

# PEC3 - Modelos predictivos

Gloria Manresa Santamaría

2022-12-23

## Lectura fichero

En primer lugar cargamos los datos y cambiamos las variables tipo integer a numeric y las tipo character a factor.

```
dat <- read.csv('datSat_Air.csv')
```

Transformamos las columnas integer a numeric y las character a factor.

```
dat$id <- as.numeric(dat$id)
dat$satisfaction <- as.factor(dat$satisfaction)
dat$Gender <- as.factor(dat$Gender)
dat$Customer_Type <- as.factor(dat$Customer_Type)
dat$Age <- as.numeric(dat$Age)
dat>Type_Travel <- as.factor(dat>Type_Travel)
dat$Class <- as.factor(dat$Class)
dat$Distance <- as.numeric(dat$Distance)
dat$Seat_comfort <- as.numeric(dat$Seat_comfort)
dat$Food_drink <- as.numeric(dat$Food_drink)
dat$Gate <- as.numeric(dat$Gate)
dat$Wifi <- as.numeric(dat$Wifi)
dat$Ent <- as.numeric(dat$Ent)
dat$Ease_booking <- as.numeric(dat$Ease_booking)
dat$Service <- as.numeric(dat$Service)
dat$Baggage_handling <- as.numeric(dat$Baggage_handling)
dat$Checkin_service <- as.numeric(dat$Checkin_service)
dat$Cleanliness <- as.numeric(dat$Cleanliness)
dat$Online_boarding <- as.numeric(dat$Online_boarding)
dat$Departure_Delay <- as.numeric(dat$Departure_Delay)
dat$Arrival_Delay <- as.numeric(dat$Arrival_Delay)

knitr::kable(sapply(dat, class), col.names="Tipo de datos")
```

| Tipo de datos |         |
|---------------|---------|
| id            | numeric |
| satisfaction  | factor  |
| Gender        | factor  |
| Customer_Type | factor  |
| Age           | numeric |

|                  | Tipo de datos |
|------------------|---------------|
| Type_Travel      | factor        |
| Class            | factor        |
| Distance         | numeric       |
| Seat_comfort     | numeric       |
| Food_drink       | numeric       |
| Gate             | numeric       |
| Wifi             | numeric       |
| Ent              | numeric       |
| Ease_booking     | numeric       |
| Service          | numeric       |
| Baggage_handling | numeric       |
| Checkin_service  | numeric       |
| Cleanliness      | numeric       |
| Online_boarding  | numeric       |
| Departure_Delay  | numeric       |
| Arrival_Delay    | numeric       |

Creamos una nueva variable “satisfaction\_re”

```
dat$satisfaction_re <- ifelse(dat$satisfaction == "neutral or dissatisfied", 0,
                                ifelse(dat$satisfaction == "satisfied", 1, NA))
```

## 1. Regresión lineal

### 1.1 Modelo regresión lineal (variables cuantitativas)

#### Apartado A

```
modelo1 <- lm(Arrival_Delay ~ Distance, data = dat )
summary(modelo1)
```

```
##
## Call:
## lm(formula = Arrival_Delay ~ Distance, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.00 -15.30 -11.74  -1.29 477.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.239e+00 2.166e-01   33.42   <2e-16 ***
## Distance    3.850e-03 9.709e-05   39.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.86 on 129444 degrees of freedom
## Multiple R-squared:  0.012, Adjusted R-squared:  0.012
## F-statistic: 1573 on 1 and 129444 DF, p-value: < 2.2e-16
```

El valor p es menor a 0.05 (recordamos que trabajamos al 95% de confianza) por lo que podemos aceptar el modelo.

Por otro lado vemos que el R cuadrado es demasiado pequeño, esto quiere decir que únicamente un 1.2% de la variabilidad de la variable predicha es explicada por la variable predictora.

### Apartado B

```
modelo2 <- lm(Arrival_Delay ~ Distance + Departure_Delay, data = dat )
summary(modelo2)
```

```
##
## Call:
## lm(formula = Arrival_Delay ~ Distance + Departure_Delay, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.422  -1.951  -0.787  -0.404 236.671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.030e-01 6.083e-02 9.913 < 2e-16 ***
## Distance    9.530e-05 2.733e-05 3.487 0.000488 ***
## Departure_Delay 9.761e-01 7.905e-04 1234.821 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## Residual standard error: 10.03 on 129443 degrees of freedom
## Multiple R-squared: 0.9227, Adjusted R-squared: 0.9227
## F-statistic: 7.724e+05 on 2 and 129443 DF, p-value: < 2.2e-16
```

El modelo mejora significativamente ya que en este caso vemos que R-squared es 0.9227 que es un porcentaje muy alto. En este caso un 92.27% de la variabilidad de la variable predicha es explicada por la variable predictora.

