15 Aug 21



Hotel Cancellation Prediction



Final Report

Sivaramakrishanan s – july 2020 E

Table of Contents

[Solution 1: introduction 1](#_Toc79952850)

[Issues in Hotel Industry 1](#_Toc79952851)

[Need of the Study/Project 1](#_Toc79952852)

[Understanding Business/Social opportunity 1](#_Toc79952853)

[Solution 2: EDA and Business Implication 2](#_Toc79952854)

[Visual inspection of data (rows, columns, descriptive details) 2](#_Toc79952855)

[Descriptive Statistics of Numerical Variables 2](#_Toc79952856)

[Descriptive Statistics of Categorical Variables 2](#_Toc79952857)

[Understanding of attributes (Variable info, Renaming if required) 3](#_Toc79952858)

[Variable info 3](#_Toc79952859)

[Duplicate Value Check. 3](#_Toc79952860)

[Univariant Analysis 3](#_Toc79952861)

[Target Variable 3](#_Toc79952862)

[Hotel Type 3](#_Toc79952863)

[Deposit Type 4](#_Toc79952864)

[Customer Type 4](#_Toc79952865)

[Meal Type 4](#_Toc79952866)

[Repeated Guest 5](#_Toc79952867)

[Market Segment 5](#_Toc79952868)

[Booking Changes 5](#_Toc79952869)

[Location 6](#_Toc79952870)

[Total Guest 6](#_Toc79952871)

[Total Stays 7](#_Toc79952872)

[Lead Time – & Lead Time – Month 7](#_Toc79952873)

[Assigned Room Type 8](#_Toc79952874)

[Reserved Room Type 9](#_Toc79952875)

[Bivariant Analysis - **After Data Cleaning** 9](#_Toc79952876)

[Previous Cancellation VS Repeated Guest 9](#_Toc79952877)

[Deposit Type VS Market Segment 9](#_Toc79952878)

[Hotel Type and Cancellation 10](#_Toc79952879)

[Total Guest and Cancellation 10](#_Toc79952880)

[Meal Type and Cancellation 10](#_Toc79952881)

[Location and Cancellation 11](#_Toc79952882)

[Market Segment and Cancellation 11](#_Toc79952883)

[Deposit Type and Cancellation 11](#_Toc79952884)

[Previous Cancellation and Cancellation 12](#_Toc79952885)

[Lead Time and Cancellation 12](#_Toc79952886)

[Total Number of Special request and Cancellation 13](#_Toc79952887)

[Total Number of Parking space request and Cancellation 13](#_Toc79952888)

[Multi-Variant Analysis 14](#_Toc79952889)

[Solution 3: Data Cleaning 15](#_Toc79952890)

[Missing Values Treatment 15](#_Toc79952891)

[Outlier Treatment 15](#_Toc79952892)

[Variable Transformations 16](#_Toc79952893)

[Pre-process the variable for the model building 16](#_Toc79952894)

[One Hot Encoding 17](#_Toc79952895)

[Solution 4: Model Building 18](#_Toc79952896)

[Validate the Relationship 18](#_Toc79952897)

[Identify the Correlation of the Predictors Using (VIF) 18](#_Toc79952898)

[Split Data for Training and Test data 19](#_Toc79952899)

[Scaling 19](#_Toc79952900)

[Classification Model - Why was a particular model(s) chosen? 20](#_Toc79952901)

[Base and Tuned Model 20](#_Toc79952902)

[Model building 21](#_Toc79952903)

[Logistic Regression 21](#_Toc79952904)

[LDR (Linear Discriminant Analysis) 22](#_Toc79952905)

[Random Forest Classifier 23](#_Toc79952906)

[Naïve Bayes 24](#_Toc79952907)

[K - Nearest Neighbour 25](#_Toc79952908)

[Artificial Neural Networks (ANN) 26](#_Toc79952909)

[XGBoost 27](#_Toc79952910)

[Boosting 28](#_Toc79952911)

[Bagging Classifier 29](#_Toc79952912)

[Efforts to improve model performance 29](#_Toc79952913)

[Solution 4: Model Validation 30](#_Toc79952914)

[How was the model validated? 30](#_Toc79952915)

[Bias / Training Error : 30](#_Toc79952916)

[Variance / Test Error : 30](#_Toc79952917)

[Just accuracy, or anything else too? 31](#_Toc79952918)

[ROC Comparison on Tuned Models 31](#_Toc79952919)

[Solution 5: Final interpretation / recommendation 32](#_Toc79952920)

[Final Interpretation 32](#_Toc79952921)

[Optimum Model Interpretation 32](#_Toc79952922)

[Business Insights 33](#_Toc79952923)

[Recommendations 34](#_Toc79952924)

[Few Strategies to increase Direct Booking 34](#_Toc79952925)

Table of Tables

[Table 1 Missing Values 15](#_Toc79952926)

[Table 2 New Variables 16](#_Toc79952927)

[Table 3 Variables Transformations 16](#_Toc79952928)

[Table 4 Variables Removed 16](#_Toc79952929)

[Table 5 Pre-Processing Binary 16](#_Toc79952930)

[Table 6 Dropped Variables 17](#_Toc79952931)

[Table 7 Higher Correlation on Target 18](#_Toc79952932)

[Table 8 High VIF Variables 19](#_Toc79952933)

[Table 9 Target Split on Train and Test 19](#_Toc79952934)

[Table 10 Choosing Classification Models 20](#_Toc79952935)

[Table 11 Business Insights 33](#_Toc79952936)

[Table 12 Recommendations 34](#_Toc79952937)

Table of Figures

[Figure 1 Cancellation Rate 1](#_Toc79952938)

[Figure 2 Rows And Columns 2](#_Toc79952939)

[Figure 3 Descriptive Statistics Numerical 2](#_Toc79952940)

[Figure 4 Descriptive Statistics Categorical 2](#_Toc79952941)

[Figure 5 Data Info 3](#_Toc79952942)

[**Figure 6 Target Distribution** 3](#_Toc79952943)

[**Figure 7 Hotel Type** 3](#_Toc79952944)

[Figure 8 Deposit Type 4](#_Toc79952945)

[**Figure 9 Customer Type** 4](#_Toc79952946)

[**Figure 10 Meal Type** 4](#_Toc79952947)

[**Figure 11 Repeated Guest** 5](#_Toc79952948)

[**Figure 12 Market Segment** 5](#_Toc79952949)

[Figure 13 Booking Changes 5](#_Toc79952950)

[Figure 14 Country 6](#_Toc79952951)

[Figure 15 Total Guest 6](#_Toc79952952)

[Figure 16 Total Stays 7](#_Toc79952953)

[Figure 17 Lead Time 7](#_Toc79952954)

[**Figure 18 Assigned Room Type** 8](#_Toc79952955)

[**Figure 19 Reversed Room Type** 9](#_Toc79952956)

[**Figure 20 Previous Cancellation VS Repeated Guest** 9](#_Toc79952957)

[Figure 21 Deposit Type VS Market Segment 9](#_Toc79952958)

[Figure 22 Hotel Type VS Cancellation 10](#_Toc79952959)

[Figure 23 Total Guest VS Cancellation 10](#_Toc79952960)

[Figure 24 Meal Type VS Cancellation 10](#_Toc79952961)

[**Figure 25 Location VS Cancellation** 11](#_Toc79952962)

[**Figure 26 Market Segment and Cancellation** 11](#_Toc79952963)

[**Figure 27 Deposit Type and Cancellation** 11](#_Toc79952964)

[**Figure 28 Non-Refund VS Group Bookings** 12](#_Toc79952965)

[Figure 29 Previous Cancellation VS Current Cancellation 12](#_Toc79952966)

[Figure 30 Lead Time VS Cancellation 12](#_Toc79952967)

[Figure 31 Lead Time and Cancellation 13](#_Toc79952968)

[Figure 32 Total Number of Parking space request and Cancellation 13](#_Toc79952969)

[Figure 33 Heat Map of Variables 14](#_Toc79952970)

[**Figure 34 Outliers** 15](#_Toc79952971)

[Figure 35 Outliers Treatment 15](#_Toc79952972)

[Figure 36 Correlation on Target Variable 18](#_Toc79952973)

[Figure 37 Correlation on Predators 19](#_Toc79952974)

[Figure 38 Logistic Regression – Base Model 21](#_Toc79952975)

[**Figure 39 - Logistic Regression – Tuned Model** 21](#_Toc79952976)

[Figure 40 LDA – Base Model 22](#_Toc79952977)

[**Figure 41 - LDA – Tuned Model** 22](#_Toc79952978)

[Figure 42 Random Forest – Base Model 23](#_Toc79952979)

[**Figure 43 – Random Forest – Tuned Model** 23](#_Toc79952980)

[Figure 44 Gaussian NB – Base Model 24](#_Toc79952981)

[**Figure 45 -** Gaussian NB **– Tuned Model** 24](#_Toc79952982)

[Figure 46 KNN – Base Model 25](#_Toc79952983)

[**Figure 47 -** KNN **– Tuned Model** 25](#_Toc79952984)

[Figure 48 ANN – Base Model 26](#_Toc79952985)

[**Figure 49 –** ANN **– Tuned Model** 26](#_Toc79952986)

[Figure 50 XG-Boost – Base Model 27](#_Toc79952987)

[**Figure 51** XG-Boost  **– Tuned Model** 27](#_Toc79952988)

[Figure 52 ADA Boost 28](#_Toc79952989)

[**Figure 53 Gradient Boosting** 28](#_Toc79952990)

[Figure 54 Bagging Classifier 29](#_Toc79952991)

[Figure 55 Model Validation 30](#_Toc79952992)

[Figure 56 Accuracy Variance 30](#_Toc79952993)

[Figure 57 Accuracy VS Recall 31](#_Toc79952994)

**Tableau Public Link for visualizations –**

<https://public.tableau.com/app/profile/sivaramakrishnan3623/viz/Sivaramakrishnan_Capstone_Hotel_Cancelation_EDA_1/SpecialRequestVSTarget>

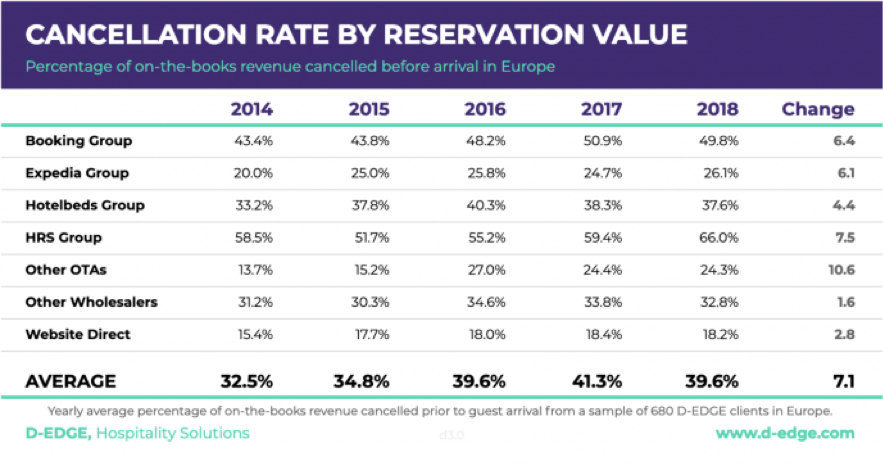
<https://public.tableau.com/app/profile/sivaramakrishnan3623/viz/ProjectNotes2_16271145779440/Test?publish=yes>

# **Solution 1: introduction**

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

## OTAs vs. Direct BookingsIssues 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)](https://triptease.com/blog/the-real-cost-of-free-cancellations/). This results, the customers have become accustomed to free cancellation policies. Based on the [Fornova](https://www.fornova.com/) 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)](https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/)



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)](https://hospitalitytech.com/global-cancellation-rate-hotel-reservations-reaches-40-average)

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)](https://triptease.com/blog/the-real-cost-of-free-cancellations/). 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)](https://www.stayntouch.com/blog/how-online-reviews-impact-hotel-revenue/).

