ALY 6015: Intermediate Analytics CRN 81176 Spring 2022

Project: Airline Passenger Satisfaction

Prof. Behzad Ahmadi
CPS Northeastern University

By Sourabh Khot
NUID 002754952

Introduction

- 'Airline Passenger Satisfaction | Kaggle' is a survey data collected by an airline
- Contains passenger details, flight characteristics, overall satisfaction
- 129,880 rows and 24 variables
  - 5 categorical (including 1 target)
  - 19 numerical (1 identifier, 14 discrete on ritcher scale, 4 continuous)
- 'satisfaction' is a binary outcome variable, indicating a passenger's overall satisfaction with the airline
- Final goal is to build a classification model to predict 'satisfaction'

id	gender	customer type	age	type of travel	customer class	flight distance	inflight wifi service	departure arrival time convenient
0	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4
1	Male	disloyal Customer	25	Business travel	Business	235	3	2
2	Female	Loyal Customer	26	Business travel	Business	1142	2	2
3	Female	Loyal Customer	25	Business travel	Business	562	2	5
4	Male	Loyal Customer	61	Business travel	Business	214	3	3

departure arrival time convenient	ease of online booking	gate location	food and drink	online boarding	seat comfort	inflight entertainment	onboard service
4	3	1	5	3	5	5	4
2	3	3	1	3	1	1	1
2	2	2	5	5	5	5	4
5	5	5	2	2	2	2	2
3	3	3	4	5	5	3	3

leg room service	baggage handling	checkin service	inflight service	cleanliness	departure delay in minutes	arrival delay in minutes	satisfaction
3	4	4	5	5	25	18	not satisfied
5	3	1	4	1	1	6	not satisfied
3	4	4	4	5	0	0	satisfied
5	3	1	4	2	11	9	not satisfied
4	4	3	3	3	0	0	satisfied

Objective

Major Questions

- What factors are highly correlated to passenger satisfaction?
- How to predict whether a new passenger will be satisfied with the airline's service?

- Minor Questions
  - Is Arrival Delay dependent upon flight distance?
  - Does Type of travel (personal/business) depends upon flight class (eco/eco plus/business)?
  - Does Seat comfort depends on

Methods to be employed

- EDA
- Hypothesis Testing
- Feature selection using Logit p-values, Lasso Regression
- Binary classification / logistic regression model to predict customer satisfaction

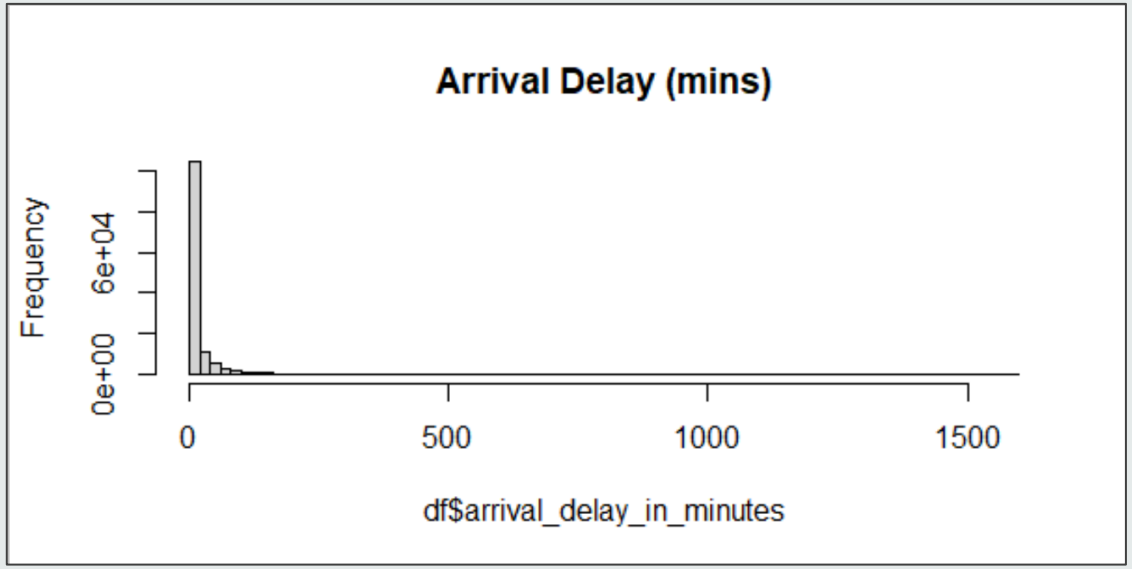
Data Preparation

- Categorical variables:
  - gender, customer type, type of travel, customer class, satisfaction
  - Converted to factors (with ordered levels for customer class)

- Missing values:
  - 'arrival delay in minutes' has 393 (0.3%) missing values
  - Half-normal distribution having 72753 (56%) values as zero
  - Imputed with median value (zero)

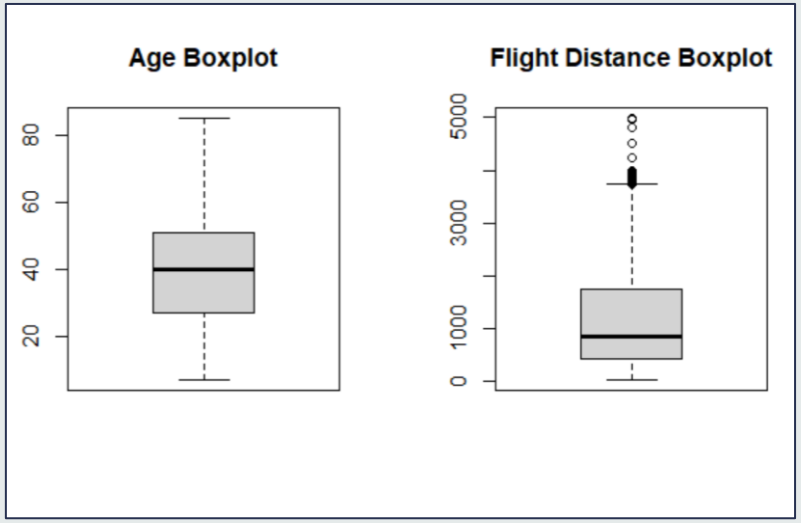
```
arrival_delay_in_minutes
Min.    :   0.00
1st Qu.:   0.00
Median :   0.00
Mean   :  15.09
3rd Qu.:  13.00
Max.   :1584.00
NA's   :393
```

```
> lapply(df[,c(2,3,5,6,24)], unique)
$gender
[1] "Male"   "Female"

$customer_type
[1] "Loyal Customer"    "Disloyal Customer"

$type_of_travel
[1] "Personal Travel" "Business travel"

$customer_class
[1] "Eco Plus" "Business" "Eco"

$satisfaction
[1] "Not Satisfied" "Satisfied"
```

