1/19/2023

Ajinkya Godbole

Predicting Hotel Reservation Cancellations

Table of Contents

[Q1. Problem Statement 4](#_Toc127712244)

[Problem Statement 4](#_Toc127712245)

[Project Need 4](#_Toc127712246)

[Understanding data 4](#_Toc127712247)

[Data Description 4](#_Toc127712248)

[Q2. Exploratory Data Analysis 7](#_Toc127712249)

[Booking Status 7](#_Toc127712250)

[Family Size 8](#_Toc127712251)

[Feature – Duration of the Stay 9](#_Toc127712252)

[Feature – Meal Plan 10](#_Toc127712253)

[Feature – Required Car Parking Space 11](#_Toc127712254)

[Feature – Room Type 13](#_Toc127712255)

[Feature – Lead Time 15](#_Toc127712256)

[Feature – Date (Arrival Year, Arrival Month, and Arrival Date) 16](#_Toc127712257)

[Feature – Market Segment Type 19](#_Toc127712258)

[Feature – Repeated Guest 19](#_Toc127712259)

[Feature – No. of Special Requests 20](#_Toc127712260)

[Correlation Plot 22](#_Toc127712261)

[Q3. Data Cleaning and Pre-processing 22](#_Toc127712262)

[Treating Outliers 22](#_Toc127712263)

[Model Building 23](#_Toc127712264)

[Removing Unnecessary Columns 23](#_Toc127712265)

[Splitting into Train and Test 23](#_Toc127712266)

[Random Forest 23](#_Toc127712267)

[Naïve Bayes 26](#_Toc127712268)

[Naïve Bayes 26](#_Toc127712269)

[Summary of Model Performances 28](#_Toc127712270)

[Business Insights 28](#_Toc127712271)

[Recommendations 28](#_Toc127712272)

List of Figures

[Figure 1 Features Information 5](#_Toc127712278)

[Figure 2 Missing Values Information 6](#_Toc127712279)

[Figure 3 Five-Point Summary 6](#_Toc127712280)

[Figure 4Histogram of Booking Status 7](#_Toc127712281)

[Figure 5 Histogram for Family Size 8](#_Toc127712282)

[Figure 6 Average Stay Period per Family 9](#_Toc127712283)

[Figure 7 Percent Cancellation per Family Size 10](#_Toc127712284)

[Figure 8 Countplot of Meal Plan 11](#_Toc127712285)

[Figure 9 Countplot of Required Parking Space 12](#_Toc127712286)

[Figure 10 Countplot of Preferred Room Type 13](#_Toc127712287)

[Figure 11 Average Price per Room 14](#_Toc127712288)

[Figure 12 Histogram of Lead Time 15](#_Toc127712289)

[Figure 13 Countplot of Bookings per Year 16](#_Toc127712290)

[Figure 14 Average Lead Time Year Wise 17](#_Toc127712291)

[Figure 15 Average Room Price Month-Wise 17](#_Toc127712292)

[Figure 16 Bookings Month Wise 18](#_Toc127712293)

[Figure 17 Number of Repeated Guests 20](#_Toc127712294)

[Figure 18 Countplot of No. of Special Requests 21](#_Toc127712295)

[Figure 19 Correlation Plot 22](#_Toc127712296)

[Figure 20 Confusion Matrix Random Forest 23](#_Toc127712297)

[Figure 21 Classification Report Random Forest 24](#_Toc127712298)

[Figure 22 Feature Importance 24](#_Toc127712299)

[Figure 23 List of Features Removed 25](#_Toc127712300)

[Figure 24 Grid Search Parameters 25](#_Toc127712301)

[Figure 25 Best Parameters 25](#_Toc127712302)

[Figure 26 Grid Search Classification Report 25](#_Toc127712303)

[Figure 27 Confusion Matrix Random Forest Laptop 26](#_Toc127712304)

[Figure 28 Classification Report Random Forest Laptop 26](#_Toc127712305)

[Figure 29 Confusion Matrix Naive Bayes 27](#_Toc127712306)

[Figure 30 Classification Report naïve Bayes 27](#_Toc127712307)

[Figure 41 Performance Summary 28](#_Toc127712308)

List of Tables

[Table 1 Data Overview 4](#_Toc127712309)

[Table 2 Family size Distribution 8](#_Toc127712310)

[Table 4 Train Test Split shape 23](#_Toc127712311)

Hotel Reservations

# Q1. Problem Statement

## Problem Statement

A renowned hotel chain wants to predict if the customer who has made a reservation will cancel the reservation or not.

## Project Need

This project is necessary not only to predict the customers who will most probably cancel the reservation, but to identify the main reasons behind their cancellation. Predicting potential cancellations will control the losses, but finding out reasons for cancellations and addressing the issues will improve the efficiency of our business. And together, our hotel can improve customer experience which will in- turn increase the business.

## Understanding data

The below table explains each feature.