## Understanding Business/Social opportunity

Artificial intelligence is playing an increasingly important role in hospitality management [(source)](https://www.revfine.com/hospitality-management/), 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.

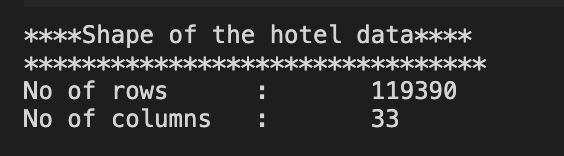


Figure 2 Rows And Columns

### Descriptive Statistics of Numerical Variables

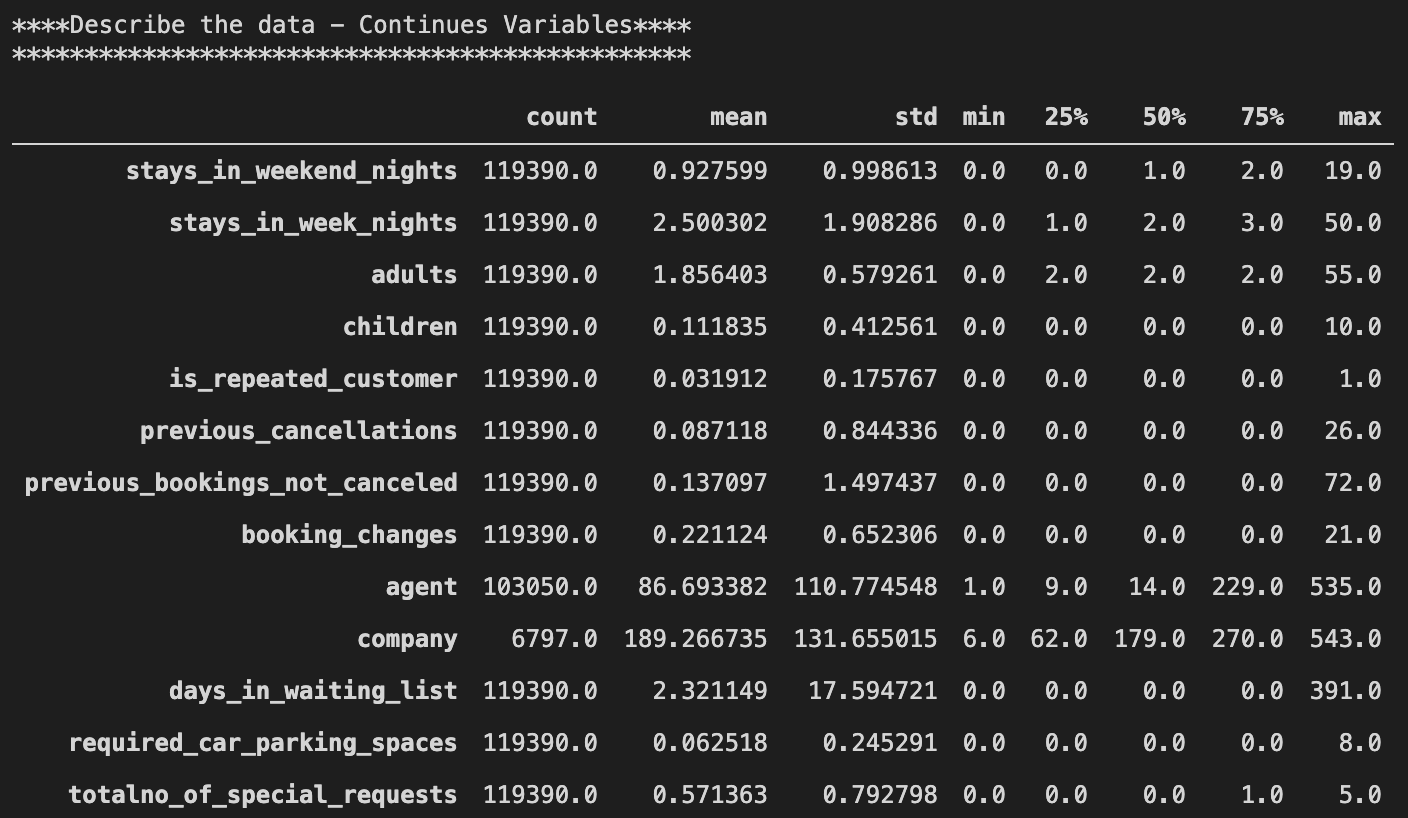
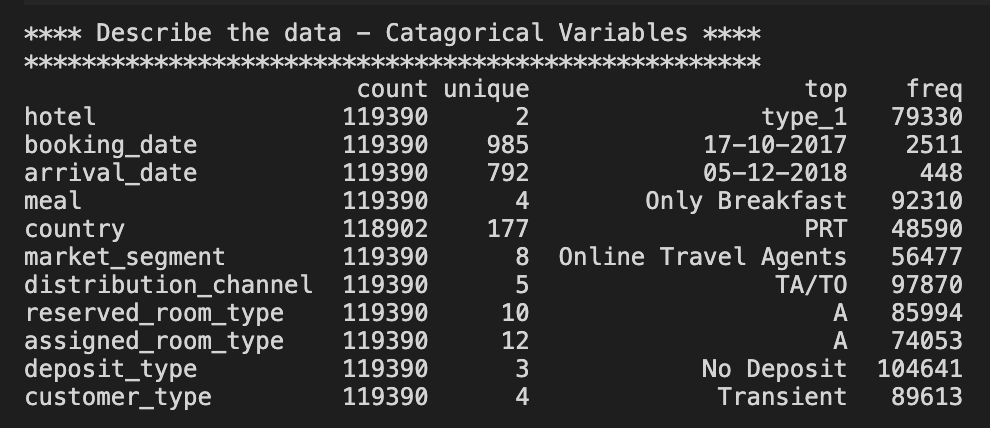


Figure 3 Descriptive Statistics Numerical

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



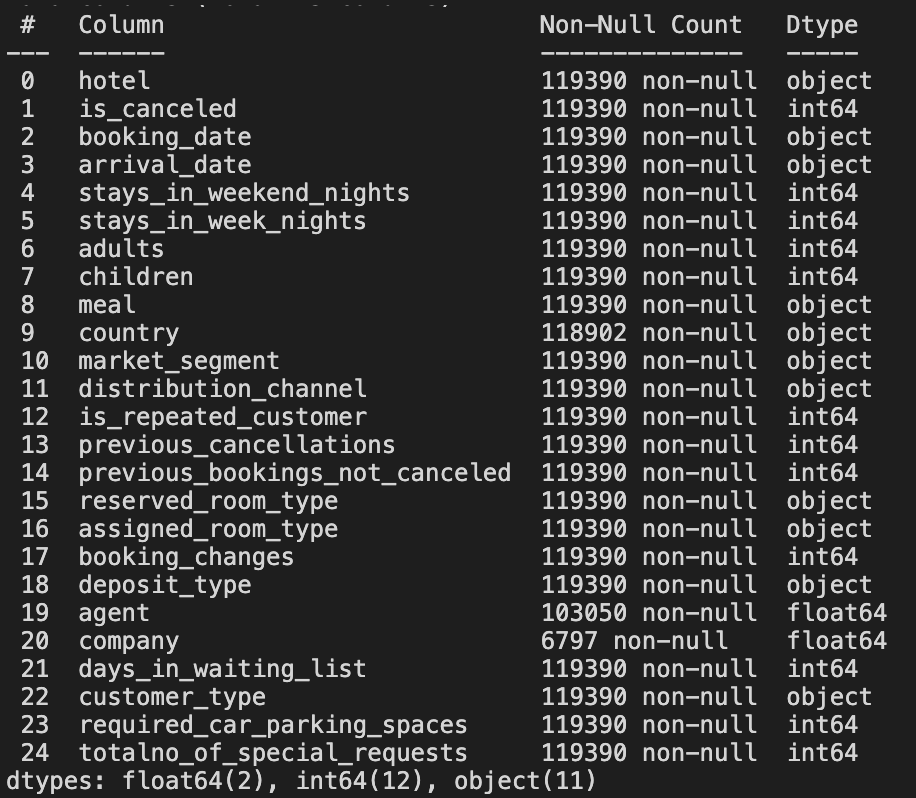
There are higher frequency for the No-Deposit which may lead to the highest cancellation since there is no loss for the customer due to the cancellation.

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



* 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

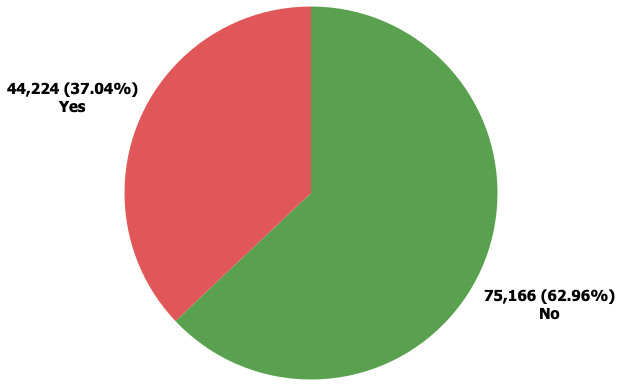
### Duplicate Value Check.