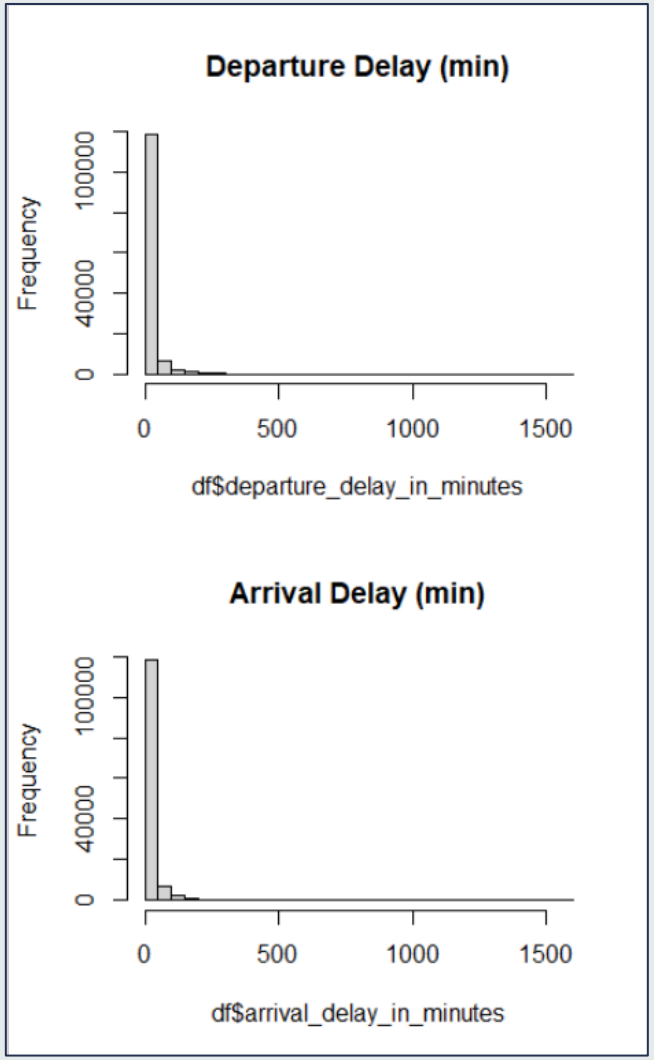


Variables' Distribution

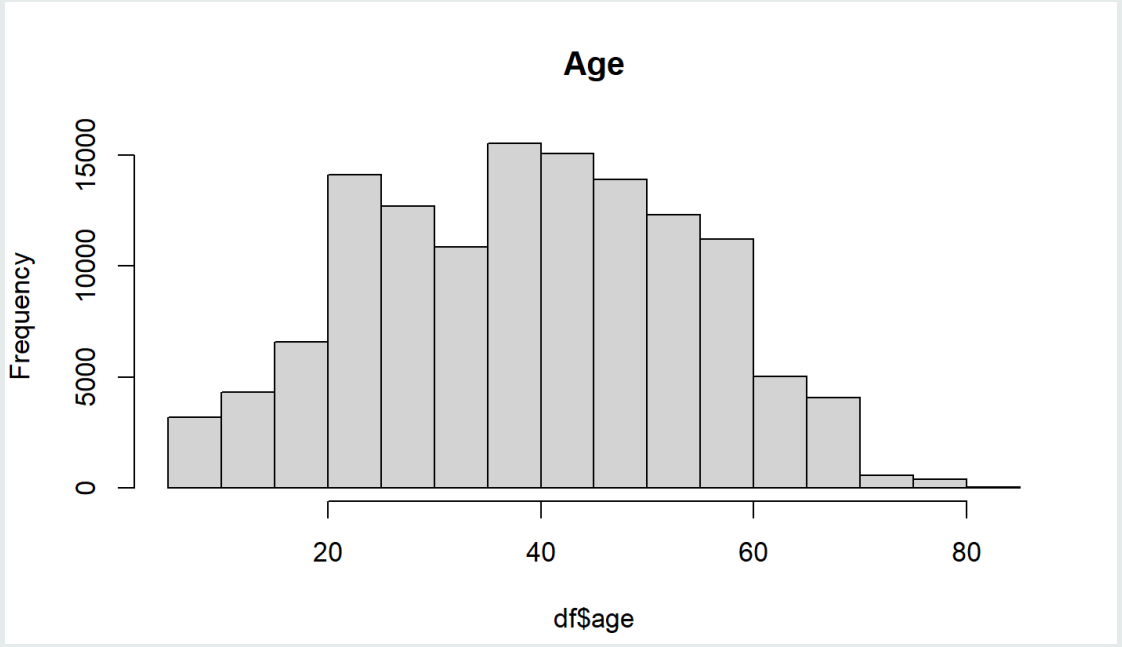
- 'Departure Delay' and 'Arrival Delay' are half-normal distribution with long right tail
- 'Flight distance' has outliers but they are explainable, hence not removed
- Ritcher scale attributes
  - Best (>3.5): inflight service, baggage handling
  - Worst (<3): Gate location, ease of online booking, inflight wifi service




```
> richter[order(-richter$mean), , drop = FALSE]
                                  mean
inflight_service                  3.64
baggage_handling                  3.63
seat_comfort                      3.44
onboard_service                   3.38
inflight_entertainment            3.36
leg_room_service                  3.35
checkin_service                   3.31
cleanliness                       3.29
online_boarding                   3.25
food_and_drink                    3.20
departure_arrival_time_convenient 3.06
gate_location                     2.98
ease_of_online_booking            2.76
inflight_wifi_service             2.73
```



