greatlearning
Learning for Life

GREAT LAKES
INSTITUTE OF MANAGEMENT

Capstone Project – Report

Hotel Booking Cancellation Prediction

Domain – Revenue Management

Mentored by,

Ms. Akhila Gurre

Submitted by,

Group – 4

Anerudh Narayanan B

Nilesh Kumar Pandey A

Radhika SB

Yaswanth Kumar M

Yuvashree P

ACKNOWLEDGEMENT

We would like to thank our mentor **MS AKHILA NAGA SAI GURRE**, Data Science Consultant, SYNITI, Bangalore for providing her valuable guidance and suggestions for our Project work. We also thank her for the continuous encouragement and the interest shown towards us to complete our Project work.

We are extremely grateful to our all teaching and non-teaching staff members of **GREAT LEARNING**, who showed keen interest and inquired our developments.

We would like to express our Gratitude to all teaching and non-teaching staff members of **Great Lakes Institute of Management**, for providing their support and guidance for our project.

We greatly admire and acknowledge the constant support we received from our friends and team members for all the effort and hard work that they have put into completing this project.

CONTENTS

ABSTRACT

This data article describes two datasets with hotel demand data. One of the hotels is a resort hotel and the other is a city hotel. Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. These datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

PROBLEM STATEMENT

1. What would you achieve by this project ?

From a business point of view, it can be very helpful if one has an idea which customer is more likely to cancel their reservation. The hotel can then devise a plan as to how they can make sure the customer doesn't cancel their reservation, and even if they do, how to not let it affect the hotel's revenue.

Here, we use a dataset which contains more than one lakh records of hotel booking demand data. Each observation represents a hotel booking. the datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. We will choose the best classification model based on different metrics and aim to build a predictive model using hotel booking demand dataset to predict the cancellation of bookings and reduce the rate of losses in hotels.

2. How would this help the business or clients ?

We will use machine learning algorithms pertaining to classification, to determine whether the client will cancel their booking or not. From a business perspective, a model like this might not seem to be explicitly useful, since the predicted category should not affect decision of cancellation from the side of company, agent or hotel itself. On the other hand, use of explainable machine learning might give us the most important factors of cancellation process and use them to reduce rate of cancelled reservations.

3. Limitation of the project

Our model is not a generalized model but a business oriented, realistic and specific model – it pertains only to the hotel management industry, and not the hospitality industry as a whole.

1. INTRODUCTION

In the hospitality industry, booking cancellations have significant impact on demand-management decisions. They limit the production of accurate forecasts, a critical tool in terms of revenue management performance. To mitigate the difficulties caused by cancellations, hotels implement rigid cancellation policies and overbooking strategies, which later can generate a negative impact on revenue and social reputation, as well as damage the hotel business performance. Overbooking forces the hotel to deny service provision, which can be a terrible experience for the customer and have a negative effect on both the hotel's reputation and immediate revenue. It can also mean future revenue loss from discontent customers who will not book again at the same hotel. On the other hand, rigid cancellation policies, especially non-refundable policies, have the potential not only to reduce the number of bookings but also to diminish revenue due to the application of significant discounts on price.

1.1 NEED FOR THE PROJECT

Research has demonstrated that using machine learning, statistics, data mining and data visualization, it is possible to predict accurate hotel booking cancellations rates, with results surpassing expectations. By identifying which bookings are likely to be cancelled, revenue managers and other members of the hotel's staff can take measures to avoid potential cancellations such as offering services, room upgrades, discounts, entrances to shows/amusement parks, or other perks.

1.2 OBJECTIVE OF THE PROJECT

Main objective of this project is to find the features/variables that affect the hotel bookings. The aim of this project is to understand the data and to build a model which can predict the cancellation chances, thus optimizing the revenue management of the hotel.

1.3 SCOPE

From a business point of view, it can be very helpful if one has an idea which customer is more likely to cancel their reservation. The hotel can then devise a plan as to how they can make sure the customer doesn't cancel their reservation, and even if they do, how to not let it affect the hotel's revenue. Thus, this project is vital from a revenue optimization standpoint.

1.4 TOOLS

- Programming Language: Python
- Visualisation: Python and Tableau

2. DATA DESCRIPTION

2.1 DESCRIPTION

The Dataset contains 1,19,391 observations of 32 variables, where each observation corresponds to a particular Booking detail.

The 32 variables in the dataset consists of 15 categorical variables, 16 numerical variables and a column of datetime.

VARIABLES	DESCRIPTION	TYPE
ADR	Average Daily Rate	Numerical
Adults	Number of adults	Numerical
Agent	ID of the travel agency that made the booking	Categorical
ArrivalDateDayOfMonth	Day of the month of the arrival date	Numerical
ArrivalDate Month	Month of arrival date with 12 categories: “January” to “December”	Categorical
ArrivalDate WeekNumber	Week number of the arrival date	Numerical
ArrivalDate Year	Year of arrival date	Numerical
AssignedRoom 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	Categorical
Babies	Number of babies	Numerical
Booking Changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation	Numerical
Children	Number of children	Numerical
	ID of the company/entity that made the booking or responsible for	

Company	paying the booking. ID is presented instead of designation for anonymity reasons	Categorical
Country	Country of origin	Categorical
CustomerType	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking	Categorical
DaysInWaiting List	Number of days the booking was in the waiting list before it was confirmed to the customer	Numerical
DepositType	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made. Non Refund – a deposit was made in the value of the total stay cost. Refundable – a deposit was made with a value under the total cost of stay.	Categorical
Distribution Channel	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.	Categorical
IsCanceled	Value indicating if the booking was canceled (1) or not (0)	Categorical
IsRepeated Guest	Value indicating if the booking name was from a repeated guest (1) or not (0)	Categorical
LeadTime	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	Numerical
Market	Market segment designation. In categories, the term “TA” means	Categorical

Segment	“Travel Agents” and “TO” means “Tour Operators”.	
Meal	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package BB – Bed & Breakfast HB – Half board (breakfast and one other meal – usually dinner) FB – Full board (breakfast, lunch and dinner).	Categorical
PreviousBookings NotCanceled	Number of previous bookings not cancelled by the customer prior to the current booking.	Numerical
Previous Cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking.	Numerical
RequiredCar ParkingSpaces	Number of car parking spaces required by the customer.	Numerical
Reservation Status	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why.	Categorical
Reservation StatusDate	Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel	DateTime
ReservedRoom Type	Code of room type reserved. Code is presented instead of designation for anonymity reasons.	Categorical
StaysInWeekndNights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.	Numerical
StaysInWeek Nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.	Numerical
TotalOfSpecial Requests	Number of special requests made by the customer (e.g. twin bed or high floor).	Numerical

Table 1 – Variables Description

VARIABLE CATEGORIZATION:

No. of Categorical Variables	15
No. of Numerical Variables	16
No. of DateTime Variables	1

Table 2 – Variable Categorization

2.2 SOURCE

The Dataset was downloaded from KAGGLE, the file downloaded was in CSV format.

3. LITERATURE SURVEY

One of the most important things which differentiates the hotel industry from other hospitality industries is advance booking. Advance booking information includes valuable insights on demand prospects, changing trends, booking patterns, etc. Therefore, models which capture the characteristics of advance bookings have always played vital roles in hotel demand forecasts.

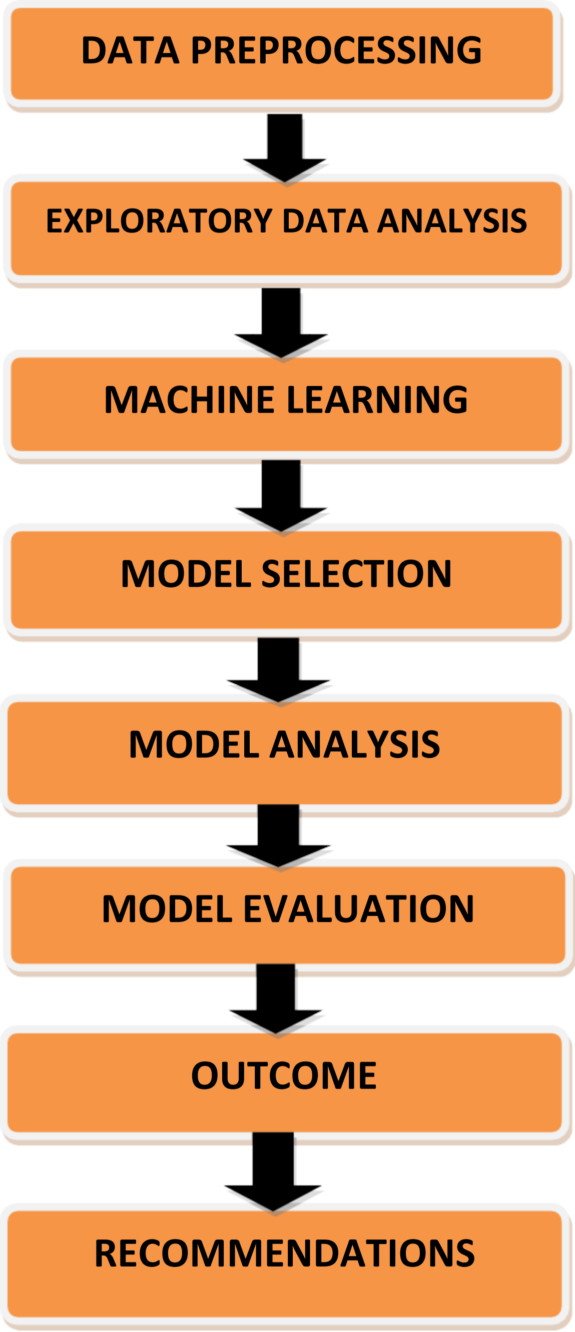
Many hotels actually have huge amounts of historical data available, but they are barely using it to its full potential. Hoteliers will base forecasts on a similar period of last year and average those results. This helps them to predict occupancy. However, this does not help them with more detailed forecasting. For example, if a hotel has 90% occupancy but the 10% of empty rooms are family suites, that impacts the number of guests in the hotel. This has significant, forecastable, effects on breakfast and dining, as well as things such as how many housekeepers are needed. Information like this is rarely used to the hotel's advantage.
Many research attempts to explore the additional effect exogenous to the system, such as local events, weather change, unemployment rate, etc. Schwartz et al. (2016) include hotel competitive set's predicted occupancy as an input of the daily occupancy forecasting. They randomly generate hotel occupancy data for the target hotel and hotels in the competitive set, and use an evolutionary algorithm to reach the lowest forecast error.

With all these various factors at play, it is critical to make accurate forecasts so that hoteliers can leverage demand 'peaks' and handle the 'dips' through smart staff allocation and optimizing hotel operations, while still leaving adequate time to make adjustments to the workforce if necessary.

With machine learning, labour demand forecasts are more accurate than ever. Machine learning models have proved their capabilities in forecasting in the last decade (Ahmed et al., 2010). These intelligent algorithms learn from past data to continuously improve the quality of the results and they open up more possibilities to scale and complete automation.