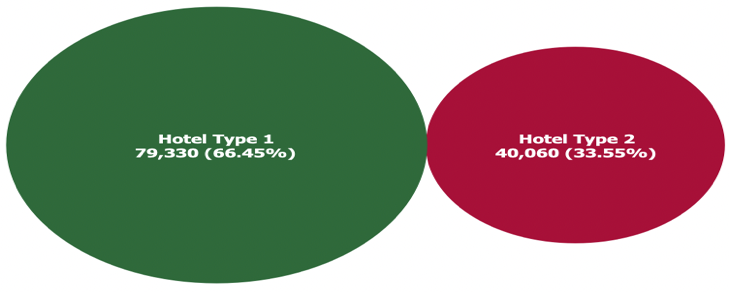
There are 33210 duplicate values in the dataset which is nearly 27.82% of total data this we can drop before the model building.

## 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](https://www.emerchantpay.com/infographic-how-can-hotels-combat-rising-cancellation-rates/" \t "_blank)**
* **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 handing**

**Figure 6 Target Distribution**

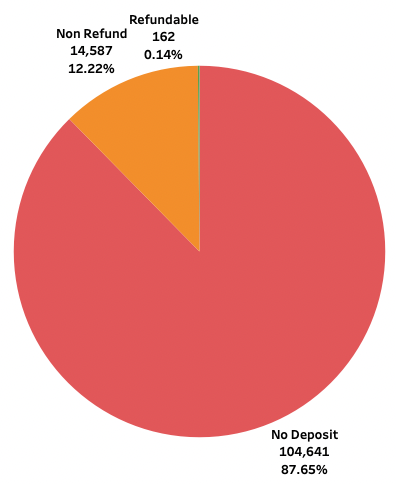
### Hotel Type

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**

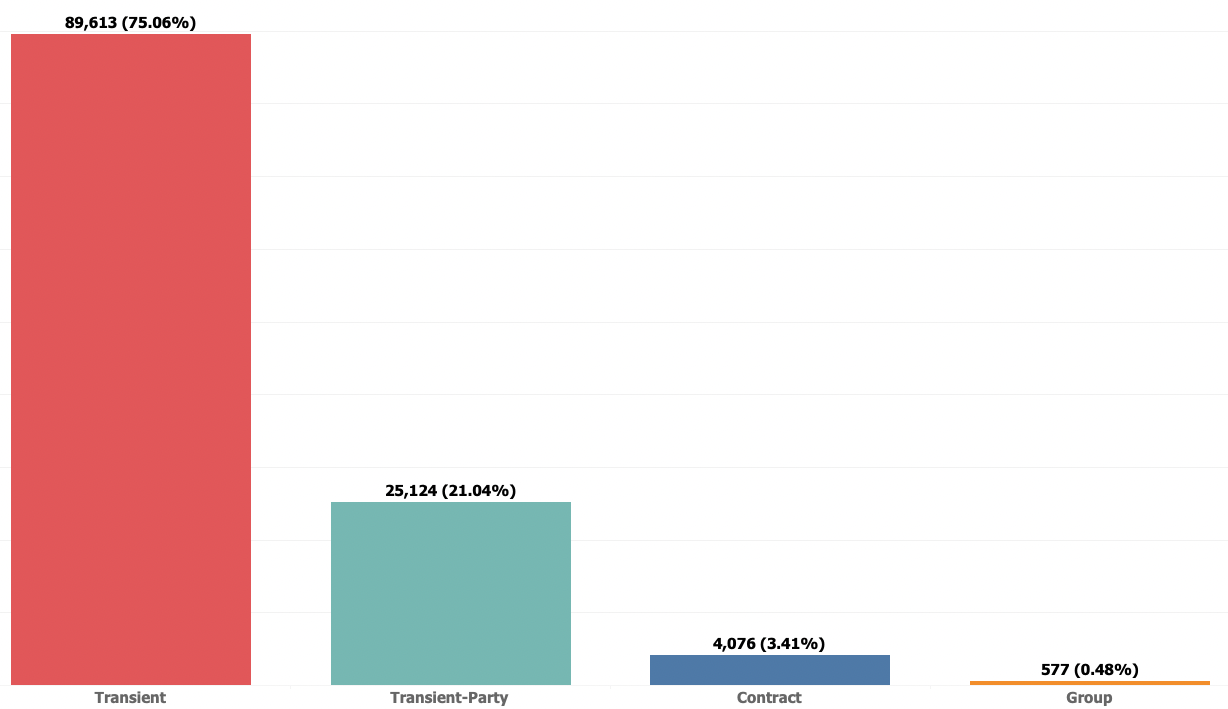
### Deposit Type

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

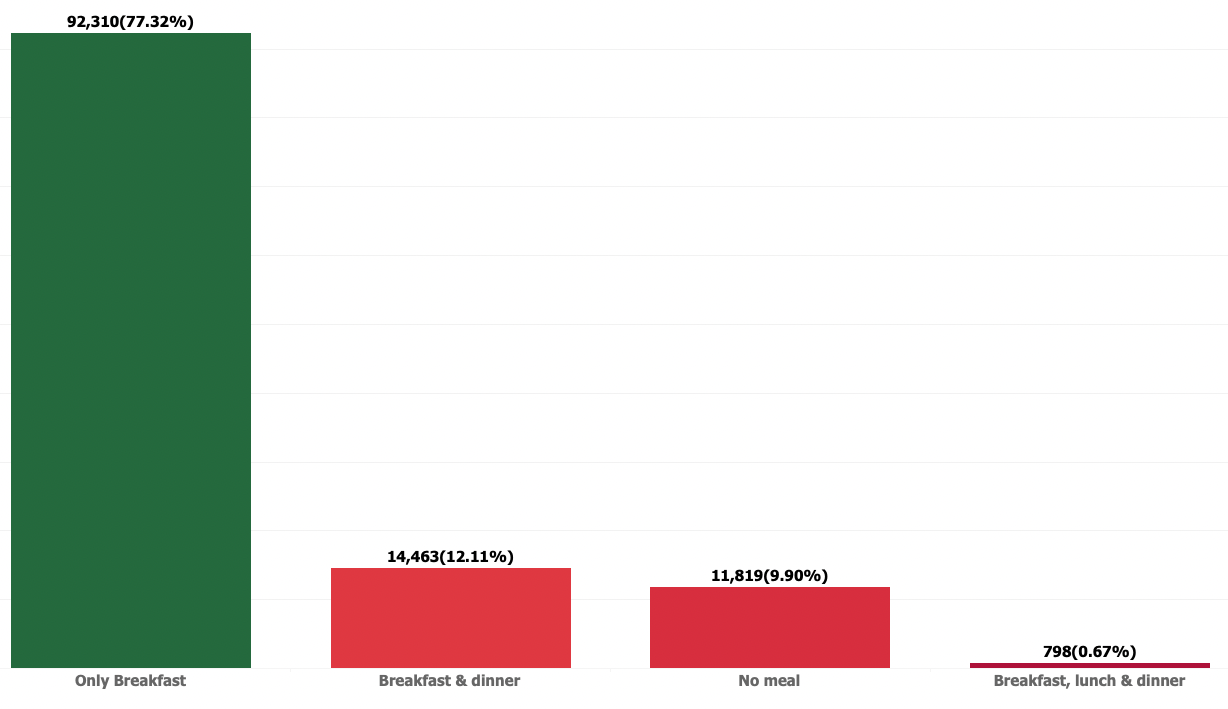
Figure Deposit Type

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**

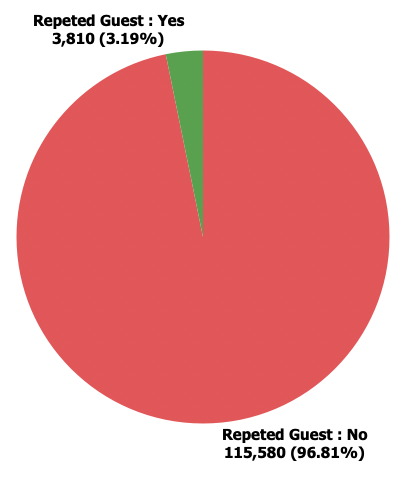
### Meal Type



* 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



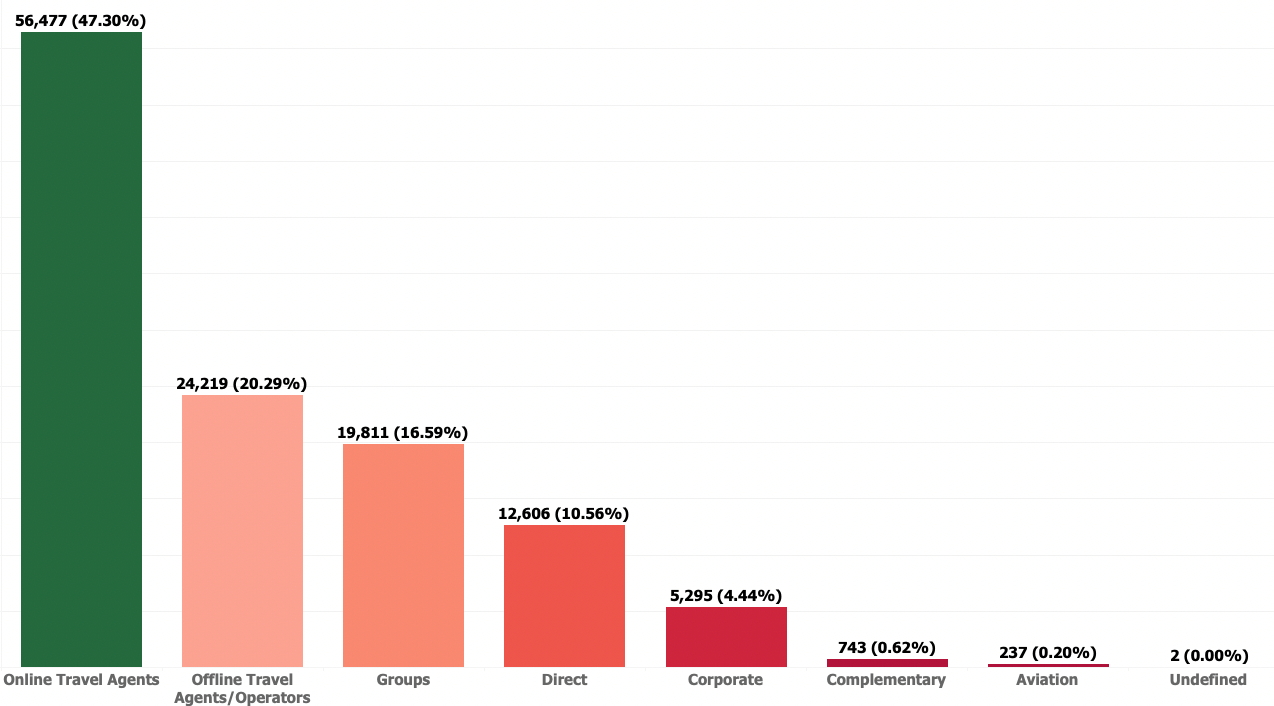
Value indicating if the booking name was from a repeated guest (1) or not (0). Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the Property Management System database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

