School of Information Studies
Syracuse University

HOTEL BOOKING CANCELATION ANALYSIS

Jugen Gawande
Yaping Wang
Shivani Kalewar
Vedant Rawat
Srushti Samant

GROUP 3

FALL 2021

Table of Contents

Table of Figures

Introduction

As a manager of a booming hospitality enterprise, it still would give you sleepless nights when your expected guest cancels. A cancelation cripples revenue cause it not only means the canceling guest will not be paying but you have deterred other potential guest you could have the opportunity to serve, but you did not have the space to. It fills you with doubt, was something lacking or are there patterns to the guests who tend to cancel. It is pertinent if there are such patterns that exists, to identify them and be on top of any problems to leave no stone unturned to maximize profitability.

With this project we will explore this idea, with a dataset containing hotel booking details of hundreds of guests from around the world. With some key attributes complied in this dataset with respect to bookings, we try to analyze if they have any influence on cancelation. If there are patterns that help us make better decisions going forward it would always be helpful.

Project Goal

Given a dataset of over a hundred bookings, identify the patterns if any which influences if a guest cancels their reservation. Understanding the factors that drives cancelations and if we have the ability to predict if a booking will eventually be canceled.

Hello dataset

We begin our analysis with understanding the dataset we will be using. Using *glimpse* and *str* functions we can get the required description of the dataset.

The dataset contains 40060 rows which is booking records. It has 20 attributes and an outcome attribute which is a binary encoded which represent if a booking record was eventually canceled.
The attributes are as listed below with their type and short description. We analyze these attributes in further sections of this report in detail.

IsCanceled < categorical >

It is a binary variable which is the outcome variable for our dataset. It indicates if a booking was cancelled or not with '1' denoting cancelled booking. The dataset contains 27.7% canceled booking data.

LeadTime <numeric>

It gives the number of days between the date of booking and arrival date. The range is from 0 to 700. But these higher values are mostly outliers as assessed in the further sections of this report.

StaysInWeekNights <numeric>

Number of weeknights booked for in the hotel by the guest. (Monday-Friday)

StaysInWeekendNights <numeric>

Number of weekend nights booked for in the hotel by the guest (Saturday-Sunday)

Adults <numeric>

The number of adults as part of the booking

Children <numeric>

The number of children in the booking

Babies <numeric>

The number of babies in the booking

Meal < categorical >

The type of meal plan chosen. There are 5 unique categories, which are – BB, HB, FB, SC, Undefined. From the dataset documentation we know that SC and Undefined are the same, hence we replace all Undefined with SC.

Country < categorical >

It contains ISO-3166 Alpha-3 abbreviations of world countries. There are 126 countries in this attribute.

MarketSegment < categorical >

It is a categorical variable that represents where the booking is coming from. There are 6 categories in this attribute which are as follows – Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups.

IsRepeatedGuest < categorical >

This is a Boolean attribute that denotes the guest has previously booked with the hotel. Only 1667 bookings the dataset are from repeated guests.

PreviousCancellations <numeric>

Records how many bookings has the guest previously canceled

PreviousBookingsNotCanceled <numeric>

Records how many booking the guest previously booked and checked in. There are 2010 bookings with guests who have not previously canceled.

ReservedRoomType < categorical >

There are 10 room types that guests can reserve as a part of the booking.

AssignedRoomType < categorical >

There are 11 room types that guests are assigned to. It usually is the ones the guests have reserved but in some cases there can be a change in the room assigned than the one reserved. It can because the guest changed his/her mind or arrival or some operational reason from the hotel end.

BookingChanges <numeric>

It is the count of number of times the booking details were changed through the course of booking and arrival.

DepositType < categorical >

It is a categorical variable that records if a deposit was paid to block the booking and if it was paid was it refundable or non-refundable.

CustomerType < categorical >

There are four categories in this variable that indicate the type of customer it is. The four categories are Transient, Contract, Transient-Party and Group.

RequiredCarParkingSpaces <numeric>

It indicates of the guest has made a request to reserve parking space. It is a numeric attribute that stores the number of parking spaces that are required.

TotalOfSpecialRequests <numeric>

It is a numeric attributes that contains the count requests made by guests that are non-standard and are something they might need to make their stay worthwhile.

Derived Columns

TOTAL GUESTS

After looking at the dataset we made a decision to have two derived columns to track the number of guests. It made more sense to have a column that tracks the total number of guest in a booking. Therefore we do the following adults + children + babies to get a new attribute called '*totalMembers'*.

IS FAMILY?

At the same time, we did not want to lose any insight about the influence of having children or infants as our patrons, hence we create a new Boolean attribute called '*isFamily'* based on the assumption that if there are children or infants accompanied by more than one adult then it is a booking for a family.

IS ROOM DIFFERENT?

There were instances in the dataset where the assigned room type was different than the one that were reserved. We create a new boolean attribute, which tracks if the room type allocated is different or the same.

TOTAL STAY DAYS

We combined the two columns of 'StayInWeekNights' and 'StayInWeekendNights' by sum, since we presume that it wil provide us with more insight than having the attributes separate.

IS WEEKEND

On combining the two columns in the above step, we lose the ability to analyze if a stay included weekend, as generally there is a high influence of when the booking is for, with weekends being the most in demand. Therefore, to address this we use a Boolean attribute which is one if there is more than one stay in weekend night. This lets us analyze if a booking included any weekend nights.

CANCEL RATE

Guests booking patterns influence cancelation majorly. The two attributes of Previously Canceled and Previous Booking not canceled provided not much insights and seemed redundant. We therefore, store the cancelation rate of the guest in a new attribute called cancel rate which is previously canceled bookings divided by total cancelation by the guest. In case the guest is new the cancel rate is infinity or stored as nan in the dataset.

[DATA CLEANING]

Analysis of Unallowed Bookings

It is important that all the bookings in the dataset are valid. We first checked if there were any anomalies in the dataset. We found several rows where the total number of members and total number of days of stay was zero. This is practically not possible and hence we dropped these rows from our dataset.

There were no other erroneous booking records found in the dataset. New clean dataset now had 39660 booking rows

Analysis of Outliers

After looking at lead time there were more than 200 entries over one year. Though an insignificant number we cannot drop these values. We instead floor these outliers to our maximum derieved from our boxplots.

Analyzing other numeric attributes like total members, we found that the number of bookings with more than 6 people was less than 1% of the dataset. At this rate, it is insignificant to even consider those bookings and therefore we decided to drop them.

Similarly, the number of total stay days above 20 is negligible (at 169) therefore, we also drop those records where the totalStayDays is more than 20.

Analyzing the variables, Required Parking Space and Total number of special requests even though they are numeric variables, since they are set for a very few bookings, we ran unique on them to get required description. It turned out there were very few unique values in a numeric column and therefore, the attribute can be dropped during modeling. Since a majority of the bookings are not canceled but they form such a small subset, it can introduce a bias in the model.