The hotel was historically not one of the industries considered to be at the forefront of technological innovation (GlobalData, 2017). There has been little discussion on applying machine learning approaches in the hotel industry. Most of the applications of machine learning techniques in the industry focus on hotel online review analysis. Gradually they started moving towards the bookings and the cancellations since that made a huge impact on the revenue and growth of the industry as a whole. They construct a function to respond to customers reservation requests and provide alternate solutions. In summary, there remains a paucity of research on applying machine learning techniques on demand forecasting in hotel settings.

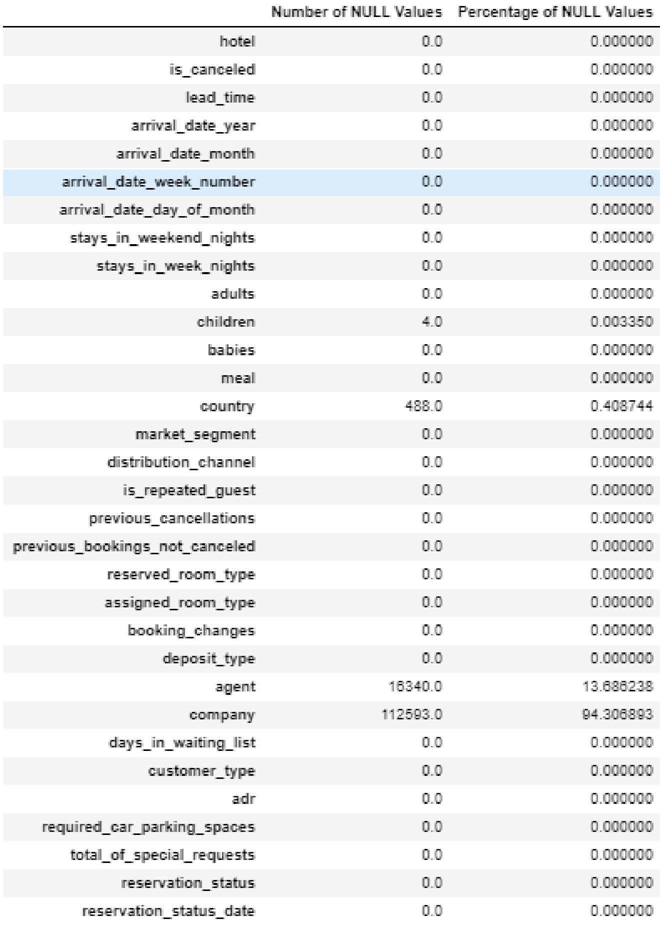
4. PROJECT METHODOLOGY



5. DATA PRE-PROCESSING

5.1. Missing/Null Values

There are Null Values present in the Dataset in some particular Variables.



	Number of NULL Values	Percentage of NULL Values
hotel	0.0	0.000000
is_canceled	0.0	0.000000
lead_time	0.0	0.000000
arrival_date_year	0.0	0.000000
arrival_date_month	0.0	0.000000
arrival_date_week_number	0.0	0.000000
arrival_date_day_of_month	0.0	0.000000
stays_in_weekend_nights	0.0	0.000000
stays_in_week_nights	0.0	0.000000
adults	0.0	0.000000
children	4.0	0.003350
babies	0.0	0.000000
meal	0.0	0.000000
country	488.0	0.408744
market_segment	0.0	0.000000
distribution_channel	0.0	0.000000
is_repeated_guest	0.0	0.000000
previous_cancellations	0.0	0.000000
previous_bookings_not_canceled	0.0	0.000000
reserved_room_type	0.0	0.000000
assigned_room_type	0.0	0.000000
booking_changes	0.0	0.000000
deposit_type	0.0	0.000000
agent	16340.0	13.686238
company	112593.0	94.306893
days_in_waiting_list	0.0	0.000000
customer_type	0.0	0.000000
adr	0.0	0.000000
required_car_parking_spaces	0.0	0.000000
total_of_special_requests	0.0	0.000000
reservation_status	0.0	0.000000
reservation_status_date	0.0	0.000000

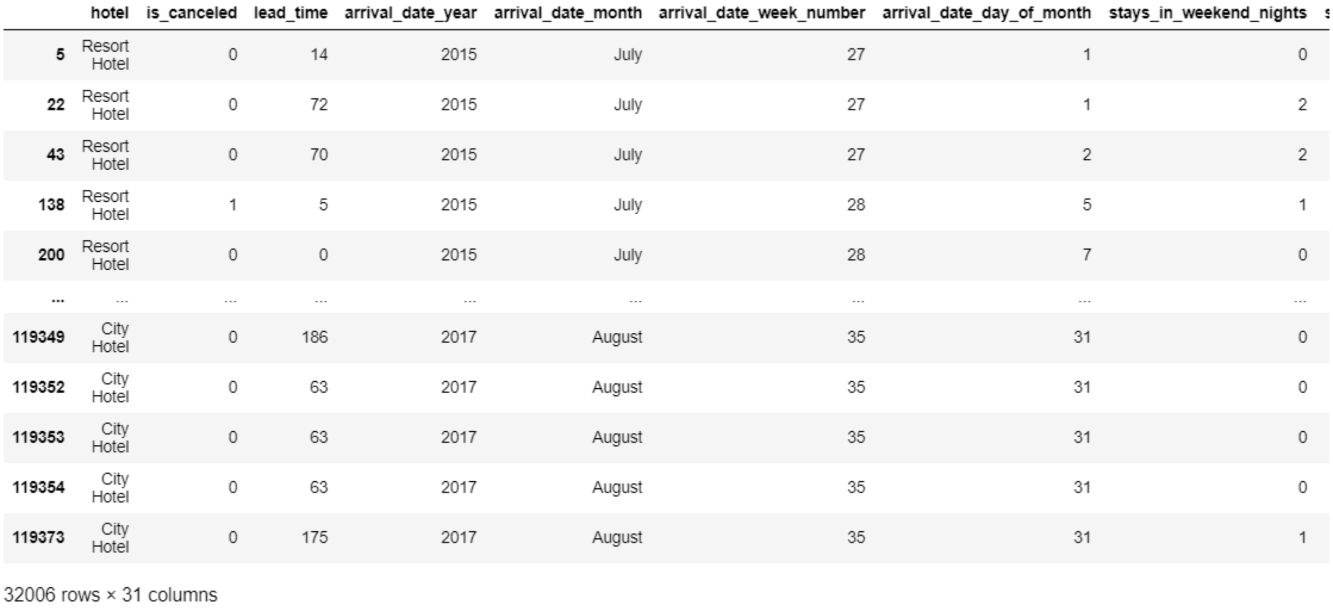
- The Null values in 'Children', 'Country', 'Agent', and 'Company' are treated as shown in the below Table no.3

Variables	No. of Missing/Null Values	Action taken
Children	4	Treated with Mode
Country	488	Treated with Mode
Agent	16340	Filled with Zero (0)
Company	112593	Column Dropped

Table-3 Processing of Null/Missing values

5.2. Duplicate Rows

There are 32006 Duplicate rows in the dataset which comprises similar values in all the variables.



	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights
5	Resort Hotel	0	14	2015	July	27	1	0
22	Resort Hotel	0	72	2015	July	27	1	2
43	Resort Hotel	0	70	2015	July	27	2	2
138	Resort Hotel	1	5	2015	July	28	5	1
200	Resort Hotel	0	0	2015	July	28	7	0
...	...	...	...	...	...	...	...	...
119349	City Hotel	0	186	2017	August	35	31	0
119352	City Hotel	0	63	2017	August	35	31	0
119353	City Hotel	0	63	2017	August	35	31	0
119354	City Hotel	0	63	2017	August	35	31	0
119373	City Hotel	0	175	2017	August	35	31	1

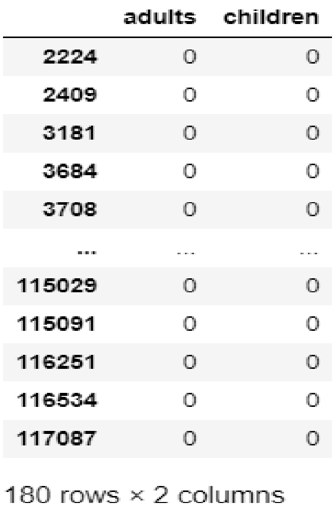
32006 rows × 31 columns

We chose not to remove them, as our dataset does not contain a unique customer ID for privacy purposes. So, there is no way to pinpoint if the record is a duplicate, or there are customers with similar attributes.

5.3. Duplicate Values

During the exploratory data analysis, some rows were found with no adults and no children. It is impossible to make a booking for a room, without any adults or children. So, we consider these records as erroneous, and drop them from the dataset for further analysis.

180 such records were thus eliminated from the dataset.



	adults	children
2224	0	0
2409	0	0
3181	0	0
3684	0	0
3708	0	0
...	...	...
115029	0	0
115091	0	0
116251	0	0
116534	0	0
117087	0	0

180 rows × 2 columns

5.4. Data Type Changes

From our observations, we have concluded that there are 2 data types hat have to be changed in the dataset. These are:

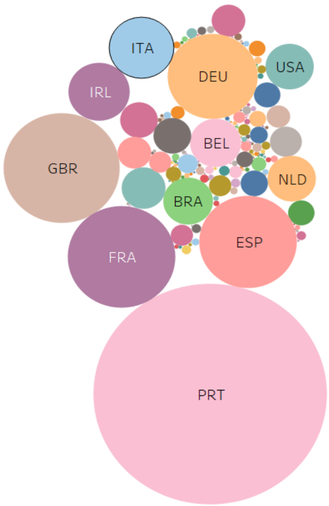
i) Children
ii) Agent

Both these columns have floating point values, which is impossible as the number of children cannot be a fraction. From our observations we have also confirmed that none of the agent IDs are fractional. So, we convert these two columns into integer data types.