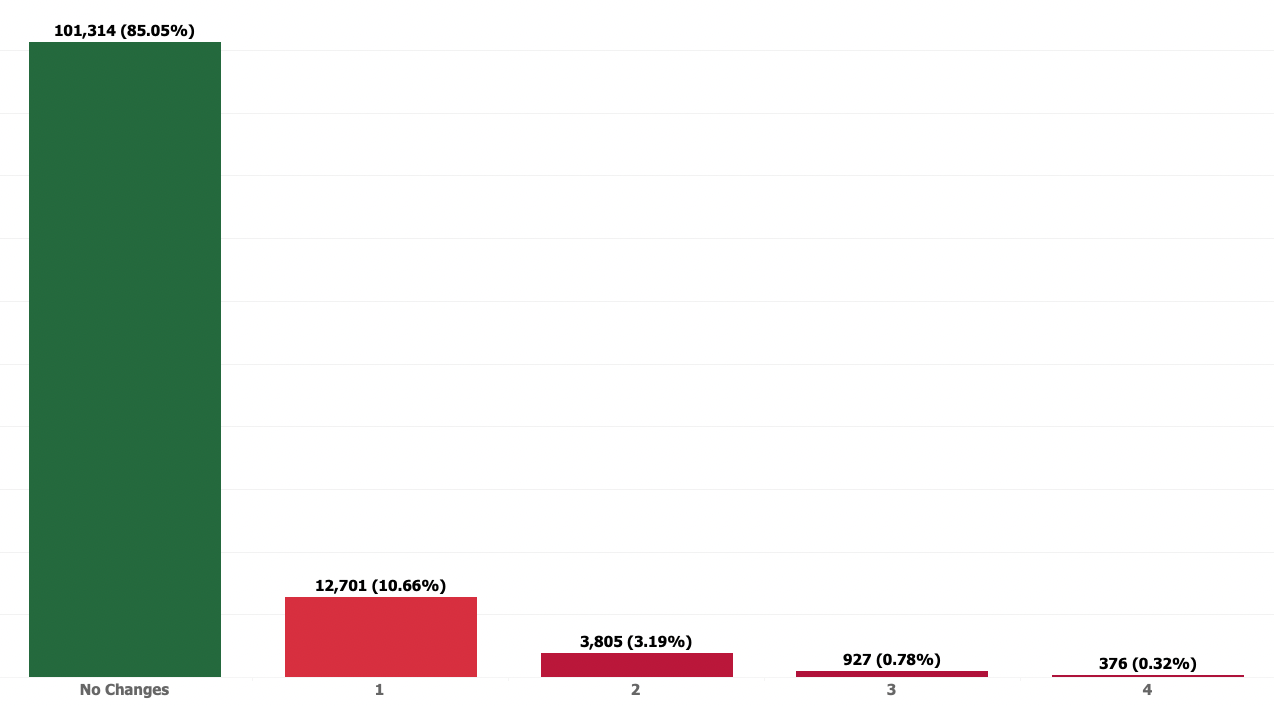


* The TA/TO (Travel Agents / Tour Operators) are the biggest booking channel compared to other channels
* The Undefined data needs to treated with most occurrence value as OTA
* The Distribution channel also have the similar distribution of the data

**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.

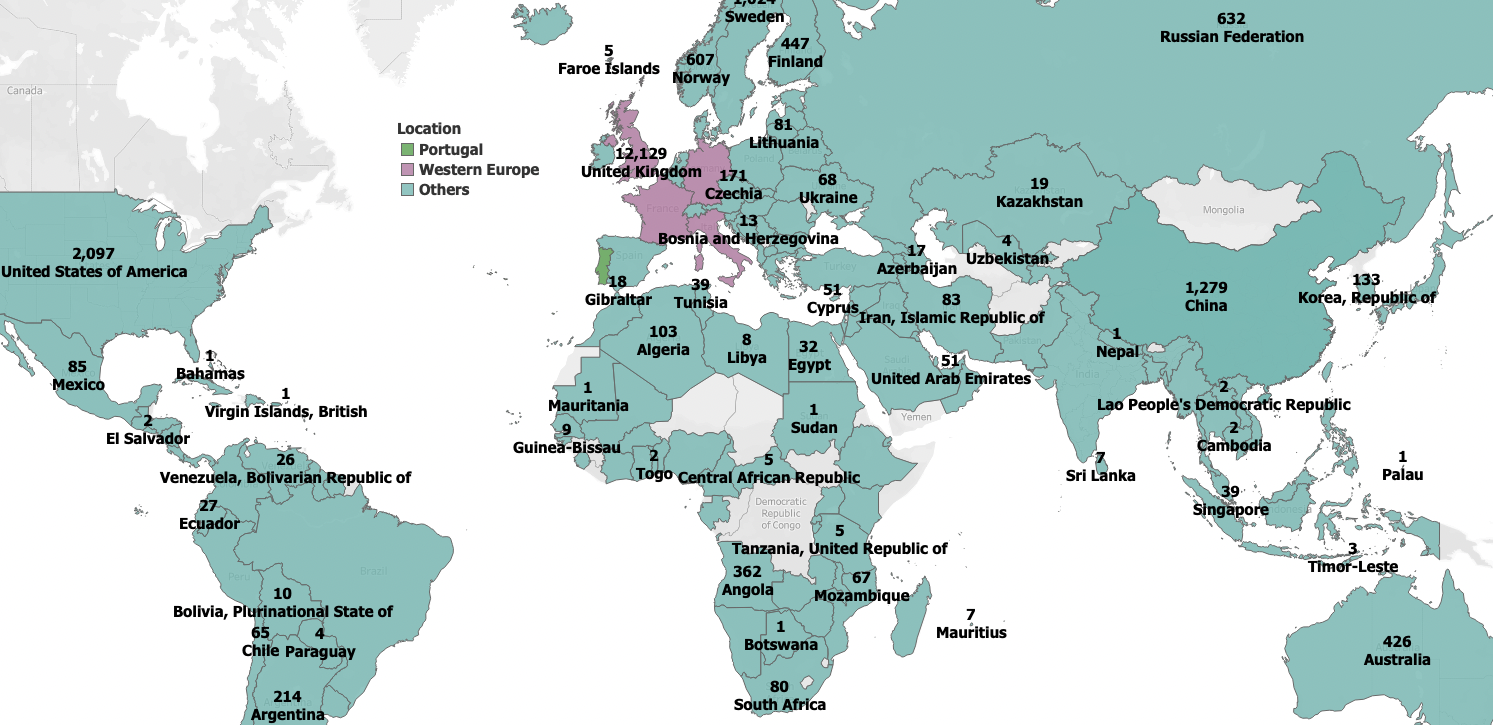
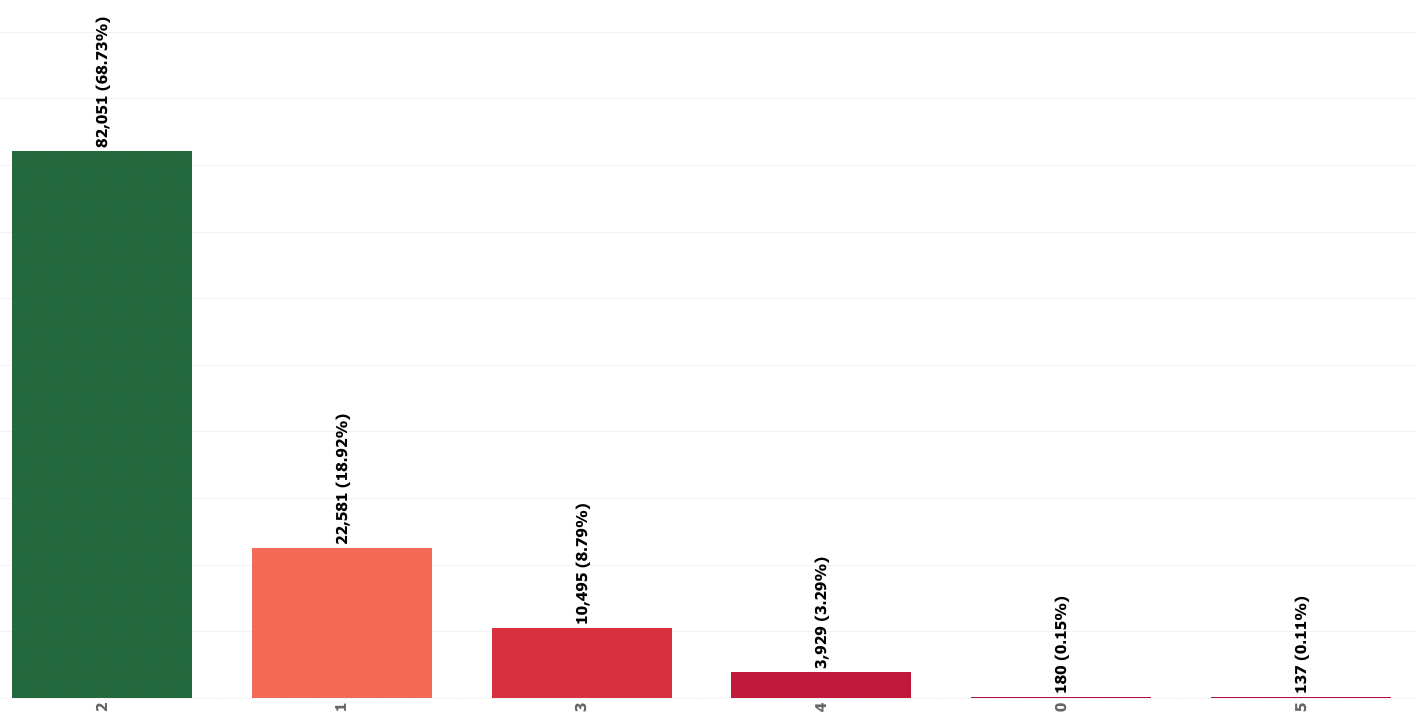


Figure 14 Country

* 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. 