This allows us to have a well distributed dataset on which we can perform further analysis.

Attribute Analysis

There are two types of attributes in the dataset – numeric and categorical. The following is the list of attributes in each category, after we have dropped the columns we have considered in-significant combined with our engineered new attributes.

Numeric Attributes

- Lead Time
- BookingChanges
- RequiredCarParkingSpaces
- TotalOfSpecialRequests
- *totalMembers*
- *totalStayDays*
- *cancelRate*

Categorical Attributes

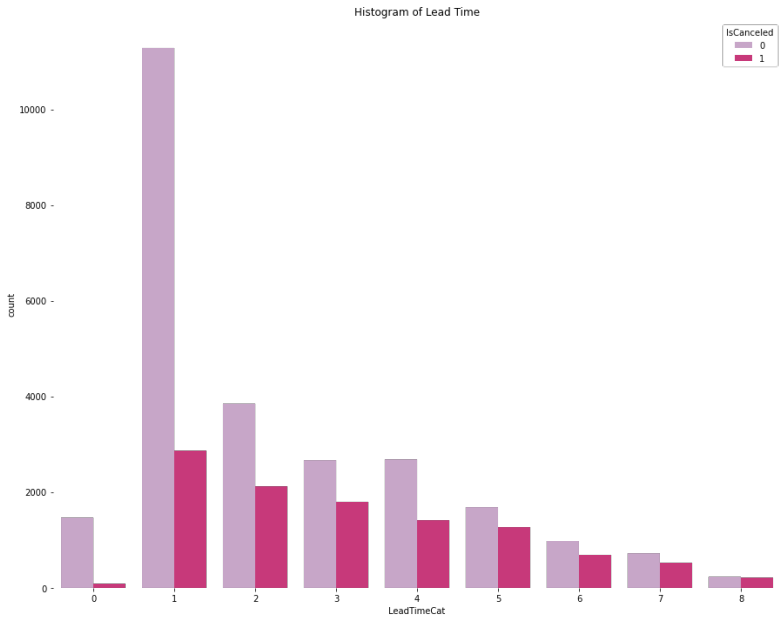
- Meal
- Country
- IsRepeatedGuest
- MarketSegment
- ReservedRoomType
- AssignedRoomType
- DepositType
- CustomerType
- *isWeekend*
- *isFamily*

In the following section, we have reported our analysis of each of these attributes. Analysis of each attribute is divided into three sub parts, where first we perform exploratory data analysis on the variable; second we perform analysis of outliers and re-visualize after the outliers have been dealt with and then we draw inferences based of how an attribute is contributing to the cancelations.

- - NUMERIC ATTRIBUTES

Lead Time

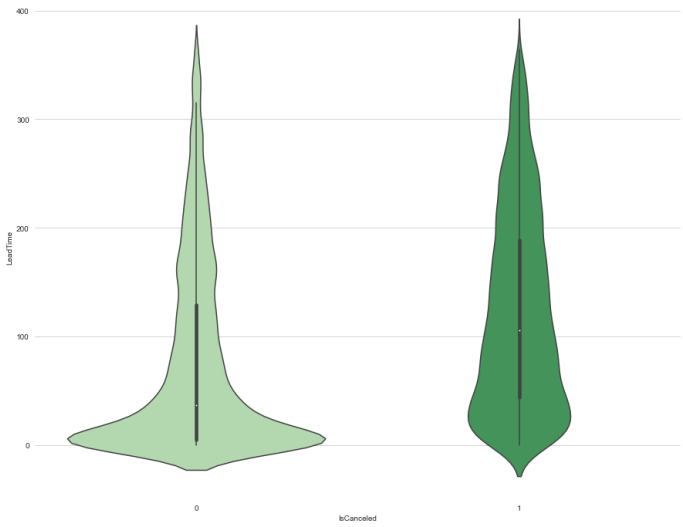
The lead time was divided into bins to improve the visualization. Here 0 represents the booking



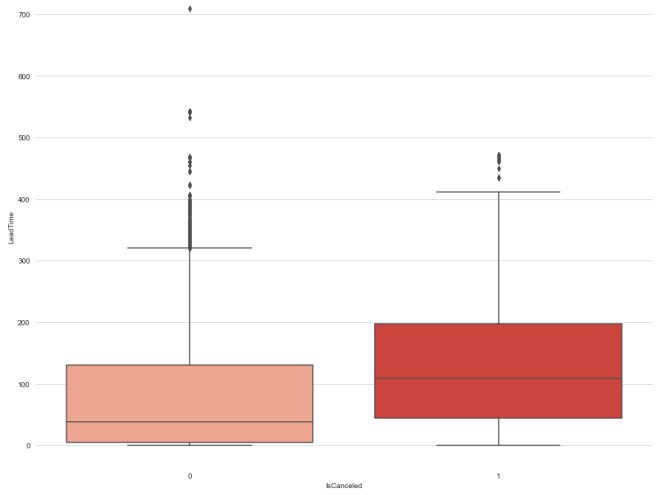
Histogram for Binned Lead Time

has a lead time of zero i.e. it was probably a walk-in booking. And we can see the cancelation is very minimal in the case of walk in. All bins from 1-7 are of size 50, so lead time 1-50, 51-100 and so on. Bin 8 is all the bookings with lead time of more than a year ago than date of arrival.

From the above visualization it is clear that a huge chunk of bookings come anywhere from 1-50 days before the date of arrival.



