



HOTEL CANCELLATION PREDICTION



FINAL REPORT

SIVARAMAKRISHNAN S – JULY 2020 E

TABLE OF CONTENTS

SOLUTION 1: INTRODUCTION.....	1
ISSUES IN HOTEL INDUSTRY.....	1
NEED OF THE STUDY/PROJECT	1
UNDERSTANDING BUSINESS/ SOCIAL OPPORTUNITY.....	1
SOLUTION 2: EDA AND BUSINESS IMPLICATION	2
VISUAL INSPECTION OF DATA (ROWS, COLUMNS, DESCRIPTIVE DETAILS).....	2
<i>Descriptive Statistics of Numerical Variables</i>	<i>2</i>
<i>Descriptive Statistics of Categorical Variables.....</i>	<i>2</i>
UNDERSTANDING OF ATTRIBUTES (VARIABLE INFO, RENAMING IF REQUIRED).....	3
Variable info	3
Duplicate Value Check.	3
UNIVARIANT ANALYSIS	3
Target Variable	3
Hotel Type.....	3
Deposit Type	4
Customer Type.....	4
Meal Type.....	4
Repeated Guest.....	5
Market Segment	5
Booking Changes.....	5
Location.....	6
Total Guest.....	6
Total Stays.....	7
Lead Time - & Lead Time - Month.....	7
Assigned Room Type.....	8
RESERVED ROOM TYPE	9
BIVARIANT ANALYSIS - AFTER DATA CLEANING	9
Previous Cancellation VS Repeated Guest	9
Deposit Type VS Market Segment	9
Hotel Type and Cancellation.....	10
Total Guest and Cancellation.....	10
Meal Type and Cancellation.....	10
Location and Cancellation.....	11
Market Segment and Cancellation	11
Deposit Type and Cancellation	11
Previous Cancellation and Cancellation.....	12
Lead Time and Cancellation	12
Total Number of Special request and Cancellation.....	13
Total Number of Parking space request and Cancellation.....	13
Multi-Variant Analysis	14
SOLUTION 3: DATA CLEANING.....	15
Missing Values Treatment	15
Outlier Treatment.....	15
Variable Transformations	16
Pre-process the variable for the model building.....	16
One Hot Encoding.....	17
SOLUTION 4: MODEL BUILDING.....	18
VALIDATE THE RELATIONSHIP	18
IDENTIFY THE CORRELATION OF THE PREDICTORS USING (VIF)	18
SPLIT DATA FOR TRAINING AND TEST DATA	19
SCALING	20
CLASSIFICATION MODEL - WHY WAS A PARTICULAR MODEL(S) CHOSEN?.....	20
Base and Tuned Model	20
MODEL BUILDING.....	21
Logistic Regression	21
LDR (Linear Discriminant Analysis).....	22
Random Forest Classifier	23
Naïve Bayes.....	24
K - Nearest Neighbour	25
Artificial Neural Networks (ANN)	26
XGBoost	27
Boosting.....	28
Bagging Classifier.....	29
EFFORTS TO IMPROVE MODEL PERFORMANCE	29
SOLUTION 4: MODEL VALIDATION	30
HOW WAS THE MODEL VALIDATED?.....	30
Bias / Training Error :	30

Variance / Test Error :.....30
 JUST ACCURACY, OR ANYTHING ELSE TOO?.....31
 ROC Comparison on Tuned Models31
SOLUTION 5: FINAL INTERPRETATION / RECOMMENDATION 32
 FINAL INTERPRETATION32
 Optimum Model Interpretation.....32
 BUSINESS INSIGHTS.....33
 RECOMMENDATIONS34
 Few Strategies to increase Direct Booking.....34

TABLE OF TABLES
 TABLE 1 MISSING VALUES 15
 TABLE 2 NEW VARIABLES 16
 TABLE 3 VARIABLES TRANSFORMATIONS 16
 TABLE 4 VARIABLES REMOVED 16
 TABLE 5 PRE-PROCESSING BINARY 16
 TABLE 6 DROPPED VARIABLES 17
 TABLE 7 HIGHER CORRELATION ON TARGET 18
 TABLE 8 HIGH VIF VARIABLES 19
 TABLE 9 TARGET SPLIT ON TRAIN AND TEST 19
 TABLE 10 CHOOSING CLASSIFICATION MODELS 20
 TABLE 11 BUSINESS INSIGHTS 33
 TABLE 12 RECOMMENDATIONS 34

TABLE OF FIGURES
 FIGURE 1 CANCELLATION RATE 1
 FIGURE 2 ROWS AND COLUMNS 2
 FIGURE 3 DESCRIPTIVE STATISTICS NUMERICAL 2
 FIGURE 4 DESCRIPTIVE STATISTICS CATEGORICAL 2
 FIGURE 5 DATA INFO 3
FIGURE 6 TARGET DISTRIBUTION 3
FIGURE 7 HOTEL TYPE 3
 FIGURE 8 DEPOSIT TYPE 4
FIGURE 9 CUSTOMER TYPE 4
FIGURE 10 MEAL TYPE 4
FIGURE 11 REPEATED GUEST 5
FIGURE 12 MARKET SEGMENT 5
 FIGURE 13 BOOKING CHANGES 5
 FIGURE 14 COUNTRY 6
 FIGURE 15 TOTAL GUEST 6
 FIGURE 16 TOTAL STAYS 7
 FIGURE 17 LEAD TIME 7
FIGURE 18 ASSIGNED ROOM TYPE 8
FIGURE 19 REVERSED ROOM TYPE 9
FIGURE 20 PREVIOUS CANCELLATION VS REPEATED GUEST 9
 FIGURE 21 DEPOSIT TYPE VS MARKET SEGMENT 9
 FIGURE 22 HOTEL TYPE VS CANCELLATION 10
 FIGURE 23 TOTAL GUEST VS CANCELLATION 10
 FIGURE 24 MEAL TYPE VS CANCELLATION 10
FIGURE 25 LOCATION VS CANCELLATION 11
FIGURE 26 MARKET SEGMENT AND CANCELLATION 11
FIGURE 27 DEPOSIT TYPE AND CANCELLATION 11
FIGURE 28 NON-REFUND VS GROUP BOOKINGS 12
 FIGURE 29 PREVIOUS CANCELLATION VS CURRENT CANCELLATION 12
 FIGURE 30 LEAD TIME VS CANCELLATION 12
 FIGURE 31 LEAD TIME AND CANCELLATION 13
 FIGURE 32 TOTAL NUMBER OF PARKING SPACE REQUEST AND CANCELLATION 13
 FIGURE 33 HEAT MAP OF VARIABLES 14
FIGURE 34 OUTLIERS 15
 FIGURE 35 OUTLIERS TREATMENT 15
 FIGURE 36 CORRELATION ON TARGET VARIABLE 18
 FIGURE 37 CORRELATION ON PREDATORS 19
 FIGURE 38 LOGISTIC REGRESSION - BASE MODEL 21
FIGURE 39 - LOGISTIC REGRESSION - TUNED MODEL 21
 FIGURE 40 LDA - BASE MODEL 22
FIGURE 41 - LDA - TUNED MODEL 22
 FIGURE 42 RANDOM FOREST - BASE MODEL 23
FIGURE 43 - RANDOM FOREST - TUNED MODEL 23
 FIGURE 44 GAUSSIAN NB - BASE MODEL 24
FIGURE 45 - GAUSSIAN NB - TUNED MODEL 24
 FIGURE 46 KNN - BASE MODEL 25
FIGURE 47 - KNN - TUNED MODEL 25

FIGURE 48 ANN - BASE MODEL	26
FIGURE 49 - ANN - TUNED MODEL	26
FIGURE 50 XG-BOOST - BASE MODEL	27
FIGURE 51 XG-BOOST - TUNED MODEL	27
FIGURE 52 ADA BOOST	28
FIGURE 53 GRADIENT BOOSTING	28
FIGURE 54 BAGGING CLASSIFIER	29
FIGURE 55 MODEL VALIDATION	30
FIGURE 56 ACCURACY VARIANCE	30
FIGURE 57 ACCURACY VS RECALL	31

Tableau Public Link for visualizations –

https://public.tableau.com/app/profile/sivaramakrishnan3623/viz/Sivaramakrishnan_Capstone_Hotel_Cancellation_EDA_1/SpecialRequestVSTarget

https://public.tableau.com/app/profile/sivaramakrishnan3623/viz/ProjectNotes2_16271145779440/Test?publish=yes

SOLUTION 1: INTRODUCTION

Question: Introduction - What did you wish to achieve while doing the project ?

ISSUES IN HOTEL INDUSTRY

The hotel industry has been transformed with a majority of bookings now made through Online Travel Agencies (OTA). These OTAs have transformed the cancellation policies from a footnote at the bottom of the page to the main selling point in their marketing campaigns ([source](#)). This results, the customers have become accustomed to free cancellation policies. Based on the [Fornova](#) research conducted on Dec 2021 on the 200K hotels for the cancellation rate across the industry the free cancellation policy hit 38% and 62% of no-refund policy on hotels where before the pandemic situation(COVID-19). The same survey ran on July 2020 the results are dramatically different as 58% of hotels now offering the free cancellation and 42% hotel are still refusing to offer the refund ([source](#))



CANCELLATION RATE BY RESERVATION VALUE						
Percentage of on-the-books revenue cancelled before arrival in Europe						
	2014	2015	2016	2017	2018	Change
Booking Group	43.4%	43.8%	48.2%	50.9%	49.8%	6.4
Expedia Group	20.0%	25.0%	25.8%	24.7%	26.1%	6.1
Hotelbeds Group	33.2%	37.8%	40.3%	38.3%	37.6%	4.4
HRS Group	58.5%	51.7%	55.2%	59.4%	66.0%	7.5
Other OTAs	13.7%	15.2%	27.0%	24.4%	24.3%	10.6
Other Wholesalers	31.2%	30.3%	34.6%	33.8%	32.8%	1.6
Website Direct	15.4%	17.7%	18.0%	18.4%	18.2%	2.8
AVERAGE	32.5%	34.8%	39.6%	41.3%	39.6%	7.1
Yearly average percentage of on-the-books revenue cancelled prior to guest arrival from a sample of 680 D-EDGE clients in Europe.						
D-EDGE, Hospitality Solutions			www.d-edge.com			

The D-Edge Hospitality survey proves that the cancellation rate over 5 years average change as 7.1 and average cancellation rate decreased from 41.3% to 39.6% ([source](#))

FIGURE 1 CANCELLATION RATE

NEED OF THE STUDY/PROJECT

When hotels try to protect themselves by using services from OTA's "Risk Free Reservations", the burden then falls on OTAs. Indeed, this service requires the OTA to pay for the reservation if the booking is cancelled and they cannot find a new guest to occupy the room ([source](#)). One thing is clear, whether you are a hotel or an OTA, cancellations have an negative financial impact on your business.

In addition to the direct financial consequences of cancellations, they also cause operational problems (such as over or understaffing). Those problems may lead to decrease customer satisfaction and negative reviews. In a world where more and more customers check online reviews before picking a hotel, those reviews can have major impacts. Indeed, TripAdvisor’s reviews and scores influenced around \$546 billions of travel spending during 2017 ([source](#)).

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

Artificial intelligence is playing an increasingly important role in hospitality management ([source](#)), primarily because of its ability to carry out traditionally human functions at any time of the day. This potentially means that hotel owners can save significant money, eliminate human error and deliver superior service.

An example of this has been seen with the Dorchester Collection hotel chain, which has made use of the Metis AI platform. By using this technology, the company has been able to sort through data collected via surveys, online reviews etc. and the AI has been able to then analyse this to draw conclusions about overall performance.

SOLUTION 2: EDA AND BUSINESS IMPLICATION

Question: EDA - Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. - Both visual and non-visual understanding of the data.

VISUAL INSPECTION OF DATA (ROWS, COLUMNS, DESCRIPTIVE DETAILS)

Booking data from both types hotels share the same structure, with 25 variables describing the 40,060 observations of type-1 and 79,330 observations of type-2 (total – **119390**). For a detailed list and description of those variables refer to the data dictionary.

```
****Shape of the hotel data****
*****
No of rows      :      119390
No of columns   :         33
```

FIGURE 2 ROWS AND COLUMNS

DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

Describe the data – Continues Variables								

	count	mean	std	min	25%	50%	75%	max
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.0	0.0	1.0	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.0	1.0	2.0	3.0	50.0
adults	119390.0	1.856403	0.579261	0.0	2.0	2.0	2.0	55.0
children	119390.0	0.111835	0.412561	0.0	0.0	0.0	0.0	10.0
is_repeated_customer	119390.0	0.031912	0.175767	0.0	0.0	0.0	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.0	0.0	0.0	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.0	0.0	0.0	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.0	0.0	0.0	0.0	21.0
agent	103050.0	86.693382	110.774548	1.0	9.0	14.0	229.0	535.0
company	6797.0	189.266735	131.655015	6.0	62.0	179.0	270.0	543.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.0	0.0	0.0	0.0	391.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.0	0.0	0.0	0.0	8.0
totalno_of_special_requests	119390.0	0.571363	0.792798	0.0	0.0	0.0	1.0	5.0

FIGURE 3 DESCRIPTIVE STATISTICS NUMERICAL

Based on the initial view the data provided are looks like a categorical we may need to convert to categorical type post detailed analysis.

DESCRIPTIVE STATISTICS OF CATEGORICAL VARIABLES

*** Describe the data – Catagorical Variables ***				

	count	unique	top	freq
hotel	119390	2	type_1	79330
booking_date	119390	985	17-10-2017	2511
arrival_date	119390	792	05-12-2018	448
meal	119390	4	Only Breakfast	92310
country	118902	177	PRT	48590
market_segment	119390	8	Online Travel Agents	56477
distribution_channel	119390	5	TA/T0	97870
reserved_room_type	119390	10	A	85994
assigned_room_type	119390	12	A	74053
deposit_type	119390	3	No Deposit	104641
customer_type	119390	4	Transient	89613

There are higher frequency for the No-Deposit which may lead to the highest cancellation since there is no loss for the customer due to the cancellation.

FIGURE 4 DESCRIPTIVE STATISTICS CATEGORICAL

In the categorical variables like Agent or Company, “NULL” is presented as one of the categories. This should not be considered a missing value, but rather as “not applicable”. For example, if a booking “Agent” is defined as “NULL” it means that the booking did not came from a travel agent." As a result, "NULL" values for agent and company will be changed to No Agent and No Company for clarity purposes.

UNDERSTANDING OF ATTRIBUTES (VARIABLE INFO, RENAMING IF REQUIRED)

VARIABLE INFO

#	Column	Non-Null Count	Dtype
0	hotel	119390 non-null	object
1	is_canceled	119390 non-null	int64
2	booking_date	119390 non-null	object
3	arrival_date	119390 non-null	object
4	stays_in_weekend_nights	119390 non-null	int64
5	stays_in_week_nights	119390 non-null	int64
6	adults	119390 non-null	int64
7	children	119390 non-null	int64
8	meal	119390 non-null	object
9	country	118902 non-null	object
10	market_segment	119390 non-null	object
11	distribution_channel	119390 non-null	object
12	is_repeated_customer	119390 non-null	int64
13	previous_cancellations	119390 non-null	int64
14	previous_bookings_not_canceled	119390 non-null	int64
15	reserved_room_type	119390 non-null	object
16	assigned_room_type	119390 non-null	object
17	booking_changes	119390 non-null	int64
18	deposit_type	119390 non-null	object
19	agent	103050 non-null	float64
20	company	6797 non-null	float64
21	days_in_waiting_list	119390 non-null	int64
22	customer_type	119390 non-null	object
23	required_car_parking_spaces	119390 non-null	int64
24	totalno_of_special_requests	119390 non-null	int64

dtypes: float64(2), int64(12), object(11)

- ⇒ There are three features have missing values.
- ⇒ Most of the features are in categorical/int64 type of variables.
- ⇒ There are 2 date type features are considered as object which may need a need to convert to date for further analysis
- ⇒ Looks like the variable names are more relevant and not required any renaming.

FIGURE 5 DATA INFO

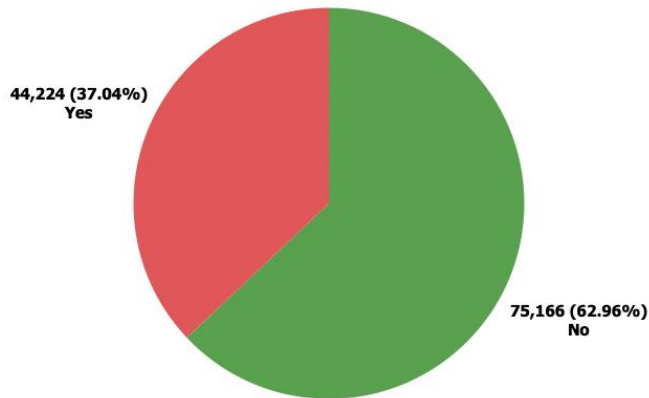
DUPLICATE VALUE CHECK.

There are 33210 duplicate values in the dataset which is nearly 27.82% of total data this we can drop before the model building.

UNIVARIANT ANALYSIS

TARGET VARIABLE

Let’s begin with the Target variable distribution on the Hotel Cancellation status. This data represented in the feature is_canceled.



