FINAL PROJECT

Andrés Padrón Quintana

2025-05-05

```
# Load libraries
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.3.2
```

```
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.3
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

```r
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

¿Qué factores influyen más en la cancelación de reservas y cómo puede el hotel reducir su tasa de cancelación?

Objetivos:

– Identificar los factores que predicen cancelaciones de reservas.

– Proponer estrategias basadas en datos para reducir cancelaciones.

– Diseñar un modelo predictivo que ayude a anticipar cancelaciones para permitir que el hotel optimice su ocupación y aumente sus ingresos.

Dataset

Extraído de: https://www.kaggle.com/datasets/mojtaba142/hotel-booking Este dataset de Kaggle proporciona información detallada sobre las reservas realizadas en dos tipos de hoteles: City Hotel y Resort Hotel. Total de registros: 119,390 Total de variables: 36

1. ANÁLISIS EXPLORATORIO DE DATOS

```r
# Load dataset
dataset <- read.csv('/Users/andrespadronquintana/Desktop/FIN CORP AV/FINAL/hotel_booking.csv')

# Convert 'is_canceled' to a factor
dataset <- dataset %>%
  mutate(is_canceled = as.factor(is_canceled))

# Summary and structure of the dataset
summary(dataset)
```

```
##     hotel           is_canceled   lead_time       arrival_date_year
##  Length:119390      0:75166     Min.   :  0.0   Min.   :2015
##  Class :character   1:44224     1st Qu.: 18.0   1st Qu.:2016
##  Mode  :character               Median : 69.0   Median :2016
##                                 Mean   :104.0   Mean   :2016
##                                 3rd Qu.:160.0   3rd Qu.:2017
##                                 Max.   :737.0   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:119390      Min.   : 1.00            Min.   : 1.0
##  Class :character   1st Qu.:16.00            1st Qu.: 8.0
##  Mode  :character   Median :28.00            Median :16.0
##                     Mean   :27.17            Mean   :15.8
##                     3rd Qu.:38.00            3rd Qu.:23.0
##                     Max.   :53.00            Max.   :31.0
##
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.0         Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.0         1st Qu.: 2.000
##  Median : 1.0000         Median : 2.0         Median : 2.000
##  Mean   : 0.9276         Mean   : 2.5         Mean   : 1.856
##  3rd Qu.: 2.0000         3rd Qu.: 3.0         3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.0         Max.   :55.000
##
##     children           babies              meal             country
##  Min.   : 0.0000   Min.   : 0.000000   Length:119390      Length:119390
##  1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
##  Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
##  Mean   : 0.1039   Mean   : 0.007949
##  3rd Qu.: 0.0000   3rd Qu.: 0.000000
##  Max.   :10.0000   Max.   :10.000000
##  NA's   :4
##  market_segment     distribution_channel is_repeated_guest
##  Length:119390      Length:119390        Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03191
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
##
```

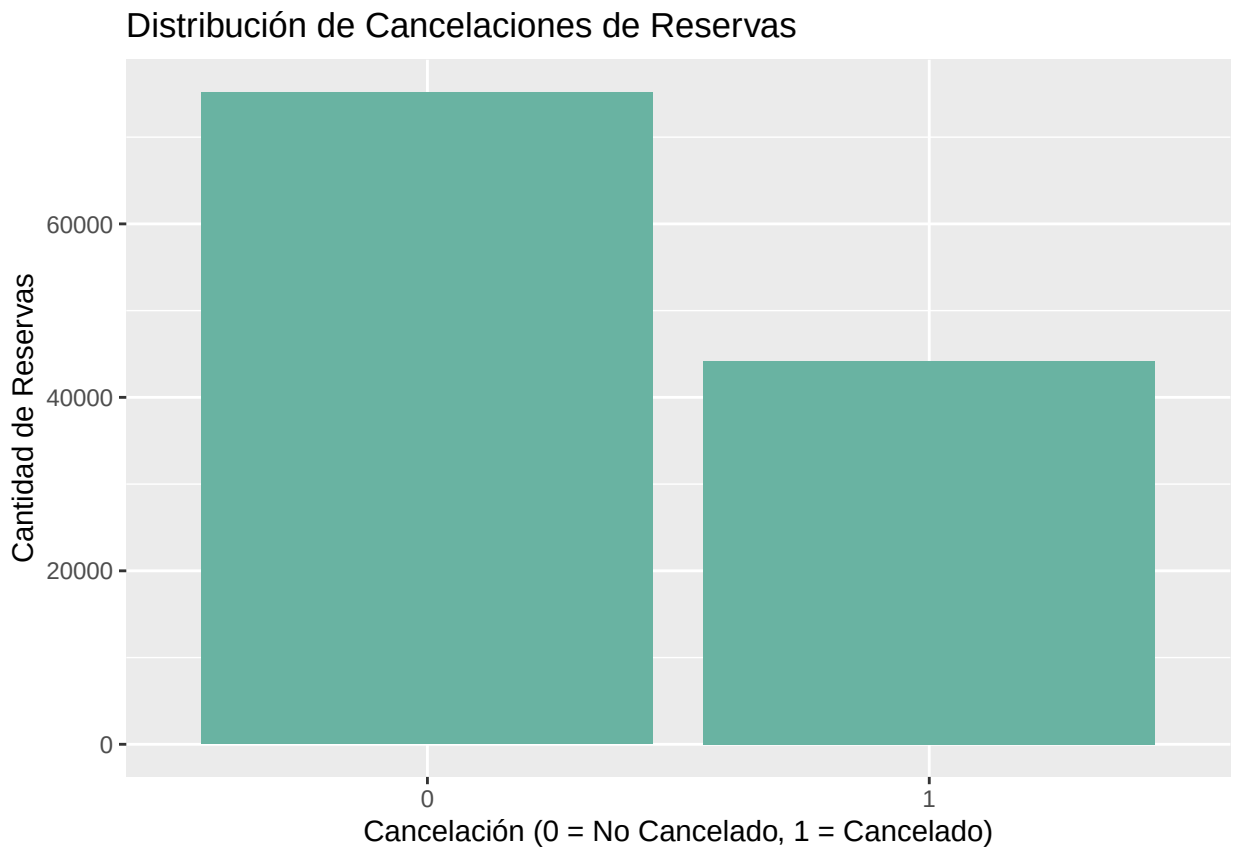
```
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.0000                Length:119390
##  1st Qu.: 0.00000       1st Qu.: 0.0000                Class :character
##  Median : 0.00000       Median : 0.0000                Mode  :character
##  Mean   : 0.08712       Mean   : 0.1371
##  3rd Qu.: 0.00000       3rd Qu.: 0.0000
##  Max.   :26.00000       Max.   :72.0000
##
##  assigned_room_type booking_changes   deposit_type           agent
##  Length:119390      Min.   : 0.0000   Length:119390      Min.   :  1.00
##  Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:  9.00
##  Mode  :character   Median : 0.0000   Mode  :character   Median : 14.00
##                     Mean   : 0.2211                      Mean   : 86.69
##                     3rd Qu.: 0.0000                      3rd Qu.:229.00
##                     Max.   :21.0000                      Max.   :535.00
##                                                          NA's   :16340
##     company       days_in_waiting_list customer_type           adr
##  Min.   :  6.0    Min.   :  0.000      Length:119390      Min.   :  -6.38
##  1st Qu.: 62.0    1st Qu.:  0.000      Class :character   1st Qu.:  69.29
##  Median :179.0    Median :  0.000      Mode  :character   Median :  94.58
##  Mean   :189.3    Mean   :  2.321                         Mean   : 101.83
##  3rd Qu.:270.0    3rd Qu.:  0.000                         3rd Qu.: 126.00
##  Max.   :543.0    Max.   :391.000                         Max.   :5400.00
##  NA's   :112593
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.0000            Length:119390
##  1st Qu.:0.00000             1st Qu.:0.0000            Class :character
##  Median :0.00000             Median :0.0000            Mode  :character
##  Mean   :0.06252             Mean   :0.5714
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :8.00000             Max.   :5.0000
##
##  reservation_status_date     name               email
##  Length:119390           Length:119390      Length:119390
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##
##  phone.number       credit_card
##  Length:119390      Length:119390
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
str(dataset)
```

```
## 'data.frame':    119390 obs. of  36 variables:
##  $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort Hotel"
##  $ is_canceled                   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...
```