6. EXPLORATORY DATA ANALYSIS

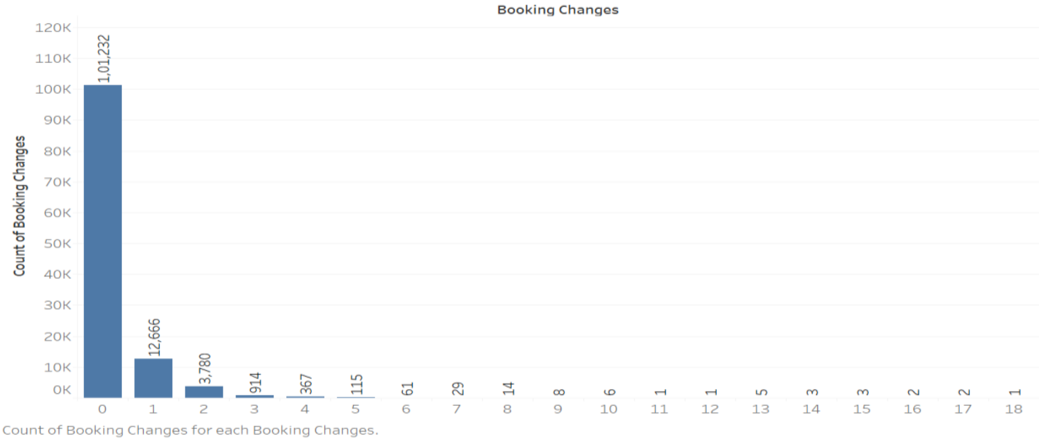
6.1 Univariate Analysis

Country:



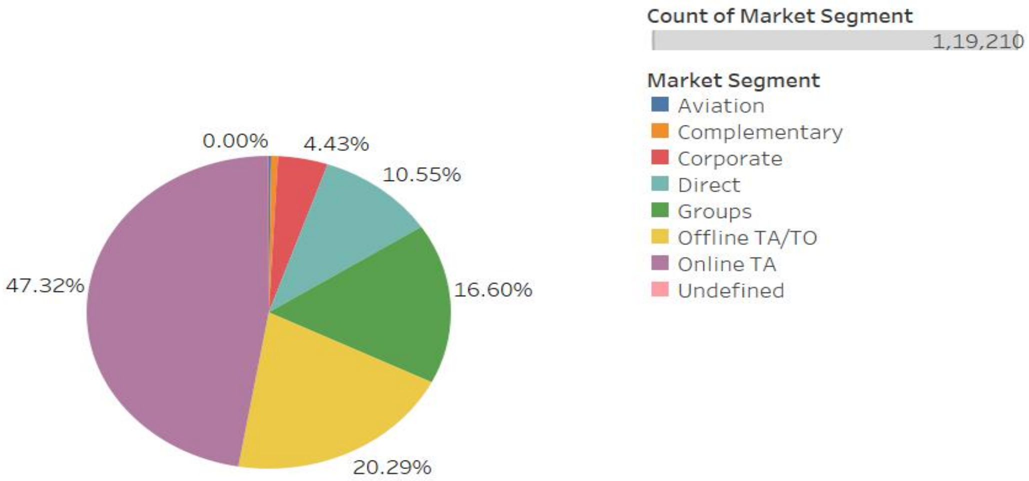
- In the dataset majority of the guests are from PRT, GBR, FRA, ESP, DEU followed by Other countries.

Booking Changes:



- Most of the people who made reservations did not opt for any changes in their booking.
- Majority of the customers haven’t made any booking changes.

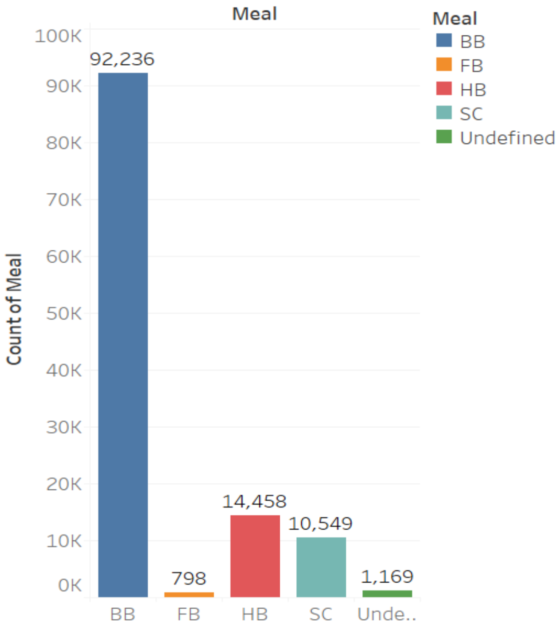
Market Segment:



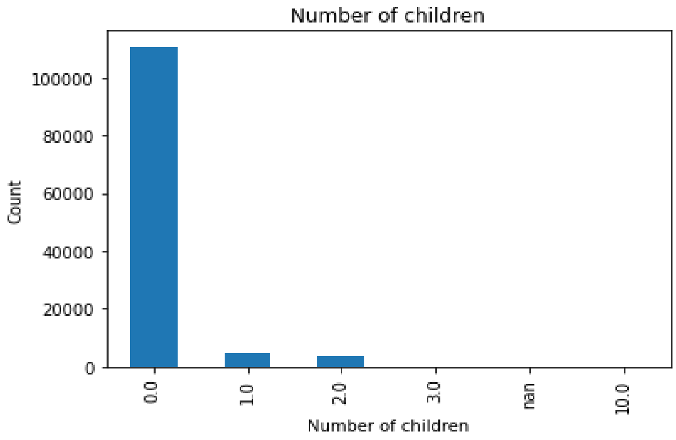
Market Segment (color) and count of Market Segment (size).

- Most of the guests reserve their rooms through Online TA followed by Offline TA/TO.

Meal and Children:



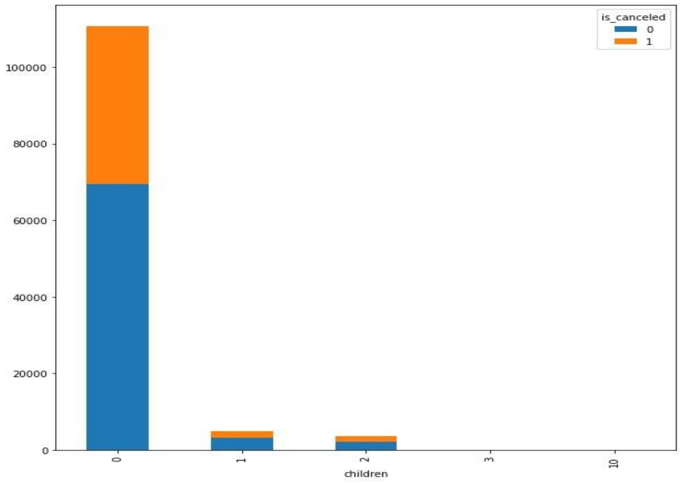
Count of Meal for each Meal. Color shows details about Meal.



- Majority of the guests prefer Bread & Breakfast (BB) over other meal options.
- Our guests are mostly people with no children.

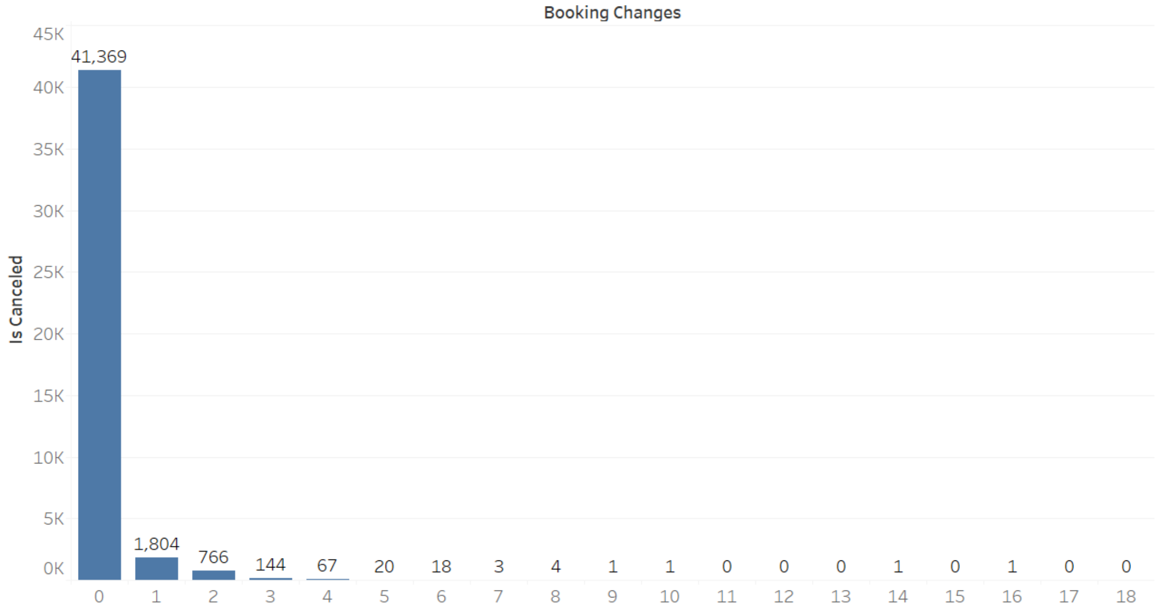
6.2 Bivariate Analysis

Iscanceled with respect to Children:



- Guests who have children are more unlikely to cancel their reservations, compared to guests with no children.
- Additionally, we have to make sure we provide family-friendly activities and spaces to make sure to retain guests with children.

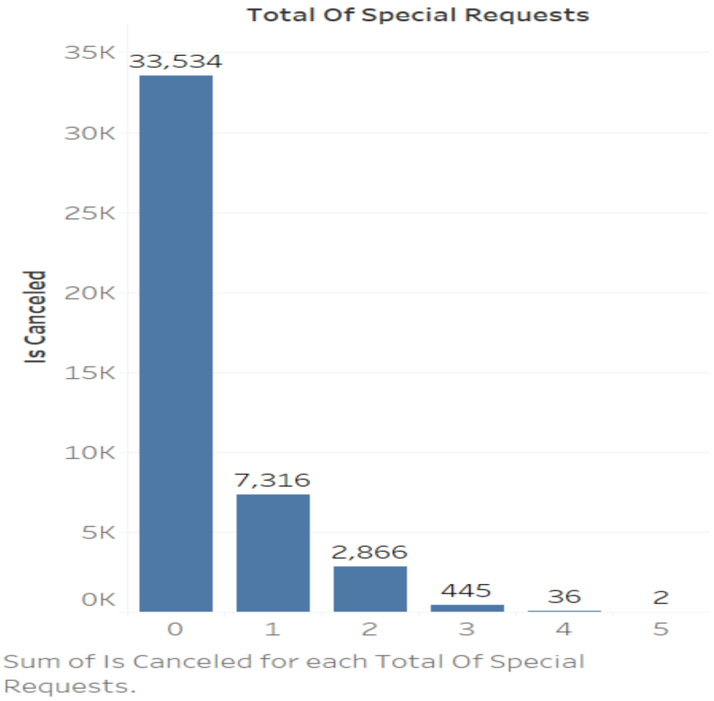
Iscanceled Vs Booking Changes:



Sum of Is Canceled for each Booking Changes.

- Most of the guests who have canceled their reservations are those who did not opt for any changes in their bookings.
- Customers who have gone for changes in the Bookings are less likely to cancel.