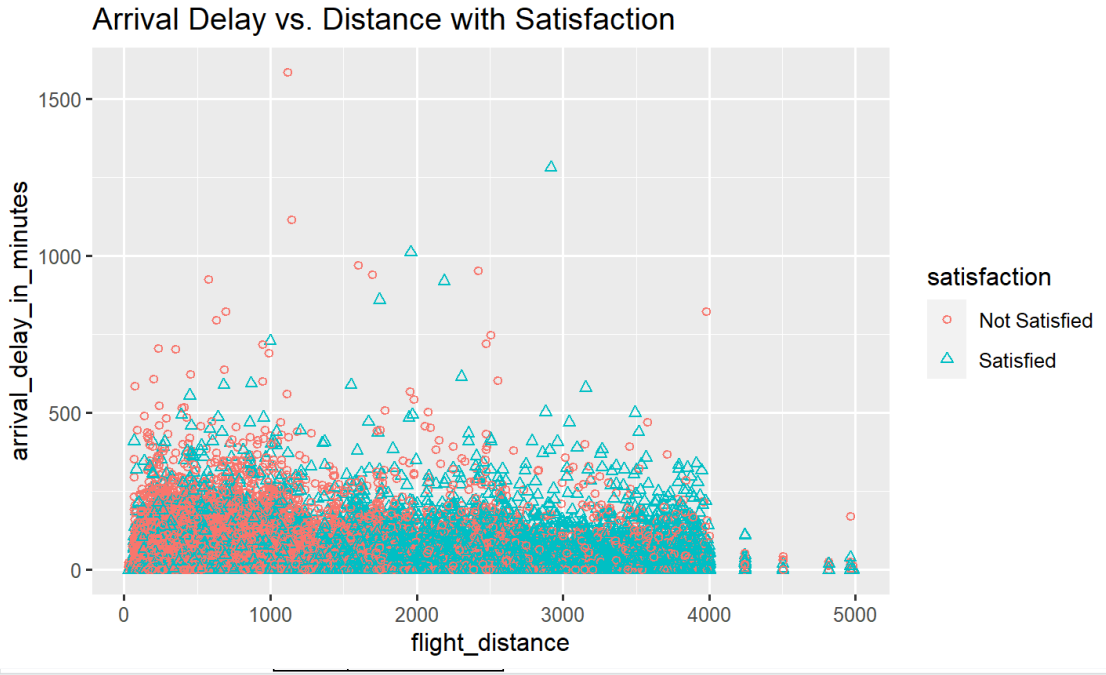
More Variable Distribution

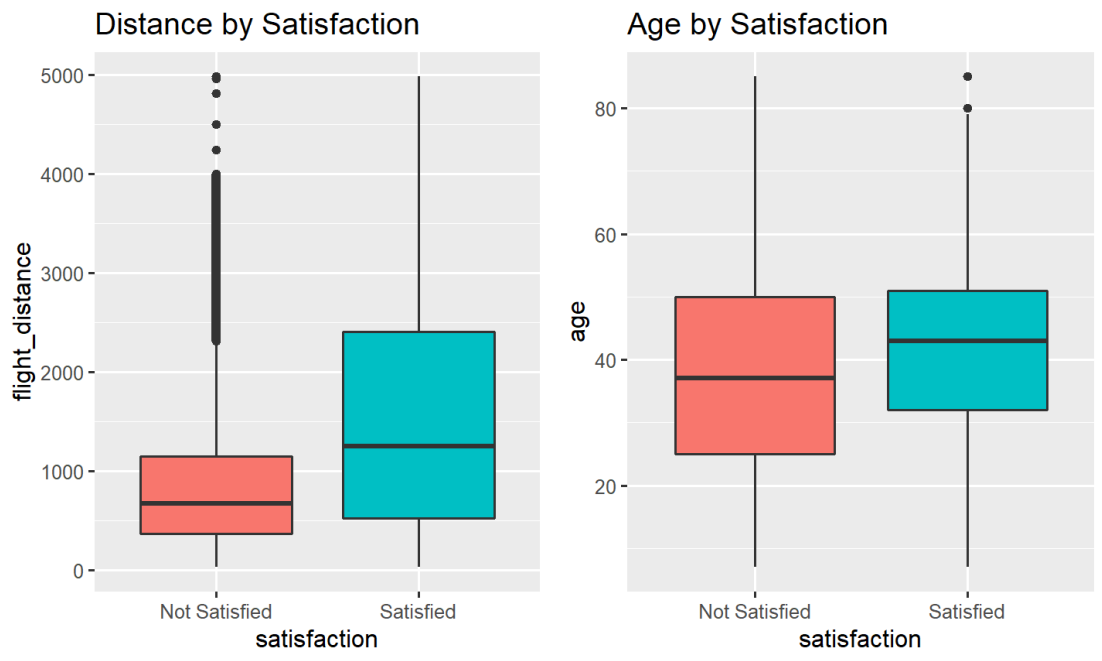



```
> summary(df2)
   gender           customer_type        age           type_of_travel   customer_class
Female:65899   Disloyal Customer: 23780   Min.   : 7.00   Business travel:89693   Eco     :62160
Male  :63981   Loyal Customer   :106100   1st Qu.:27.00   Personal Travel:40187   Eco Plus:58309
                                          Median :40.00                           Business: 9411
                                          Mean   :39.43
                                          3rd Qu.:51.00
                                          Max.   :85.00

 flight_distance inflight_wifi_service departure_arrival_time_convenient ease_of_online_booking
 Min.   :  31    Min.   :0.000         Min.   :0.000                     Min.   :0.000
 1st Qu.: 414    1st Qu.:2.000         1st Qu.:2.000                     1st Qu.:2.000
 Median : 844    Median :3.000         Median :3.000                     Median :3.000
 Mean   :1190    Mean   :2.729         Mean   :3.058                     Mean   :2.757
 3rd Qu.:1744    3rd Qu.:4.000         3rd Qu.:4.000                     3rd Qu.:4.000
 Max.   :4983    Max.   :5.000         Max.   :5.000                     Max.   :5.000

 gate_location   food_and_drink  online_boarding  seat_comfort   inflight_entertainment onboard_service
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000          Min.   :0.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000          1st Qu.:2.000
 Median :3.000   Median :3.000   Median :3.000   Median :4.000   Median :4.000          Median :4.000
 Mean   :2.977   Mean   :3.205   Mean   :3.253   Mean   :3.441   Mean   :3.358          Mean   :3.383
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000          3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000          Max.   :5.000

 leg_room_service baggage_handling checkin_service inflight_service  cleanliness
 Min.   :0.000    Min.   :1.000    Min.   :0.000   Min.   :0.000    Min.   :0.000
 1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.000    1st Qu.:2.000
 Median :4.000    Median :4.000    Median :3.000   Median :4.000    Median :3.000
 Mean   :3.351    Mean   :3.632    Mean   :3.306   Mean   :3.642    Mean   :3.286
 3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:5.000    3rd Qu.:4.000
 Max.   :5.000    Max.   :5.000    Max.   :5.000   Max.   :5.000    Max.   :5.000

 departure_delay_in_minutes arrival_delay_in_minutes        satisfaction
 Min.   :   0.00            Min.   :   0.00            Not Satisfied:73452
 1st Qu.:   0.00            1st Qu.:   0.00            Satisfied    :56428
 Median :   0.00            Median :   0.00
 Mean   :  14.71            Mean   :  15.05
 3rd Qu.:  12.00            3rd Qu.:  13.00
 Max.   :1592.00            Max.   :1584.00
```