## 1.2 Modelo de regresión lineal

### Apartado A

```
modelo3 <- lm(Arrival_Delay ~ Distance + Departure_Delay + Service + Food_drink + satisfaction + Customer_Type, data = dat)
summary(modelo3)
```

```
##
## Call:
## lm(formula = Arrival_Delay ~ Distance + Departure_Delay + Service +
##     Food_drink + satisfaction + Customer_Type, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.770  -2.186  -0.668  -0.194 237.002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```

## (Intercept)           1.143e+00  1.239e-01   9.227 < 2e-16 ***
## Distance             8.383e-05  2.733e-05   3.067  0.00216 **
## Departure_Delay      9.753e-01  7.922e-04  1231.181 < 2e-16 ***
## Service              -6.457e-02  2.344e-02   -2.755  0.00588 **
## Food_drink            -2.467e-02  1.945e-02   -1.268  0.20464
## satisfactionsatisfied -7.060e-01  6.285e-02   -11.233 < 2e-16 ***
## Customer_TypeLoyal Customer 2.137e-01  7.537e-02    2.836  0.00457 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.03 on 129439 degrees of freedom
## Multiple R-squared:  0.9228, Adjusted R-squared:  0.9228
## F-statistic: 2.579e+05 on 6 and 129439 DF,  p-value: < 2.2e-16

```

La variable Food\_drink la quitamos del modelo por tener un p valor mayor a 0.05. También eliminamos del modelo la variable Customer\_Type por ser una variable confusa.

```

ModelF <- lm(Arrival_Delay ~ Distance + Departure_Delay + Service + satisfaction, data = dat )
res <- resid(ModelF)
summary(ModelF)

```

```

##
## Call:
## lm(formula = Arrival_Delay ~ Distance + Departure_Delay + Service +
##     satisfaction, data = dat)
##
## Residuals:
##    Min      1Q      Median      3Q      Max
## -53.712  -2.186  -0.643  -0.248 237.063
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.230e+00  9.910e-02 12.413 < 2e-16 ***
## Distance             8.301e-05  2.733e-05   3.038  0.00239 **
## Departure_Delay      9.754e-01  7.921e-04  1231.403 < 2e-16 ***
## Service              -6.577e-02  2.344e-02   -2.806  0.00501 **
## satisfactionsatisfied -6.647e-01  5.997e-02   -11.085 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.03 on 129441 degrees of freedom
## Multiple R-squared:  0.9228, Adjusted R-squared:  0.9228
## F-statistic: 3.868e+05 on 4 and 129441 DF,  p-value: < 2.2e-16