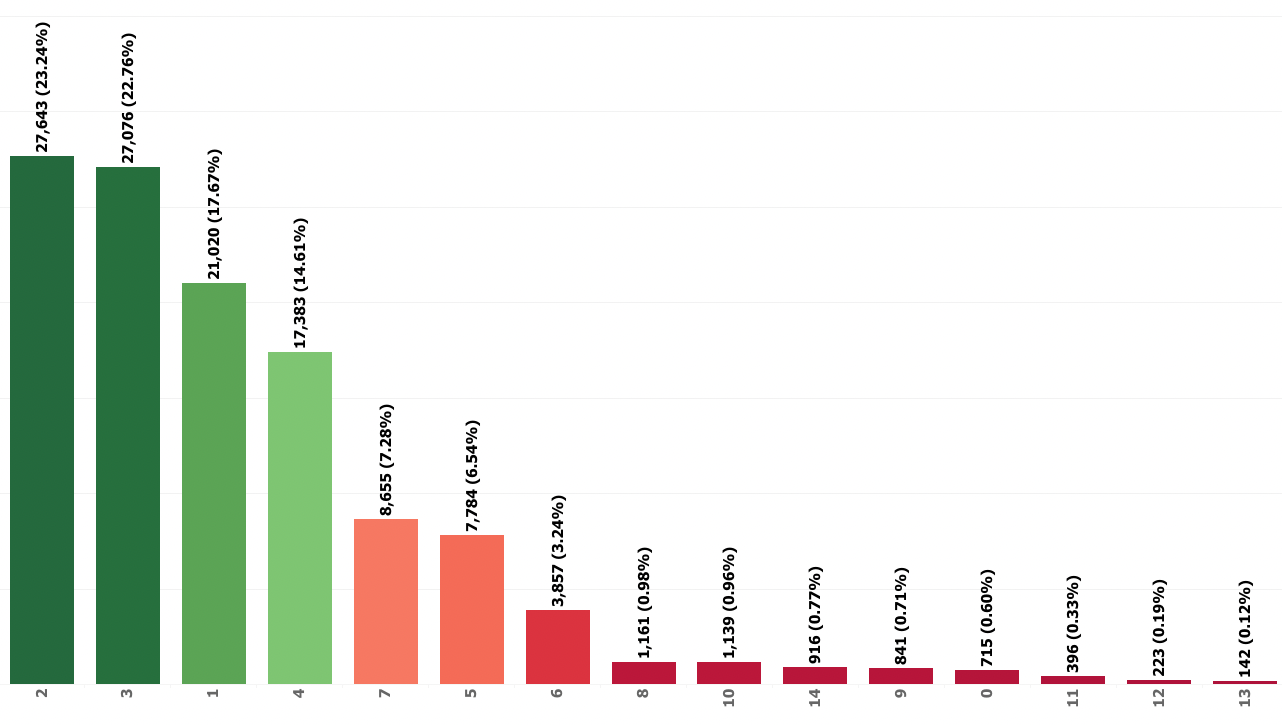
* The Total Guest should not be as 0 this looks like missing data we can drop this values since this is only 0.15% of total data
* The highest count of Total Guest is 2 we assume that mostly couples stayed.

Figure Total Guest

### 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.**



* Total Stays of 2,3,1 and 4 days are the most popular total stays
* There are customers to book hotel on the same day and check out (not an overnight stay) so we will not be dropping the rows that has 0 total stays (not an overnight) stay

Figure 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

* The majority of lead-time booking and arrival on the same month and some bookings made a year or more ahead before the arrival
* We need to check this longer values how affect the cancellation rate

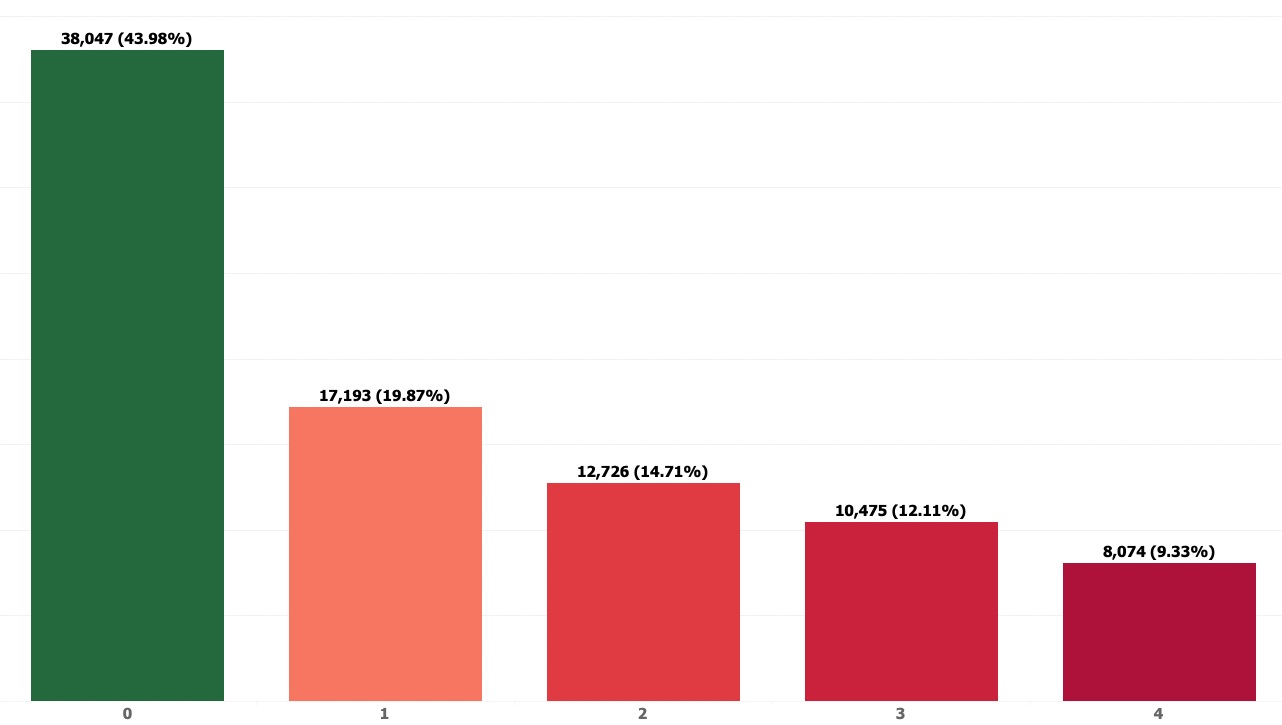
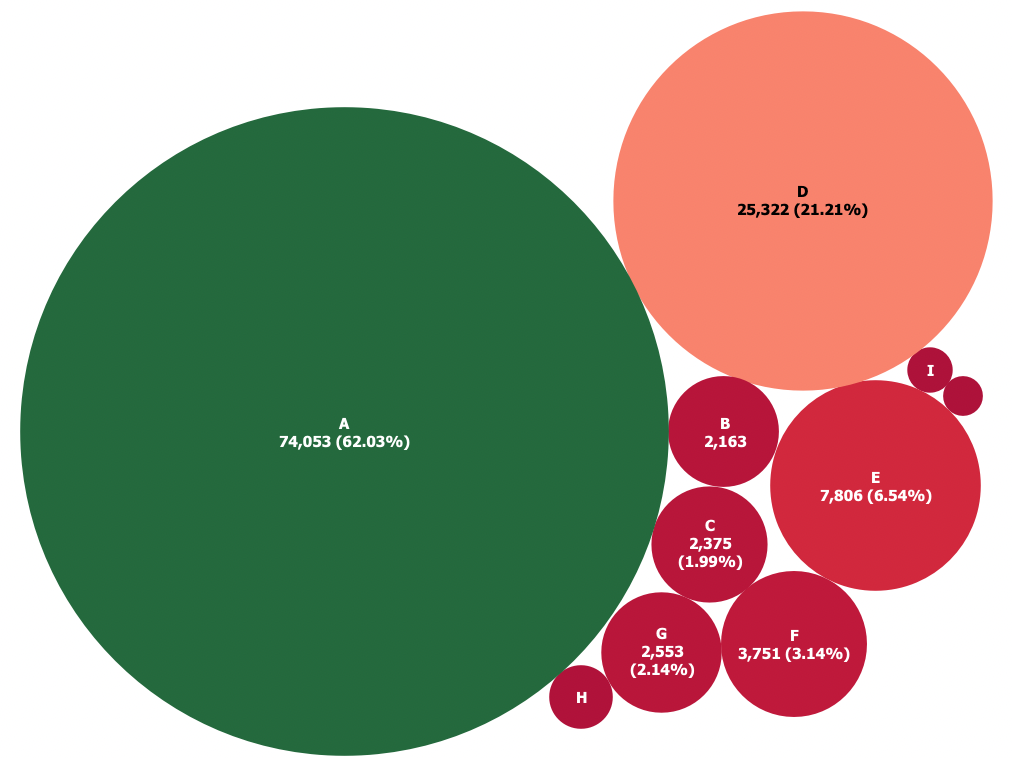


Figure Lead Time

|  |
| --- |
| 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](https://www.usatoday.com/story/travel/hotels/2014/08/14/hotel-booking-tips/14006883/) |
| 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.   Almost 95% of the booking never been cancelled before in this data setWe 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 |
| Required Special Request Number of special requests made by the customer (e.g. twin bed or high floor).     * 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 |

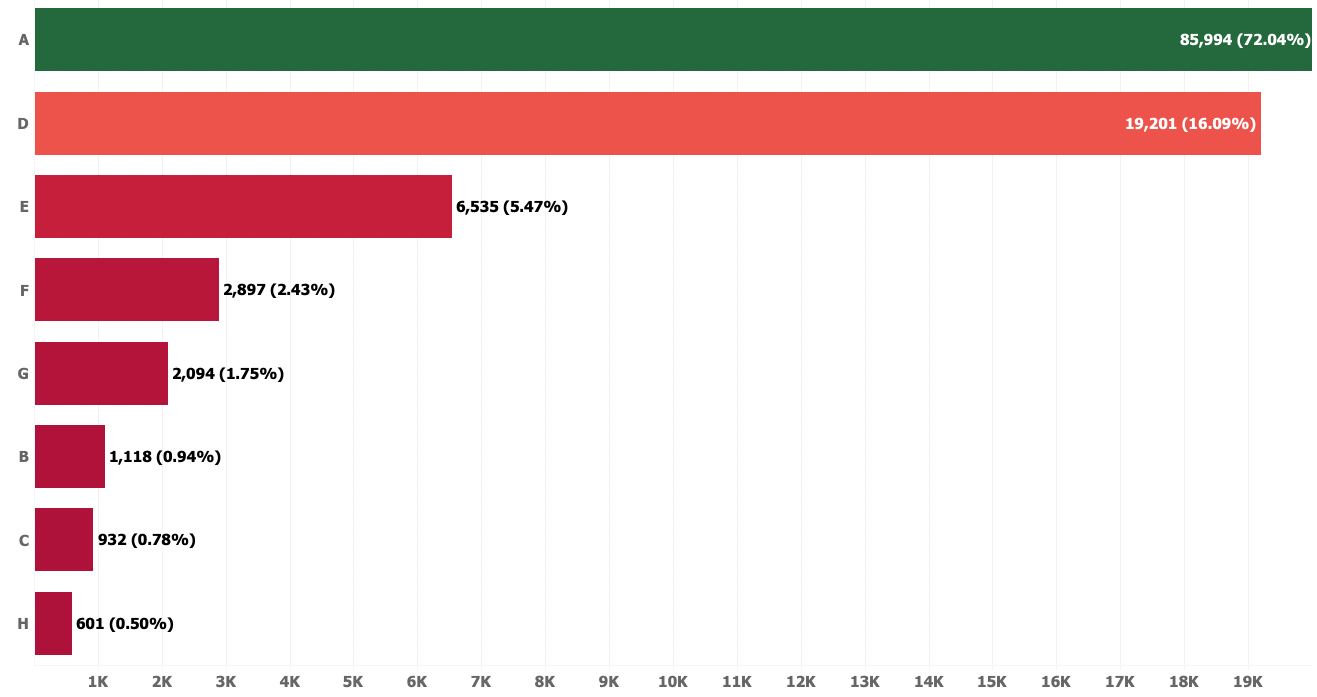
### Assigned Room Type



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

**Figure 18 Assigned Room Type**

Reserved Room Type 

* The Reversed Room Type A is most popular room type since this has a highest number of booking
* The reason for this could be the cheapest room in the both hotel types, we may do later analysis to validate this assumption