Violin Plot for Lead Time

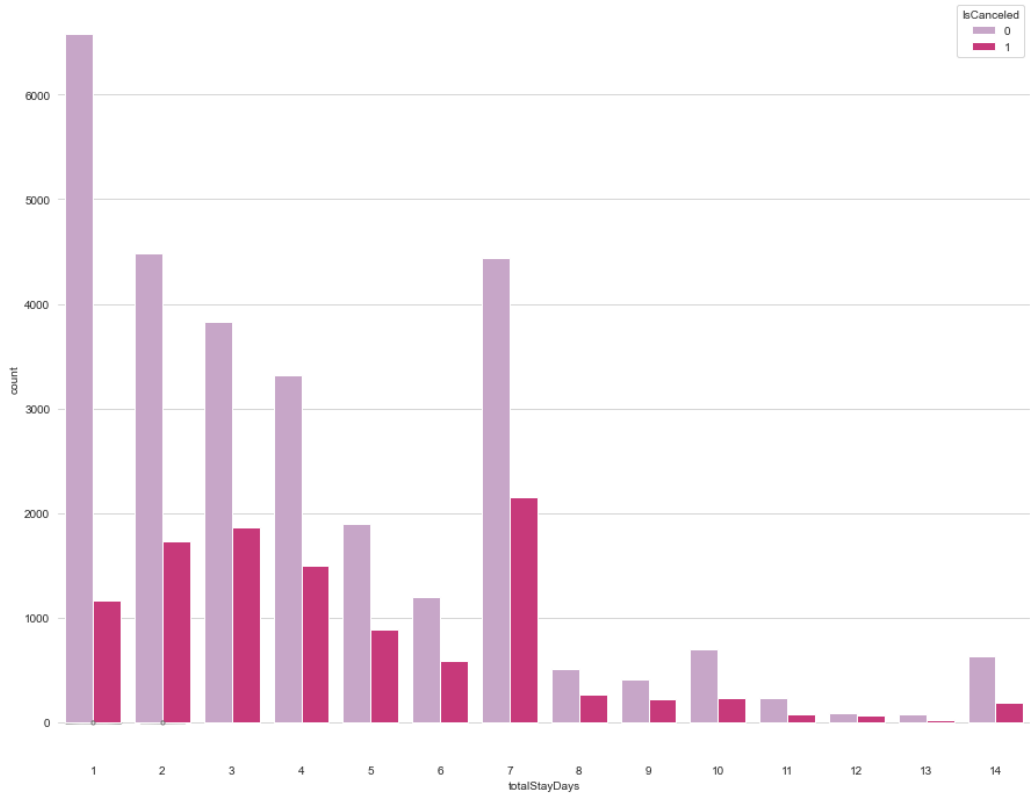


Box Plot for Lead Time

A violin plot helps us look at the information provided by a box plot and histogram together. The violin plot further confirms our findings from histogram. The IQR for not canceled bookings is from 5 days to 130 days with a mean of 79 days (~2.5 months before). Where as, for the canceled bookings it's from 44 days to 198 days with a mean of 128 days (~4 months before).

We can therefore conclude that lead time has a perceivable influence on the cancelation. The bookings made well in advance, from 3-6 months before have a higher chance of cancelation. But it's not a linear relationship. On the other hand, we observe from our analysis of box plot for outliers there are bookings made more than one year in advance, where the chance of cancelation is more than 50% but it only skewed by a group booking of 52 bookings from the orign country of GBR.

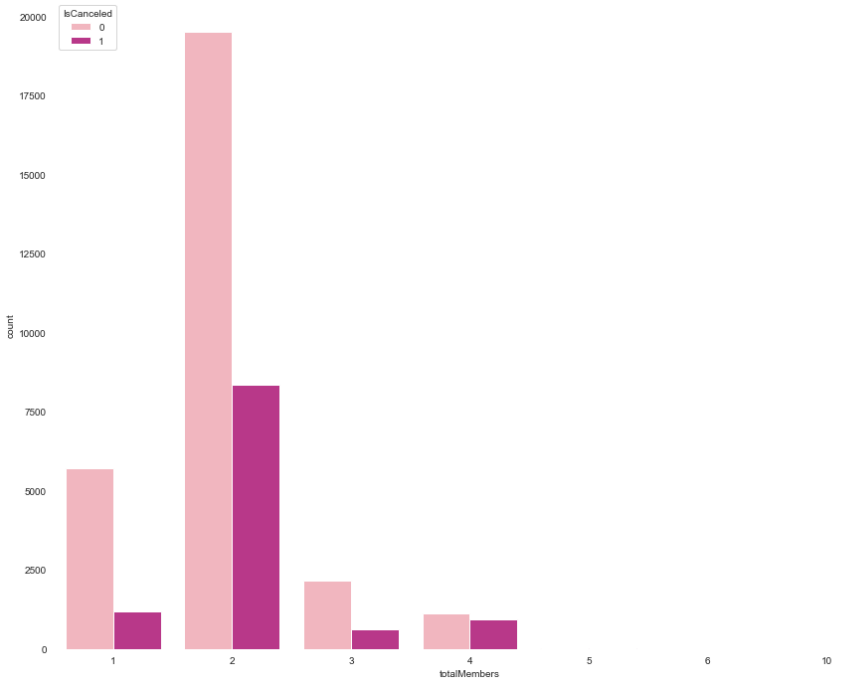
Total Days of Stay



Histogram of total stay days

Looking at this attribute we found the average length of total stay is 3 for single guest and 4 for all guest from couple to families up to size 5. Groups with 6-10 members had an average of 7 total stay days. We could see a higher concentration of single guest staying for one day, which when furthered analyzed revealed came from corporate market segment. Which is in line with our other analysis that checked for market segment against the number of people.

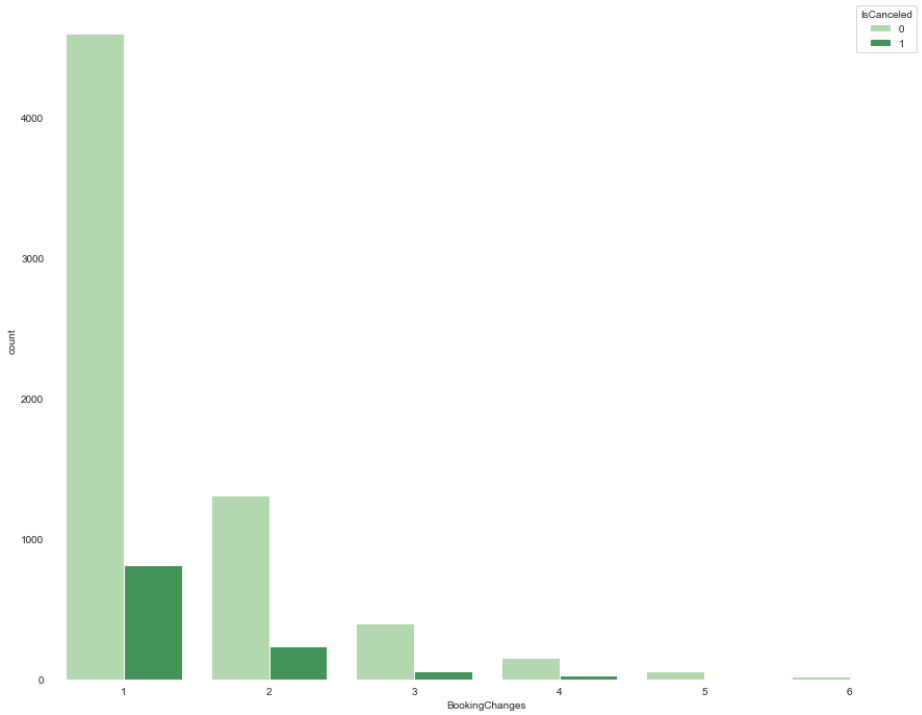
Total Guests



Histogram of total guests in every booking

Looking at the dataset we could clearly infer those bookings with 2 guest is predominant segment of the hotel. Followed by single guests at 17% of the guests. The hospitality can be seen better suited for couple than families. Also, we can see the cancelation rate from the graph above that it is very less for single guest as compared to couples which have the most cancellations.