Scatterplot, Boxplot

- Are long distance flights more delayed?
  - No, short flights slightly more delayed
- Segregate by Satisfaction level
  - More pax not satisfied with shorter flights
- Satisfaction levels vs. Distance and Age
  - More pax satisfied with longer flights
  - Younger pax are not satisfied





Hypothesis Testing <1/3>

Correlation between Flight Distance & Arrival Delay

- Ho: ρ = 0
- H1: ρ ≠ 0

Result: Not enough evidence to reject Ho: no correlation

```
        Pearson's product-moment correlation

data:  df2$arrival_delay_in_minutes and df2$flight_distance
t = -0.711, df = 129878, p-value = 0.4771
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.007411295  0.003465626
sample estimates:
         cor
-0.001972893
```

Dependence betw. Class of Travel and Satisfaction

- Chi-square Test of Independence
- Ho: Satisfaction proportion is independent of travel class
- H1: Satisfaction proportion differs on travel class

Result: Satisfaction dependent upon class of travel

	Not Satisfied	Satisfied
Eco	18994	43166
Eco Plus	47366	10943
Business	7092	2319

```
        Pearson's Chi-squared test

data:  ctable
X-squared = 32906, df = 2, p-value < 0.00000000000000022

> ifelse(result1$p.value > alpha, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Sufficient evidence to reject H0"
```

Hypothesis Testing <2/3>

Seat Comfort depends on Travel Class

VARS	IV
online_boarding	1.8521
inflight_wifi_service	1.6215
customer_class	1.1312
type_of_travel	1.0492
inflight_entertainment	0.8015
seat_comfort	0.6554
leg_room_service	0.4939
onboard_service	0.4831
cleanliness	0.4498
flight_distance	0.4234
ease_of_online_booking	0.3680
age	0.3640
baggage_handling	0.3554
inflight_service	0.3404
checkin_service	0.2690
food_and_drink	0.2264
customer_type	0.1538
arrival_delay_in_minutes	0.0515
departure_delay_in_minutes	0.0208
departure_arrival_time_convenient	0.0191
gender	0.0004
gate_location	0.0000

- One-Way ANOVA
- Ho: 3 means are same
- H1: At least 1 is different

Result: Reject Ho: At least one is different

Seat Comfort

Eco	Eco Plus	Business
3	2	3
1	1	5
4		2
		3

```
                  Df Sum Sq Mean Sq F value               Pr(>F)
customer_class     2  12393    6197    3767 <0.0000000000000002 ***
Residuals     129877 213664       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ifelse(result2[[1]][[1,"Pr(>F)"]] > alpha, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Sufficient evidence to reject H0"
```

Scheffe Test to find which different Mean

- Not Tukey since sample sizes are different

Result: We find that Eco has different mean, while Business and Eco Plus may have same means

```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$customer_class
                          diff       lwr.ci      upr.ci                  pval
Eco Plus-Eco      -0.62189312 -0.639993423 -0.60379282 <0.0000000000000002 ***
Business-Eco      -0.59549915 -0.630226095 -0.56077220 <0.0000000000000002 ***
Business-Eco Plus  0.02639397 -0.008483437  0.06127139                 0.1798

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Testing <3/3>

Seat Comfort (1-5)		Travel Class (A)		
		Eco	Eco Plus	Business
Travel Purpose (B)	Personal	...	...	...
	Business	...	...	...

Two-way ANOVA to test if seat comfort depends on Travel Class (A) and Travel Purpose (B)

- Ho, A*B: There is NO interaction between travel class and travel purpose
- H1, A*B: There is significant interaction between travel class and travel purpose
- Ho, A: Seat comfort ratings of the three travel classes are equal
- H1, A: Seat comfort ratings at least one travel class is different
- Ho, B: Seat comfort ratings of two travel purposes are equal
- H1, B: Seat comfort ratings of two travel purposes are different

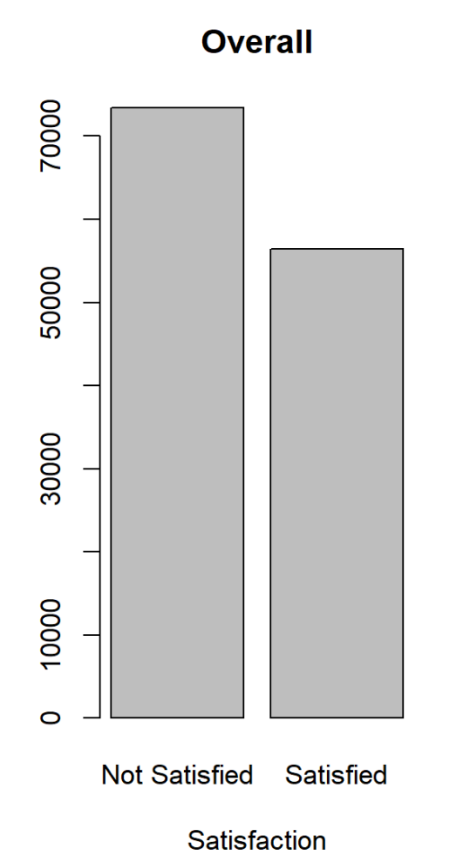
Results:

- There is significant interaction between travel class and travel purpose
- Seat comfort ratings of at least one travel class (Eco, Eco Plus, Business) is different
- Seat comfort ratings of two travel purposes (Personal, Business) are equal