- ✓ The rate of cancellation is likely matching with the industry standard which is around 37% - 40% Source: [Emerchantpay Link](#)
- ✓ The problem that hospitality industries are facing that there are almost 4 cancellation in every 10 bookings
- ✓ The target data is almost balance, so later on for the machine learning process we won’t need to do an imbalance handling

FIGURE 6 TARGET DISTRIBUTION

HOTEL TYPE

The data collected from the two different types of hotels the hotel types are named as “type_1” and “type_2”

- ✓ There are more booking from the type1 hotel booking compared to the type2 hotel booking in this case we will see it later on how this affect cancellation

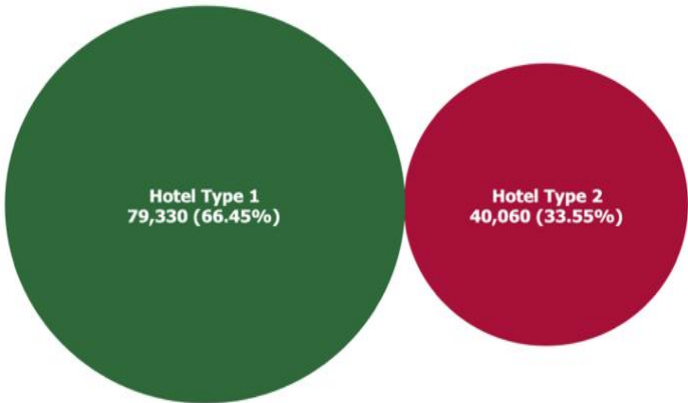
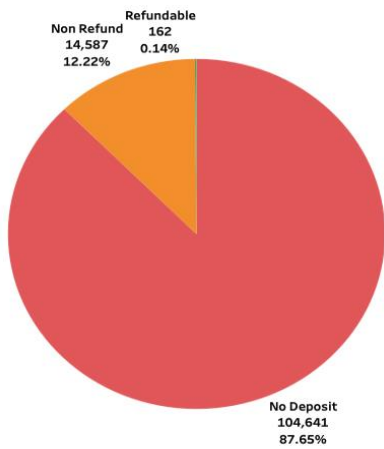


FIGURE 7 HOTEL TYPE

DEPOSIT TYPE



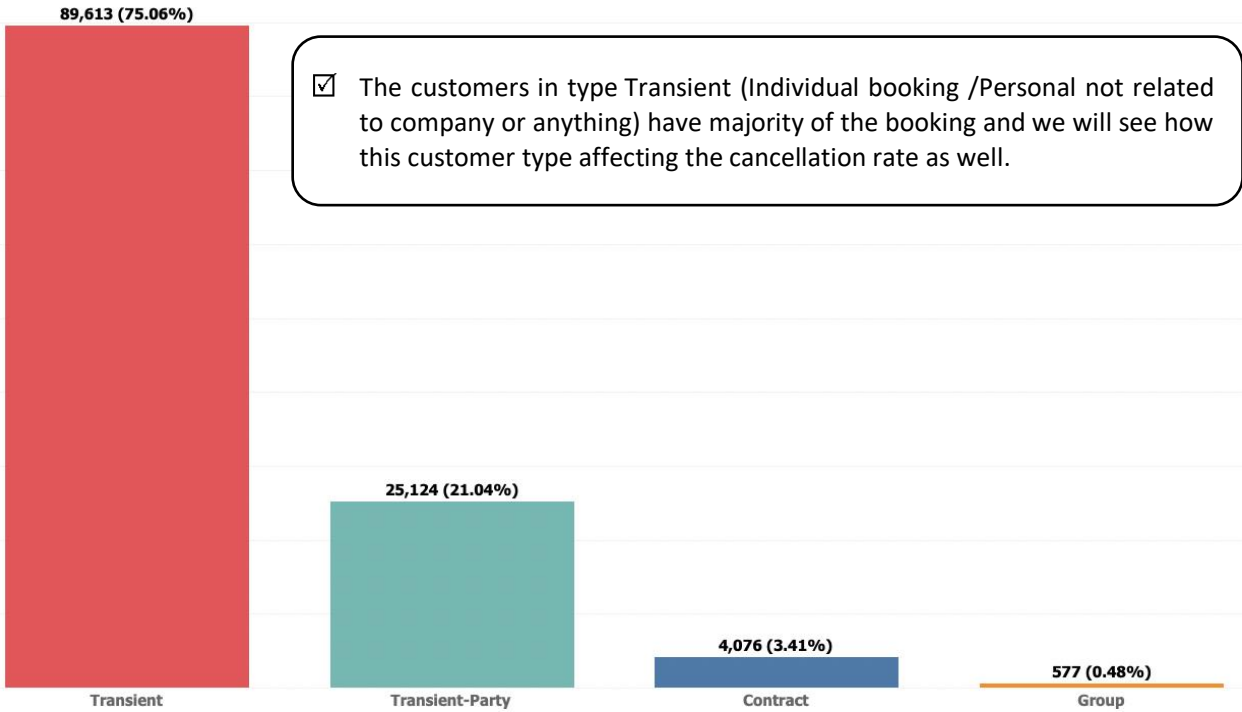
Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:

Value calculated based on the payments identified for the booking in the transaction (TR) table before the bookings arrival or cancellation date. In case no payments were found the value is “No Deposit”. If the payment was equal or exceeded the total cost of stay, the value is set as “Non Refund”. Otherwise the value is set as “Refundable”.

☑ The No Deposit type is the highest count this deposit type may impact highly affect the cancellation rate in both type of hotels

FIGURE 8 DEPOSIT TYPE

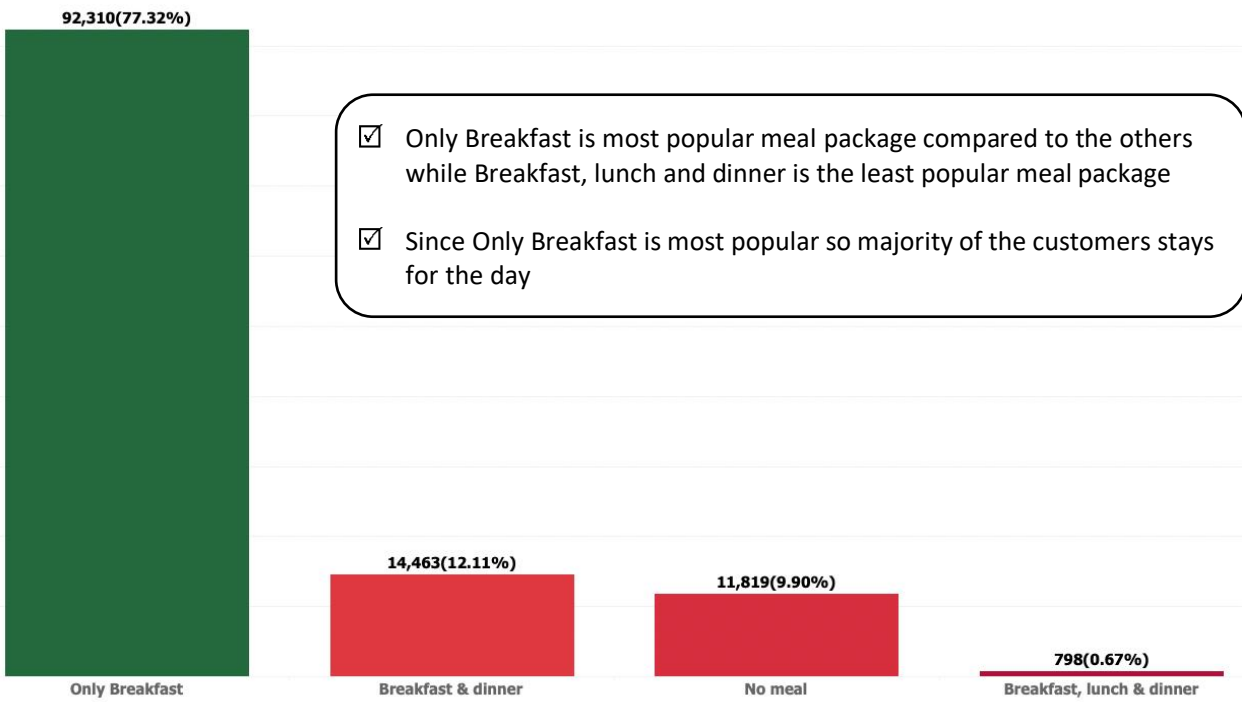
CUSTOMER TYPE



☑ The customers in type Transient (Individual booking /Personal not related to company or anything) have majority of the booking and we will see how this customer type affecting the cancellation rate as well.

FIGURE 9 CUSTOMER TYPE

MEAL TYPE

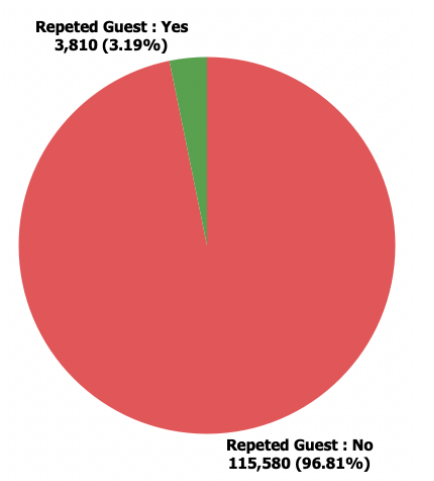


☑ Only Breakfast is most popular meal package compared to the others while Breakfast, lunch and dinner is the least popular meal package

☑ Since Only Breakfast is most popular so majority of the customers stays for the day

FIGURE 10 MEAL TYPE

REPEATED GUEST



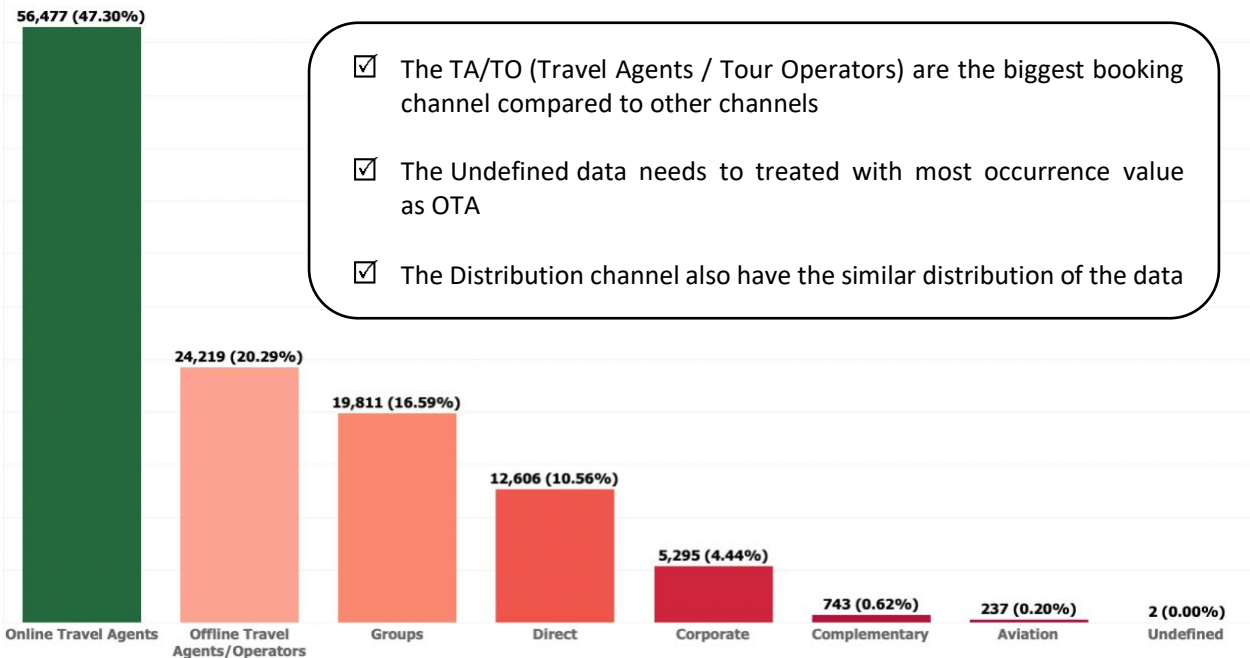
Value indicating if the booking name was from a repeated guest (1) or not (0). Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the Property Management System database it was assumed the booking was from a repeated guest.

- ☑ The Repeated customers are only 3.19% so there are very low rate of loyal customers
- ☑ The loyal customers are the most profitable then new customers, Below are some reasons
 - ⇒ They are familiar with hotels offered services
 - ⇒ Loyal guests usually spend more money at your hotel
 - ⇒ The stay period for loyal guests is usually longer than that of new guests

FIGURE 11 REPEATED GUEST

MARKET SEGMENT

The hotel market segmentation shall help to identify the purpose of the trip: either business or leisure. The price does not decide the market segmentation. A clear distinction must also be achieved between individual and group business.

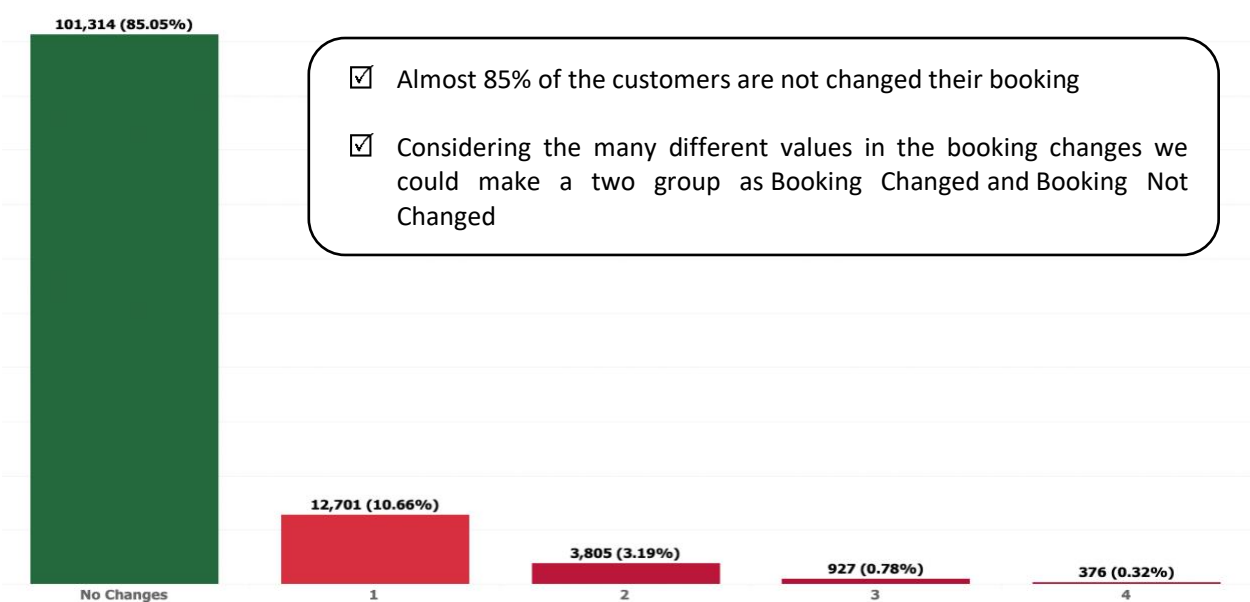


- ☑ The TA/TO (Travel Agents / Tour Operators) are the biggest booking channel compared to other channels
- ☑ The Undefined data needs to treated with most occurrence value as OTA
- ☑ The Distribution channel also have the similar distribution of the data

FIGURE 12 MARKET SEGMENT

BOOKING CHANGES

Number of changes / amendments made to the booking from the moment the booking was entered on the Property Management System until the moment of check-in or cancellation. Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal.



- ☑ Almost 85% of the customers are not changed their booking
- ☑ Considering the many different values in the booking changes we could make a two group as Booking Changed and Booking Not Changed

FIGURE 13 BOOKING CHANGES

LOCATION

Country of origin. Categories are represented in the International Standards Organization (ISO) 3155–3:2013 format.

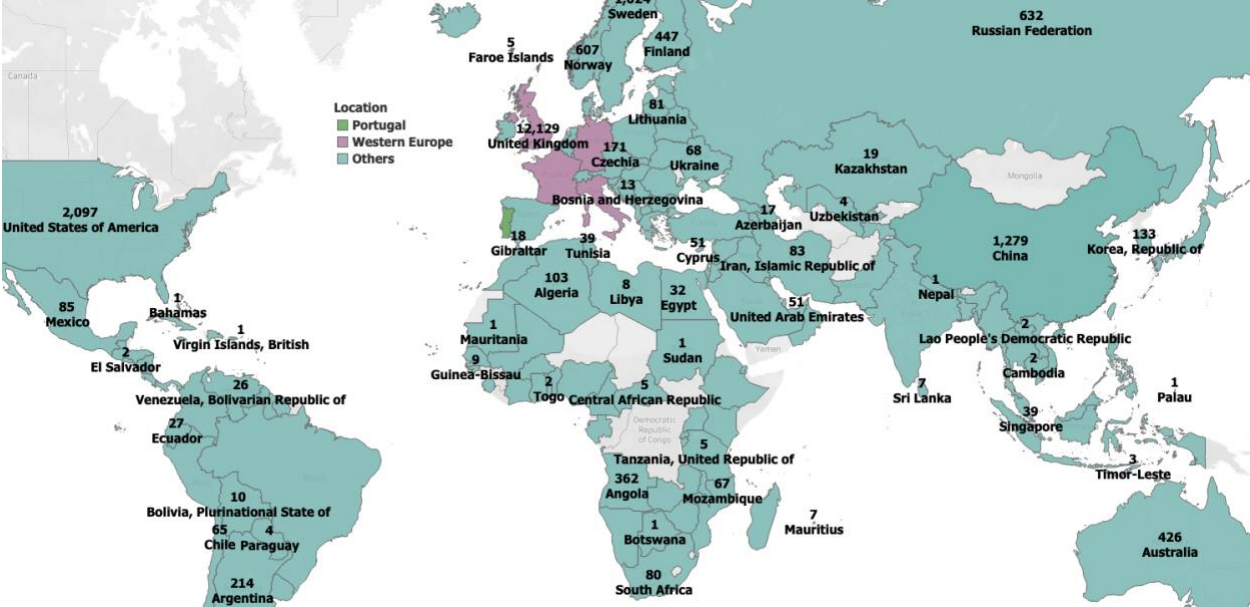


FIGURE 14 COUNTRY

Country	Count	Percent
Portugal	48590	40.699
United Kingdom	12129	10.159
France	10415	8.724
Spain	8568	7.176
Germany	7287	6.104
Italy	3766	3.154
Ireland	3375	2.827
Belgium	2342	1.962

- ✓ Almost half of the booking is made from Portugal
- ✓ The new variable created by splitting the bookings into 3 groups Portugal, Weston Europe (UK, France, Spain, Germany and Italy) and Others.

TOTAL GUEST

We are creating the variable as total_guest by combining the value of adults and children in the booking.

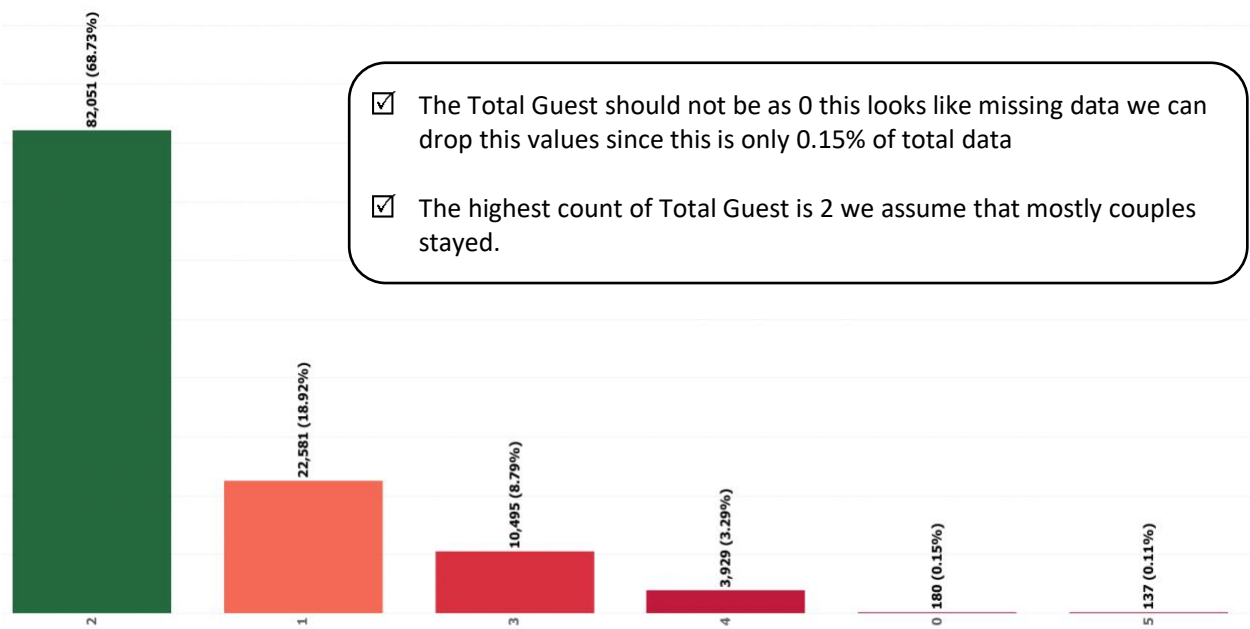


FIGURE 15 TOTAL GUEST

- ✓ The Total Guest should not be as 0 this looks like missing data we can drop this values since this is only 0.15% of total data
- ✓ The highest count of Total Guest is 2 we assume that mostly couples stayed.

TOTAL STAYS

The Total stays calculated based on the number of Stays in Weekend Night and number of Stays in Weekday Nights.

Stays in Weekend Night + Stays in Weekday Nights.

