Batch details	DSE-FT-G-Jan22-Group 6
Team members	1. Akshit 2. Ashish Bhardwaj 3. Aryan 4. Gaurav Manori 5. Haider 6. Raj Sonkar
Domain of Project	Hospitality and Tourism
Proposed project title	Hotel Booking Demand
Group Number	6
Team Leader	Gaurav Manori
Mentor Name	PV Subramanian

Date:

Table of Contents

OVERVIEW

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impact a hotel on various fronts:

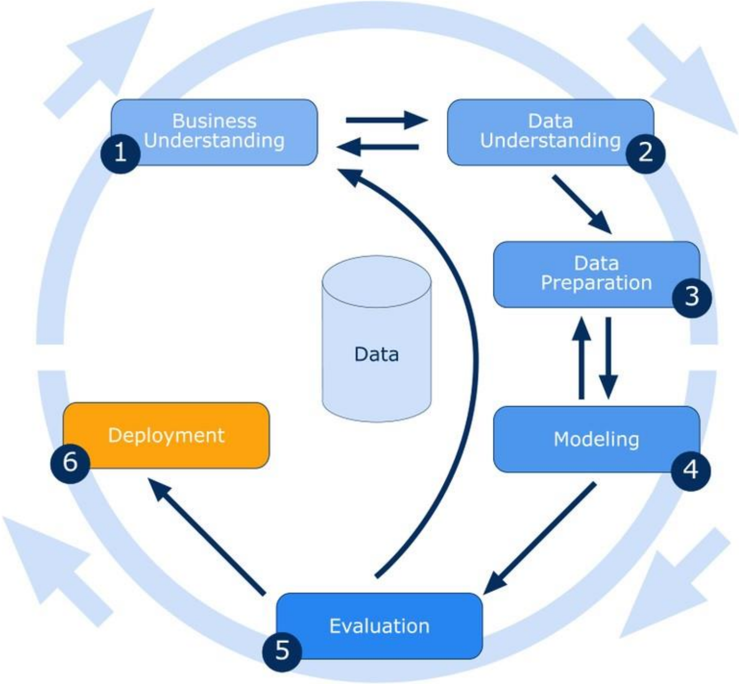
1. Loss of resources (revenue) when the hotel cannot resell the room.
2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
4. Human resources to make arrangements for the guests.

Business problem statement (GOALS)

A Hotel Group chain is facing problems with the high number of booking cancellations. You as a data scientist have to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be cancelled in advance, and help in formulating profitable policies for cancellations and refunds.

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be cancelled .

METHODOLOGY TO BE FOLLOWED (Follow 1-2-3-4-5)



1. Business Understanding: focuses on understanding the objectives and requirements of the project
2. Data Understanding:
   This phase also has four tasks:

   - **Collect initial data:** Acquire the necessary data and (if necessary) load it into your analysis tool.
   - **Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
   - **Explore data:** Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
   - **Verify data quality:** How clean/dirty is the data? Document any quality issues.

3. Data Preparation: This phase, which is often referred to as "data munging", prepares the final data set(s) for modeling.

We'll perform following tasks for Data Preparation:

1. Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.
2. Clean data: We'll perform these steps for Data Cleaning process:

   - Remove irrelevant data
   - Deduplicate your data
   - Fix Structural Errors
   - Deal with Missing data
   - Filter out Outliers
   - Validating Data

3. Construct data: Derive new attributes that will be helpful.
4. Format data: Re-format data as necessary

4. Modeling: We'll perform following steps here:
   - Descriptive Data Analysis:
     - Univariate Analysis: It includes following techniques-:
       - Measure of frequency
       - Measure of Central tendency: Mean, Median and Mode
       - Measure of dispersion: These are the statistics that come under the measure of dispersion.
         - Range
         - Percentiles or Quartiles
         - Standard Deviation
         - Variance
         - Skewness
     - Bivariate Analysis: We'll create Contingency Table here
   - Exploratory Data Analysis:
     - Univariate Analysis:
       - For Numerical variables: Distplots, Histograms, Boxplots
       - For Categorical variables: Frequency plot using countplot
     - Bivariate Analysis:
       - Numerical vs Numerical Variables: ScatterPlot
       - Categorical vs Numerical Variables: Boxplot
       - Category vs Category: Countplot using crosstab
   - Predictive Data Analysis: Select modeling techniques:
     - Logistic Regression
     - Decision Tree
     - Random Forest

5. Evaluation: Here we'll evaluate which model has performed better based on ROC AUC, F1score, Precision and recall values
6. Deployment: It is outside the scope of our project

Variable information/Data description

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.
The target variable is the status of booking whether the booking is cancelled or not

Data Dictionary(Name & Description)

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

- **hotel**: Types of Hotel (H1 = Resort Hotel or H2 = City Hotel)
- **is_cancelled**: Value indicating if the booking was cancelled (1) or not (0)
- **lead_time**: Number of days that elapsed between the entered date of the booking and the arrival date
- **arrival_date_year**: Year of arrival date
- **arrival_date_month**: Month of arrival date
- **arrival_date_week_number**: Week number of year for arrival date
- **arrival_date_day_of_month**: Day of arrival date
- **stays_in_weekend_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- **stays_in_week_nights**: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **adults**: Number of adults
- **children**: Number of children
- **babies**: Number of babies
- **meal**: Type of meal plan booked by the customer:
    1. Not Selected – No meal plan selected
    2. Meal Plan 1 – Breakfast
    3. Meal Plan 2 – Half board (breakfast and one other meal)
    4. Meal Plan 3 – Full board (breakfast, lunch, and dinner)
- **Country**: Country of origin.
- **market_segment**: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

- **distribution_channel**: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

- **is_repeated_guest**: Value indicating if the booking name was from a repeated guest (1) or not (0)

- **previous_cancellations**: Number of previous bookings that were cancelled by the customer prior to the current booking

- **previous_bookings_not_canceled**: Number of previous bookings not cancelled by the customer prior to the current booking.

- **reserved_room_type:** Code of room type reserved. Code is presented instead of designation for anonymity reasons.

- **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.

- **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.

- **deposit_type**: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.

- **Agent** :ID of the travel agency that made the booking.

- **Company**: ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons

- **days_in_waiting_list**:Number of days the booking was in the waiting list before it was confirmed to the customer.

- **customer_type**: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.

- **Adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.

- **required_car_parking_spaces**: Number of car parking spaces required by the customer.

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

- **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

- **reservation_status_dat**e: 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

Exploratory Data Analysis

General Observations:

- Dataset has 119390 rows and 32 columns

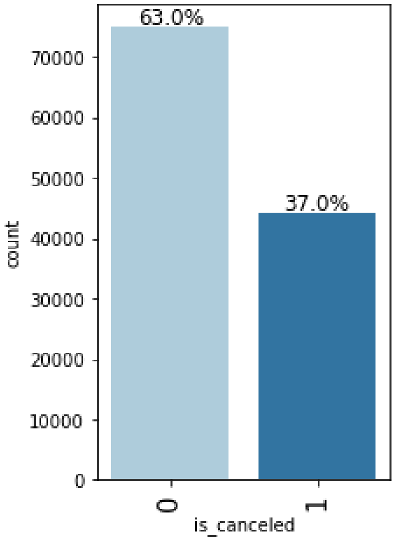
- Count of Missing Values

```
company                112593
agent                   16340
country                   488
children                    4
```

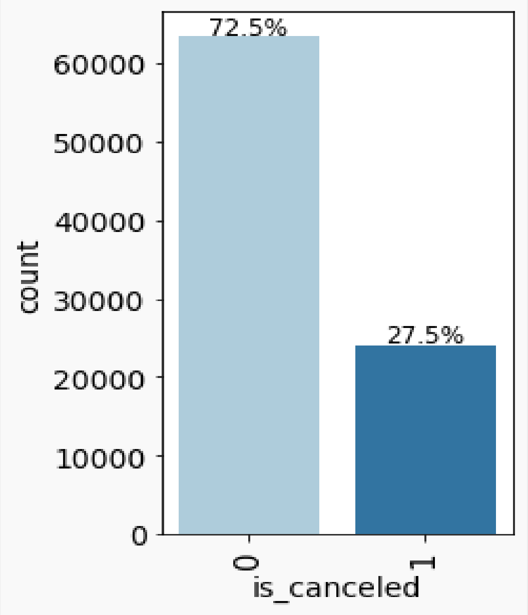
- Percentage of Missing Values(Null values):

```
company                            94.306893
agent                              13.686238
country                             0.408744
children                            0.003350
```

- There is imbalance in the data



- There are around 32000 duplicate rows present in the dataset. So, imbalance after removing duplicate values from the data



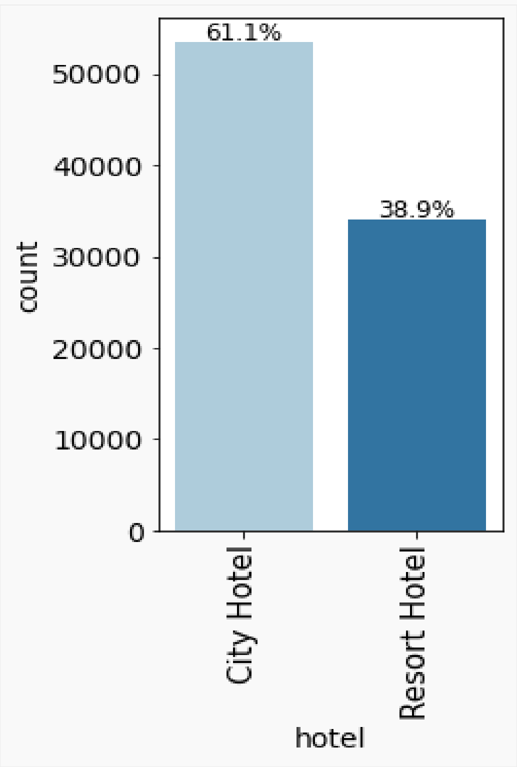
Univariate Analysis

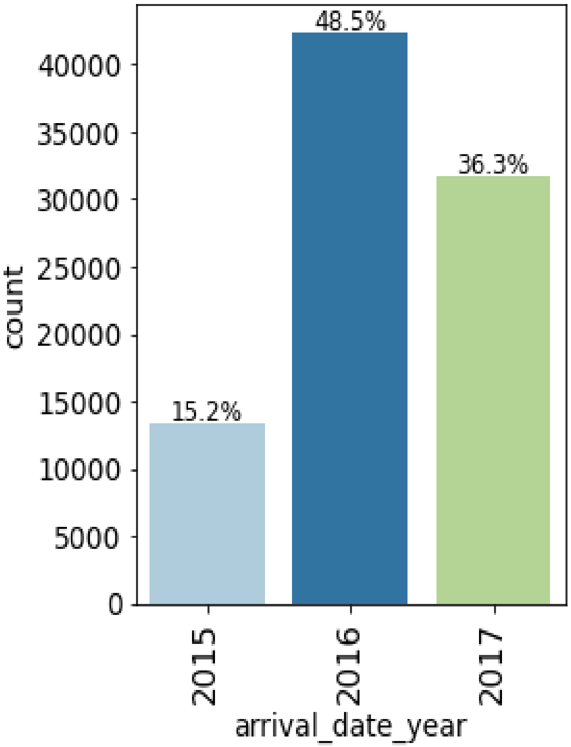
Hotel

- There are larger number of City Hotels than the Resort Hotels (Ratio- 3:2)

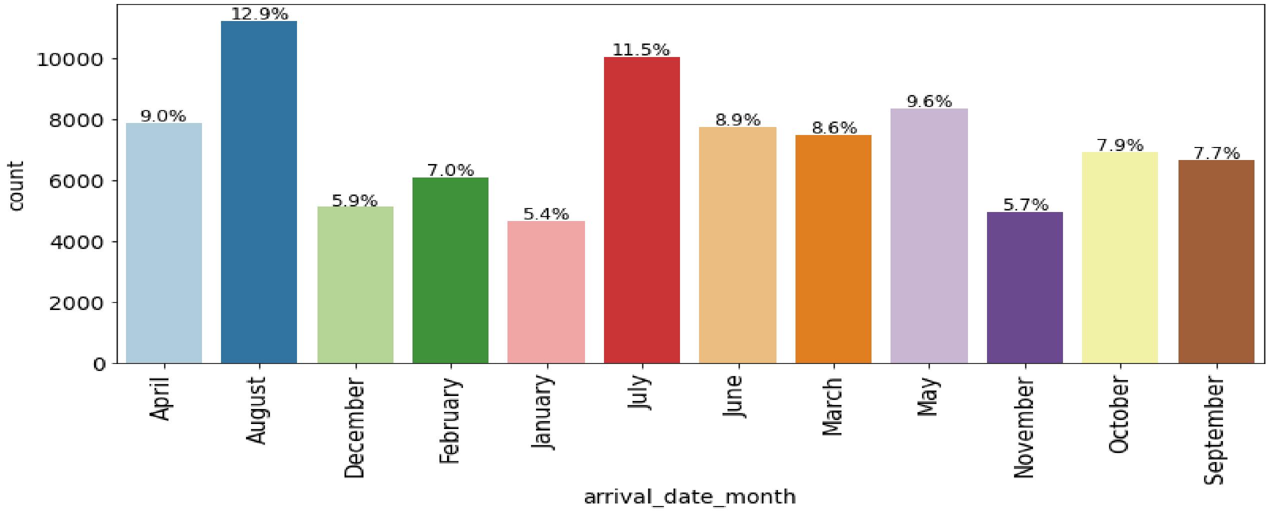
Arrival_date_year

- Most of the bookings were in 2016 followed by 2017 and 2015





- More number of people prefer to visit during July and August.
- November, December and January seems to be the off season where around 5% people prefer to visit.