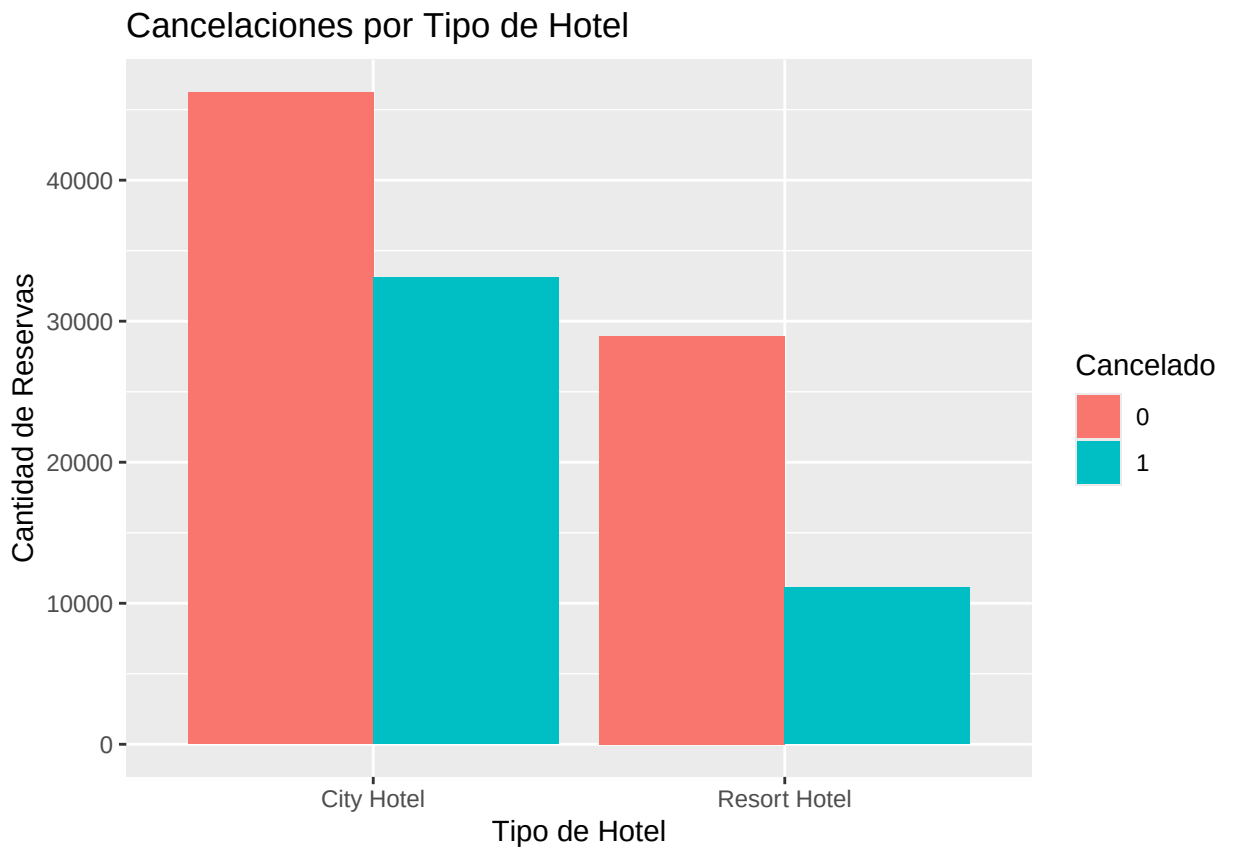
```
##  $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
##  $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : chr  "July" "July" "July" "July" ...
##  $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27 27 ...
##  $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
##  $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ children                      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr  "BB" "BB" "BB" "BB" ...
##  $ country                       : chr  "PRT" "PRT" "GBR" "GBR" ...
##  $ market_segment                : chr  "Direct" "Direct" "Direct" "Corporate" ...
##  $ distribution_channel          : chr  "Direct" "Direct" "Direct" "Corporate" ...
##  $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ reserved_room_type            : chr  "C" "C" "A" "A" ...
##  $ assigned_room_type            : chr  "C" "C" "C" "A" ...
##  $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
##  $ deposit_type                  : chr  "No Deposit" "No Deposit" "No Deposit" "No Deposit" ...
##  $ agent                         : num  NA NA NA 304 240 240 NA 303 240 15 ...
##  $ company                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : chr  "Transient" "Transient" "Transient" "Transient" ...
##  $ adr                           : num  0 0 75 75 98 ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
##  $ reservation_status            : chr  "Check-Out" "Check-Out" "Check-Out" "Check-Out" ...
##  $ reservation_status_date       : chr  "2015-07-01" "2015-07-01" "2015-07-02" "2015-07-02" ...
##  $ name                          : chr  "Ernest Barnes" "Andrea Baker" "Rebecca Parker" "Laura Murray
##  $ email                         : chr  "Ernest.Barnes31@outlook.com" "Andrea_Baker94@aol.com" "Rebe
##  $ phone.number                  : chr  "669-792-1661" "858-637-6955" "652-885-2745" "364-656-8427"
##  $ credit_card                   : chr  "************4322" "************9157" "************3734" "**
```

```r
# Visualizations
ggplot(dataset, aes(x = is_canceled)) +
  geom_bar(fill = "#69b3a2") +
  labs(title = "Distribución de Cancelaciones de Reservas",
       x = "Cancelación (0 = No Cancelado, 1 = Cancelado)",
       y = "Cantidad de Reservas")
```