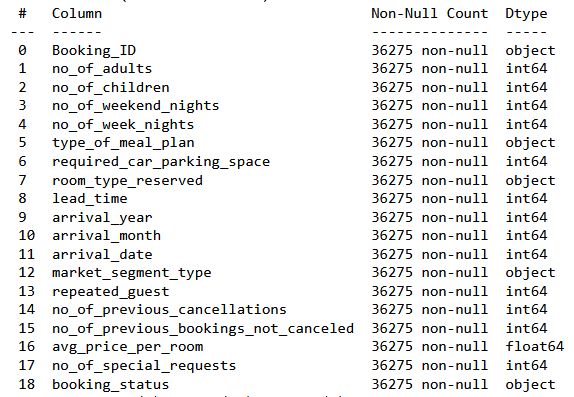
Table 1 Data Overview

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Booking\_ID | Unique ID for a transaction |
| no\_of\_adults | Number of adult members |
| no\_of\_children | Number of children |
| preferred\_device | Through which device user preferred to do login |
| no\_of\_weekend\_nights | Number of weekend nights in a booking |
| no\_of\_week\_nights | Number of week nights in a booking |
| type\_of\_meal\_plan | Type of meal selected |
| required\_car\_parking\_space | Does customer require car parking |
| room\_type\_reserved | Type of room selected |
| lead\_time | Time between the reservation and check-in date |
| arrival\_year | Year of arrival. |
| arrival\_month | Arrival month |
| arrival\_date | Arrival Date |
| repeated\_guest | Is the customer a repeated guest |
| market\_segment\_type | How customer did a reservation. |
| no\_of\_previous\_cancellations | Number of previous cancellations |
| no\_of\_previous\_bookings\_not\_canceled | Number of previous bookings not cancelled. |
| avg\_price\_per\_room | Average price per room |
| no\_of\_special\_requests | Number of special requests |
| booking\_status | If the booking is cancelled or not (Target column) |

## Data Description

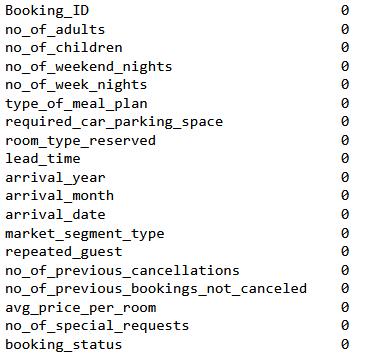
The data set has **19** features and **36275** observations. Most of the features are of float or integer datatype. The below image shows the feature information.

Figure 1 Features Information



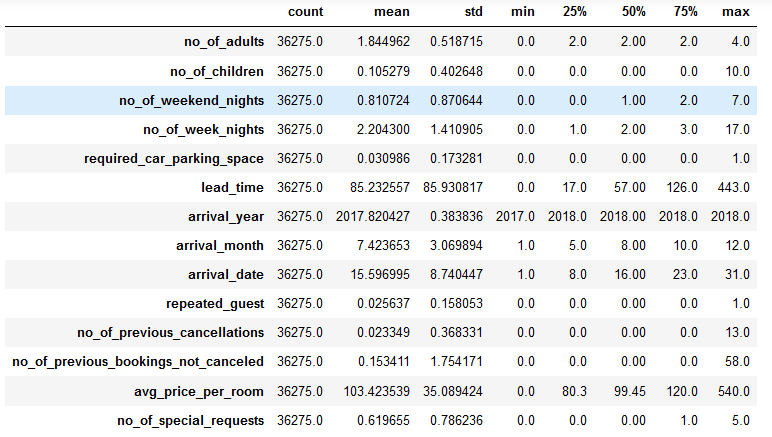
There are 5 features of object datatype and need to be converted into integer or float datatype. Also, it can be seen that there are no missing values in the dataset.

Figure 2 Missing Values Information



The below figure shows the five point summary.

Figure 3 Five-Point Summary



Two features are continuous in nature, Average price per room and Lead Time. From the five point summary, it is clear that mean is greater than median in both the features. This means both the features are right-skewed or positively skewed. Feature “lead-time” is more right skewed than “average price per room”.

We will check the distribution of each variable graphically for better understanding.

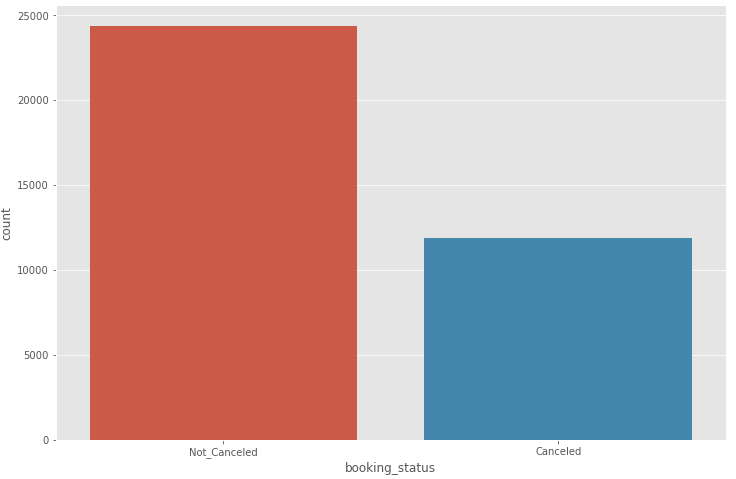
# Q2. Exploratory Data Analysis

We will do the EDA feature-wise and also, based on the EDA we will clean and treat the variables.

## Booking Status

This is the target column. We will check if the data is balanced or not.

Figure 4Histogram of Booking Status





**11885** customers cancelled their books in 2017 and 2018. Almost **33%** bookings were cancelled in the 2017 and 2018.

For machine learning model, data is not highly imbalanced.

**Encoding:** Not\_Cancelled – 0, Cancelled – 1.

## Family Size

This is a new column derived by adding No. of adults and no. of children. Below figure shows what is the most common family size of our customers.