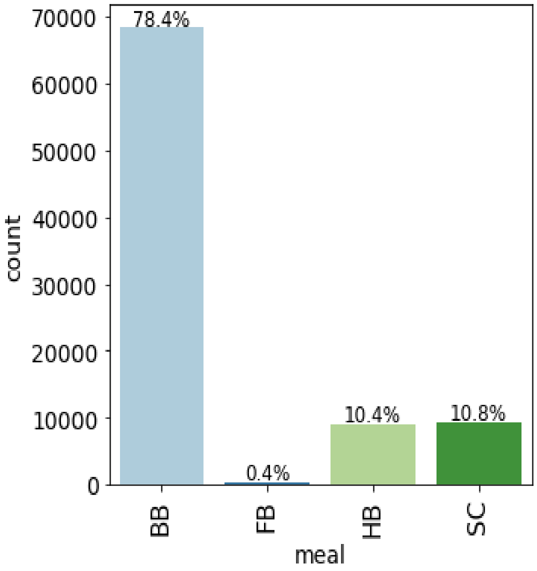


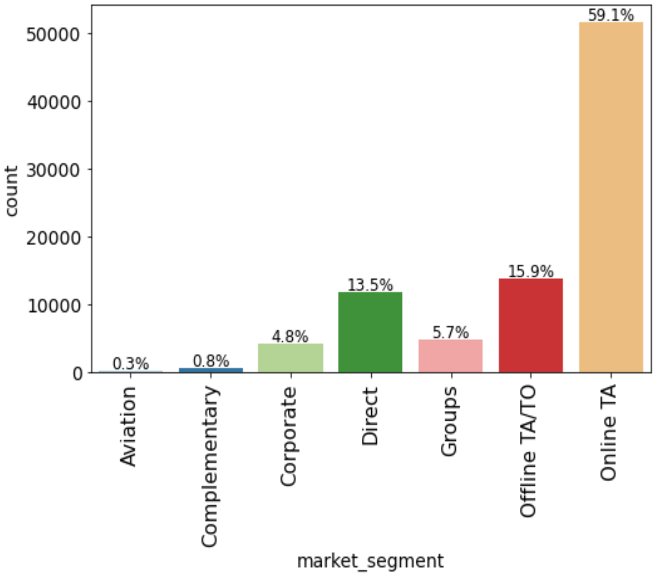
Meal

- Most of the people prefer breakfast only in their meal plan
- There are very few cases where people chose Full Board i.e Breakfast,Lunch as well Dinner around 0.4%

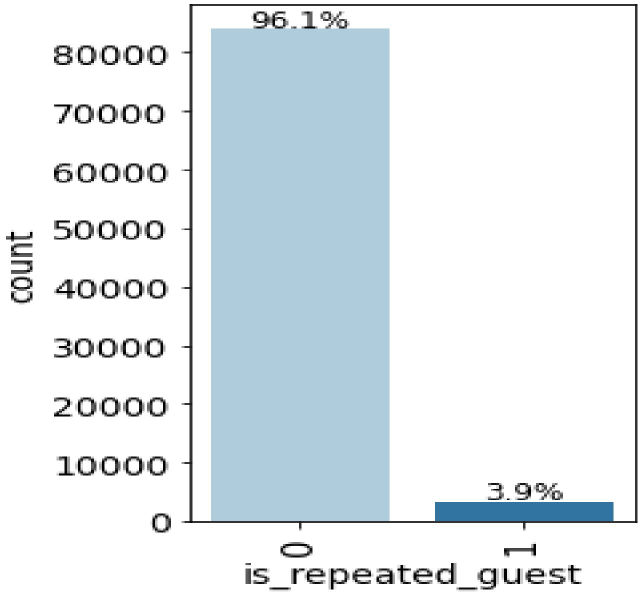
Market_segment

- Most of the people booked online through Travel Agent(for instance: MakeMyTrip)
- Very few cases of booking for the hotel by Aviation company.
- There are very few cases where visitors were offered free stays at the hotel

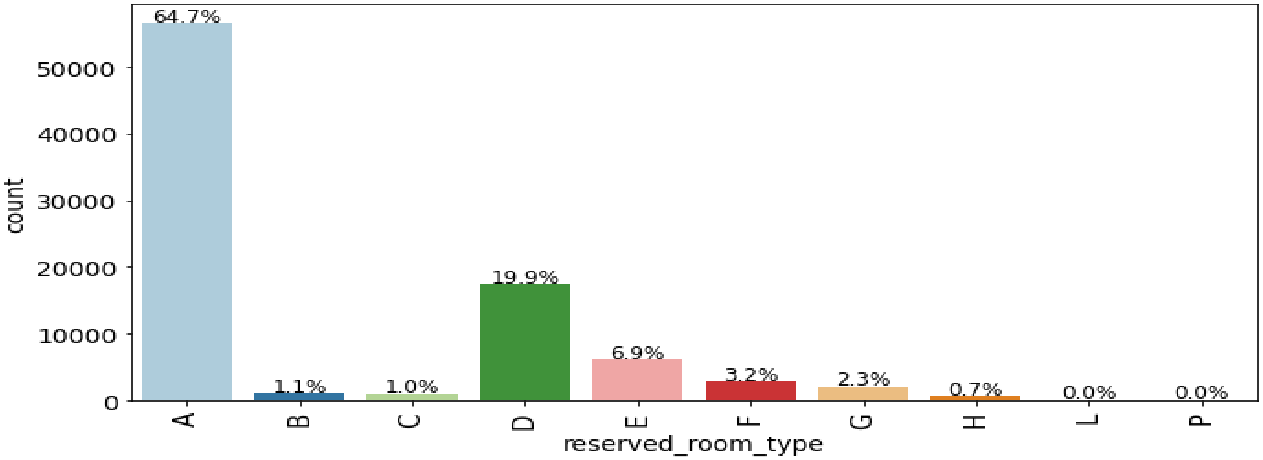




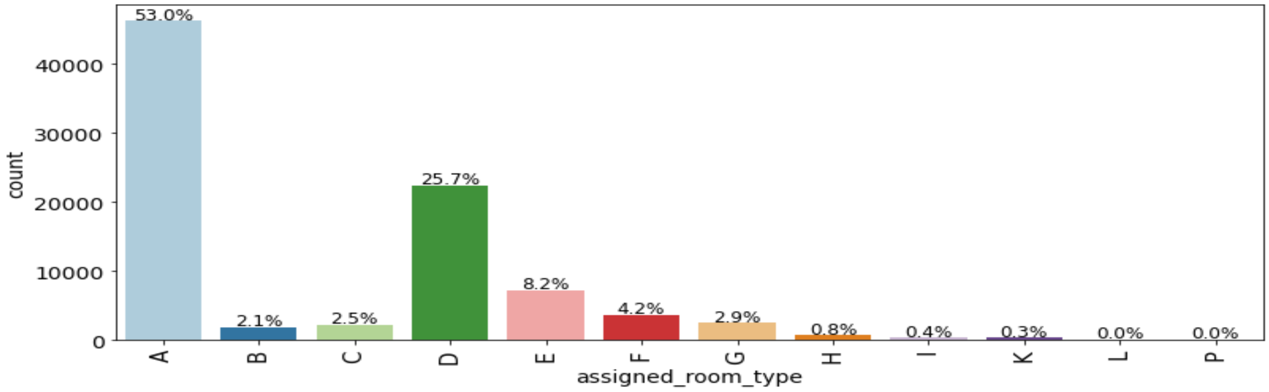
- There are very few cases of repeated visitors who made the booking. Otherwise mostly people booked for the first time to the hotel



- Around 64.7% of the customers booked room type A. It is the most demanded room type followed by D.



- Around 53% bookings were assigned room type A.
- 11.7% of the customers who opted for room type A were assigned another room type due to high chances of unavailability of room type A.

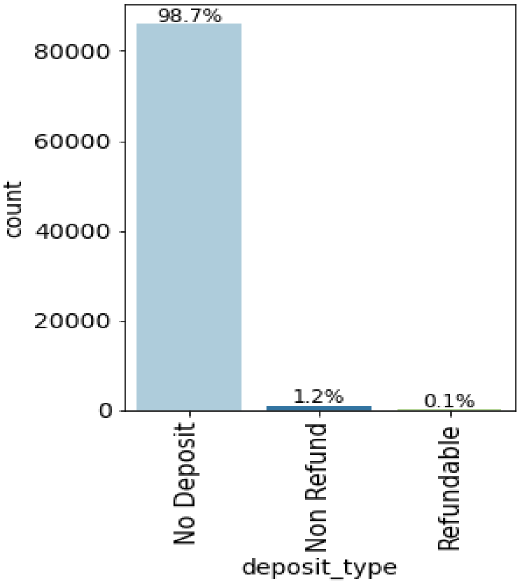


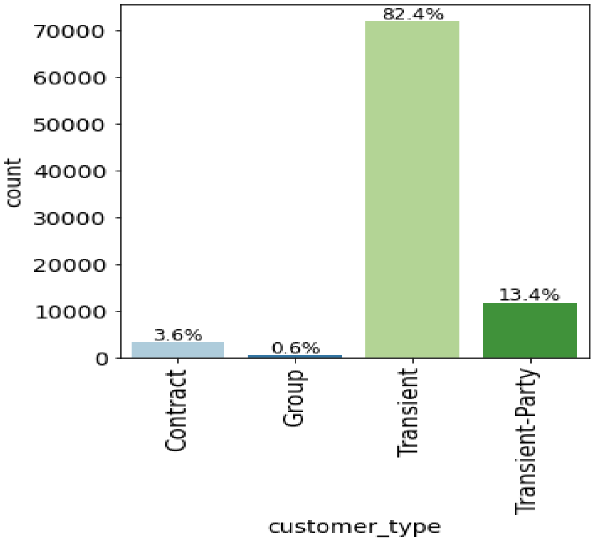
Deposit_type

- Around 98.7% people made no deposit to guarantee their booking.