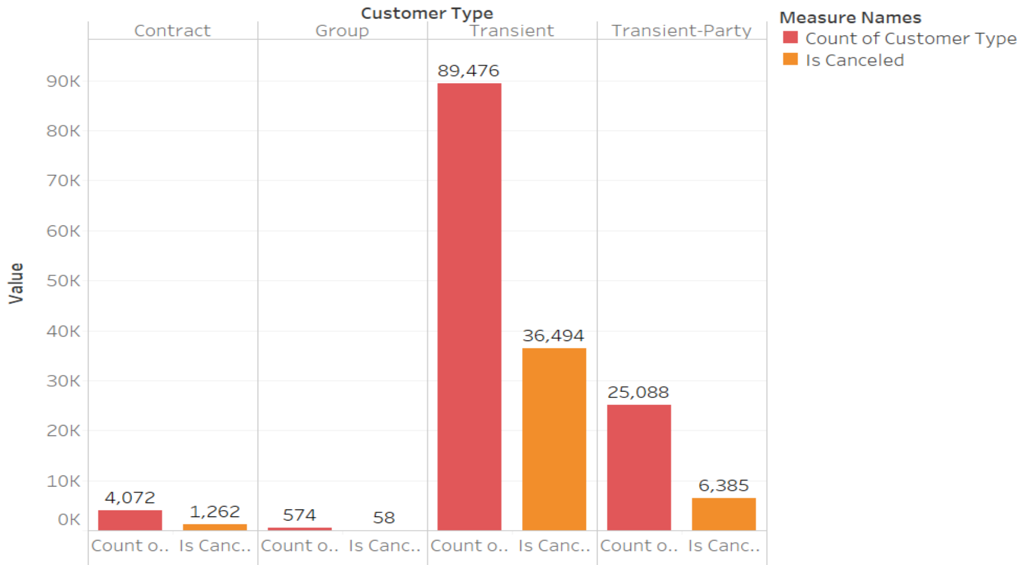
Iscanceled Vs Total no of special requests:



Sum of Is Canceled for each Total Of Special Requests.

- People who made more special requests were more unlikely to cancel their reservation.
- People who made no special requests were the most likely to cancel, and this rate of cancellation steadily decreases as the number of special requests increases.

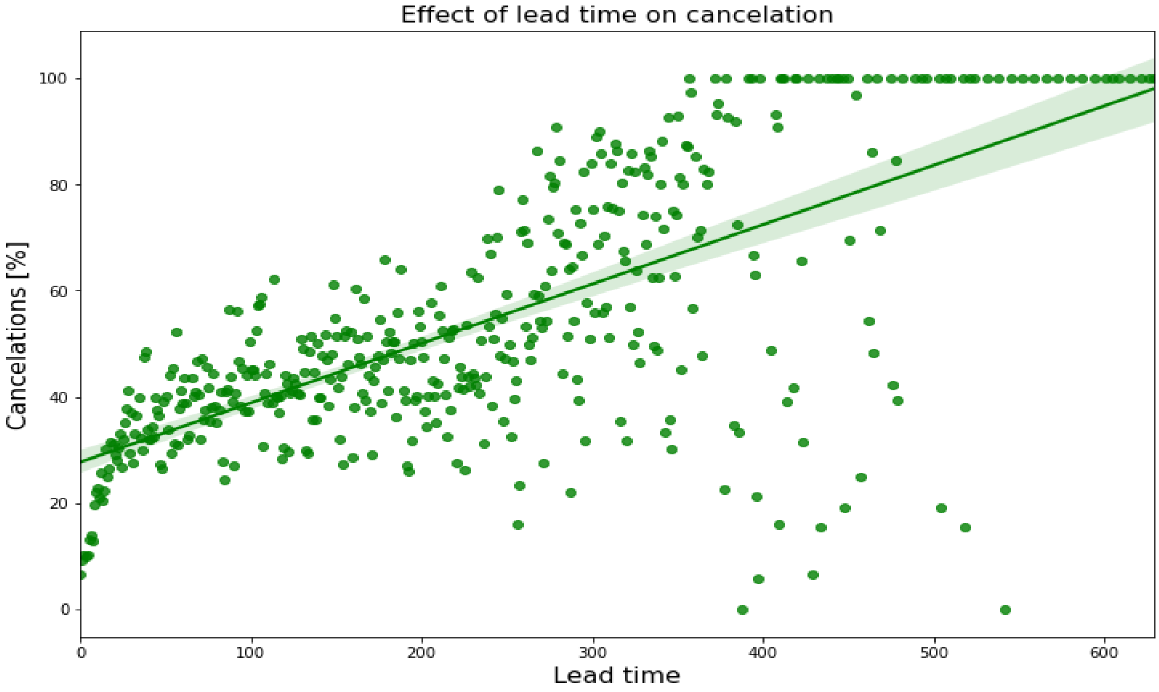
Iscanceled Vs Customer Type:



Count of Customer Type and Is Canceled for each Customer Type. Color shows details about count of Customer Type and Is Canceled.

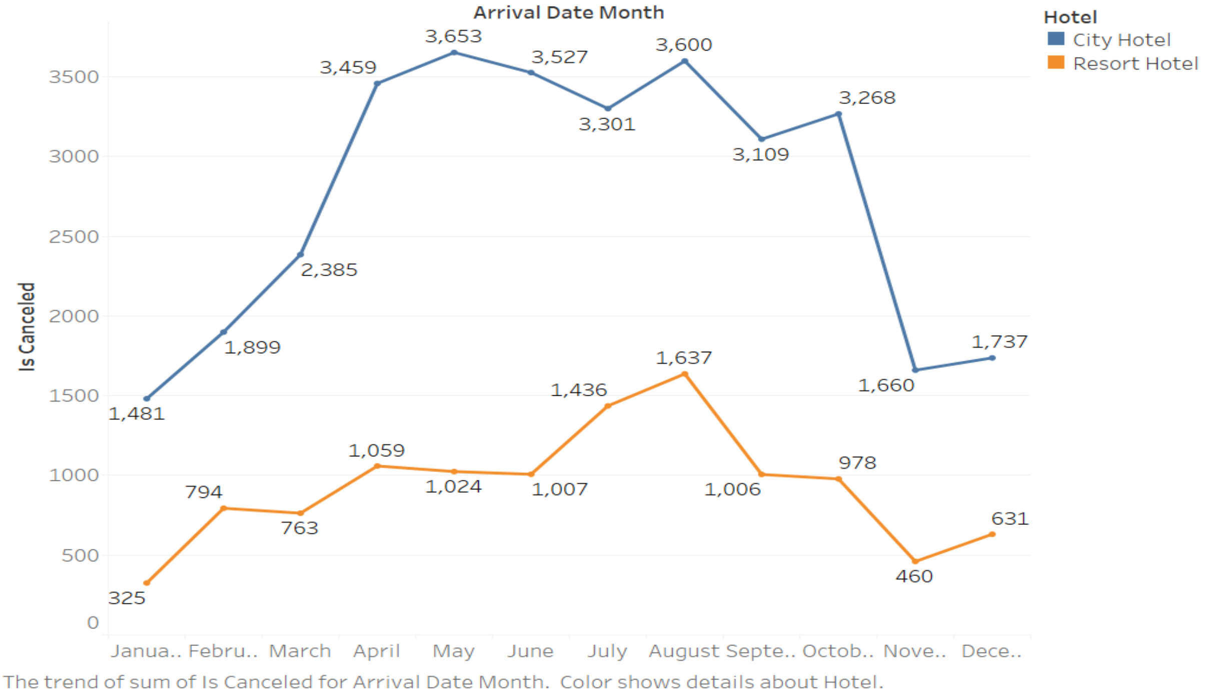
- Transient and transient-party type makeup the majority of the guests.
- Group bookings are very less compared to the other type of bookings.

Iscanceled Vs Lead time:



- Higher the lead time, more the number of cancellations.
- Bookings made a few days before the arrival date are rarely canceled, whereas bookings made over one year in advance are canceled very often.

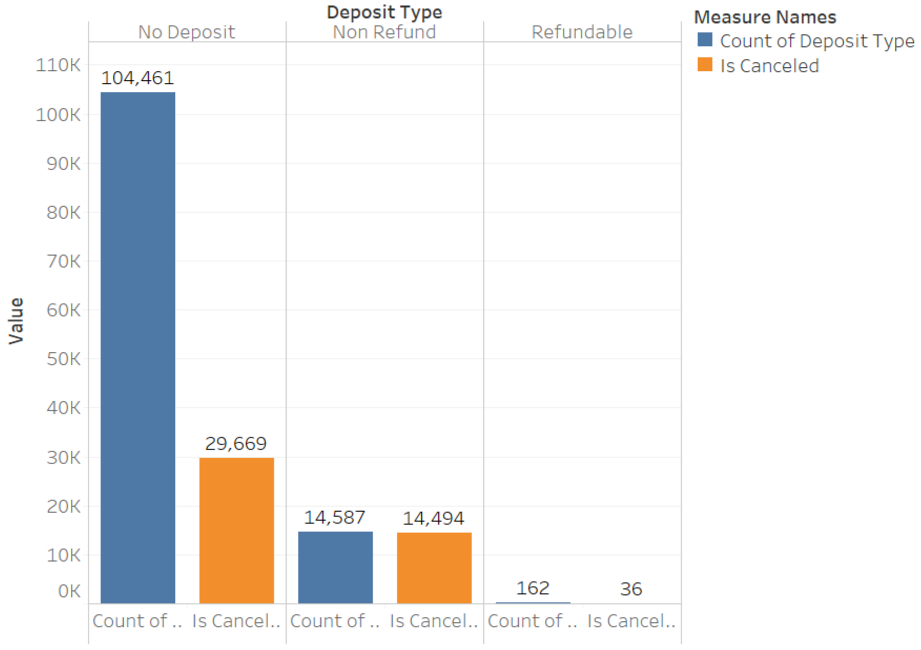
Iscanceled Vs Arrival Date Month:



The trend of sum of Is Canceled for Arrival Date Month. Color shows details about Hotel.

- The month of august accounts for highest cancellations being 5037 in which 3600 is city hotels and 1637 being resort hotels.

Iscancelled vs Deposit type:

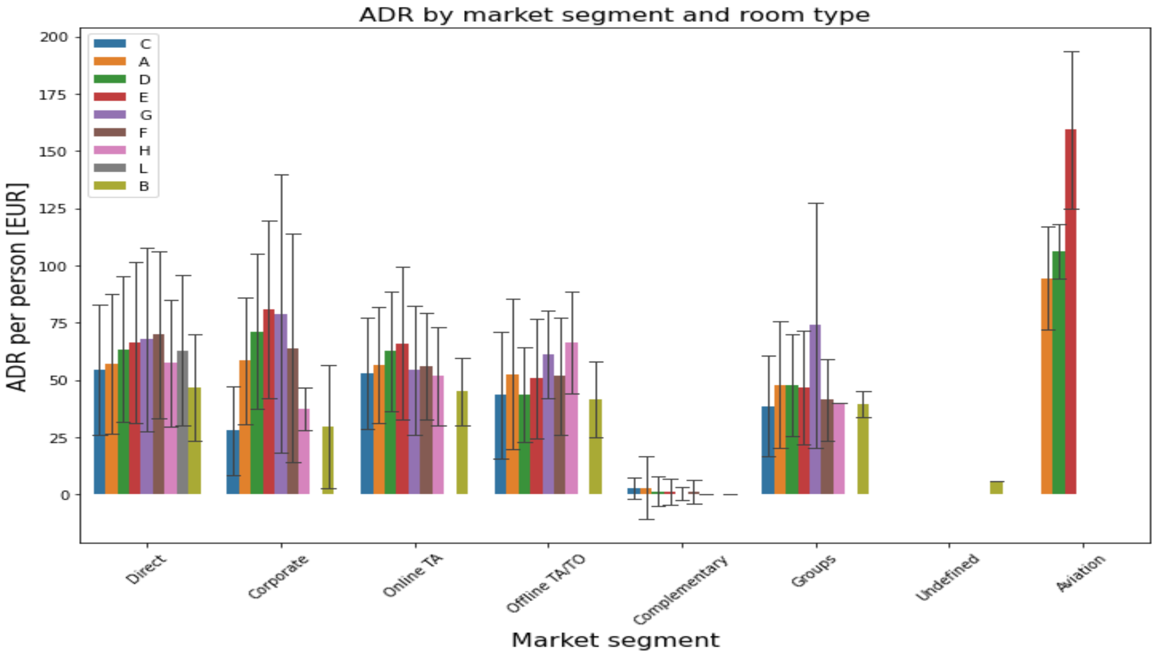


Count of Deposit Type and Is Canceled for each Deposit Type. Color shows details about count of Deposit Type and Is Canceled.