**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

* OTA customers are majority on the No Deposit Scheme which needs to further analysis.
* Non Refund made by Groups books this needs further investigation on cancellation rate.

Figure Deposit Type VS Market Segment

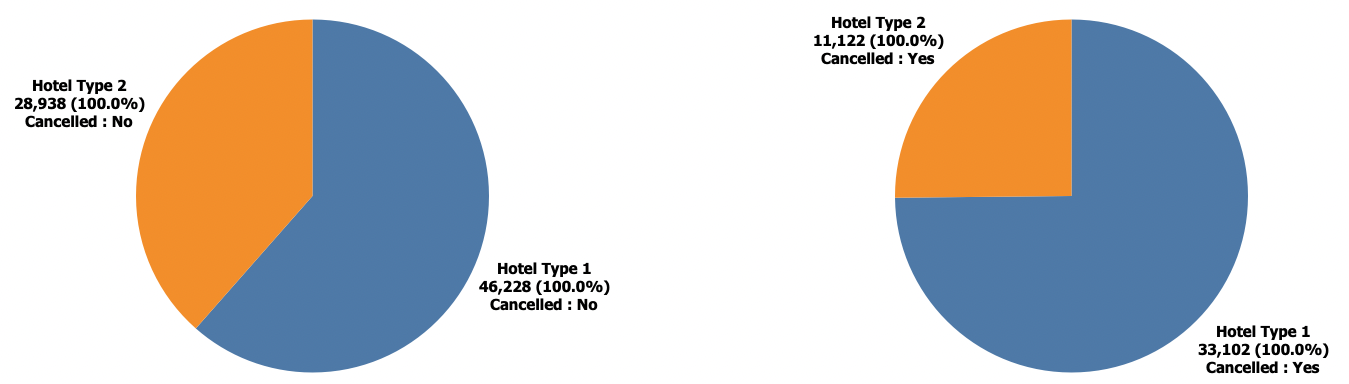
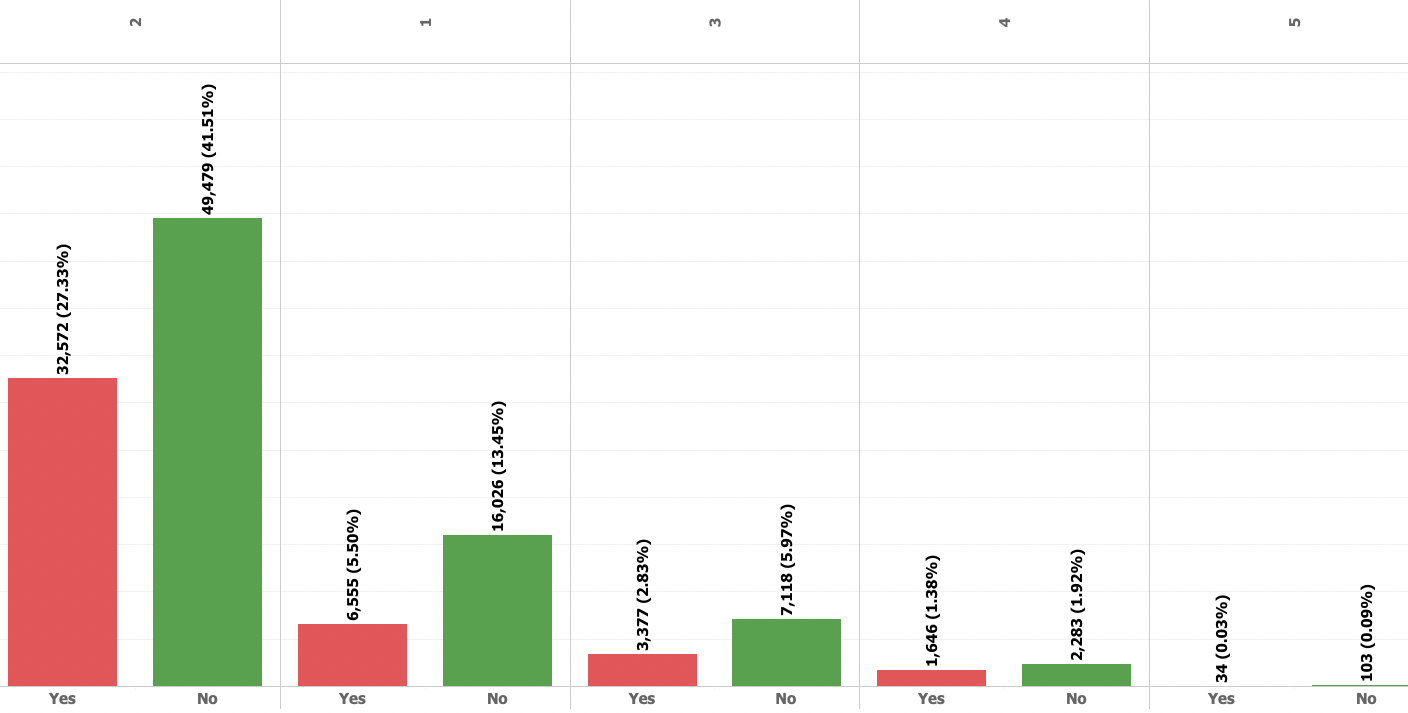
Hotel Type and Cancellation 

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

* 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%

Figure 24 Meal Type VS Cancellation

### Location and Cancellation

**Figure 25 Location VS Cancellation**

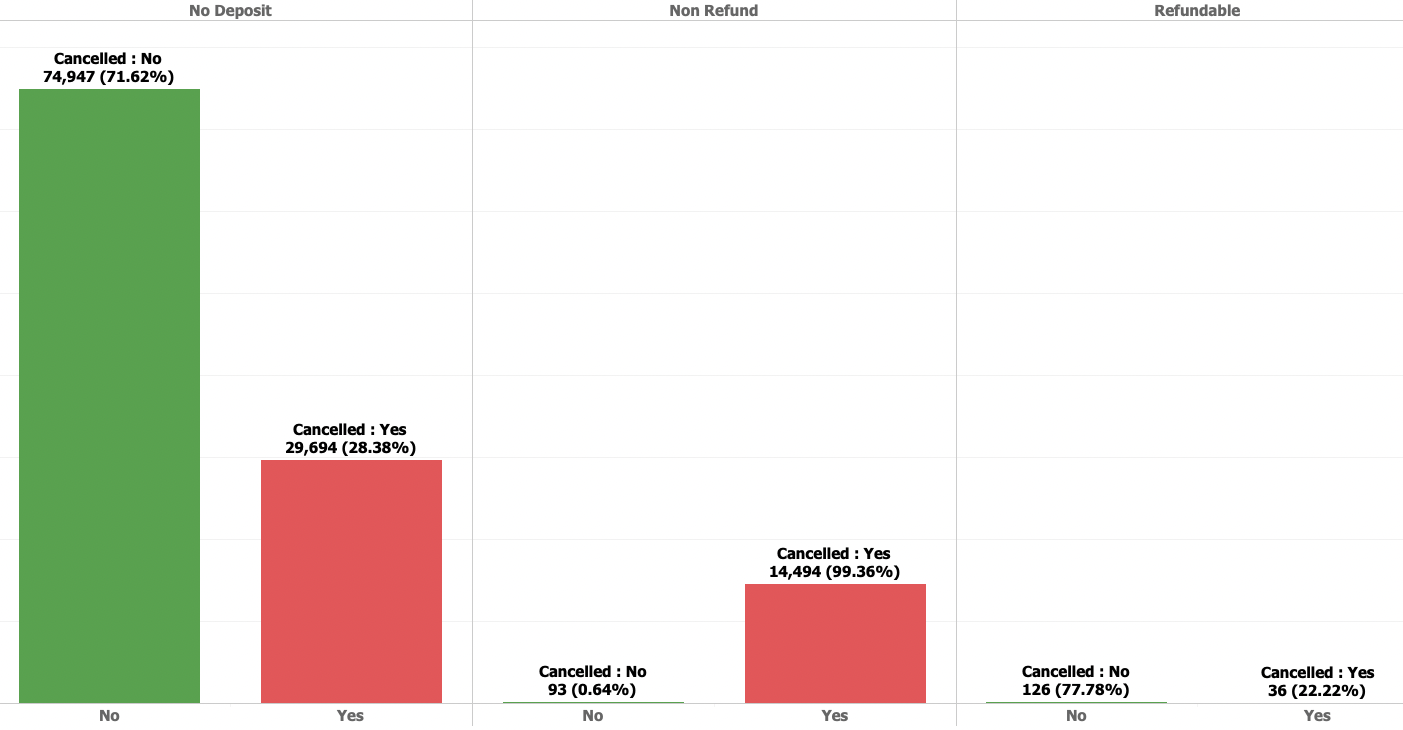
* The booking made in Portugal are almost 2.5 X more likely to be cancelled compared to booking that's made outside Portugal

### 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

* Whoever cancelled the booking before have a high impact 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.