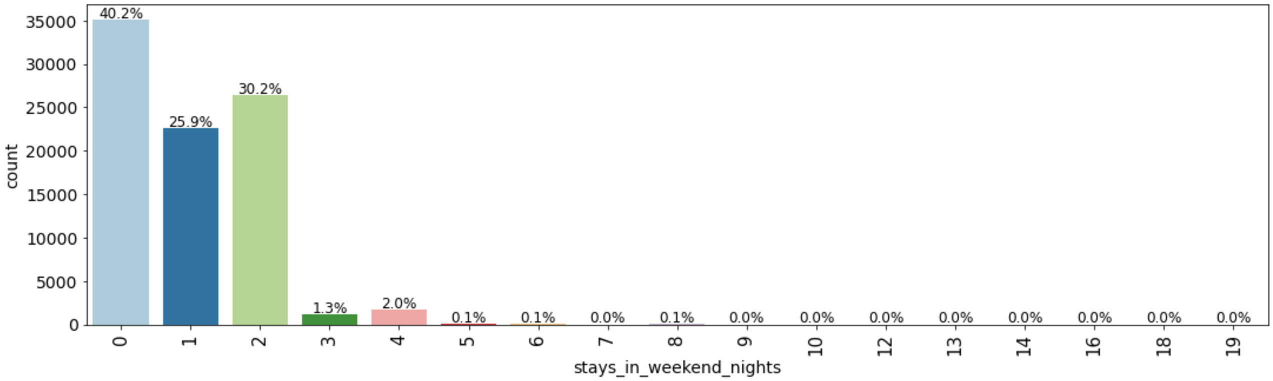
Customer_type

- Most of the bookings (around 82.4%) are Transient (customers who made last minute booking or came for short stays at the hotel) followed by Transient party.

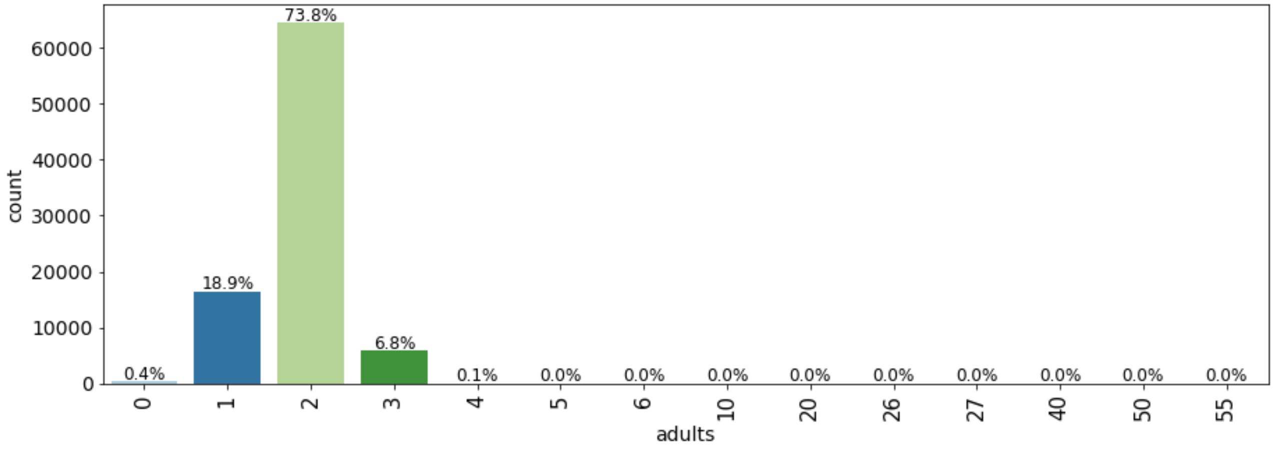




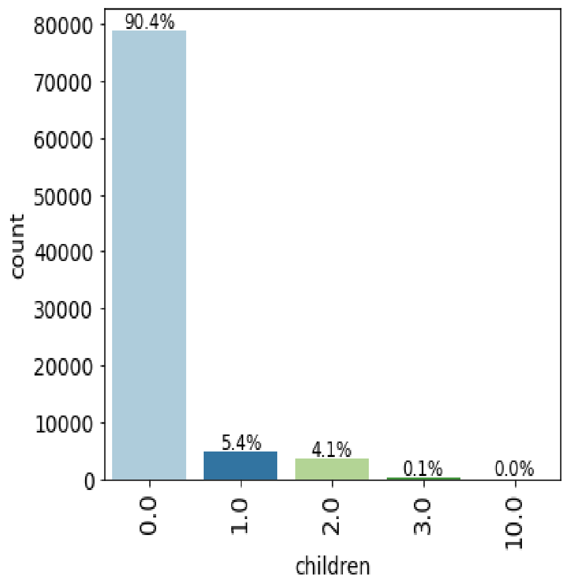
- Half of the bookings are for a weekend
- Around 40% visitors did not plan to spend weekend on the hotel

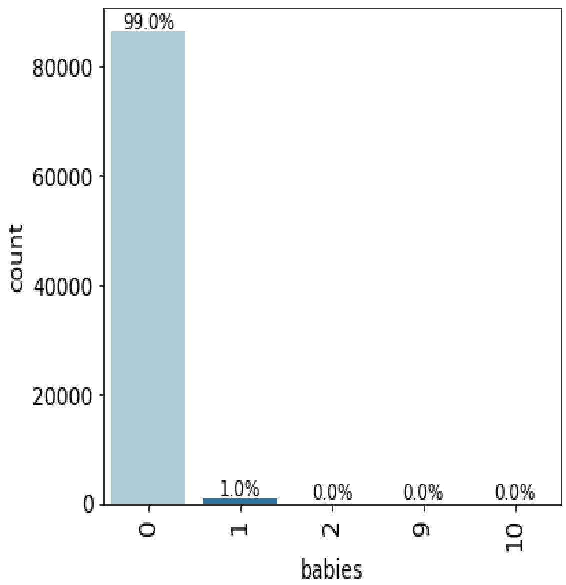


- Mostly 2 people planned to visit the hotel.

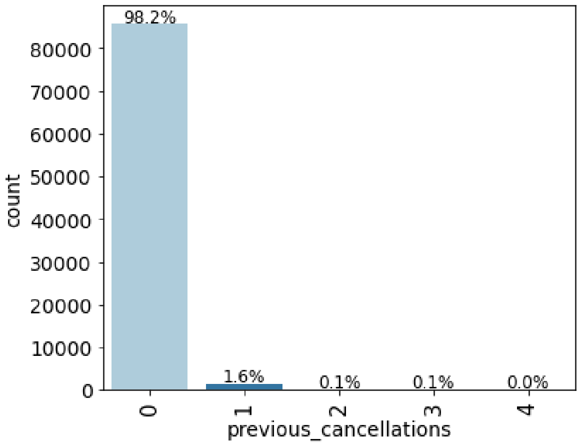


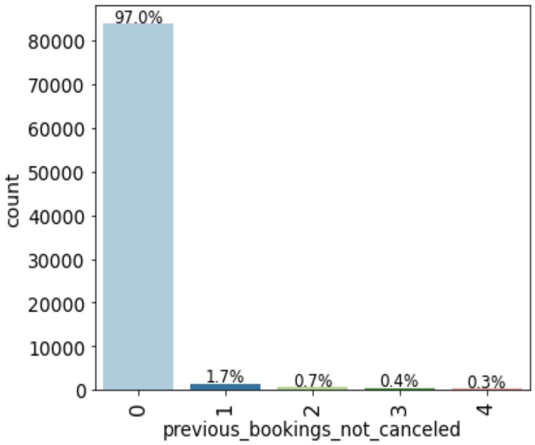
- Mostly people didn't plan to bring their children/babies



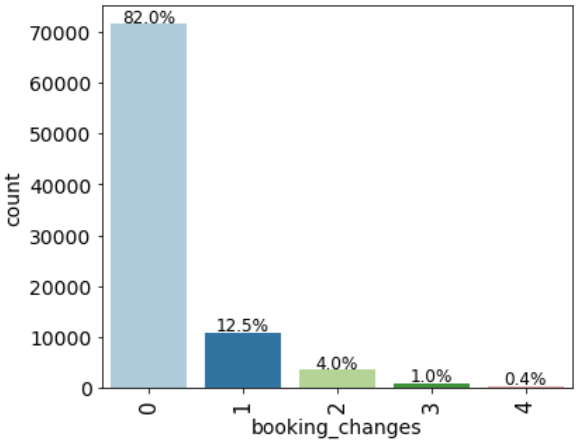


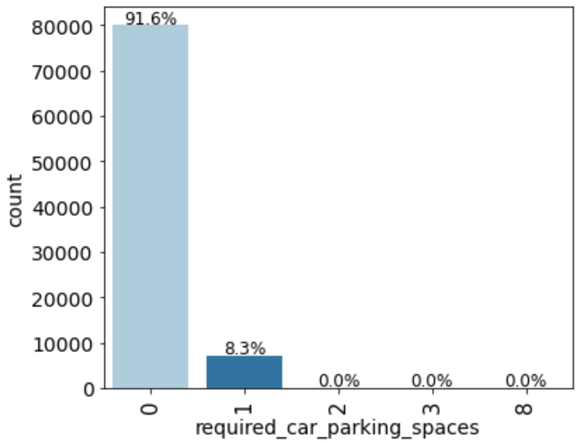
- Mostly customers did not cancel their bookings prior to the current booking
- Mostly number of bookings(around 96.4%) are not cancelled by the customer prior to the current booking



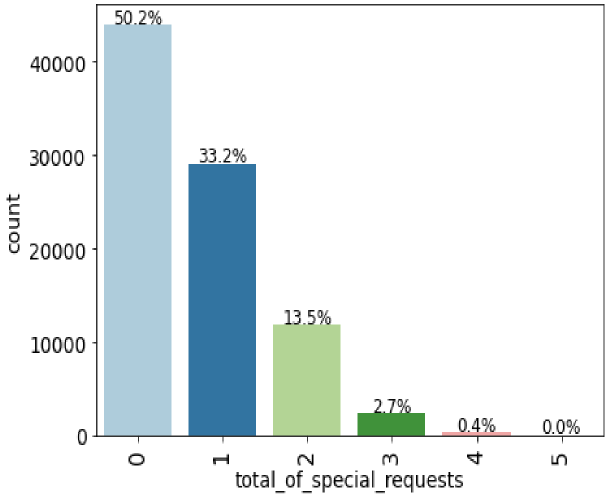


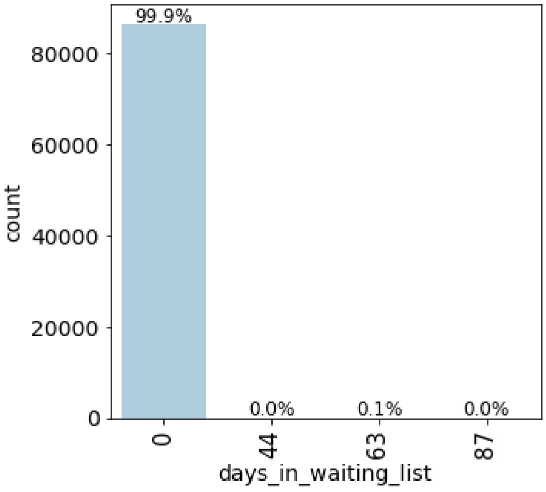
- Mostly customers did not make any amendments in their bookings
- Mostly customers did not require any parking spaces
- Around 8.3% customers require one parking spaces





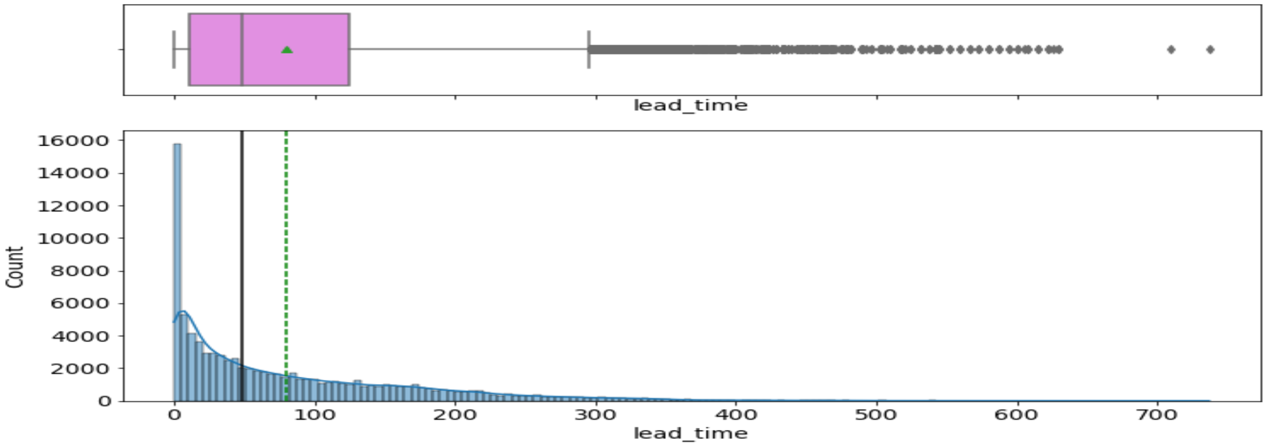
- Around half of the bookings do not have any special requests
- Around one-third of bookings have one special request
- Around 99.9% bookings were booked instantly by the hotel



