checkinservice                   4.052e-01  5.654e-02   7.166 7.69e-13 ***
Inflightservice                  8.069e-02  7.631e-02   1.057 0.290331
Cleanliness                      3.324e-01  7.616e-02   4.364 1.27e-05 ***
DepartureDelayinMinutes         -1.088e-02  6.496e-03  -1.675 0.093954 .
ArrivalDelayinMinutes            2.300e-03  6.405e-03   0.359 0.719527
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression Model

By looking at the coefficients of the variables, it was decided to remove these variables to increase the performance of the model:

- id
- TypeofTravel
- Inflightentertainment
- FlightDistance

Logistic Regression Model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 332  47
         1  28 193

               Accuracy : 0.875
                 95% CI : (0.8458, 0.9004)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7361

 Mcnemar's Test P-Value : 0.03767

            Sensitivity : 0.9222
            Specificity : 0.8042
         Pos Pred Value : 0.8760
         Neg Pred Value : 0.8733
             Prevalence : 0.6000
         Detection Rate : 0.5533
   Detection Prevalence : 0.6317
      Balanced Accuracy : 0.8632

       'Positive' Class : 0
```

Random Forests

```
Random Forest

2722 samples
  19 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2450, 2450, 2450, 2450, 2449, 2449, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  1     0.9129323  0.8258728
  2     0.9364455  0.8728965
  3     0.9397517  0.8795057
  4     0.9404897  0.8809834

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4.
```

Predicting RandomForests Model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 342  32
         1  18 208

               Accuracy : 0.9167
                 95% CI : (0.8916, 0.9375)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8247

 Mcnemar's Test P-Value : 0.06599

            Sensitivity : 0.9500
            Specificity : 0.8667
         Pos Pred Value : 0.9144
         Neg Pred Value : 0.9204
             Prevalence : 0.6000
         Detection Rate : 0.5700
   Detection Prevalence : 0.6233
      Balanced Accuracy : 0.9083

       'Positive' Class : 0
```

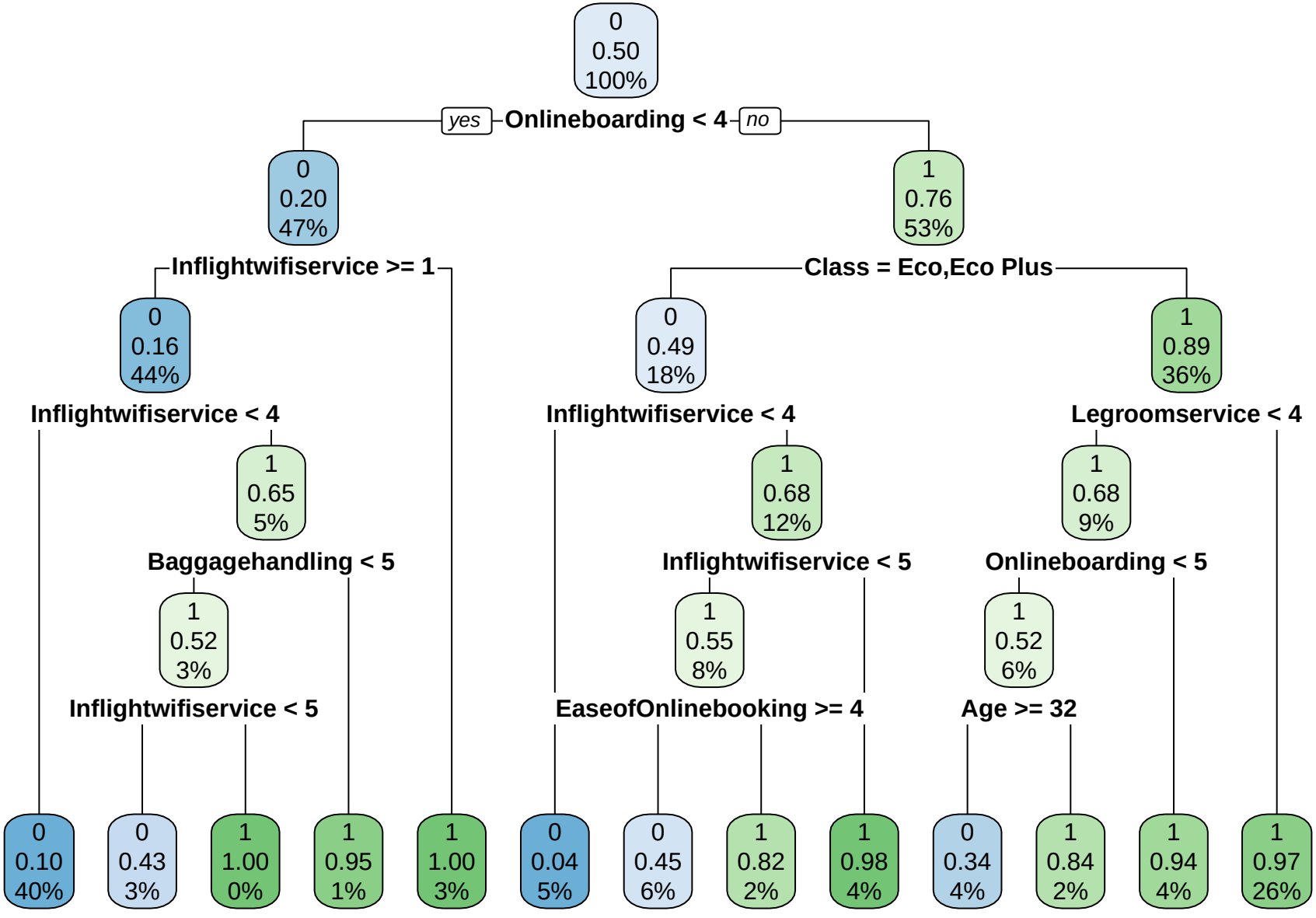
Decision Tree

```
> grid <- expand.grid(cp = seq(0,0.1,0.01))
> control <- trainControl(method = "cv", number = 10, search = "grid")
> set.seed(333)
> dt_model_tuned <- train(target ~ .,
+                         data = balanced_train,
+                         method = "rpart",
+                         tuneGrid = grid,
+                         trControl = control)
Something is wrong; all the Accuracy metric values are missing:
    Accuracy       Kappa
 Min.   : NA   Min.   : NA
 1st Qu.: NA   1st Qu.: NA
 Median : NA   Median : NA
 Mean   :NaN   Mean   :NaN
 3rd Qu.: NA   3rd Qu.: NA
 Max.   : NA   Max.   : NA
 NA's   :11    NA's   :11
Error: Stopping
In addition: There were 11 warnings (use warnings() to see them)
> warnings()
Warning messages:
1: model fit failed for Fold01: cp=0 Error in `[.data.frame`(m, labs) : undefined columns selected

2: model fit failed for Fold02: cp=0 Error in `[.data.frame`(m, labs) : undefined columns selected

3: model fit failed for Fold03: cp=0 Error in `[.data.frame`(m, labs) : undefined columns selected

4: model fit failed for Fold04: cp=0 Error in `[.data.frame`(m, labs) : undefined columns selected

5: model fit failed for Fold05: cp=0 Error in `[.data.frame`(m, labs) : undefined columns selected

6: model fit failed for Fold06: cp=0 Error in `[.data.frame`(m, labs) : undefined columns selected
```