```

## Apartado B

Comprobamos si existe colinealidad en las variables del modelo

```
car::vif(ModelF)
```

```

##          Distance Departure_Delay        Service      satisfaction
## 1.013852       1.017933       1.142322       1.147393

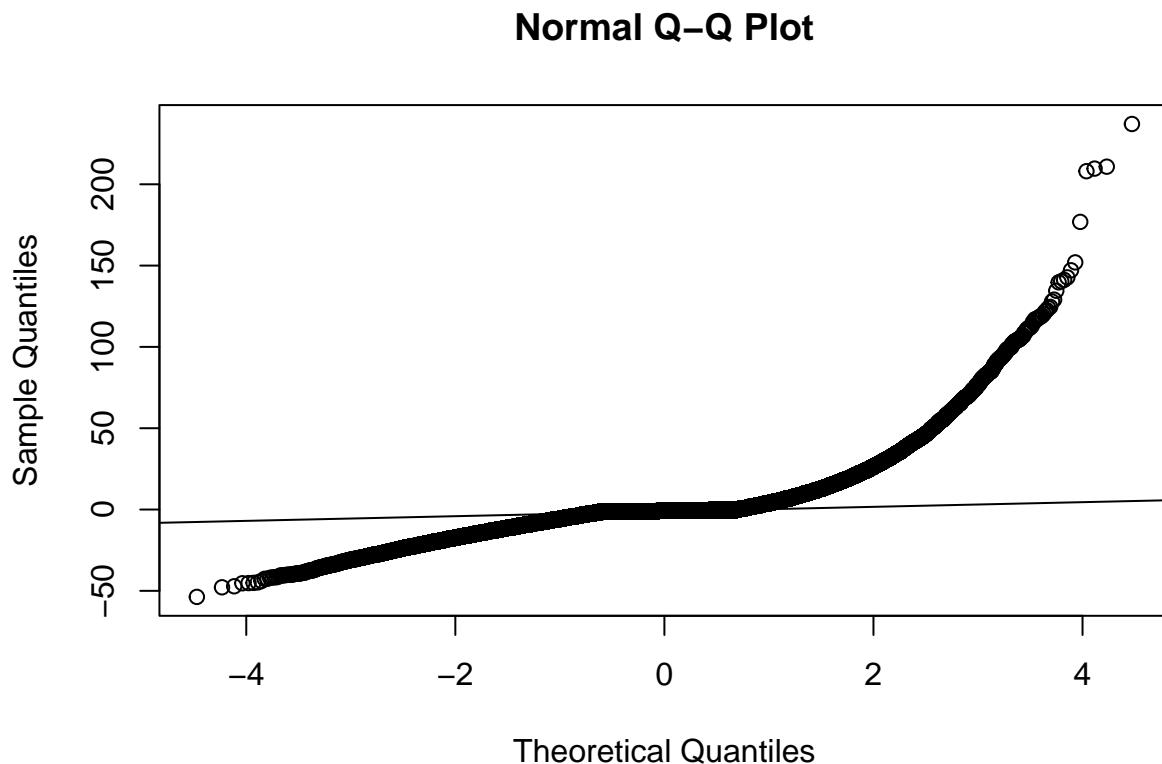
```

Confirmamos que no existe colinealidad en ningún caso.

### 1.3 Diagnosis del modelo

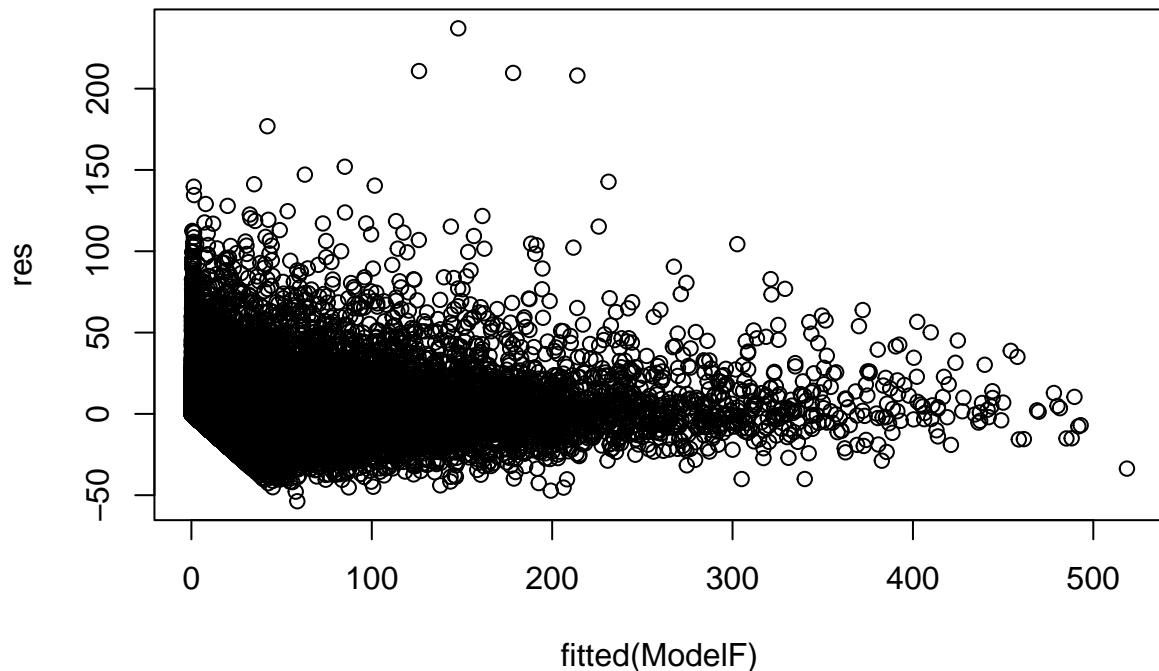
En el gráfico cuantil-cuantil siguiente observamos que los residuos no se ajustan a la recta por lo que no cumplen con una distribución normal.

```
qqnorm(residuals(ModelF))  
qqline(residuals(ModelF))
```