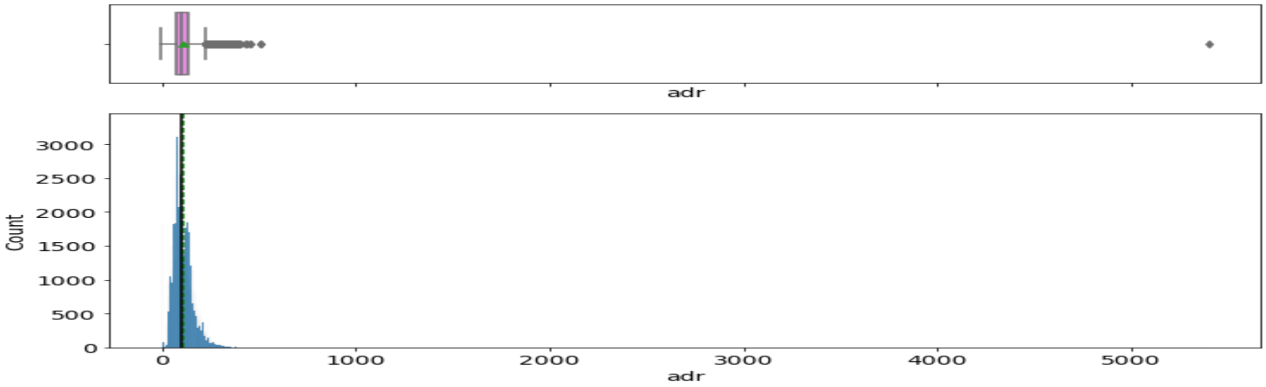


Numerical Columns

- There are large number of outliers above the upper whisker.That's why mean has shifted towards right
- The Distribution is skewed to the right
- Large number of people made their booking on the same day of arrival

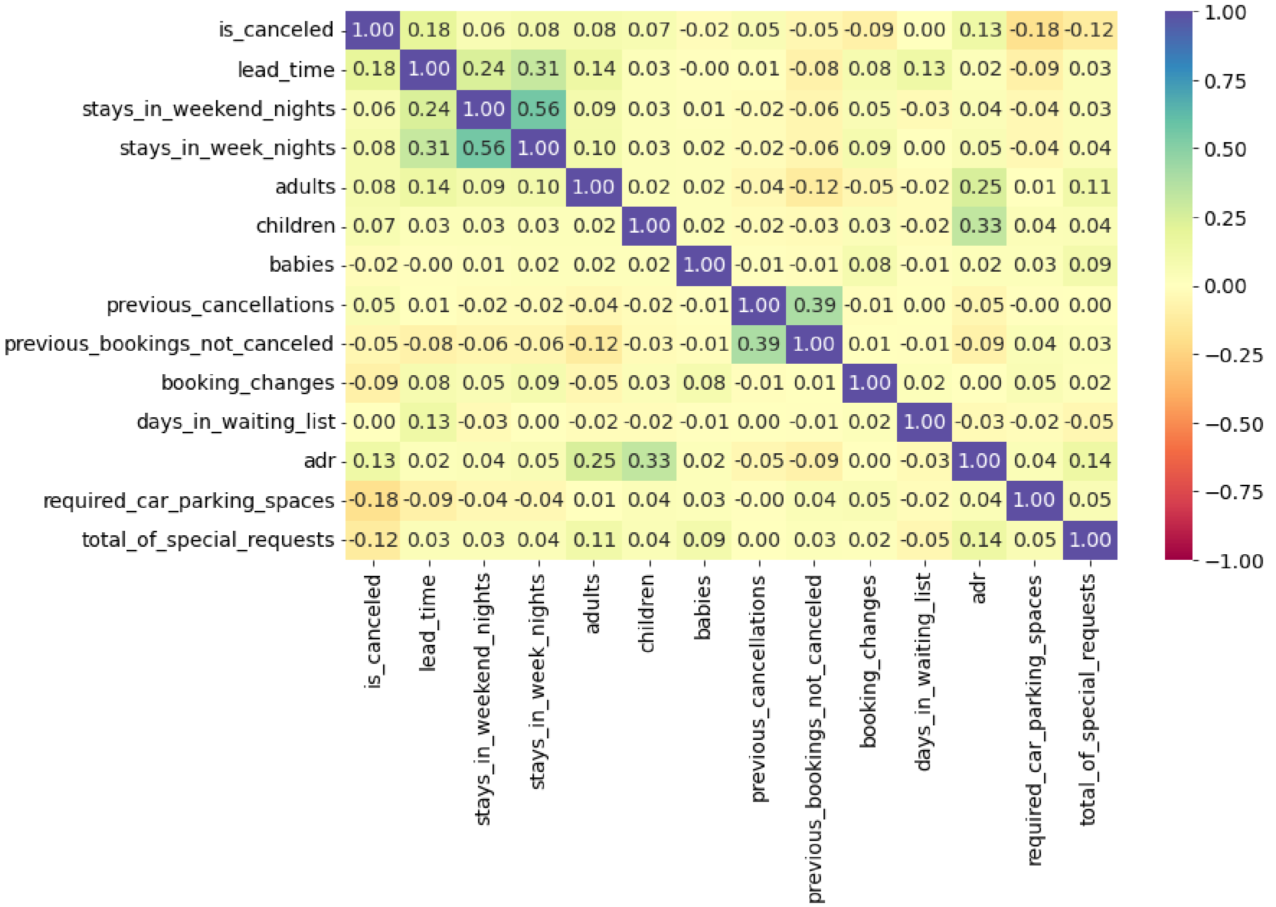


- Distribution is right skewed.
- There are significant number of bookings where average daily rate is 0. These are complimentary bookings.



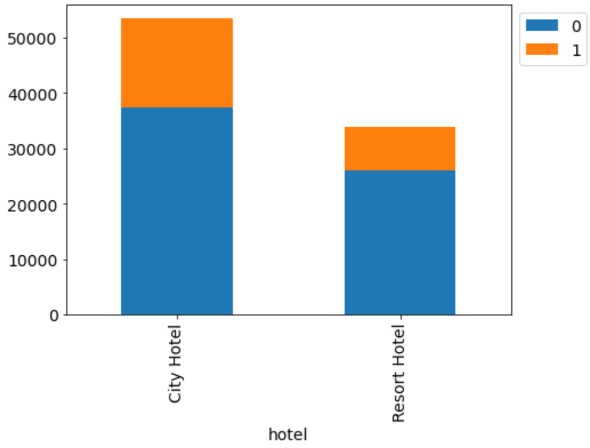
Multivariate Analysis:

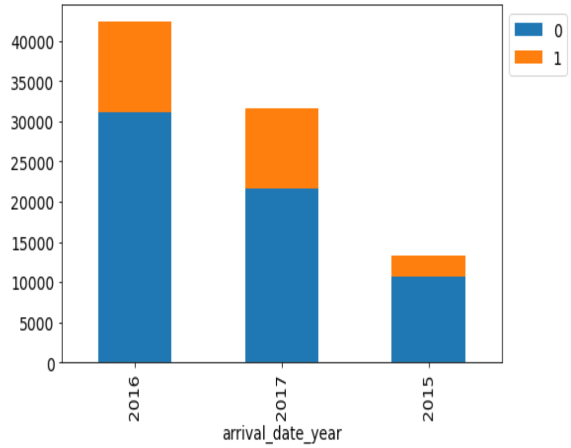
- Number of stays in weekend nights and number of stays in weekday nights are moderate positively correlated



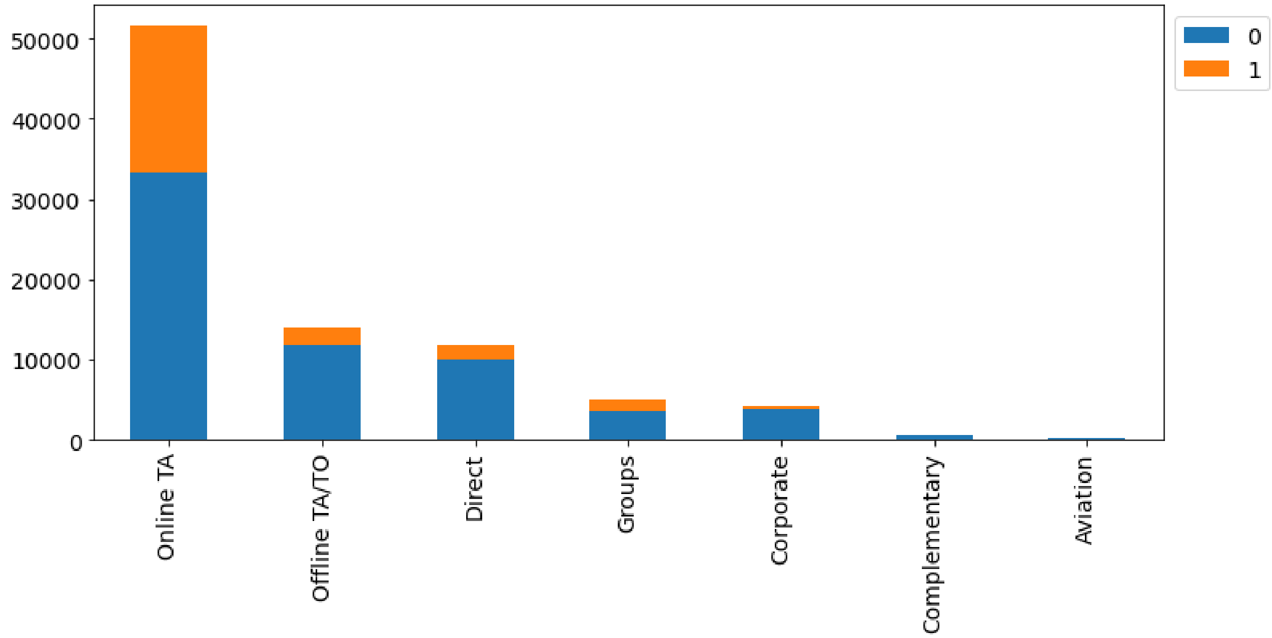
	is_canceled	lead_time	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	previous_cancellations	previous_bookings_not_canceled	booking_changes	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests
is_canceled	1.00	0.18	0.06	0.08	0.08	0.07	-0.02	0.05	-0.05	-0.09	0.00	0.13	-0.18	-0.12
lead_time	0.18	1.00	0.24	0.31	0.14	0.03	-0.00	0.01	-0.08	0.08	0.13	0.02	-0.09	0.03
stays_in_weekend_nights	0.06	0.24	1.00	0.56	0.09	0.03	0.01	-0.02	-0.06	0.05	-0.03	0.04	-0.04	0.03
stays_in_week_nights	0.08	0.31	0.56	1.00	0.10	0.03	0.02	-0.02	-0.06	0.09	0.00	0.05	-0.04	0.04
adults	0.08	0.14	0.09	0.10	1.00	0.02	0.02	-0.04	-0.12	-0.05	-0.02	0.25	0.01	0.11
children	0.07	0.03	0.03	0.03	0.02	1.00	0.02	-0.02	-0.03	0.03	-0.02	0.33	0.04	0.04
babies	-0.02	-0.00	0.01	0.02	0.02	0.02	1.00	-0.01	-0.01	0.08	-0.01	0.02	0.03	0.09
previous_cancellations	0.05	0.01	-0.02	-0.02	-0.04	-0.02	-0.01	1.00	0.39	-0.01	0.00	-0.05	-0.00	0.00
previous_bookings_not_canceled	-0.05	-0.08	-0.06	-0.06	-0.12	-0.03	-0.01	0.39	1.00	0.01	-0.01	-0.09	0.04	0.03
booking_changes	-0.09	0.08	0.05	0.09	-0.05	0.03	0.08	-0.01	0.01	1.00	0.02	0.00	0.05	0.02
days_in_waiting_list	0.00	0.13	-0.03	0.00	-0.02	-0.02	-0.01	0.00	-0.01	0.02	1.00	-0.03	-0.02	-0.05
adr	0.13	0.02	0.04	0.05	0.25	0.33	0.02	-0.05	-0.09	0.00	-0.03	1.00	0.04	0.14
required_car_parking_spaces	-0.18	-0.09	-0.04	-0.04	0.01	0.04	0.03	-0.00	0.04	0.05	-0.02	0.04	1.00	0.05
total_of_special_requests	-0.12	0.03	0.03	0.04	0.11	0.04	0.09	0.00	0.03	0.02	-0.05	0.14	0.05	1.00

