Data Science

Machine Learning

Predicting Hotel Booking Cancellation

Akash Kandarkar

TABLE OF CONTENTS

INTRODUCTION	2
BUSINESS UNDERSTANDING	2
BUSINESS QUESTIONS	3
TECHNICAL DETAILS	4
GOAL OF ANALYSIS	6
DATA ACQUISITION	6
PACKAGES USED	7
DATA CLEANING	7
VARIABLE ANALYSIS	11
MODEL PREDICTION	18
CONCLUSION	47
RECOMMENDATION	48

Introduction

The Hotel Industry has changed over the years, with most of the booking by third-party companies. These Online Travel Agencies have changed the cancellation policy from footnote at the bottom of the page to making it the main selling point in their marketing campaign. As a result, customers have become accustomed to free cancellation policies. Back in 2019, D-Edge Hospitality Solutions reported that the global cancelation rate of hotel reservation reached 40%.

The dataset contains hotel booking data where it is necessary to evaluate numerous factors that can lead to bookings being cancelled. Few of these factors that can determine cancellations include the lag period between booking made and date booked, type of customer, is he/she a regular customer, location etc. It is necessary to predict if the bookings will still prevail by analyzing the past data which shows which bookings are being cancelled and recording a pattern which can help the hotel industry make a better prediction.

Our Project aims to provide valuable insights on hotel cancellations using analytics. To successfully address the needs of the project, we have adopted Kanban Project Methodology. We found this useful as it enables agility and prevents overloading the development process.

Business Understanding

Booking Cancellation is a challenge faced by the hospitality industry because it has a direct impact on their revenue generation. When customers cancel the booking, there are severe implications for the hotels, it affects their occupancy rates.

Revenue management strategies, such as dynamic pricing, overbooking, and strict cancellation policies, are employed to address booking cancellation and maximize occupancy rates. However, when done based on intuition only, these strategies might backfire, such as loss of sales, deteriorated business reputation and fall in customer loyalty.

- ## Business Questions

1.Do people who cancel their booking tend to make booking changes?

2.Can we predict a pattern based on previous cancellations?

3.What type of customers usually cancel the booking?

4.Which type of deposit accounted for more cancellations?

5.Determine a threshold number of days after which if a customer cancels their booking, they need to pay a convenience fee?

6.Which country has the maximum cancellation and what needs to be done there?

7.Which market segment has the maximum cancellation and which market segment needs to be focused on?

- ## Technical Details

There are 40060 rows and 20 columns in our dataset

Looking at the columns we can see that 'IsCanceled' is the column that tells us if a booking is cancelled or not and this is the column we will be trying to predict.

1. **LeadTime** (Integer) is an Integer which tells us the Number of days that elapsed between the entering date of the booking into and the arrival date.
2. **StaysInWeekendNights** (Integer) is an Integer which is Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
3. **StaysInWeekNights** (Integer) is a variable which is Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel.
4. **Adults (**Integer) which gives Number of adults.
5. **Children** (Integer) which gives Number of children.
6. **Babies** (Integer) which gives Number of babies.
7. **Meal** (Categorical) have of meal booked. Categories are presented in standard hospitality meal packages:
    a. Undefined/SC – no meal package.
    b. BB – Bed & Breakfast.

c. HB – Half board (breakfast and one other meal – usually dinner);
d. FB – Full board (breakfast, lunch and dinner)

8. **Country** (Categorical) represents data about country of origin. Categories are represented in the ISO 3155– 3:2013 format
9. **MarketSegment (**Categorical) Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
10. **IsRepeatedGuest** (Categorical), Value indicating if the booking name was from a repeated guest (1) or not (0)
11. **PreviousCancellations** (Integer), Number of previous bookings that were cancelled by the customer prior to the current booking
12. **PreviousBookingsNotCanceled** (Integer) Number of previous bookings not cancelled by the customer prior to the current booking
13. **ReservedRoomType** (Categorical), Code of room type reserved. Code is presented instead of designation for anonymity reasons
14. **AssignedRoomType** (Categorical) Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g., overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
15. **BookingChanges** (Integer) Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
16. **DepositType**: Categorical, Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made. Non-Refund – a deposit was made in the value of the total stay cost. Refundable – a deposit was made with a value under the total cost of stay.
17. **CustomerType** (Categorical) consists of type of booking, assuming one of four categories:
    a. Contract - when the booking has an allotment or other type of contract associated to it;
    b. Group – when the booking is associated to a group;
    c. Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;
    d. Transient-party – when the booking is transient, but is associated to at least other transient booking
18. **RequiredCardParkingSpaces** (Integer) Number of car parking spaces required by the customer

19. **TotalOfSpecialRequests** (Integer) Number of special requests made by the customer (e.g. twin bed or high floor)

- # Goal of our Analysis

Using different Models like Support Vector Machine, Regression, Association Rule, we aim to provide the best model to provide insights into what affects booking cancellation and provide a recommendation on how to improve cancellation rates.

- # Data Acquisition

> glimpse(file)

```
Rows: 38,886
Columns: 22
$ IsCanceled                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, …
$ LeadTime                   <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, 23, 35, 68, 18, 37, 68, 37, 12, 0, 7, 37, 72, 72, 7…
$ StaysInWeekendNights       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2, 2, 2, 2, 2, 2, 2, …
$ StaysInWeekNights          <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 1, 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, …
$ Adults                     <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, …
$ Children                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ Babies                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ Meal                       <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB", "FB", "BB", "HB", "BB", "HB", "BB", "HB", "BB", …
$ Country                    <chr> "Portugal", "Portugal", "United Kingdom", "United Kingdom", "United Kingdom", "United King…
$ MarketSegment              <chr> "Direct", "Direct", "Direct", "Corporate", "Online TA", "Online TA", "Direct", "Direct", "…
$ IsRepeatedGuest            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ PreviousCancellations      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ PreviousBookingsNotCanceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ ReservedRoomType           <chr> "C", "C", "A", "A", "A", "A", "C", "C", "A", "D", "E", "D", "D", "G", "E", "D", "E", "A", …
$ AssignedRoomType           <chr> "C", "C", "C", "A", "A", "A", "C", "C", "A", "D", "E", "D", "E", "G", "E", "E", "E", "E", …
$ BookingChanges             <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, …
$ DepositType                <chr> "No Deposit", "No Deposit", "No Deposit", "No Deposit", "No Deposit", "No Deposit", "No De…
$ CustomerType               <chr> "Transient", "Transient", "Transient", "Transient", "Transient", "Transient", "Transient",…
$ RequiredCarParkingSpaces   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, …
$ TotalOfSpecialRequests     <dbl> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, 1, 0, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 2, 0, 1, …
$ family                     <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
$ IsCan_Factor               <fct> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, …
```

- # PACKAGES USED:

The following packages were used:

- Tidyverse – collection of R packages
- caret- build machine learning models
- ggplot2- primarily used for data visualization
- party- create decision trees
- ggpubr- produce production-quality visualizations
- kernlab- Used for kernel-based machine Learning
- arules- represent, manipulate, and analyze transaction data and patterns.
- arulesViz- handling and mining association rules.
- Maps- used to make map outlines and points
- ggmap- functions to visualize spatial data and models
- mapproj- convert latitude/longitude to coordinates
- rworldmap- maps global data

- # Data Cleaning

```
Cancellations=read_csv("https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv");

>>view(Cancellations)
```

	IsCanceled	LeadTime	StaysInWeekendNights	StaysInWeekNights	Adults	Children	Babies	Meal	Country	MarketSegment
1	0	342	0	0	2	0	0	BB	PRT	Direct
2	0	737	0	0	2	0	0	BB	PRT	Direct
3	0	7	0	1	1	0	0	BB	GBR	Direct
4	0	13	0	1	1	0	0	BB	GBR	Corporate
5	0	14	0	2	2	0	0	BB	GBR	Online TA
6	0	14	0	2	2	0	0	BB	GBR	Online TA
7	0	0	0	2	2	0	0	BB	PRT	Direct
8	0	9	0	2	2	0	0	FB	PRT	Direct
9	1	85	0	3	2	0	0	BB	PRT	Online TA
10	1	75	0	3	2	0	0	HB	PRT	Offline TA/TO
11	1	23	0	4	2	0	0	BB	PRT	Online TA
12	0	35	0	4	2	0	0	HB	PRT	Online TA
13	0	68	0	4	2	0	0	BB	USA	Online TA
14	0	18	0	4	2	1	0	HB	ESP	Online TA
15	0	37	0	4	2	0	0	BB	PRT	Online TA
16	0	68	0	4	2	0	0	BB	IRL	Online TA
17	0	37	0	4	2	0	0	BB	PRT	Offline TA/TO
18	0	12	0	1	2	0	0	BB	IRL	Online TA
19	0	0	0	1	2	0	0	BB	FRA	Corporate
20	0	7	0	4	2	0	0	BB	GBR	Direct

IsRepeatedGuest	PreviousCancellations	PreviousBookingsNotCanceled	ReservedRoomType	AssignedRoomType	BookingChanges	DepositType	CustomerType	RequiredCarParkingSpaces	TotalOfSpecialRequests
0	0	0	C	C	3	No Deposit	Transient	0	0
0	0	0	C	C	4	No Deposit	Transient	0	0
0	0	0	A	C	0	No Deposit	Transient	0	0
0	0	0	A	A	0	No Deposit	Transient	0	0
0	0	0	A	A	0	No Deposit	Transient	0	1
0	0	0	A	A	0	No Deposit	Transient	0	1
0	0	0	C	C	0	No Deposit	Transient	0	0
0	0	0	C	C	0	No Deposit	Transient	0	1
0	0	0	A	A	0	No Deposit	Transient	0	1
0	0	0	D	D	0	No Deposit	Transient	0	0
0	0	0	E	E	0	No Deposit	Transient	0	0
0	0	0	D	D	0	No Deposit	Transient	0	0
0	0	0	D	E	0	No Deposit	Transient	0	3
0	0	0	G	G	1	No Deposit	Transient	0	1
0	0	0	E	E	0	No Deposit	Transient	0	0
0	0	0	D	E	0	No Deposit	Transient	0	3
0	0	0	E	E	0	No Deposit	Contract	0	0
0	0	0	A	E	0	No Deposit	Transient	0	0
0	0	0	A	G	0	No Deposit	Transient	0	0
0	0	0	G	G	0	No Deposit	Transient	0	1