En el gráfico siguiente observamos que la varianza no es constante ya que los datos no se encuentran ordenados entre dos líneas paralelas.

```
plot(fitted(ModelF), res)
```



## 1.4 Predicción del modelo

Asignando los valores especificados al modelo nos predice un retraso de 29.8 minutos.

```
pred14 <- predict(ModelF, data.frame(Distance=2500, Departure_Delay=30, Service=3, satisfaction="satisfied"))
pred14
```

```
##          1
## 29.83705
```

## 2 Regresión logística

### 2.1 Generación de los conjuntos de entrenamiento y de test

Se dividen los datos en train (80%) y test (20%).

```
set.seed(111)

dt = sort(sample(nrow(dat), nrow(dat)*0.8))
train <- dat[dt,]
test <- dat[-dt,]
```

## 2.2 Estimación del modelo con el conjunto de entrenamiento e interpretación

### Apartado A

Modelo de regresión logística con la variable dependiente satisfaction\_re

```
modelo5 <- glm(satisfaction_re ~ Gender + Customer_Type + Age + Type_Travel + Class +
                  Distance + Seat_comfort + Food_drink + Gate + Wifi + Ent + Ease_booking +
                  Service + Baggage_handling + Checkin_service + Cleanliness + Online_boarding +
                  Departure_Delay + Arrival_Delay, data=train, family = binomial(link=logit))
summary(modelo5)

##
## Call:
## glm(formula = satisfaction_re ~ Gender + Customer_Type + Age +
##      Type_Travel + Class + Distance + Seat_comfort + Food_drink +
##      Gate + Wifi + Ent + Ease_booking + Service + Baggage_handling +
##      Checkin_service + Cleanliness + Online_boarding + Departure_Delay +
##      Arrival_Delay, family = binomial(link = logit), data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.9741 -0.5881  0.2044  0.5415  3.5579
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.296e+00  7.033e-02 -89.520 < 2e-16 ***
## GenderMale                -1.011e+00  1.823e-02 -55.467 < 2e-16 ***
## Customer_TypeLoyal Customer  1.912e+00  2.752e-02  69.466 < 2e-16 ***
## Age                      -7.123e-03  6.311e-04 -11.287 < 2e-16 ***
## Type_TravelPersonal Travel -8.375e-01  2.585e-02 -32.394 < 2e-16 ***
## ClassEco                 -7.696e-01  2.359e-02 -32.624 < 2e-16 ***
## ClassEco Plus            -8.535e-01  3.619e-02 -23.584 < 2e-16 ***
## Distance                 -1.180e-04  9.481e-06 -12.442 < 2e-16 ***
## Seat_comfort              2.766e-01  1.003e-02  27.582 < 2e-16 ***
## Food_drink                -2.796e-01  9.972e-03 -28.038 < 2e-16 ***
## Gate                      3.587e-02  7.966e-03   4.503 6.69e-06 ***
## Wifi                      -8.935e-02  9.618e-03  -9.290 < 2e-16 ***
## Ent                       7.175e-01  8.848e-03  81.096 < 2e-16 ***
## Ease_booking               3.264e-01  1.226e-02  26.610 < 2e-16 ***
## Service                   3.243e-01  9.141e-03  35.480 < 2e-16 ***
## Baggage_handling          1.284e-01  1.028e-02  12.500 < 2e-16 ***
## Checkin_service            2.881e-01  7.649e-03  37.672 < 2e-16 ***
## Cleanliness                9.221e-02  1.067e-02   8.645 < 2e-16 ***
## Online_boarding            1.687e-01  1.053e-02  16.020 < 2e-16 ***
## Departure_Delay            3.047e-03  8.909e-04   3.421 0.000625 ***
## Arrival_Delay              -8.455e-03  8.771e-04  -9.640 < 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: 142594  on 103555  degrees of freedom
## Residual deviance: 81316  on 103535  degrees of freedom
```

```

## AIC: 81358
##
## Number of Fisher Scoring iterations: 5

```

Observamos que todas las variables son significativas ya que el p valor de cada una de ellas es menor a 0.05.