- Ratio of canceled to not canceled in city hotel is 1:2 and 1:3 in resort hotel
- Cancellations in 2015 were around 20% which is less as compared to 2016 and 2017 where it was more than 30%

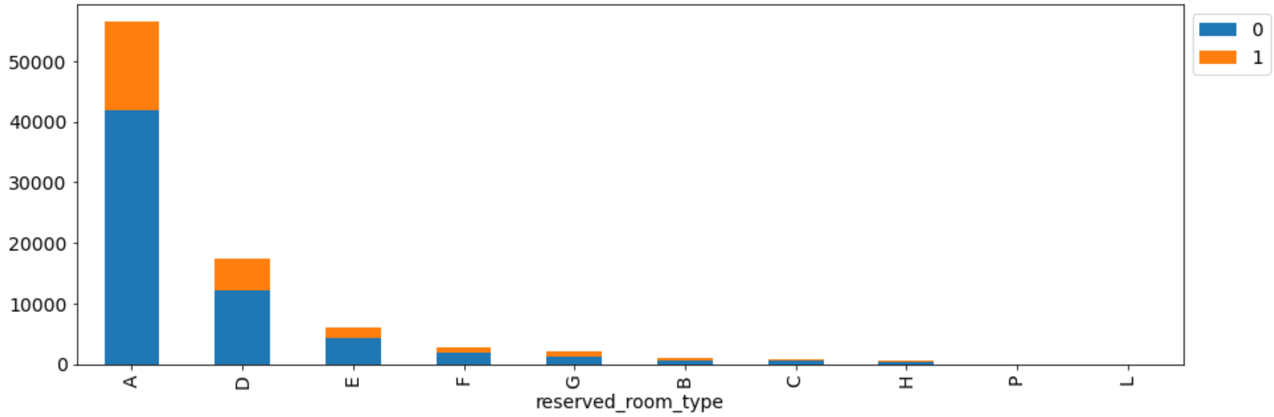




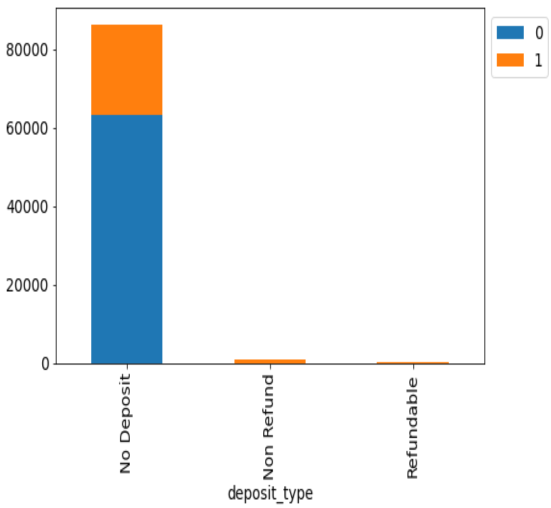
- Mostly cancellations are in the case of bookings via online travel agent

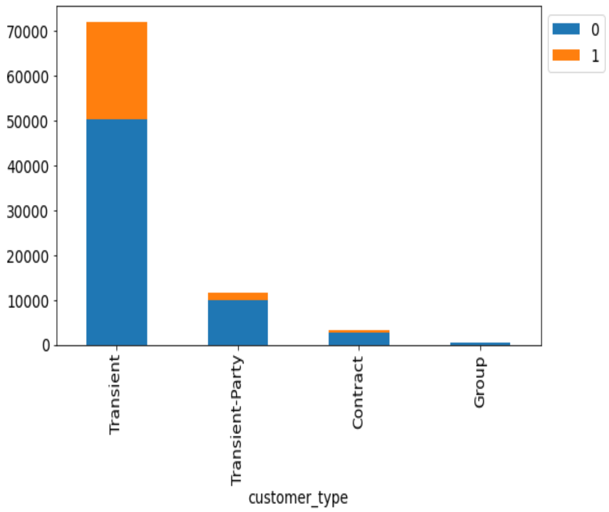


- Around 16% of the total bookings are cancelled for room type A

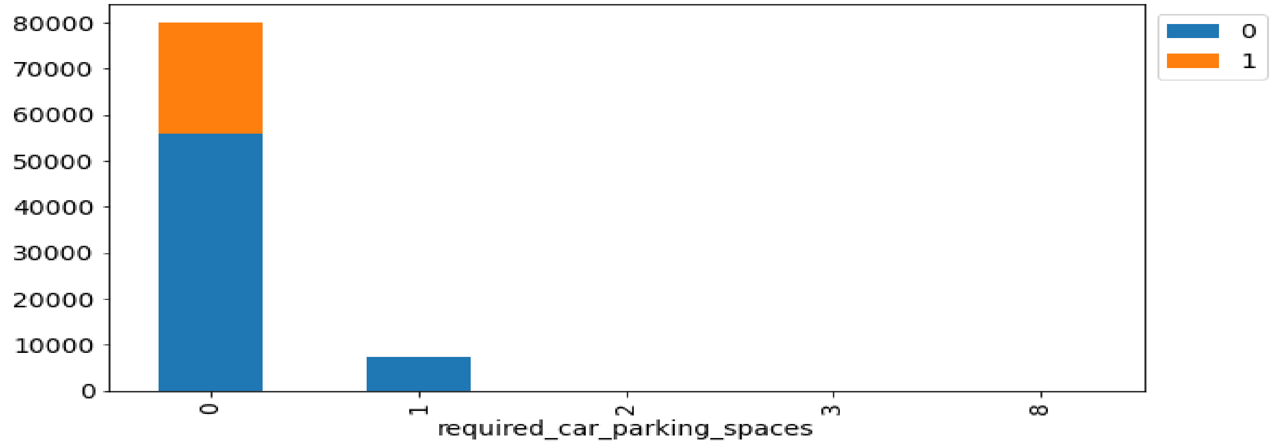


- Cancellations in case where customer made no deposit to guarantee their booking are higher
- Cancellations in case of Transient booking are higher

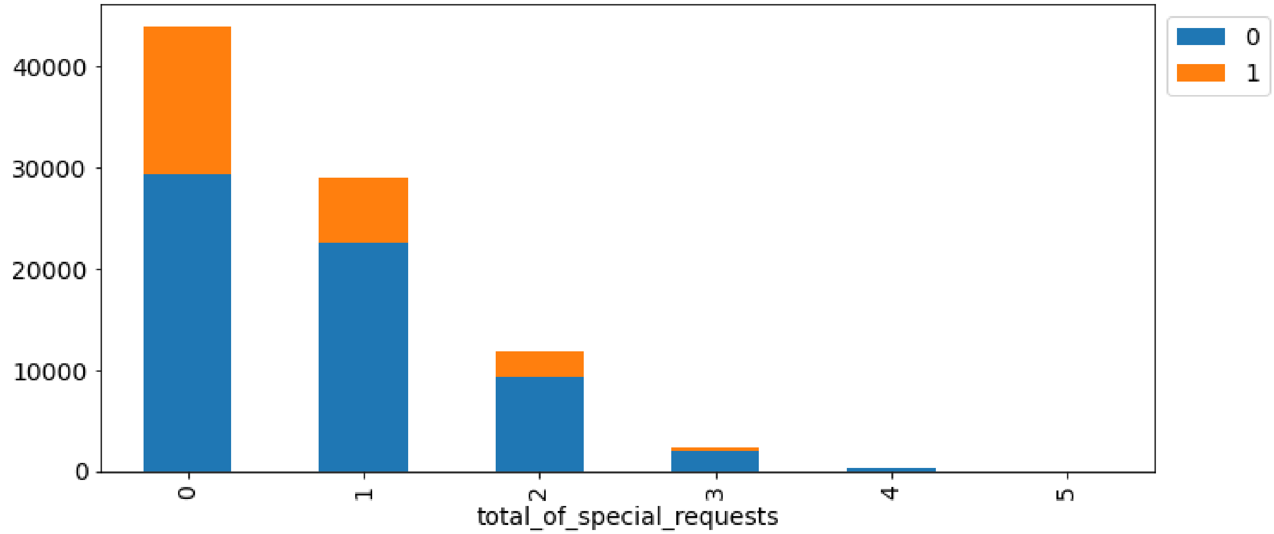




- Cancellations are only in the case where car parking space is not required



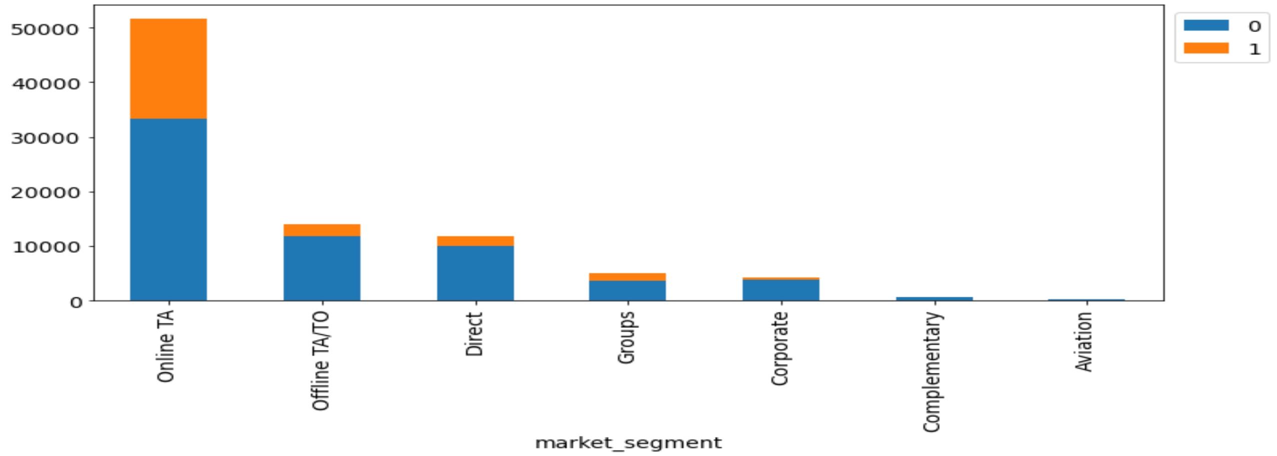
- As number of special requests rises chances of cancellation drops