>>dim(Cancellation)

```
> dim(Cancellations)
[1] 40060    20
```

>>str(Cancellations)

```
spec_tbl_df [40,060 × 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ IsCanceled                 : num [1:40060] 0 0 0 0 0 0 0 0 1 1 ...
 $ LeadTime                   : num [1:40060] 342 737 7 13 14 14 0 9 85 75 ...
 $ StaysInWeekendNights       : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ StaysInWeekNights          : num [1:40060] 0 0 1 1 2 2 2 2 3 3 ...
 $ Adults                     : num [1:40060] 2 2 1 1 2 2 2 2 2 2 ...
 $ Children                   : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ Babies                     : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ Meal                       : chr [1:40060] "BB" "BB" "BB" "BB" ...
 $ Country                    : chr [1:40060] "PRT" "PRT" "GBR" "GBR" ...
 $ MarketSegment              : chr [1:40060] "Direct" "Direct" "Direct" "Corporate" ...
 $ IsRepeatedGuest            : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ PreviousCancellations      : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ PreviousBookingsNotCanceled: num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ ReservedRoomType           : chr [1:40060] "C" "C" "A" "A" ...
 $ AssignedRoomType           : chr [1:40060] "C" "C" "C" "A" ...
 $ BookingChanges             : num [1:40060] 3 4 0 0 0 0 0 0 0 0 ...
 $ DepositType                : chr [1:40060] "No Deposit" "No Deposit" "No Deposit" "No Deposit" ...
 $ CustomerType               : chr [1:40060] "Transient" "Transient" "Transient" "Transient" ...
 $ RequiredCarParkingSpaces   : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ TotalOfSpecialRequests     : num [1:40060] 0 0 0 0 1 1 0 1 1 0 ...
 - attr(*, "spec")=
```

```
 - attr(*, "spec")=
  .. cols(
  ..   IsCanceled = col_double(),
  ..   LeadTime = col_double(),
  ..   StaysInWeekendNights = col_double(),
  ..   StaysInWeekNights = col_double(),
  ..   Adults = col_double(),
  ..   Children = col_double(),
  ..   Babies = col_double(),
  ..   Meal = col_character(),
  ..   Country = col_character(),
  ..   MarketSegment = col_character(),
  ..   IsRepeatedGuest = col_double(),
  ..   PreviousCancellations = col_double(),
  ..   PreviousBookingsNotCanceled = col_double(),
  ..   ReservedRoomType = col_character(),
  ..   AssignedRoomType = col_character(),
  ..   BookingChanges = col_double(),
  ..   DepositType = col_character(),
  ..   CustomerType = col_character(),
  ..   RequiredCarParkingSpaces = col_double(),
  ..   TotalOfSpecialRequests = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

>>summary(Cancellation)

```
  IsCanceled        LeadTime      StaysInWeekendNights StaysInWeekNights     Adults          Children          Babies
Min.   :0.0000   Min.   :  0.00   Min.   : 0.00        Min.   : 0.000    Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.: 10.00   1st Qu.: 0.00        1st Qu.: 1.000    1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.:0.0000
Median :0.0000   Median : 57.00   Median : 1.00        Median : 3.000    Median : 2.000   Median : 0.0000   Median :0.0000
Mean   :0.2776   Mean   : 92.68   Mean   : 1.19        Mean   : 3.129    Mean   : 1.867   Mean   : 0.1287   Mean   :0.0139
3rd Qu.:1.0000   3rd Qu.:155.00   3rd Qu.: 2.00        3rd Qu.: 5.000    3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :737.00   Max.   :19.00        Max.   :50.000    Max.   :55.000   Max.   :10.0000   Max.   :2.0000
    Meal             Country          MarketSegment      IsRepeatedGuest   PreviousCancellations PreviousBookingsNotCanceled
Length:40060       Length:40060       Length:40060       Min.   :0.00000   Min.   : 0.0000       Min.   : 0.0000
Class :character   Class :character   Class :character   1st Qu.:0.00000   1st Qu.: 0.0000       1st Qu.: 0.0000
Mode  :character   Mode  :character   Mode  :character   Median :0.00000   Median : 0.0000       Median : 0.0000
                                                         Mean   :0.04438   Mean   : 0.1017       Mean   : 0.1465
                                                         3rd Qu.:0.00000   3rd Qu.: 0.0000       3rd Qu.: 0.0000
                                                         Max.   :1.00000   Max.   :26.0000       Max.   :30.0000
ReservedRoomType   AssignedRoomType   BookingChanges   DepositType        CustomerType       RequiredCarParkingSpaces
Length:40060       Length:40060       Min.   : 0.000   Length:40060       Length:40060       Min.   :0.0000
Class :character   Class :character   1st Qu.: 0.000   Class :character   Class :character   1st Qu.:0.0000
Mode  :character   Mode  :character   Median : 0.000   Mode  :character   Mode  :character   Median :0.0000
                                      Mean   : 0.288                                        Mean   :0.1381
                                      3rd Qu.: 0.000                                        3rd Qu.:0.0000
                                      Max.   :17.000                                        Max.   :8.0000
TotalOfSpecialRequests
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.6198
3rd Qu.:1.0000
Max.   :5.0000
```

>>sapply(Cancellations,function(x) sum(is.null(x)))

```
> sapply(Cancellations,function(x) sum(is.null(x)))
                IsCanceled                   LeadTime        StaysInWeekendNights          StaysInWeekNights
                         0                          0                           0                          0
                    Adults                   Children                      Babies                       Meal
                         0                          0                           0                          0
                   Country                MarketSegment            IsRepeatedGuest       PreviousCancellations
                         0                          0                           0                          0
PreviousBookingsNotCanceled          ReservedRoomType            AssignedRoomType             BookingChanges
                         0                          0                           0                          0
                DepositType                CustomerType    RequiredCarParkingSpaces      TotalOfSpecialRequests
                         0                          0                           0                          0
```

>>sapply(Cancellations,function(x) sum(is.na(x)))

```
> sapply(Cancellations,function(x) sum(is.na(x)))
                IsCanceled                   LeadTime        StaysInWeekendNights          StaysInWeekNights
                         0                          0                           0                          0
                    Adults                   Children                      Babies                       Meal
                         0                          0                           0                          0
                   Country                MarketSegment            IsRepeatedGuest       PreviousCancellations
                         0                          0                           0                          0
PreviousBookingsNotCanceled          ReservedRoomType            AssignedRoomType             BookingChanges
                         0                          0                           0                          0
                DepositType                CustomerType    RequiredCarParkingSpaces      TotalOfSpecialRequests
                         0                          0                           0                          0
```

>>table(Cancellations$Country)

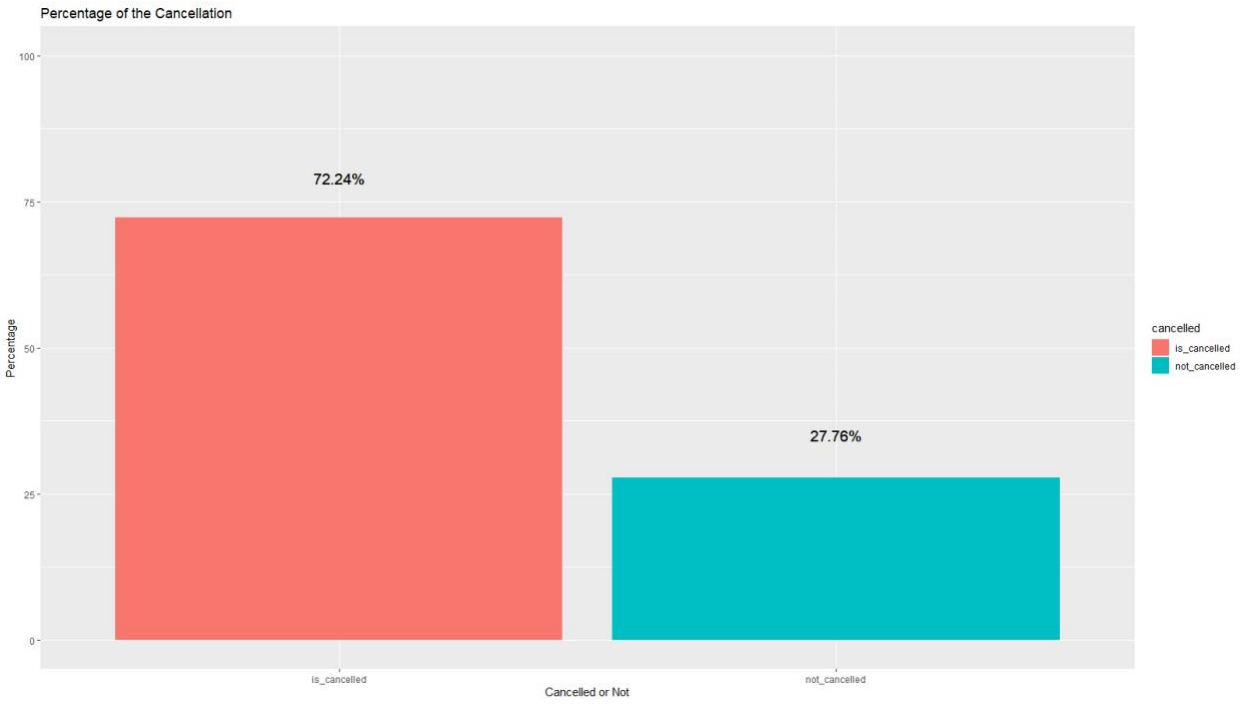
```
> table(Cancellations$Country)