Figure 29 Previous Cancellation VS Current Cancellation

### Lead Time and Cancellation

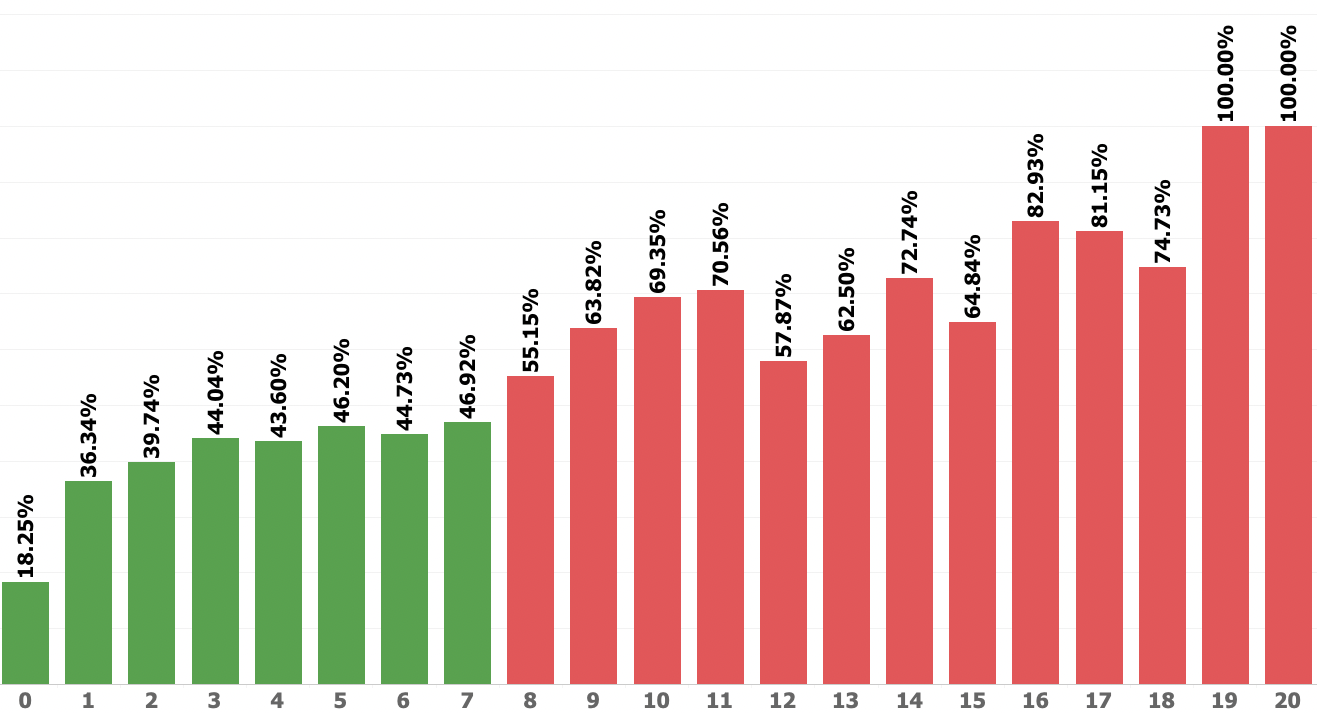


Figure 30 Lead Time VS Cancellation

* Cancelled bookings have a longer lead time on average.
* There are 2 bookings not Cancelled with higher lead time these bookings could 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

* Customers who cancel their bookings make on average fewer special requests. The potential reasons what were discussed above.

Figure 31 Lead Time and Cancellation

### Total Number of Parking space request and Cancellation

* The customers are not cancelled their bookings tend to require more parking spaces.
* There are around **7383 (6.2 %)** that required car parking space(s) that require a parking space **there not a single booking that’s Cancelled (0 Cancellation)**

Figure 32 Total Number of Parking space request and Cancellation

### Multi-Variant Analysis

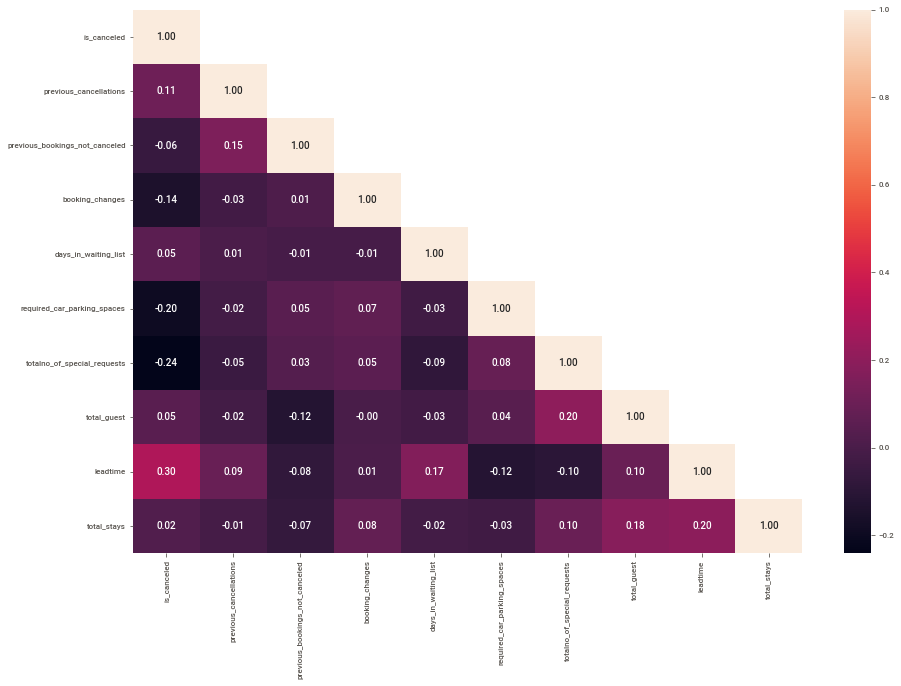


Figure 33 Heat Map of Variables

**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   * 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   * 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](https://statisticsbyjim.com/glossary/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

|  |  |
| --- | --- |
| Outliers | After Outlier Treatment |
|  |  |
|  |  |
|  |  |
|  |  |
| **Figure 34 Outliers** | Figure 35 Outliers Treatment |

### 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 gropping |

### 

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

# 

# 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:

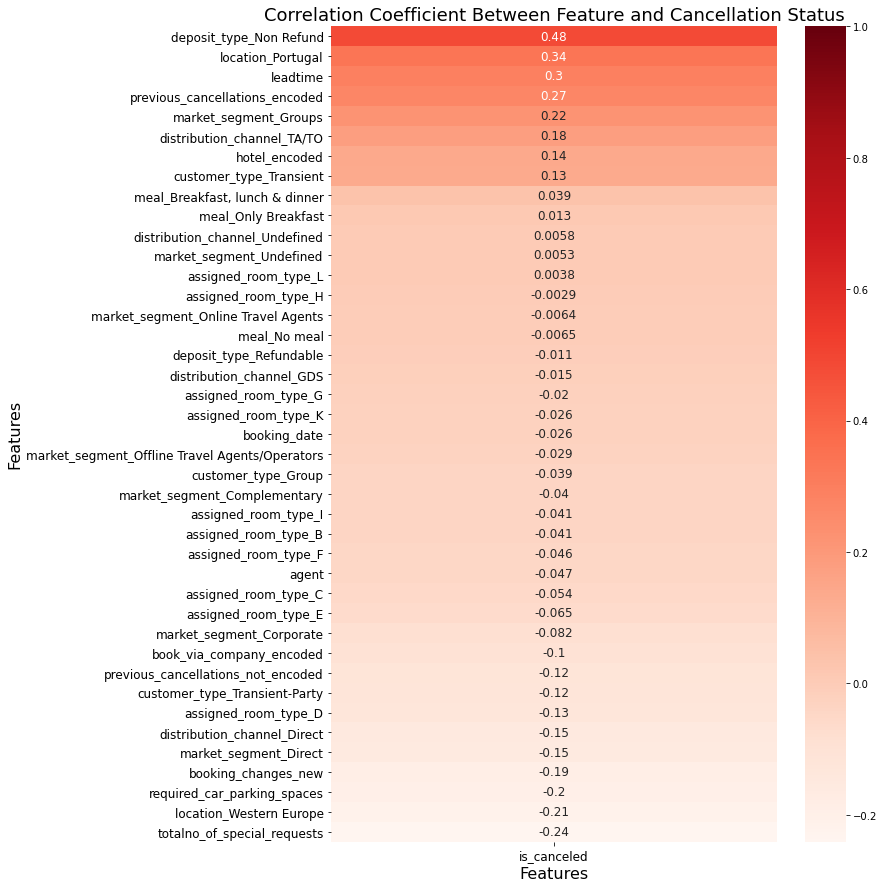


Figure 36 Correlation on Target Variable

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)

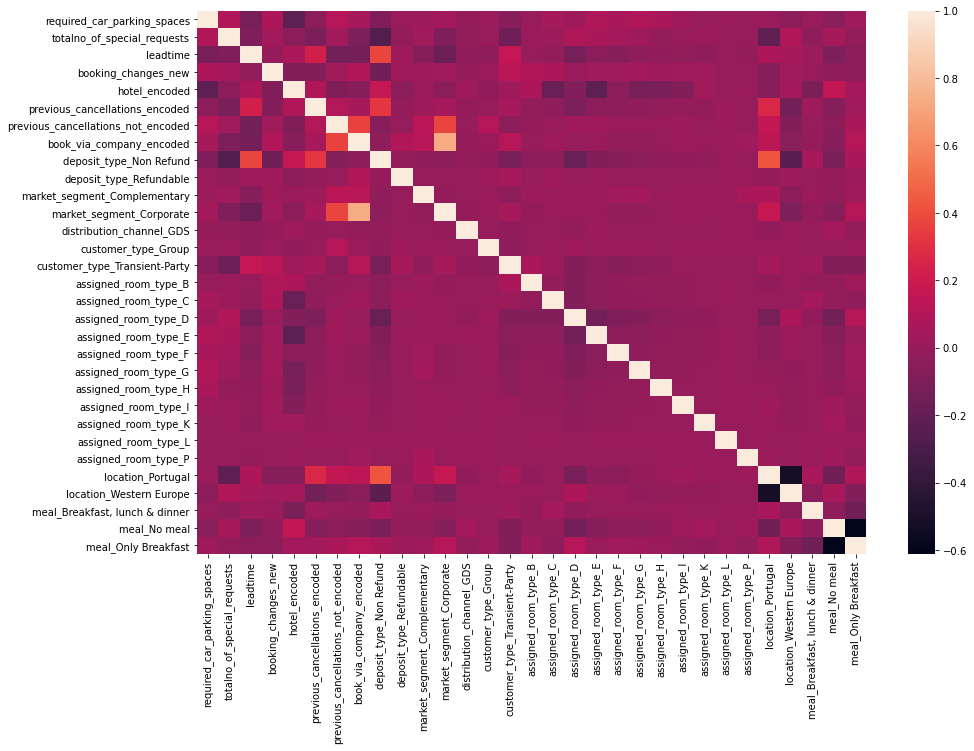


Figure 37 Correlation on Predators

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](https://intellipaat.com/blog/tutorial/machine-learning-tutorial/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 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(p̂) 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](https://whatis.techtarget.com/definition/dependent-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','11','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 |
| Figure 38 Logistic Regression – Base Model | **Figure 39 - Logistic Regression – Tuned Model** |
| Inferences:  * 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:  * 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 |
| Figure 40 LDA – Base Model | **Figure 41 - LDA – Tuned Model** |
| Inferences:  * 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:  * 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:** | **Best Estimators:** | | * max\_depth' : [22,23] * 'max\_features' : [18,19] * 'min\_samples\_leaf' : 3, * 'min\_samples\_split' : 6 * 'n\_estimators' : [50,60] | * max\_depth' : 22 * 'max\_features' : 18 * 'min\_samples\_leaf' : 3 * 'min\_samples\_split' : 6 * 'n\_estimators' : 60 | |
| Figure 42 Random Forest – Base Model | **Figure 43 – Random Forest – Tuned Model** |
| Inferences:  * This model is a overfit model and the training score is greater than testing scores. * The model is performing the baseline with a testing accuracy of 87.72%. | Inferences:  * There is no over fit after tuned * This model has a highest accuracy upon all models. |