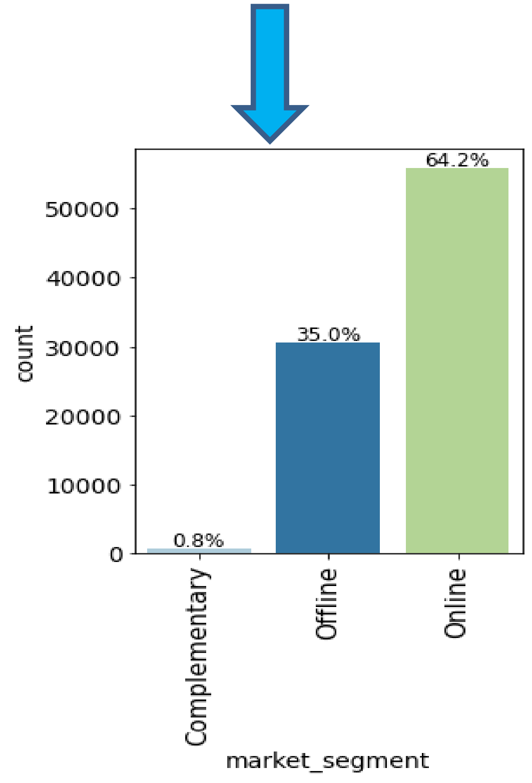


Derive New Features:

There are mostly categorical features in the dataset. We have derived new features in order to reduce complexity while creating dummy variables

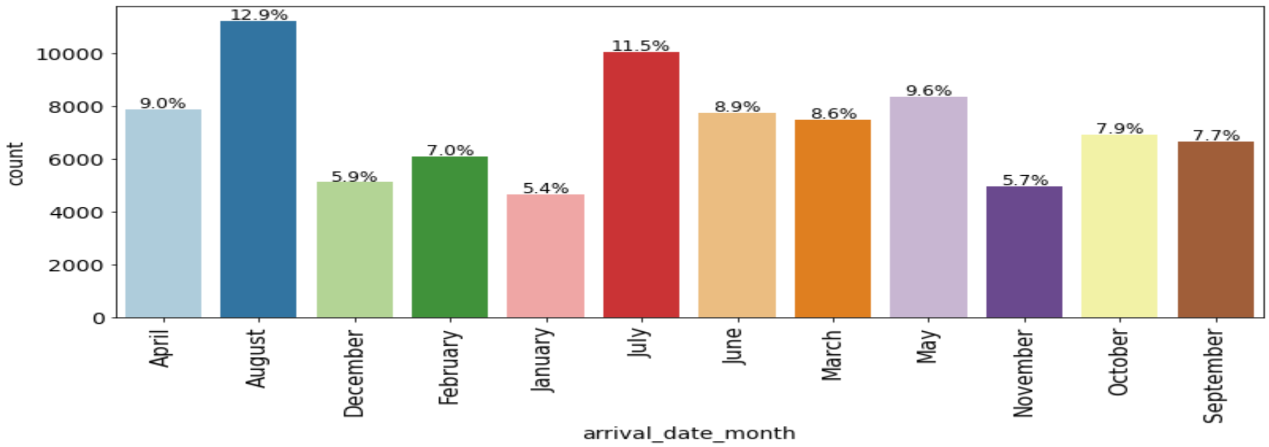
market_segment

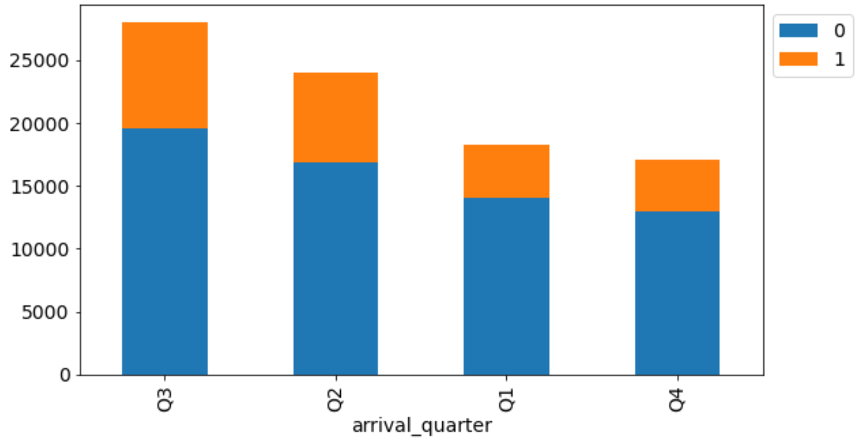




- Online: Aviation + Online TA + Corporate
- Offline: Offline TA/TO + Direct + Groups

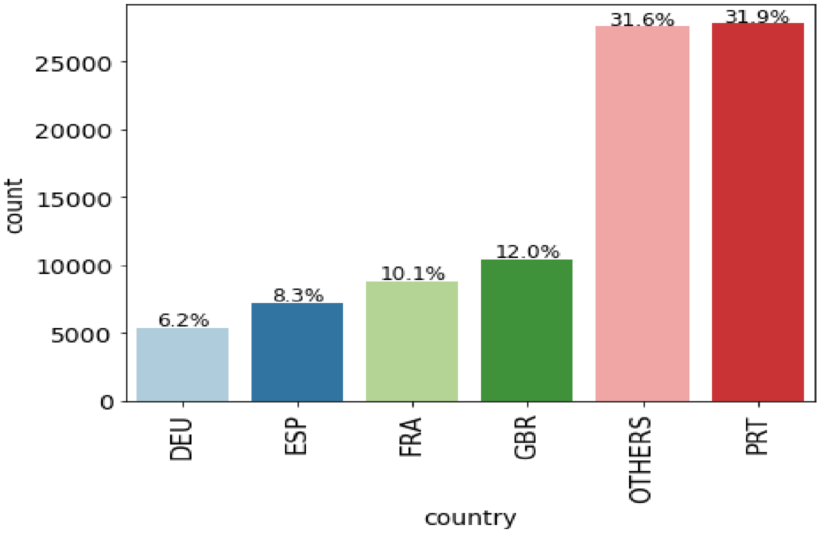
arrival_quarter





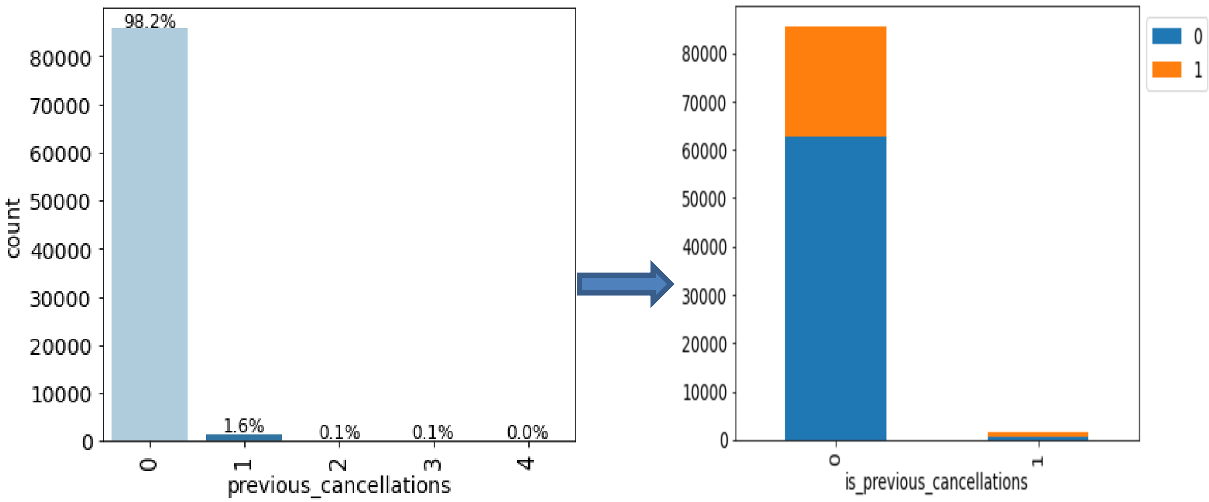
- Q3 has the highest bookings as well as the highest cancellations followed by Q2

country

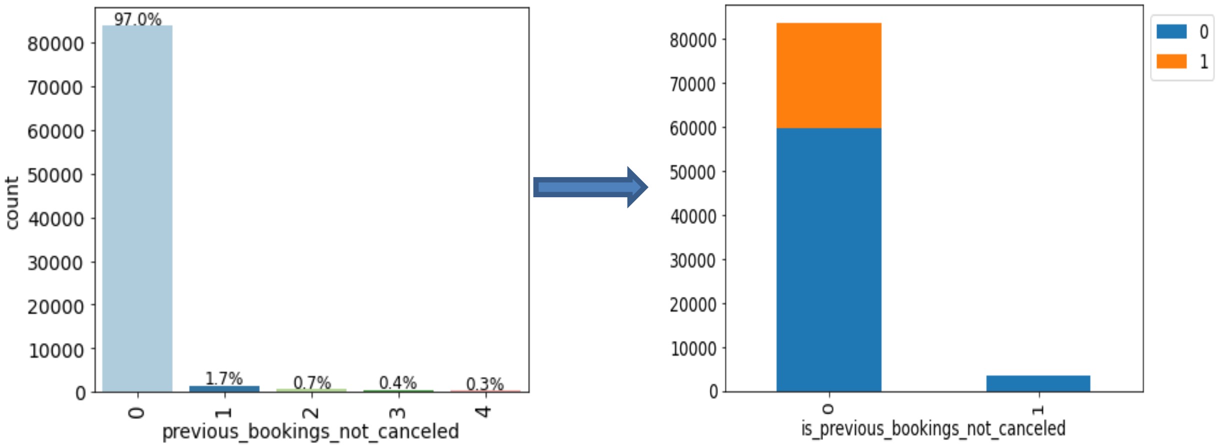


- Top five countries has around 70% data. So, rest of the countries are classified as ‘OTHERS’