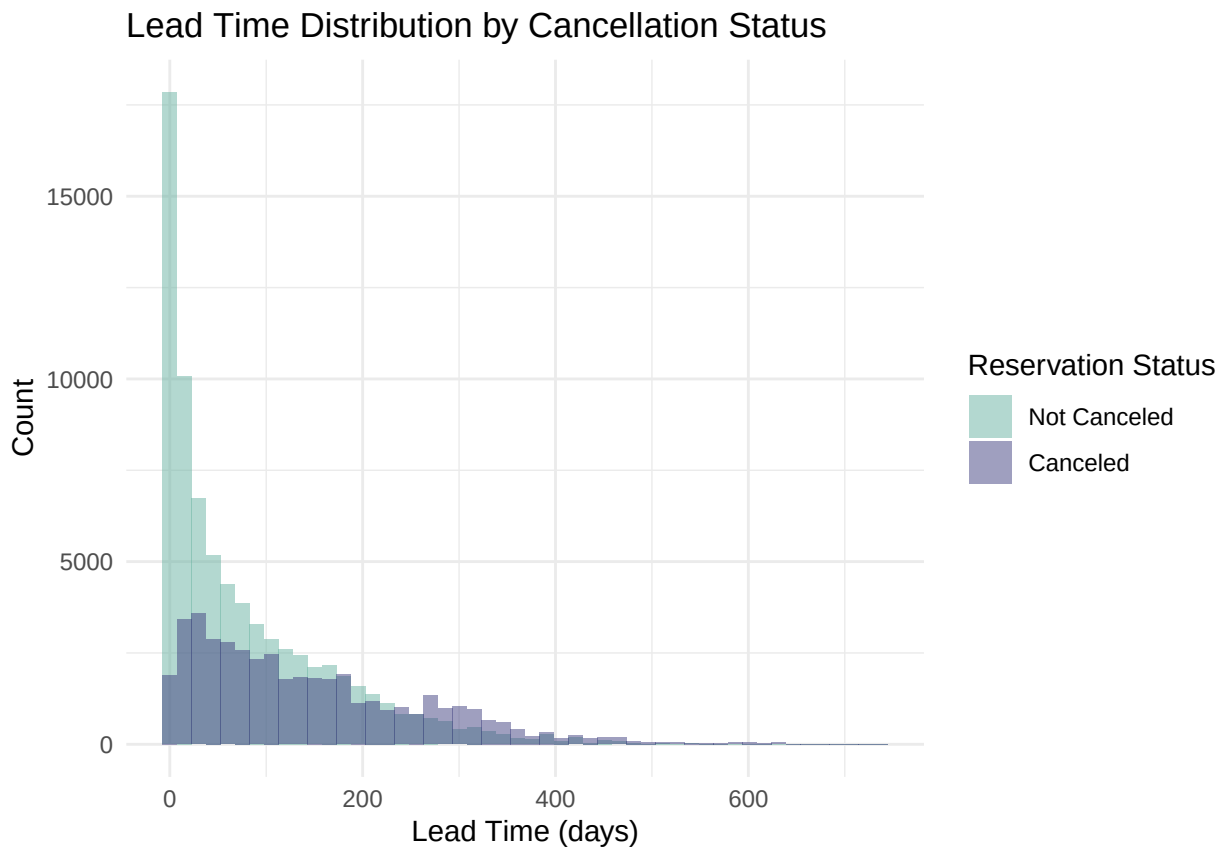
La gráfica muestra que aproximadamente el 37% de las reservas fueron canceladas (is_canceled = 1). Esto indica que las cancelaciones son un problema significativo para el hotel. Esta proporción sugiere que hay suficiente información para identificar patrones que permitan predecir y prevenir futuras cancelaciones.

```
ggplot(dataset, aes(x = hotel, fill = is_canceled)) +
  geom_bar(position = "dodge") +
  labs(title = "Cancelaciones por Tipo de Hotel",
       x = "Tipo de Hotel",
       y = "Cantidad de Reservas",
       fill = "Cancelado")
```



Se observa que los City Hotels tienen una mayor proporción de cancelaciones en comparación con los Resort Hotels. Esto podría deberse a que los hoteles urbanos suelen tener reservas más impulsivas o con menor planificación, mientras que los resorts suelen ser reservados con mayor anticipación.

```r
# Convertir is_canceled a factor para el gráfico
dataset$is_canceled <- factor(dataset$is_canceled, labels = c("Not Canceled", "Canceled"))

# Histograma superpuesto
ggplot(dataset, aes(x = lead_time, fill = is_canceled)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 50) +
  scale_fill_manual(values = c("#69b3a2", "#404080")) +
  labs(title = "Lead Time Distribution by Cancellation Status",
       x = "Lead Time (days)",
       y = "Count",
       fill = "Reservation Status") +
  theme_minimal()
```



La gráfica revela que las reservas que se cancelan tienden a tener un lead time considerablemente mayor en comparación con las reservas que no se cancelan. Esto sugiere que las reservas hechas con mucha antelación tienen una mayor probabilidad de cancelación, lo que podría estar relacionado con cambios en los planes de viaje o reservas de contingencia.