Decision Tree

```
+ }
> # Bayesian optimizasyon
> opt_result <- BayesianOptimization(opt_function,
+                                    bounds = list(cp = c(0, 0.1)),
+                                    init_points = 10,
+                                    n_iter = 30)
Error: Stopping
In addition: There were 11 warnings (use warnings() to see them)
Timing stopped at: 0.33 0.01 0.39
> warnings()
Warning messages:
1: model fit failed for Fold01: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

2: model fit failed for Fold02: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

3: model fit failed for Fold03: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

4: model fit failed for Fold04: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

5: model fit failed for Fold05: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

6: model fit failed for Fold06: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

7: model fit failed for Fold07: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

8: model fit failed for Fold08: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

9: model fit failed for Fold09: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

10: model fit failed for Fold10: cp=0.0467 Error in `[.data.frame`(m, labs) : undefined columns selected

11: In nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,  ... :
  There were missing values in resampled performance measures.
```

Decision Tree



Predicting Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 343  45
         1  17 195

               Accuracy : 0.8967
                 95% CI : (0.8695, 0.9199)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7805

 Mcnemar's Test P-Value : 0.0006058

            Sensitivity : 0.9528
            Specificity : 0.8125
         Pos Pred Value : 0.8840
         Neg Pred Value : 0.9198
             Prevalence : 0.6000
         Detection Rate : 0.5717
   Detection Prevalence : 0.6467
      Balanced Accuracy : 0.8826

       'Positive' Class : 0
```

Comparing Models

- Although random forests and logistic regression models have high-performance, they are disadvantageous in terms of interpretability. For this reason, it was decided to use the decision trees model, which provides easy interpretability.

General Interpretation

- **Onlineboarding**: Serving as the first split, the online boarding process significantly impacts customer satisfaction. Customers with lower online boarding scores are generally dissatisfied.
- **Inflightwifiservice**: The second most important factor is the quality of in-flight Wi-Fi service, which greatly influences satisfaction.
- **Class**: The flight class (economy or economy plus) also affects satisfaction, with economy class customers typically showing more dissatisfaction.
- **Baggagehandling**: The quality of baggage handling is another determinant of customer satisfaction.
- **EaseofOnlinebooking**: The ease of online booking is particularly important for customers with certain Wi-Fi service scores.
- **Age**: The age of the customer can also impact satisfaction in some scenarios.

Recommendations to Improve Customer Satisfaction

1. **Improve Online Boarding Process**: Making the online boarding process more user-friendly can enhance customer satisfaction.
2. **Enhance In-Flight Wi-Fi Service**: Improving the quality of Wi-Fi service can significantly boost satisfaction levels.
3. **Streamline Baggage Handling**: Efficient and smooth baggage handling processes can enhance customer satisfaction.
4. **Simplify Online Booking System**: Making the booking process easier can improve satisfaction, especially for those who rely on the internet.
5. **Improve Legroom Service**: Enhancing legroom service quality can particularly boost satisfaction for specific customer segments.

These insights can help the airline company identify areas for improvement to enhance overall customer satisfaction.