## Apartado B

```
car::vif(modelo5)
```

|                     | GVIF      | Df | GVIF^(1/(2*Df)) |
|---------------------|-----------|----|-----------------|
| ## Gender           | 1.067948  | 1  | 1.033416        |
| ## Customer_Type    | 1.474395  | 1  | 1.214247        |
| ## Age              | 1.227158  | 1  | 1.107772        |
| ## Type_Travel      | 1.968739  | 1  | 1.403118        |
| ## Class            | 1.679844  | 2  | 1.138459        |
| ## Distance         | 1.198094  | 1  | 1.094575        |
| ## Seat_comfort     | 2.243326  | 1  | 1.497774        |
| ## Food_drink       | 2.532589  | 1  | 1.591411        |
| ## Gate             | 1.394937  | 1  | 1.181075        |
| ## Wifi             | 1.971263  | 1  | 1.404017        |
| ## Ent              | 1.421677  | 1  | 1.192341        |
| ## Ease_booking     | 2.906035  | 1  | 1.704710        |
| ## Service          | 1.600442  | 1  | 1.265086        |
| ## Baggage_handling | 1.821174  | 1  | 1.349509        |
| ## Checkin_service  | 1.161735  | 1  | 1.077838        |
| ## Cleanliness      | 1.945538  | 1  | 1.394825        |
| ## Online_boarding  | 2.257069  | 1  | 1.502354        |
| ## Departure_Delay  | 12.839314 | 1  | 3.583199        |
| ## Arrival_Delay    | 12.857237 | 1  | 3.585699        |

Parece ser que las variables Departure\_Delay y Arrival\_Delay presentan una fuerte correlación entre sí. Probaremos a eliminar una de ellas del modelo y estudiar de nuevo la colinealidad.

```

ModelgF <- glm(satisfaction_re ~ Gender + Customer_Type + Age + Type_Travel + Class + Distance +
                  Seat_comfort + Food_drink + Gate + Wifi + Ent + Ease_booking + Service +
                  Baggage_handling + Checkin_service + Cleanliness + Online_boarding + Departure_Delay,
                  data=dat, family = binomial(link = logit))
summary(ModelgF)

##
## Call:
## glm(formula = satisfaction_re ~ Gender + Customer_Type + Age +
##       Type_Travel + Class + Distance + Seat_comfort + Food_drink +
##       Gate + Wifi + Ent + Ease_booking + Service + Baggage_handling +
##       Checkin_service + Cleanliness + Online_boarding + Departure_Delay,
##       family = binomial(link = logit), data = dat)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.9647   -0.5899    0.2045    0.5419    3.5821
##
## Coefficients:

```

```

##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -6.278e+00  6.281e-02 -99.952 < 2e-16 ***
## GenderMale                  -1.019e+00  1.629e-02 -62.571 < 2e-16 ***
## Customer_TypeLoyal Customer  1.895e+00  2.457e-02  77.119 < 2e-16 ***
## Age                         -7.204e-03  5.635e-04 -12.784 < 2e-16 ***
## Type_TravelPersonal Travel -8.379e-01  2.312e-02 -36.250 < 2e-16 ***
## ClassEco                    -7.639e-01  2.110e-02 -36.202 < 2e-16 ***
## ClassEco Plus               -8.519e-01  3.239e-02 -26.300 < 2e-16 ***
## Distance                     -1.207e-04  8.487e-06 -14.225 < 2e-16 ***
## Seat_comfort                 2.763e-01  8.975e-03  30.790 < 2e-16 ***
## Food_drink                   -2.796e-01  8.921e-03 -31.338 < 2e-16 ***
## Gate                         3.299e-02  7.142e-03   4.619 3.85e-06 ***
## Wifi                         -9.108e-02  8.625e-03 -10.560 < 2e-16 ***
## Ent                          7.154e-01  7.918e-03  90.359 < 2e-16 ***
## Ease_booking                 3.371e-01  1.098e-02  30.684 < 2e-16 ***
## Service                      3.195e-01  8.181e-03  39.051 < 2e-16 ***
## Baggage_handling             1.250e-01  9.193e-03  13.598 < 2e-16 ***
## Checkin_service              2.958e-01  6.834e-03  43.292 < 2e-16 ***
## Cleanliness                  8.846e-02  9.537e-03   9.275 < 2e-16 ***
## Online_boarding              1.668e-01  9.428e-03  17.692 < 2e-16 ***
## Departure_Delay              -5.108e-03  2.263e-04 -22.575 < 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: 178281  on 129445  degrees of freedom
## Residual deviance: 101774  on 129426  degrees of freedom
## AIC: 101814
##
## Number of Fisher Scoring iterations: 5