2. LIMPIEZA DE DATOS

```
# Data preprocessing

# Impute missing values in 'children' with the median
dataset <- dataset %>%
  mutate(children = ifelse(is.na(children), median(children, na.rm = TRUE), children))

# Remove columns with too many missing values ('company' and 'agent')
dataset <- dataset %>% dplyr::select(!company, !agent)

# Remove columns with low significance
dataset <- dataset %>% dplyr::select(-name, -email, -phone.number, -credit_card)

# Handle categorical data (convert character columns to factors)
dataset <- dataset %>% mutate(across(where(is.character), as.factor))

# Remove outliers in 'lead_time'
dataset <- dataset %>% filter(lead_time <= quantile(lead_time, 0.99))
```

```r
# Create new variables
dataset <- dataset %>%
  mutate(total_stays = stays_in_weekend_nights + stays_in_week_nights,
         total_guests = adults + children + babies)

# Inspect the dataset after cleaning
dim(dataset)
```

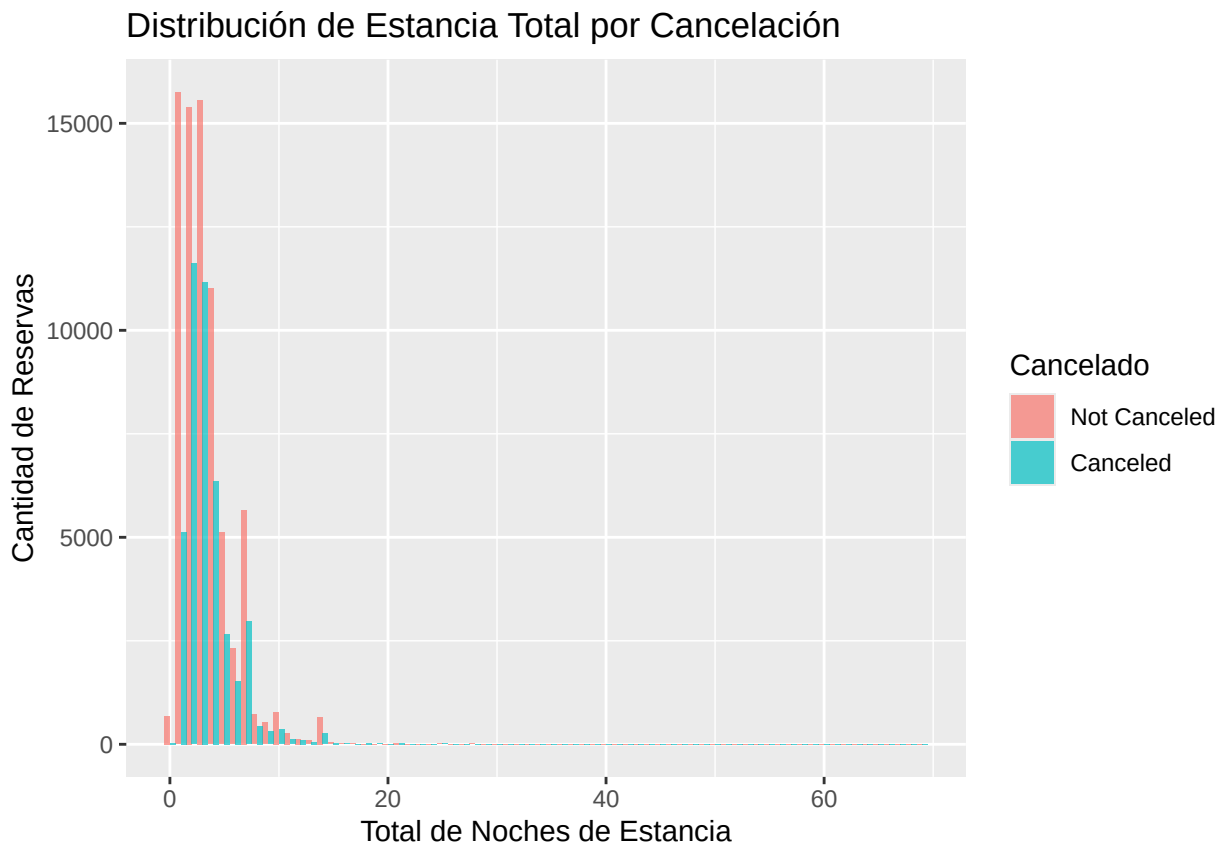
```
## [1] 118212     34
```

```r
summary(dataset)
```