Booking Changes



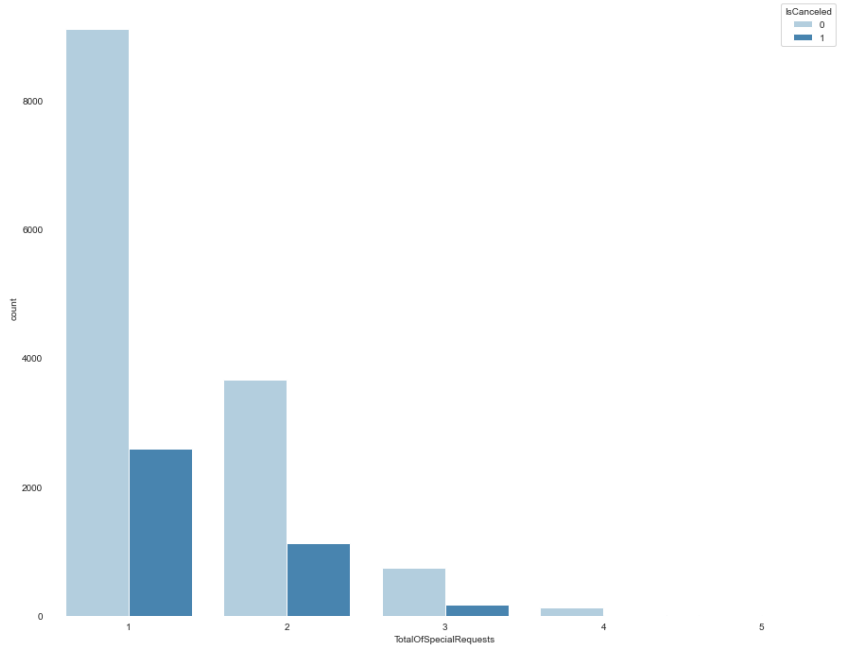
Histogram for Booking Changes

From the above graph, we can see as the booking changes increases the cancelation rate decreases considerably. We can infer that as booking changes is increasing there is a sort of commitment from the guest that they will follow through on their booking.

Required Parking Space

After looking at the data we could see that for those who reserved a parking space which was usually one, was by guests who made a reservation close to check-in date. Therefore, there was no cancelation found in bookings. There were 5475 bookings that required parking spaces which is 13% of the dataset.

Total Number of Special Request



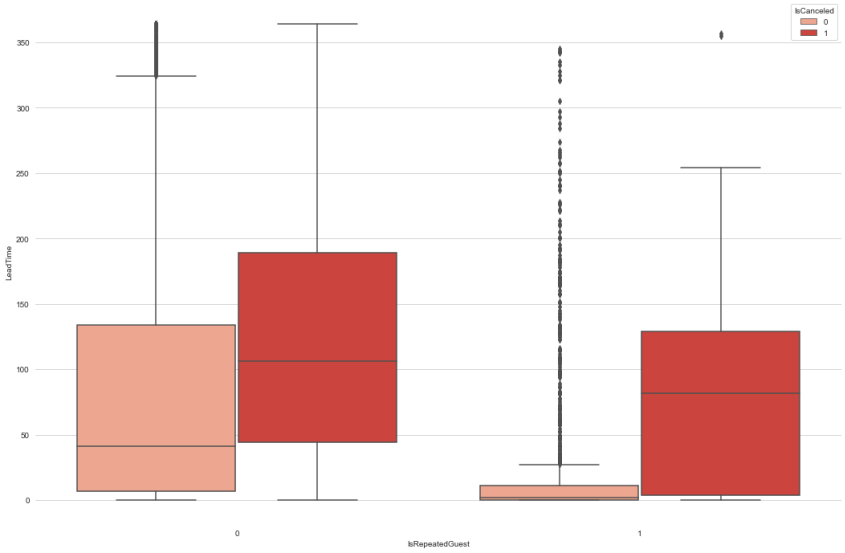
Histogram for Total Special Request

Similar to parking spaces we could see that having special request resembled to a certain degree a commitment to bookings by the guests. As the number of request went up we could see a drop in the cancelations.

--CATERGORICAL ATRRIBUTES

IsRepeatedGuest

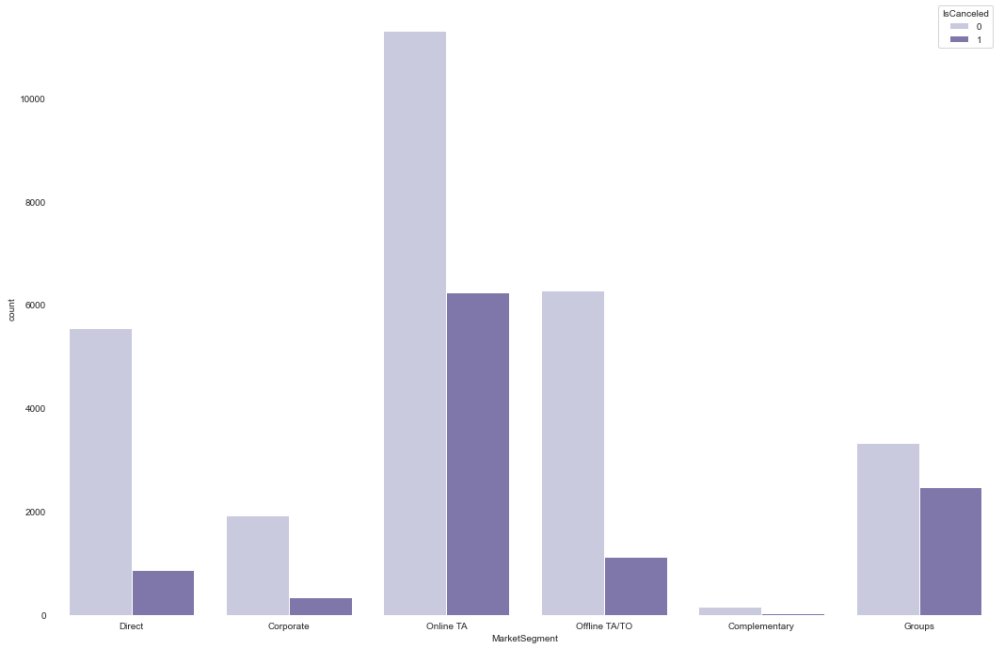
We further looked at how having repeated guests affected our cancelations. The following graph we can see that the repeated guests tend to book close to check in time (on an average within one month) who tend to not cancel. Amongst repeated guest it is a very small percentage that cancel, which we can hold insignificant when we compare it against the whole dataset.