```
                                Df Sum Sq Mean Sq  F value              Pr(>F)
customer_class                   2  12393    6197 3771.335 <0.0000000000000002 ***
type_of_travel                   1      1       1    0.626               0.429
customer_class:type_of_travel    2    273     137   83.195 <0.0000000000000002 ***
Residuals                   129874 213390       2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ifelse(result3[[1]][[3,"Pr(>F)"]] > alpha, "Not enough evidence to reject H0.A*B",
+        "Sufficient evidence to reject H0.A*B")
[1] "Sufficient evidence to reject H0.A*B"
> ifelse(result3[[1]][[1,"Pr(>F)"]] > alpha, "Not enough evidence to reject H0.A",
+        "Sufficient evidence to reject H0.A")
[1] "Sufficient evidence to reject H0.A"
> ifelse(result3[[1]][[2,"Pr(>F)"]] > alpha, "Not enough evidence to reject H0.B",
+        "Sufficient evidence to reject H0.B")
[1] "Not enough evidence to reject H0.B"
```

Splitting data for cross validation

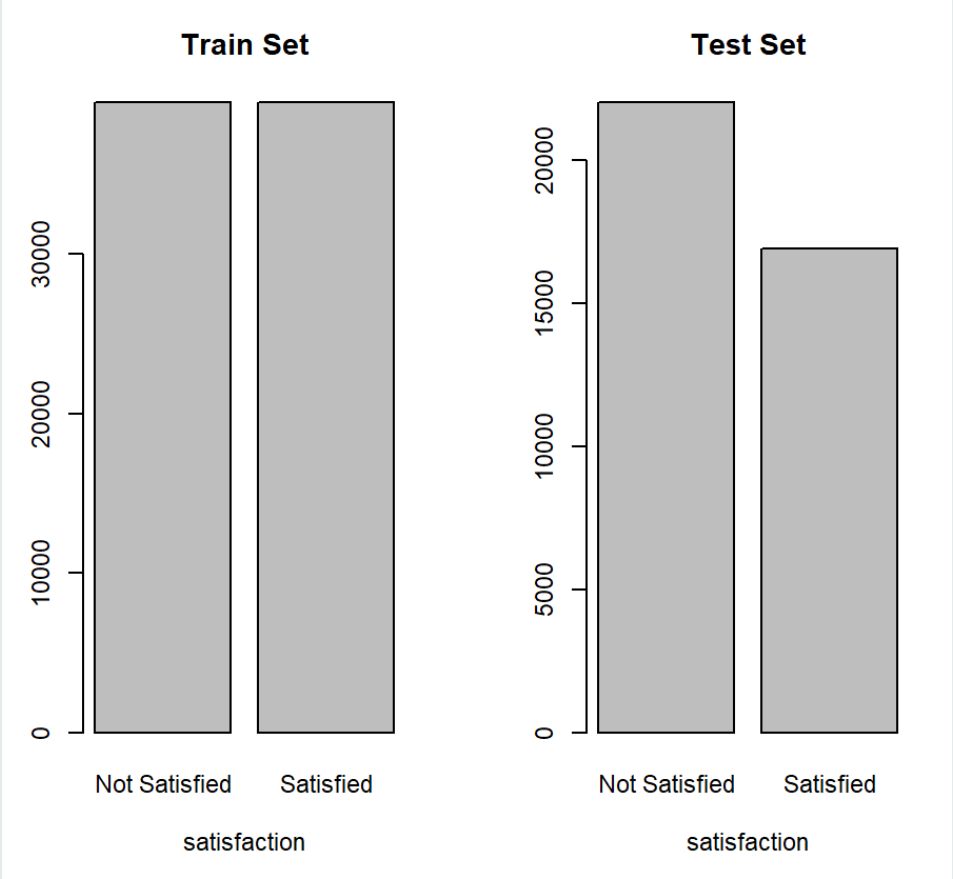
Checking class bias



- Proportion of 'Not satisfied' class is 30.2% more than 'satisfied'
- We can sample a balanced training data by taking equal number of classes, based on 70% of the less frequent class

```
> # checking class bias
> table(df2$satisfaction)

Not Satisfied     Satisfied
        73452         56428
```

Test vs Train Set



Model 1: Feature Selection using Logit p-values, Multicollinearity

- Initially, fitted a logistic regression model with all 22 input variables
- Iteratively, removed highest p-value features. Similarly, iteratively removed features with high multicollinearity
- Removed 5 variables:
  - Flight distance, gate location, departure delay, food and drink, inflight entertainment
- Probability threshold taken as 0.5 to predict 'Satisfied

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9458  -0.5056   0.0363   0.4103   3.8869

Coefficients:
                                  Estimate  Pr(>|z|)
(Intercept)                     -7.273e+00   < 2e-16 ***
genderMale                       6.655e-02  0.002422 **
customer_typeLoyal Customer      1.998e+00   < 2e-16 ***
age                             -8.434e-03   < 2e-16 ***
type_of_travelPersonal Travel   -2.651e+00   < 2e-16 ***
customer_classEco Plus          -7.189e-01   < 2e-16 ***
customer_classBusiness          -7.962e-01   < 2e-16 ***
flight_distance                 -2.085e-05  0.104563
inflight_wifi_service            4.302e-01   < 2e-16 ***
departure_arrival_time_convenient -1.315e-01  < 2e-16 ***
ease_of_online_booking          -1.747e-01   < 2e-16 ***
gate_location                    2.581e-02  0.011474 *
food_and_drink                  -3.415e-02  0.004941 **
online_boarding                  5.885e-01   < 2e-16 ***
seat_comfort                     5.326e-02  1.79e-05 ***
inflight_entertainment           5.774e-02  0.000283 ***
onboard_service                  2.877e-01   < 2e-16 ***
leg_room_service                 2.355e-01   < 2e-16 ***
baggage_handling                 1.241e-01   < 2e-16 ***
checkin_service                  3.182e-01   < 2e-16 ***
inflight_service                 1.257e-01   < 2e-16 ***
cleanliness                      2.236e-01   < 2e-16 ***
departure_delay_in_minutes       3.193e-03  0.001957 **
arrival_delay_in_minutes        -7.730e-03  4.66e-14 ***
    Null deviance: 109514  on 78997  degrees of freedom
Residual deviance:  54373  on 78974  degrees of freedom
AIC: 54421
```

Removed 3 variables with high p-values

	GVIF
gender	1.006943
customer_type	1.505452
age	1.176277
type_of_travel	1.838374
customer_class	1.497517
inflight_wifi_service	2.348562
departure_arrival_time_convenient	1.513685
ease_of_online_booking	2.443940
food_and_drink	2.076887
online_boarding	1.449473
seat_comfort	2.048620
inflight_entertainment	3.295040
onboard_service	1.630474
leg_room_service	1.208632
baggage_handling	1.817230
checkin_service	1.201877
inflight_service	1.981584
cleanliness	2.527977
departure_delay_in_minutes	12.776028
arrival_delay_in_minutes	12.793927

Removed 2variables with high VIFs

Model 1: p-value Model & Diagnostics

Train Set

```
Confusion Matrix and Statistics

                Reference
Prediction      Not Satisfied Satisfied
  Not Satisfied         34635      5616
  Satisfied              4864     33883

               Accuracy : 0.8673
                 95% CI : (0.865, 0.8697)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7347

 Mcnemar's Test P-Value : 0.0000000000002201

            Sensitivity : 0.8769
            Specificity : 0.8578
         Pos Pred Value : 0.8605
         Neg Pred Value : 0.8745
             Prevalence : 0.5000
         Detection Rate : 0.4384
   Detection Prevalence : 0.5095
      Balanced Accuracy : 0.8673

       'Positive' Class : Not Satisfied
```

