HOTEL BOOKING DEMAND

ASIM AKÇAY

```
#install.packages("readr")
#install.packages("magrittr")
library(readr)
```

```
library(readxl)

hotel_bookings <- read.csv("hotel_bookings.csv", na.strings = "NULL")
View(hotel_bookings)
```

Task 1 - Problem: Conduct an analysis on the Hotel Booking Demand dataset to gain insights on hotel reservations and use this information to predict future demand.

Features: Location of the hotel (City Hotel or Resort Hotel) Information provided by customers when making reservations When the reservation was made Whether the reservation was cancelled

Objective: Understand when customers make hotel reservations and which features are in higher demand Predict future demand and provide recommendations to hotel owners Use regression analysis and time series analysis to predict future demand.

Task 2 — The Hotel Booking Demand dataset contains over 32,000 hotel reservations and 31 variables (columns). The following terms apply: Size: The dataset has 119,390 rows and 31 variables (columns) Variable Types: The variables in the dataset have the following types: Hotel Location (categorical) Hotel Type (categorical) Customer Information (categorical and numerical) Reservation Information (categorical and numerical) Cancellation Status (categorical)

Lead time: The number of days between the date of booking and the arrival date

ADR: Average Daily Rate per occupied room

Deposit Type: The type of deposit made for the reservation

```r
str(hotel_bookings)
```

```
'data.frame':   119390 obs. of  32 variables:
 $ hotel                         : chr  "Resort Hotel" "Resort Hotel" "Resort Hotel" "Resort
 $ is_canceled                   : int  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 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                      : chr  "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                         : int  NA NA NA 304 240 240 NA 303 240 15 ...
 $ company                       : int  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"
```

To list the name, type, and first few values of all variables in the dataset:

```r
head(hotel_bookings)
```

```
         hotel is_canceled lead_time arrival_date_year arrival_date_month
1 Resort Hotel           0       342              2015               July
2 Resort Hotel           0       737              2015               July
3 Resort Hotel           0         7              2015               July
4 Resort Hotel           0        13              2015               July
5 Resort Hotel           0        14              2015               July
6 Resort Hotel           0        14              2015               July
  arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
1                       27                         1                      0
2                       27                         1                      0
3                       27                         1                      0
4                       27                         1                      0
5                       27                         1                      0
6                       27                         1                      0
  stays_in_week_nights adults children babies meal country market_segment
1                    0      2        0      0   BB     PRT         Direct
2                    0      2        0      0   BB     PRT         Direct
3                    1      1        0      0   BB     GBR         Direct
4                    1      1        0      0   BB     GBR      Corporate
5                    2      2        0      0   BB     GBR      Online TA
6                    2      2        0      0   BB     GBR      Online TA
  distribution_channel is_repeated_guest previous_cancellations
1               Direct                 0                      0
2               Direct                 0                      0
3               Direct                 0                      0
4            Corporate                 0                      0
5                TA/TO                 0                      0
6                TA/TO                 0                      0
  previous_bookings_not_canceled reserved_room_type assigned_room_type
1                              0                 C                 C
2                              0                 C                 C
3                              0                 A                 C
4                              0                 A                 A
5                              0                 A                 A
6                              0                 A                 A
  booking_changes deposit_type agent company days_in_waiting_list customer_type
1               3   No Deposit    NA      NA                    0     Transient
2               4   No Deposit    NA      NA                    0     Transient
3               0   No Deposit    NA      NA                    0     Transient
4               0   No Deposit   304      NA                    0     Transient
5               0   No Deposit   240      NA                    0     Transient
6               0   No Deposit   240      NA                    0     Transient
  adr required_car_parking_spaces total_of_special_requests reservation_status
```

```
1   0                          0                          0          Check-Out
2   0                          0                          0          Check-Out
3  75                          0                          0          Check-Out
4  75                          0                          0          Check-Out
5  98                          0                          1          Check-Out
6  98                          0                          1          Check-Out
  reservation_status_date
1              2015-07-01
2              2015-07-01
3              2015-07-02
4              2015-07-02
5              2015-07-03
6              2015-07-03
```

Code to see statistical summary of variables in dataset:

```
summary(hotel_bookings)
```

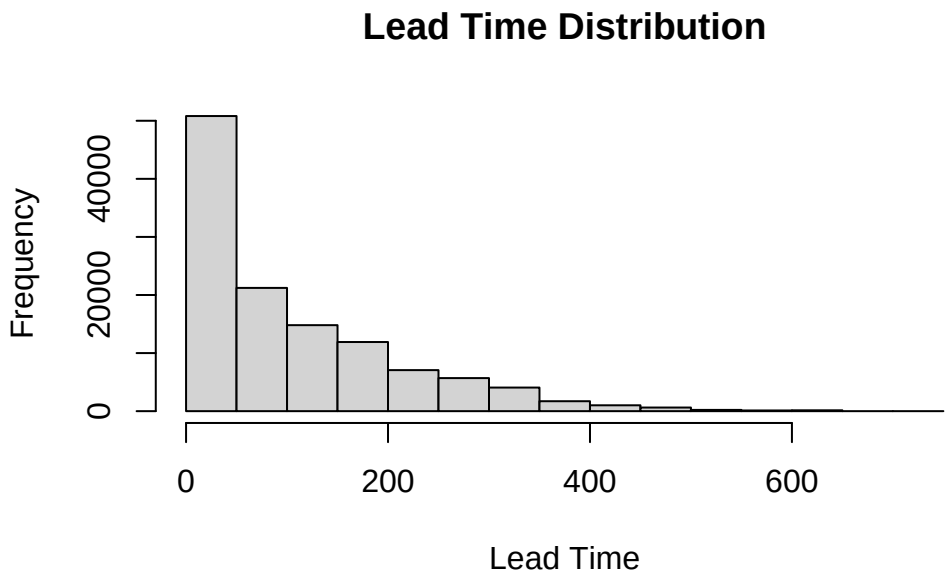
```
   hotel            is_canceled       lead_time   arrival_date_year
Length:119390      Min.   :0.0000   Min.   :  0   Min.   :2015
Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
Mode  :character   Median :0.0000   Median : 69   Median :2016
                   Mean   :0.3704   Mean   :104   Mean   :2016
                   3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017
                   Max.   :1.0000   Max.   :737   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
Length:119390      Min.   : 0.000000   Length:119390      Length:119390
Class :character   1st Qu.: 0.000000   Class :character   Class :character
Mode  :character   Median : 0.000000   Mode  :character   Mode  :character
                   Mean   : 0.007949
                   3rd Qu.: 0.000000
                   Max.   :10.000000

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
Length:119390
Class :character
Mode  :character
```

visualize the distribution of variables in the dataset:

```
hist(hotel_bookings$lead_time, main="Lead Time Distribution", xlab="Lead Time")
```

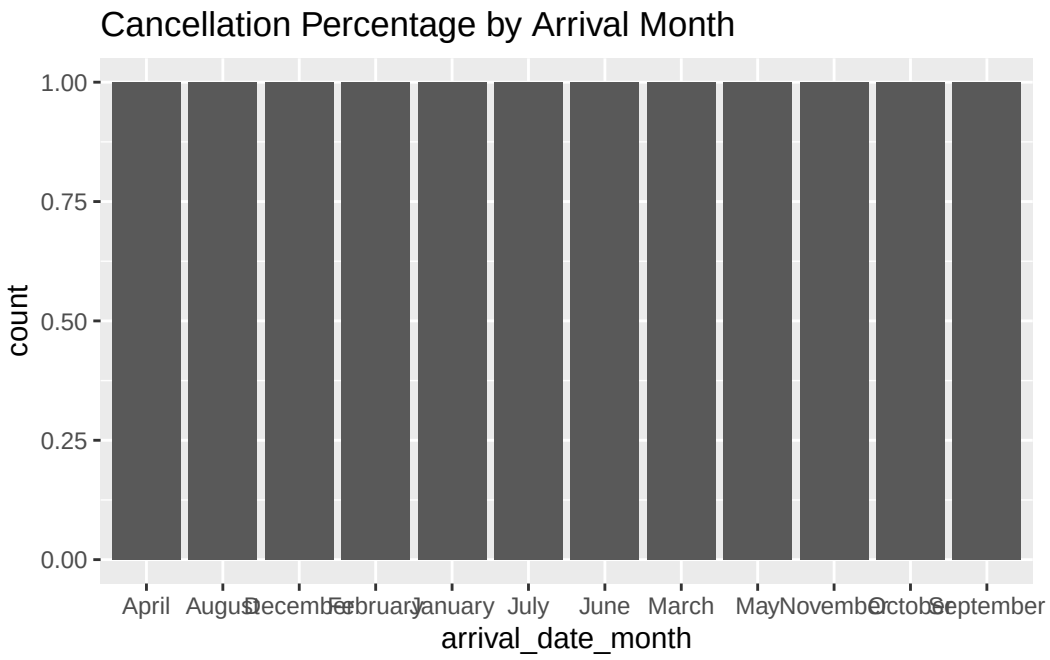


A bar chart showing the percentage of the variable "is_canceled" relative to the variable "arrival_date_month"

```
library(ggplot2)

ggplot(hotel_bookings, aes(x = arrival_date_month, fill = is_canceled, group = arrival_dat
```

```
    geom_bar(position="fill") +
    ggtitle("Cancellation Percentage by Arrival Month")
```

```
Warning: The following aesthetics were dropped during statistical transformation: fill
i This can happen when ggplot fails to infer the correct grouping structure in
  the data.
i Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```




```
set.seed(123) # for reproducibility
index <- sample(1 : nrow(hotel_bookings), round(nrow(hotel_bookings) * 0.80))
train <- hotel_bookings[index, ]
test <- hotel_bookings[-index, ]
```

TASK 3 In this dataset, we have a binary response variable named "is_canceled", which indicates whether a reservation has been canceled. Therefore, logistic regression to model the probability of cancellation given predictors

```
otel_verisi <- read.csv("hotel_bookings.csv", stringsAsFactors = FALSE, na.strings = "NULL
# Fitting a logistic regression model
model <- glm(is_canceled ~ lead_time + arrival_date_month + stays_in_weekend_nights + stay
```

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

```r
# Model summary
summary(model)
```

```
Call:
glm(formula = is_canceled ~ lead_time + arrival_date_month +
    stays_in_weekend_nights + stays_in_week_nights + adults +
    children + babies + meal + market_segment + distribution_channel +
    is_repeated_guest + previous_cancellations + previous_bookings_not_canceled +
    reserved_room_type + assigned_room_type + deposit_type +
    customer_type + adr + required_car_parking_spaces + total_of_special_requests,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.7435  -0.3564   0.2051   6.0062

Coefficients: (1 not defined because of singularities)
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.589e+00  2.068e-01 -12.518  < 2e-16 ***
lead_time                     3.711e-03  1.095e-04  33.885  < 2e-16 ***
arrival_date_monthAugust     -1.544e-01  3.845e-02  -4.015 5.94e-05 ***
arrival_date_monthDecember    1.308e-01  4.698e-02   2.785 0.005355 **
arrival_date_monthFebruary    1.231e-01  4.461e-02   2.760 0.005775 **
arrival_date_monthJanuary     2.985e-04  5.089e-02   0.006 0.995320
arrival_date_monthJuly       -2.311e-01  3.832e-02  -6.031 1.63e-09 ***
arrival_date_monthJune       -1.596e-01  4.028e-02  -3.963 7.40e-05 ***
arrival_date_monthMarch      -1.618e-01  4.197e-02  -3.855 0.000116 ***
arrival_date_monthMay        -1.099e-01  3.894e-02  -2.821 0.004788 **
arrival_date_monthNovember    1.595e-02  4.792e-02   0.333 0.739270
arrival_date_monthOctober    -3.411e-02  4.143e-02  -0.823 0.410308
arrival_date_monthSeptember  -2.046e-01  4.279e-02  -4.781 1.74e-06 ***
stays_in_weekend_nights       3.659e-02  9.965e-03   3.672 0.000241 ***
stays_in_week_nights          4.256e-02  5.231e-03   8.136 4.09e-16 ***
adults                        1.655e-01  1.926e-02   8.595  < 2e-16 ***
children1                     2.426e-01  4.366e-02   5.557 2.75e-08 ***
children10                    1.597e+01  5.354e+02   0.030 0.976199
children2                     4.481e-01  6.510e-02   6.882 5.89e-12 ***
children3                    -5.030e-01  3.671e-01  -1.370 0.170606
```

```
childrenNA                     -5.969e+02  8.570e+05  -0.001 0.999444
babies                          1.680e-01  9.883e-02   1.700 0.089214 .
mealFB                          6.443e-01  1.198e-01   5.377 7.59e-08 ***
mealHB                         -1.379e-01  2.955e-02  -4.668 3.05e-06 ***
mealSC                          4.847e-02  2.868e-02   1.690 0.091053 .
mealUndefined                  -5.897e-01  1.098e-01  -5.369 7.92e-08 ***
market_segmentComplementary     9.661e-01  2.575e-01   3.752 0.000175 ***
market_segmentCorporate         9.786e-02  2.010e-01   0.487 0.626333
market_segmentDirect            2.532e-01  2.222e-01   1.140 0.254388
market_segmentGroups            2.988e-01  2.093e-01   1.428 0.153341
market_segmentOffline TA/TO    -2.969e-01  2.101e-01  -1.413 0.157530
market_segmentOnline TA         9.386e-01  2.092e-01   4.486 7.25e-06 ***
market_segmentUndefined         1.077e+00  5.227e+02   0.002 0.998356
distribution_channelDirect     -5.835e-01  1.060e-01  -5.506 3.67e-08 ***
distribution_channelGDS        -1.102e+00  2.178e-01  -5.058 4.24e-07 ***
distribution_channelTA/TO      -8.511e-02  7.805e-02  -1.091 0.275492
distribution_channelUndefined   6.127e+02  8.570e+05   0.001 0.999430
is_repeated_guest              -5.965e-01  9.582e-02  -6.225 4.81e-10 ***
previous_cancellations          2.824e+00  6.821e-02  41.400  < 2e-16 ***
previous_bookings_not_canceled -5.098e-01  2.803e-02 -18.188  < 2e-16 ***
reserved_room_typeB             5.802e-01  1.157e-01   5.014 5.33e-07 ***
reserved_room_typeC             1.264e+00  1.471e-01   8.593  < 2e-16 ***
reserved_room_typeD             1.138e+00  5.113e-02  22.255  < 2e-16 ***
reserved_room_typeE             2.010e+00  1.018e-01  19.739  < 2e-16 ***
reserved_room_typeF             2.050e+00  1.520e-01  13.488  < 2e-16 ***
reserved_room_typeG             2.974e+00  2.221e-01  13.394  < 2e-16 ***
reserved_room_typeH             2.073e+00  4.962e-01   4.177 2.95e-05 ***
reserved_room_typeL             1.617e+00  1.216e+00   1.329 0.183763
reserved_room_typeP             1.474e+01  2.127e+02   0.069 0.944749
assigned_room_typeB            -8.103e-01  9.216e-02  -8.791  < 2e-16 ***
assigned_room_typeC            -1.408e+00  1.112e-01 -12.658  < 2e-16 ***
assigned_room_typeD            -1.327e+00  4.862e-02 -27.299  < 2e-16 ***
assigned_room_typeE            -2.010e+00  9.803e-02 -20.501  < 2e-16 ***
assigned_room_typeF            -2.628e+00  1.419e-01 -18.515  < 2e-16 ***
assigned_room_typeG            -3.345e+00  2.154e-01 -15.528  < 2e-16 ***
assigned_room_typeH            -2.235e+00  4.840e-01  -4.617 3.90e-06 ***
assigned_room_typeI            -4.498e+00  5.518e-01  -8.153 3.55e-16 ***
assigned_room_typeK            -2.835e+00  3.652e-01  -7.764 8.21e-15 ***
assigned_room_typeL             1.209e+01  5.354e+02   0.023 0.981991
assigned_room_typeP                    NA         NA      NA       NA
deposit_typeNon Refund          5.468e+00  1.292e-01  42.309  < 2e-16 ***
deposit_typeRefundable          2.902e-01  2.286e-01   1.269 0.204407
customer_typeGroup             -1.167e-01  1.887e-01  -0.619 0.536184
```

```
customer_typeTransient          7.995e-01  6.015e-02  13.293  < 2e-16 ***
customer_typeTransient-Party    2.668e-01  6.384e-02   4.180 2.92e-05 ***
adr                             4.585e-03  2.691e-04  17.042  < 2e-16 ***
required_car_parking_spaces    -6.246e+02  8.570e+05  -0.001 0.999419
total_of_special_requests      -7.333e-01  1.297e-02 -56.541  < 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: 125929  on 95511  degrees of freedom
Residual deviance:  80768  on 95445  degrees of freedom
AIC: 80902

Number of Fisher Scoring iterations: 12
```

4 task

```
hotel_data <- read.csv("hotel_bookings.csv")
```

arguments and target variable

```
X <- hotel_data[,1:27]
y <- hotel_data$is_canceled
```

prediction of the model:

```
predicted <- predict(model, newdata = na.omit(test), type = "response")
```

```
Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
prediction from a rank-deficient fit may be misleading
```

```
predicted_class <- ifelse(predicted > 0.5, 1, 0)
```

performance of the model

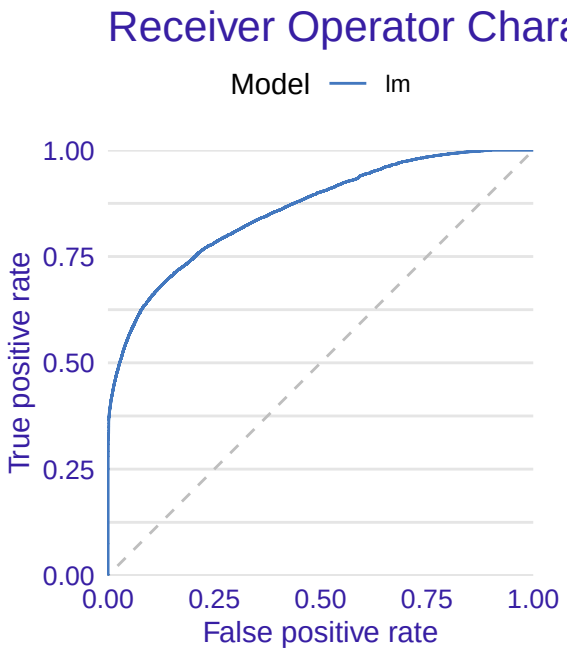
```
library(DALEX)
```

```
Welcome to DALEX (version: 2.4.2).
Find examples and detailed introduction at: http://ema.drwhy.ai/
Additional features will be available after installation of: ggpubr.
Use 'install_dependencies()' to get all suggested dependencies
```

```r
explain_lr <- explain(model = model,
                      data = hotel_bookings,
                      y = hotel_bookings$is_canceled == 1,
                      type = "classification",
                      verbose = FALSE)
```

```
Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
prediction from a rank-deficient fit may be misleading

Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
prediction from a rank-deficient fit may be misleading

Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
prediction from a rank-deficient fit may be misleading
```

```r
performace_lr <- model_performance(explain_lr)
plot(performace_lr, geom = "roc")
```



Task 5 calculate the class distribution

```
table(hotel_bookings$is_canceled)
```

```
    0     1
75166 44224
```

controls the class distribution of the dataset and to balance it using over/under sampling methods:

```
table(hotel_bookings$is_canceled)
```

```
    0     1
75166 44224
```

```
# Balancing with oversampling
#install.packages("ROSE")

library(ROSE)
```

```
Loaded ROSE 0.0-4
```

```
balanced_data_over <- ovun.sample(is_canceled ~ ., data = hotel_data, method = "over")$dat
table(balanced_data_over$is_canceled)
```

```
    0     1
75166 75112
```

```
# Balancing with undersampling
balanced_data_under <- ovun.sample(is_canceled ~ ., data = hotel_data, method = "under")$d
table(balanced_data_under$is_canceled)
```

```
    0      1
44137 44220
```

After the balancing process was completed, there were two new datasets (balanced_data_over and balanced_data_under) that were balanced with oversampling and undersampling methods. performance of the model

A model for the 1st oversampling dataset