- Most of the people who booked Non-refundable rooms have cancelled.
- Customers mostly prefer the no deposit type since they don't have to pay in advance for the booking.
- The cancellation rate is around 30% for no deposit category.

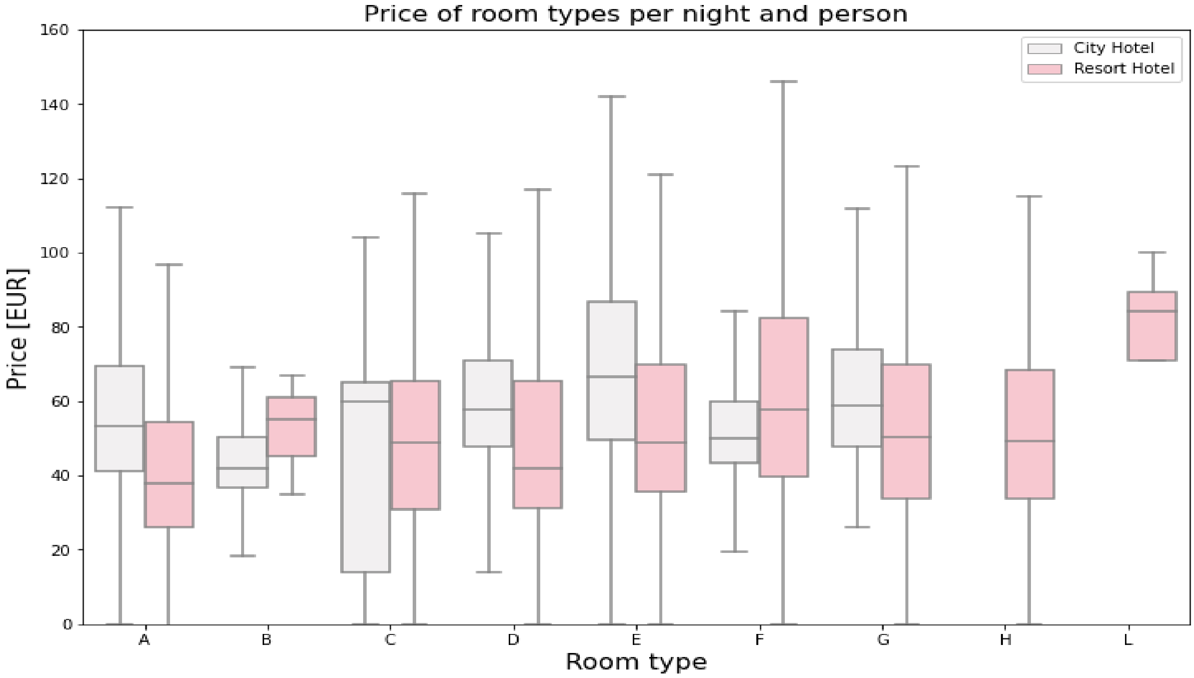
6.3 Multivariate Analysis

ADR – Market Segment – Room Type



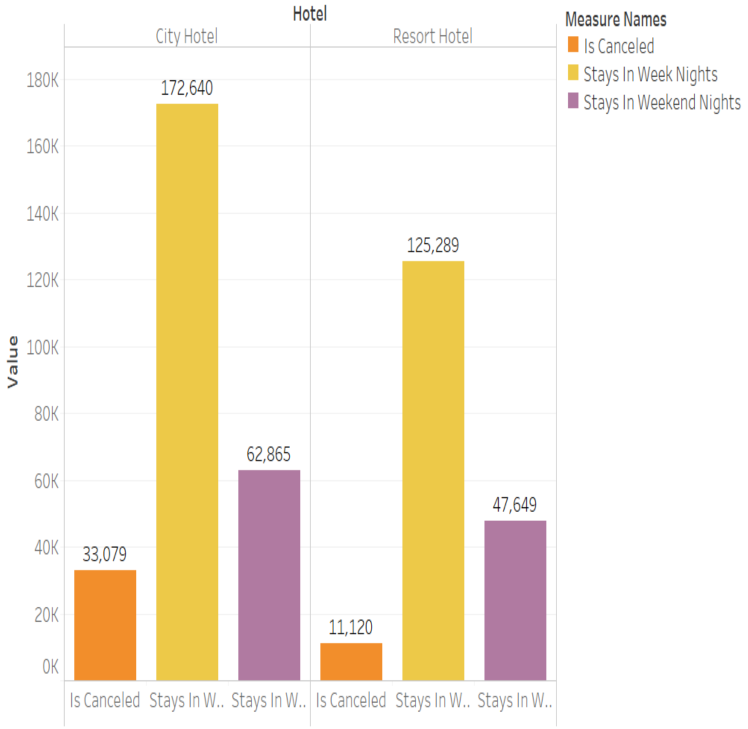
- On average, groups get the best prices and Airlines pay approximately twice as much.

Price – Room Type - Hotel



- This figure shows the average price per room, depending on its type and the standard deviation.
- On an average City Hotel rooms were priced higher than Resort Hotel rooms.

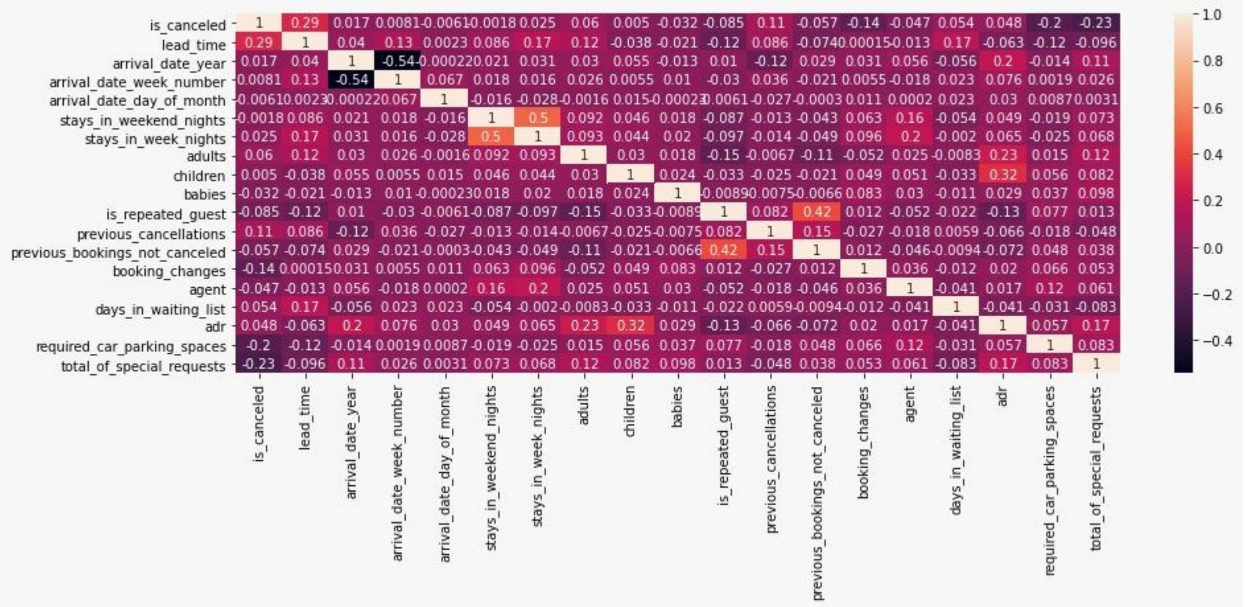
Iscanceled – Stays in Week nights – Stays in weekend nights



Is Canceled, Stays In Week Nights and Stays In Weekend Nights for each Hotel. Color shows details about Is Canceled, Stays In Week Nights and Stays In Weekend Nights.

- More guests reserved rooms for week nights.
- The number of cancellations is more for city hotels compared to resort hotels.

Relationship between variables:



- we can see that previous bookings not cancelled and repeated guests have a positive correlation. Lead time has a positive effect on cancellation too.
- Total no of special requests and Iscanceled has a negative correlation.

Feature Engineering

Features can be engineered by decomposing or splitting features, from external data sources, or aggregating or combining features to create new features.
Feature engineering involves extracting features from raw data with the help of data mining techniques and domain knowledge.
The intention of feature engineering is to achieve two primary goals:
1. Preparing an input dataset that is compatible with and best fits the machine learning algorithm.
2. Improving the performance of machine learning models.

Room:

We have created a new feature called 'Room' which was given values from 2 columns namely Reserved room type and Assigned Room type.

Reserved room type is the one which was assigned to the guest while booking. Assigned room type is the one which was assigned to the guest when the person visits the hotel. If both the room types match, we have assigned 1 to the 'room' column else given 0.

Days Waiting:

Days Waiting has been derived from a column called days in waiting list in the dataset which shows the number of days the booking was in the waiting list before it was confirmed to the customer. Days Waiting has variables like high, medium, low and No waiting. If the number of waiting days is 0, then it was taken as 'No waiting', if the days were between 1 and 50, then low, if the days were between 51 and 130, medium and others are taken as high.

Net Canceled:

Net cancelled will tell us whether previous bookings cancellations are greater than previous bookings not cancelations or not and given values accordingly.

Previous bookings not canceled tells the number of bookings not canceled by the customer prior to the current booking and previous bookings canceled the vice versa.

Lead Category:

Lead category is derived from the column called 'lead time' from the dataset.
Lead time is the number of days that elapsed between the booking date and the arrival date.

If the lead time is 0, the lead category is taken as no lead time. If the lead time is less than 101, low. If the lead time is between 101 and 400, its medium. If the lead time is more than 400, the lead category is high.

Booking Changes:

We have converted the numerical column into a categorical column. If there are no booking changes, replace 0 with 'No changes', if there are less than 6 changes, replace the values with 'few changes' and if the values are more than 6, we have replaced the values with 'more changes'.

Country:

We have around 177 countries in the ‘country’ column. We have taken the percentage of the entries in each country.

We have assigned those countries to respective continents using the country code provided in the dataset.
Now the total unique entries in the country column has reduced from 177 to 7.

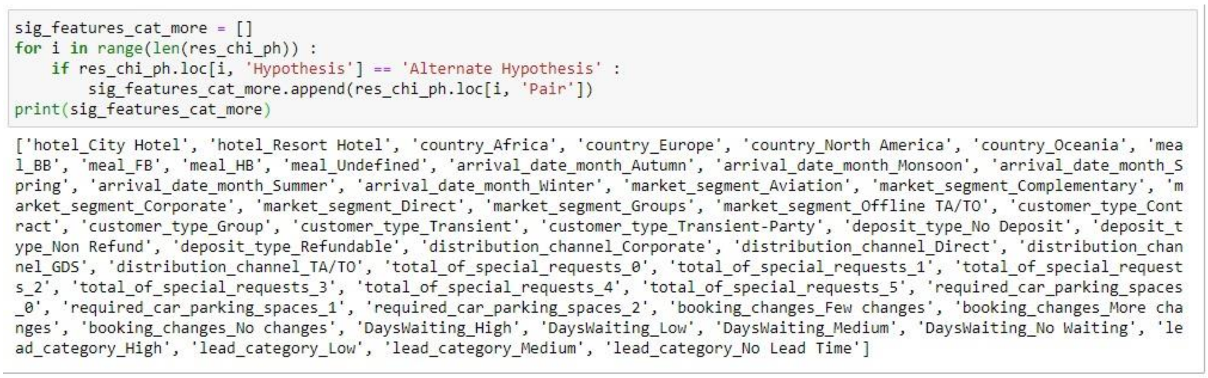
Insignificant Columns:

There are columns in the dataset like ‘arrival date year’, ‘arrival date week number’, ‘arrival date day of month’, ‘reservation status’ and ‘reservation status date’ since they don’t convey much information. We can’t do much analysis or get insights from the above - mentioned columns so we dropped them.

Statistical test for categorical vs categorical variables

Chisquare test for independence:

Null hypothesis - The two variables are independent
Alternate hypothesis - The two variables are dependent



```
sig_features_cat_more = []
for i in range(len(res_chi_ph)) :
    if res_chi_ph.loc[i, 'Hypothesis'] == 'Alternate Hypothesis' :
        sig_features_cat_more.append(res_chi_ph.loc[i, 'Pair'])
print(sig_features_cat_more)
```

```
['hotel_City Hotel', 'hotel_Resort Hotel', 'country_Africa', 'country_Europe', 'country_North America', 'country_Oceania', 'mea
l_BB', 'meal_FB', 'meal_HB', 'meal_Undefined', 'arrival_date_month_Autumn', 'arrival_date_month_Monsoon', 'arrival_date_month_S
pring', 'arrival_date_month_Summer', 'arrival_date_month_Winter', 'market_segment_Aviation', 'market_segment_Complementary', 'm
arket_segment_Corporate', 'market_segment_Direct', 'market_segment_Groups', 'market_segment_Offline TA/TO', 'customer_type_Cont
ract', 'customer_type_Group', 'customer_type_Transient', 'customer_type_Transient-Party', 'deposit_type_No Deposit', 'deposit_t
ype_Non Refund', 'deposit_type_Refundable', 'distribution_channel_Corporate', 'distribution_channel_Direct', 'distribution_chan
nel_GDS', 'distribution_channel_TA/TO', 'total_of_special_requests_0', 'total_of_special_requests_1', 'total_of_special_request
s_2', 'total_of_special_requests_3', 'total_of_special_requests_4', 'total_of_special_requests_5', 'required_car_parking_spaces
_0', 'required_car_parking_spaces_1', 'required_car_parking_spaces_2', 'booking_changes_Few changes', 'booking_changes_More cha
nges', 'booking_changes_No changes', 'DaysWaiting_High', 'DaysWaiting_Low', 'DaysWaiting_Medium', 'DaysWaiting_No Waiting', 'le
ad_category_High', 'lead_category_Low', 'lead_category_Medium', 'lead_category_No Lead Time']
```

The above attributes have been found to have a significant influence by their P-values using chisquare test. (p-values are lesser than 5%)

Applying Transformations:

While checking the normality of the numerical features, it is observed that they are highly skewed. So, we apply transformation techniques on the numerical columns to fit them into a Gaussian distribution to make statistical analysis easier.

A power transform will make the probability distribution of a variable more Gaussian.

This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution.

We can apply a power transform directly by calculating the log or square root of the variable, although this may or may not be the best power transform for a given variable. Instead, we can use a generalized version of the transform that finds a parameter (lambda) that best transforms a variable to a Gaussian probability distribution.

There are two popular approaches for such automatic power transforms; they are:

- Box-Cox Transform
- Yeo-Johnson Transform

In our case we use the yeo-johnson technique because our dataset consists of zero values which cannot be processed by the box-cox technique as it accepts strictly positive values only.

The transformed training dataset can then be fed to a machine learning model to learn a predictive modeling task.

After transformation, we observe that 'children', 'babies' and 'adults' column are still highly skewed. We club them together into a new feature, 'People', and re-apply the transformation techniques.

Finally, an almost normal distribution of numerical features is obtained, and we proceed to scale our data.

One-Way ANOVA :

To perform the ANOVA test, we check for the following two conditions :

1. Test of normality

2. Test of Variance

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

We Use ANOVA for:

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

$$H_O: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

Where µ = group mean and k = number of groups.

If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (HA), which is that there are at least two group means that are statistically significantly different from each other.

Scaling the data:

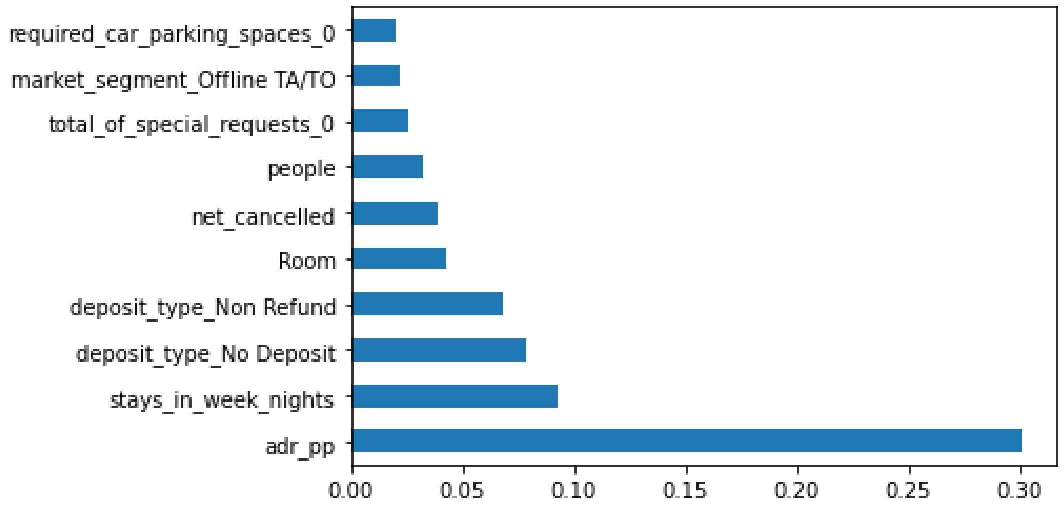
Feature Scaling refers to putting the values in the same range or same scale so that no variable is dominated by the other.

Most of the time, our dataset will contain features highly varying in magnitudes, units and range. Since some of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem.

If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

We perform Min Max Scaling to scale the data in our dataset.

Best features obtained through XGBoost Classifier:



- Adr_pp
- Stays_in_week_nights
- Deposit_type_No_Deposit
- Deposit_type_Non_Refund
- Room
- Net_cancelled

7. MACHINE LEARNING

1. Train-Test Split :

The Dataset is split into train and test in the ratio of 70:30

2. Base Model

Logistic Regression

It is a predictive algorithm using independent variables to predict the dependent variable. It is similar to Linear Regression, but with a difference that the dependent variable should be categorical variable.
Independent variables can be numeric or categorical variables, but the dependent variable will always be categorical.
Logistic regression is a statistical model that uses Logistic function to model the conditional probability.

Representation of Logistic Regression

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

Performance Metrics

Accuracy

Accuracy is the sum of true positives and true negatives divided by the sum of all the observations.
Accuracy = (TP+TN)/(TP+FP+FN+TN)
The accuracy of our base model is around 81.43%.

Precision

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are actually negative.
Precision = TP/(TP+FP)
The precision of our base model is around 83.29%.

Recall

The ratio of correct positive predictions to the total positive examples. It is also called Sensitivity or True Positive Rate.

Recall = TP/(TP+FN)

The recall score of our model is around 62.39%.

F1Score

F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall and when there is an uneven class distribution, F1 score performs better than accuracy.

So F1 score is the preferred metric.

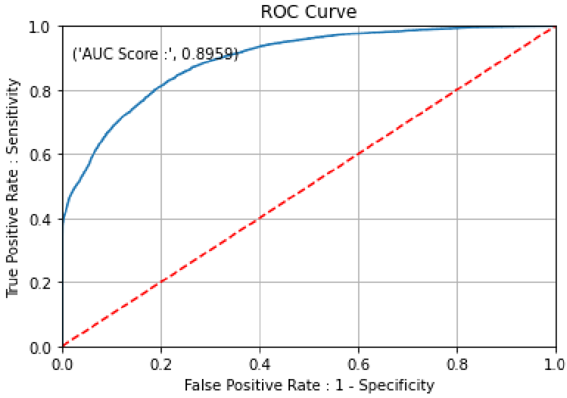
F1 score = 2*(P*R/(P+R))

P - Precision

R – Recall

The base model’s F1-Score is around 71.34%.

ROC-AUC Curve



The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Our base model has an AUC Score of 89.59%, which indicates our model can correctly classify between the classes almost 90% of the times.

2. Pros and Cons:

Pros of Logistic Regression:

i) Logistic Regression performs well when the dataset is linearly separable.

ii) Logistic regression is less prone to over-fitting.

iii) Logistic Regression not only gives a measure of how relevant a predictor is, but also its direction of association.

iv) Logistic regression is easier to implement, interpret and very efficient to train.

Cons of Logistic Regression:

i) Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables

ii) If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit

iii) Logistic Regression can only be used to predict discrete functions.

Therefore, the dependent variable of Logistic Regression is restricted
to the discrete number set.

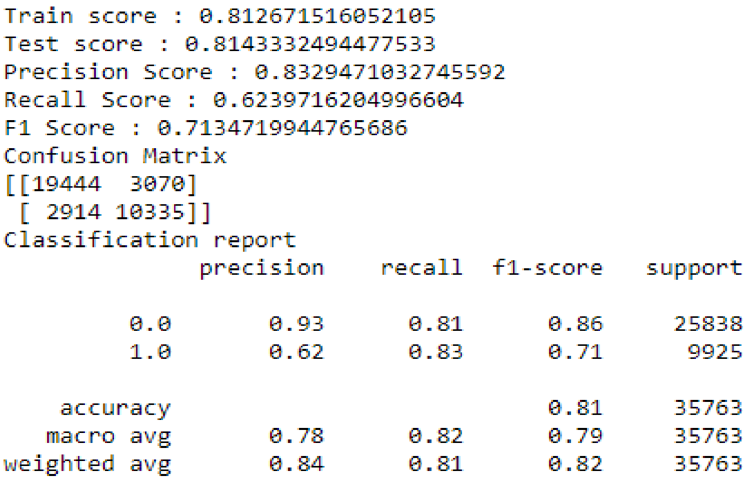
To improve our prediction capability, we fit our train dataset with different models and compare their performances.

Reselecting our Base Model :

MODEL	TRAIN	TEST	PRECISION	RECALL	F1
Logistic Regression	0.81	0.81	0.83	0.62	0.71
Decision Tree	0.96	0.82	0.76	0.74	0.75
Random Forest	0.96	0.84	0.80	0.74	0.77
KNN	0.87	0.82	0.78	0.73	0.75
Naïve Bayes	0.53	0.54	0.44	0.97	0.60
XG Boost	0.85	0.84	0.84	0.70	0.76
Ada Boost	0.81	0.81	0.84	0.61	0.71

Base Model Results:

The base Logistic Regression model gives the following prediction:



```
Train score : 0.812671516052105
Test score : 0.8143332494477533
Precision Score : 0.8329471032745592
Recall Score : 0.6239716204996604
F1 Score : 0.7134719944765686
Confusion Matrix
[[19444  3070]
 [ 2914 10335]]
Classification report
              precision    recall  f1-score   support

         0.0       0.93      0.81      0.86     25838
         1.0       0.62      0.83      0.71      9925

    accuracy                           0.81     35763
   macro avg       0.78      0.82      0.79     35763
weighted avg       0.84      0.81      0.82     35763
```

Base model's precision, recall and F1-score are not so appealing so further techniques such as SMOTE, feature selection and featuretools are applied to improve the metrics of the predictive model.

F1 Score is 77.54%. We try to improve this score using SMOTE, an oversampling technique, to deal with class imbalance.

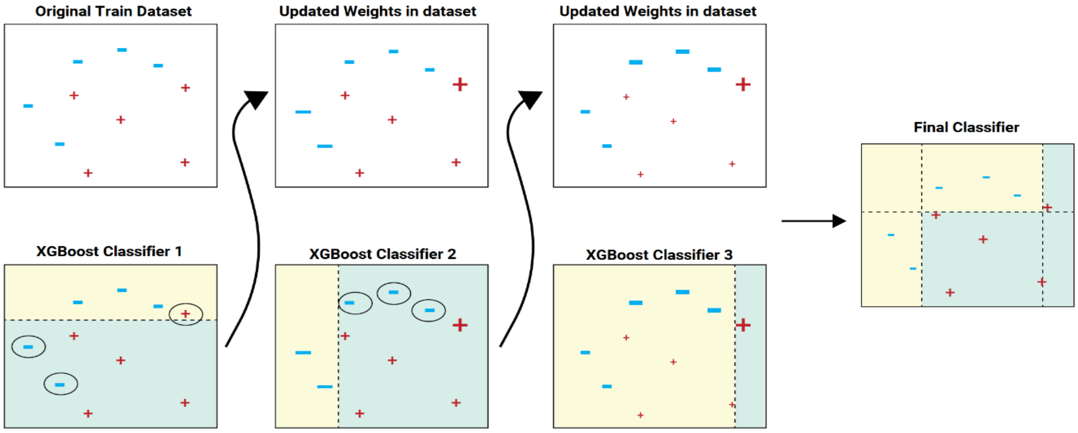
XGBoost Classifier :

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

XGBoost is a gradient boosting library with focus on tree model, which means inside XGBoost, there are 2 distinct parts:

1. The model consisting of trees and
2. Hyperparameters and configurations used for building the model.

The sequential ensemble methods, also known as "boosting", creates a sequence of models that attempt to correct the mistakes of the models before them in the sequence. The first model is built on training data, the second model improves the first model, the third model improves the second, and so on.



In the above image example, the train dataset is passed to the classifier 1. The yellow background indicates that the classifier predicted hyphen and blue background indicates that it predicted plus. The classifier 1 model incorrectly predicts two hyphens and one plus. These are highlighted with a circle. The weights of these incorrectly predicted data points are increased and sent to the next classifier. That is to classifier 2. The classifier 2 correctly predicts the two hyphen which classifier 1 was not able to. But classifier 2 also makes some other errors. This process continues and we have a combined final classifier which predicts all the data points correctly.

The classifier models can be added until all the items in the training dataset is predicted correctly or a maximum number of classifier models are added. The optimal maximum number of classifier models to train can be determined using hyperparameter tuning.

In gradient boosting while combining the model, the loss function is minimized using gradient descent. Technically speaking, a loss function can be said as an error, ie the difference between the predicted value and the actual value. Of course, the less the error, the better is the machine learning model.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction.

The objective of the XGBoost model is given as:

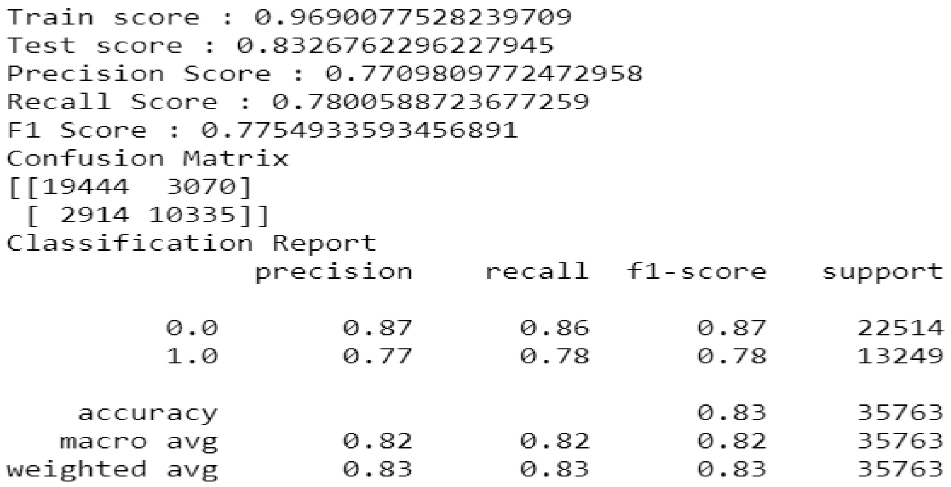
$$\text{Obj} = L + \Omega$$

Where L is the loss function which controls the predictive power, and Ω is regularization component which controls simplicity and overfitting

The loss function (L) which needs to be optimized can be Root Mean Squared Error for regression, Logloss for binary classification, or mlogloss for multi-class classification.

The regularization component (Ω) is dependent on the number of leaves and the prediction score assigned to the leaves in the tree ensemble model.

It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. The Gradient boosting algorithm supports both regression and classification predictive modelling problems.

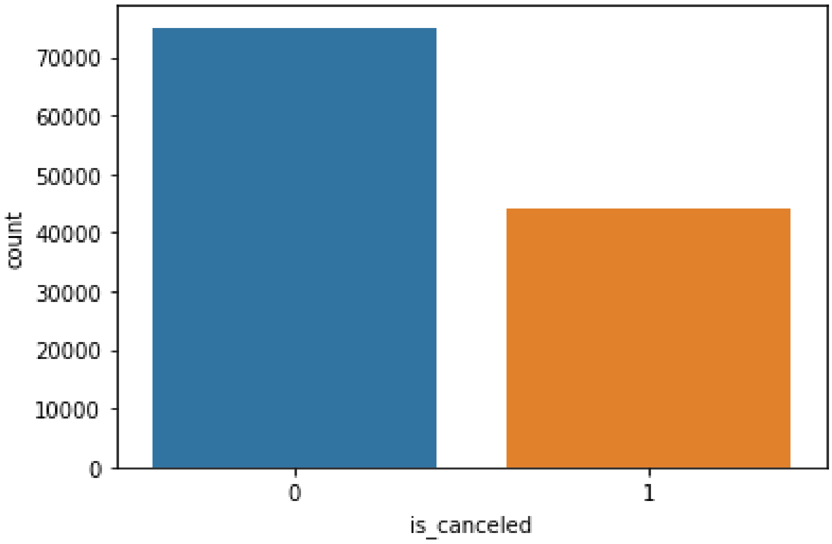


```
Train score : 0.9690077528239709
Test score : 0.8326762296227945
Precision Score : 0.7709809772472958
Recall Score : 0.7800588723677259
F1 Score : 0.7754933593456891
Confusion Matrix
[[19444  3070]
 [ 2914 10335]]
Classification Report
              precision    recall  f1-score   support

         0.0       0.87      0.86      0.87     22514
         1.0       0.77      0.78      0.78     13249

    accuracy                           0.83     35763
   macro avg       0.82      0.82      0.82     35763
weighted avg       0.83      0.83      0.83     35763
```

Synthetic Minority Oversampling Technique (SMOTE):

This two-class dataset is imbalanced (63% vs 37%). As a result, there is a possibility that the model built might be biased towards to the majority and over-represented class. After applying Synthetic Minority Oversampling Technique (SMOTE) to over sample the minority class, we obtain good F1-score & Recall score with selected features.

In order to rectify this problem, we can do the following techniques to sample the data: 1. Up Sampling 2. Down Sampling 3. Up Sampling or Down Sampling using SMOTE



Even though smote gives the lesser f1 score we proceed with it since the class imbalance is treated here. So we proceed with the new train data.

Tuning the HyperParameters:

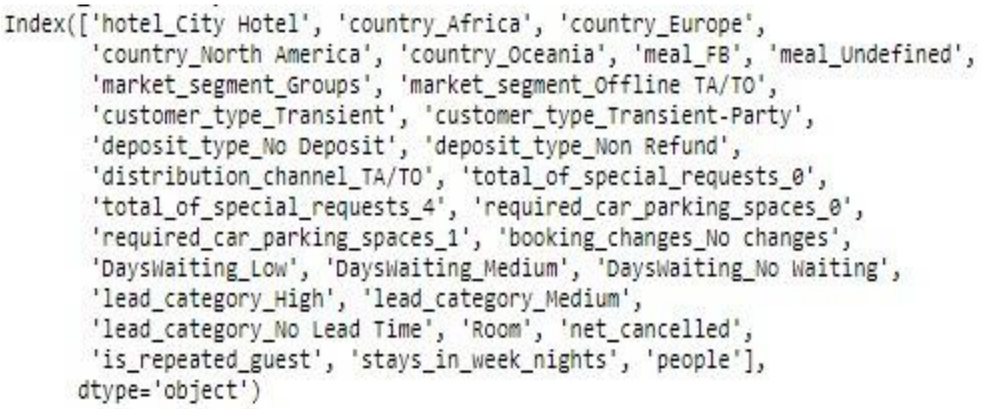
A hyperparameter is a parameter that is set before the learning process begins. It is given by the user. We use the GridSearchCV technique to select the hyperparameters that enable our model to provide the best results. We give a range of different hyperparamater values, from which the GridSearchCV algorithm selects the optimum hyperparameters. GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. We then pass these hyperparameters to our base model, i.e.; we re-fit our base model before performing any further optimization techniques.

We fit the base model with the tuned parameters and we notice an increase in performance so we proceed with this base model with the tuned hyper parameters.
Now the new F1 score is 78.32%

Feature Selection:

Recursive Feature Elimination (RFE):

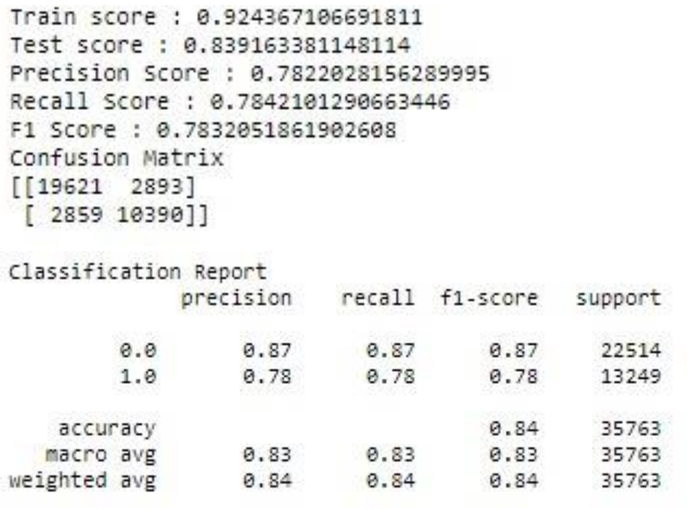
After the application of RFE, the original features were reduced from 58 to 30 Features.



```
Index(['hotel_City Hotel', 'country_Africa', 'country_Europe',
       'country_North America', 'country_Oceania', 'meal_FB', 'meal_Undefined',
       'market_segment_Groups', 'market_segment_Offline TA/TO',
       'customer_type_Transient', 'customer_type_Transient-Party',
       'deposit_type_No Deposit', 'deposit_type_Non Refund',
       'distribution_channel_TA/TO', 'total_of_special_requests_0',
       'total_of_special_requests_4', 'required_car_parking_spaces_0',
       'required_car_parking_spaces_1', 'booking_changes_No changes',
       'DaysWaiting_Low', 'DaysWaiting_Medium', 'DaysWaiting_No Waiting',
       'lead_category_High', 'lead_category_Medium',
       'lead_category_No Lead Time', 'Room', 'net_cancelled',
       'is_repeated_guest', 'stays_in_week_nights', 'people'],
      dtype='object')
```

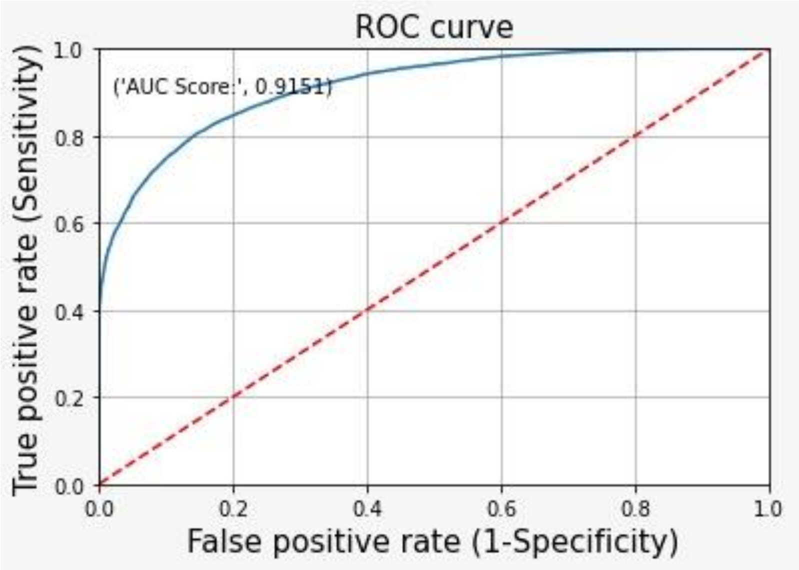
These are the significant features in the model but when we built a model only with these significant features our model performance decreases. So we select our final model with tuned parameters and all the features.

FINAL MODEL:



```
Train score : 0.924367106691811
Test score : 0.839163381148114
Precision Score : 0.782202815628995
Recall Score : 0.7842101290663446
F1 Score : 0.7832051861902608
Confusion Matrix
[[19621  2893]
 [ 2859 10390]]

Classification Report
              precision    recall  f1-score   support

         0.0       0.87      0.87      0.87     22514
         1.0       0.78      0.78      0.78     13249

    accuracy                           0.84     35763
   macro avg       0.83      0.83      0.83     35763
weighted avg       0.84      0.84      0.84     35763
```

It is observed that XGBoost Classifier along with the features yielded the best results of all the other predictive models.



- ROC AUC: 0.91
- Train accuracy score: 0.92
- Test accuracy score: 0.84

The ROC AUC score, precision, F1-score have all improved comparatively and hence this is adopted as our final model.

8. BUSINESS SUGGESTIONS

A business insight combines data and analysis to find meaning in and increase understanding of a situation, resulting in some competitive advantage for a business. Simply performing exploratory data analysis, building models and deriving insights won't be of any help if we are not able to leverage these insights into business solutions.

Here, we discuss some of the business solutions that the hotels can implement and improvise upon, based on the insights gathered from our dataset.

Insight: Guests who made more special requests were less likely to cancel their booking.
Business Solution: In order to ensure lesser number of cancellations, the hotel must cater to as many special requests as possible, and even encourage guests to state any preferences they have. This could add a hint of personalization to the guests' hotel rooms and their overall stay experience, and they could be less likely to cancel.

Insight: The most popular distribution channel was online TA/TO.
Business Solution: Most of the bookings were done online. So, if the hotel invests in increasing their online presence, they might get more guests. Also, their online marketing must be relevant with respect to the different countries.

Insight: Higher the lead time, i.e.; the time between booking and checking in, more the cancellations.
Business Solution: People who booked just before checking in, were more likely to not cancel their booking. But for bookings which were made well in advance, chances are high that they will be cancelled. To tackle this, hotels can make advanced bookings non-refundable, or at least only refund a part of the deposit. This could reduce the number of cancellations, or in the case of cancellations, reduction in the revenue lost by the hotel.

Insight: Bed and Breakfast was the most preferred meal option.
Business Solution: Since breakfast is the most preferred option, the hotels can make sure to cater to people with different preferences by ensuring that a variety of meal options, such as vegan, vegetarian, low-carb options are available. The hotels can also offer a discount on the other meals if the guest opts for a BB plan, thus encouraging guests to opt for the other meals as well, contributing to the overall revenue of the hotel.

Insight: Most of the guests are from European countries.
Business Solution: Offer discounts and set up special hotel plans, conduct events and themed parties during the European holidays and holiday seasons. Different marketing strategies relevant to the other countries must be discussed and implemented to encourage people from other countries to reserve bookings.

Insight: The arrival season is really important in determining cancellation.
Business Solution: Different seasonal events ranging from heritage walks to harvesting events can be held to boost bookings. The hotels can also work closely with the locals to provide guests with the workings of the local society during the different seasons.

Insight: People with children are less likely to cancel bookings.
Business solution: Children-friendly spaces and events can be conducted in certain parts of the hotel. This must be heavily advertised during the holiday seasons when families are most likely to take a vacation.

9. PROJECT OUTCOME

There are different metrics that help in quantify performance of the model such as accuracy, recall, precision, F1 score. For our model, F1 score is the most accurate metric to judge the performance.

Before explaining the reason why, we will talk about the FP(False Positive) and False Negative(FN) predictions made by the model. With respect to our dataset, FP implies that the model has incorrectly labeled a customer who has not cancelled to has cancelled. FN implies that the customer who has cancelled as not cancelled.

FP condition may lead to overbooking in hotels and subsequent loss of reputation, whereas as FN may lead to underbooking and loss of revenue for the hotel. Both of these are unfortunate and have

a high cost attached to it. While precision gives weightage to FP, recall gives weightage of FN. F1 score is the harmonic mean of precision and recall, thus giving weightage to both the conditions. Meanwhile accuracy doesn't give any weightage to False Positive and False Negative.

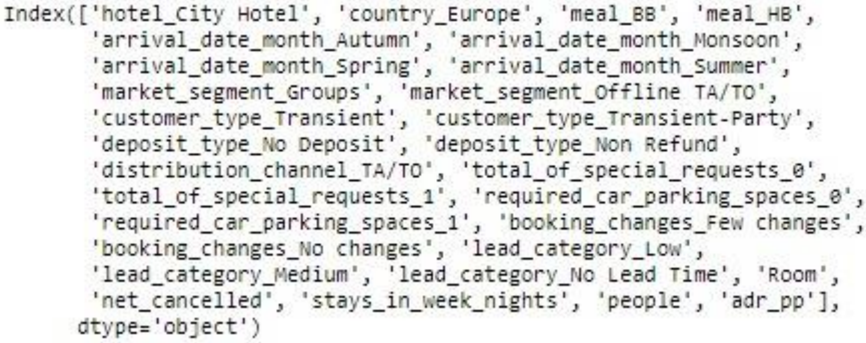
Hence, F1 score is the best score for our model. We have obtained an F1 score of 0.7832

An AUC Score of 0.915 is closer to 1, which indicates that our model is able to distinguish between the two classes quite correctly.

F1 score is given as follows: (2 * Precision * Recall) / (Precision + Recall)

The train score of our model is 92.43%, whereas the test score is 83.91%. The model is only slightly overfit.

<u>The 30 significant features which helps us predict the cancellation</u>



```
Index(['hotel_City Hotel', 'country_Europe', 'meal_BB', 'meal_HB',
       'arrival_date_month_Autumn', 'arrival_date_month_Monsoon',
       'arrival_date_month_Spring', 'arrival_date_month_Summer',
       'market_segment_Groups', 'market_segment_Offline TA/TO',
       'customer_type_Transient', 'customer_type_Transient-Party',
       'deposit_type_No Deposit', 'deposit_type_Non Refund',
       'distribution_channel_TA/TO', 'total_of_special_requests_0',
       'total_of_special_requests_1', 'required_car_parking_spaces_0',
       'required_car_parking_spaces_1', 'booking_changes_Few changes',
       'booking_changes_No changes', 'lead_category_Low',
       'lead_category_Medium', 'lead_category_No Lead Time', 'Room',
       'net_cancelled', 'stays_in_week_nights', 'people', 'adr_pp'],
      dtype='object')
```

=

10. CONCLUSION

By using hotel bookings dataset from city and resort hotels, we found the main objectives of the research with the application of data science skills such as data visualization and machine learning, It was found that different features differ in importance depending on the hotel, and some features are not required for some of the hotels. So, we built a model to classify bookings likely to be cancelled.

This demonstrates that machine learning algorithms, in this case, the XGBoost Classifier algorithm with tuned hyperparameters, is a good technique to build booking cancellations prediction models. Model development revealed that features had different weights and different importance accordingly to the hotel, so we tried building a model which can satisfy all the criteria.

These prediction models enable hotel management to mitigate revenue loss derived from booking cancellations. Booking cancellations models also allow hotel managers to implement less rigid cancellation policies, without increasing uncertainty. This has the potential to translate into more sales, since less rigid cancellation policies generate more bookings. These models allow hotel industries to take actions on bookings identified as potentially going to be cancelled, but also to produce more accurate demand forecasts.

Concurrently, development of these models should contribute to improve hotel revenue management as its use of technology and mathematical/machine learning models is in accordance with the work of hotel bookings and its cancellations.