Figure 5 Histogram for Family Size

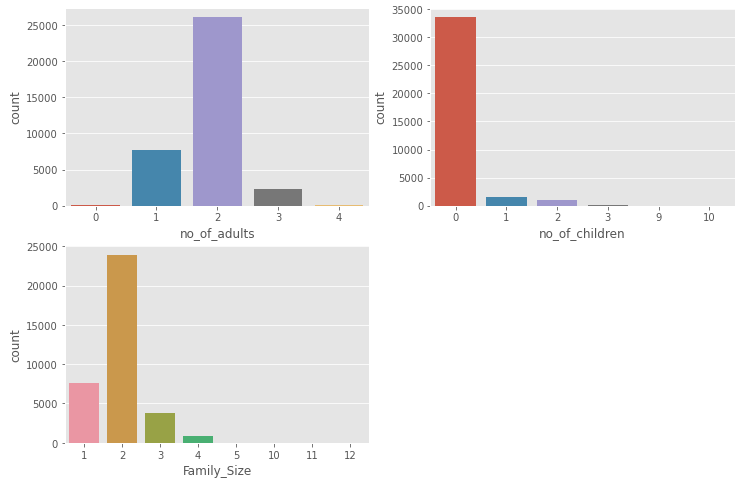


Table 2 Family size Distribution

|  |  |
| --- | --- |
| **Family Size** | **Family Count** |
| 1 | 7552 |
| 2 | 23942 |
| 3 | 3851 |
| 4 | 912 |
| More than 4 | 17 |

Most of our customers come in pairs. We have very less customers with family size greater than 4 member, less that 0.05%. So at the time of model building, we will replace these 17 records to family size with 4 members.

We must check age of the customers. So that we can check if most our customers are honeymoon couples. If this turns out to be true, hotel must come up with new offerings for honeymoon couples.

## Feature – Duration of the Stay

This is also a new column derived by adding No. weekend nights and no. of weeknights. Below figure shows what is the average stay period per family and percentage of cancellation.

Figure 6 Average Stay Period per Family

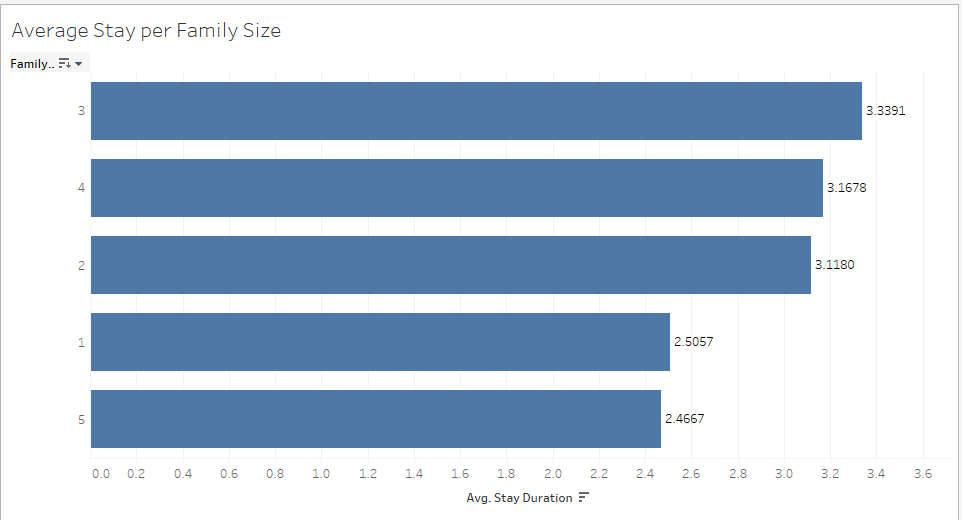
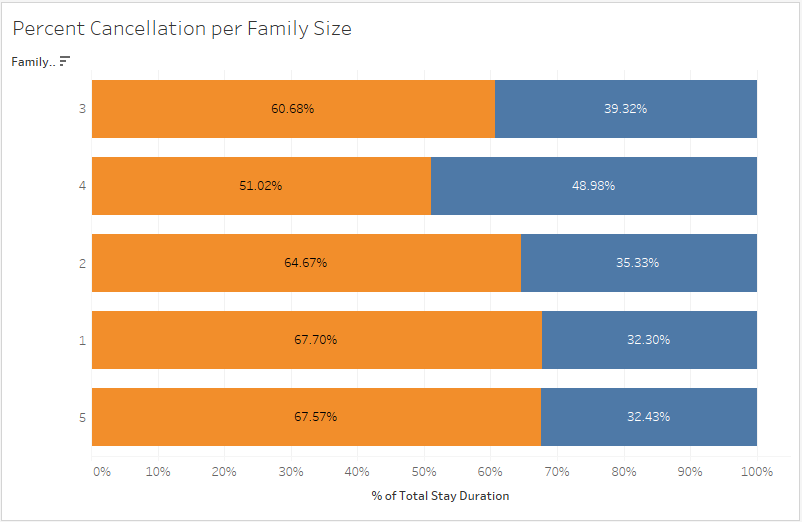


Figure 7 Percent Cancellation per Family Size

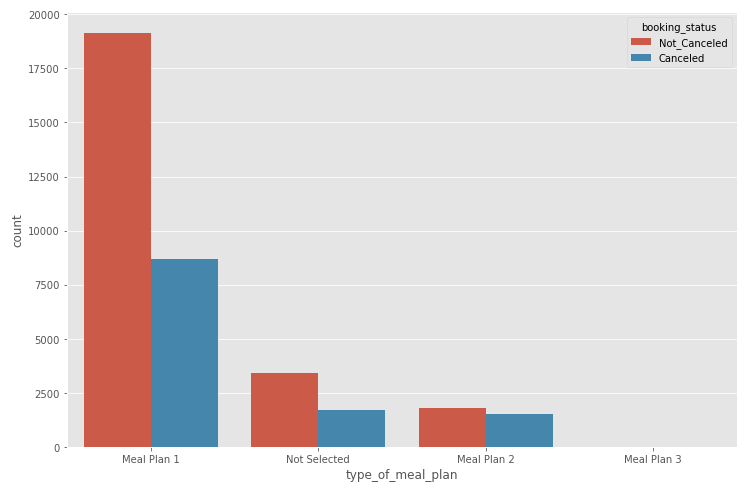


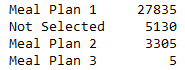
From the above two figures, it can be concluded that average stay duration for different family sizes is more or less same, 3 days. And, the cancellation percentage is highest among families with 4 members, almost 50%.

## Feature – Meal Plan

This is also a categorical variable. Below figure shows the most popular meal plan is.

Figure 8 Countplot of Meal Plan





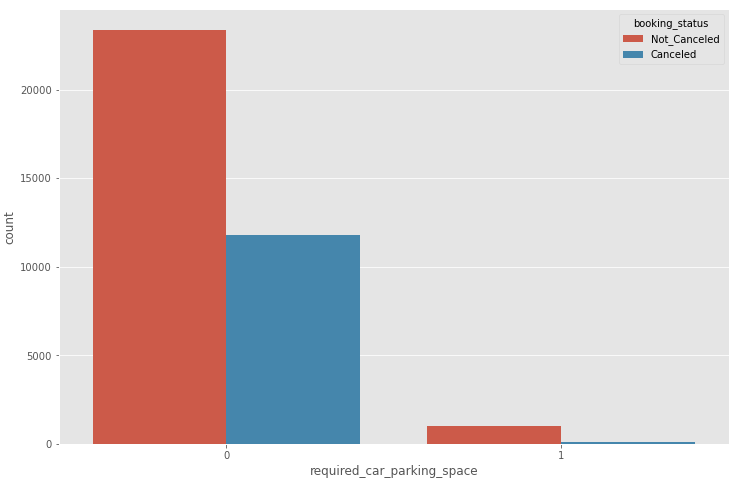
Almost 76% of the customers selected Meal Plan 1. Only 5 customers selected Meal plan 3.

**ENCODING** – Meal Plan 1 – 1, Meal Plan 2 – 2, Meal Plan 3 – 3, Not Selected – -1.

## Feature – Required Car Parking Space

Required car parking space is a categorical variable with values 0 and 1. 0 indicates No and 1 indicates Yes.

Figure 9 Countplot of Required Parking Space

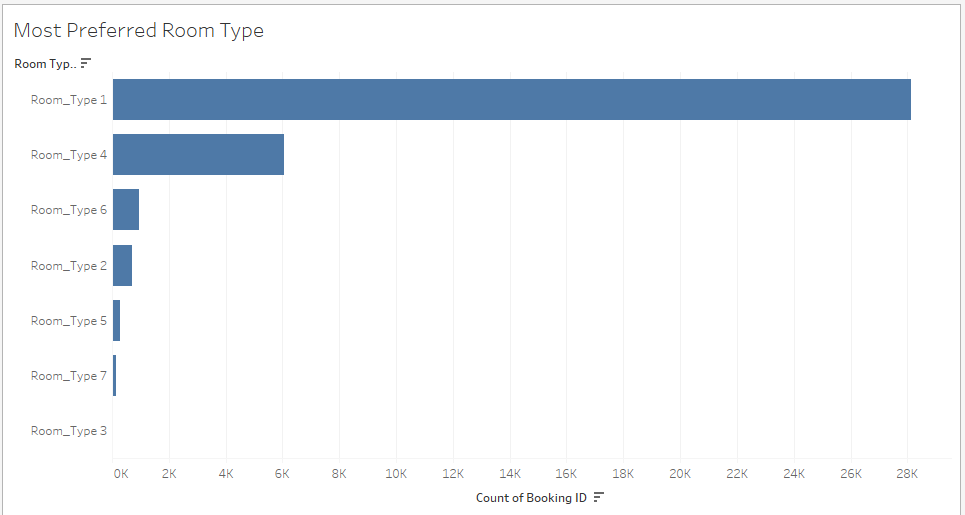


Most of the customers do not require car parking space. This is may be because customers are coming from far away. If that is the case, hotel can provide cab facility for pick-up and drop and sightseeing.

## Feature – Room Type

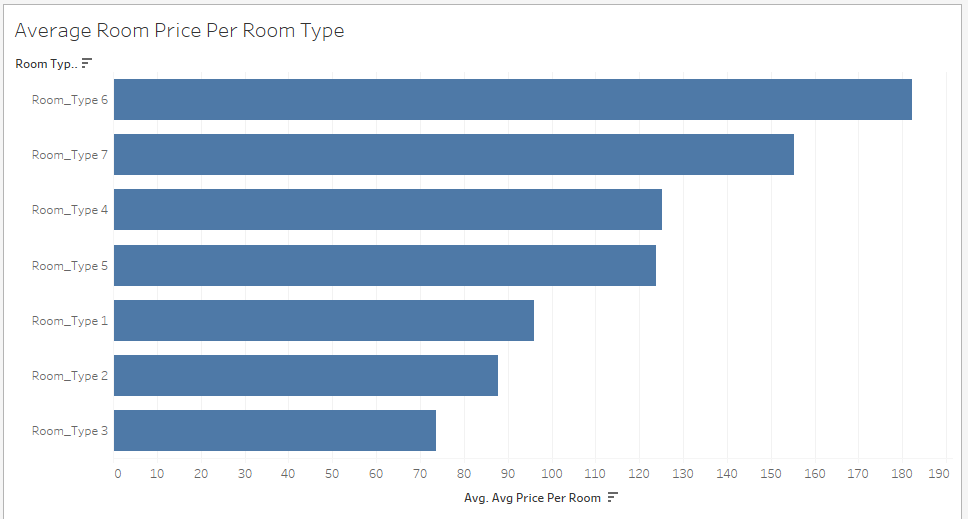
Room Type is a categorical variable. There are seven room types. Let’s see the most preferred room type.

Figure 10 Countplot of Preferred Room Type



Demand for room type 1 is very high. There is almost no demand for room type 3. Let’s check the average price per room.

Figure 11 Average Price per Room



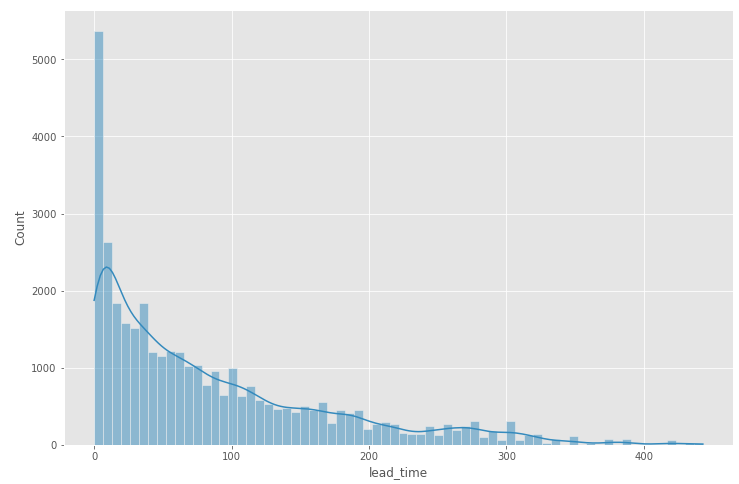
Average price of room type 6 is the highest. While average price for room type 3 is the lowest. Customers are opting for middle ranged room types.

**ENCODING** – Room\_Type 1 – 1, Room\_Type 2 – 2, Room\_Type 3– 3, Room\_Type 4 – 4, Room\_Type 5 – 5, Room\_Type 6 – 6, Room\_Type 7 –7 .

## Feature – Lead Time

This is a continuous variable and there is a direct relationship between cancellation and higher lead times.

Figure 12 Histogram of Lead Time



As we have seen in the five point summary, mean is greater that median. Now, it is evident that the data is right skewed. Also, there are many outliers in this feature that we need to fix.



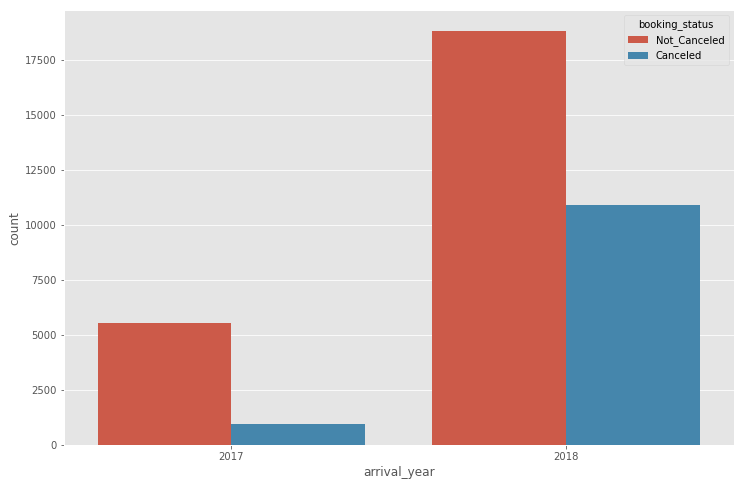
One more thing to notice, average lead time for the cancelled bookings is significantly greater than bookings that are not cancelled. So, there is some relationship between lead time and cancellations.

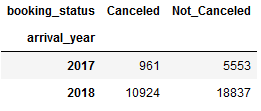
NOTE: We have removed outliers before building machine learning model.

## Feature – Date (Arrival Year, Arrival Month, and Arrival Date)

This is a date time variable. Let’s visualize various features against date variable.

Figure 13 Countplot of Bookings per Year

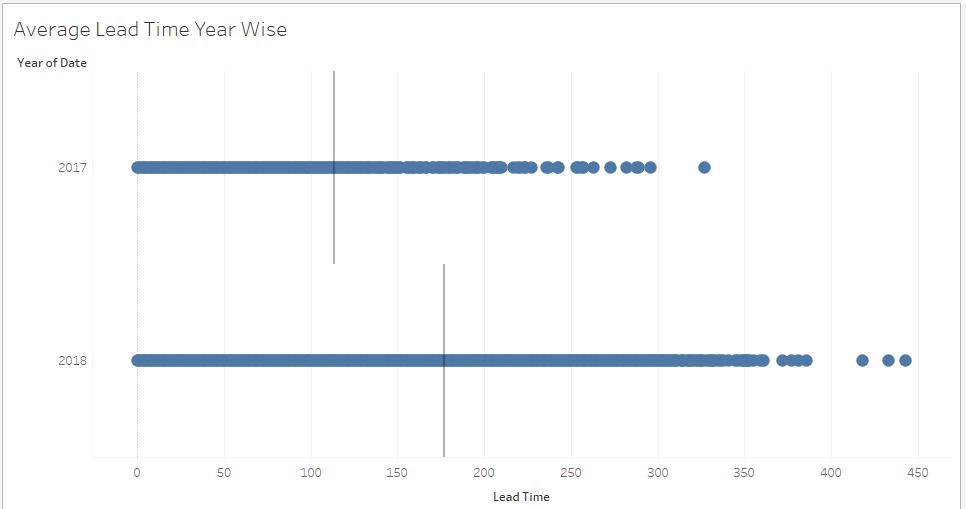




We can see that number of bookings grew exponentially, but number of cancellations also grew exponentially. In 2018, almost **58%** of the bookings were cancelled. While in 2017, less than **1%** of the bookings were cancelled.

As we know, lead time has some relation with cancellations. Let’s check the average lead time in 2017 and 2018.