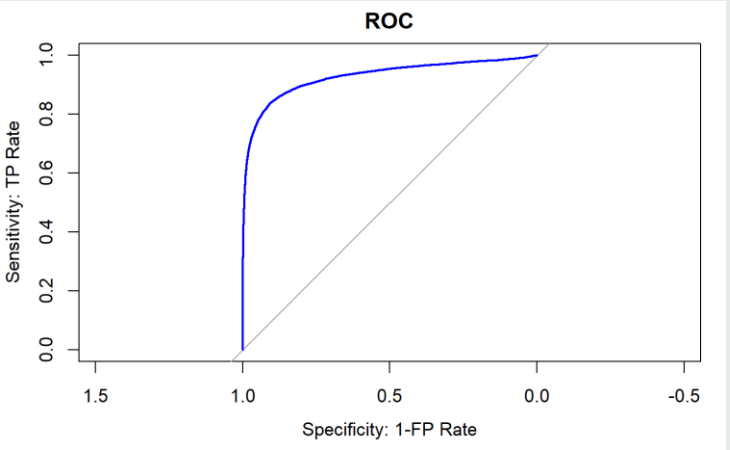
Recall = 0.8769
Precision = 0.8605

Test Set

```
Confusion Matrix and Statistics

                Reference
Prediction      Not Satisfied Satisfied
  Not Satisfied         19279      2362
  Satisfied              2757     14567

               Accuracy : 0.8686
                 95% CI : (0.8652, 0.872)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7334

 Mcnemar's Test P-Value : 0.00000003653

            Sensitivity : 0.8749
            Specificity : 0.8605
         Pos Pred Value : 0.8909
         Neg Pred Value : 0.8409
             Prevalence : 0.5655
         Detection Rate : 0.4948
   Detection Prevalence : 0.5554
      Balanced Accuracy : 0.8677

       'Positive' Class : Not Satisfied
```

Recall = 8749
Precision = 0.8909

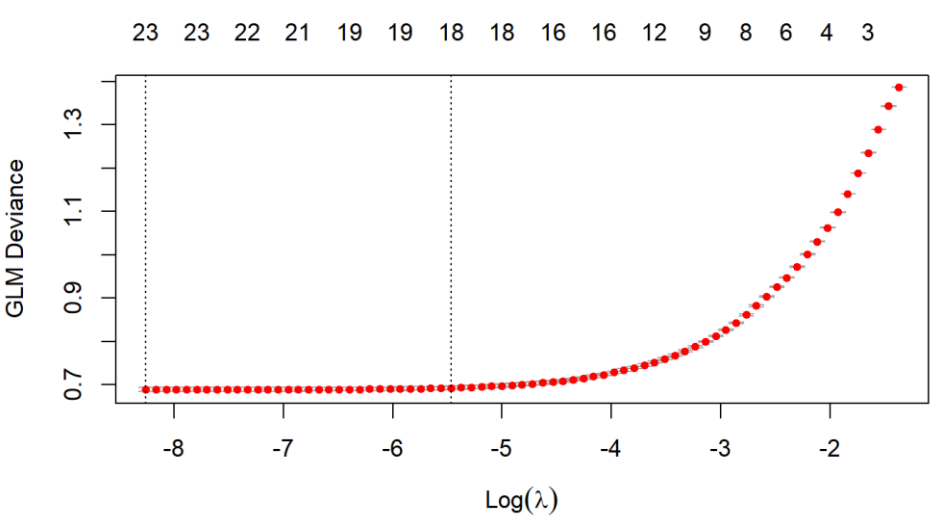
```
Coefficients:
                                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)                       -7.3439770  0.0800886 -91.698 < 0.0000000000000002 ***
genderMale                         0.0630761  0.0219186   2.878              0.00401 **
customer_typeLoyal Customer        2.0022616  0.0318664  62.833 < 0.0000000000000002 ***
age                               -0.0081098  0.0007926 -10.232 < 0.0000000000000002 ***
type_of_travelPersonal Travel     -2.6685233  0.0340247 -78.429 < 0.0000000000000002 ***
customer_classEco Plus            -0.6957687  0.0274278 -25.367 < 0.0000000000000002 ***
customer_classBusiness            -0.7710135  0.0448330 -17.197 < 0.0000000000000002 ***
inflight_wifi_service              0.4354681  0.0128174  33.975 < 0.0000000000000002 ***
departure_arrival_time_convenient -0.1259420  0.0086239 -14.604 < 0.0000000000000002 ***
ease_of_online_booking            -0.1670459  0.0122069 -13.685 < 0.0000000000000002 ***
online_boarding                    0.5789934  0.0109882  52.692 < 0.0000000000000002 ***
seat_comfort                       0.0605202  0.0116311   5.203          0.000000196 ***
onboard_service                    0.2975405  0.0109822  27.093 < 0.0000000000000002 ***
leg_room_service                   0.2382599  0.0094419  25.234 < 0.0000000000000002 ***
baggage_handling                   0.1325803  0.0124923  10.613 < 0.0000000000000002 ***
checkin_service                    0.3117742  0.0093904  33.201 < 0.0000000000000002 ***
inflight_service                   0.1380867  0.0128159  10.775 < 0.0000000000000002 ***
cleanliness                        0.2359191  0.0112103  21.045 < 0.0000000000000002 ***
arrival_delay_in_minutes          -0.0046939  0.0002907 -16.144 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109514  on 78997  degrees of freedom
Residual deviance:  54408  on 78979  degrees of freedom
AIC: 54446
```




```
Area under the curve: 0.9273
```

- Great model but too many variables (17)

Model 2: Feature Selection using LASSO Regularization




```
> cv.lamdas$lambda.min
[1] 0.0002592474
> cv.lamdas$lambda.1se
[1] 0.004225087
```

- Evaluated lambda using CV with k = 10
- Selected λ.1se for LASSO Regression
- Simpler model with fewer variables and error is within 1 standard error of the minimum error

```
Call:  glmnet(x = train_x, y = train_y, family
mbda = cv.lamdas$lambda.1se, standadize = TRUE)

  Df  %Dev   Lambda
1 18 50.06 0.004225
> coef(model)
24 x 1 sparse Matrix of class "dgCMatrix"
                                                s0
(Intercept)                          -6.823350256
genderMale                            .
customer_typeLoyal Customer           1.731102733
age                                  -0.003477848
type_of_travelPersonal Travel        -2.464617465
customer_classEco Plus               -0.648662096
customer_classBusiness               -0.607003073
flight_distance                       .
inflight_wifi_service                 0.321726800
departure_arrival_time_convenient    -0.106018006
ease_of_online_booking               -0.076128708
gate_location                         .
food_and_drink                        .
online_boarding                       0.553962138
seat_comfort                          0.034537461
inflight_entertainment                0.078284522
onboard_service                       0.264157270
leg_room_service                      0.212994052
baggage_handling                      0.111326973
checkin_service                       0.282130087
inflight_service                      0.108176650
cleanliness                           0.183651081
departure_delay_in_minutes            .
arrival_delay_in_minutes             -0.003369552
```

- Variables with coefficients reduced to zero:
  - Flight distance, gate location, departure delay, food and drink, gender
- Still many features (17)

Model 2: LASSO Model & Diagnostics

```
Call:  glmnet(x = train_x, y = train_y, family
mbda = cv.lamdas$lambda.1se, standadize = TRUE)

  Df  %Dev   Lambda
1 18 50.06 0.004225
> coef(model)
24 x 1 sparse Matrix of class "dgCMatrix"
                                              s0
(Intercept)                          -6.823350256
genderMale                            .
customer_typeLoyal Customer           1.731102733
age                                  -0.003477848
type_of_travelPersonal Travel        -2.464617465
customer_classEco Plus               -0.648662096
customer_classBusiness               -0.607003073
flight_distance                       .
inflight_wifi_service                 0.321726800
departure_arrival_time_convenient    -0.106018006
ease_of_online_booking               -0.076128708
gate_location                         .
food_and_drink                        .
online_boarding                       0.553962138
seat_comfort                          0.034537461
inflight_entertainment                0.078284522
onboard_service                       0.264157270
leg_room_service                      0.212994052
baggage_handling                      0.111326973
checkin_service                       0.282130087
inflight_service                      0.108176650
cleanliness                           0.183651081
departure_delay_in_minutes            .
arrival_delay_in_minutes             -0.003369552
```

Train Set

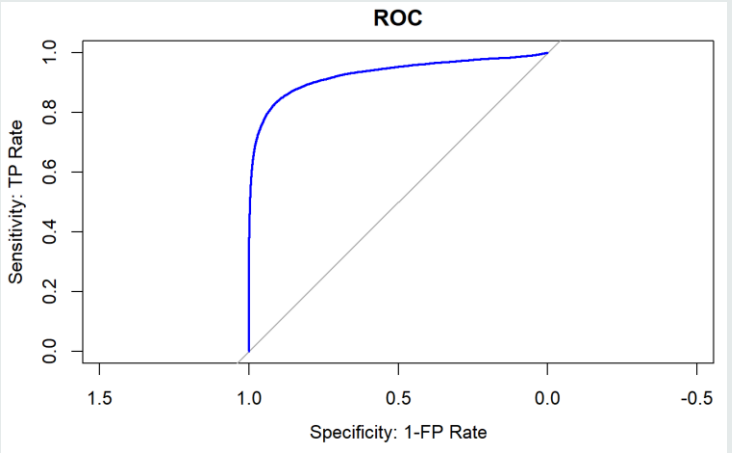
```
Confusion Matrix and Statistics

               Reference
Prediction      Not Satisfied Satisfied
  Not Satisfied         34594      5609
  Satisfied              4905     33890

               Accuracy : 0.8669
                 95% CI : (0.8645, 0.8693)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7338

 Mcnemar's Test P-Value : 0.000000000007081

            Sensitivity : 0.8758
            Specificity : 0.8580
         Pos Pred Value : 0.8605
         Neg Pred Value : 0.8736
             Prevalence : 0.5000
         Detection Rate : 0.4379
   Detection Prevalence : 0.5089
      Balanced Accuracy : 0.8669

       'Positive' Class : Not Satisfied
```

Recall = 0.8758
Precision = 0.8605

Test Set

```
Confusion Matrix and Statistics

               Reference
Prediction      Not Satisfied Satisfied
  Not Satisfied         19237      2370
  Satisfied              2799     14559

               Accuracy : 0.8673
                 95% CI : (0.8639, 0.8707)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7308

 Mcnemar's Test P-Value : 0.000000002632

            Sensitivity : 0.8730
            Specificity : 0.8600
         Pos Pred Value : 0.8903
         Neg Pred Value : 0.8387
             Prevalence : 0.5655
         Detection Rate : 0.4937
   Detection Prevalence : 0.5545
      Balanced Accuracy : 0.8665

       'Positive' Class : Not Satisfied
```

Recall = 0.8730
Precision = 0.8903




```
Area under the curve: 0.9273
```

- Great model but too many variables (17)

Model 3: Feature Selection using IV and WOE

```
> iv_df <- iv_df[order(-iv_df$IV), ]  # sort
> iv_df
                                  VARS     IV
12                  online_boarding 1.8521
7               inflight_wifi_service 1.6215
4                   customer_class 1.1312
3                   type_of_travel 1.0492
14            inflight_entertainment 0.8015
13                     seat_comfort 0.6554
16                leg_room_service 0.4939
15                onboard_service 0.4831
20                    cleanliness 0.4498
6                   flight_distance 0.4234
9             ease_of_online_booking 0.3680
5                               age 0.3640
17                baggage_handling 0.3554
19                inflight_service 0.3404
18                 checkin_service 0.2690
11                  food_and_drink 0.2264
2                    customer_type 0.1538
22          arrival_delay_in_minutes 0.0515
21        departure_delay_in_minutes 0.0208
8  departure_arrival_time_convenient 0.0191
1                           gender 0.0004
10                    gate_location 0.0000
```

- Information Value of input variables against Satisfaction
- Selected top 5 variables in my linear regression model

```
> vif(model)
                              GVIF Df GVIF^(1/(2*Df))
online_boarding           1.113629  1        1.055286
inflight_wifi_service     1.187785  1        1.089856
customer_class            1.325355  2        1.072959
type_of_travel            1.244362  1        1.115510
inflight_entertainment    1.016004  1        1.007970
```