```

```
car::vif(ModelgF)
```

```

##                               GVIF Df GVIF^(1/(2*Df))
## Gender                  1.067152  1      1.033030
## Customer_Type            1.472572  1      1.213496
## Age                      1.225325  1      1.106944
## Type_Travel              1.966684  1      1.402385
## Class                     1.682419  2      1.138895
## Distance                 1.197555  1      1.094329
## Seat_comfort              2.240768  1      1.496919
## Food_drink                2.534499  1      1.592011
## Gate                      1.399149  1      1.182856
## Wifi                      1.978454  1      1.406575
## Ent                       1.421618  1      1.192316
## Ease_booking               2.912012  1      1.706462
## Service                   1.600701  1      1.265188
## Baggage_handling           1.824541  1      1.350756
## Checkin_service             1.160441  1      1.077238
## Cleanliness                1.950613  1      1.396644
## Online_boarding             2.258499  1      1.502830
## Departure_Delay             1.031588  1      1.015671

```

Eliminando la variable Arrival\_Delay hemos solucionado el problema de la colinealidad.

### Apartado C

Según los cálculos realizados en el apartado 2.3 las variables que pueden considerarse factores de riesgo o protección son:

Ease\_booking, Baggage\_handling, Cleanliness , Customer\_Type, Seat\_comfort, Gate, Ent, Service, Check-in\_Service, Online\_boarding

### 2.3 Cálculo de las OR (Odds-Ratio)

Los odds-ratio correspondientes al modelo ModelgF son los siguientes.

```
OR_ModelgF <- exp(coef(ModelgF))
OR_ModelgF
```

|    |                             |                  |
|----|-----------------------------|------------------|
| ## | (Intercept)                 | GenderMale       |
| ## | 0.001877625                 | 0.360803284      |
| ## | Customer_TypeLoyal Customer | Age              |
| ## | 6.651404719                 | 0.992822126      |
| ## | Type_TravelPersonal Travel  | ClassEco         |
| ## | 0.432597087                 | 0.465861177      |
| ## | ClassEco Plus               | Distance         |
| ## | 0.426610285                 | 0.999879277      |
| ## | Seat_comfort                | Food_drink       |
| ## | 1.318301650                 | 0.756117115      |
| ## | Gate                        | Wifi             |
| ## | 1.033540316                 | 0.912946251      |
| ## | Ent                         | Ease_booking     |
| ## | 2.045081386                 | 1.400813649      |
| ## | Service                     | Baggage_handling |
| ## | 1.376432243                 | 1.133157516      |
| ## | Checkin_service             | Cleanliness      |
| ## | 1.344258165                 | 1.092490490      |
| ## | Online_boarding             | Departure_Delay  |
| ## | 1.181521879                 | 0.994905084      |

En el caso de la variable Class observamos que el OR es menor a 0 por lo que el suceso (estar satisfecho) es menos probable en presencia de esta variable (estar en las clases ECO y ECO PLUS).

En el caso de la variable Customer\_Type, que presenta un OR mayor a 1 (6.55) el suceso (estar satisfecho) es más probable en presencia de esta variable (Customer type = Loyal).

La probabilidad de que un hombre esté satisfecho es de 0.358671059 veces con respecto a las mujeres. Es decir, la probabilidad de que un hombre esté satisfecho es 35.8% mayor con respecto a las mujeres.

Por último, los clientes es más probable que estén satisfecho en presencia de entretenimiento.

### 2.4 Matriz confusión

La matriz de confusión quedaría de la siguiente manera:

```

library("caret")

results <- predict(ModelgF,newdata=test,type='response')
results <- ifelse(results > 0.5,1,0)

test$satisfaction_re <- as.factor(test$satisfaction_re)
results <- as.factor(results)

matriz_confusion <- confusionMatrix(data=results, reference = test$satisfaction_re)
matriz_confusion

## Confusion Matrix and Statistics
##
##             Reference
## Prediction      0      1
##           0 9551  2128
##           1 2241 11970
##
##             Accuracy : 0.8312
##             95% CI : (0.8266, 0.8358)
##   No Information Rate : 0.5445
##   P-Value [Acc > NIR] : < 2e-16
##
##             Kappa : 0.6595
##
##   Mcnemar's Test P-Value : 0.09018
##
##             Sensitivity : 0.8100
##             Specificity  : 0.8491
##   Pos Pred Value : 0.8178
##   Neg Pred Value : 0.8423
##   Prevalence    : 0.4555
##   Detection Rate : 0.3689
##   Detection Prevalence : 0.4511
##   Balanced Accuracy : 0.8295
##
##   'Positive' Class : 0
##

```

Observamos que tenemos una precisión en el modelo del 83.12%.