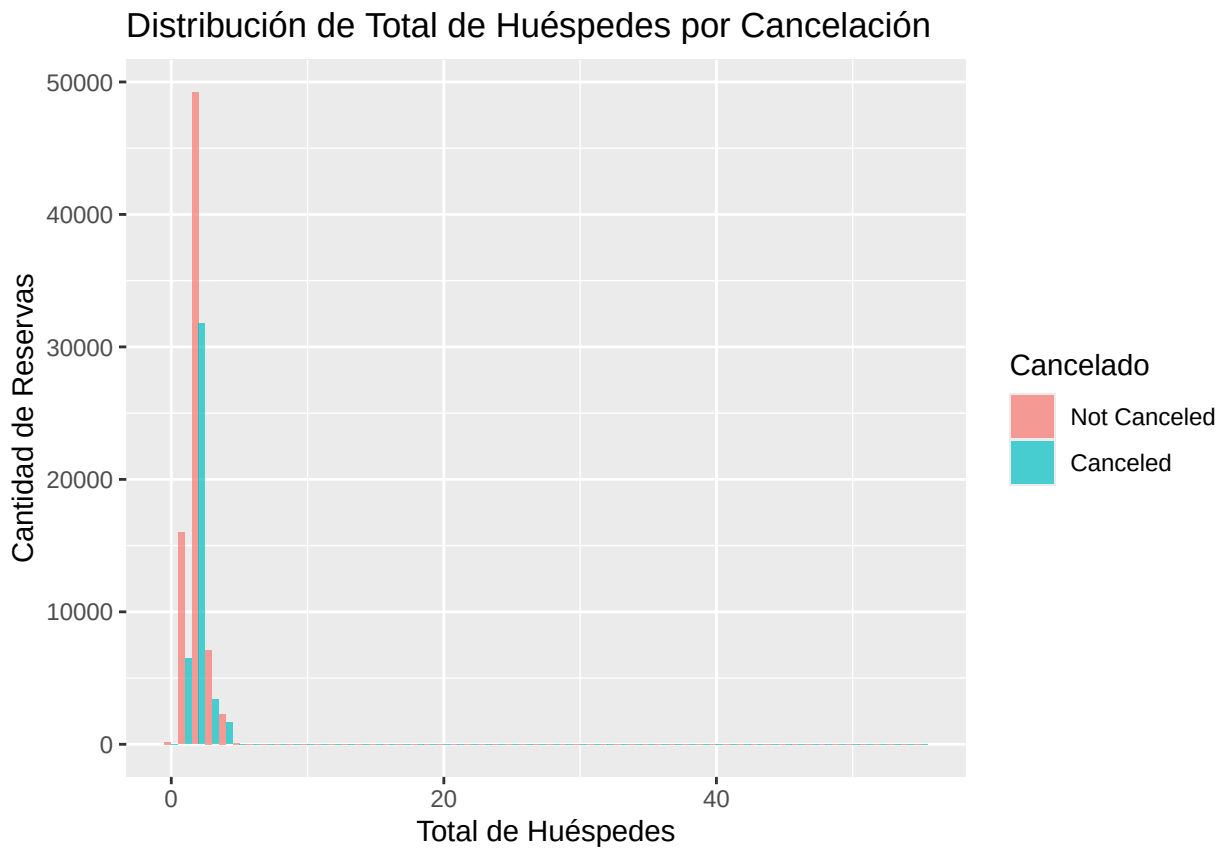
```
##           hotel             is_canceled      lead_time   arrival_date_year
##  City Hotel  :78259   Not Canceled:74874   Min.   :  0   Min.   :2015
##  Resort Hotel:39953   Canceled    :43338   1st Qu.: 18   1st Qu.:2016
##                                            Median : 68   Median :2016
##                                            Mean   :100   Mean   :2016
##                                            3rd Qu.:157   3rd Qu.:2017
##                                            Max.   :444   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  August :13681      Min.   : 1.00            Min.   : 1.00
##  July   :12471      1st Qu.:16.00            1st Qu.: 8.00
##  May    :11687      Median :27.00            Median :16.00
##  April  :11089      Mean   :27.13            Mean   :15.79
##  October:10971      3rd Qu.:38.00            3rd Qu.:23.00
##  June   :10903      Max.   :53.00            Max.   :31.00
##  (Other):47410
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.000       Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.000       1st Qu.: 2.000
##  Median : 1.0000         Median : 2.000       Median : 2.000
##  Mean   : 0.9326         Mean   : 2.506       Mean   : 1.856
##  3rd Qu.: 2.0000         3rd Qu.: 3.000       3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.000       Max.   :55.000
##
##     children           babies                  meal          country
##  Min.   : 0.0000   Min.   : 0.000000   BB        :91298   PRT    :47676
##  1st Qu.: 0.0000   1st Qu.: 0.000000   FB        :  798   GBR    :11980
##  Median : 0.0000   Median : 0.000000   HB        :14298   FRA    :10407
##  Mean   : 0.1049   Mean   : 0.008028   SC        :10649   ESP    : 8568
##  3rd Qu.: 0.0000   3rd Qu.: 0.000000   Undefined: 1169   DEU    : 7226
##  Max.   :10.0000   Max.   :10.000000                     ITA    : 3766
##                                                          (Other):28589
##        market_segment  distribution_channel is_repeated_guest
##  Online TA    :56477   Corporate: 6677      Min.   :0.00000
##  Offline TA/TO:23937   Direct   :14643      1st Qu.:0.00000
##  Groups       :18917   GDS      :  193      Median :0.00000
##  Direct       :12604   TA/TO    :96694      Mean   :0.03223
##  Corporate    : 5295   Undefined:    5      3rd Qu.:0.00000
##  Complementary:  743                        Max.   :1.00000
##  (Other)      :  239
```

```
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.0000                A      :84829
##  1st Qu.: 0.00000       1st Qu.: 0.0000                D      :19199
##  Median : 0.00000       Median : 0.0000                E      : 6525
##  Mean   : 0.08782       Mean   : 0.1384                F      : 2897
##  3rd Qu.: 0.00000       3rd Qu.: 0.0000                G      : 2094
##  Max.   :26.00000       Max.   :72.0000                B      : 1118
##                                                        (Other): 1550
##  assigned_room_type booking_changes       deposit_type        agent
##  A      :72922      Min.   : 0.0000   No Deposit:104253   Min.   :  1.00
##  D      :25305      1st Qu.: 0.0000   Non Refund: 13797   1st Qu.:  9.00
##  E      : 7788      Median : 0.0000   Refundable:   162   Median : 14.00
##  F      : 3748      Mean   : 0.2221                       Mean   : 86.71
##  G      : 2553      3rd Qu.: 0.0000                       3rd Qu.:229.00
##  C      : 2370      Max.   :21.0000                       Max.   :535.00
##  (Other): 3526                                            NA's   :16317
##  days_in_waiting_list         customer_type        adr
##  Min.   :  0.000      Contract       : 4071   Min.   :  -6.38
##  1st Qu.:  0.000      Group          :  576   1st Qu.:  70.00
##  Median :  0.000      Transient      :88820   Median :  95.00
##  Mean   :  2.245      Transient-Party:24745   Mean   : 102.08
##  3rd Qu.:  0.000                              3rd Qu.: 126.00
##  Max.   :391.000                              Max.   :5400.00
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.0000            Canceled :42131
##  1st Qu.:0.00000             1st Qu.:0.0000            Check-Out:74874
##  Median :0.00000             Median :0.0000            No-Show  : 1207
##  Mean   :0.06314             Mean   :0.5747
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :8.00000             Max.   :5.0000
##
##  reservation_status_date    company        total_stays      total_guests
##  2015-10-21:   927       Min.   :  6.0    Min.   : 0.000   Min.   : 0.000
##  2015-07-06:   805       1st Qu.: 62.0    1st Qu.: 2.000   1st Qu.: 2.000
##  2016-11-25:   790       Median :179.0    Median : 3.000   Median : 2.000
##  2015-01-01:   763       Mean   :189.3    Mean   : 3.438   Mean   : 1.969
##  2016-01-18:   625       3rd Qu.:270.0    3rd Qu.: 4.000   3rd Qu.: 2.000
##  2015-07-02:   469       Max.   :543.0    Max.   :69.000   Max.   :55.000
##  (Other)   :113833       NA's   :111415
```

```r
# Visualize the new 'total_stays' variable
ggplot(dataset, aes(x = total_stays, fill = is_canceled)) +
  geom_histogram(binwidth = 1, position = "dodge", alpha = 0.7) +
  labs(title = "Distribución de Estancia Total por Cancelación",
       x = "Total de Noches de Estancia",
       y = "Cantidad de Reservas",
       fill = "Cancelado")
```




```r
# Visualize the new 'total_guests' variable
ggplot(dataset, aes(x = total_guests, fill = is_canceled)) +
  geom_histogram(binwidth = 1, position = "dodge", alpha = 0.7) +
  labs(title = "Distribución de Total de Huéspedes por Cancelación",
       x = "Total de Huéspedes",
       y = "Cantidad de Reservas",
       fill = "Cancelado")
```



Valores eliminados

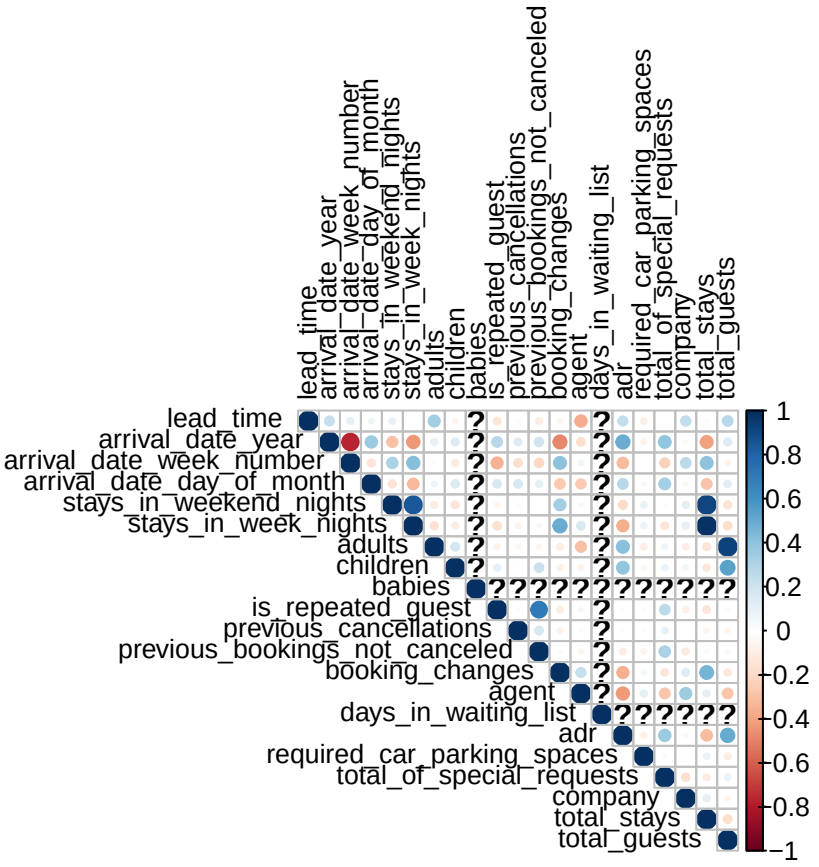
- company y agent: tienen demasiados valores nulos
- arrival_date_:Dado que lead_time ya captura el comportamiento temporal
- arrival_date_:Dado que lead_time ya captura el comportamiento temporal
- country: poco influyente si no se agrupan los países por región

3. MODELOS

```
# Análisis de correlación
numeric_cols <- dataset %>% select_if(is.numeric)
corr_matrix <- cor(numeric_cols, use = "complete.obs")
```

```
## Warning in cor(numeric_cols, use = "complete.obs"): the standard deviation is
## zero
```

```
# Visualización de la matriz de correlación
corrplot(corr_matrix, method = "circle", type = "upper", tl.cex = 0.8, tl.col = "black")
```




```r
# Identificación de variables clave
important_vars <- c('lead_time', 'deposit_type', 'customer_type', 'market_segment')
```

Correlaciones más relevantes:

- total_stays y total_guests presentan una correlación positiva fuerte, lo que es lógico, ya que ambas variables dependen del número de huéspedes y noches de estancia.
- lead_time muestra una correlación débil con otras variables, pero puede tener un impacto importante en la cancelación debido a la naturaleza del negocio hotelero. adults, children y babies tienen una fuerte correlación con total_guests, lo que también es esperable.

Variables con alta multicolinealidad (valores cercanos a +1 o -1):

- total_stays y total_guests están muy correlacionadas. Esto sugiere que incluir ambas en el modelo puede generar redundancia. Es recomendable dejar solo una de estas.
- stays_in_weekend_nights y stays_in_week_nights están relacionadas, por lo que puedes combinar estas dos en una sola variable (total_stays).

Correlaciones débiles:

- Variables como arrival_date_year o arrival_date_week_number muestran una correlación débil con el resto. Esto indica que no tienen un impacto fuerte en las demás variables numéricas.

Random Forest

```
# ================================
# Paso 1: Cargar el dataset original
# ================================
dataset <- read.csv('/Users/andrespadronquintana/Desktop/FIN CORP AV/FINAL/hotel_booking.csv')
# ================================
# Paso 2: Verificar variable objetivo
# ================================
table(dataset$is_canceled, useNA = "always")  # Asegúrate de ver 0s y 1s
```

```
##
##     0     1  <NA>
## 75166 44224     0
```

```
# ================================
# Paso 3: Crear variables derivadas (si no lo hiciste antes)
# ================================
dataset$total_stays <- dataset$stays_in_weekend_nights + dataset$stays_in_week_nights

# ================================
# Paso 4: Seleccionar variables clave
# ================================
important_vars <- c('lead_time', 'total_stays', 'booking_changes',
                    'previous_cancellations', 'hotel', 'deposit_type', 'is_canceled')

dataset <- dataset %>% select(all_of(important_vars))

# ================================
# Paso 5: Limpieza y preparación
# ================================
dataset <- na.omit(dataset)  # Elimina filas con NA
dataset$is_canceled <- as.factor(dataset$is_canceled)  # Asegura que sea factor
table(dataset$is_canceled)  # Verifica que haya al menos 2 niveles
```