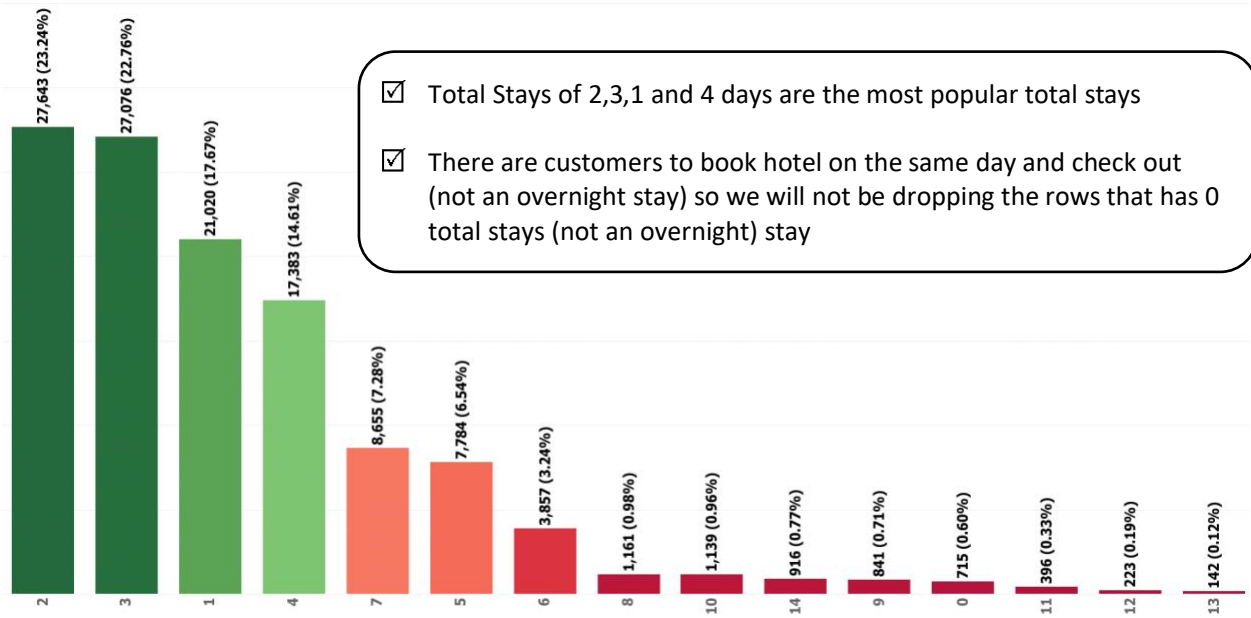


FIGURE 16 TOTAL STAYS

LEAD TIME – & LEAD TIME – MONTH

Number of days that elapsed between the entering date of the booking into the Property Management System and the arrival date. Calculated by subtracting the entering date from the arrival date.

- ⇒ Lead Time Days = (Date of Arrival - Date of Booking)
- ⇒ There are many unique values in the lead time and we can group it as months to get more insights on the trend
- ⇒ Lead Time Months = ((lead-time days) // 30) (// Returns Rounding off value)

The table displayed the top 5 values of Lead time of Month

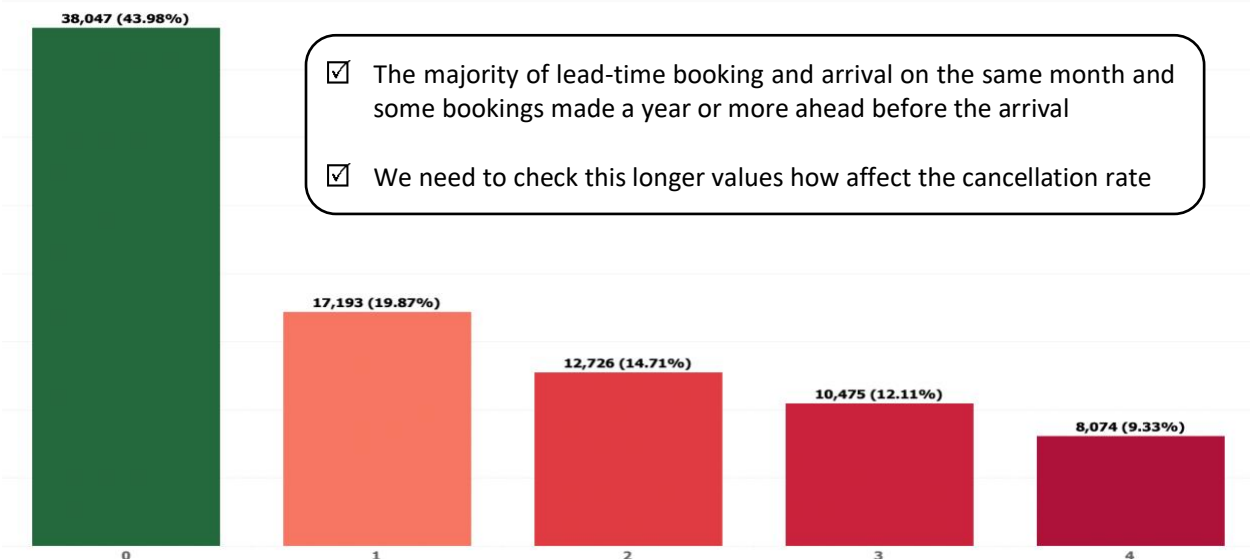


FIGURE 17 LEAD TIME

Waiting List Days

Number of days the booking was in the waiting list before it was confirmed to the customer. Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the Property Management System.

The table has a top 3 of more than 150 days of waiting list.

- ✓ Almost 97% customers are got the rooms without any waiting list
- ✓ The cause of the waiting list could be the reason the customer booking at the wrong time(Last minute travel) during days in high occupancy (Important Country festivals)
- ✓ This can be avoided book the rooms 40 days before Source: [USA Today](#)

Days In Waiting List	Count	Percent
0	115692	96.903
39	227	0.19
58	164	0.137

Previous Bookings Cancelled

Number of previous bookings that were cancelled by the customer prior to the current booking. In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and cancelled.

No Previous Cancellation	Count	Percent
No	115770	96.968
Yes	3620	3.032

- ✔ Almost 95% of the booking never been cancelled before in this data set
- ✔ We will group this into booking that's never been cancelled or have been cancelled before

Required Car Parking

No of Car Parking space requested by customer during the booking of the Room.

- ✔ Over 94% customer not requested for the car parking
- ✔ There is 6% of customer required 1 car parking
- ✔ We need to do further analysis on the effect of the cancellation rate

# Of Car Parking Required	Count	Percent
0	111974	93.788
1	7383	6.184
2	28	0.023
3	3	0.003
8	2	0.002

Required Special Request

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

# Of Special Request Required	Count	Percent
0	70318	58.898
1	33226	27.83
2	12969	10.863
3	2497	2.091
4	340	0.285
5	40	0.034

- ✔ Almost 59% of customers not requested for special request
- ✔ Over 28% customers are requested for 1 special request and 11% customers are requested for 2 special request

ASSIGNED ROOM TYPE

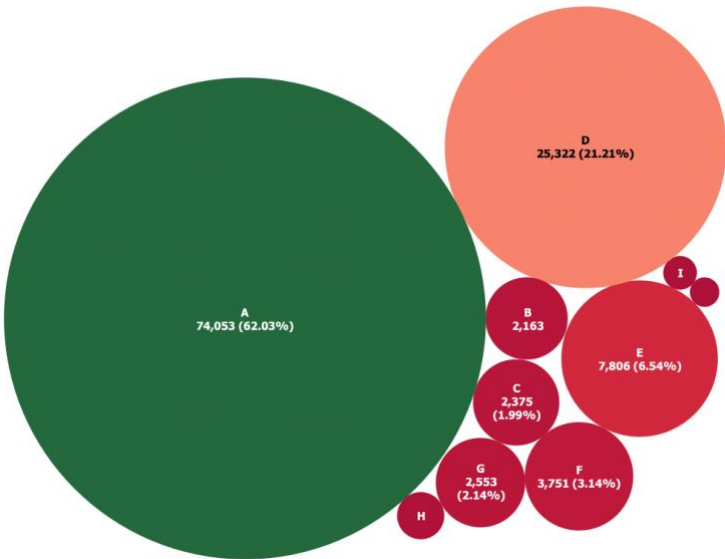


FIGURE 18 ASSIGNED ROOM TYPE

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.

- ✔ There are moderate difference from Reversed Room Type to Assigned Room Type
- ✔ The Room types(I & K) are not booked by customers but there are assignments in the dataset, these assignments may be due to the Reversed rooms are assigned to early arrived loyal customers visit on last minute

RESERVED ROOM TYPE

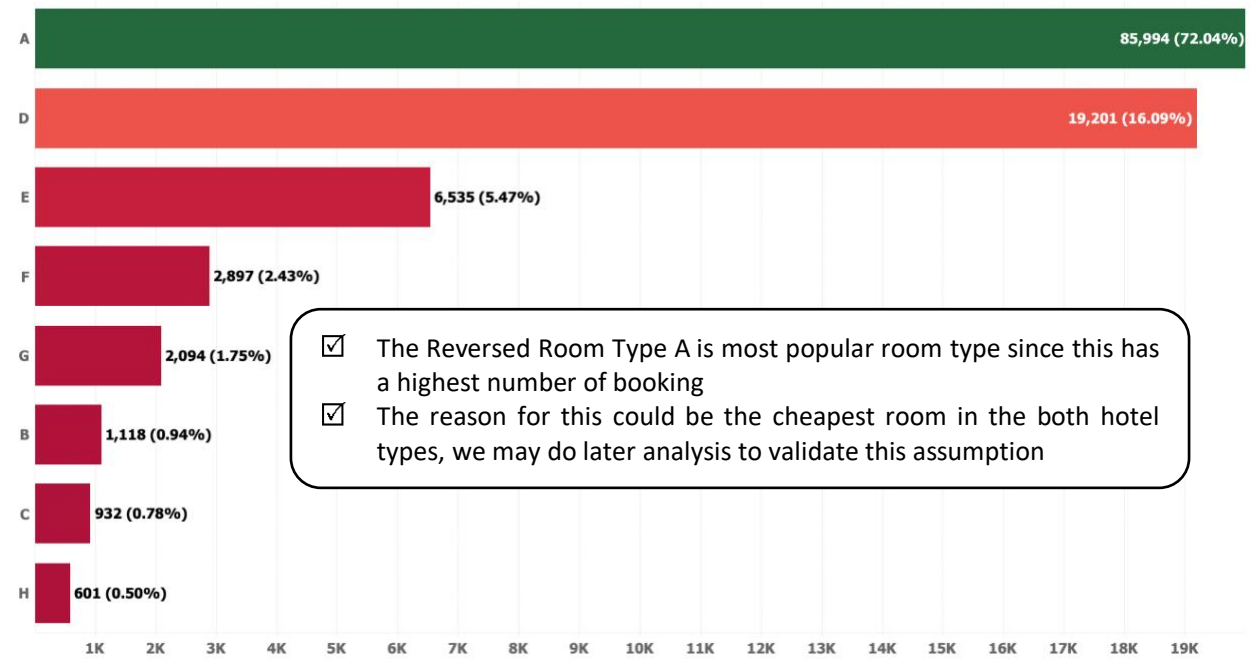
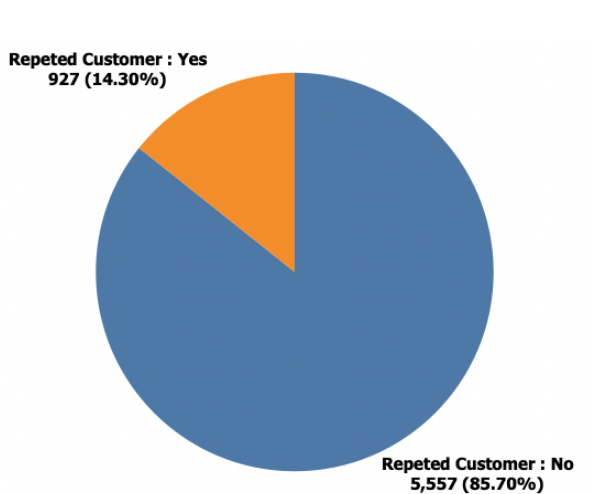


FIGURE 19 REVERSED ROOM TYPE

BIVARIANT ANALYSIS - AFTER DATA CLEANING



PREVIOUS CANCELLATION VS REPEATED GUEST

- ✓ Almost 86% of not cancelled the previous bookings.
- ✓ This indicates the low revenue loss on the repeated customers.

FIGURE 20 PREVIOUS CANCELLATION VS REPEATED GUEST

DEPOSIT TYPE VS MARKET SEGMENT

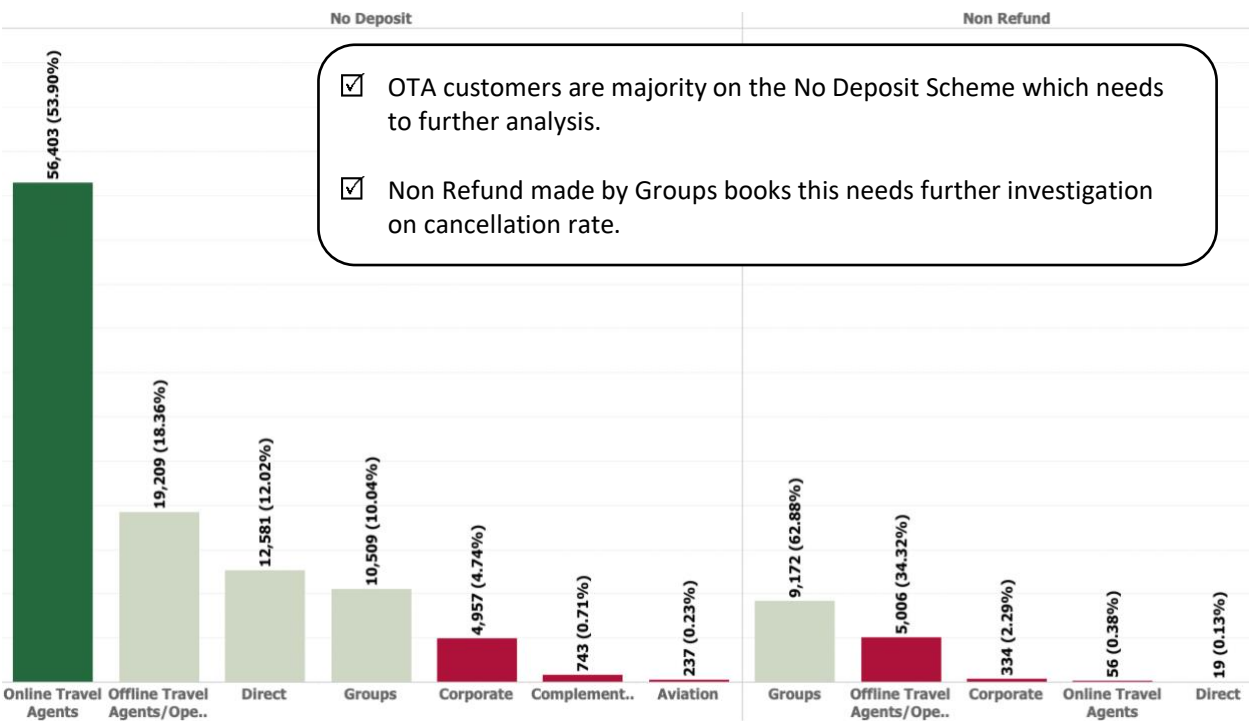


FIGURE 21 DEPOSIT TYPE VS MARKET SEGMENT

HOTEL TYPE AND CANCELLATION

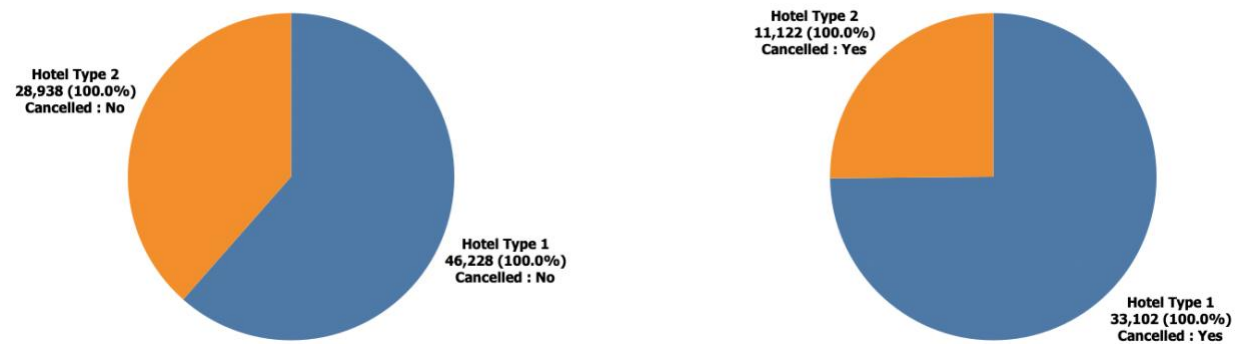
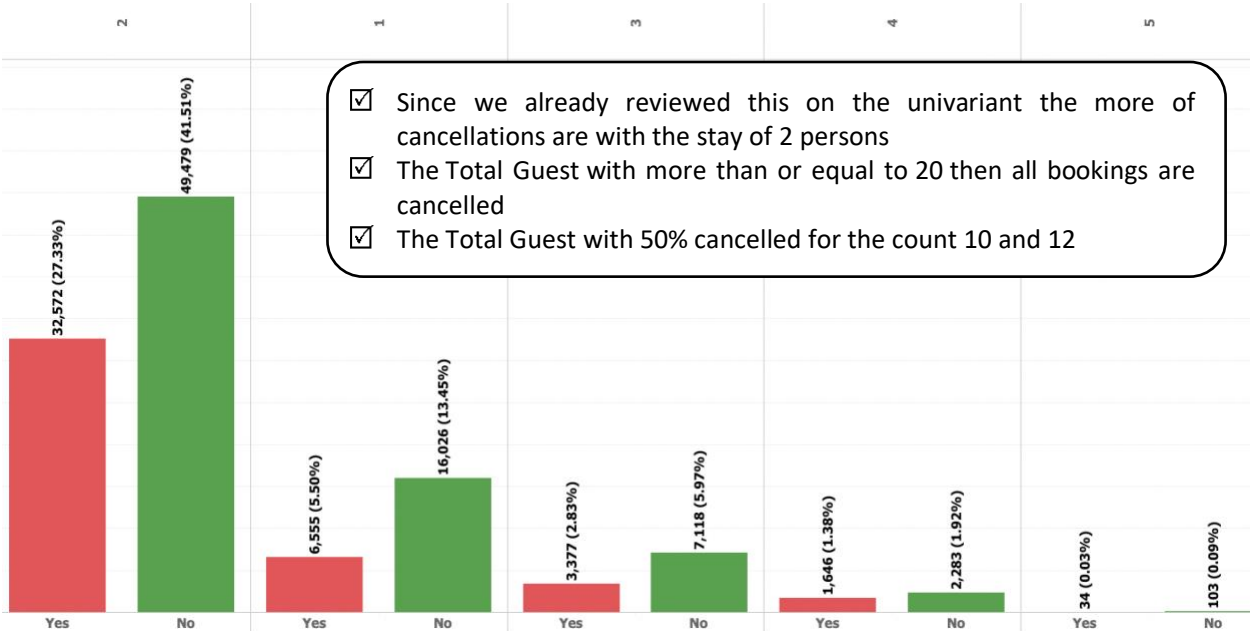