### 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.

|  |  |
| --- | --- |
| BASE MODEL | *TUNED MODEL* |
| Default Parameters: priors=None, var\_smoothing=1e-9 | ***Grid Search Parameters:***  'var\_smoothing': np.logspace(0,-9, num=100)  ***Best Parameters:***  var\_smoothing': 1.873817422860383e-05 |
| Figure 44 Gaussian NB – Base Model | **Figure 45 -** Gaussian NB **– 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 78.92% which is less then Logistic Regression & LDA | Inferences:  * There is no much improvement in the tuned model, Still the training and testing models are close together. |

### 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** |
| Figure 46 KNN – Base Model | **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:** | **Best Estimators:** | | * 'hidden\_layer\_sizes' : [200,350], * 'max\_iter' : [500,750], * 'solver' : ['sgd','adam'], * 'tol' : [0.01,0.001], | * hidden\_layer\_sizes : 350, * max\_iter : 500, * random\_state : 9, * tol : 0.001 | |
| Figure 48 ANN – Base Model | **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](https://analyticsindiamag.com/gradient-descent-everything-you-need-to-know-with-implementation-in-python/)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:  |  |  | | --- | --- | | * 'n\_estimators': [18], * 'colsample\_bytree': [0.7], * 'max\_depth': [19], * 'reg\_alpha': [1.9], | * 'min\_child\_weight': [2.5], * 'gamma': [4.2], * 'subsample': [0.9], * 'objective':['binary:hinge']50], | |
| Figure 50 XG-Boost – Base Model | **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.

|  |  |
| --- | --- |
| ADA Boosting | *Gradient Boosting* |
| Parameters:  * n\_estimator : 10, * random\_state : 0, | Parameters:  * random\_state : 0, |
| Figure 52 ADA Boost | **Figure 53 Gradient Boosting** |
| Inferences:  * This model is same not overfit or underfit (the training and testing scores are close together) | Inferences:  * This model performance mostly matching with Random Forest * This model has better accuracy then Ada Boosting but recall reduced. |

### 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](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html" \l "sklearn.tree.DecisionTreeClassifier" \o "sklearn.tree.DecisionTreeClassifier). ***We are going use the RF model for Bagging***

|  |
| --- |
| Bagging Classfier |
| Parameters:  * base\_estimator : rfcl (This is the tuned RF model), * n\_estimator : 50, * random\_state : 0, |
| Figure 54 Bagging Classifier |
| Inferences:  * Since we used base model as a Random Forest so we got the accuracy more similar to that. * There is fall on the recall in the test model. |

## 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 Model Validation





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 VARIENCE 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**

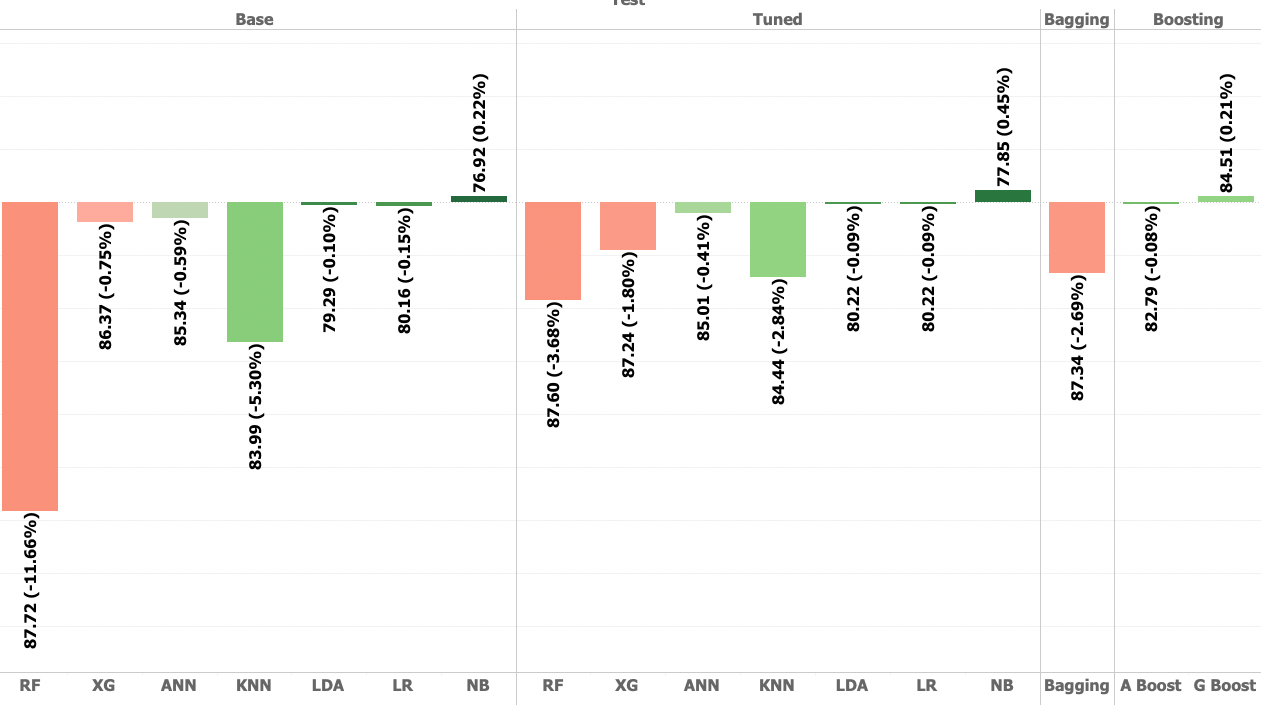


Figure Accuracy Variance

## 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.**

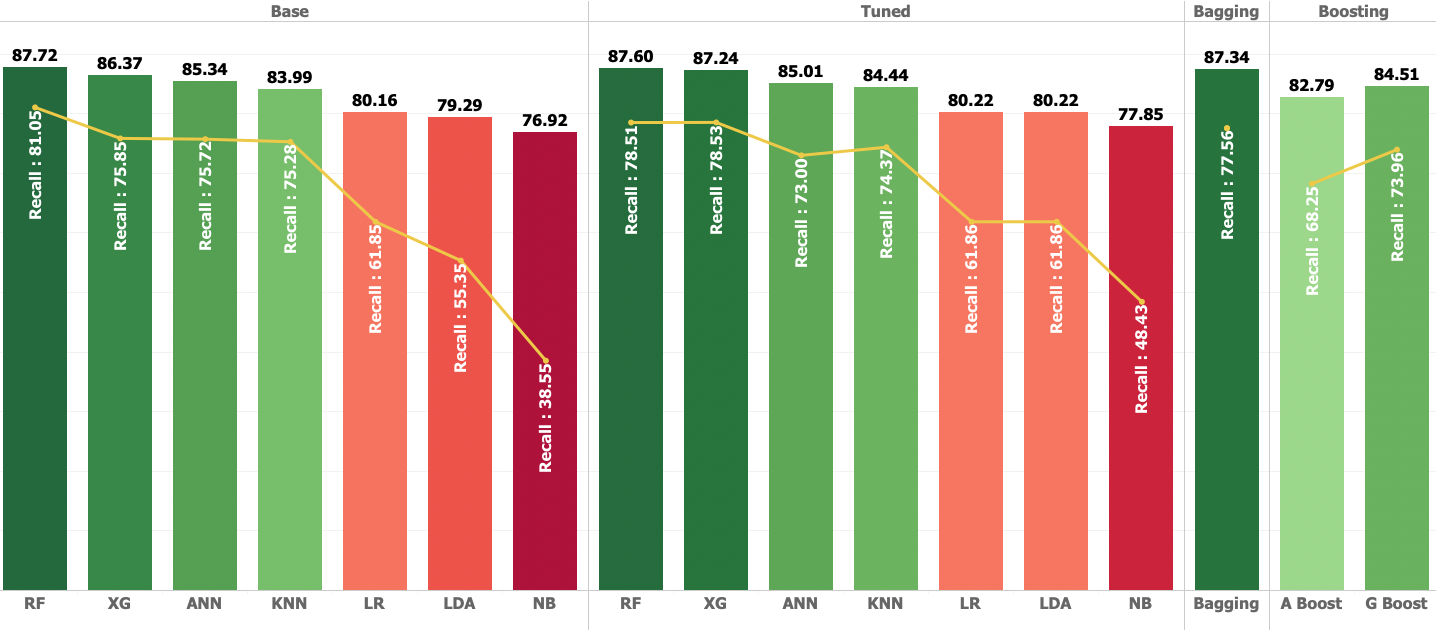
****

Figure 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 Cancelled but classified as Not Cancelled**.
  + **4% (792)** of the Total test data is classified as False Negative which is **Actually Not Cancelled but classified as Cancelled**.

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 Business Insights

|  |  |  |
| --- | --- | --- |
| SNo | Feature Name | Insights |
| 1 | Lead Time | 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   * 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 | There are 3 kinds of deposit type in the data set are **NO Deposit, NO Refund, and Refundable**   * **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%**   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?** |
| 3 | Market Segment | * 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 %**   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) |
| 4 | Location | 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)**   * 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 | * 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 | 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).   * **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 | 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   * **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

1. 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.

1. 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](https://www.ezeeabsolute.com/blog/increase-direct-hotel-bookings/" \t "_blank)