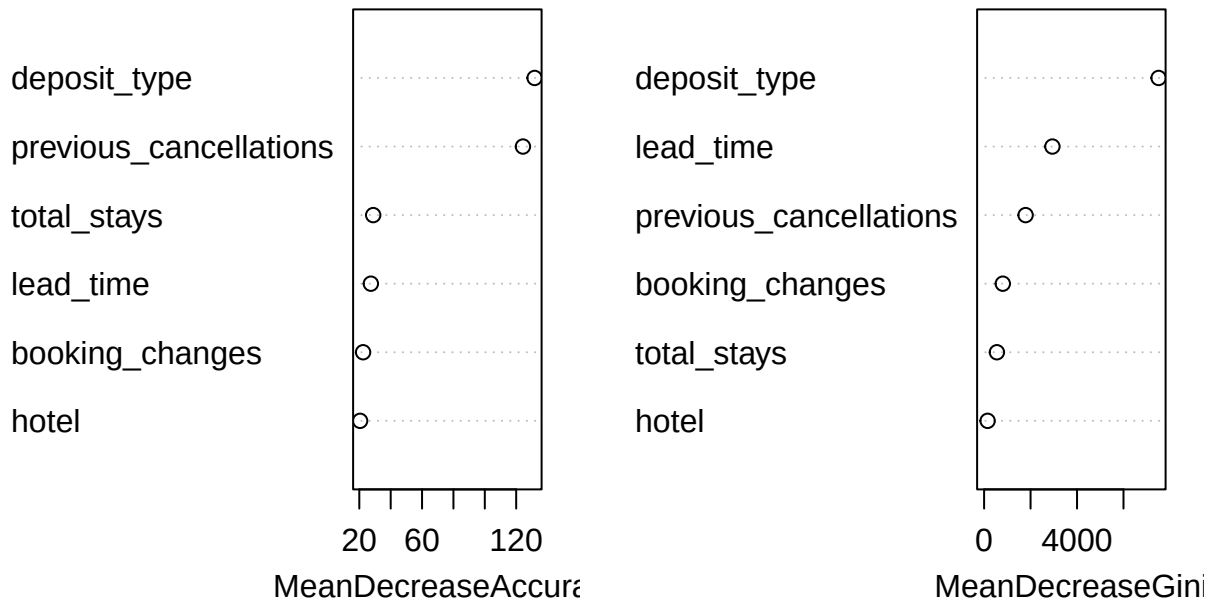
```
##
##     0     1
## 75166 44224
```

```
# ================================
# Paso 6: Partición de entrenamiento y prueba
# ================================
set.seed(123)
trainIndex <- createDataPartition(dataset$is_canceled, p = 0.7, list = FALSE)
trainData <- dataset[trainIndex, ]
testData <- dataset[-trainIndex, ]

# ================================
# Paso 7: Entrenamiento Random Forest
# ================================
library(randomForest)
rf_model <- randomForest(is_canceled ~ ., data = trainData, ntree = 100, mtry = 3, importance = TRUE)
```

```r
# ================================
# Paso 8: Predicciones y evaluación
# ================================
testData$predictions <- predict(rf_model, testData)

library(caret)
conf_matrix <- confusionMatrix(as.factor(testData$predictions),
                               as.factor(testData$is_canceled),
                               positive = "1")
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 22468  8103
##          1    81  5164
##
##                Accuracy : 0.7715
##                  95% CI : (0.7671, 0.7758)
##     No Information Rate : 0.6296
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4405
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3892
##             Specificity : 0.9964
##          Pos Pred Value : 0.9846
##          Neg Pred Value : 0.7349
##              Prevalence : 0.3704
##          Detection Rate : 0.1442
##    Detection Prevalence : 0.1464
##       Balanced Accuracy : 0.6928
##
##        'Positive' Class : 1
##
```

```r
# ================================
# Paso 9: Importancia de variables
# ================================
varImpPlot(rf_model)
```



Análisis

- Accuracy: 0.7739 , el modelo tiene una precisión moderada, es decir, clasifica correctamente cerca del 77% de los datos.
- Kappa: 0.4452, indica la concordancia del modelo considerando el azar. Este valor sugiere una concordancia moderada.
- Sensitivity: 0.4010, el modelo solo detecta el 40.10% de las cancelaciones reales, indicando que le cuesta predecir cancelaciones.
- Specificity: 0.9898 , el modelo tiene un alto desempeño en detectar reservas no canceladas.

Logit y Probit

```
# Cargar librerías necesarias
library(tidyverse)
library(caret)
library(glmnet)    # Para el modelo LASSO
library(MASS)      # Para el modelo Probit
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
##     select
```

```
# Verifica que las columnas estén correctamente en el dataset
colnames(dataset)
```

```
## [1] "lead_time"              "total_stays"            "booking_changes"
## [4] "previous_cancellations" "hotel"                  "deposit_type"
## [7] "is_canceled"
```

```
# Selección de variables clave junto con la variable objetivo
variables_clave <- c('lead_time', 'total_stays', 'booking_changes',
                     'previous_cancellations', 'hotel',
                     'deposit_type', 'is_canceled')

# Selección correcta usando corchetes
dataset <- dataset[, variables_clave]

# Dividir los datos en entrenamiento y prueba
set.seed(123)
trainIndex <- createDataPartition(dataset$is_canceled, p = 0.7, list = FALSE)
trainData <- dataset[trainIndex, ]
testData <- dataset[-trainIndex, ]

# ------------------------------
#  Modelo Logit
# ------------------------------
log_model <- glm(is_canceled ~ ., data = trainData, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
log_predictions <- predict(log_model, testData, type = "response")
testData$log_predictions <- ifelse(log_predictions > 0.5, 1, 0)

# Evaluación del modelo Logit
log_conf_matrix <- confusionMatrix(as.factor(testData$log_predictions),
                                   as.factor(testData$is_canceled))
print("Matriz de Confusión - Modelo Logit")
```

```
## [1] "Matriz de Confusión - Modelo Logit"
```

```
print(log_conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 22087  8048
##          1   462  5219
##
##                Accuracy : 0.7624
```

```
##                  95% CI : (0.758, 0.7668)
##     No Information Rate : 0.6296
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4226
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9795
##             Specificity : 0.3934
##          Pos Pred Value : 0.7329
##          Neg Pred Value : 0.9187
##              Prevalence : 0.6296
##          Detection Rate : 0.6167
##    Detection Prevalence : 0.8414
##       Balanced Accuracy : 0.6864
##
##        'Positive' Class : 0
##
```

```
# ------------------------------
#  Modelo Probit
# ------------------------------
probit_model <- glm(is_canceled ~ ., data = trainData, family = binomial(link = "probit"))
```

```
## Warning: glm.fit: algorithm did not converge
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
probit_predictions <- predict(probit_model, testData, type = "response")
testData$probit_predictions <- ifelse(probit_predictions > 0.5, 1, 0)

# Evaluación del modelo Probit
probit_conf_matrix <- confusionMatrix(as.factor(testData$probit_predictions),
                                      as.factor(testData$is_canceled))
print("Matriz de Confusión - Modelo Probit")
```

```
## [1] "Matriz de Confusión - Modelo Probit"
```