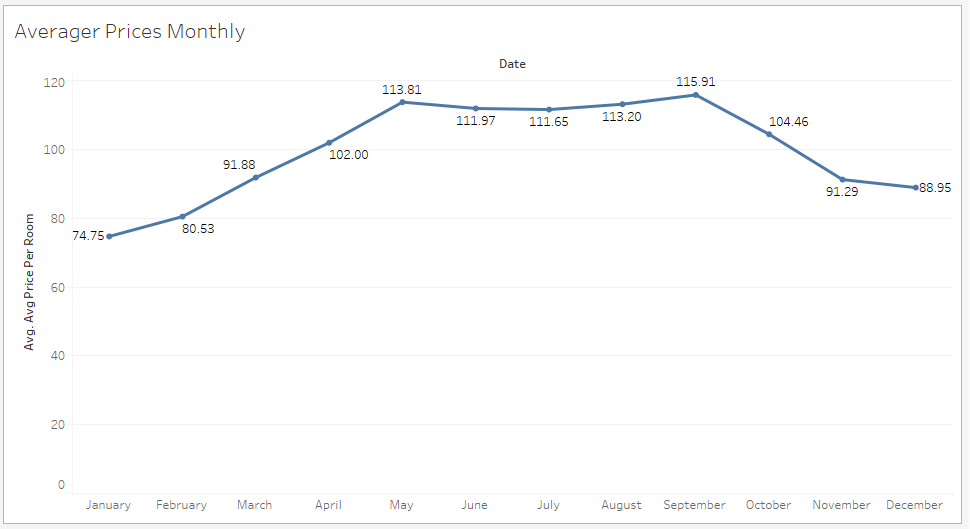
Figure 14 Average Lead Time Year Wise



People booked way in advance in 2018. And we know, higher lead time leads to higher cancellation.

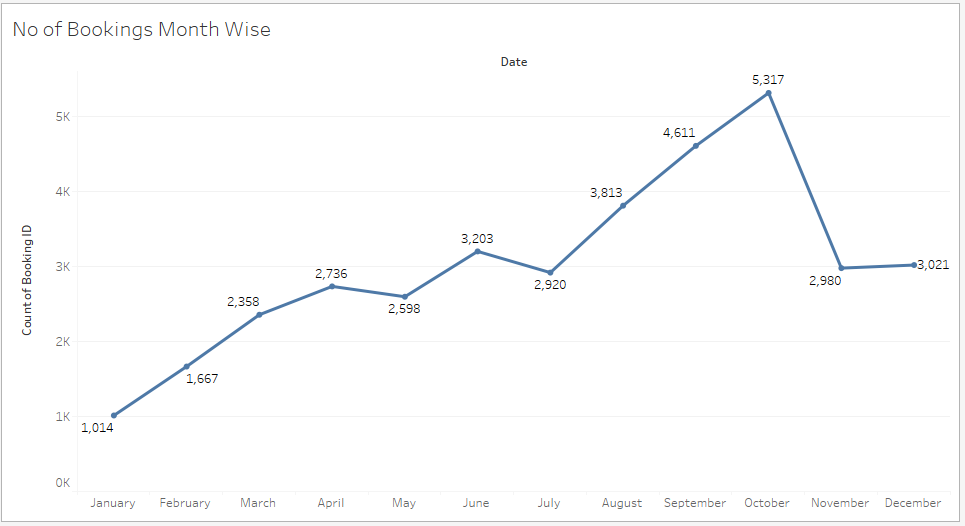
Let’s check average room price per month.

Figure 15 Average Room Price Month-Wise



Room prices are highest in October. It must be related to demand. Let’s check the month wise demand.

Figure 16 Bookings Month Wise

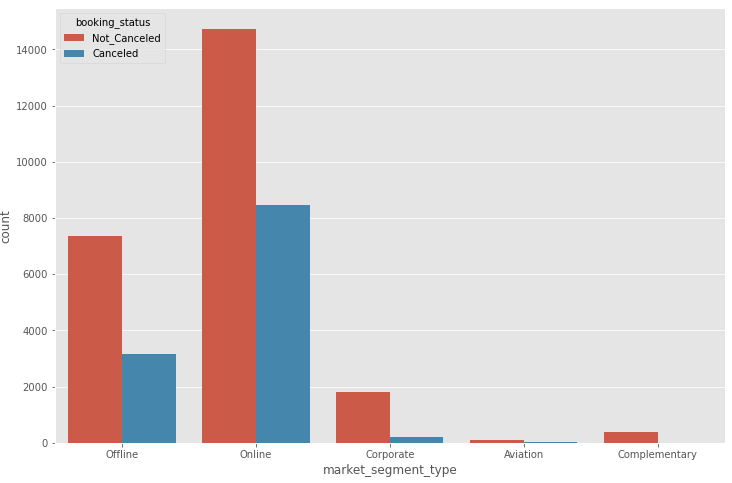


From the figure, we can see that at the start of the year hotel bookings are at the lowest. As the year progresses, demand goes on increasing. The demand is highest in October month.

**Hotel can offer discounts on bookings in November, December, and January months.**

## Feature – Market Segment Type

This is a categorical variable. It has five categories namely Aviation, Complementary, Corporate, Offline, and Online.