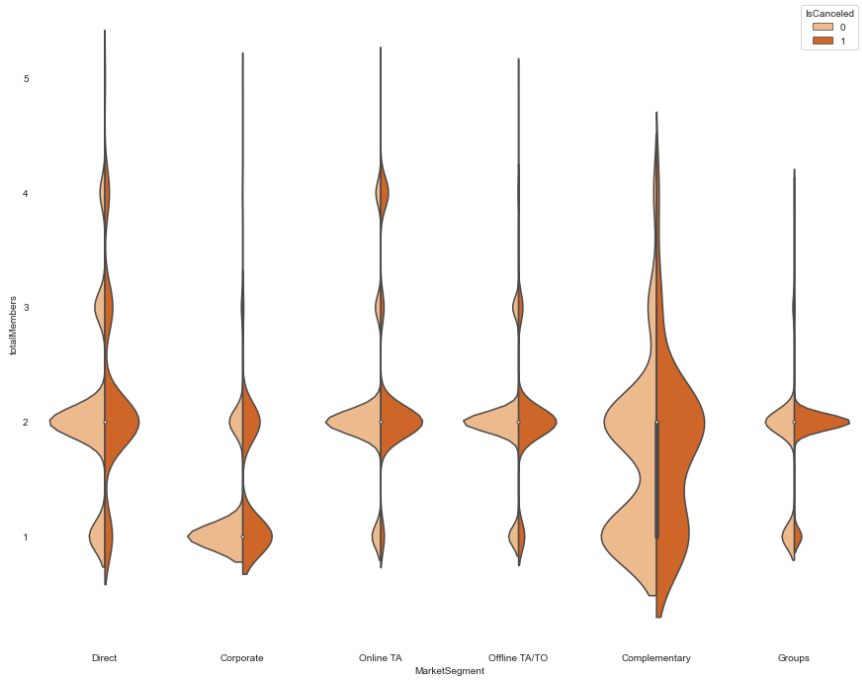
Box Plot for Repeated Guests

Market Segment

We could see that major bookings come from online Travel Agents followed by direct bookings with the hotel. From the analysis we could very clearly infer that ‘Online TA’ is the prime share of bookings but at the same time it also has the highest cancelation rate. Compared, to direct which has a very low cancelation rate, followed by corporate. For corporate, though it has a very low cancelation rate we, the total number of bookings coming from it is very low compared to others.



Histogram for market segment



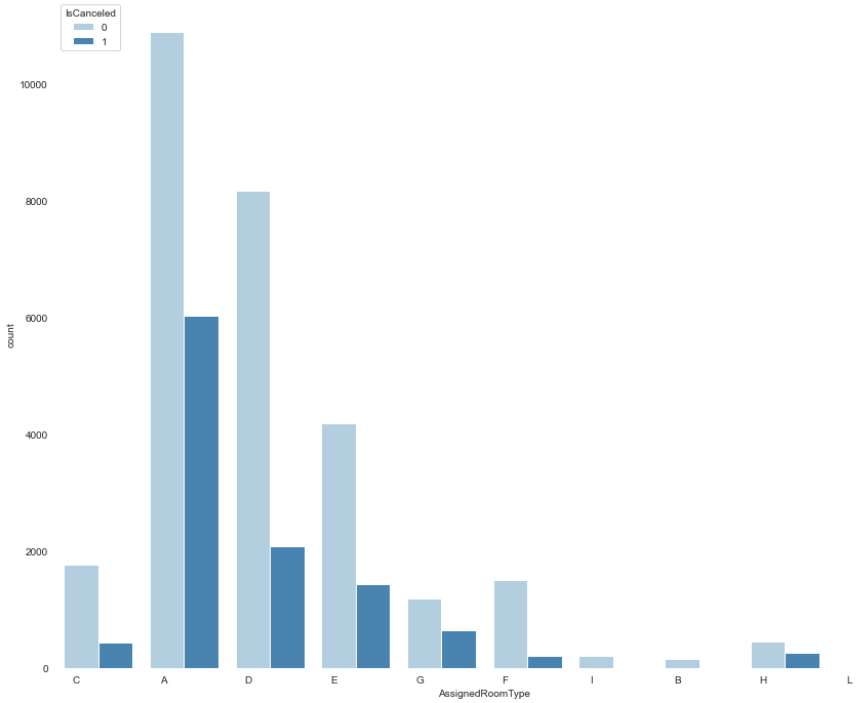
Violin Plot for Market Segment compared with Total number of Guests

In the above graph we looked at how different market segments bring in what number guests. It was important to identify which market segment is bringing in our most guest, i.e two guests which forms a huge chunk of our dataset. Corporate brings in single guest where as both online and offline TA combined with Direct category brings in the couple guests which are predominant in our dataset.

Assigned Room Type

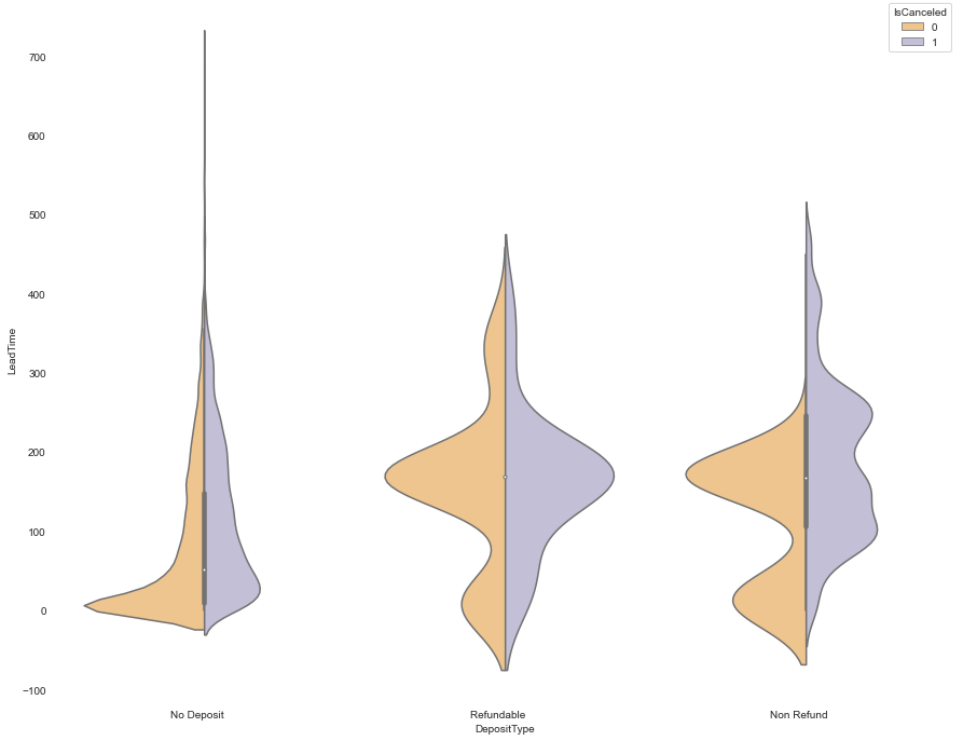
We found that a huge chunk of guests preferred room type ‘A’. In fact, comparing it against reserved room type, ‘A’ was the most booked. And whenever there were was a change in the room assigned than reserved it was often that it was room type ‘A’ that was assigned to a new type, typically ‘D’ or ‘I’ which is the reason we have the spike for them when we analyze the histogram of assigned room types.

Since, room type A was the most reserved and assigned, there was a proportional high cancelation rate which is in harmony with the cancelation rate of the whole dataset.



Histogram for Assigned Room Type

Deposit Type

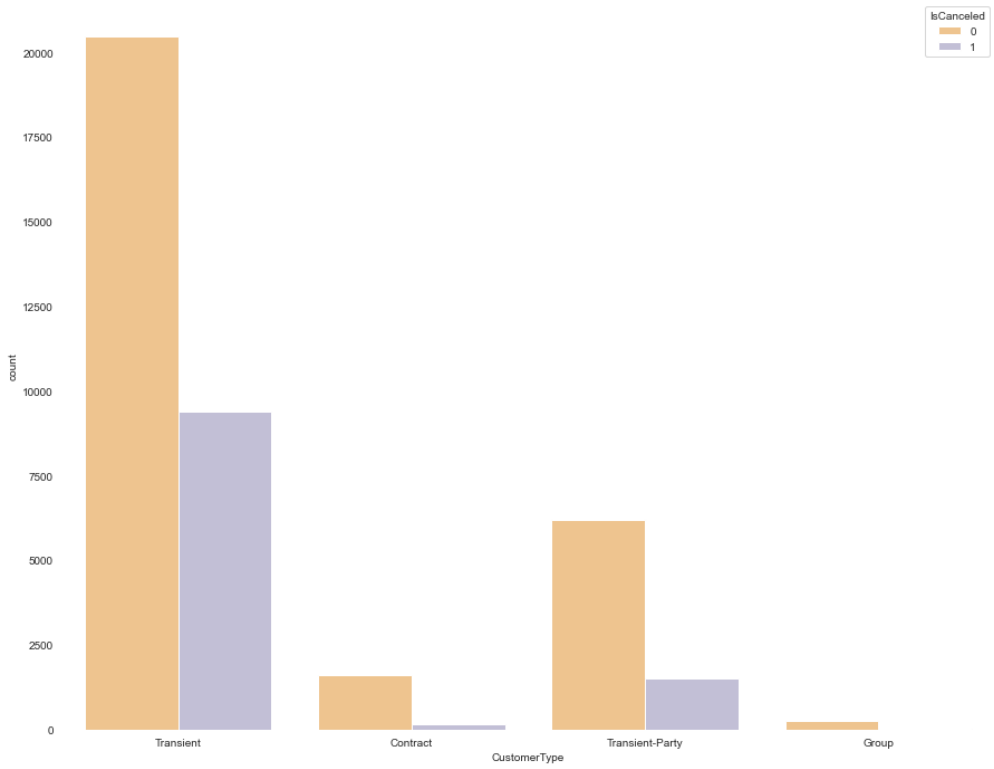


Violin Plot for Deposit Type

There are three categories for deposit type – No deposit, Refundable, Non- Refundable. From the violin plot we could see a deposit was taken when a booking was made more than 5 months before. There more when guests made booking six month or an year in advance there was a deposit that was levied.

It was astonishing to see that there was considerable cancelation in the non-refundable deposit category. This goes against the intuition that paying a non-refundable fee up front is a commitment of the tallest order. In light of this finding, we analyze it against the lead time and figure out the cancelations are spread evenly across a long period of time, meaning it should be some sort of personal or professional emergencies that push guests to cancel.

Customer Type



Histogram for Customer Type

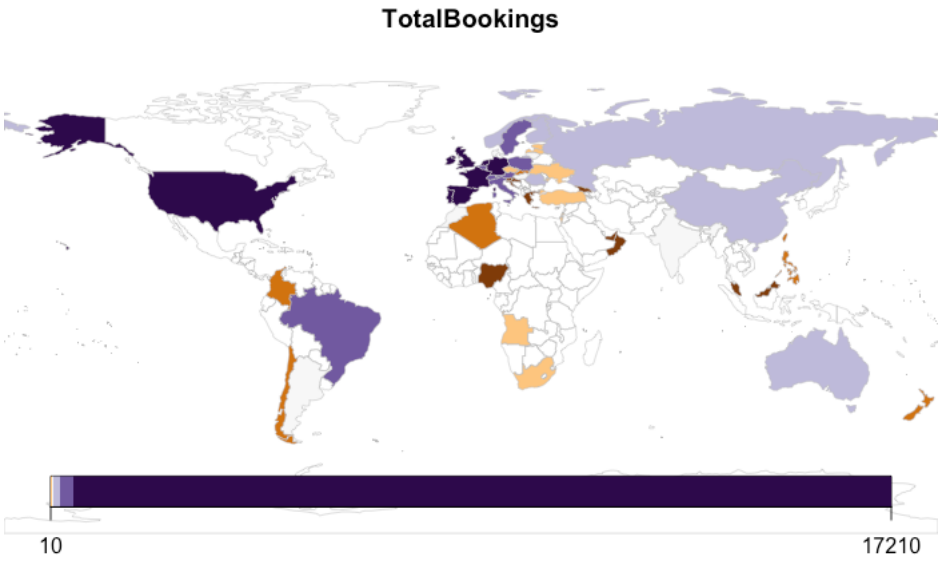
Looking at the dataset, we found there was major group of transient guests and a very small percentage of Groups. Also we can see that corporate has the lowest cancelation rate, a pattern we have seen over and over now after analysis of various attributes.
Transient guests which form a major part of our dataset have a cancelation rate of close to 50% which is in harmony with our dataset.

From our analysis of other categorical variable, we couldn't see much effect of them on our cancelation rate. Meal plan choice did not influence the cancelations, we found that 'BB' which is Bed and Breakfast is the most preferred by all guests. Though country is a categorical variable in this dataset, we have chosen to analyze it independently. We felt that the influence of the origin country of guest has little to no effect on the cancelation and introducing it in the equation would only skew our analysis. Therefore, we decided to perform independent analysis on the country attribute to identify good performing countries and understand which market segment is dominated by which country.

Country

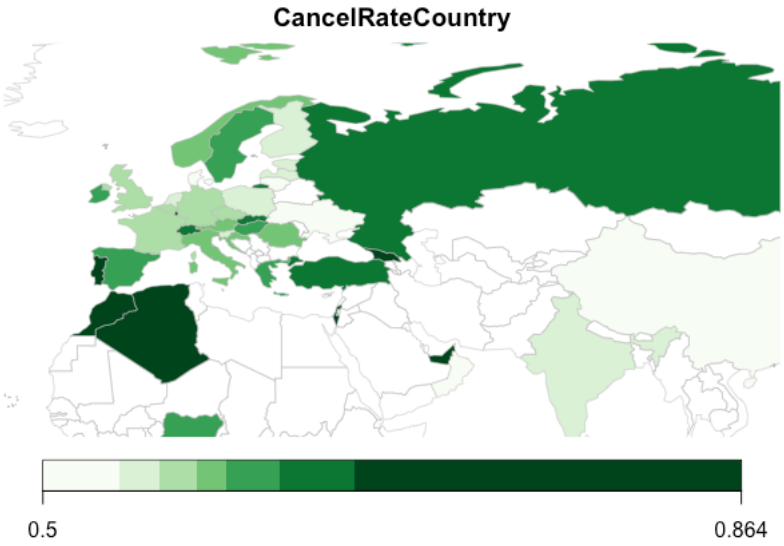
We found that there are 464 entries of countries where the country was NULL. And further more when combining with ISO-3166 there were three abbreviations which were not part of the standard that were present in the dataset, and therefore only 51 countries could be mapped.

Below is the total bookings plotted on the world map. We could see that the most bookings came



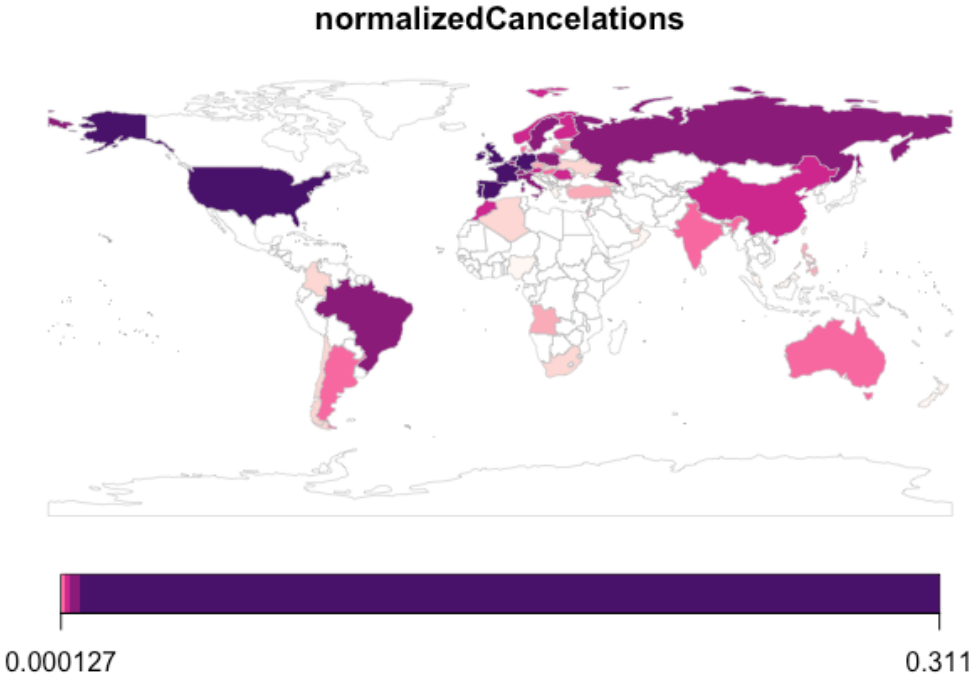
Total Bookings plotted on world map

from Portugal, Great Britain, Spain, Ireland, France in that order. To get a clearer picture it was necessary to plot the cancelation rate of a country, i.e how many cancelations were there for every booking from that country. This factor would be a better judge of a country's performance. On analysis we could see that United Arab Emirates, Georgia, Philippines and Morocco had the most cancelations compared to the number of bookings from that country.



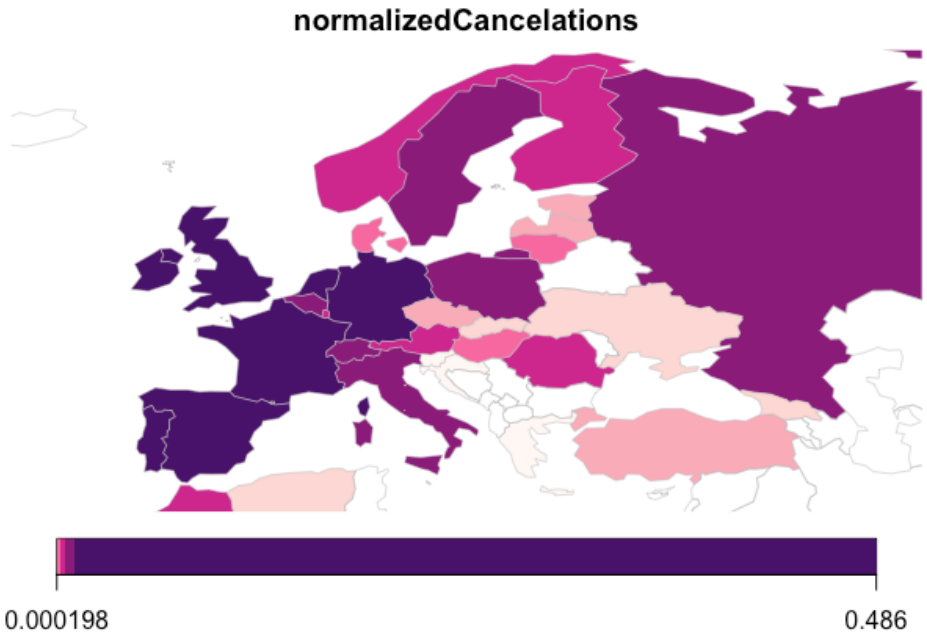
Country-wise Cancelation Rate for Eurasia

But it doesn't paint a clearer picture because it is important to compare these against how the whole dataset performs. Therefore, the following map plot is of all countries with normalized cancelation rate.



Normalized Cancelation Rate of Country On World Map

After normalization we have the same sequence of top 5 countries which we identified with highest booking numbers. Therefore it was important to normalize the cancelation rate as we can see that the results are different when we compare it to countries with highest cancelation rate.



Normalized Cancelation Rate of Country For Eurasia

Attribute Importance

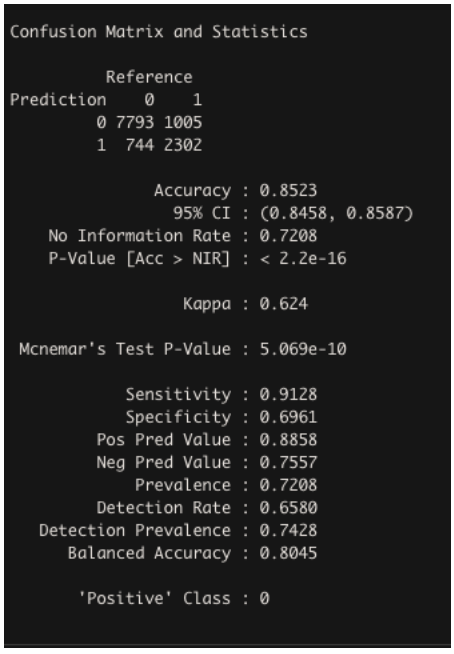
Since the scope of the project is to identify what is causing cancelation and how to reduce it, we found that creating a model to identify customers who will cancel is redundant. Machine Learning models to the most part are black-boxes and therefore, they do not align with the scope of this project.

But one part of ML model that allows a little transparency in the black box is feature importance. When a model is created, we can analyze the influence of each feature on predicting the outcome. This in combination of our previous analysis through data visualization can give further insights as to which variables are supposed to be focused on, to bring about a significant change in the outcome (here improve cancelation).

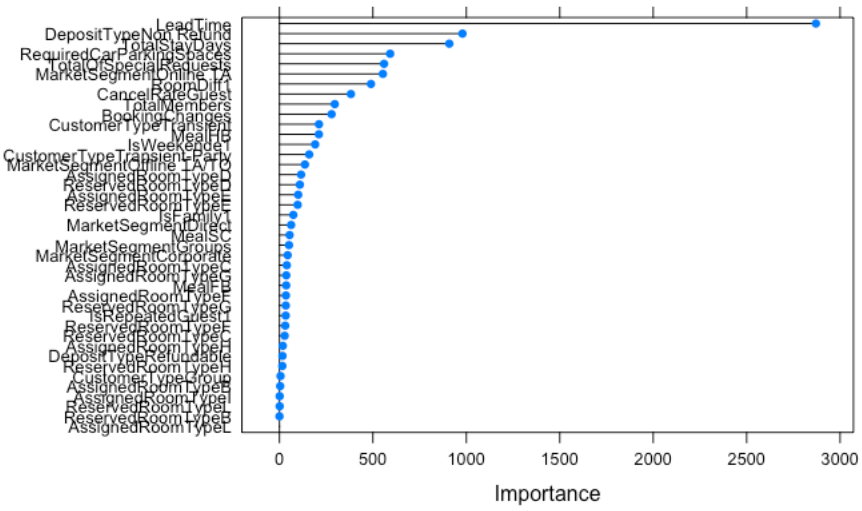
Since we only need to derive feature importance, we will use a classification machine learning algorithm. We chose to use Support Vector Machine. It is because our outcome variable is binary categorical and a classification algorithm like SVM will gives us a output in form of categories.
SVM generates an optimal hyperplane between the features such that there is a distinct separation with the outcome categories.

We use a 70-30% split for our training and testing dataset. The model was initially trained on all original attributes, then combined with our derived attributes and then attributes that weren't contributing any substantial information to the prediction model were dropped based on the feature importance gained from the initial models.

We got the model with an accuracy of 85%. The columns used were {"IsCanceled","LeadTime", "Meal","MarketSegment","IsRepeatedGuest", "ReservedRoomType","AssignedRoomType", "BookingChanges","DepositType","CustomerType", "RequiredCarParkingSpaces", "TotalOfSpecialRequests", "TotalMembers", "IsFamily", "TotalStayDays", "IsWeekende", "CancelRateGuest" , "RoomDiff" }



We further generate an feature importance plot that indicate to us the influence of each attribute passed to the model. It is evident that lead time has the most influence on cancelation. Therefore



Feature Importance Plot

if we have to improve on cancelation we will have to work with lead time and compare all other attributes with it. Followed by that is Nonrefundable deposit type and total stay days along with market segment of Online TA. In the earlier sections we have identified how skewed and dominating all of these factors already are in our dataset.

After this we go back and compare all our attributes against lead time and therefore all conclusions and inference discussed in the earlier sections were drawn after conclusion of this model analysis.

Insights and Recommendation

From the analysis we could see that the main demographics of the hotel is bookings with two guests, followed by single and then families. The main attribute which showed strong influence on cancelation was the lead time. As the time between booking and check-in date increased there was an increase in probability of a booking getting canceled.

Against intuition when deposit type was analyzed there was a high rate of cancelation when the deposit type was non-refundable. This was spread across the entire lead time as compared to a certain period.

There is also a huge influx coming from market segment of Online TA, which is the majority share of bookings followed by direct bookings. There is no significant cancelations when there is a change in the room assigned compared to the one reserved.

Given our numerous analysis and visualizations of the attributes provided to us, we have the following recommendations which we believe will help the hotel reduce their cancelation.

Restricting lead time variable

We could see there was a spike in cancelation when the booking was made more than 4 month in advance. We recommend the booking period be capped at four months before check-in date. This according to our analysis is the perfect spot to optimize cancelation rate.

At the same time there can be a bigger cut off period for group booking, as groups tend to cancel less even if the reservation was made 8-12 months in advance. This allows the hospitality to maximize the reservation without losing any potential customers.

Providing deals to guests who made non-refundable deposits

A major chunk of guests that canceled were of the type who paid non-refundable deposit. Though this is good for business has we earnt the deposit money from the plight of our cancelling guests we could use a fraction of the money to provide these guest with deals and coupon or further persuade them to return to your hospitality when they can.

Researching further into cancelation from Online TA

There is a huge influx from the online travel agent market segment but we also saw equally high cancelation from that segment. It is pertinent to figure out the reason behind it. It could be two fold, where in the mechanisms of online TA do not punish guests to cancel or provide them with influential deals which in the bigger picture are more worthwhile and therefore cancelation can be

affordable; or there is a possibility of false advertising where there is disparity between what was presented and promised and what was delivered by the hotel.

Surveying guests

The analysis had shown that a very small percentage of guests are repeated and its predominately single guests from corporate sector which intuitively makes sense. Since, the hotel isn't able to convert other spectrum of guests into returning customers, it is important to get feedback about what can be done to better improve the guests experience that the hospitality is worth returning to.

Cancelations influence the total revenue made. From our analysis, we recommend the following that could help boost the total revenue and not just reduce cancelation per se.

Identifying the traits of room type A

Room type A was predominantly the one that was requested in close to 50% of the bookings. If we can further analyze why room type A is preferred and increase its capacity to avoid the room being unavailable to guest who have reserved it requiring a change in the room we can further reduce cancelation. If the hotel can emulate the offerings of room A they can see a boost in their bookings.

Expanding on contract customer type

Exploring the option of collaborating with corporates and event management companies to expand on contractual based guest will further boost bookings. Currently, the contract customer type is extremely undervalued and has a potential to expand. If tapped into we can have a increase in bookings and if the cancelation pattern continues we will have less cancelations in this customer type.

Focusing on corporates

Expanding on the above idea, corporates delivers a huge chunk of the solo guests in the hotel. Exploring the options to better serve these business individuals with business centres, faster internet, lounge and private meeting rooms,, would help boost customer rentention. The factor for focusing on this market segment is it is the segment with least cancelation ratio and a lot of potential.