FIGURE 22 HOTEL TYPE VS CANCELLATION

- ✓ The type1 hotels has a cancelling rate then type2. Based on the dataset the type1 more records so this may be cause of this
- ✓ One assumption can be made that increase number of booking will increase number of cancellation (Positively Correlated)

TOTAL GUEST AND CANCELLATION



- ✓ Since we already reviewed this on the univariant the more of cancellations are with the stay of 2 persons
- ✓ The Total Guest with more than or equal to 20 then all bookings are cancelled
- ✓ The Total Guest with 50% cancelled for the count 10 and 12

FIGURE 23 TOTAL GUEST VS CANCELLATION

MEAL TYPE AND CANCELLATION

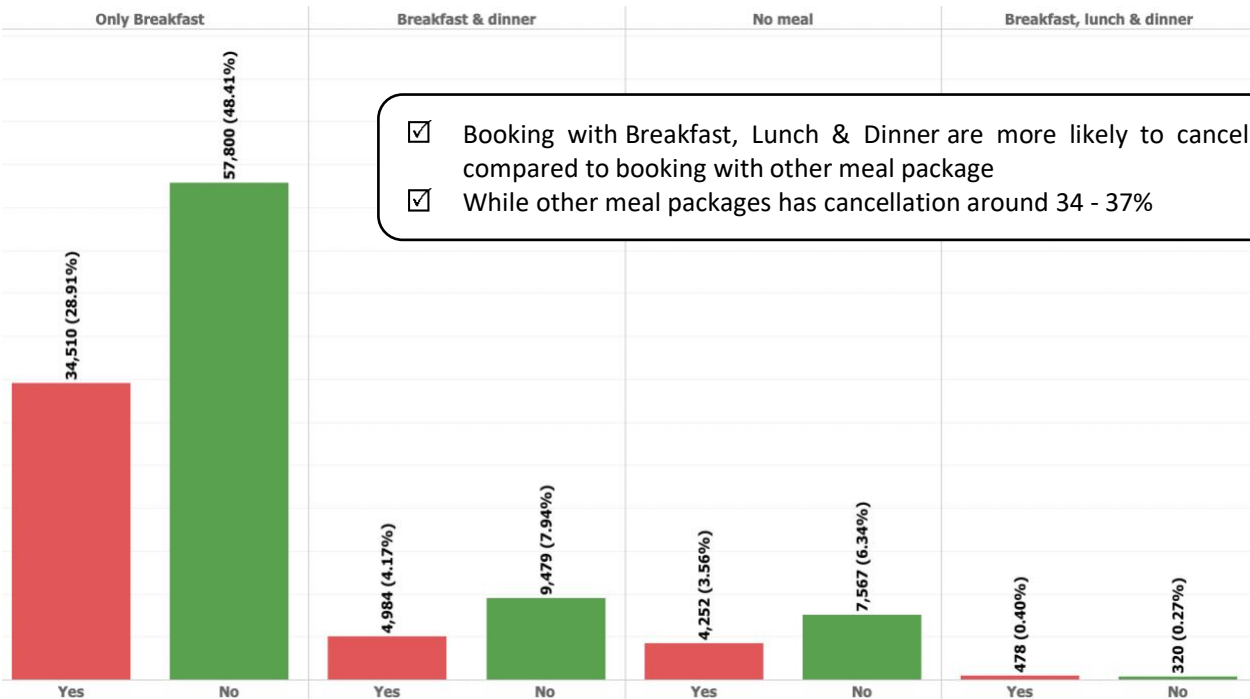


FIGURE 24 MEAL TYPE VS CANCELLATION

- ✓ Booking with Breakfast, Lunch & Dinner are more likely to cancel compared to booking with other meal package
- ✓ While other meal packages has cancellation around 34 - 37%

LOCATION AND CANCELLATION

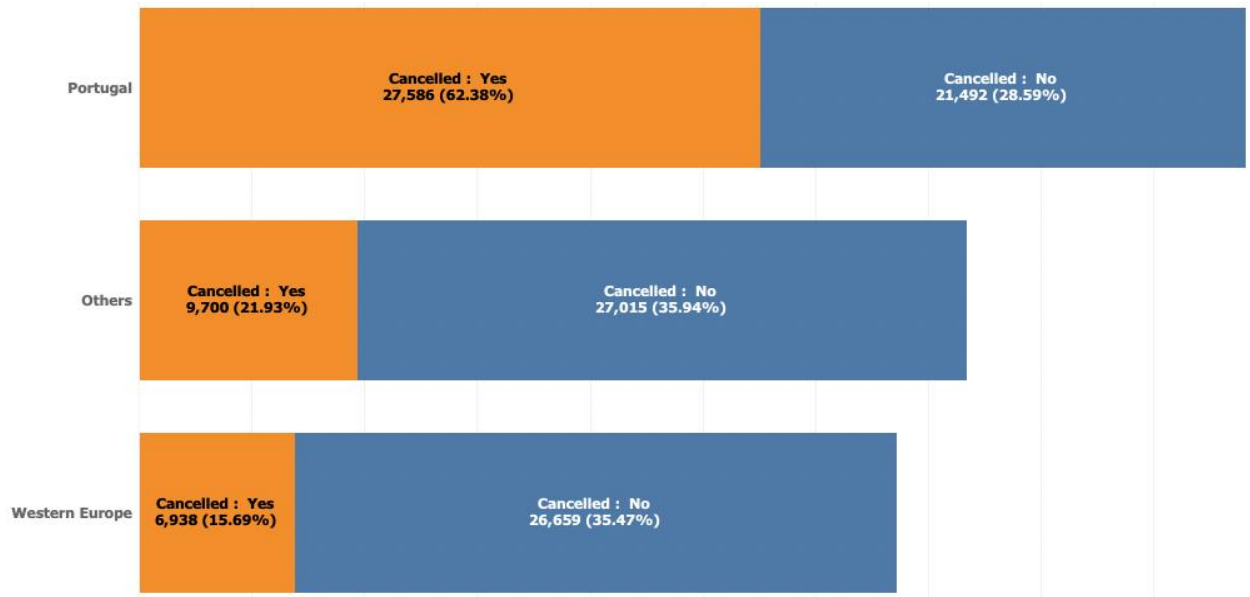
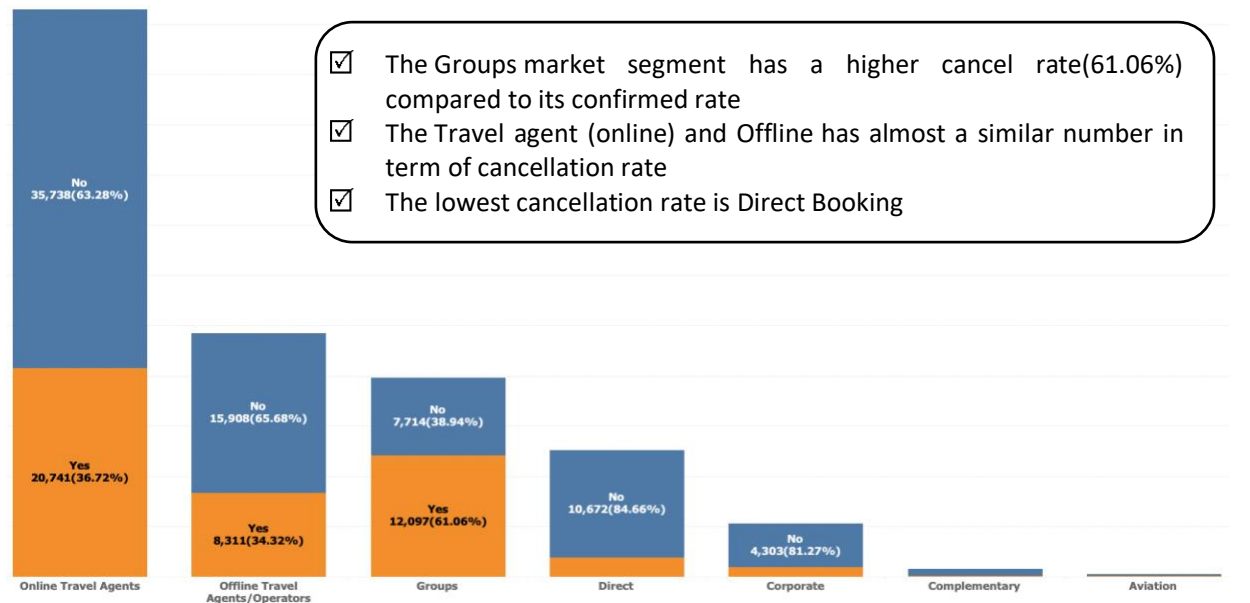


FIGURE 25 LOCATION VS CANCELLATION

- ☑ The booking made in Portugal are almost 2.5 X more likely to be cancelled compared to booking that's made outside Portugal

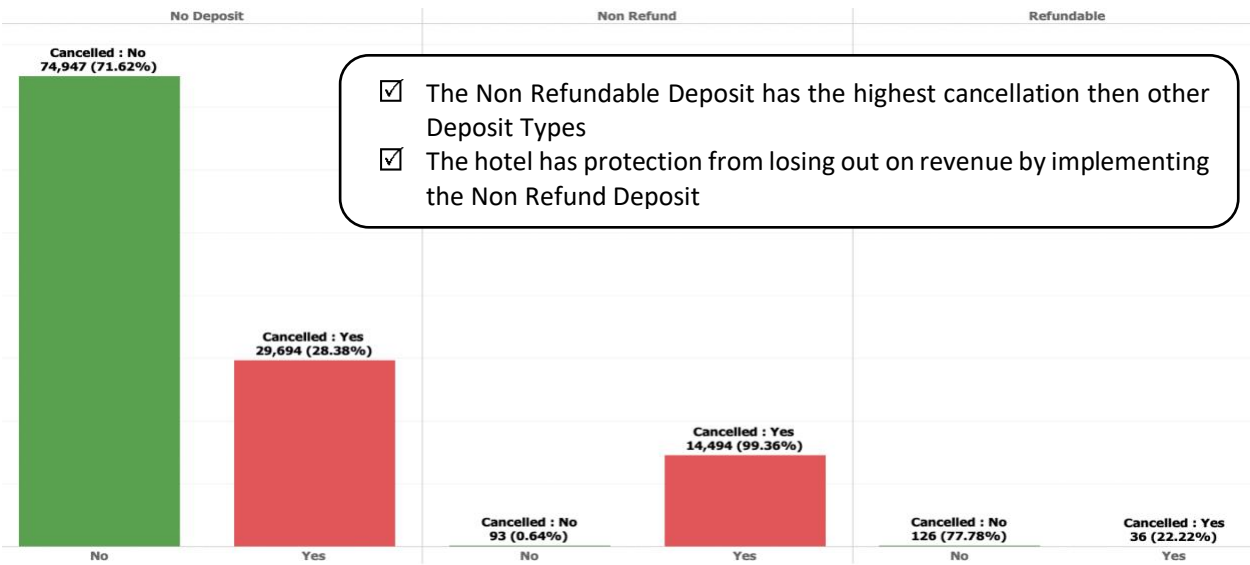
MARKET SEGMENT AND CANCELLATION



- ☑ The Groups market segment has a higher cancel rate(61.06%) compared to its confirmed rate
- ☑ The Travel agent (online) and Offline has almost a similar number in term of cancellation rate
- ☑ The lowest cancellation rate is Direct Booking

FIGURE 26 MARKET SEGMENT AND CANCELLATION

DEPOSIT TYPE AND CANCELLATION



- ☑ The Non Refundable Deposit has the highest cancellation then other Deposit Types
- ☑ The hotel has protection from losing out on revenue by implementing the Non Refund Deposit

FIGURE 27 DEPOSIT TYPE AND CANCELLATION

The Group Bookings of Non-Refund Deposit type has a high impact on the cancellation rate.

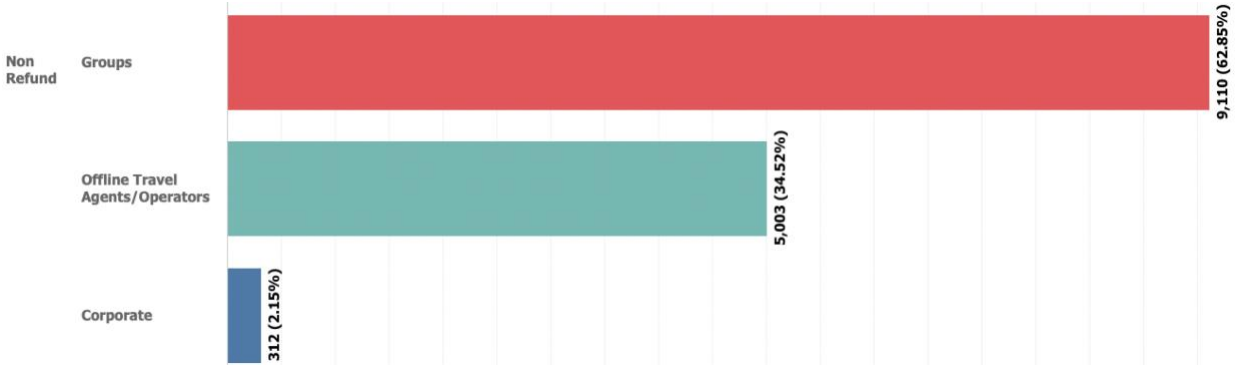


FIGURE 28 NON-REFUND VS GROUP BOOKINGS

PREVIOUS CANCELLATION AND CANCELLATION

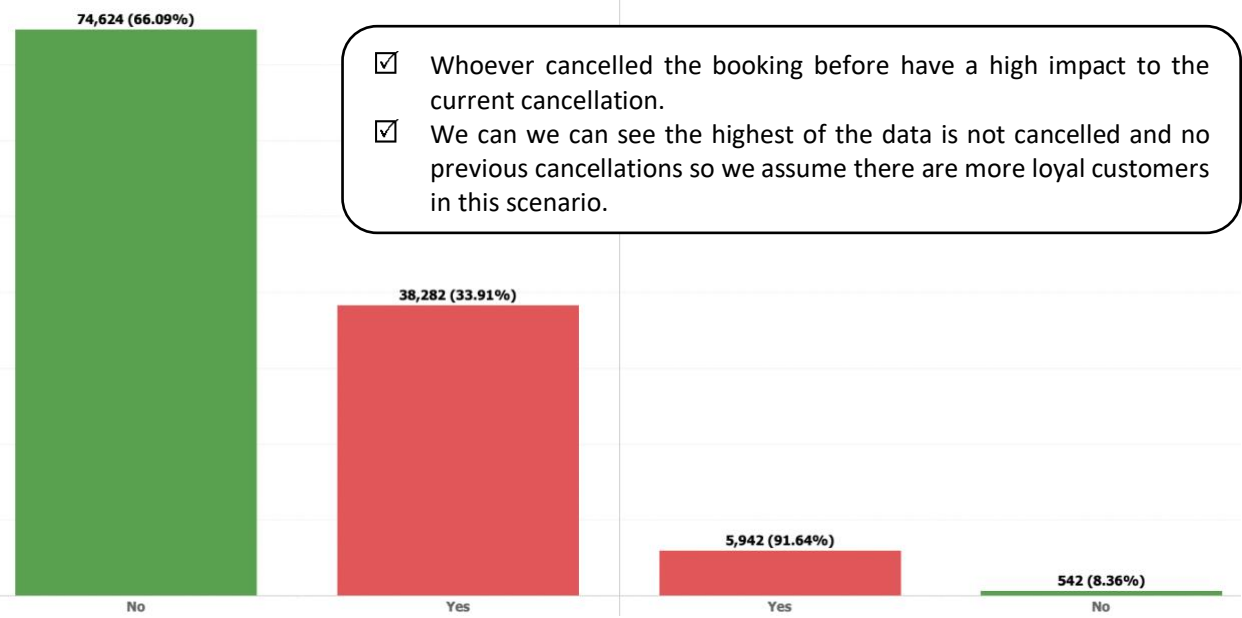


FIGURE 29 PREVIOUS CANCELLATION VS CURRENT CANCELLATION

LEAD TIME AND CANCELLATION

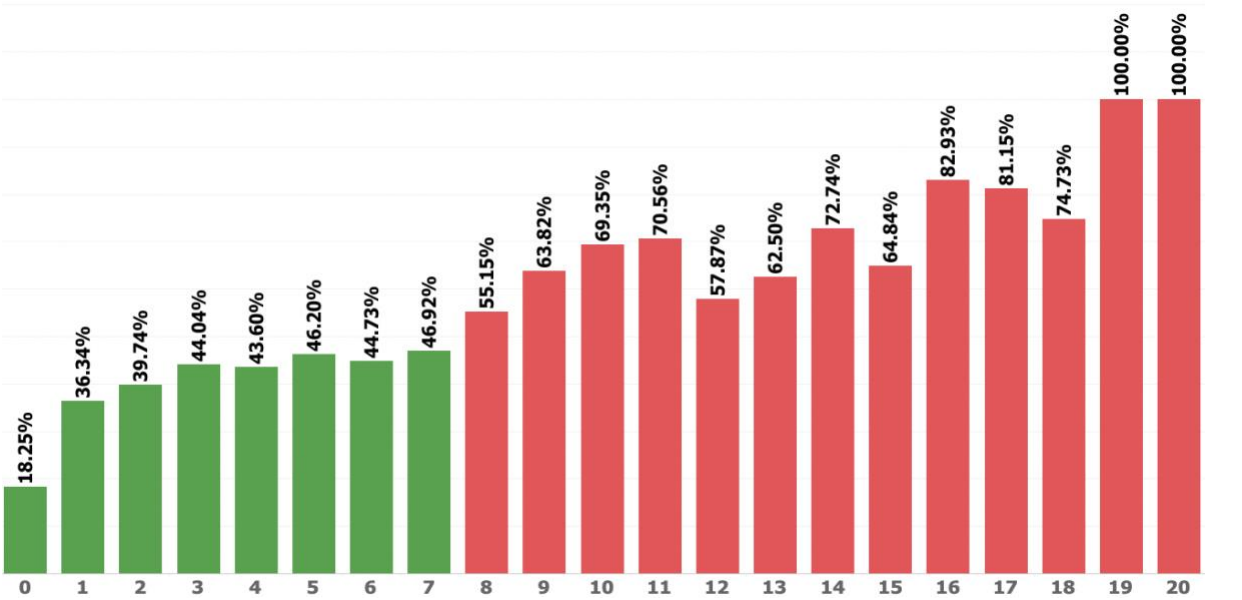


FIGURE 30 LEAD TIME VS CANCELLATION

- ✓ Cancelled bookings have a longer lead time on average.
- ✓ There are 2 bookings not Cancelled with higher lead time these bookings could be by the loyal customers
- ✓ Bookings that has more than 7 months lead time are more likely to be cancelled compared to confirmed