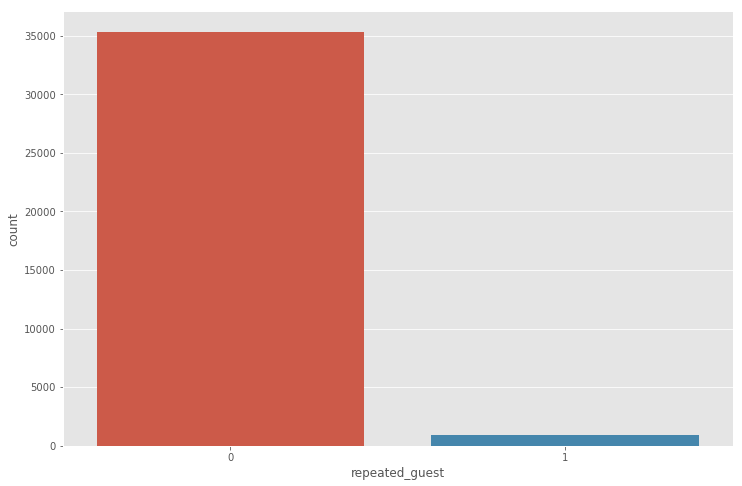
It can be seen that online bookings have the most number of cancellations. More than 50% of the bookings were cancelled.

One more thing to note, more than 20,000 customers used “Online” method for booking in 2018 than in 2017. There is some relation between online bookings and cancellations.

## Feature – Repeated Guest

This is also a categorical variable. Below image shows number of repeated customers.

Figure 17 Number of Repeated Guests



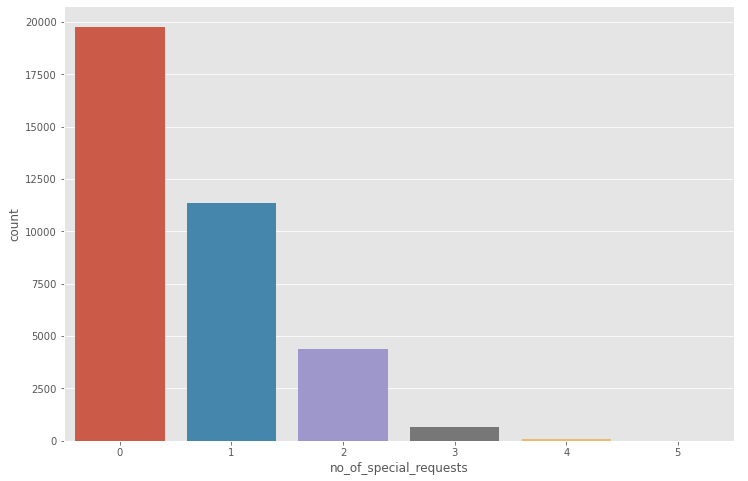
Number of repeat customer is less than 1%. Customer retention is a very big part of modern day business.

Loyalty programs must be launched to retain customers.

## Feature – No. of Special Requests

Let’s see how many special requests are received from the guests.

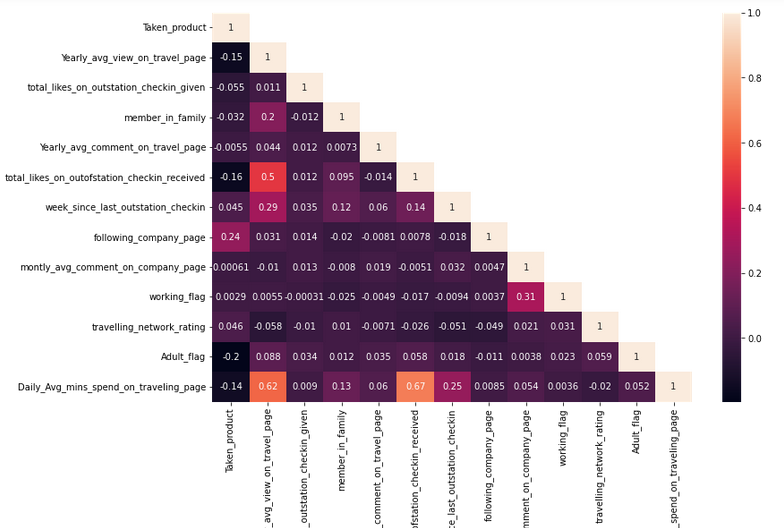
Figure 18 Countplot of No. of Special Requests



Most of the customers did not make any special request.

## Correlation Plot

Figure 19 Correlation Plot



There are very few features that have strong positive and negative relationships.

The correlation between dependent and independent variable is very less. Also, there is no much multi-collinearity in the dataset.

# Q3. Data Cleaning and Pre-processing

## Treating Outliers

We will use an empirical rule for detecting outliers. Whenever a data point is outside the range of Q1 – (1.5\*IQR) and Q3 + (1.5\*IQR), it will be considered as an outlier.

Where,

IQR = Inter Quartile Range (The range between 25th percentile and 75th percentile)

Q1 = 25th percentile value

Q3 = 75th percentile value

Below image shows the boxplots of variables after treating outliersScaling

Scaling is mandatory for building distance based algorithms or algorithm that uses gradient descent to find slope and intercept values.

# Model Building

## Removing Unnecessary Columns

We will remove “Booking ID" column from the dataset. As this column will not add any value while building models.

## Splitting into Train and Test

We have split the data into train and test using test size as 30%. Below table shows the shape and size of train and test data.

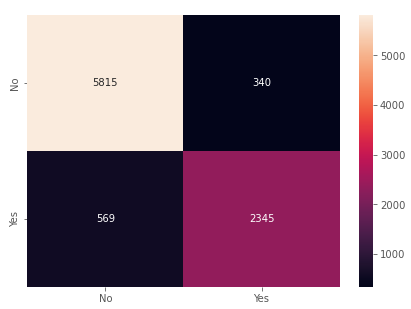
Table 4 Train Test Split shape

|  |  |  |
| --- | --- | --- |
| **Type** | **Rows** | **Column** |
| X\_train | 27206 | 19 |
| X\_test | 9069 | 19 |
| y\_train | 27206 | - |
| y\_test | 9069 | - |

## Random Forest

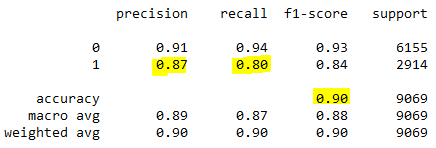
Random Forest is a decision tree based classifier. Here we are using confusion matrix and classification report as an evaluation metric.