## 2.5 Predicción

El cliente encuestado número 3 presenta los siguientes datos:

```

id_3 <- dat[dat$id==3,]
id_3

##
##      id satisfaction Gender Customer_Type Age      Type_Travel     Class
## 51303  3      satisfied Male Loyal Customer  41 Business travel Business
##          Distance Seat_comfort Food_drink Gate Wifi Ent Ease_booking Service
## 51303     879                 4           4     5     5                  3       3

```

```

##      Baggage_handling Checkin_service Cleanliness Online_boarding
## 51303            3             4            3            5
##      Departure_Delay Arrival_Delay satisfaction_re
## 51303            0             0            1

pred25 <- predict(ModelgF,data.frame(Gender="Male", Customer_Type="Loyal Customer", Age=41,
                                         Type_Travel="Business travel", Class="Business", Distance=879,
                                         Seat_comfort=4, Food_drink=4, Gate=4, Wifi=5, Ent=5, Ease_booking=5,
                                         Baggage_handling=3, Checkin_service=4, Cleanliness=3, Online_boarding=5,
                                         Departure_Delay =0),type = "response")

pred25

##      1
## 0.887406

```

El cliente encuestado con ID = 3 tiene un 88'7% de estar satisfecho con la aerolínea.

## 2.6 Bondad de ajuste

### Apartado A

Según podemos observar en nuestro modelo la desvianza residual es menor que la desvianza nula:

Null deviance: 178281 on 129445 degrees of freedom

Residual deviance: 101774 on 129426 degrees of freedom

### Apartado B

```
chisq.test(as.matrix(matriz_confusion))
```

```

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: as.matrix(matriz_confusion)
## X-squared = 11260, df = 1, p-value < 2.2e-16

```

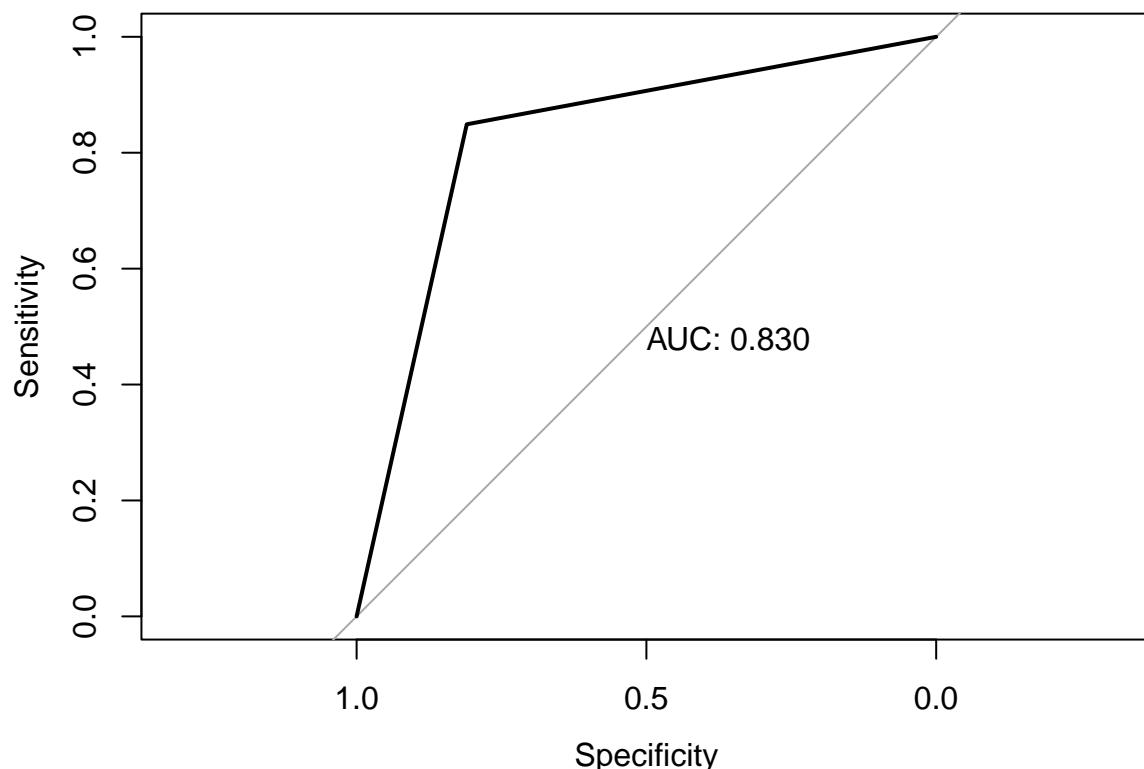
El p valor es menor a 0.05 por lo que no podemos afirmar que exista dependencia entre las variables.

