Airline Passenger Satisfaction - Random Forest Exploration

Leif Berg

2023-04-29

About

This script is an exploration of using a random forest to train and classify satisfaction ratings of airline passengers.

Data source: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction)

Setup

Read in Data

Read in Training Data

```
training_df <- read.csv("train.csv")
```

Read in Testing Data

```
test_df <- read.csv("test.csv")
```

View structure of data

```
str(training_df)
```

```
## '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 65725 ...
##  $ Gender                          : chr  "Male" "Male" "Female" "Female" ...
##  $ Customer.Type                   : chr  "Loyal Customer" "disloyal Customer" "Loyal Custom
er" "Loyal Customer" ...
##  $ Age                             : int  13 25 26 25 61 26 47 52 41 20 ...
##  $ Type.of.Travel                  : chr  "Personal Travel" "Business travel" "Business trav
el" "Business travel" ...
##  $ Class                           : chr  "Eco Plus" "Business" "Business" "Business" ...
##  $ 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                    : chr  "neutral or dissatisfied" "neutral or dissatisfie
d" "satisfied" "neutral or dissatisfied" ...
```

View first few rows of data

```
head(training_df)
```

```
##   X     id Gender     Customer.Type Age  Type.of.Travel    Class
## 1 0  70172   Male    Loyal Customer  13 Personal Travel Eco Plus
## 2 1   5047   Male disloyal Customer  25 Business travel Business
## 3 2 110028 Female    Loyal Customer  26 Business travel Business
## 4 3  24026 Female    Loyal Customer  25 Business travel Business
## 5 4 119299   Male    Loyal Customer  61 Business travel Business
## 6 5 111157 Female    Loyal Customer  26 Personal Travel      Eco
##   Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
## 1             460                     3                                 4
## 2             235                     3                                 2
## 3            1142                     2                                 2
## 4             562                     2                                 5
## 5             214                     3                                 3
## 6            1180                     3                                 4
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1                      3             1              5               3
## 2                      3             3              1               3
## 3                      2             2              5               5
## 4                      5             5              2               2
## 5                      3             3              4               5
## 6                      2             1              1               2
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1            5                      5                4                3
## 2            1                      1                1                5
## 3            5                      5                4                3
## 4            2                      2                2                5
## 5            5                      3                3                4
## 6            1                      1                3                4
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1                4               4                5           5
## 2                3               1                4           1
## 3                4               4                4           5
## 4                3               1                4           2
## 5                4               3                3           3
## 6                4               4                4           1
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes            satisfaction
## 1                         25                       18 neutral or dissatisfied
## 2                          1                        6 neutral or dissatisfied
## 3                          0                        0               satisfied
## 4                         11                        9 neutral or dissatisfied
## 5                          0                        0               satisfied
## 6                          0                        0 neutral or dissatisfied
```

Clean data

Remove any rows with NA (for simplicty) #remove na values (any row with na)

```
training_df <- na.omit(training_df)
test_df <- na.omit(test_df)
```

Ensure satisfaction is a factor

```
training_df$satisfaction <- as.factor(training_df$satisfaction)
test_df$satisfaction <- as.factor(test_df$satisfaction)
```

Model - Random Forest

Create Random Forest (limit items to ensure it finishes)

```
rf <- randomForest(satisfaction ~., data=training_df[1:5000,])
rf
```

```
##
## Call:
##  randomForest(formula = satisfaction ~ ., data = training_df[1:5000,      ])
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 5.72%
## Confusion matrix:
##                          neutral or dissatisfied satisfied class.error
## neutral or dissatisfied                    2716       107  0.03790294
## satisfied                                   179      1998  0.08222324
```

Create Predictions using random forest against test data

```
pred <- predict(rf, test_df)
head(pred)
```

```
##                       1                       2                       3
##               satisfied               satisfied neutral or dissatisfied
##                       4                       5                       6
##               satisfied neutral or dissatisfied               satisfied
## Levels: neutral or dissatisfied satisfied
```

Create Confusion Matrix to display results

```
confusionMatrix(pred, test_df$satisfaction,positive='satisfied')
```

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                neutral or dissatisfied satisfied
##   neutral or dissatisfied                   13975       864
##   satisfied                                   553     10501
##
##                Accuracy : 0.9453
##                  95% CI : (0.9424, 0.948)
##     No Information Rate : 0.5611
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8886
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9240
##             Specificity : 0.9619
##          Pos Pred Value : 0.9500
##          Neg Pred Value : 0.9418
##              Prevalence : 0.4389
##          Detection Rate : 0.4056
##    Detection Prevalence : 0.4269
##       Balanced Accuracy : 0.9430
##
##        'Positive' Class : satisfied
##
```