Figure 20 Confusion Matrix Random Forest



**5815** are True Negatives and **2345** are True positives Total number of wrongly classified data points are 909.

Figure 21 Classification Report Random Forest



Let’s fine tune the model. We will use ExtraTreeClassifier for Feature selection. Here is the list of feature importance.

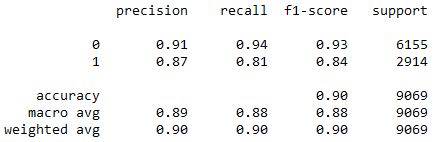
Figure 22 Feature Importance



We will remove features with the lowest importance. We will remove one by one and check if there is any improvement in accuracy.

Figure 23 List of Features Removed

|  |
| --- |
| no\_of\_previous\_cancellations |
| no\_of\_previous\_bookings\_not\_canceled |
| repeated\_guest |
| required\_car\_parking\_space |
| no\_of\_children |
| no\_of\_adults |



Only Recall got improved by 1% by removing couple features. We will use GridSearch to try to improve efficiency.

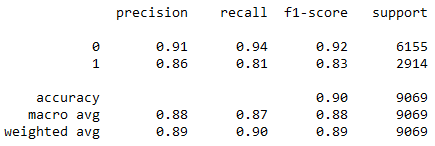
Figure 24 Grid Search Parameters

|  |  |
| --- | --- |
| Parameters | Values |
| Criterion | "gini", "entropy", "log\_loss" |
| max\_depth | 10,20,30,40,50 |
| max\_features | 6,8,12 |

Figure 25 Best Parameters

|  |  |
| --- | --- |
| Parameters | Values |
| Criterion | "log\_loss" |
| max\_depth | 20 |
| max\_features | 6 |

Figure 26 Grid Search Classification Report



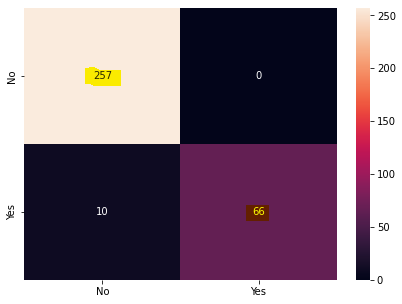
Here, precision for 1 got reduced by 1%.

We will try different models as well.

### Naïve Bayes

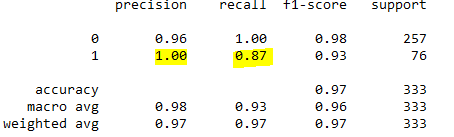
This model is based on Bayes theorem.

Figure 27 Confusion Matrix Random Forest Laptop



Here True positives are **66** and True Negatives are **257**.

Figure 28 Classification Report Random Forest Laptop

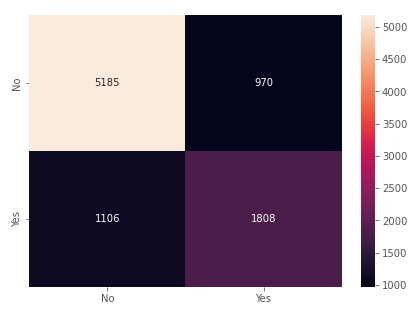


In this report both precision and recall is good.

## Naïve Bayes

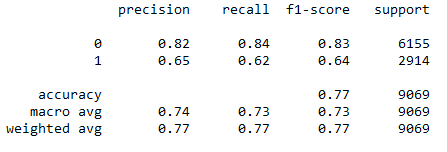
Naïve Bayes works on Bayes theorem.

Figure 29 Confusion Matrix Naive Bayes



Here True positives are **1808** and True Negatives are **5185**.

Figure 30 Classification Report naïve Bayes



Precision and Recall is very bad. Potential reasons for bad precision and recall are:

1. Data is highly imbalanced.
2. All numerical and continuous variables are not properly normally distributed.

# Summary of Model Performances

Figure 41 Performance Summary

|  |  |  |
| --- | --- | --- |
| Model | Precision | Recall |
| Random Forest | 87% | 80% |
| Random Forest Fine Tuned | 87% | 81% |
| Naïve Bayes | 65% | 62% |

Random Forest has performed better than all the models. The main reasons are:

1. Most of the variables are categorical in nature.
2. Random Forest doesn’t have any assumption of normality for continuous variables.
3. Random Forest usually performs better for binary classification problems.

# Business Insights

* Most of the customers come in pairs. We must check if they are honeymoon couples.
* Most of the customers doesn’t need car parking space.
* Since customers doesn’t require car parking that means they are coming from far away.
* Almost 60% of the bookings in 2018 were cancelled as compared to less than 1% in 2017.
* Lead time has direct relation with cancellation. Higher the lead time, cancellation chances are high.
* Room type 1 and Meal type 1 is the most preferred choice.
* People who booked the room online have more chances of cancellation.

# Recommendations

* If most of the customers are honeymoon couple, hotel must come up with honeymoon specials package.
* Since people don’t come by their own vehicle, hotel canprovide free pick and drop service to hotel. Also, hotel can provide chargeable cab service for sightseeing.
* Usually October is the rush period and three months after that is a slack period. Hotel can offer discounts on bookings in November, December, and January,
* If hotel is planning on providing discounts in the month of November, December, and January, Offers must be made3 public after August month because higher lead time leads to cancellation.
* As we know, October is a rush period. Hotel can provide discounts on not so popular rooms to attract more customers.