  AGO   ALB   AND   ARE   ARG   ARM   AUS   AUT   AZE   BDI   BEL   BGR   BHR   BHS   BIH   BLR   BRA   BWA   CAF   CHE
   24     3     5    11    57     2    87   210     3     1   448     5     1     1     1     7   430     1     3   435
  CHL   CHN   CIV   CMR    CN   COL   COM   CPV   CRI   CUB   CYM   CYP   CZE   DEU   DJI   DNK   DOM   DZA   ECU   EGY
   17   134     2     2   710    16     1     5     2     4     1     8    27  1203     1    65     3    12     2     1
  ESP   EST   FIN   FJI   FRA   GBR   GEO   GGY   GIB   GRC   HKG   HRV   HUN   IDN   IND   IRL   IRN   ISL   ISR   ITA
 3957    33   151     1  1611  6814    11     1    13    10     4    11    47     5    37  2166     5     6    28   459
  JAM   JEY   JOR   JPN   KAZ   KOR   KWT   LBN   LKA   LTU   LUX   LVA   MAC   MAR   MDG   MDV   MEX   MKD   MLT   MOZ
    5     3     2     9     5     9     3     6     1    46    80    33     1    75     1     6     6     1     2     6
  MUS   MWI   MYS   NGA   NLD   NOR   NPL  NULL   NZL   OMN   PAK   PER   PHL   PLW   POL   PRI   PRT   QAT   ROU   RUS
    1     2    10    10   514   123     1   464    14    11     4     1    16     1   333     9 17630     1   177   189
  SAU   SEN   SGP   SMR   SRB   SUR   SVK   SVN   SWE   SYC   SYR   TGO   THA   TUN   TUR   TWN   UGA   UKR   URY   USA
    1     1     4     1     7     4    12    11   304     1     1     1     6     1    23    12     1    23     8   479
  UZB   VEN   VNM   ZAF   ZMB   ZWE
    1     3     2    18     1     2
```

We checked the data for missing or null value and found that there are few dummy values in countries and meals which the model does not consider, and it does not affect accuracy.

- **Variable analysis**

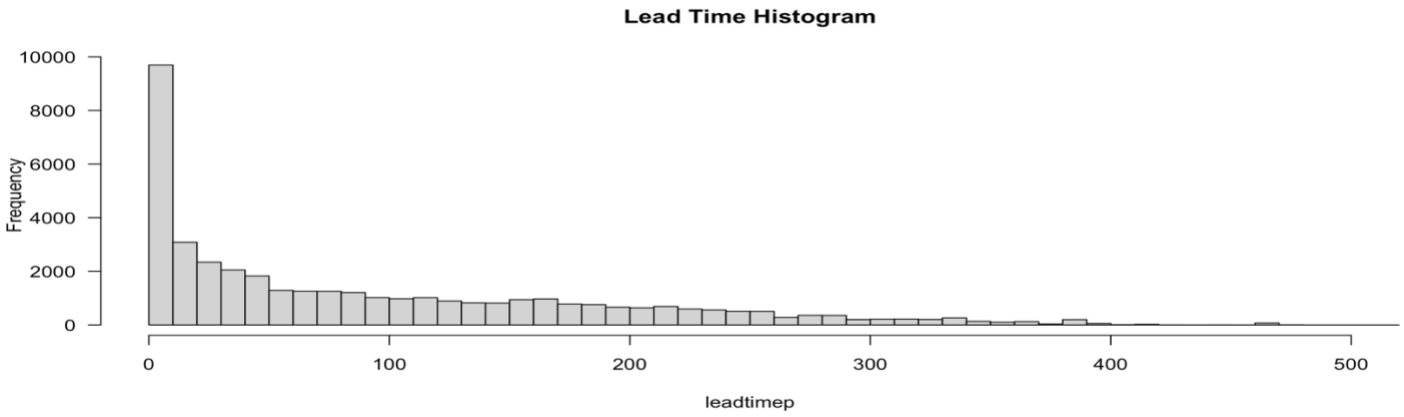
a) Number of cancellation vs Number of bookings

Row Labels	Count of IsCanceled	Percentage
0	28938	72.24%
1	11122	27.76%
Grand Total	**40060**	



From the table and bar plot, we see that there are 11,122 cancellations and 28,938 bookings. IsCanceled is the response variable for our model. We see that there are 27.76% of cancellation and 72.24% of bookings in our dataset.

b) Lead Time before checking in

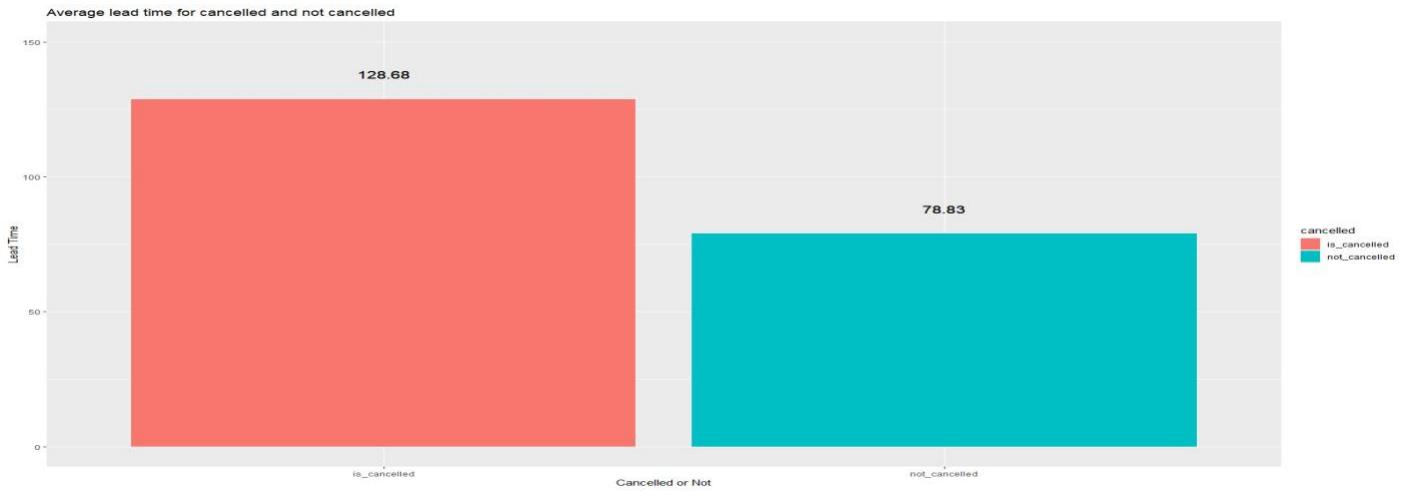


Insight about how lead Time is distributed

```
> summary(Cancellations$LeadTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   10.00   57.00   92.68  155.00  737.00
```

From the graph of Leadtime, the number of days between booking the hotel and checking in, we can see that the distribution is skewed right. The median is 57 days (about 2 months), meaning people book their stays almost 2 months in advance. The interquartile range is 145 days (about 5 months), so there is variability in terms of how far in advance a client reserves a hotel. It is surprising to see that a vast number of bookings were made on the same day or within 10 days of check-in. The maximum for the distribution is 737 days (about 2 years).

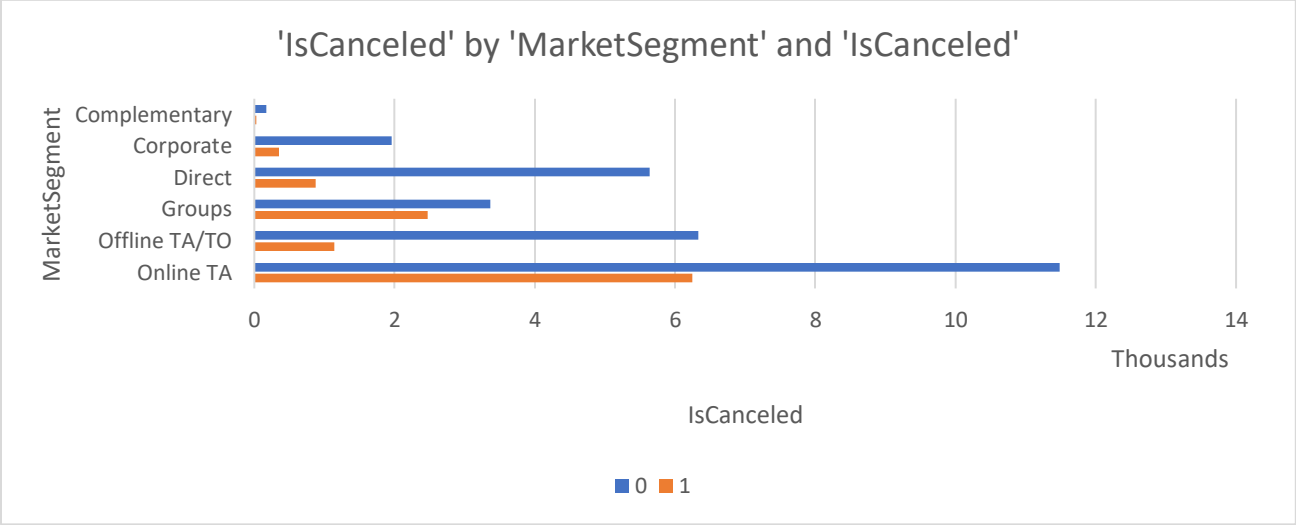
Average LeadTime for cancellation:



From the above bar plot, we can see that when the clients are cancelling their bookings, their average Lead Time before the check in is 129 days (about 4 months) and the average Lead Time when the clients are not cancelling their booking is 79 days (about 2 and a half months).

c) Market Segment

Count of IsCanceled	Column Labels			
Row Labels	**0**	**1**	**Grand Total**	**Percentage of Cancellation**
Complementary	168	33	201	19.64%
Corporate	1958	351	2309	17.93%
Direct	5635	878	6513	15.58%
Groups	3362	2474	5836	73.59%
Offline TA/TO	6334	1138	7472	17.97%
Online TA	11481	6248	17729	54.42%
Grand Total	**28938**	**11122**	**40060**	



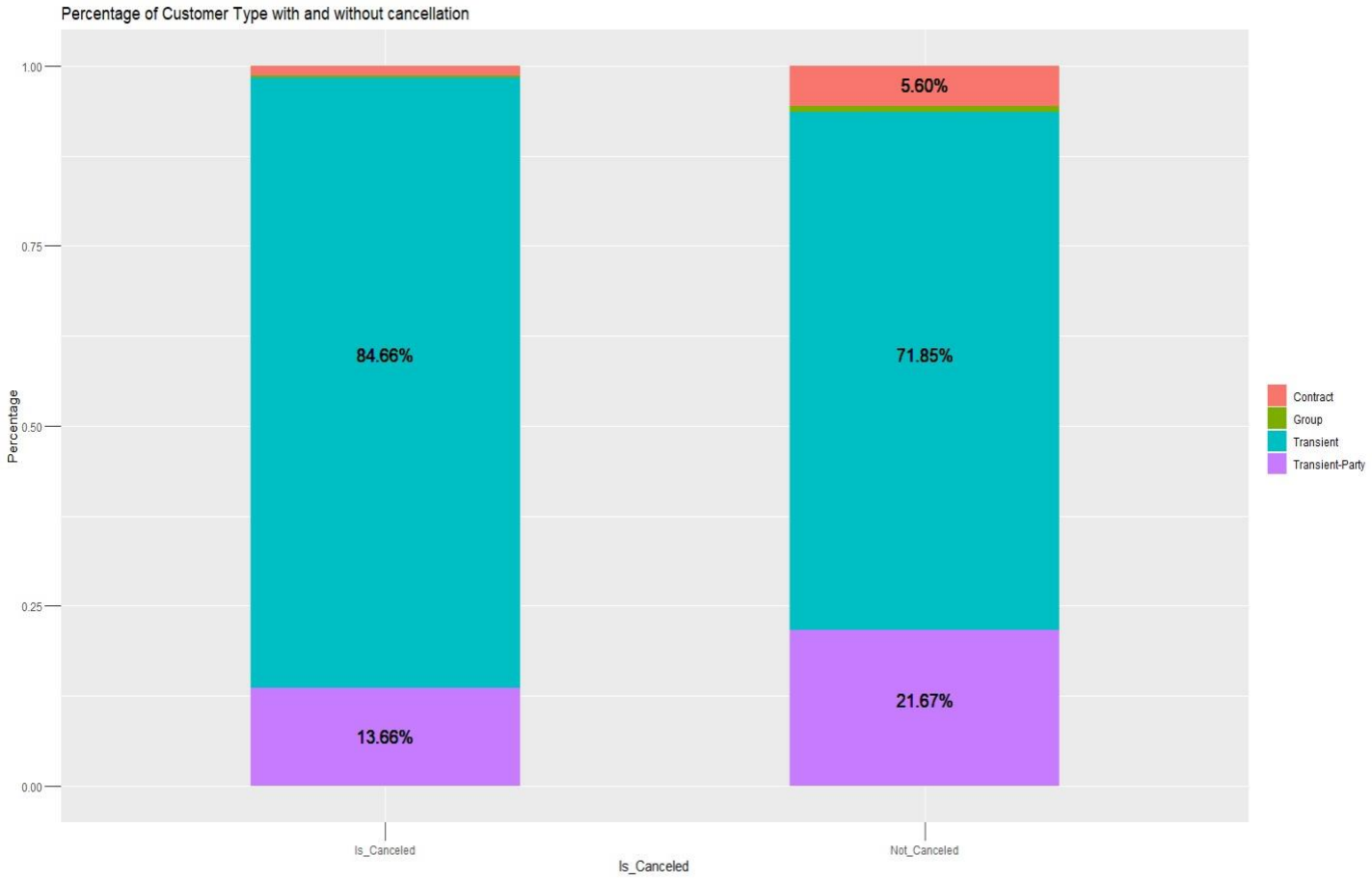
From the table and bar chart of market segment, we can see that most bookings in the dataset are made by Online Travel Agent, followed by Offline Travel Agent/Tour Operators, then clients who have booked directly and then by groups.

Similarly, we can see that most cancellation in the dataset are done by group bookings where 73.59% of the bookings are canceled, followed by Online Travel Agents where 54.42% of the bookings are cancelled.

d) Customer Type

Count of IsCanceled

Row Labels	Column Labels			
	0	1	Grand Total	Percentage of Cancellation
Contract	1619	157	1776	9.70%
Group	254	30	284	11.81%
Transient	20793	9416	30209	45.28%
Transient-Party	6272	1519	7791	24.22%
Grand Total	**28938**	**11122**	**40060**	

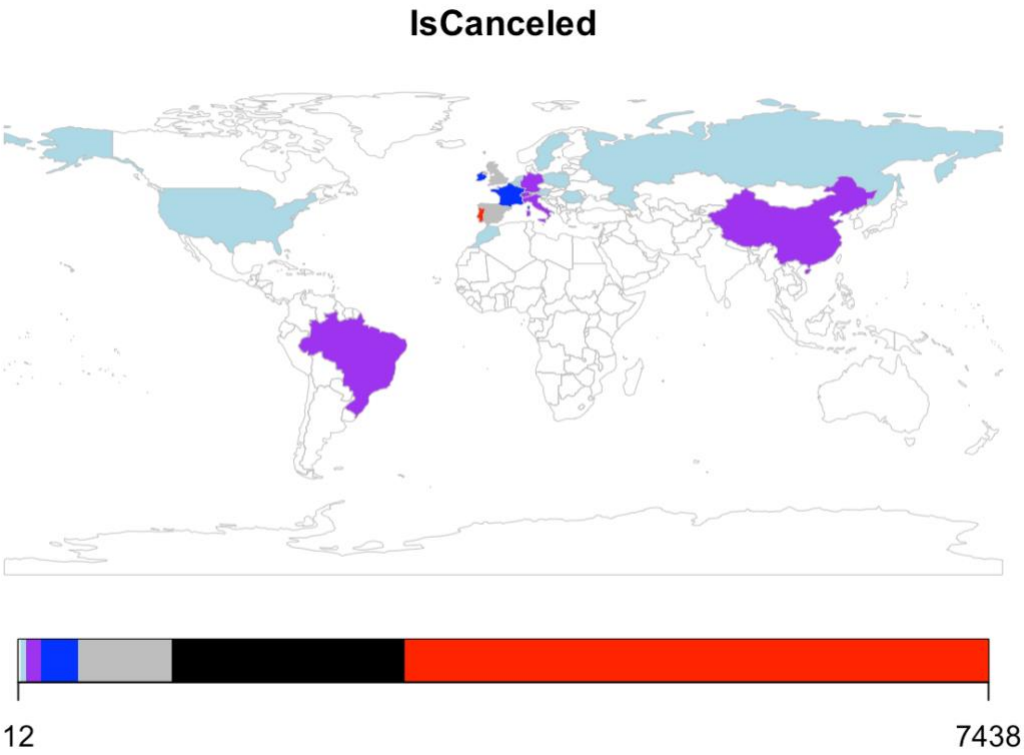


From the above table and bar plot, we can see that the maximum bookings have been done by transient customer type, where transient booking type is 94.85% of total booking. Transient Customer type has 45.28% of cancellation on all the bookings they made, which is almost half of the bookings.

e) Countries

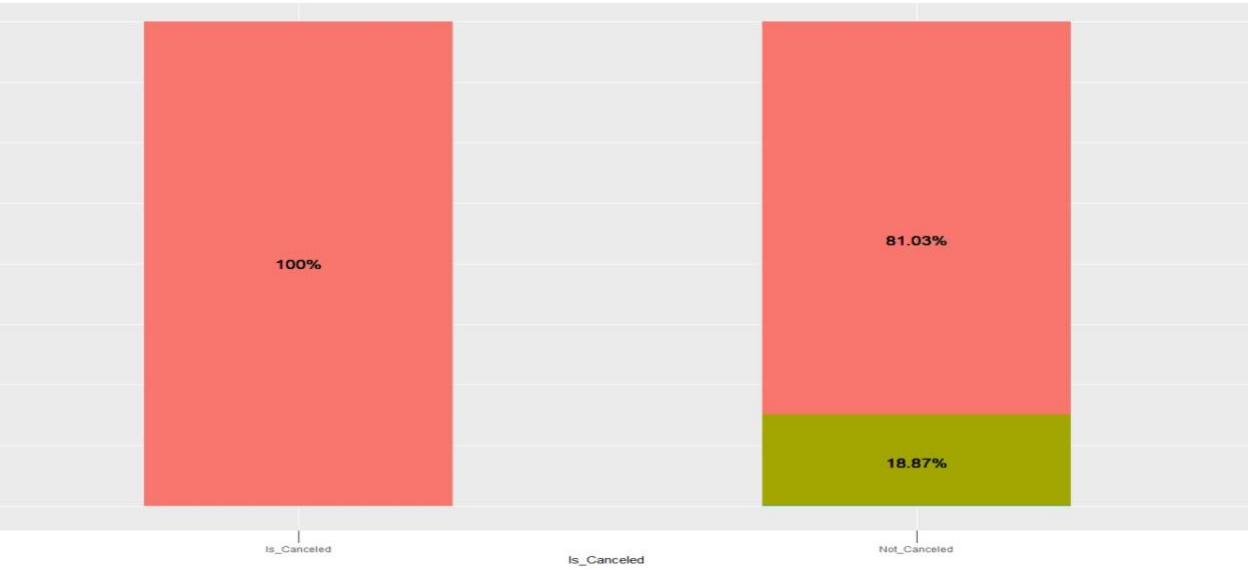
Which country and what factors are in that country?

```
> table(Cancellations$Country) %>% as.data.frame() %>% arrange(desc(Freq))
    Var1  Freq
1    PRT 17630
2    GBR  6814
3    ESP  3957
4    IRL  2166
5    FRA  1611
6    DEU  1203
7     CN   710
8    NLD   514
9    USA   479
10  NULL   464
11   ITA   459
12   BEL   448
13   CHE   435
14   BRA   430
15   POL   333
16   SWE   304
17   AUT   210
18   RUS   189
19   ROU   177
20   FIN   151
21   CHN   134
22   NOR   123
23   AUS    87
24   LUX    80
25   MAR    75
26   DNK    65
27   ARG    57
28   HUN    47
29   LTU    46
30   IND    37
31   EST    33
```



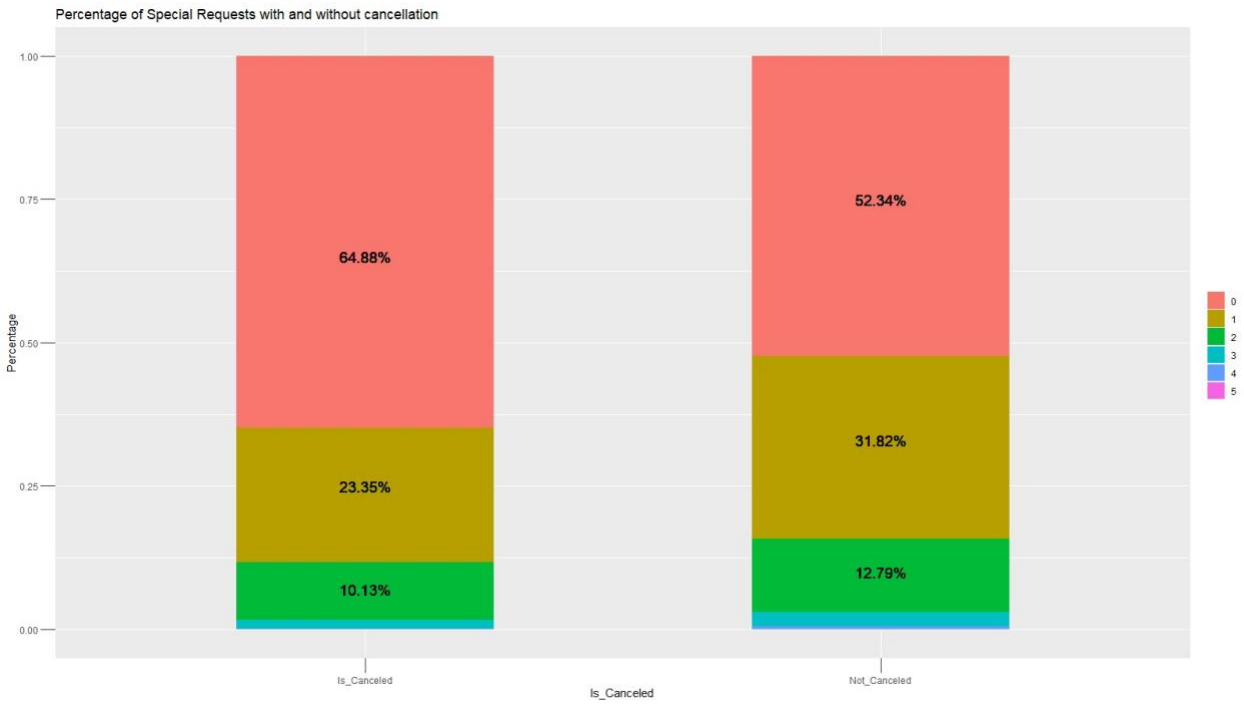
As we can see the from the stats with top countries contributing to cancellation: PRT-17630, GBR - 6814, ESP-3957, IRL-2166, FRA-1611, DEU-1203, CN-710, NLD-514, USA 479 where Portugal (PRT) is country having most of the cancellations as shown in map.

f) Required Parking Space



From here, we can see that if the customer is likely to reserve a parking space in a hotel, the chances of that customer cancelling decrease. There is 18.87% chance that the customer will not cancel his cancellation.

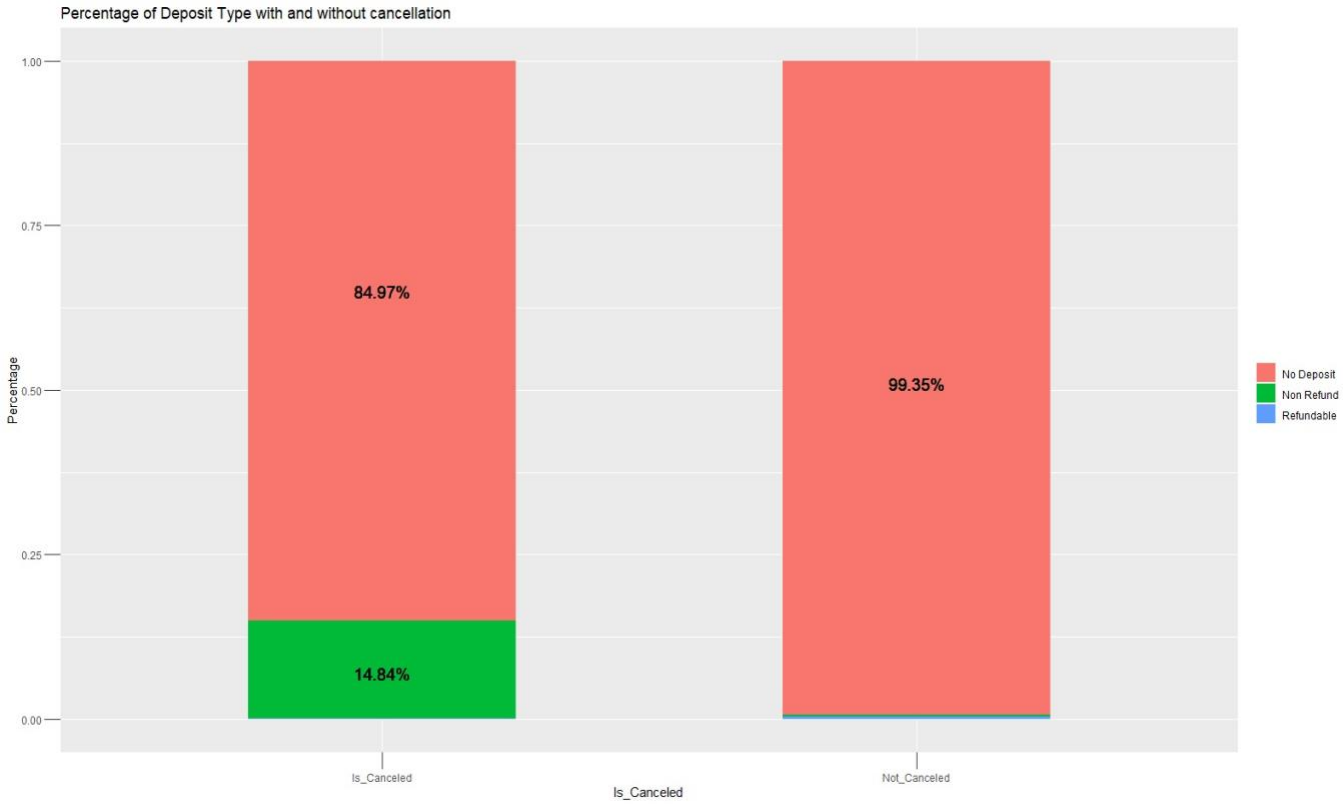
g) Special Requests



From here, we can see that when there are no special requests, the percentage of bookings cancelled is higher than the percentage of bookings not cancelled. And as the no of special request increases the percentage of cancellation decreases. With only one special request, there is 31.82% chance of booking not being canceled. Hence, if the hotel takes care of special requests made by the customer, the cancellation percentage will be decreased.

h) Deposit Type

Count of IsCanceled	**Column Labels**			
Row Labels	**No Deposit**	**Non-Refund**	**Refundable**	**Grand Total**
0	28749	69	120	28938
1	9450	1650	22	11122
Grand Total	**38199**	**1719**	**142**	**40060**



From the above plot, we can see that when a customer puts in a non-refundable deposit, there are still chances of the customer cancelling their reservation. Customers do not mind cancelling their trip even if they pay a refundable deposit.

- **Different Models**

After analyzing the data and making sure that we do not have any null values. We produced a set of 6 features which we thought would be the best for running the model. We ran varIMP function on our dataset to see which features have the highest importance leading to accurate prediction of booking cancellations. The higher the value, the more its significance is of that feature. We realized Lead Time, Required Parking Space, Market Segment, Country is a few of the most significant variables and used those features in our CART and SVM Model.

Note: The train() command that computed the Regression Trees and SVMs were not included in these results because their ksvm() and rpart() counterparts had better performances (like accuracy).

a) Linear model

- The dataset was modeled using linear regression only using the important variables:

```
'data.frame':   40026 obs. of  19 variables:
 $ IsCanceled                : int  0 0 0 0 0 0 0 0 1 1 ...
 $ LeadTime                  : int  342 737 7 13 14 14 0 9 85 75 ...
 $ StaysInWeekendNights      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ StaysInWeekNights         : 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                      : num  1 1 1 1 1 1 1 2 1 3 ...
 $ Country                   : num  97 97 46 46 46 46 97 97 97 97 ...
 $ MarketSegment             : num  3 3 3 2 6 6 3 3 6 5 ...
 $ IsRepeatedGuest           : num  1 1 1 1 1 1 1 1 1 1 ...
 $ PreviousCancellations     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PreviousBookingsNotCanceled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ ReservedRoomType          : num  3 3 1 1 1 1 3 3 1 4 ...
 $ AssignedRoomType          : num  3 3 3 1 1 1 3 3 1 4 ...
 $ BookingChanges            : num  4 5 1 1 1 1 1 1 1 1 ...
 $ DepositType               : num  1 1 1 1 1 1 1 1 1 1 ...
 $ CustomerType              : num  3 3 3 3 3 3 3 3 3 3 ...
 $ TotalOfSpecialRequests    : int  0 0 0 0 1 1 0 1 1 0 ...
```

```
> LinearModel=lm(IsCanceled~.,data=CIV)
> summary(LinearModel)

Call:
lm(formula = IsCanceled ~ ., data = CIV)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99722 -0.26860 -0.09507  0.29716  2.12103

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.272e-01  8.165e-02   2.783 0.005396 **
LeadTime                   1.022e-03  2.368e-05  43.150  < 2e-16 ***
RequiredCarParkingSpaces  -2.737e-01  5.581e-03 -49.039  < 2e-16 ***
MarketSegmentCorporate    -5.507e-02  2.795e-02  -1.970 0.048790 *
MarketSegmentDirect       -7.580e-03  2.720e-02  -0.279 0.780511
MarketSegmentGroups        1.406e-01  2.744e-02   5.125 2.99e-07 ***
MarketSegmentOffline TA/TO -3.224e-02  2.730e-02  -1.181 0.237517
MarketSegmentOnline TA     2.077e-01  2.701e-02   7.690 1.50e-14 ***
PreviousCancellations      2.568e-02  1.436e-03  17.886  < 2e-16 ***
CountryALB                -2.259e-03  2.323e-01  -0.010 0.992243
CountryAND                 2.054e-01  1.865e-01   1.101 0.270717
CountryARE                 2.769e-01  1.381e-01   2.004 0.045024 *
CountryARG                -2.095e-01  9.231e-02  -2.270 0.023232 *
CountryARM                -3.223e-01  2.792e-01  -1.154 0.248401
CountryAUS                -2.424e-01  8.748e-02  -2.771 0.005597 **
CountryAUT                -2.441e-01  8.178e-02  -2.984 0.002845 **
CountryAZE                -2.847e-01  2.323e-01  -1.225 0.220400
CountryBDI                -4.210e-01  3.872e-01  -1.087 0.276878
CountryBEL                -3.375e-01  7.951e-02  -4.244 2.20e-05 ***
CountryBGR                -4.073e-01  1.865e-01  -2.184 0.028969 *
CountryBHR                 4.639e-01  3.871e-01   1.198 0.230760
CountryBHS                -4.239e-01  3.872e-01  -1.095 0.273530
CountryBIH                -2.196e-01  3.871e-01  -0.567 0.570538
CountryBLR                -1.407e-02  1.629e-01  -0.086 0.931199
CountryBRA                -1.991e-01  7.958e-02  -2.502 0.012366 *
CountryBWA                -5.545e-01  3.871e-01  -1.432 0.152080
CountryCAF                -3.921e-01  2.323e-01  -1.688 0.091418 .
CountryCHE                -1.776e-01  7.956e-02  -2.232 0.025633 *
CountryCHL                -2.293e-01  1.203e-01  -1.907 0.056498 .
CountryCHN                -3.624e-01  8.411e-02  -4.308 1.65e-05 ***
CountryCIV                -2.902e-01  2.792e-01  -1.039 0.298665
CountryCMR                -5.949e-01  2.792e-01  -2.131 0.033105 *
CountryCN                 -3.418e-01  7.878e-02  -4.339 1.43e-05 ***
CountryCOL                -1.943e-01  1.225e-01  -1.587 0.112551
CountryCOM                -4.349e-01  3.871e-01  -1.123 0.261303
CountryCPV                -3.666e-01  1.865e-01  -1.966 0.049330 *
CountryCRI                -3.614e-01  2.792e-01  -1.294 0.195537
CountryCUB                -3.477e-01  2.049e-01  -1.697 0.089689 .
CountryCYM                -5.688e-01  3.871e-01  -1.469 0.141767
CountryCYP                -6.701e-02  1.549e-01  -0.433 0.665265
CountryCZE                -3.603e-01  1.065e-01  -3.385 0.000713 ***
CountryDEU                -3.564e-01  7.825e-02  -4.555 5.26e-06 ***
CountryDJI                -4.482e-01  3.871e-01  -1.158 0.246985
CountryDNK                -3.658e-01  9.064e-02  -4.035 5.46e-05 ***
CountryDOM                -5.593e-01  2.323e-01  -2.408 0.016060 *
CountryDZA                 2.185e-02  1.341e-01   0.163 0.870564
CountryECU                -9.810e-02  2.792e-01  -0.351 0.725279
CountryEGY                -4.492e-01  3.871e-01  -1.160 0.245909
CountryESP                -1.806e-01  7.768e-02  -2.325 0.020087 *
CountryEST                -3.455e-01  1.018e-01  -3.393 0.000691 ***
CountryFIN                -3.194e-01  8.337e-02  -3.831 0.000128 ***
CountryFJI                 1.962e-01  3.872e-01   0.507 0.612253
CountryFRA                -2.756e-01  7.803e-02  -3.532 0.000413 ***
```

```
CountryGBR               -3.678e-01  7.764e-02  -4.737 2.17e-06 ***
CountryGEO                2.084e-02  1.382e-01   0.151 0.880098
CountryGGY                5.355e-01  3.871e-01   1.383 0.166606
CountryGIB               -1.095e-01  1.307e-01  -0.838 0.401804
CountryGRC               -2.385e-01  1.428e-01  -1.670 0.094864 .
CountryHKG                2.108e-01  2.049e-01   1.029 0.303386
CountryHRV               -2.391e-01  1.381e-01  -1.731 0.083462 .
CountryHUN               -3.075e-01  9.519e-02  -3.230 0.001239 **
CountryIDN                1.251e-01  1.865e-01   0.671 0.502230
CountryIND               -3.088e-01  9.943e-02  -3.106 0.001899 **
CountryIRL               -3.365e-01  7.792e-02  -4.318 1.58e-05 ***
CountryIRN               -4.949e-01  1.865e-01  -2.654 0.007962 **
CountryISL               -4.252e-01  1.732e-01  -2.455 0.014079 *
CountryISR               -8.793e-02  1.056e-01  -0.833 0.404845
CountryITA               -2.629e-01  7.945e-02  -3.309 0.000936 ***
CountryJAM               -4.055e-01  1.865e-01  -2.174 0.029677 *
CountryJEY                4.251e-01  2.323e-01   1.830 0.067279 .
CountryJOR               -4.487e-01  2.792e-01  -1.607 0.108010
CountryJPN               -3.662e-01  1.483e-01  -2.470 0.013522 *
CountryKAZ               -2.362e-01  1.865e-01  -1.267 0.205239
CountryKOR               -1.322e-01  1.483e-01  -0.892 0.372536
CountryKWT               -1.193e-01  2.323e-01  -0.514 0.607447
CountryLBN                2.241e-02  1.731e-01   0.129 0.897037
CountryLKA               -2.420e-01  3.871e-01  -0.625 0.531882
CountryLTU               -4.149e-01  9.555e-02  -4.342 1.42e-05 ***
CountryLUX               -1.361e-01  8.831e-02  -1.541 0.123303
CountryLVA               -3.265e-01  1.018e-01  -3.208 0.001338 **
CountryMAC               -4.778e-01  3.871e-01  -1.234 0.217115
CountryMAR                5.693e-02  8.897e-02   0.640 0.522270
CountryMDG               -2.196e-01  3.871e-01  -0.567 0.570538
CountryMDV                1.584e-01  1.731e-01   0.915 0.360253
CountryMEX               -4.171e-01  1.731e-01  -2.409 0.016014 *
CountryMKD               -5.546e-01  3.871e-01  -1.432 0.152009
CountryMLT               -1.450e-01  2.792e-01  -0.519 0.603482
CountryMOZ               -4.268e-03  1.731e-01  -0.025 0.980336
CountryMUS               -5.903e-01  3.871e-01  -1.525 0.127312
CountryMWI               -1.755e-01  2.792e-01  -0.629 0.529626
CountryMYS               -5.083e-01  1.428e-01  -3.560 0.000371 ***
CountryNGA               -1.983e-01  1.428e-01  -1.389 0.164984
CountryNLD               -3.187e-01  7.924e-02  -4.022 5.77e-05 ***
CountryNOR               -3.195e-01  8.470e-02  -3.772 0.000162 ***
CountryNPL               -4.512e-01  3.871e-01  -1.166 0.243767
CountryNULL              -1.321e-01  7.948e-02  -1.662 0.096577 .
```

```
CountryNZL                -4.096e-01  1.276e-01  -3.211 0.001324 **
CountryOMN                -4.431e-01  1.381e-01  -3.208 0.001339 **
CountryPAK                 6.888e-02  2.049e-01   0.336 0.736680
CountryPER                -1.755e-01  3.871e-01  -0.453 0.650329
CountryPHL                 7.854e-02  1.225e-01   0.641 0.521295
CountryPLW                -6.741e-01  3.871e-01  -1.741 0.081659 .
CountryPOL                -3.484e-01  8.021e-02  -4.344 1.41e-05 ***
CountryPRI                -2.424e-01  1.484e-01  -1.633 0.102442
CountryPRT                 2.469e-02  7.751e-02   0.319 0.750076
CountryQAT                -4.522e-01  3.871e-01  -1.168 0.242734
CountryROU                -3.104e-01  8.255e-02  -3.759 0.000171 ***
CountryRUS                -1.742e-01  8.223e-02  -2.118 0.034148 *
CountrySAU                -1.898e-01  3.871e-01  -0.490 0.623906
CountrySEN                 2.554e-01  3.871e-01   0.660 0.509474
CountrySGP                -4.641e-01  2.049e-01  -2.266 0.023481 *
CountrySMR                -3.694e-01  3.872e-01  -0.954 0.340140
CountrySRB                -3.458e-01  1.630e-01  -2.122 0.033824 *
CountrySUR                -4.289e-01  2.049e-01  -2.094 0.036282 *
CountrySVK                -2.134e-01  1.341e-01  -1.591 0.111663
CountrySVN                -3.315e-01  1.381e-01  -2.400 0.016384 *
CountrySWE                -2.502e-01  8.047e-02  -3.109 0.001880 **
CountrySYC                 3.883e-01  3.871e-01   1.003 0.315811
CountrySYR                -4.615e-01  3.871e-01  -1.192 0.233227
CountryTGO                -6.271e-01  3.871e-01  -1.620 0.105268
CountryTHA                -2.979e-01  1.731e-01  -1.720 0.085355 .
CountryTUN                 4.987e-01  3.871e-01   1.288 0.197679
CountryTUR                -1.631e-01  1.107e-01  -1.473 0.140731
CountryTWN                -5.677e-01  1.342e-01  -4.231 2.33e-05 ***
CountryUGA                -4.543e-01  3.871e-01  -1.174 0.240580
CountryUKR                -4.917e-01  1.107e-01  -4.440 9.02e-06 ***
CountryURY                -9.073e-02  1.549e-01  -0.586 0.558026
CountryUSA                -2.348e-01  7.936e-02  -2.958 0.003094 **
CountryUZB                -2.329e-01  3.871e-01  -0.602 0.547412
CountryVEN                -6.316e-02  2.323e-01  -0.272 0.785695
CountryVNM                -5.795e-01  2.792e-01  -2.076 0.037925 *
CountryZAF                -2.224e-01  1.183e-01  -1.880 0.060111 .
CountryZMB                -5.862e-01  3.871e-01  -1.514 0.129983
CountryZWE                -3.620e-01  2.792e-01  -1.297 0.194797
StaysInWeekNights          1.331e-02  8.808e-04  15.109  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3793 on 39925 degrees of freedom
Multiple R-squared:  0.2851,    Adjusted R-squared:  0.2827
F-statistic: 118.8 on 134 and 39925 DF,  p-value: < 2.2e-16
```

The adjusted R-squared from this model is 0.2827. The adjusted R-squared of 0.2827 means that the "LeadTime", "RequiredCarParkingSpaces", "MarketSegment", "PreviousCancellations", "Country", and "StaysInWeekNights" variables explain 28.27% of the "IsCanceled" (whether or not the client cancelled their booking) variable's variation. The p-value of the model is 2.2e-16, so it is highly likely that the changes caused by these variables do not occur by chance.

The "LeadTime", "RequiredCarParkingSpaces", "PreviousCancellations", and "StaysInWeekNights" are statistically significant without any conditions because they are metric variables. The "MarketSegment" variable is statistically significant when its value is equal to either "Corporate", "Groups", or "Online TA". Similarly, the "Country" variable is statistically significant when its value is equal to either "ARE", "ARG", "AUS", "AUT", "BEL", "BGR", "BRA", "CHE", "CHN", "CMR", "CN", "CPV", "CZE", "DEU", "DNK", "DOM", "ESP", "EST", "FIN", "FRA", "GBR", "HUN", "IND", "IRL", "IRN", "ISL", "ITA", "JAM", "JPN", "LTU", "LVA", "MEX", "MYS", "NLD", "NOR", "NZL", "OMN", "POL", "ROU", "RUS", "SGP", "SRB", "SUR", "SVN", "SWE", "TWN", "UKR", "USA", or "VNM".

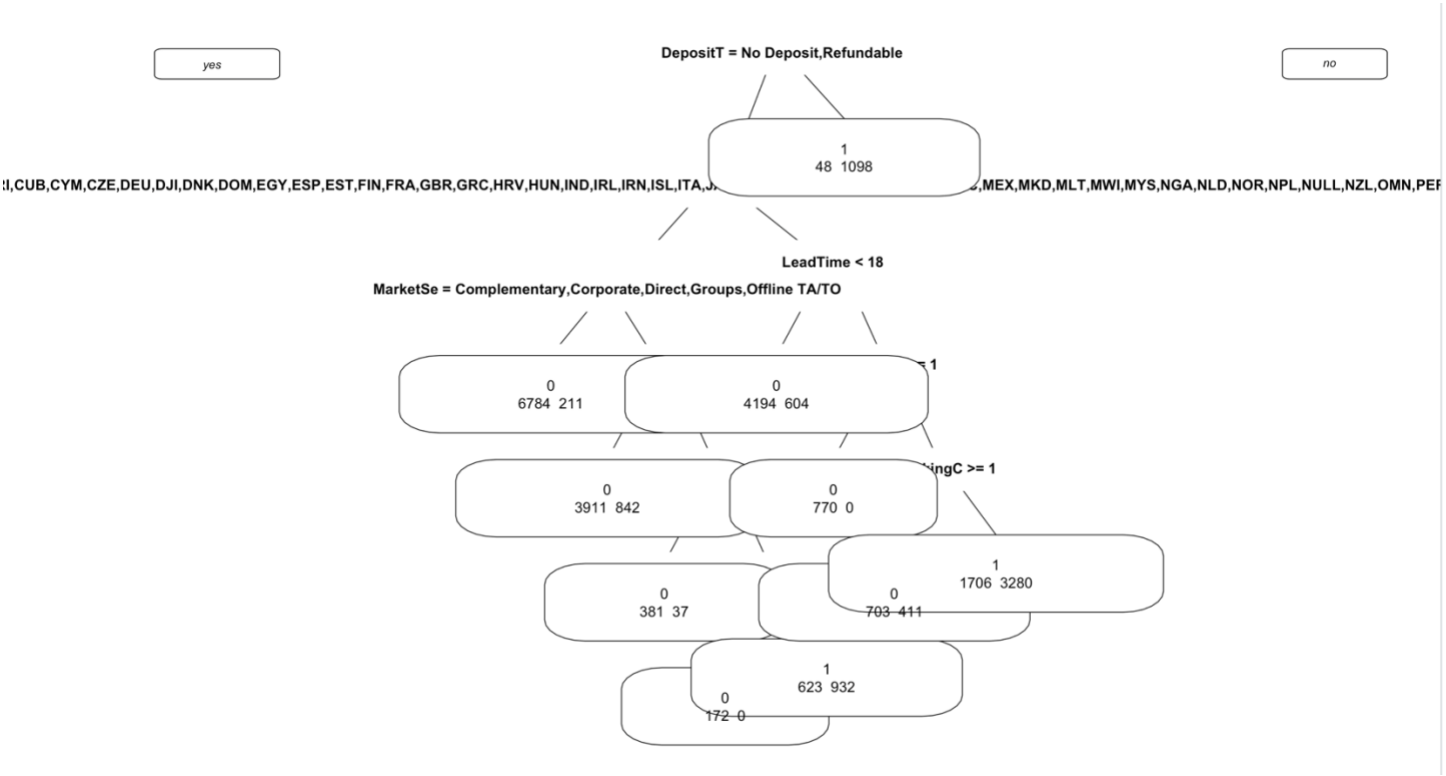
These statistical significances mean that, for every unit increase in the "LeadTime" variable, the score for a hotel booking cancellation increases by 0.001021611. For every unit increase in the "RequiredCarParkingSpaces" variable (for every additional day), the score for a hotel booking cancellation decreases by 0.273710333.

If the value for the "MarketSegment" variable is equal to "Corporate", the score for a hotel booking cancellation decreases by 0.055068980. If the value for the "MarketSegment" variable is equal to "Groups", the score for a hotel booking cancellation increases by 0.140619578. If the value for the "MarketSegment" variable is equal to "Online TA", the score for a hotel booking cancellation increases by 0.207684142. For every unit increase in the "PreviousCancellations" variable (for every additional previous cancellation), the score for a hotel booking cancellations increases by 0.025679557. If the value for the "Country" variable is equal to "ARE", the score for a hotel booking cancellation increases by 0.276874182. If the value for the "Country" variable is equal to "ARG", the score for a hotel booking cancellation decreases by 0.209516692. If the value for the "Country" variable is equal to "AUS", the score for a hotel cancellation decreases by 0.242363906. If the value for the "Country" variable is equal to "AUT", the score for a hotel booking cancellation decreases by 0.24405749. If the value for the "Country" variable is equal to "BEL", the score for a hotel booking cancellations decreases by 0.33745448. If the value for the "Country" variable is equal to "BGR", the score for a hotel booking cancellation decreases by 0.40725660. If the value for the "Country" variable is equal to "BRA", the score for a hotel booking cancellation decreases by 0.19908075. If the value for the "Country" variable is equal to "CHE", the score for a hotel booking cancellation decreases by 0.17756059. If the value for the "Country" variable is equal to "CHN", the score for a hotel booking cancellation decreases by 0.36236998. If the value for the "Country" variable is equal to "CMR", the score for a hotel booking cancellation decreases by 0.59487863. If the value for the "Country" variable is equal to "CN", the score for a hotel booking cancellation decreases by 0.34184074. If the value for the "Country" variable is equal to "CPV", the score for a hotel booking cancellation decreases by 0.36656955. If the value for the "Country" variable is equal to "CZE", the score for a hotel booking cancellation decreases by 0.36031245. If the value for the "Country" variable is equal to "DEU", the score for a hotel booking cancellation decreases by 0.35642467. If the value for the "Country" variable is equal to "DNK", the score for a hotel booking cancellation decreases by 0.36575231. If the value for the "Country" variable is equal to "DOM", the score for a hotel booking cancellation decreases by 0.55926935. If the value for the "Country" variable is equal to "ESP", the score for a hotel booking cancellation decreases by 0.18059302. If the value for the "Country" variable is equal to "EST", the score for a hotel booking cancellation decreases by 0.34546621. If the value for the "Country" variable is equal to "FIN", the score for a hotel booking cancellation decreases by 0.31936623. If the value for the "Country" variable is equal to "FRA", the score for a hotel booking cancellation decreases by 0.27557225. If the value for the "Country" variable is equal to "GBR", the score for a hotel booking cancellation decreases by 0.36779672. If the value for the "Country" variable is equal to "HUN", the score for a hotel booking cancellation decreases by 0.30747982. If the value for the "Country" variable is equal to "IND", the score for a hotel booking cancellation decreases by 0.30882464. If the value for the "Country" variable is equal to "IRL", the score for a hotel booking cancellation decreases by 0.33645439. If the value for the "Country" variable is equal to "IRN", the score for a hotel booking cancellation decreases by 0.49488468. If the value for the "Country" variable is equal to "ISL", the score for a hotel booking cancellation decreases by 0.42515268. If the value for the "Country" variable is equal to "ITA", the score for a hotel booking cancellation decreases by 0.26291873. If the value for the "Country" variable is equal to "JAM", the score for a hotel booking cancellation decreases by 0.40549994. If the value for the "Country" variable is equal to "JPN", the score for a hotel booking cancellation decreases by 0.36619693. If the value for the "Country" variable is equal to "LTU", the score for a hotel booking cancellation decreases by 0.41486293. If the value for the "Country" variable is equal to "LVA", the score for a hotel booking cancellation decreases by 0.32654568. If the value for the "Country" variable is equal to "MEX", the score for a hotel booking cancellation decreases by 0.41705237. If the value for the "Country" variable is equal to "MYS", the score for a hotel booking cancellation decreases by 0.50832586. If the value for the "Country" variable is equal to "NLD", the score for a hotel booking cancellation decreases by 0.31872093. If the value for the "Country" variable is equal to "NOR", the

score for a hotel booking cancellation decreases by 0.31954483. If the value for the "Country" variable is equal to "NZL", the score for a hotel booking cancellation decreases by 0.40964462. If the value for the "Country" variable is equal to "OMN", the score for a hotel booking cancellation decreases by 0.44310324. If the value for the "Country" variable is equal to "POL", the score for a hotel booking cancellation decreases by 0.34839527. If the value for the "Country" variable is equal to "ROU", the score for a hotel booking cancellation decreases by 0.31035796. If the value for the "Country" variable is equal to "RUS", the score for a hotel booking cancellation decreases by 0.17419130. If the value for the "Country" variable is equal to "SGP", the score for a hotel booking cancellation decreases by 0.46411686. If the value for the "Country" variable is equal to "SRB", the score for a hotel booking cancellation decreases by 0.34582290. If the value for the "Country" variable is equal to "SUR", the score for a hotel booking cancellation decreases by 0.42893594. If the value for the "Country" variable is equal to "SVN", the score for a hotel booking cancellation decreases by 0.33152122. If the value for the "Country" variable is equal to "SWE", the score for a hotel booking cancellation decreases by 0.25015330. If the value for the "Country" variable is equal to "TWN", the score for a hotel booking cancellation decreases by 0.56768091. If the value for the "Country" variable is equal to "UKR", the score for a hotel booking cancellation decreases by 0.49165922. If the value for the "Country" variable is equal to "USA", the score for a hotel booking cancellation decreases by 0.23477465. If the value for the "Country" variable is equal to "VNM", the score for a hotel booking cancellation decreases by 0.57948500. Lastly, for every unit increase in the "StaysInWeekNights" variable (for every additional week night a customer stayed in the hotel), the score for a hotel booking cancellation increases by 0.01330873.

a) CART Model: Decision Tree

The working of decision tree models is based on repeated partitioning the data into multiple sub-spaces, so that the outcome in each final sub-space is as homogeneous as possible. This approach is technically called recursive partitioning.



Classification And Regression Trees (CART) algorithm is a classification algorithm for building a decision tree based on Gini's impurity index as splitting criterion. CART is a binary tree build by splitting node into two child nodes repeatedly.

```
> cartdata <-data.frame(varImp(cartTree))
> cartdata
                              Overall
AssignedRoomType             231.72289
BookingChanges               386.71823
Children                     135.01671
Country                     1551.79289
CustomerType                 279.67768
DepositType                 1108.87652
LeadTime                    2895.74226
MarketSegment               2273.17627
Meal                          91.26208
PreviousCancellations        248.44296
RequiredCarParkingSpaces    2731.55957
StaysInWeekendNights         386.44532
StaysInWeekNights            373.12705
TotalOfSpecialRequests       222.89146
Adults                         0.00000
Babies                         0.00000
IsRepeatedGuest                0.00000
PreviousBookingsNotCanceled    0.00000
ReservedRoomType               0.00000
```

From the above decision tree data we can see that LeadTime, RequiredCarParkingSpaces, MarketSegment, Country are top contributors to the model's accuracy.

```
> confusionMatrix(as.factor(CartPrediction[,2]),testSet$IsCanceled)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8498 1105
         1 1148 2602

               Accuracy : 0.8313
                 95% CI : (0.8248, 0.8376)
    No Information Rate : 0.7224
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5808

 Mcnemar's Test P-Value : 0.3762

            Sensitivity : 0.8810
            Specificity : 0.7019
         Pos Pred Value : 0.8849
         Neg Pred Value : 0.6939
             Prevalence : 0.7224
         Detection Rate : 0.6364
   Detection Prevalence : 0.7192
      Balanced Accuracy : 0.7915

       'Positive' Class : 0
```

We can see that we get an accuracy of 0.8313 with an NIR value of 0.7224 along with p value being less than alpha level of 0.05 and we can see that we have a good model.

b) SVM

```
> library(kernlab)
> SVM=ksvm(IsCanceled~.,data=trainSet,C=5,cross=3,prob.model=TRUE)
> SVM
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.217830415292759

Number of Support Vectors : 10014

Objective Function Value : -46017.28
Training error : 0.163815
Cross validation error : 0.170629
Probability model included.
> Prediction=predict(SVM,newdata=testSet,type="response")
> confusionMatrix(Prediction,testSet$IsCanceled)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8925 1470
         1  721 2237

               Accuracy : 0.8359
                 95% CI : (0.8295, 0.8422)
    No Information Rate : 0.7224
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5638

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9253
            Specificity : 0.6035
         Pos Pred Value : 0.8586
         Neg Pred Value : 0.7563
             Prevalence : 0.7224
         Detection Rate : 0.6684
   Detection Prevalence : 0.7785
      Balanced Accuracy : 0.7644

       'Positive' Class : 0
```

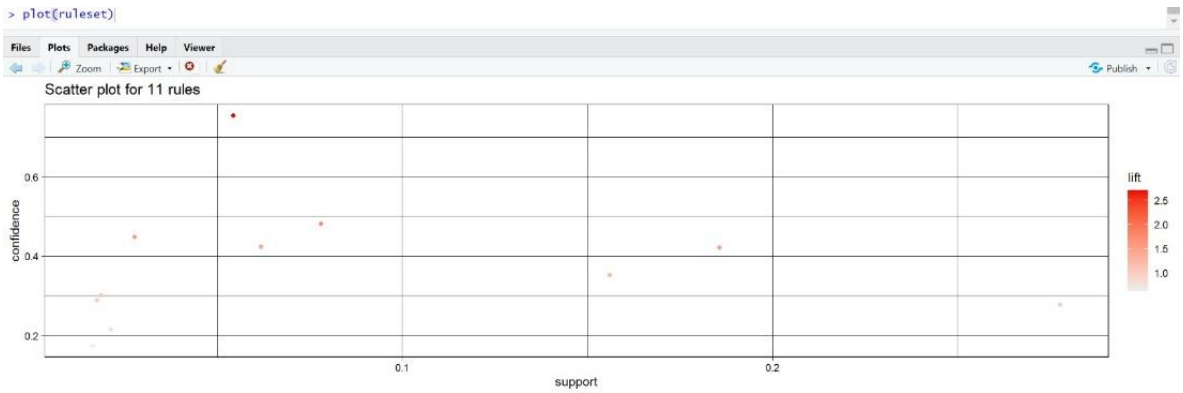
A p-value of <2.2e-16 for Accuracy > NIR, the null hypothesis is rejected, therefore, the accuracy is statistically significantly better than the No Information Rate.

c) Association Rules

```
> ruleset=apriori(Transaction,parameter=list(supp=0.012,conf=0.172),control=list(verbose=FALSE),appearance=list(default="lhs",rhs="IsCanceled=1"))
> inspect(ruleset)
     lhs                                      rhs             support    confidence coverage   lift      count
[1]  {}                                    => {IsCanceled=1} 0.27763355 0.2776335  1.00000000 1.0000000 11122
[2]  {Country=ESP}                         => {IsCanceled=1} 0.02124314 0.2150619  0.09877683 0.7746251   851
[3]  {MarketSegment=Groups}                => {IsCanceled=1} 0.06175736 0.4239205  0.14568148 1.5269066  2474
[4]  {Country=PRT}                         => {IsCanceled=1} 0.18567149 0.4218945  0.44008987 1.5196092  7438
[5]  {MarketSegment=Online TA}             => {IsCanceled=1} 0.15596605 0.3524169  0.44256116 1.2693601  6248
[6]  {MarketSegment=Online TA,Country=ESP} => {IsCanceled=1} 0.01857214 0.3001210  0.06188218 1.0809969   744
[7]  {MarketSegment=Groups,Country=PRT}    => {IsCanceled=1} 0.05434348 0.7545927  0.07201697 2.7179450  2177
[8]  {MarketSegment=Direct,Country=PRT}    => {IsCanceled=1} 0.01640040 0.1736258  0.09445831 0.6253776   657
[9]  {MarketSegment=Online TA,Country=GBR} => {IsCanceled=1} 0.01752371 0.2880591  0.06083375 1.0375514   702
[10] {MarketSegment=Offline TA/TO,Country=PRT} => {IsCanceled=1} 0.02773340 0.4481646  0.06188218 1.6142306  1111
[11] {MarketSegment=Online TA,Country=PRT} => {IsCanceled=1} 0.07798303 0.4815785  0.16193210 1.7345834  3124
> AcceptableLiftRuleset=ruleset[quality(ruleset)$lift>1]
> inspect(AcceptableLiftRuleset)
     lhs                                      rhs             support    confidence coverage   lift     count
[1]  {MarketSegment=Groups}                => {IsCanceled=1} 0.06175736 0.4239205  0.14568148 1.526907 2474
[2]  {Country=PRT}                         => {IsCanceled=1} 0.18567149 0.4218945  0.44008987 1.519609 7438
[3]  {MarketSegment=Online TA}             => {IsCanceled=1} 0.15596605 0.3524169  0.44256116 1.269360 6248
[4]  {MarketSegment=Online TA,Country=ESP} => {IsCanceled=1} 0.01857214 0.3001210  0.06188218 1.080997  744
[5]  {MarketSegment=Groups,Country=PRT}    => {IsCanceled=1} 0.05434348 0.7545927  0.07201697 2.717945 2177
[6]  {MarketSegment=Online TA,Country=GBR} => {IsCanceled=1} 0.01752371 0.2880591  0.06083375 1.037551  702
[7]  {MarketSegment=Offline TA/TO,Country=PRT} => {IsCanceled=1} 0.02773340 0.4481646  0.06188218 1.614231 1111
[8]  {MarketSegment=Online TA,Country=PRT} => {IsCanceled=1} 0.07798303 0.4815785  0.16193210 1.734583 3124
```

```
> crossTable(Transaction,sort=TRUE)[1:5,1:5]
                          IsCanceled=0 MarketSegment=Online TA Country=PRT IsCanceled=1 MarketSegment=Offline TA/TO
IsCanceled=0                     28938                   11481       10192           0                         6334
MarketSegment=Online TA          11481                   17729        6487        6248                            0
Country=PRT                      10192                    6487       17630        7438                         2479
IsCanceled=1                         0                    6248        7438       11122                         1138
MarketSegment=Offline TA/TO       6334                       0        2479        1138                         7472
```

```
> itemFrequencyPlot(Transaction,topN=10)
```



There's a 5.43438% chance that a customer has "Groups" as their Market Segment, "PRT" as their Country, and cancelled their hotel booking.

If a customer has "Groups" as its Market Segment and "PRT" as their Country, there's a 75.45927% chance the customer will cancel their hotel booking.

There's a 7.798303% chance that a customer has "Online TA" as their Market Segment, "PRT" as their Country, and cancelled their hotel booking.

If a customer has "Online TA" as its Market Segment, and "PRT" as their Country, there's a 48.15785% chance the customer will cancel their hotel booking.

- # Conclusion

We initially started with all 20 variables and then reduced to 11 variables which we thought would be important. After running through CART, we realized the 6 most important variables and went through with them in our completed model.

In our introduction, we stated the goal of our analysis is to answer two key questions

- Does Lead Time and Market Segment related to booking cancelation?
- Can we predict with accuracy if a booking will be cancelled based on the attributes?

Based on our data analysis, these are the significant and potential casual relationships that affect booking cancellation

a) Lead Time

Based on our analysis, we learned Lead Time is one of the most significant factors that affect cancellation. When lead time increases, the customers are more likely to cancel. We observed this from CART analysis.

We see people book their reservations 2 months in advance and if someone cancels their booking when there are less than 60 days (about 2 months) left for their check in, the hotel can charge them as convenience fee.

b) Market Segment

One factor we found from the linear model, which is significant, is Market Segment and CART analysis gave us a similar result. We see that the maximum percentage of cancellation is done by groups and travel agents. The hotel needs to rethink their terms with online travel agents and penalty for group cancellations.

c) Deposit Type

We see customers are not bothered by the monetary penalty when deciding to cancel. For hotels, making Non-Refundable rooms will not prevent cancellation.

d) Required Car Parking/Special Requests

We see that when customers reserve a parking space when they are booking their hotel reservation, the chances of them cancelling their reservation decreases. Similarly, with special requests. Giving a customer what he wants makes the cancellation rate less.

e) Can we predict with accuracy if a booking will be cancelled based on its attributes

Through our analysis, we think the Support Vector Machine model works well to predict hotel cancellations. We tested two other models to see how accurate our prediction models are. CART Model gives us an accuracy of 81.08% whereas Support Vector Machine gives us the highest accuracy of 83.59%

• Recommendation

- Determining a threshold number of days for customers to pay convenience fee if booking is cancelled

- Customers who book 2 months or more time ago have less chance of cancellation.
- The cancellations in the category of market segment are done by online travel agents, the hotel can change terms and conditions they have with online agents to avoid more cancellation from online agents
- The recommendation we have is the booking deposit should be dependent on how many days ago booking is made.

Limitation

- Lack of customer behavior data
- From our analysis, we discovered that variables relating to the customer's behavior, such as special requests, show significance in leading to booking cancellation. Unfortunately, our dataset is limited in providing variables that relate to customer behavior. As we do not know if the special requests made by the customer were taken care of or not. However, we understand such data is not readily available and might be difficult to monitor.

Future Studies

Include more customer behavior data

One recommendation that can improve our study is to source customer behavior data. Examples include data which provides information about customers' browsing sessions, type of special request and if that special request was fulfilled or not. With this additional data, we might understand the customer's booking journey in greater details and identify alternate good predictors for booking cancellation.

Cross reference results and findings from different hotels

To improve our model and increase its relevance to different hotels, one recommendation is to perform a similar analysis of data from other hotels and cross-reference the results. Doing this allows us to verify our insights.