## 2.7 Curva ROC

```

library(pROC)
test_roc = roc(as.numeric(test$satisfaction_re)~as.numeric(results),plot=TRUE, print.auc=TRUE)

```



El área bajo la curva es 0.83 por lo que parece que el modelo trabaja correctamente.

### 3 Informe ejecutivo

#### 3.1 Presentación de los principales resultados del estudio en una tabla

| Pregunta  | Conclusión  |
|---|---|
| 1.1.a Regresión lineal.<br>Arrival_Delay - Distance   | P valor acepta el modelo. Sin embargo el modelo explica un porcentaje muy bajo (1.2%) de la variabilidad de la variable predicha. |
| 1.1.b Regresión lineal.<br>Arrival_Delay - Distance+Departure_Delay   | El modelo mejora notablemente ya que R-squared aumenta a 0.9227   |
| 1.2.a Regresión lineal.<br>Arrival_Delay - Distance+Departure_Delay<br>+Service+Food_drink+satisfaction+Customer_Type | Eliminamos del modelo Food_drink ( $p>0.05$ ) y Customer_Type por ser una variable confusa  |
| 1.2.b ¿Existen problemas de colinealidad?   | Comprobamos que no existen problemas de colinealidad en las variables explicativas  |

| Pregunta  | Conclusión   |
|---|--|
| 1.3 Interpretación de gráficos  | Los residuos no siguen una distribución normal. La varianza no es constante.   |
| 1.4 Predicción del modelo   | El retraso predicho por nuestro modelo es de 29.83 minutos.  |
| 2.1 Generación de los conjuntos de entrenamiento y test                                       | Se divide el dataset original en train (80%) y en test (20%)   |
| 2.2.a Estimad el modelo de regresión logística siendo la variable dependiente satisfaction_re | Las variables explicativas presentan todas un p valor mayor a 0.05   |
| 2.2.b Estudiad la presencia de colinealidad   | Eliminamos Arrival_Delay ya que parece que presenta colinealidad   |
| 2.2.c ¿Hay variables que puedan considerarse factores de riesgo?                              | Existen 10 variables que pueden considerarse variables de riesgo.  |
| 2.3 Cálculo de las OR   | Es menos probable que los clientes de las clases ECO y ECO PLUS estén satisfechos frente a clientes de otras clases. Los clientes tipo "Loyal" es más probable que estén satisfechos frente a otros tipos de clientes. La probabilidad de que un hombre esté satisfecho es de 35.8% mayor frente a las mujeres. Los clientes es más probable que estén satisfecho en presencia de entretenimiento. |
| 2.4 Matriz confusión  | La precisión del modelo es 0.83  |
| 2.5 Predicción  | La probabilidad de que el encuestado número tres esté satisfecho es del 88.7%  |
| 2.6 Bondad de ajuste  | Confirmamos que la devianza residual es menor a la nula. El p valor del test Chi cuadrado es menor a 0.05 por lo que acepta el modelo.   |
| 2.7 Curva ROC   | El área bajo la curva es 0.83  |

### 3.2 Resumen ejecutivo.

En primer lugar se ha observado que en un 92.27% de los casos el retraso en la llegada del vuelo es explicado por un retraso en la salida del mismo. Con el modelo realizado se puede predecir el retraso en la llegada de un vuelo a partir otros valores, el retraso en la salida del mismo entre otros.

Se ha estudiado también la satisfacción de los clientes. La probabilidad de que un hombre esté satisfecho es 35.8% mayor con respecto a las mujeres. La presencia de entretenimiento mejora la satisfacción general de los clientes. El modelo creado es capaz de predecir la probabilidad de que un cliente quede satisfecho.