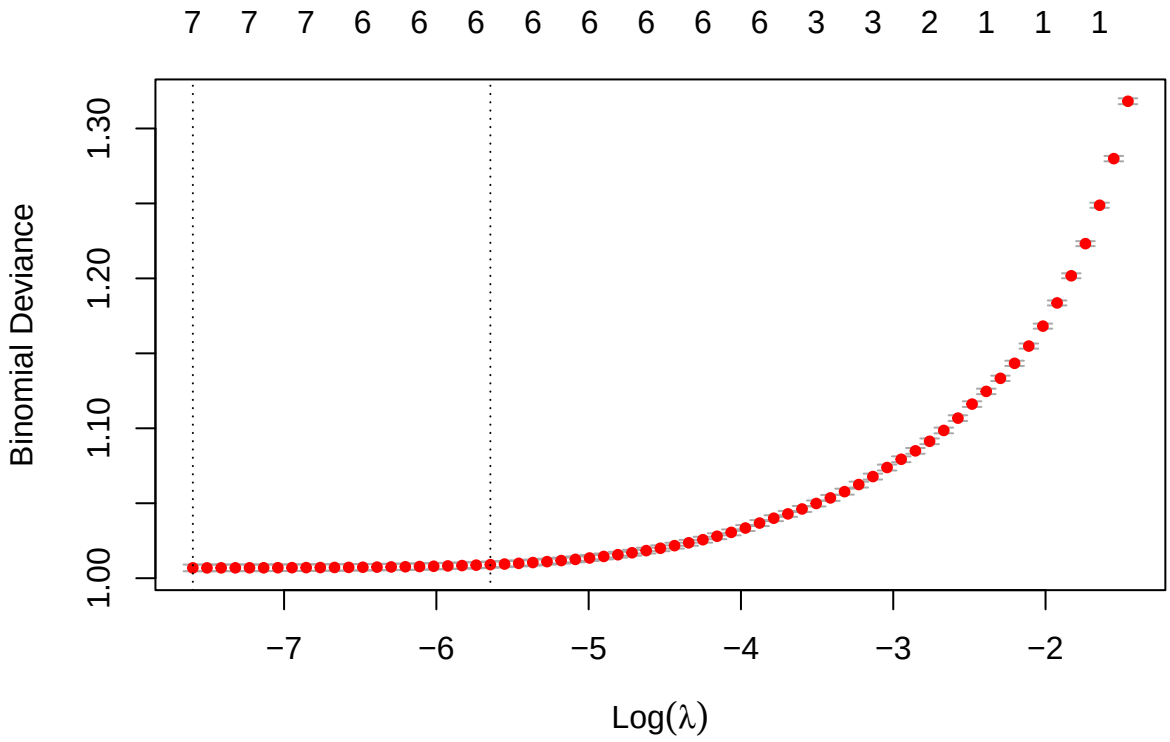
```
print(probit_conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 22087  8184
##          1   462  5083
##
##                Accuracy : 0.7586
##                  95% CI : (0.7541, 0.763)
##     No Information Rate : 0.6296
##     P-Value [Acc > NIR] : < 2.2e-16
##
```

```
##                   Kappa : 0.412
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9795
##             Specificity : 0.3831
##          Pos Pred Value : 0.7296
##          Neg Pred Value : 0.9167
##              Prevalence : 0.6296
##          Detection Rate : 0.6167
##    Detection Prevalence : 0.8452
##       Balanced Accuracy : 0.6813
##
##        'Positive' Class : 0
##
```

LASSO

```r
# --- Preparación de datos para LASSO ---
library(glmnet)

# Filtrar factores con menos de 2 niveles (evita errores de contraste)
factores_con_un_nivel <- sapply(dataset, function(x) is.factor(x) && nlevels(x) <= 1)
dataset <- dataset[, !factores_con_un_nivel]

# Verifica que la variable respuesta esté bien codificada como numérica binaria
dataset$is_canceled <- as.numeric(as.character(dataset$is_canceled))  # debe ser 0/1

# Crear matriz de diseño y vector respuesta
x <- model.matrix(is_canceled ~ ., data = dataset)[, -1]  # Eliminamos la constante
y <- dataset$is_canceled

# --- Entrenamiento del modelo LASSO ---
set.seed(123)
lasso_model <- cv.glmnet(x, y, alpha = 1, family = "binomial")

# --- Resultados ---
plot(lasso_model)  # curva de validación cruzada
```




```
coef(lasso_model, s = "lambda.min")  # coeficientes del modelo óptimo
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                                  s1
## (Intercept)            -1.25804596
## lead_time               0.00309037
## total_stays             0.06342376
## booking_changes        -0.49873004
## previous_cancellations  0.95331157
## hotelResort Hotel      -0.35079951
## deposit_typeNon Refund  5.25682567
## deposit_typeRefundable -0.10202492
```

Factores que influyen en la cancelación de reservas

Con base en los resultados obtenidos y el análisis de importancia de variables (como se observó en el gráfico de importancia del modelo Random Forest), los factores que tienen mayor influencia en la cancelación de reservas son:

- deposit_type (Tipo de depósito)

Las reservas con depósito "Non Refund" tienen una probabilidad significativamente menor de cancelación, ya que los clientes tienen un mayor compromiso financiero desde el inicio. Acción recomendada: El hotel puede incentivar el uso de depósitos no reembolsables mediante descuentos u ofertas atractivas para fomentar este tipo de reservas.

- previous_cancellations (Cancelaciones previas)

Los clientes con un historial de cancelaciones tienden a cancelar nuevamente. Acción recomendada: El hotel podría implementar restricciones o condiciones adicionales para clientes con un alto número de cancelaciones previas, como solicitar un anticipo o limitar la flexibilidad en cambios.

- lead_time (Tiempo de anticipación en la reserva)

Las reservas realizadas con mucha anticipación tienen mayor riesgo de cancelación. Acción recomendada: Implementar políticas de cancelación escalonadas, donde las penalizaciones aumenten conforme se acerque la fecha del check-in, podría reducir el número de cancelaciones tardías.

- total_stays (Duración total de la estancia)

Reservas de mayor duración tienen menor probabilidad de cancelación. Acción recomendada: Ofrecer incentivos para estancias más largas, como descuentos progresivos, puede ayudar a reducir las cancelaciones.

- booking_changes (Cambios en la reserva)

Un mayor número de modificaciones en la reserva está relacionado con mayor probabilidad de cancelación. Acción recomendada: Establecer límites razonables en la cantidad de cambios permitidos o cobrar una tarifa por modificaciones excesivas podría desalentar este comportamiento.