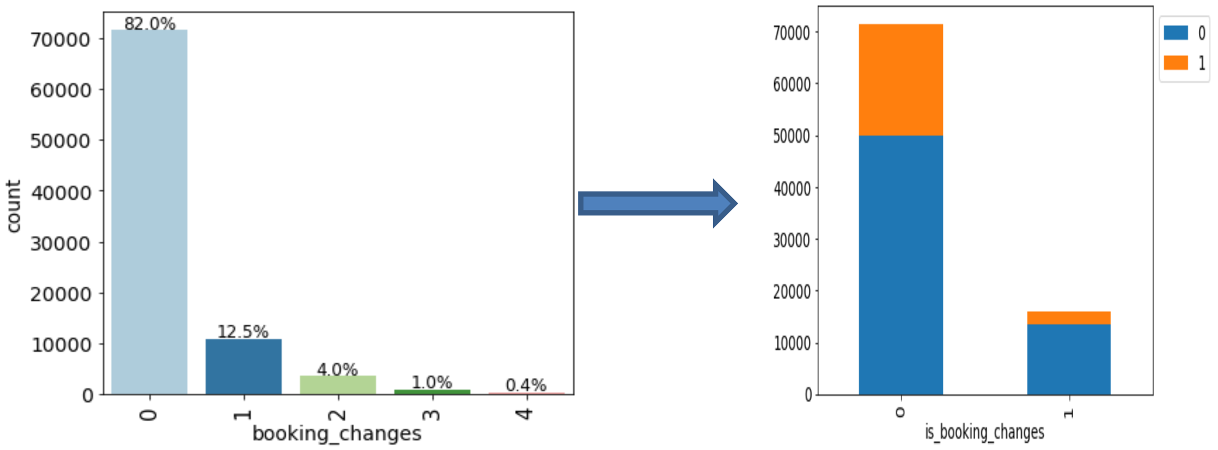
is_previous_cancellations



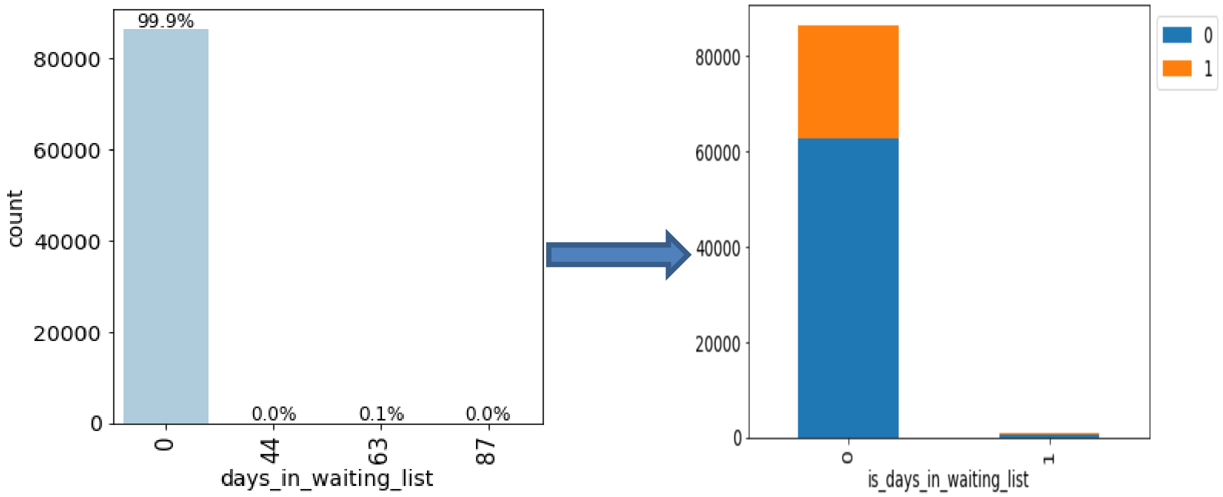
is_previous_bookings not cancelled



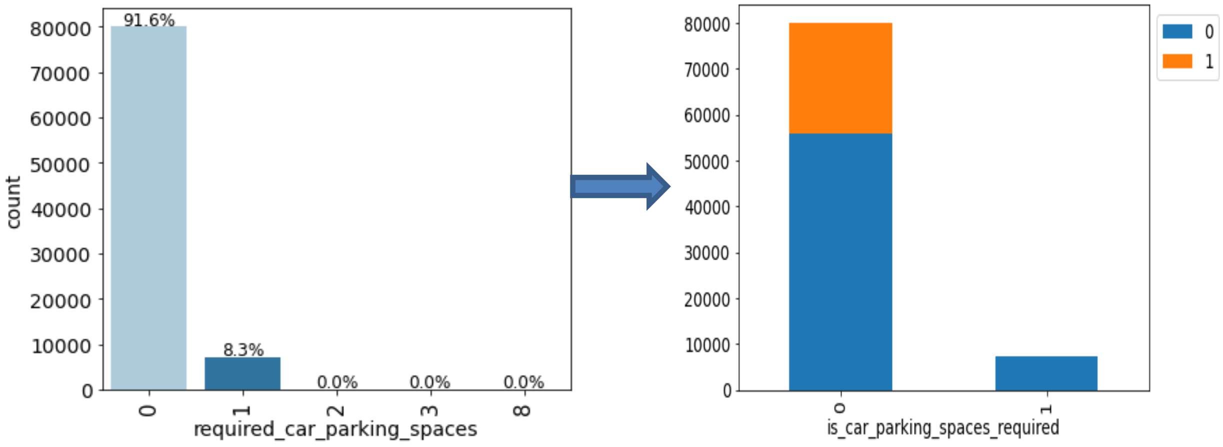
is_booking Changes



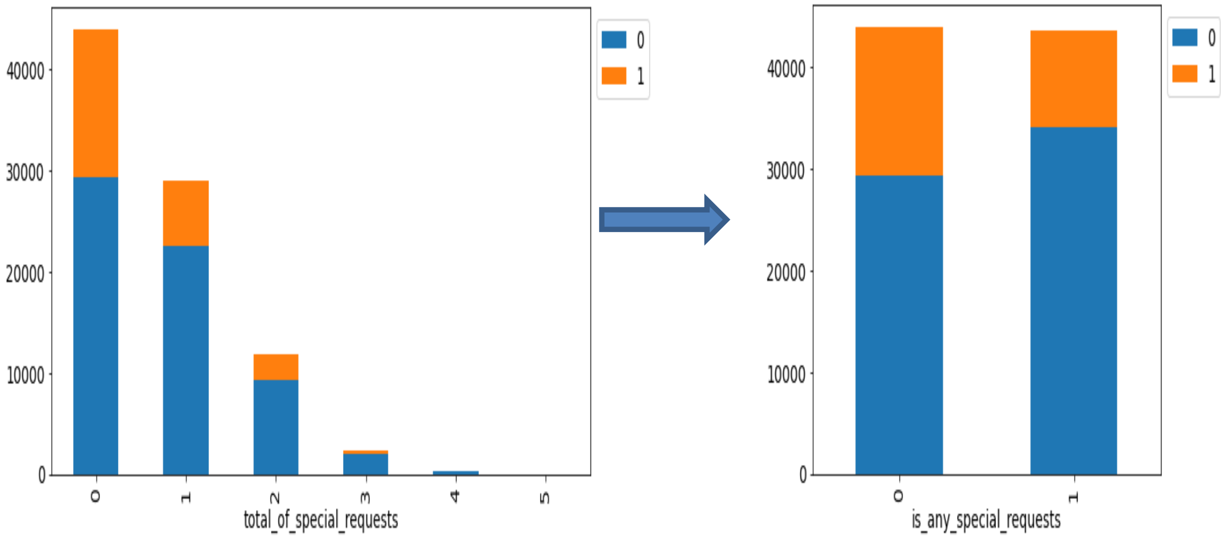
Is_days_in_waiting_list



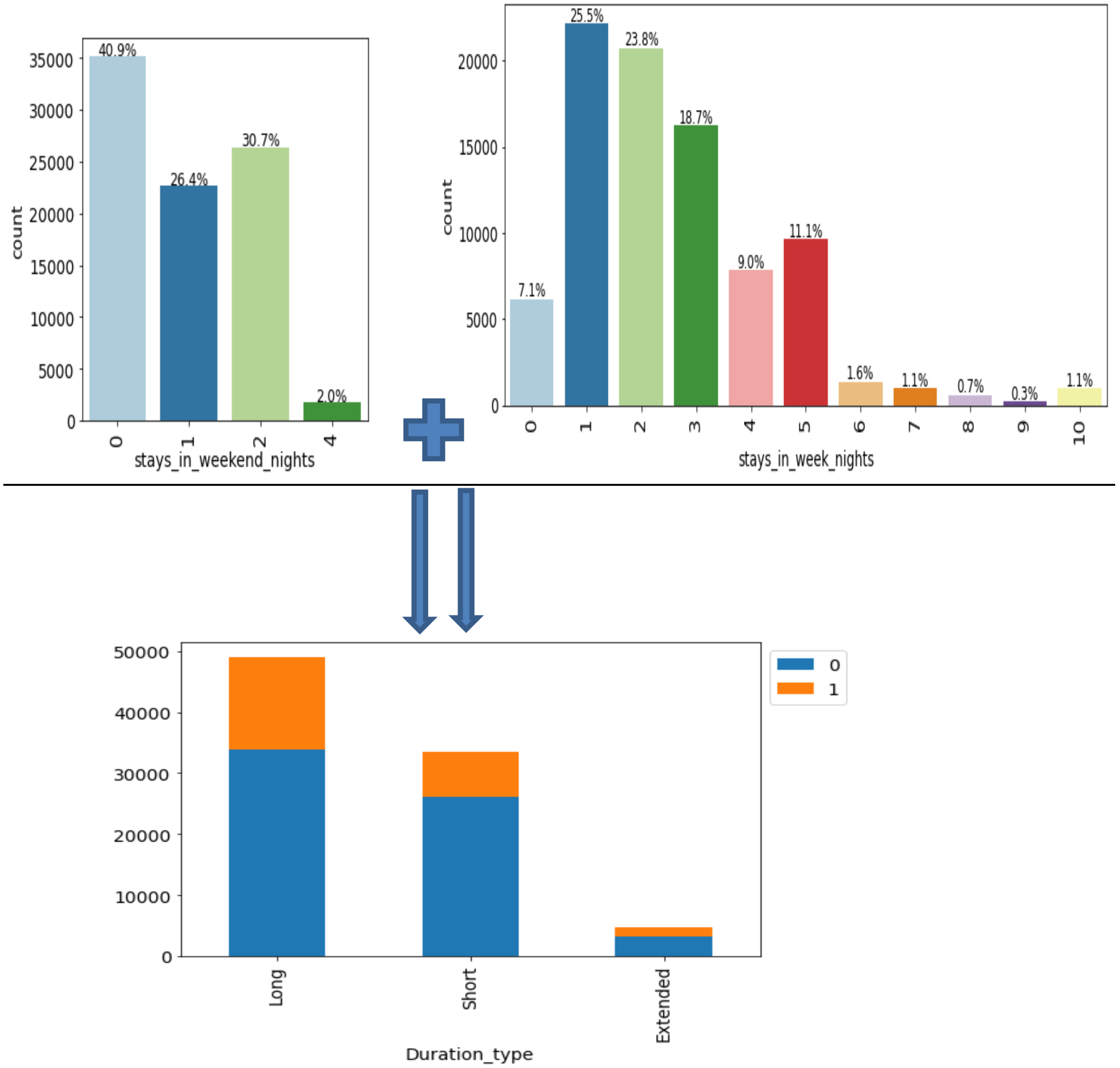
Is_car_parking_space_required



Is_any_special request

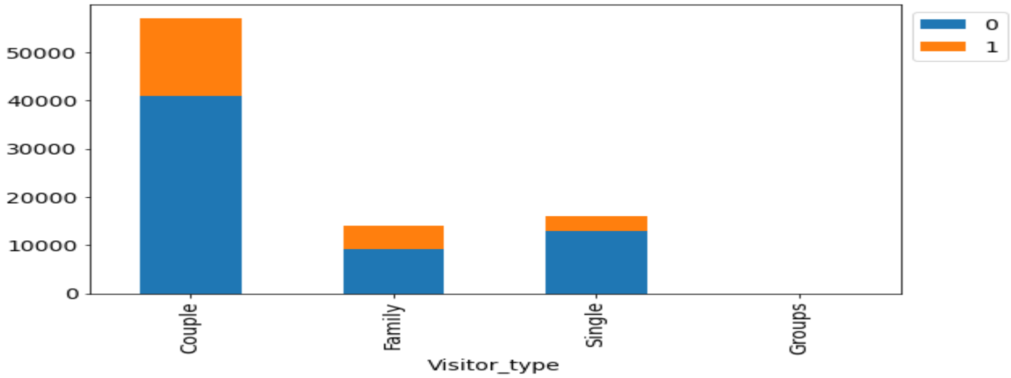


Duration_type



- Total_nights = stays_in_weekend_nights + stays_in_week_nights
- Duration_type :
  1. Short: Total_nights <3
  2. Long: Total_nights >=3 and <=7
  3. Extended: Total_nights >7

Visitor_type



- Total_Visitors = adults + children + babies
- Visitor_type:
    1. Single: Total_Visitors=1
    2. Couple: Total_Visitors=2
    3. Family: Total_Visitors >=3 and <=6
    4. Groups: Total_Visitors >6. In the dataset ,minimum number of visitors in case of groups are 10

Statistical Analysis

Statistical tests were performed to see the whether each of the 24 independent variables have a significant relationship with the dependent variable, is_canceled

- For making any statistical inference, the confidence level has been chosen as 95 percent. This implies level of significance would be 1-0.95=0.05 percent
- If p-value is greater than level of significance, we fail to reject the null hypothesis
- If p-value is less than or equal to level of significance, we reject the null hypothesis
- P-value: It is the probability of getting a sample if the null hypothesis is true (or alternate hypothesis is false)

Chi-square Test:

For the Categorical Columns, a Chi-square Test of independence was performed with the target variable, DEFAULT which is also a categorical column.
Here **Null Hypothesis H0:** There is NO association between the two variables
And **Alternate Hypothesis Ha:** There is an association between the two variables.

