Demo Day Módulo 2 RStudio

PERLA CONCHITA PEÑA CAMPOS

Contenido

1. Data

2. EDA

3. Pruebas estadísticas

4. Métodos de clasificación

Data set

Decribe los datos de dos hoteles del 01 julio 2015 al 31 Agosto 2017

- H1: Hotel resort

- H2: Hotel ciudad

La variable que se estudia es la cancelación de la reservación con respecto al resto.

Se hizo limpieza de la base de datos y transformación de variables (revisar pdf DemoDay_M2)

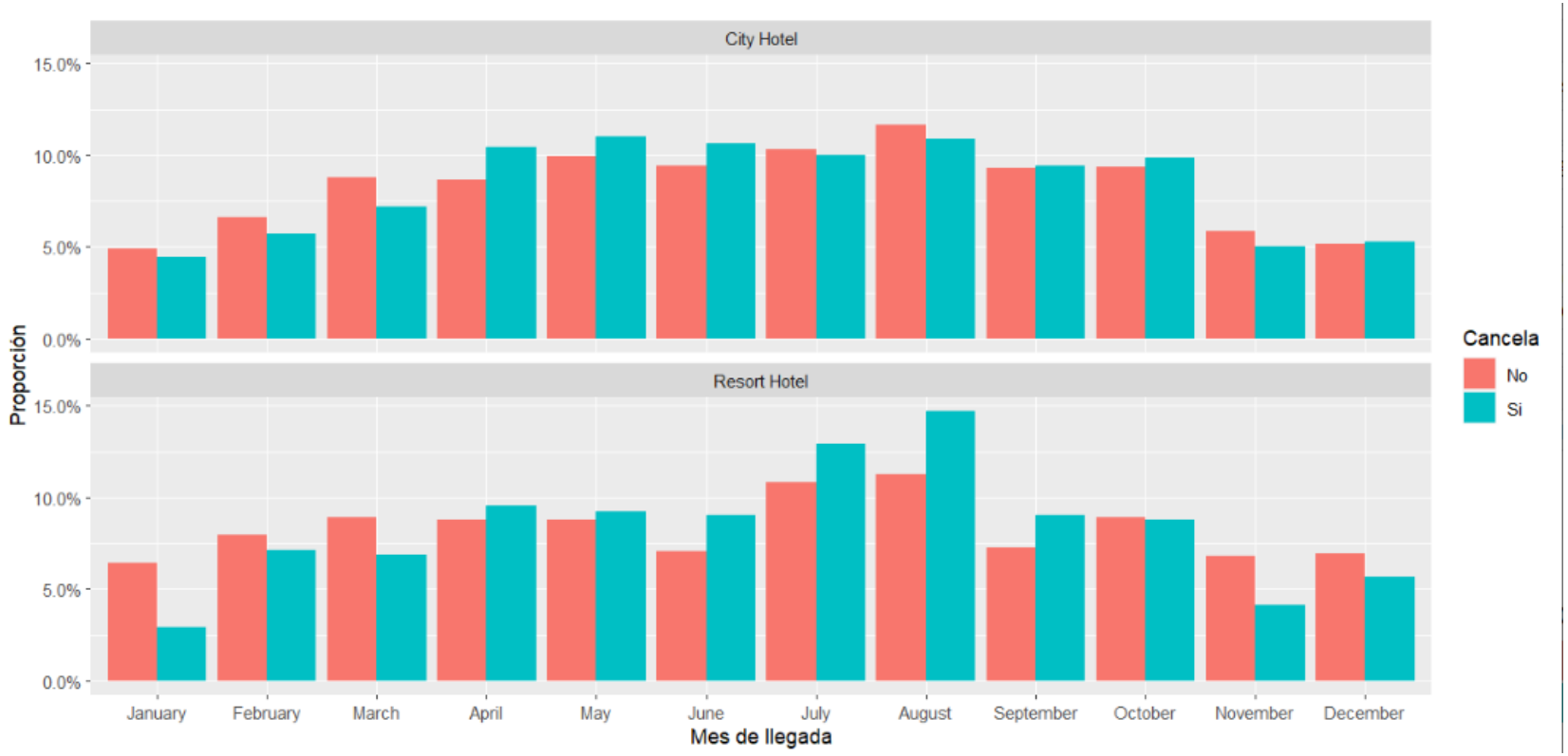
https://www.sciencedirect.com/science/article/pii/S2352340918315191

```
> str(booking)
'data.frame': 119390 obs. of  32 variables:
 $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "R
 $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
 $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
 $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 20
 $ 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                      : int  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" "Corp
 $ distribution_channel          : chr  "Direct" "Direct" "Direct" "Corp
 $ 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 De
```

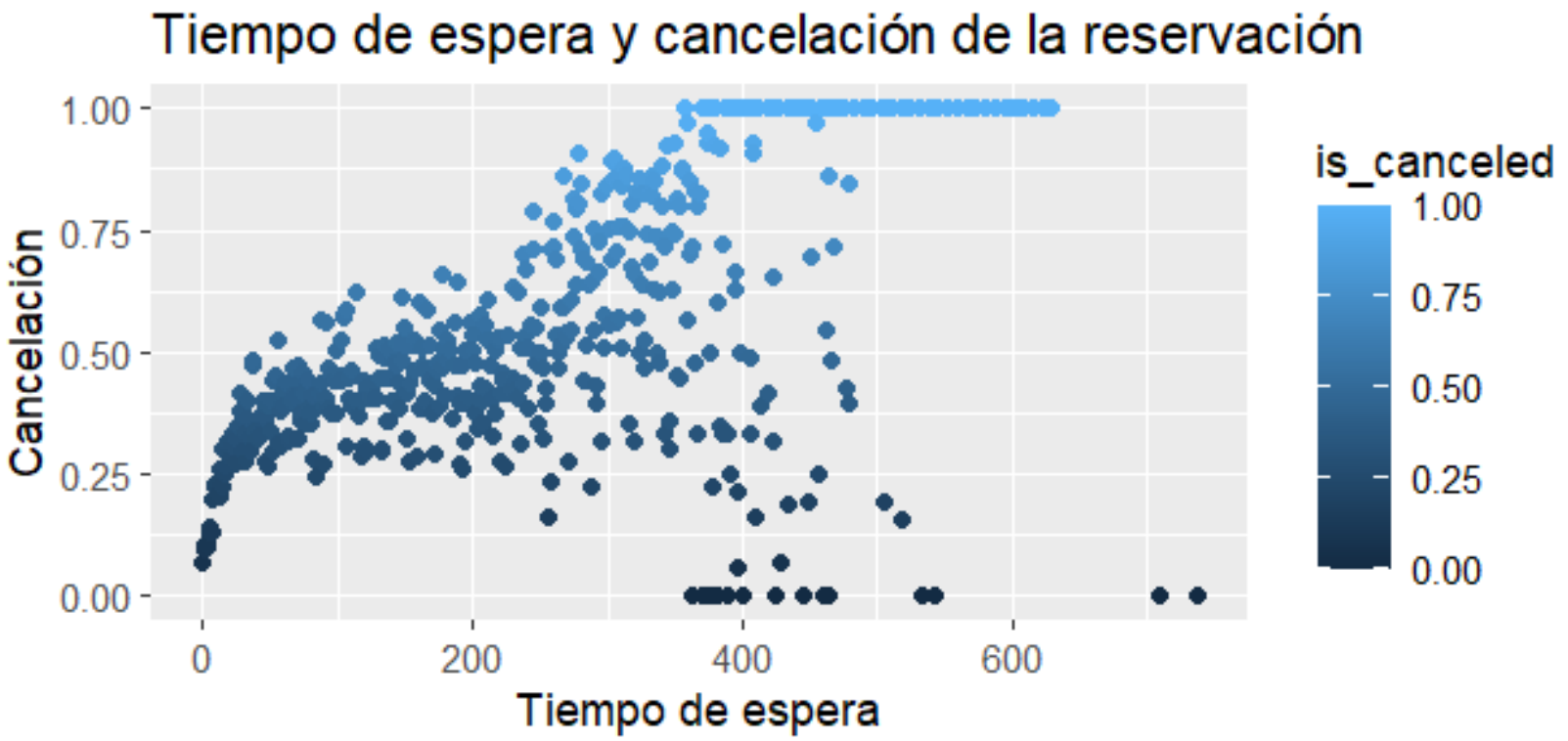
EDA

EJEMPLOS

Reservaciones por mes



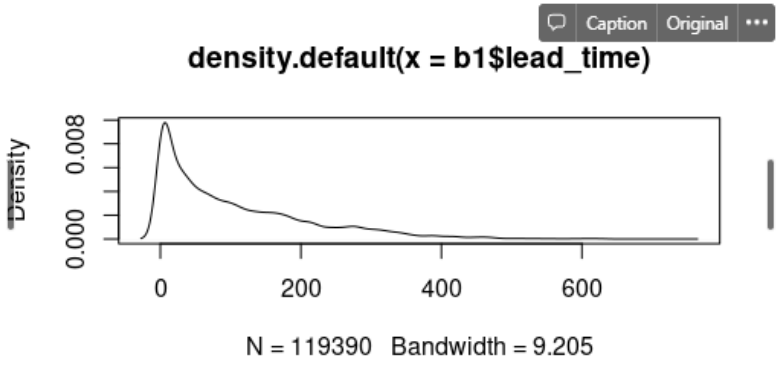
Tiempo de espera y cancelación



Prueba estadística de hipótesis

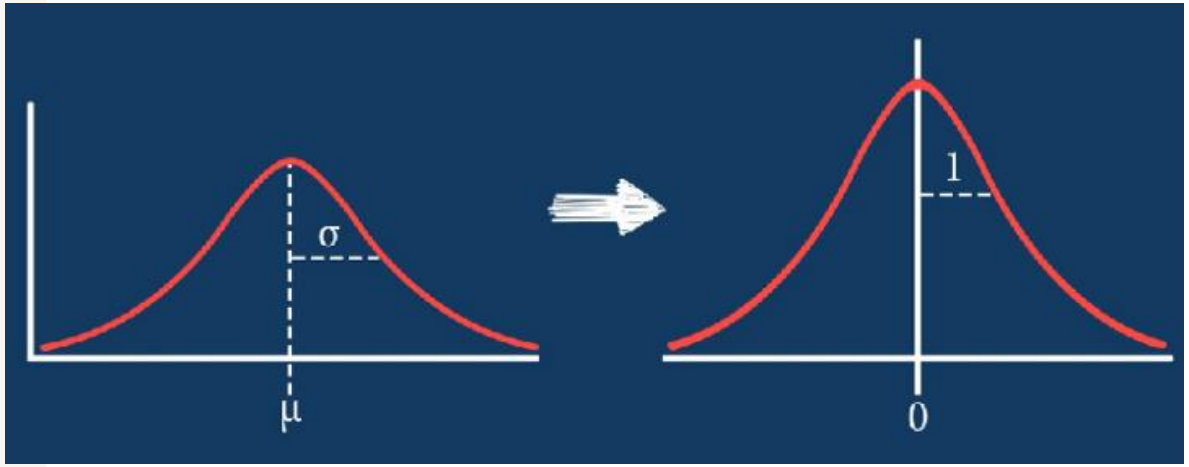
Prueba de normalidad *ejemplo de una variable

```r
#Verificar la normalidad de los datos con un histograma
hist(b1$lead_time)

#Densidad
d <- density(b1$lead_time)
plot(d)

#Prueba de  shapiro test
lt.test <- shapiro.test(b1$lead_time[0:5000])
lt.test

Shapiro-Wilk normality test

data:  b1$lead_time[0:5000]
W = 0.88581, p-value < 2.2e-16
```



Normalización datos

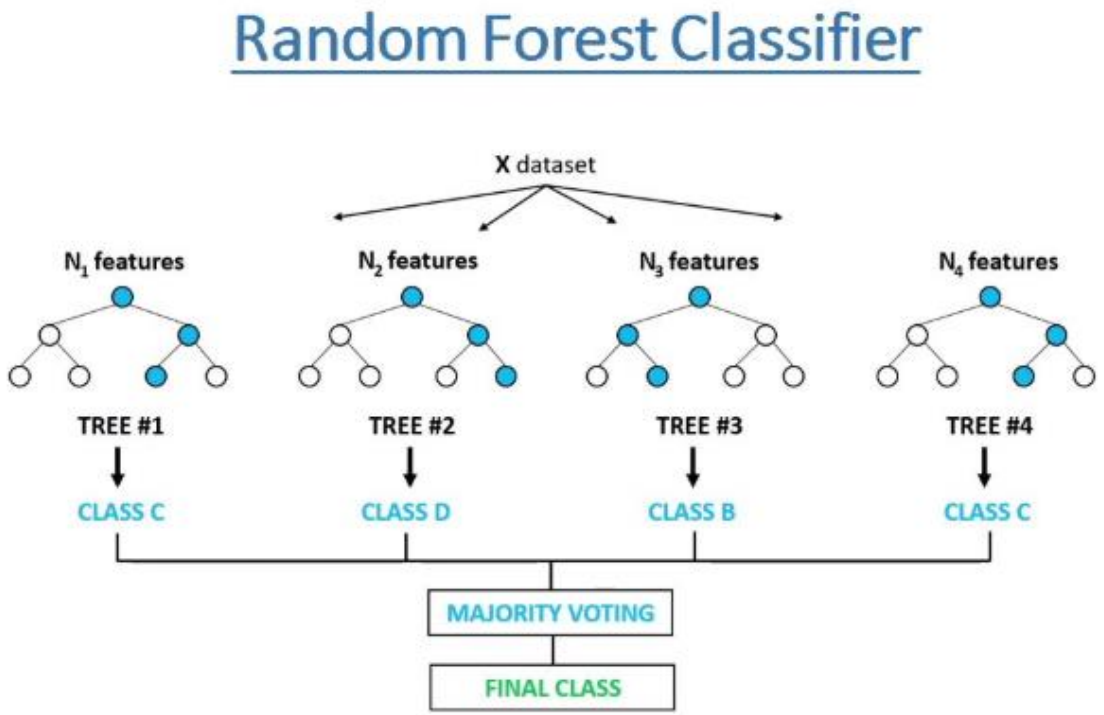
```r
#Estandarización de la base de datos (usada para SVM)
## Preparar los datos
hotel_df <- hotel_stays%>%
  select(is_canceled, hotel, arrival_date_month, meal,
         adr, deposit_type, lead_time, adults, required_car_parking_spaces,
         total_of_special_requests, market_segment,
         stays_in_week_nights, stays_in_weekend_nights)%>%
  mutate_if(is.character,factor)

#Instalar biblioteca
install.packages("tidymodels")
library(tidymodels)

#Normalización base de datos
hotel_rec <- recipe(is_canceled ~., data = hotel_train) %>%
  step_dummy(all_nominal(), -all_outcomes()) %>%
  step_zv(all_numeric())%>%
  step_normalize(all_numeric()) %>%
  prep()
```



Clasificación

RANDOM FOREST, REGRESIÓN LOGÍSTICA Y SVM

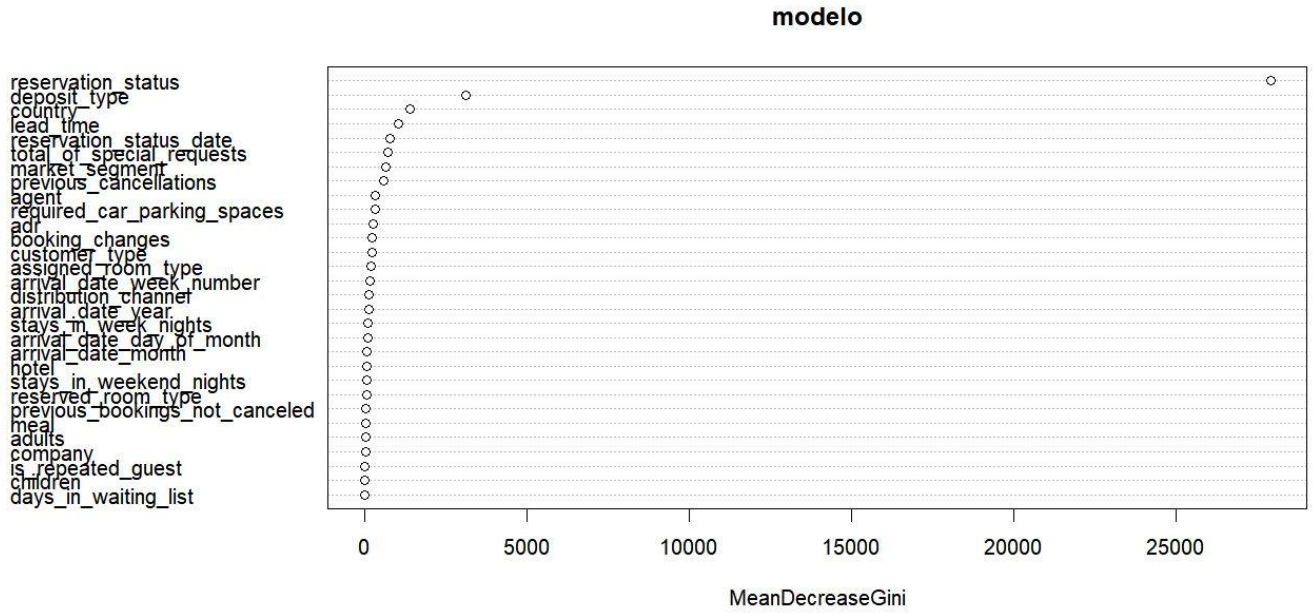
Random Forest



Importancia de las variables

- Se rescatan del modelo:
  - Tipo de depósito
  - País
  - Tiempo de espera
  - Total de requisites especiales
  - Precio promedio habitación