No variables with high VIF

```
Coefficients:
                                Estimate Std. Error z value             Pr(>|z|)
(Intercept)                    -3.770583   0.042920  -87.85 <0.0000000000000002 ***
online_boarding                 0.680021   0.009130   74.48 <0.0000000000000002 ***
inflight_wifi_service           0.240464   0.008342   28.82 <0.0000000000000002 ***
customer_classEco Plus         -1.410080   0.024021  -58.70 <0.0000000000000002 ***
customer_classBusiness         -1.210892   0.041535  -29.15 <0.0000000000000002 ***
type_of_travelPersonal Travel  -1.624714   0.027626  -58.81 <0.0000000000000002 ***
inflight_entertainment          0.538245   0.008231   65.39 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109514  on 78997  degrees of freedom
Residual deviance:  63174  on 78991  degrees of freedom
AIC: 63188
```

Model 3: IV Model & Diagnostics

```
Coefficients:
                                Estimate Std. Error z value             Pr(>|z|)
(Intercept)                   -3.770583   0.042920  -87.85 <0.0000000000000002 ***
online_boarding                0.680021   0.009130   74.48 <0.0000000000000002 ***
inflight_wifi_service          0.240464   0.008342   28.82 <0.0000000000000002 ***
customer_classEco Plus        -1.410080   0.024021  -58.70 <0.0000000000000002 ***
customer_classBusiness        -1.210892   0.041535  -29.15 <0.0000000000000002 ***
type_of_travelPersonal Travel -1.624714   0.027626  -58.81 <0.0000000000000002 ***
inflight_entertainment         0.538245   0.008231   65.39 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109514  on 78997  degrees of freedom
Residual deviance:  63174  on 78991  degrees of freedom
AIC: 63188
```

Train Set

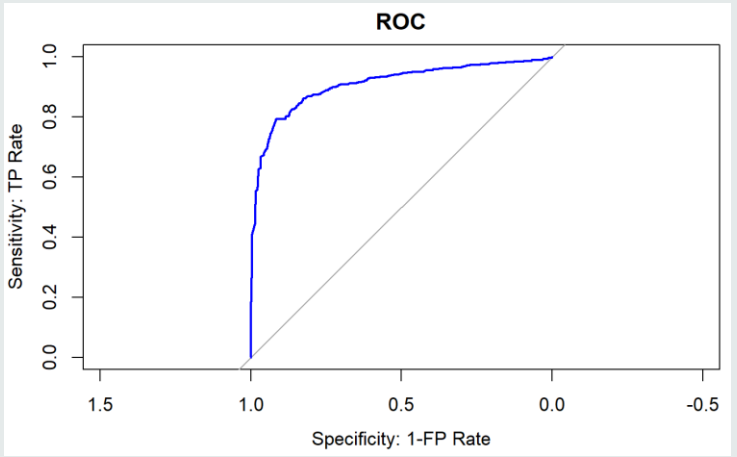
```
Confusion Matrix and Statistics

              Reference
Prediction      Not Satisfied Satisfied
  Not Satisfied         32805      6119
  Satisfied              6694     33380

               Accuracy : 0.8378
                 95% CI : (0.8352, 0.8404)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6756

 Mcnemar's Test P-Value : 0.0000003959

            Sensitivity : 0.8305
            Specificity : 0.8451
         Pos Pred Value : 0.8428
         Neg Pred Value : 0.8330
             Prevalence : 0.5000
         Detection Rate : 0.4153
   Detection Prevalence : 0.4927
      Balanced Accuracy : 0.8378

       'Positive' Class : Not Satisfied
```

Recall = 0.8305
Precision = 0.8428

Test Set

```
Confusion Matrix and Statistics

              Reference
Prediction      Not Satisfied Satisfied
  Not Satisfied         18295      2521
  Satisfied              3741     14408

               Accuracy : 0.8393
                 95% CI : (0.8356, 0.8429)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6757

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.8302
            Specificity : 0.8511
         Pos Pred Value : 0.8789
         Neg Pred Value : 0.7939
             Prevalence : 0.5655
         Detection Rate : 0.4695
   Detection Prevalence : 0.5342
      Balanced Accuracy : 0.8407

       'Positive' Class : Not Satisfied
```

Recall = 0.8302
Precision = 0.8789



Area under the curve: 0.9085

- Good model with just 5 variables!

Model Comparison

Model 1: p-value

```
               Accuracy : 0.8686
                 95% CI : (0.8652, 0.872)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7334

 Mcnemar's Test P-Value : 0.00000003653

            Sensitivity : 0.8749
            Specificity : 0.8605
         Pos Pred Value : 0.8909
         Neg Pred Value : 0.8409
             Prevalence : 0.5655
         Detection Rate : 0.4948
   Detection Prevalence : 0.5554
      Balanced Accuracy : 0.8677

       'Positive' Class : Not Satisfied
```

Area under the curve: 0.9273

Recall = 8749
Precision = 0.8909

- 17 variables, removed 5 variables:
  - Flight distance, gate location, departure delay, food and drink, inflight entertainment

Model 2: LASSO

```
               Accuracy : 0.8673
                 95% CI : (0.8639, 0.8707)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7308

 Mcnemar's Test P-Value : 0.000000002632

            Sensitivity : 0.8730
            Specificity : 0.8600
         Pos Pred Value : 0.8903
         Neg Pred Value : 0.8387
             Prevalence : 0.5655
         Detection Rate : 0.4937
   Detection Prevalence : 0.5545
      Balanced Accuracy : 0.8665

       'Positive' Class : Not Satisfied
```

Area under the curve: 0.9273

Recall = 0.8730
Precision = 0.8903

- 17 variables, removed 5 variables:
  - Flight distance, gate location, departure delay, food and drink, gender

Model 3: IV WOE

```
               Accuracy : 0.8393
                 95% CI : (0.8356, 0.8429)
    No Information Rate : 0.5655
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6757

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.8302
            Specificity : 0.8511
         Pos Pred Value : 0.8789
         Neg Pred Value : 0.7939
             Prevalence : 0.5655
         Detection Rate : 0.4695
   Detection Prevalence : 0.5342
      Balanced Accuracy : 0.8407

       'Positive' Class : Not Satisfied
```

Area under the curve: 0.9085

Recall = 0.8302
Precision = 0.8789

- Best model since just 5 variables:
  - Online boarding, inflight WiFi service, customer class, type of travel, inflight entertainment

Thank you!

References

- Dataset: Airline Passenger Satisfaction | Kaggle
- Slide 1 Image: The Best U.S. Airlines for Business Travel 2015 | Fortune
- Logistic Regression: Module 3 Introduction: ALY6015 81176 Intermediate Analytics SEC 20 Spring 2022 CPS [SJO-A-HY] (instructure.com)
- Splitting Bias Data and Information Value: Logistic Regression With R (r-statistics.co)