TOTAL NUMBER OF SPECIAL REQUEST AND CANCELLATION



FIGURE 31 LEAD TIME AND CANCELLATION

TOTAL NUMBER OF PARKING SPACE REQUEST AND CANCELLATION

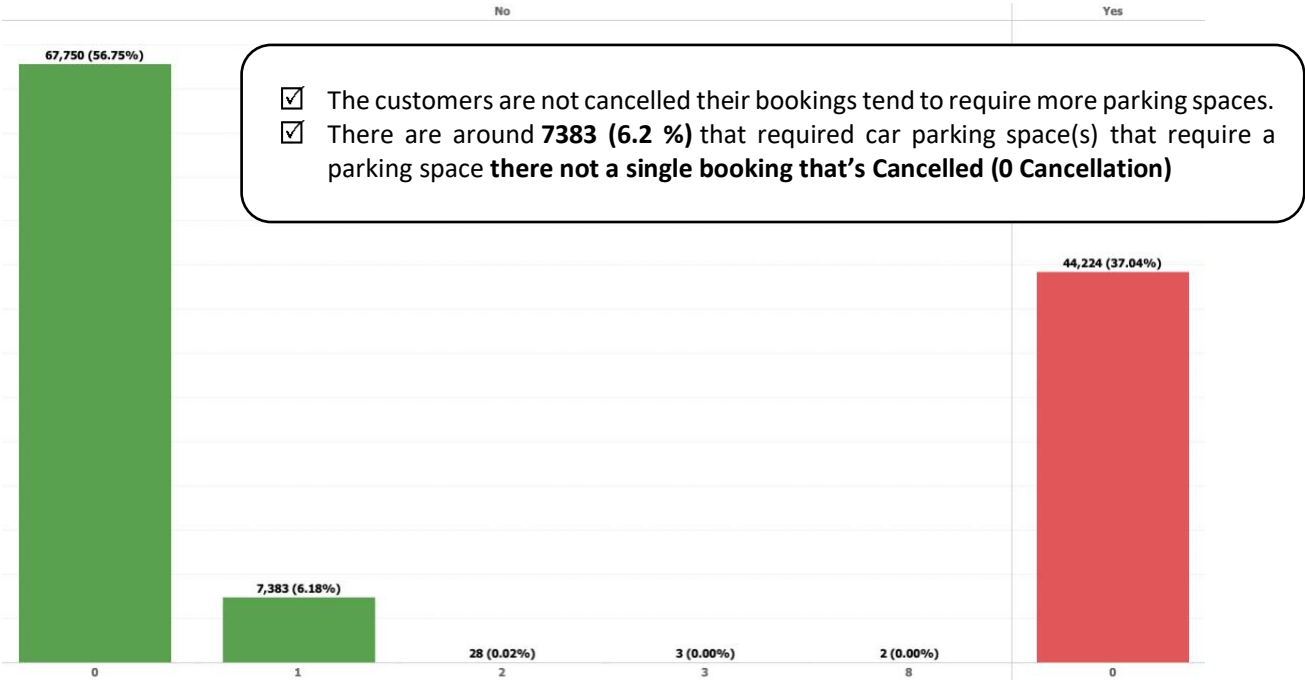


FIGURE 32 TOTAL NUMBER OF PARKING SPACE REQUEST AND CANCELLATION

MULTI-VARIANT ANALYSIS

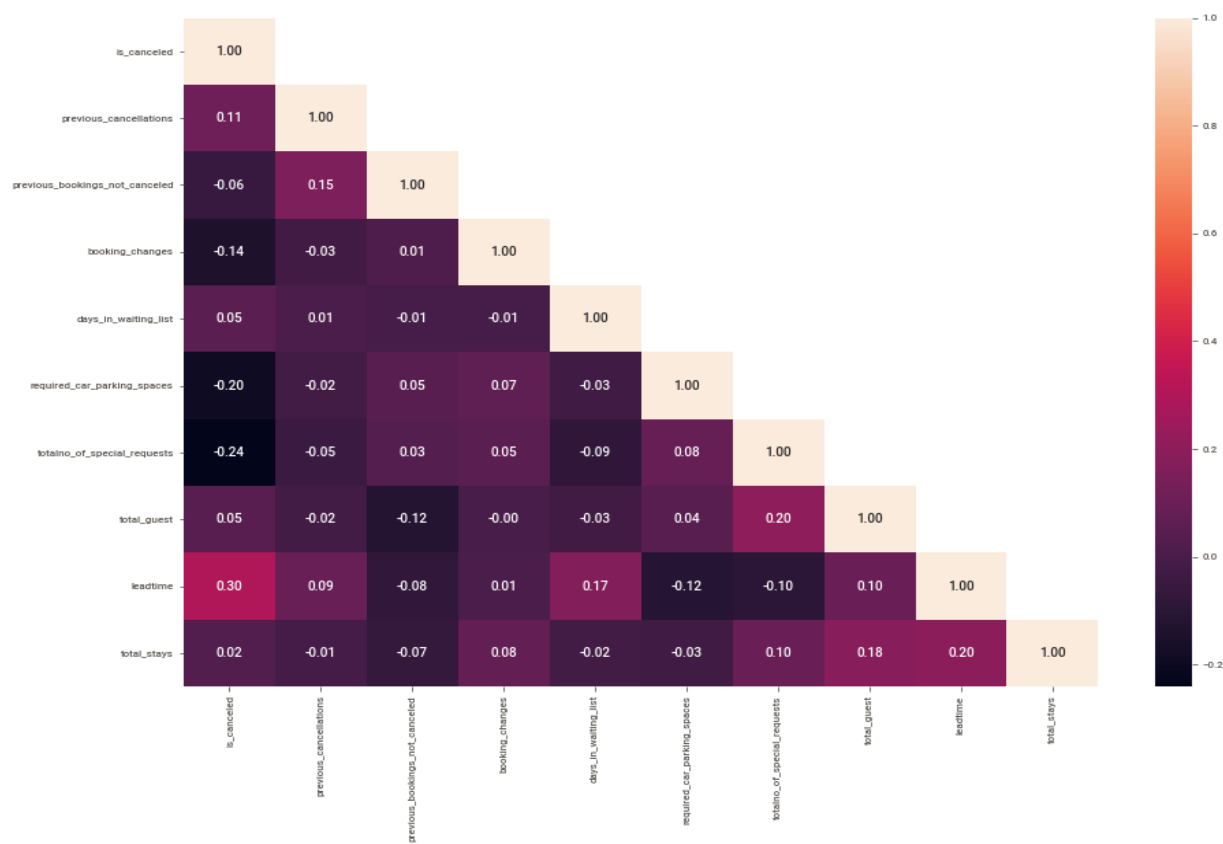


FIGURE 33 HEAT MAP OF VARIABLES

Relationship on Target Variable - Top 3

- ✓ Lead time is the most highly correlated(0.30) feature with whether or not a booking is_canceled. It makes sense that as the number of days between when the booking is made and the supposed arrival date increases, customers have more time to cancel the reservation and there is more time for an unforeseen circumstance derailing travel plans to arise.
- ✓ The total number of special requests is the second highest feature with the strongest correlation (-0.24) to our is_canceled target. As the number of special requests made increases, the likelihood that a booking is cancelled decreases. This suggests that engagement with the hotel prior to arrival and feeling like their needs are heard may make a customer less likely to cancel their reservation.
- ✓ The number of required car parking spaces is the third highest feature with the strongest correlation of (-0.20) to the is_canceled target. As the number of parking spaces requests increases, the likelihood that a booking is cancelled decreases. There is a potential reasons for this relationship are discussed later on.

Relationship Between Predictors

- ✓ There is a moderate correlation(0.17) between days_in_waiting_list and the lead_time. Since both are related to no of days so can this could be the moderate correlation. We need further investigation on the multicollinearity and decide on the feature selection.
- ✓ We also see more features have moderate correlation with lead_time
 - lead_time VS total_stays - 0.16
 - lead_time VS required_car_parking - 0.12
 - lead_time VS required_car_parking - 0.10

SOLUTION 3: DATA CLEANING

Question : Data Cleaning and Pre-processing - Approach used for identifying and treating missing values and outlier treatment (and why) - Need for variable transformation (if any) - Variables removed or added and why (if any)

MISSING VALUES TREATMENT

In the categorical variables like Agent or Company, “NULL” is presented as one of the categories. This should not be considered a missing value, but rather as “not applicable”. For example, if a booking “Agent” is defined as “NULL” it means that the booking did not came from a travel agent." As a result, "NULL" values for agent and company will be changed to No Agent and No Company for clarity purposes.

The missing value in the country can be updates as the maximum occurrence value since there is very minimal missing values.

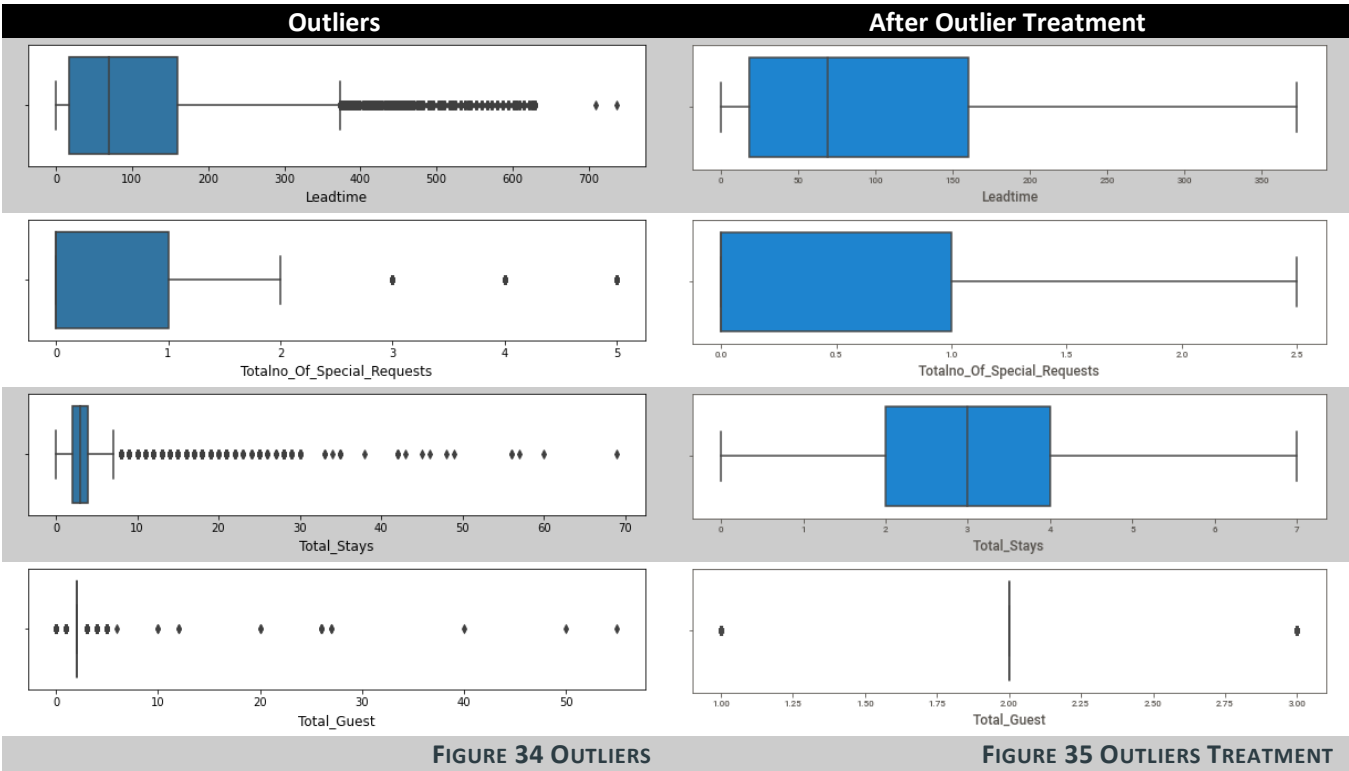
TABLE 1 MISSING VALUES

SNo.	Feature Name	# of Missing value		Possible Action
1	Country	488	(0.41%)	Can be filled with maximum occurrence
2	Agent	16340	(13.69%)	This feature is related to ID information details which would not require for ML for further analysis we are creating the new variables as booking_by_agent by <ul style="list-style-type: none">✓ Agent_Booking – the data has the values✓ Not Agent Booking – where the agent id exists
3	Company	112593	(94.31%)	
				This feature is related to ID information details which would not require for ML for further analysis we are creating the new variables as booking_via_company by <ul style="list-style-type: none">✓ Booking Via Company – the data has the values✓ Booking Not Via Company – where the agent id exists

OUTLIER TREATMENT

Outliers are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions. Unfortunately, all analysts will confront outliers and be forced to make decisions about what to do with them.

- ✓ There are outliers present in all the variables.
- ✓ The variable "Lead time","Total_Stays","Total_Guest" will be treated the outliers



VARIABLE TRANSFORMATIONS

Addition of New Variables

We have created the new feature to analyse more insights on the data and to identify how the new feature has impact to the target variable.

TABLE 2 NEW VARIABLES

Sno.	Variable	Reason
1	a_year	Year of Arrival date of the Booking
2	a_month	Month of the Arrival date of the Booking(January, February.... December)
3	a_day_of_week	Weekday of the Arrival date of the booking(Sunday, Monday....)
4	a_weekno	Week of the Year for the Arrival date of the booking
5	lead_time	Identify the difference of Booking Date and Arrival Date
6	Country_name	Created new feature to more idea about the country code provided in the data this achieved as extracted the country details using iso3166 python package.
7	total_guest	The customer classified as adult and children to identify the total guest we have added both the features.
8	total_stays	The data provided as number of week day nights and week end night days so we have created this variable to calculate the total number of nights stayed by the customer
9	country_new	We assume the all the local booking is from Portugal since nearly 40% of booking from Portugal so we split the data based on the Local(Portugal) and International

Variables Modified

TABLE 3 VARIABLES TRANSFORMATIONS

Sno.	Variable	Reason
1	booking_changes_new	Created new variable to split changes happened after booking or not.
2	booking_via_company	Whether booking done via company
3	booking_via_agent	Where the booking done by travel agent
4	Lead Time Month	This feature created to find average month the lead time for each booking

Variables Removed

TABLE 4 VARIABLES REMOVED

Sno.	Variable	Reason
1	booking_date	This has been transferred to lead time and since this is a date variable which not required for classification algorithm
2	arrival_date	
3	stays_in_weekend_nights	Transformed to total stays
4	stays_in_week_nights	Transformed to total stays
5	adults	Transformed to total_guests
6	children	Transformed to total_guests
8	agent	Transformed to agent_by_booking
9	company	Transferred to booking_via_company
10	booking_changes	Transformed to booking_changes_new for groping

PRE-PROCESS THE VARIABLE FOR THE MODEL BUILDING

The variables not as numeric needs to be pre-processed as numeric before scaling and model building. There are difference kinds of encoding technics are used to convert the variables.

TABLE 5 PRE-PROCESSING BINARY

Features	Value to 1	Value to 0
hotel	Type1	Type2
previous_cancellations_new	Yes	No
previous_bookings_not_canceled_new	Yes	No
previous_cancellations_encoded	Booking Via Company	Booking Not Via Company
book_by_agent	Agent Booking	Not Agent Booking

ONE HOT ENCODING

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

Below categorical variables are one hot encoding.

- ☒ Deposit Type
- ☒ Market Segment
- ☒ Distribution Channel
- ☒ Customer Type
- ☒ Assigned Room Type
- ☒ Location
- ☒ Meal

The “booking date” and “arrival date” are converted using the string replace function.

Example : 2017-07-24 converted as **20170724**

TABLE 6 DROPPED VARIABLES

Deposit Type	Book By Agent Encoded	Meal
Market Segment	Country Type	Total Guest
Distribution Channel	C Name	Days In Waiting List
Customer Type	Country	A Day Of Week
Assigned Room Type	Leadtime Month	A Month
Location	Previous Cancellations	A Day
Hotel	Is Repeated Customer	A Weekno
Previous Cancellations New	Previous Bookings Not Canceled	A Year
Previous Bookings Not Canceled New		Reserved Room Type
Book Via Company		Total Stays
Book By Agent		

SOLUTION 4: MODEL BUILDING

Question : Model building - Clear on why was a particular model(s) chosen. - Effort to improve model performance.

VALIDATE THE RELATIONSHIP

Visualizing correlation coefficients between features and cancellation:

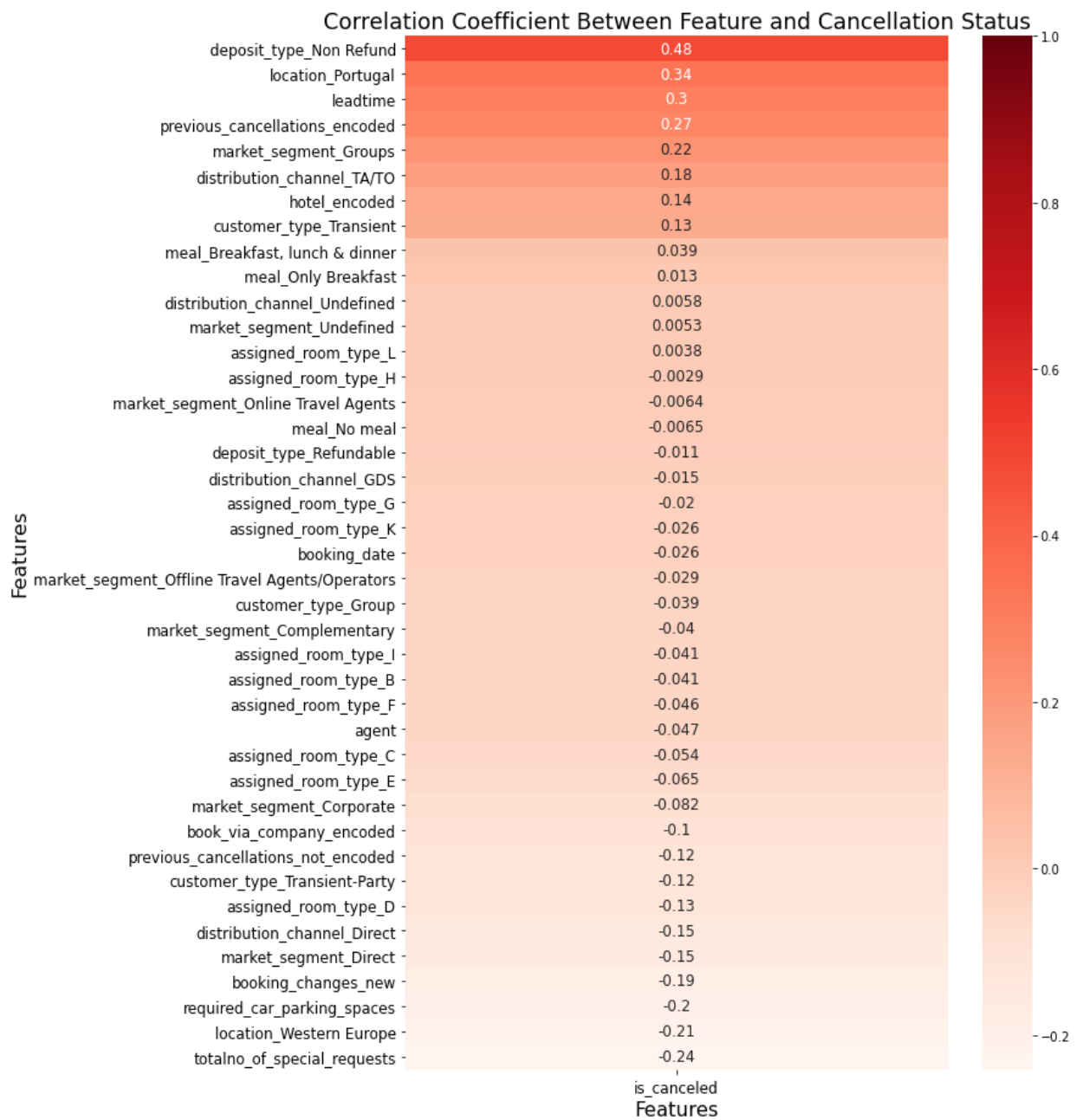


FIGURE 36 CORRELATION ON TARGET VARIABLE

The below variables are have optimum correlation on the target variable. We have taken the threshold as 0.2 to generate the below table.

TABLE 7 HIGHER CORRELATION ON TARGET

SNO.	FEATURES	CORR(IS_CANCELED)
1	deposit_type_Non Refund	0.482033
2	location_Portugal	0.337683
3	leadtime	0.295044
4	previous_cancellations_encoded	0.271239
5	market_segment_Groups	0.222251
6	location_Western Europe	-0.212844
7	totalno_of_special_requests	-0.240975

IDENTIFY THE CORRELATION OF THE PREDICTORS USING (VIF)

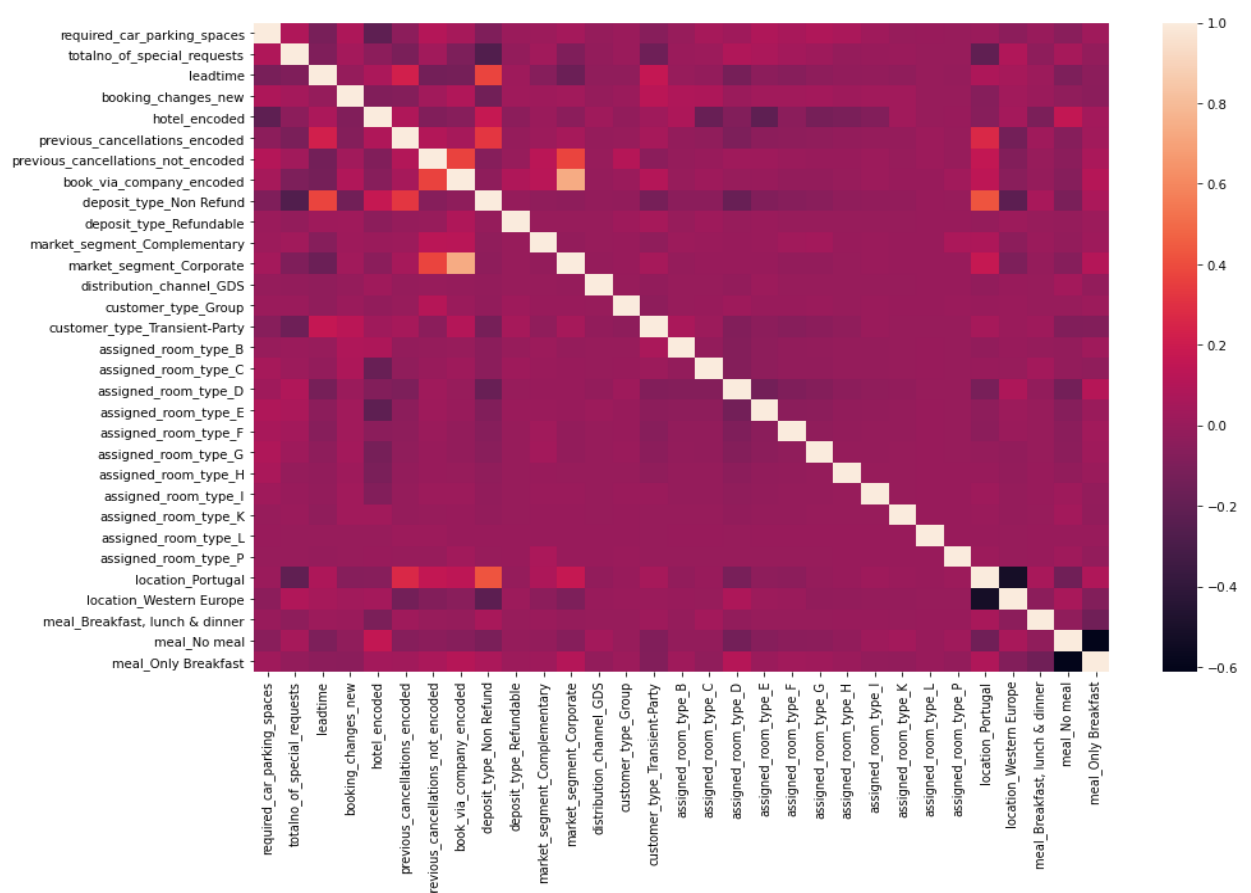


FIGURE 37 CORRELATION ON PREDATORS

From the heatmap there is no multi correction in the majority of variables. Below are variables have high correction between predictors this will kept for model building.

TABLE 8 HIGH VIF VARIABLES

Features	Multicollinearity
market_segment_Direct	16.37
market_segment_Groups	18.66
market_segment_Offline Travel Agents/Operators	22.51
market_segment_Online Travel Agents	54.64
distribution_channel_Direct	12.43
distribution_channel_TA/TO	56.53
customer_type_Transient	23.63

SPLIT DATA FOR TRAINING AND TEST DATA

To build the machine learning models required to split the dataset into Training and Testing Data with the ratio of 85:15. Then the sliced datasets are stored in two variables as X_train and X_test. The Random State “9” is used for split the data.

Made the parameter “stratify” = “y” to equally distribute the predictors and target variables.

The target variable “is_canceled” is dropped in X dataset and stored in y dataset for the verification of the model performance.

The X_train has 101138 values and X_test as 17849 and below is the target distribution.

TABLE 9 TARGET SPLIT ON TRAIN AND TEST

Class	Train	Test
Not Cancelled (0)	63641 (63%)	11231(63%)
Cancelled (1)	37497 (37%)	6618 (37%)

SCALING

Standardization or scaling is an important aspect of data pre-processing, it is applied to independent variables which helps to normalise the data in a particular range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

The Machine Learning algorithms that require the feature scaling are mostly KNN (K-Nearest Neighbours), Neural Networks, Linear Regression, and Logistic Regression.

The machine learning algorithms that do not require feature scaling is mostly non-linear ML algorithms such as Decision trees, Random Forest, AdaBoost, Naïve Bayes, etc.

CLASSIFICATION MODEL - WHY WAS A PARTICULAR MODEL(S) CHOSEN?

Our goal here is to rightly classify the cancellation status (is_canceled) of the hotel bookings on the data set. There are two classes provided as

- ⇒ **Cancelled as (1)**
- ⇒ **Confirmed / Not Cancelled as (0)**

TABLE 10 CHOOSING CLASSIFICATION MODELS

Model	Why Chosen and How it's Works on the Classification Model?
Logistic Regression	This used when class is in binary in nature. Logistic Regression uses sigmoid function which resembles an “S” shaped curve on the graph it “squishes” them towards the margins at the top and bottom, labelling them as 0 and 1.
LDA (Linear Discriminant Analysis)	The linear Discriminant analysis estimates the probability that a new set of inputs belongs to every class. The output class is the one that has the highest probability. That is how the LDA makes its prediction
Naïve Bayes	It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable
Random Forest	Random Forest uses multiple trees to average (regression) or compute majority votes (classification) in the terminal leaf nodes when making a prediction.
KNN	KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).
ANN	In binary classification tasks, it is common to classify all the predictions of a neural network to the positive class(1) if the estimated probability(\hat{p}) is greater than a certain threshold, and similarly, to the negative class(0) if the estimated probability.
XG-Boost	By default, the predictions made by XGBoost are probabilities . Because this is a binary classification problem, each prediction is the probability of the input pattern belonging to the first class. We can easily convert them to binary class values by rounding them to 0 or 1.

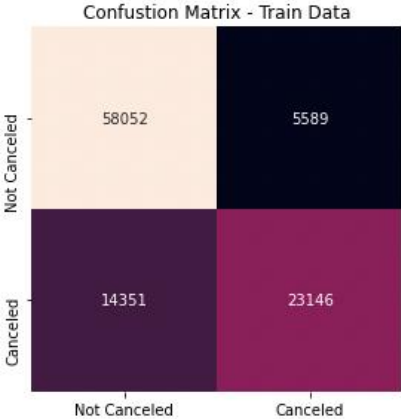

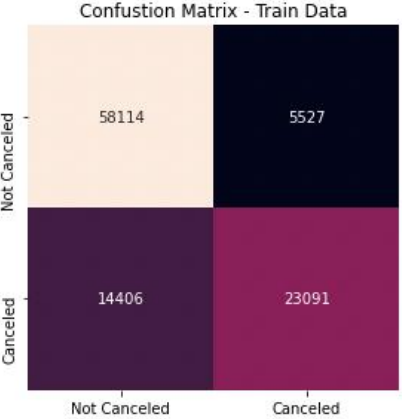
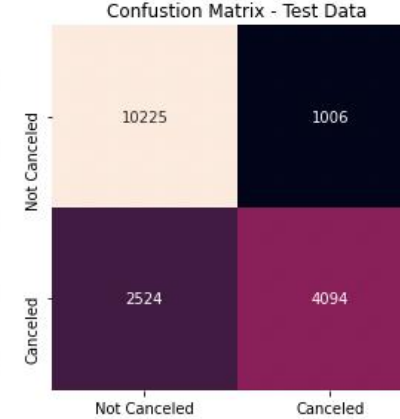
BASE AND TUNED MODEL

All the above models are built based using default parameters as “Base Model” and based on the performance results the model has been tuned using hyper parameters and built the “Tuned Model” to improve the model performance.

MODEL BUILDING

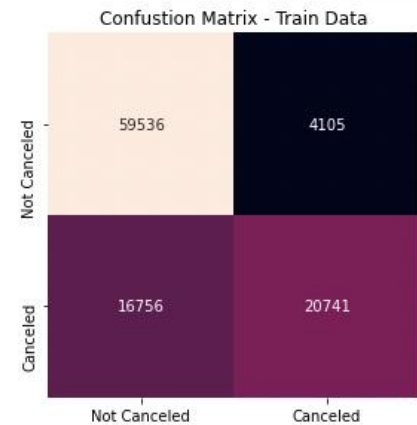
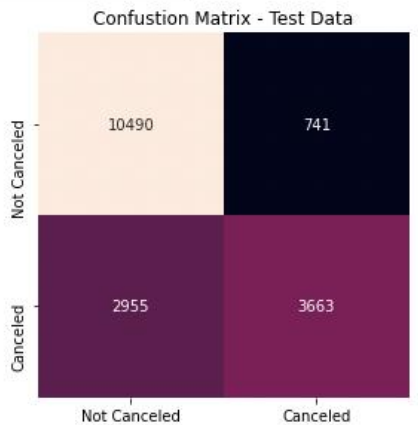
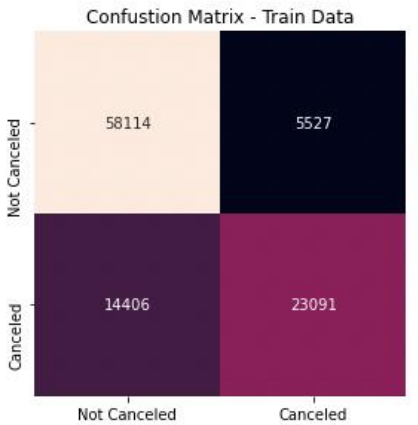
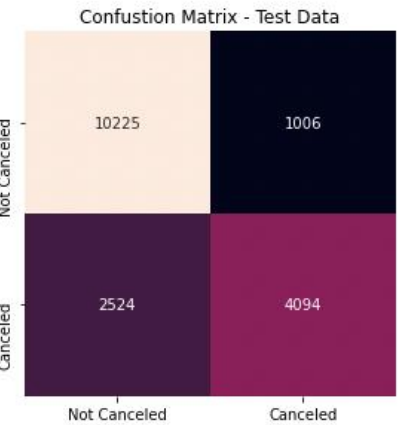
LOGISTIC REGRESSION

A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables.

BASE MODEL						TUNED MODEL					
Base Mode - Default Parameters: penalty='l2', dual=False, tol=0.0001, C=1, fit_intercept=True, intercept_scaling=1, class_weight=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None						Hyper Parameters: penalty=['none','l1','l2'], solver=['newton-cg','liblinear', tol=[0.01,0.001] , C=[0.1,0.01], maxit = [2000], njobs = [8,9]					
Best Estimators: C': 0.01, 'max_iter': 2000, 'n_jobs': 8, 'penalty': 'l2', 'solver': 'liblinear', 'tol': 0.01											
Scores	Precision	Recall	F-Score	Accuracy	AUC	Scores	Precision	Recall	F-Score	Accuracy	AUC
LogisticRegression_Base_Train	80.55	61.73	69.89	80.28	88.24	LogisticRegression_Tuned_Train	80.69	61.58	69.85	80.29	88.23
LogisticRegression_Base_Test	80.10	61.85	69.80	80.16	88.37	LogisticRegression_Tuned_Test	80.27	61.86	69.88	80.22	88.36
											
FIGURE 38 LOGISTIC REGRESSION – BASE MODEL											
FIGURE 39 - LOGISTIC REGRESSION – TUNED MODEL											
Inferences: <ul style="list-style-type: none">☑ This model is not overfit or underfit (the training and testing scores are close together)☑ The model is outperforming the baseline with a testing accuracy of 80.16%						Inferences: <ul style="list-style-type: none">☑ There is no much improvement in the tuned model, Still the training and testing models are close together.					


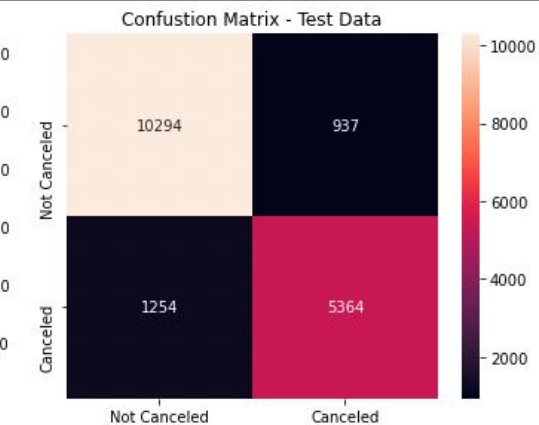

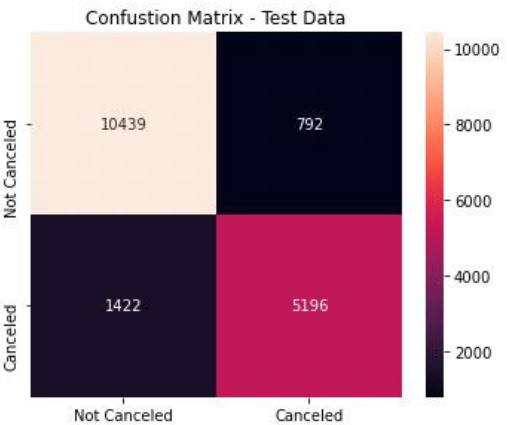
LDR (LINEAR DISCRIMINANT ANALYSIS)

Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis is a dimensionality reduction technique which is commonly used for the supervised classification problems.

BASE MODEL						TUNED MODEL					
Default Parameters: solver='svd', shrinkage=None, priors=None, n_components=None, store_covariance=False, tol=0.0001, covariance_estimator=None						Hyper Parameters: solver=['svd','eigen'] tol=[0.01,0.01,0.00001] Best Estimators: 'n_components': 1, 'solver': 'svd', 'tol': 0.01					
Scores	Precision	Recall	F-Score	Accuracy	AUC	Scores	Precision	Recall	F-Score	Accuracy	AUC
LDA_Base_Train	83.48	55.31	66.54	79.37	87.88	LDA_Tuned_Train	80.69	61.58	69.85	80.29	87.88
LDA_Base_Test	83.17	55.35	66.47	79.29	88.00	LDA_Tuned_Test	80.27	61.86	69.88	80.22	88.00
											
FIGURE 40 LDA – BASE MODEL											
											
						FIGURE 41 - LDA – TUNED MODEL					
Inferences: <ul style="list-style-type: none">☑ This model is same as Logistic regression and not overfit or underfit (the training and testing scores are close together)☑ The model is outperforming the baseline with a testing accuracy of 79.29% which is less then Logistic Regression						Inferences: <ul style="list-style-type: none">☑ There is no much improvement in the tuned model, Still the training and testing models are close together.					

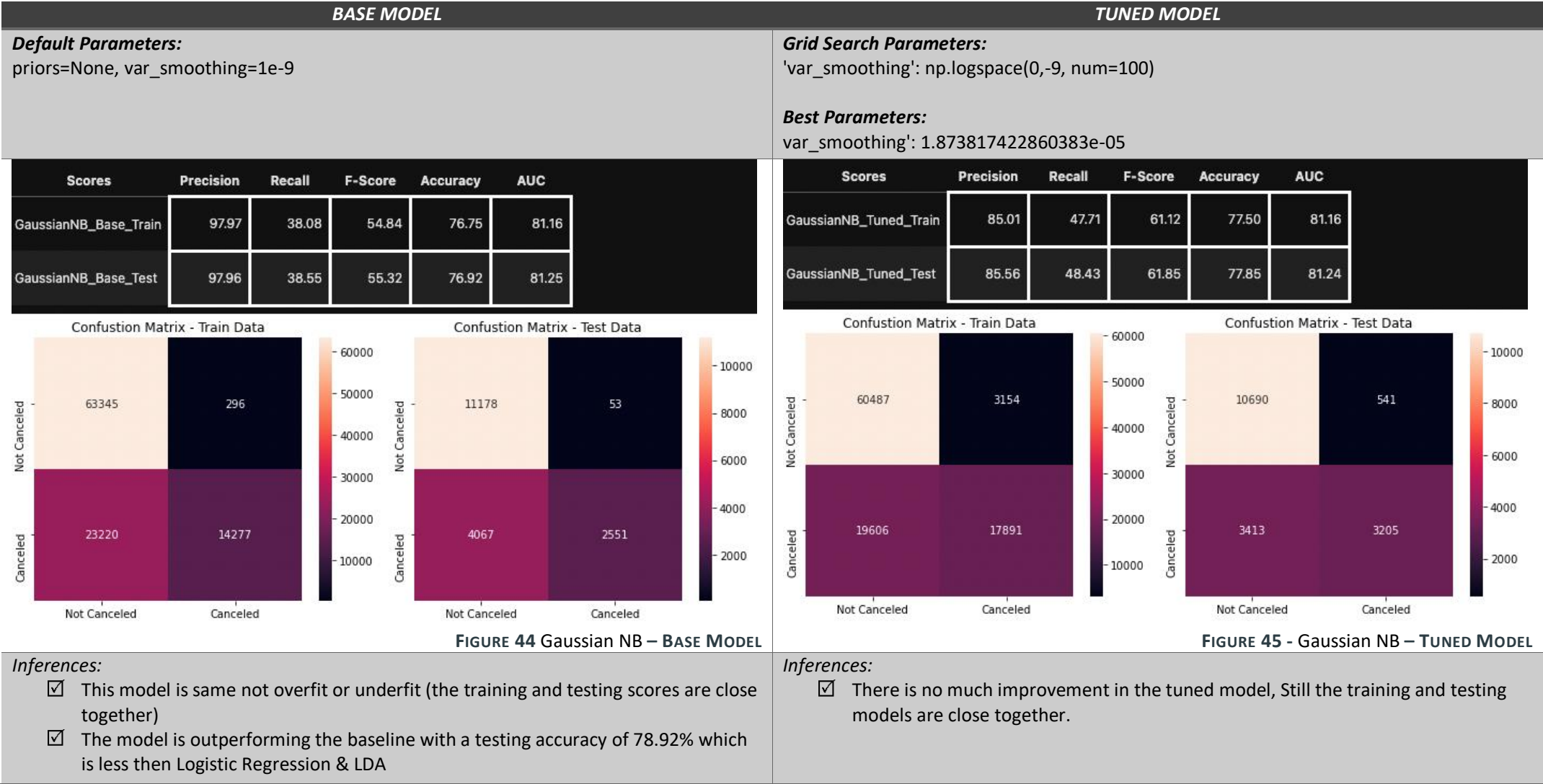
RANDOM FOREST CLASSIFIER

Random Forest is Supervised Learning Technique used in Machine Learning which consists of many decision trees that helps in predictions using individual trees and selects the best output from them.

BASE MODEL						TUNED MODEL					
Default Parameters: n_estimators=100, criterion="gini", max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0, max_features="auto", max_leaf_nodes=None, min_impurity_decrease=0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0, max_samples=None						Hyper Parameters: <input checked="" type="checkbox"/> max_depth' : [22,23] <input checked="" type="checkbox"/> 'max_features' : [18,19] <input checked="" type="checkbox"/> 'min_samples_leaf' : 3, <input checked="" type="checkbox"/> 'min_samples_split' : 6 <input checked="" type="checkbox"/> 'n_estimators' : [50,60]			Best Estimators: <input checked="" type="checkbox"/> max_depth' : 22 <input checked="" type="checkbox"/> 'max_features' : 18 <input checked="" type="checkbox"/> 'min_samples_leaf' : 3 <input checked="" type="checkbox"/> 'min_samples_split' : 6 <input checked="" type="checkbox"/> 'n_estimators' : 60		
Scores	Precision	Recall	F-Score	Accuracy	AUC	Scores	Precision	Recall	F-Score	Accuracy	AUC
RandomForest_Base_Train	99.30	98.80	99.05	99.30	99.93	RandomForest_Tuned_Train	90.90	84.01	87.32	90.95	97.79
RandomForest_Base_Test	85.13	81.05	83.04	87.72	94.69	RandomForest_Tuned_Test	86.77	78.51	82.44	87.60	94.93
											
FIGURE 42 RANDOM FOREST – BASE MODEL											
Inferences: <input checked="" type="checkbox"/> This model is a overfit model and the training score is greater than testing scores. <input checked="" type="checkbox"/> The model is performing the baseline with a testing accuracy of 87.72%.											
Inferences: <input checked="" type="checkbox"/> There is no over fit after tuned <input checked="" type="checkbox"/> This model has a highest accuracy upon all models.						FIGURE 43 – RANDOM FOREST – TUNED MODEL					

NAÏVE BAYES

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable.



K - NEAREST NEIGHBOUR

KNN is a lazy learning, non-parametric algorithm. It uses data with several classes to predict the classification of the new sample point. KNN is non-parametric since

BASE MODEL						TUNED MODEL					
Default Parameters: n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None,						Since KNN is works based on the n_neighbors parameter the model built based by iterating the KNN from 3 to 15 with increment value as 3 , Here 4 models built. The MCE (Minimum Classification Error) calculated by subtracting the model accuracy with the value 1 (100%) and the best model with less MCE is KNN=9					
Scores	Precision	Recall	F-Score	Accuracy	AUC	Scores	Precision	Recall	F-Score	Accuracy	AUC
KNN_Base_Train	86.94	81.78	84.28	88.69	96.01	KNN_Tuned_Train	85.44	78.00	81.55	86.92	94.73
KNN_Base_Test	80.32	75.28	77.72	83.99	90.25	KNN_Tuned_Test	81.99	74.37	78.00	84.44	91.27

Confusion Matrix - Train Data

Confusion Matrix for Base Model Train Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 4607, True Negatives (TN) are 59034, False Positives (FP) are 6832, and False Negatives (FN) are 30665. A color scale on the right ranges from 0 to 100,000.

Not Canceled	59034	4607
Canceled	6832	30665

Confusion Matrix - Test Data

Confusion Matrix for Base Model Test Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 1221, True Negatives (TN) are 10010, False Positives (FP) are 1636, and False Negatives (FN) are 4982. A color scale on the right ranges from 0 to 10,000.

Not Canceled	10010	1221
Canceled	1636	4982

FIGURE 46 KNN – BASE MODEL

Confusion Matrix - Train Data

Confusion Matrix for Tuned Model Train Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 4983, True Negatives (TN) are 58658, False Positives (FP) are 8250, and False Negatives (FN) are 29247. A color scale on the right ranges from 0 to 100,000.

Not Canceled	58658	4983
Canceled	8250	29247

Confusion Matrix - Test Data

Confusion Matrix for Tuned Model Test Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 1081, True Negatives (TN) are 10150, False Positives (FP) are 1696, and False Negatives (FN) are 4922. A color scale on the right ranges from 0 to 10,000.

Not Canceled	10150	1081
Canceled	1696	4922

FIGURE 47 - KNN – TUNED MODEL

Inferences:

- ☑ This model is same as medium variance on accuracy which is good to proceed with model. The training and testing scores are not closer or not much far.
- ☑ The model is outperforming the baseline with a testing accuracy of 83.99% which is better than Logistic and LDA Models

Inferences:

- ☑ There is no significant change in the this model on the accuracy after tuning.

ARTIFICIAL NEURAL NETWORKS (ANN)

An artificial neural network (ANN) is the component of artificial intelligence that is meant to simulate the functioning of a human brain. Processing units make up ANNs, which in turn consist of inputs and outputs.

BASE MODEL						TUNED MODEL					
Default Parameters: hidden_layer_sizes=(100,), activation="relu", *, solver='adam', alpha=0.0001, batch_size='auto', learning_rate="constant", learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-8, n_iter_no_change=10, max_fun=15000						Hyper Parameters: 'hidden_layer_sizes' : [200,350], 'max_iter' : [500,750], 'solver' : ['sgd','adam'], 'tol' : [0.01,0.001],		Best Estimators: <input checked="" type="checkbox"/> hidden_layer_sizes : 350, <input checked="" type="checkbox"/> max_iter : 500, <input checked="" type="checkbox"/> random_state : 9, <input checked="" type="checkbox"/> tol : 0.001			
Scores	Precision	Recall	F-Score	Accuracy	AUC	Scores	Precision	Recall	F-Score	Accuracy	AUC
ANN_Base_Train	83.69	76.76	80.07	85.84	93.50	ANN_Tuned_Train	84.68	73.86	78.90	85.36	93.05
ANN_Base_Test	83.21	75.66	79.26	85.32	92.79	ANN_Tuned_Test	84.46	73.00	78.31	85.01	92.68

Confusion Matrix - Train Data

Confusion Matrix for Base Model Train Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 28782, True Negatives (TN) are 58032, False Positives (FP) are 8715, and False Negatives (FN) are 5609. The color scale ranges from 10000 to 50000.

	Not Canceled	Canceled
Not Canceled	58032	5609
Canceled	8715	28782

Confusion Matrix - Test Data

Confusion Matrix for Base Model Test Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 5007, True Negatives (TN) are 10221, False Positives (FP) are 1611, and False Negatives (FN) are 1010. The color scale ranges from 2000 to 10000.

	Not Canceled	Canceled
Not Canceled	10221	1010
Canceled	1611	5007

FIGURE 48 ANN – BASE MODEL

Confusion Matrix - Train Data

Confusion Matrix for Tuned Model Train Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 27696, True Negatives (TN) are 58631, False Positives (FP) are 9801, and False Negatives (FN) are 5010. The color scale ranges from 10000 to 50000.

	Not Canceled	Canceled
Not Canceled	58631	5010
Canceled	9801	27696

Confusion Matrix - Test Data

Confusion Matrix for Tuned Model Test Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 4831, True Negatives (TN) are 10342, False Positives (FP) are 1787, and False Negatives (FN) are 889. The color scale ranges from 2000 to 10000.

	Not Canceled	Canceled
Not Canceled	10342	889
Canceled	1787	4831

FIGURE 49 – ANN – TUNED MODEL

Inferences:

- ☒ This model is same not overfit or underfit (the training and testing scores are close together)
- ☒ The model is outperforming the baseline with a testing accuracy of 85.32%

Inferences:

- ☒ There is no much improvement in the tuned model, Still the training and testing models are close together.

XGBOOST

XGBoost or extreme gradient boosting is one of the well-known gradient boosting techniques(ensemble) having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms.

BASE MODEL						TUNED MODEL					
Default Parameters: objective="binary:logistic", use_label_encoder=True						Hyper Parameters: <div><div><input checked="" type="checkbox"/> 'n_estimators': [18], <input checked="" type="checkbox"/> 'colsample_bytree': [0.7], <input checked="" type="checkbox"/> 'max_depth': [19], <input checked="" type="checkbox"/> 'reg_alpha': [1.9],</div><div><input checked="" type="checkbox"/> 'min_child_weight': [2.5], <input checked="" type="checkbox"/> 'gamma': [4.2], <input checked="" type="checkbox"/> 'subsample': [0.9], <input checked="" type="checkbox"/> 'objective':['binary:hinge']50],</div></div>					
Scores	Precision	Recall	F-Score	Accuracy	AUC	Scores	Precision	Recall	F-Score	Accuracy	AUC
XGBoost_Base_Train	86.53	76.96	81.47	87.02	84.95	XGBoost_Tuned_Train	87.86	81.22	84.41	88.88	87.31
XGBoost_Base_Test	85.75	75.85	80.50	86.37	84.21	XGBoost_Tuned_Test	85.67	78.21	81.77	87.07	85.25

Confusion Matrix - Train Data

Confusion Matrix for Base Model Train Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 28858, True Negatives (TN) are 59149, False Positives (FP) are 8639, and False Negatives (FN) are 4492. A color scale on the right ranges from 10000 to 50000.

	Not Canceled	Canceled
Not Canceled	59149	4492
Canceled	8639	28858

Confusion Matrix - Test Data

Confusion Matrix for Base Model Test Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 5020, True Negatives (TN) are 10397, False Positives (FP) are 1598, and False Negatives (FN) are 834. A color scale on the right ranges from 2000 to 10000.

	Not Canceled	Canceled
Not Canceled	10397	834
Canceled	1598	5020

FIGURE 50 XG-Boost – BASE MODEL

Confusion Matrix - Train Data

Confusion Matrix for Tuned Model Train Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 30456, True Negatives (TN) are 59433, False Positives (FP) are 7041, and False Negatives (FN) are 4208. A color scale on the right ranges from 10000 to 50000.

	Not Canceled	Canceled
Not Canceled	59433	4208
Canceled	7041	30456

Confusion Matrix - Test Data

Confusion Matrix for Tuned Model Test Data. The matrix shows counts for 'Not Canceled' and 'Canceled' classes. True Positives (TP) are 5176, True Negatives (TN) are 10365, False Positives (FP) are 1442, and False Negatives (FN) are 866. A color scale on the right ranges from 2000 to 10000.

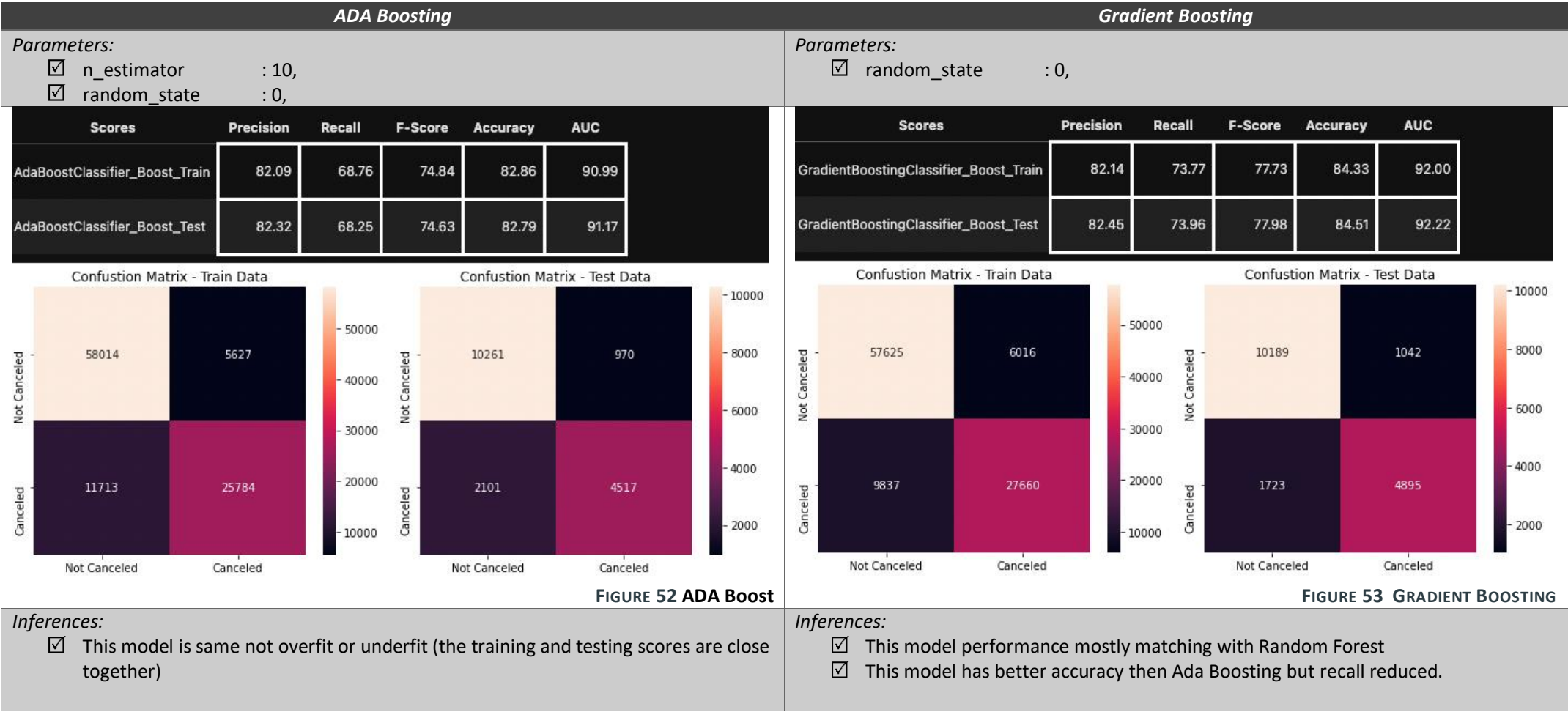
	Not Canceled	Canceled
Not Canceled	10365	866
Canceled	1442	5176

FIGURE 51 XG-Boost – TUNED MODEL

| Inferences: ☒ This model is same not overfit or underfit (the training and testing scores are close together) ☒ The model is outperforming the baseline with a testing accuracy of 86.37% | | | | | | Inferences: ☒ This model performance mostly matching with Random Forest ☒ There is an improvement in the train score after tuning on accuracy as 87.07% | | | | | |

BOOSTING

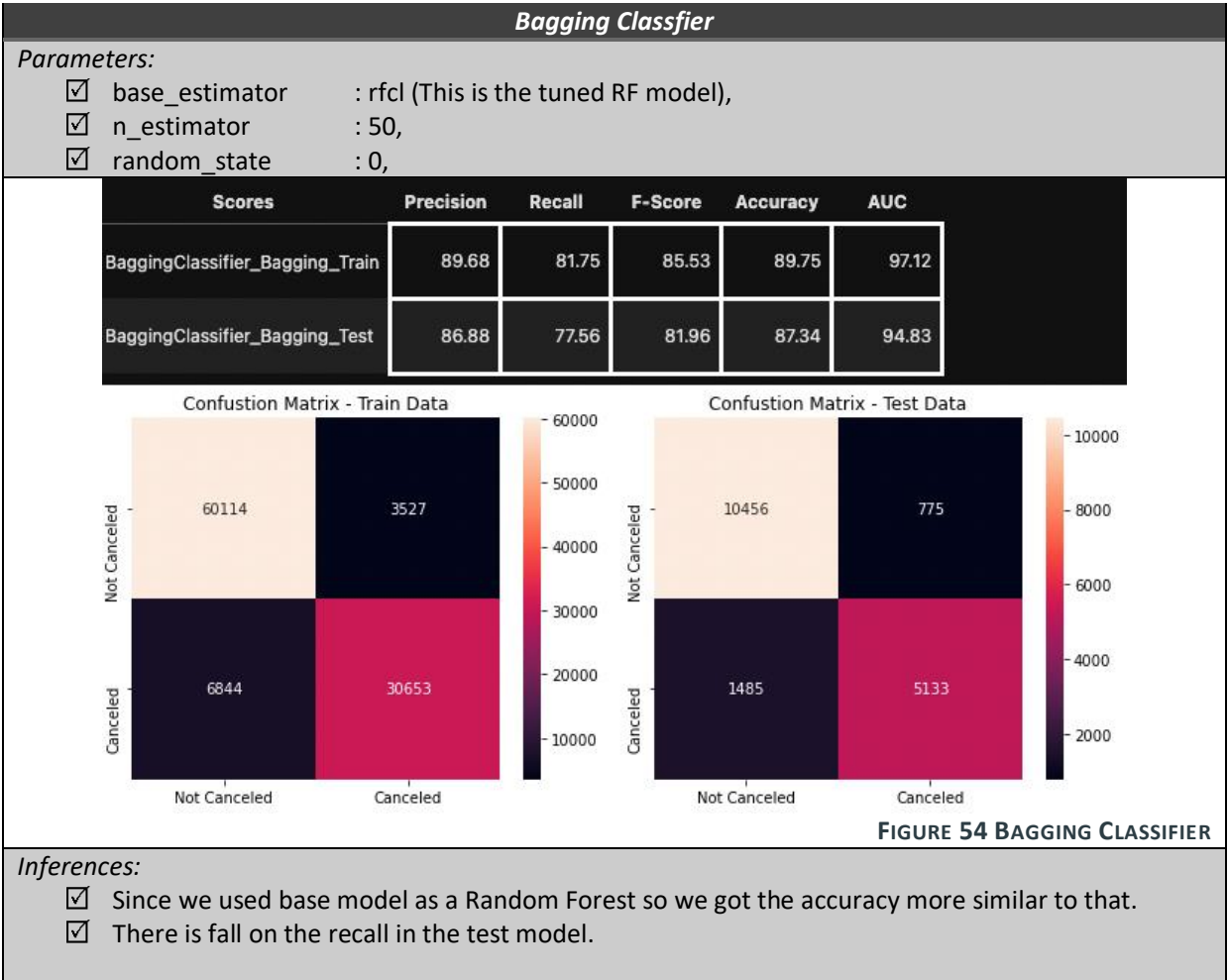
Boosting is a sequential ensemble method that in general decreases the bias error and builds strong predictive models. The term ‘Boosting’ refers to a family of algorithms which converts a weak learner to a strong learner.



BAGGING CLASSIFIER

Bagging, a Parallel ensemble method (stands for Bootstrap Aggregating), is a way to decrease the variance of the prediction model by generating additional data in the training stage

The base estimator to fit on random subsets of the dataset. If None, then the base estimator is a DecisionTreeClassifier. **We are going use the RF model for Bagging**



EFFORTS TO IMPROVE MODEL PERFORMANCE

Hyper-parameters are parameters that are not learnt within model by itself. Hyper-parameters are passed as arguments to the constructor of the steps in pipeline. Based on the cross validation score it is possible to fetch the best possible parameters.

For hyper-parameter tuning **GridSearchCV** is one of the options. It performs exhaustive search over specified parameter values for the model. Here we are passing the hyper-parameters to steps in pipeline using **param_grid**. **cv** is set to **3** since we are to perform 3-fold cross-validation. **scoring** is set to **accuracy** since we want to predict accuracy of the model.

For Example : In the Random Forest – The model is overfitting and so our goal to reduce the bias and maintain or reduce the variance on the model tuning. The parameters **n_estimators**, **max_features**, **min_sample_split** and **min_sample_leaf** on multiple iterations and the best estimator has been identified. The **n_estimators** is a number of trees on increasing this value will make the model complex and it will make the grid search slow also increasing the **min_sample_leaf** will increase the model bias and move the type2 error to type1 error so model become reliable after tuning.

SOLUTION 4: MODEL VALIDATION

Question : How was the model validated ? Just accuracy, or anything else too ?

HOW WAS THE MODEL VALIDATED?

Model validation is the process by which model outputs are (systematically) compared to independent real-world observations to judge the quantitative and qualitative correspondence with reality.

FIGURE 55 MODEL VALIDATION

Model	RF				ANN				KNN				XG			
Base_Tuned	Base		Tuned		Base		Tuned		Base		Tuned		Base		Tuned	
Train_Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Precision	99.30	85.13	90.90	86.77	83.73	83.24	84.68	84.46	86.94	80.30	85.44	81.99	86.53	85.75	87.91	85.84
Recall	98.80	81.05	84.01	78.51	76.76	75.72	73.86	73.00	81.78	75.28	78.00	74.37	76.96	75.85	81.03	78.53
F-Score	99.05	83.04	87.32	82.44	80.09	79.30	78.90	78.31	84.28	77.71	81.55	78.00	81.47	80.50	84.33	82.02
Accuracy	99.30	87.72	90.95	87.60	85.85	85.34	85.36	85.01	88.69	83.99	86.91	84.44	87.02	86.37	88.84	87.24
AUC	99.93	94.69	97.79	94.93	93.50	92.79	93.05	92.68	96.01	90.25	94.73	91.27	84.95	84.21	87.23	85.45

Model	LDA				LR				NB			
Base_Tuned	Base		Tuned		Base		Tuned		Base		Tuned	
Train_Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Precision	83.48	83.17	80.69	80.27	80.55	80.10	80.69	80.27	97.97	97.96	85.01	85.56
Recall	55.31	55.35	61.58	61.86	61.73	61.85	61.58	61.86	38.08	38.55	47.71	48.43
F-Score	66.54	66.47	69.85	69.88	69.89	69.80	69.85	69.88	54.84	55.32	61.12	61.85
Accuracy	79.37	79.29	80.29	80.22	80.28	80.16	80.29	80.22	76.75	76.92	77.50	77.85
AUC	87.88	88.00	87.88	88.00	88.24	88.37	88.23	88.36	81.16	81.25	81.16	81.24

Model	Bagging		Boosting			
Base_Tuned	Bagging		Ada Boost		Gradient Boost	
Train_Test	Train	Test	Train	Test	Train	Test
Precision	89.68	86.88	82.09	82.32	82.14	82.45
Recall	81.75	77.56	68.76	68.25	73.77	73.96
F-Score	85.53	81.96	74.84	74.63	77.73	77.98
Accuracy	89.75	87.34	82.86	82.79	84.33	84.51
AUC	97.12	94.83	90.99	91.17	92.00	92.22

The train and test data has been validated on the Based and Tuned models to identify the best model. Above is the heat map of the all the models validations represented as Green is high score and Red as low score and Yellow is average score based on the score.

BIAS / TRAINING ERROR :

- ✓ The comparison of the actual classes and trained classes are BIAS or training error. For example from the above table the Train accuracy of the RF Base model is 99.30 which is $100-99.30 = 0.60$ which is low bias.
- ✓ The accuracy of the NB based model is 76.75 which is $100-76.75 = 23.25$ which is high bias and this is low bias based on all other models.

VARIANCE / TEST ERROR :

- ✓ The difference of the train metrics and test metrics is VARIANCE or test error. again the Base RF model accuracy has a high variance as $87.72 - 99.30 = -11.55$. Since this model has a low bias and high variance the model is overfitting.
- ✓ The Tuned RF model has a low variance(-3.68) and low bias(9.05) which could be the best model.

Accuracy Variance compression by Models on Base and Tuned

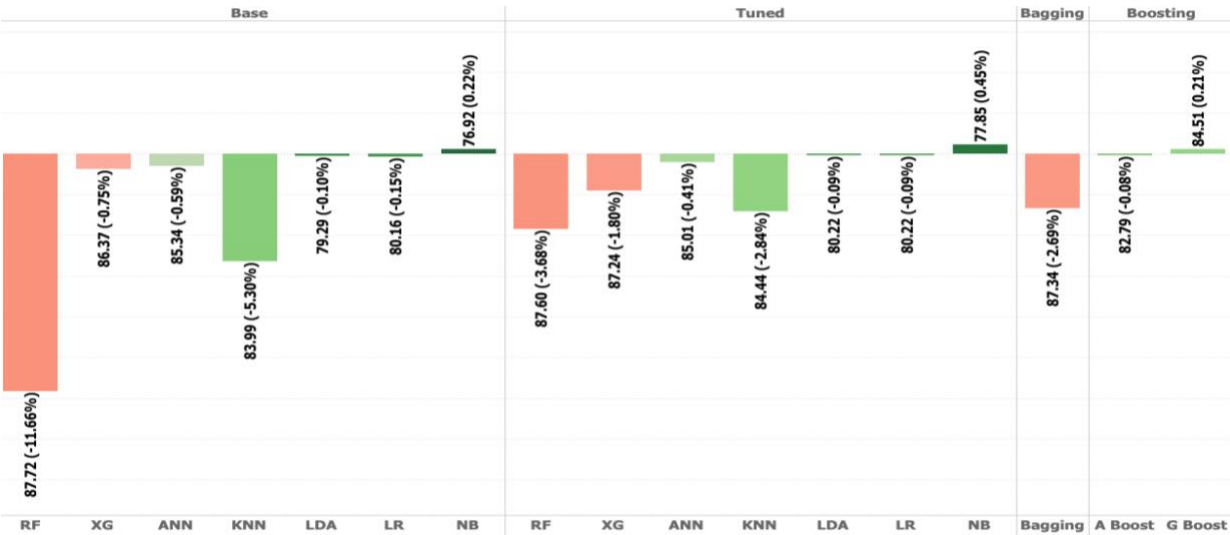


FIGURE 56 ACCURACY VARIANCE

JUST ACCURACY, OR ANYTHING ELSE TOO?

Accuracy is one of the common evaluation metrics in classification problems, Accuracy is useful when the target class is *well balanced* but is not a good choice with unbalanced classes.

From the initial analysis there is no class imbalance (37% : 63%) so the standard approach of the accuracy has been taken as primary metric for this problem.

Recall: The ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives.

Since the problem is related to cancellation rate which is calculating the negative impact of the business so the model should also consider as secondary score as recall to avoid the cancellations which is actually cancelled and the model has classified as not cancelled which have a revenue impact.

The Trend of Accuracy and Recall for Models broken down by Base, Tuned & Ensemble.

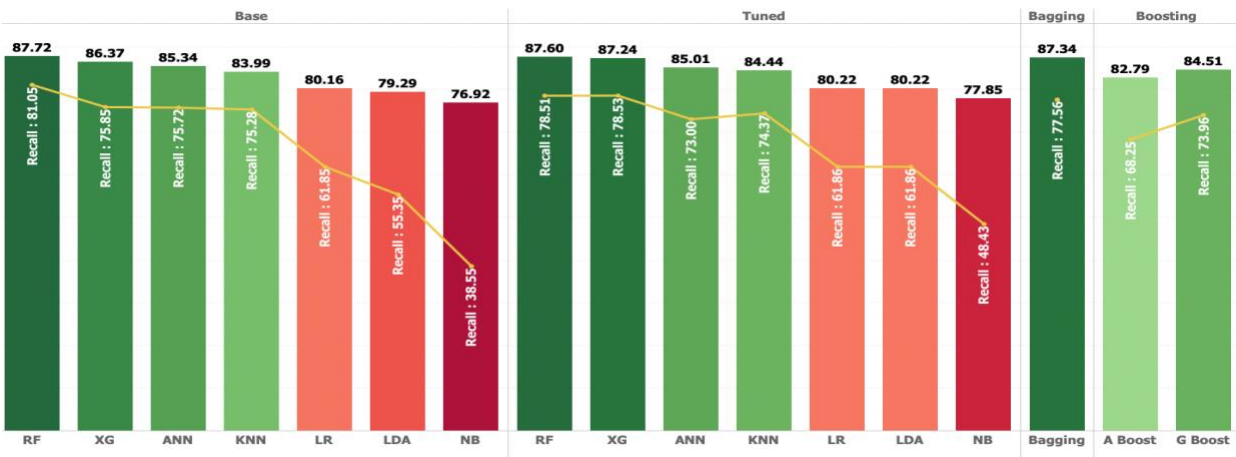
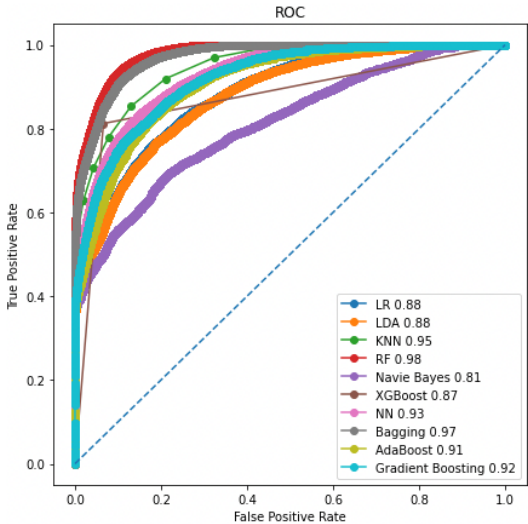


FIGURE 57 ACCURACY VS RECALL

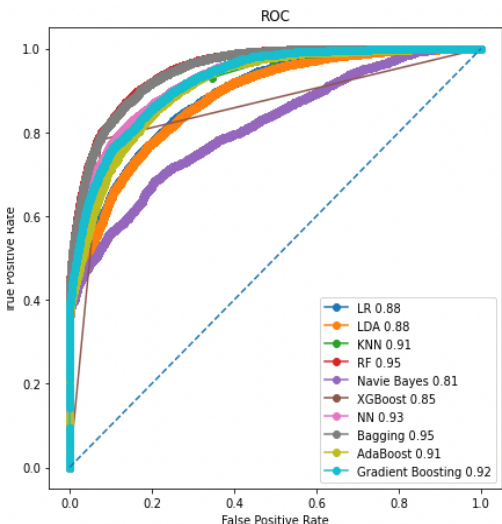
- ✓ Here we can see there is tough fight between Tuned RF model, XG and Bagging where the Accuracy is nearly 87% for all three models and recall as 78% unlike for Bagging as 77%.
- ✓ The next good fit model is Base model of Artificial Neural Networks which has very low variance and recall score as 75.72.
- ✓ The Boosting models have accuracy of more than 80% but the recall is not performed so not so good to selecting these models for this problem.
- ✓ Even though the Precision of the Navis Bayes model is good but the Accuracy and Recall is very low which is the worst model among all models. There no major improvement on after tuning.
- ✓ We can skip the Bagging from the further comparison since we the base estimator for the Bagging we have used Random Forest model the XGBoost VS Random Forest will compare to find the optimum model for Hotel Cancellation.
- ✓ Even though the test scores of the Random Forest are mostly matching with the XG-Boost model but it has a high bias then Random Forest Model which is not performing in train set. So it might not perform when new data comes in.

ROC COMPARISON ON TUNED MODELS

Train Models



Test Models



SOLUTION 5: FINAL INTERPRETATION / RECOMMENDATION

Question : Very clear and crisp on what recommendations do you want to give to the management / client.

FINAL INTERPRETATION

Important Features

- ☑ Lead Time, Deposit Type, Special Request, Location (Portugal) and Previous Cancellations.

Base Models

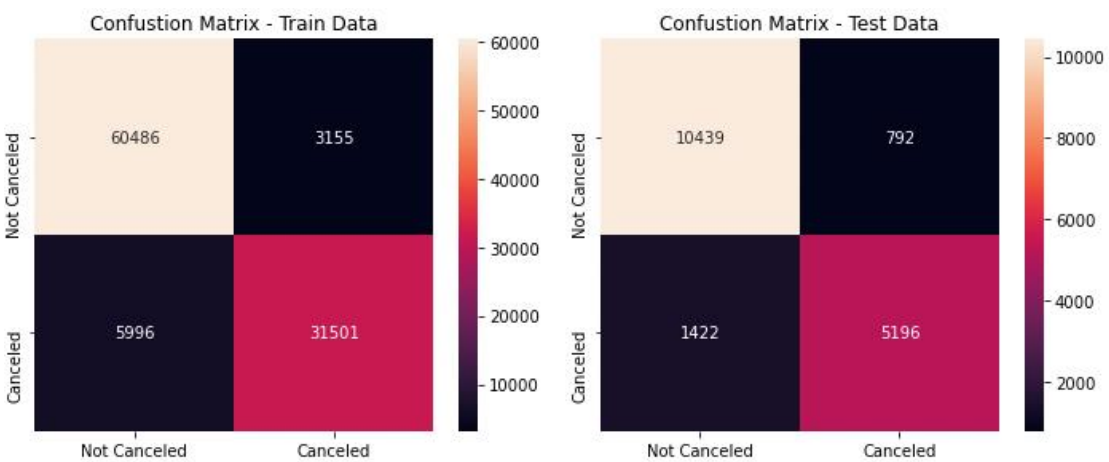
- ☑ Random Forest Overfitting as high Bias: **12.40%** and high Variance : **11.58%**
- ☑ LR and LDA has High Bias(app **20%**) and Low Variance(**0.10%** - **0.15%**)
- ☑ The XGBoost and ANN are the good fit models with Medium Bias (**15%**) and Low Variance(**0.75%**)
- ☑ GaussianNB has high precision and unable to predict which is actually predict.

Tuned Models

- ☑ No more overfitting after tuning on the Random Forest.
- ☑ Random Forest, XGBoost and Bagging models almost the accuracy as a **87%** upon the XGBoost has a low variance(**1.8%**).
- ☑ KNN has a good fit even in Tuned, but no major change in the Bias and Variance

OPTIMUM MODEL INTERPRETATION

- ☑ Overfitting
 - ⇒ For the base model we see that other than Random Forest algorithms that doesn't have an overfitting condition
 - ⇒ After Hyperparameter Tuning on the Random Forest have not overfitting and the variance as the significant difference which is less (10%), and after hyperparameter tuning (Random Forest) has the highest accuracy score
 - ⇒ Tuned Random Forest Confusion Matrix



- ☑ Random Forest Model able to predict booking cancellation with **87%** of accuracy
 - ⇒ After tuning the Bias (**12.4%**) and Variance (**3.68%**)
 - ⇒ Still the Recall score as **78.5%** so there is a risk of **21.5%** of misclassifications.
 - **8% (1422)** of the Total test data is classified as True Negative which is **Actually Canceled but classified as Not Canceled**.
 - **4% (792)** of the Total test data is classified as False Negative which is **Actually Not Canceled but classified as Canceled**.

Based on the all-models comparison and accepting all above constraints we are good to choose the **Random Forest** as the **Final Model**.

BUSINESS INSIGHTS

TABLE 11 BUSINESS INSIGHTS

SNO	FEATURE NAME	INSIGHTS
1	Lead Time	<p>We have grouped the lead time into monthly (30 days month) lead time to make it more general to analyse compared to a specific number of days</p> <ul style="list-style-type: none">☑ Booking that has more than 7 months of lead time are more likely to be cancelled than confirmed☑ Cancellation is positively correlated with lead time (the higher the lead time the higher the cancellation rate)
2	Deposit Type	<p>There are 3 kinds of deposit type in the data set are NO Deposit, NO Refund, and Refundable</p> <ul style="list-style-type: none">☑ No Refund Booking has the highest cancellation rate at 99.4%☑ No Deposit has cancellation rate of 28.3 %☑ While Refundable has cancellation rate around 22% <p>From the hotels point of view there is nothing alarming since they don't lose revenue when no refund booking is cancelled, but it's always a good practice to question something is extraordinary, why does non-refundable booking are most likely to be cancelled?</p>
3	Market Segment	<ul style="list-style-type: none">☑ Based on the analysis the corporate , Direct, and Aviation has a cancellation rate around 18 - 22 % of their booking☑ Travel Agent (Online / Offline) has a cancellation rate around 34 - 36 %☑ Lastly Group has the highest cancellation rate around 61 % <p>Based on this can be concluded the group booking are the market segment that's most likely to be cancelled compared to other market segment while Direct has the lowest cancellation rate at 15% (Outside Complimentary)</p>
4	Location	<p>The booking location in this dataset we originally have 178 countries (including Portugal), it's not efficient and not effective to aggregating every country with Portugal in this one we split the booking location into 3 Local (Booking that is from Portugal) and International (Booking Outside Portugal)</p> <ul style="list-style-type: none">☑ Nearly 40% of the data has the Portugal data and which has a 62.38% of cancellation rate.☑ Western Europe has 21.93% cancellation Rate.☑ Others Booking have 15.69% cancellation Rate
5	Repeated Guests	<ul style="list-style-type: none">☑ Whoever cancelled the booking before have a high impact(92%) to the current cancellation.☑ We can we can see the highest of the data is not cancelled and no previous cancellations so we assume there are more loyal customers in this scenario.☑ Booking that's originally wasn't cancelled has 34% Cancellation rate
6	Parking Space	<p>This is one of the not common metrics to look at when it comes to predicting cancellation and analysing cancellation, however in this data set there are around 7383 (6.2 %) that required car parking space(s).</p> <ul style="list-style-type: none">☑ 7383 Bookings that require a parking space there not a single booking that's Cancelled (0 Cancellation)☑ This conclude that booking that required a parking space will high likely to be confirmed
7	Special Request	<p>The number of special request(s) in a booking apparently affecting the cancellation rate of a booking from our analysis we see that booking that has no special request are more likely to cancelled compared to booking that has a special request</p> <ul style="list-style-type: none">☑ The cancellation rate of booking that has a special request is ranging from 5 - 22 % with booking with 5 special requests has the lowest cancellation rate☑ While Booking with no special request has cancellation rate of 48%

RECOMMENDATIONS

TABLE 12 RECOMMENDATIONS

SNO	HEAD	RECOMMENDATIONS
1	Only Non Refundable Deposit For Group Booking	This analysis results that the group booking has the highest cancellation rate among all market segment, only allowing non-refundable deposit for group booking will help protect the hotel from losing revenue due to last minute cancellation and not able to find replacement. Only Allowing Non Refundable Rates might result in fewer bookings for Group , however it might protect the hotel from losing revenue
2	Setting Maximum Lead Time for Booking	The pattern of the booking that has more than 210 days of lead time are more likely to be cancelled, setting up maximum lead time means it won't be able to make booking that's too far in advance (> 210 days), and setting maximum advance reservation will help you to reduce cancellation
3	Increase Direct Booking Market Segment	The dataset has direct booking has the least cancellation rate 15% (outside complimentary) compared to other market segment, with only being 10% of total booking market segment having more booking from direct market segment will likely to reduce the number of cancellation.

FEW STRATEGIES TO INCREASE DIRECT BOOKING

- 1. **Have a mobile-friendly hotel website**
 - ☑ Website should be accessible to any device
 - ☑ Offer & Ensure Best Rate Guarantee
 - ☑ Highlight the unique selling services
- 2. **Optimize website to rank on Google**
 - ☑ Nowadays, your guests would always explore your hotel and more options on search engines like Google
 - ☑ Need to perform search engine optimization (SEO) of your hotel website this will help to increase rank in the search engine so your hotel appears in first page of search.
- 3. **Implement a live chatbot to attend guest inquiries.**
 - ☑ Its becomes easy for you to provide instant replies to your website visitors.
 - ☑ When they will get their answer in a fraction of seconds, they will be able to make a decision instantly.

Source : [Ezeeabsolute](#)