  - OOB 0.01%



Regresión logística

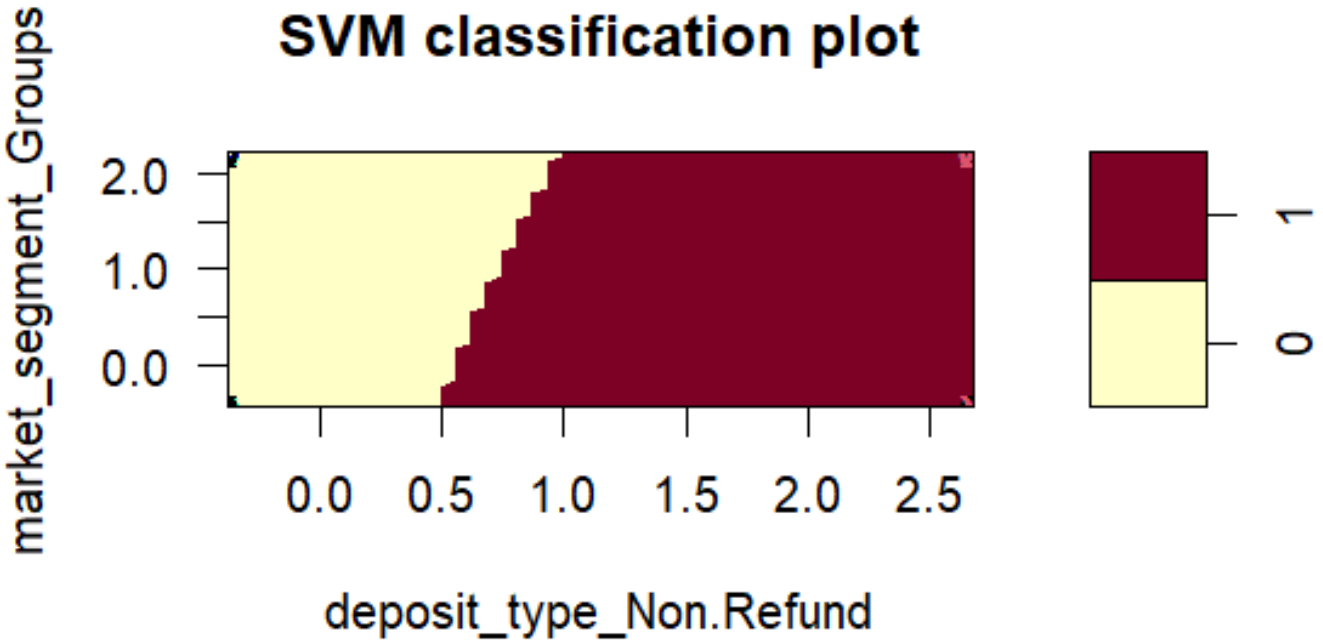
Regresión logística. Coeficientes

Las variables con mayor significancia son:

- Tipo de depósito
- Tiempo de espera
- Segmento de mercado
- Precio promedio de la habitación

```
                            Pr(>|z|)
(Intercept)                  < 2e-16 ***
deposit_typeNon Refund       < 2e-16 ***
deposit_typeRefundable      0.198390
lead_time                    < 2e-16 ***
market_segmentComplementary 0.037156 *
market_segmentCorporate     0.000517 ***
market_segmentDirect        6.59e-06 ***
market_segmentGroups        0.130061
market_segmentOffline TA/TO 4.41e-06 ***
market_segmentOnline TA     0.081027 .
market_segmentUndefined     0.829173
adr                          < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SVM

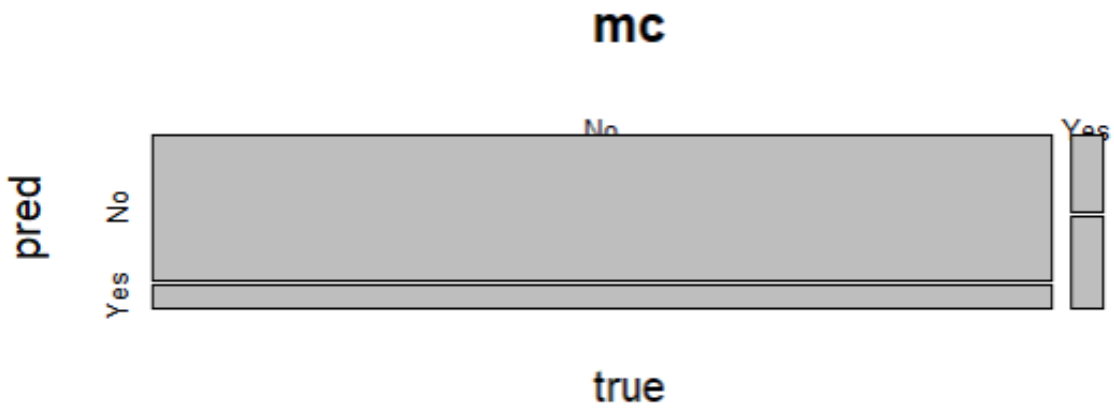

SVM classification plot
market_segment_Groups
2.0
1.0
0.0
0.0
0.5
1.0
1.5
2.0
2.5
deposit_type_Non.Refund
0
1

SVM

Variables.

- Cancelación (VD)
- Tipo de hotel
- Segmento de mercado
- Mes de llegada
- Comidas
- Precio promedio habitación
- Total de requerimientos especiales
- Total de noches reservadas

Accurancy: 85.22%



Dashboard (proceso)

HTTPS://TN3JN6-PERLA-CONCHITA.SHINYAPPS.IO/HOTEL_BOOKING/

Conclusiones

Proyecto

- SVM mejor modelo. Ajustar el error.
- Probar Nearest Neighbors
- Terminar dashboard
- Graficar de manera atractiva los resultados del modelo

Personales

- Agregar la documentación al entregable
- Probar gráficos de comparación
- Intentar otros problemas