Variable	P-value	Significant
hotel	1.5219752505471974e-99	yes
arrival_date_year	1.3085557935705007e-146	yes
arrival_quarter	5.867865786651421e-94	yes
meal	5.037896681861171e-72	yes
country	7.335438053713435e-104	yes
market_segment	0.0	yes
distribution_channel	0.0	yes
is_repeated_guest	5.166277981932753e-150	yes
is_previous_cancellations	2.939461189259926e-306	yes
is_previous_bookings_not_canceled	2.2643683773209235e-204	yes
reserved_room_type	7.366950629174164e-29	yes
assigned_room_type	1.8455256208956267e-111	yes

is_booking_changes	3.534605340933596e-303	yes
deposit_type	0.0	yes
is_days_in_waiting_list	3.792658631749707e-05	yes
customer_type	3.554407830443141e-304	yes
is_car_parking_spaces_required	0.0	yes
is_any_special_requests	0.0	yes
Duration_type	1.6490264836610937e-182	yes
Visitor_type	1.7450038806652548e-192	yes

All variables have significant relationship with the target variable

ANOVA test:

For all the numeric variables, ANOVA test was performed between values of the variable for two classes of target variables to compare their means.

Here **Null Hypothesis H0:** The means of the two samples are EQUAL
And **Alternate Hypothesis Ha:** The means of the two samples are NOT EQUAL.

If the means of the two samples are significantly different form each other, then we can conclude that the variable does have a significant relationship with the target variable.

Variables	P-value	Significant
lead_time	0.00000	yes
adr	0.00000	yes

Base Models

Before proceeding with the predictive modelling for the problem statement, we start with some baseline models to give some sense of how various classification models perform on our dataset. In the due process we have deployed Logistic Regression, Decision Trees, Random Forest, K Nearest Neighbors and Light gradient boosted machine.

We mainly concentrated on the following metrics:

- **Accuracy**: Overall, how often is the classifier correct?
- **Precision**: It is used to depict what proportion of bookings that the model predicted to be cancelled, actually were cancelled.
- **Recall**: It is used to depict what proportion of bookings that were actually cancelled, were correctly predicted to be cancelled by the model.
- **F1-score**: It is the harmonic mean of precision and recall.
- **ROC AUC**: The area under the curve obtained for True Positive Rate and False Positive Rate tells us how much the model is capable of distinguishing between the classes.
- **Cohen's Kappa**: It is used to measure inter-annotator agreement. It simply highlights prediction power of an imbalanced target class.

Training data (before SMOTE)

```
Before OverSampling, the shape of train_X: (61042, 56)
Before OverSampling, the shape of train_y: (61042,)

Before OverSampling, counts of label '1': 16841
Before OverSampling, counts of label '0': 44201
```

Model		Accuracy	Precision	Recall	F1 Score	ROC AUC	COHEN'S KAPPA
Logistic Regression	Train	0.73	0.76	0.01	0.01	0.50	0.0089
	Test	0.727	0.77	0.01	0.01		
Decision Tree	Train	0.84	0.73	0.65	0.69	0.75	0.51
	Test	0.81	0.68	0.61	0.64		
Random Forest	Train	0.90	0.86	0.75	0.80	0.765	0.56
	Test	0.84	0.75	0.61	0.67		
KNN	Train	0.83	0.64	0.91	0.75	0.726	0.46
	Test	0.74	0.63	0.58	0.61		
LGBM Classifier	Train	0.84	0.76	0.63	0.69	0.77	0.56
	Test	0.84	0.74	0.62	0.68		

Inferences:

- From above metrics we can clearly see that LGBM classifier and Random Forest Classifier is giving out the best results
- ROC AUC is similar between them

Final Models

As we know that there is an imbalance between both the classes of target variable. So, we'll rebuild the model using proper imbalanced data handling techniques
Various approaches to handle imbalanced classification problem :-

1. Data level approach: Resampling Technique
   - Random-Under Sampling:

     Advantages: It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.

     Disadvantages:
     - It can discard potentially useful information which could be important for building rule classifiers.
     - The sample chosen by random under sampling may be a biased sample. And it will not be an accurate representative of the population. Thereby, resulting in inaccurate results with the actual test data set.

   - Random-Over Sampling:

     Advantages: Unlike under sampling this method leads to no information loss.Outperforms under sampling

     Disadvantages: It increases the likelihood of overfitting since it replicates the minority class events.

2. **Cluster-Based Over Sampling:**

   Advantages:
   - This clustering technique helps overcome the challenge between class imbalance. Where the number of examples representing positive class differs from the number of examples representing a negative class.
   - Also, overcome challenges within class imbalance, where a class is composed of different sub clusters. And each sub cluster does not contain the same number of examples.

   Disadvantages: The main drawback of this algorithm, like most oversampling techniques is the possibility of over-fitting the training data.

3. **Informed Over Sampling: Synthetic Minority Over-sampling Technique for imbalanced data(SMOTE)**

   Advantages:
   - Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances.
   - No loss of useful information

   Disadvantages:

   - While generating synthetic examples SMOTE does not take into consideration neighbouring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise.
   - SMOTE is not very effective for high dimensional data

   So, we'll apply SMOTE to the training data as it is the best among all the other imbalanced data handling technique.

Training Data (After SMOTE)

```
After OverSampling, the shape of train_X: (88402, 56)
After OverSampling, the shape of train_y: (88402,)

After OverSampling, counts of label '1': 44201
After OverSampling, counts of label '0': 44201
```

Model		Accuracy	Precision	Recall	F1 Score	ROC AUC	COHEN'S KAPPA
Logistic Regression	Train	0.76	0.75	0.79	0.77	0.734	0.417
	Test	0.74	0.52	0.70	0.60		
Decision Tree	Train	0.85	0.82	0.88	0.85	0.72	0.46
	Test	0.78	0.58	0.68	0.63		
Random Forest	Train	0.91	0.86	0.96	0.91	0.80	0.557
	Test	0.81	0.61	0.80	0.69		
KNN	Train	0.88	0.83	0.95	0.89	0.748	0.432
	Test	0.74	0.52	0.77	0.62		
LGBM Classifier	Train	0.85	0.82	0.88	0.85	0.80	0.555
	Test	0.81	0.61	0.80	0.69		

Results have improved after applying smote since there's more data for the machine to learn pattern for cancellations.

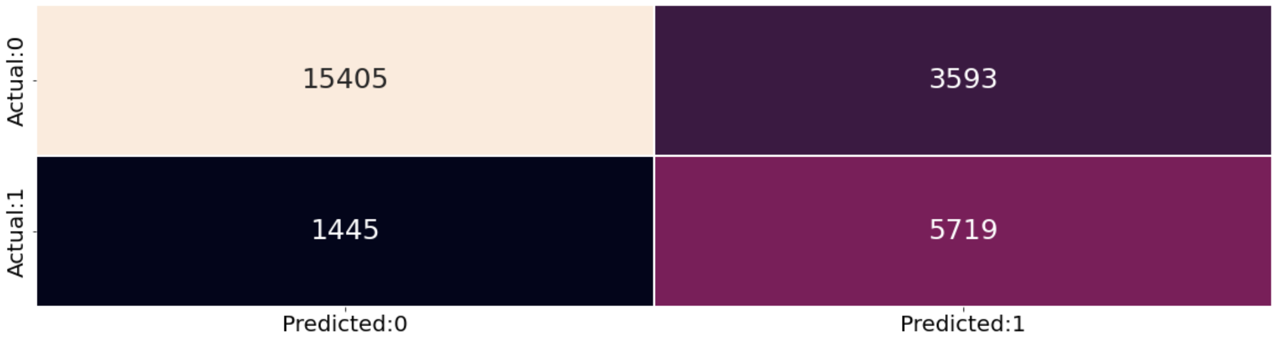
Hyperparameter Tuning:

For tuning the hyperparameter of various models we used GridSearchCV.

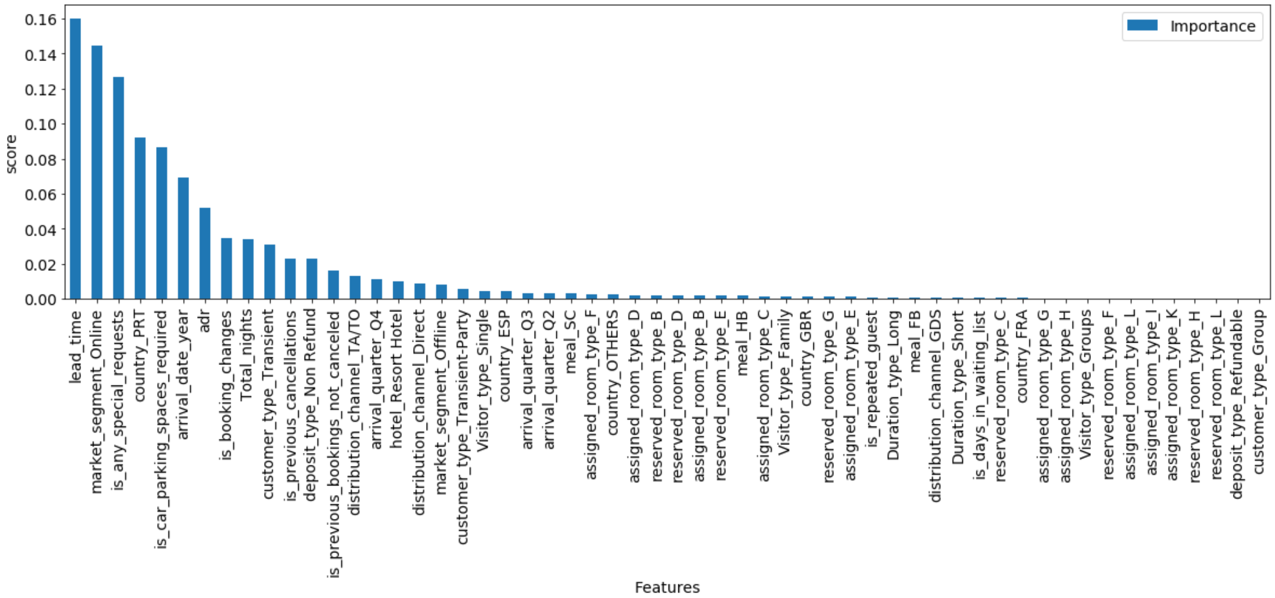
Below is the best parameter for Random Forest Classifier which is so far the best model:

'criterion': 'gini', 'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200

Confusion matrix of Random Forest Classifier(Best Model)



Feature Importance from Random Forest Model:



Our best model has shown lead_time, market_segment(Online), is_any_special_requests as the most important features .

Closing Reflections

Our learning throughout the journey of developing this model has been very insightful. In feature engineering, we tried to derive the new features without losing the importance of them. Also, we tried different algorithms on different thresholds. Our main concern is to reduce the number of False Negatives. If our model would have predicted that the booking is not going to cancel and it gets cancelled, then the hotel has no choice but lowering the price resulting in loss of revenue. So, Recall value has been our major factor to decide the model performance as per our needs.

Limitations

While working on the dataset, we noted few limitations which impacted the expectation of our result

- There were evidences of minority class been already up sampled/synthetically sampled. Possible use of SMOTE and over sampling applied previously.
- Dataset was not oriented towards the actual scenario of hotels in India.
- We could see further scope of improvement. we could see that our developed model had some restrictions to reach its full capacity in generating the desired results.

Business Recommendations

- There are very few repeated guests visited to the hotel. Hotel managers need to focus on increasing repeated customers. Retaining old visitors is much affordable than acquiring new ones.

- Booking channel origin makes a huge amount of difference to whether a guest cancels or not, and the data consistently comes out in favor of direct bookings over OTAs. There were higher cancellations on bookings via OTAs. So, direct bookings avoid the chances of commissions taken by different travel portals thus helping in generating more revenue

- In our dataset, there are mostly cases of No-deposit as deposit type and cancellations are also higher in that case. Hotel Managers should avoid such type of bookings during on season or they should draft new policies for No-deposit bookings to avoid No-shows.

- Cancellations are only in the case where car parking space is not required. So, bank managers should not hold bookings of customers who actually require car parking spaces

- 70% data belongs to the five countries. Therefore, if we want to increase the number of customers from other countries, we can optimize SEO from other sources based on the place, community, and language.

References

Source of data: https://www.kaggle.com/jessemostipak/hotel-booking-demand
Journel source: https://www.sciencedirect.com/science/article/pii/S2352340918315191

https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/

https://towardsdatascience.com/stop-using-smote-to-treat-class-imbalance-take-this-intuitive-approach-instead-9cb822b8dc45

https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet