

A Guide to Reduce the Number of Cancellations within the Hotel Industry

Project by:
Jay Maru, Colin Hardy, Simran Pathak, Khadijah Sabu, Manasi Todankar



Table of Contents:

Table of Contents:	2
1. Abstract	3
2. Introduction	4
2.1 Business Questions:	4
3. Examining the Data Set	5
4. Workflow	13
4.1 Data Cleaning:	123
4.2 Exploratory Data Analysis:	123
4.3 Feature Selection for Machine Learning Process:	133
4.4 Applying Rules	133
5. Coding Methods	134
5.1 Visualizations	135
5.1.1 Histogram	135
5.1.2 Box Plot	145
5.1.3 Bar Plot	145
5.1.4 Linear Model	145
5.1.5 Support Vector Machine	156
5.1.6 Maps	156
5.1.7 Decision Tree	167
6. Model Results	178
7. Suggestions to Ownership	199
8. Further Research and Future Uses	21
9. Conclusion	244



1. Abstract

Understanding the customers and their interaction with the business is very essential for the growth of the organization. Knowing the behavior of customers can help the business understand what the customers want, what they are getting and hence make further strategies. Sometimes, this can also change the business objectives, missions, and goals of an organization. Understanding the customers helps the business tailor the facilities on a granular level which in turn results in strong customer relationships and new sales through positive word of recommendation.

Customer satisfaction is the key to business growth in this highly competitive world. A customer has got a lot of options to choose from and thus becoming that one option which is the best for customers is the end goal of all the businesses operating across the globe. Even companies like Google, Microsoft, Amazon, and Facebook have got competitions and hence they invest a lot in understanding their customers and their feedback. Because of the data revolution happening across the globe, it is becoming more and more easy to gather the data from various channels to perform analysis and build up on strategies further. The aim of this project is to determine why people cancel hotel reservations and/or better predict who will cancel. We have used one dataset that contains real-life hotel stay data, with each row representing a hotel booking. The dataset was processed further to utilize it for getting deeper insights into the data and understand the data by preparing visualizations. Strategies were formed based on the business questions we obtained from the insights depending on their level of importance and their impact on the business. We conclude by giving out recommendations to the hotel manager on ways the reservation process can be improved to minimize chances of cancellations.



2. Introduction

This project was undertaken for the course of IST 687 – Introduction to Data Science. The goal of it was to intake a dataset provided by a hotel booking database, and find solutions to lessen the chances that a booking is canceled. Due to the ease of cancellations via online booking websites and unavoidable pandemic related issues, vacation plans are getting changed and reservations canceled at an increased rate. This creates discomfort for many institutions and creates a desire to take precautions. Therefore, predicting reservations that can be canceled and preventing these cancellations by some form of solutions provided will create a surplus value for the institutions. By creating models and finding solutions for this problem, we can give suggestions to the hotel's upper-management in hopes of lessening the cancellation issue.

Business Questions:

Below are the list of questions to help achieve the project goals:

- How does the market segment of booking affect cancellation?
- How does the lead time of a booking affect cancellations?
- How does the different deposit type affect cancellation of a booking?
- What are the other factors that affect cancellation of booking?
- What machine learning algorithm has the highest accuracy when it comes to predicting hotel booking cancellations?

3. Examining the Data Set

The provided data set shows the records of 40,060 reservations with hotels. We did not get specifics about whether or not it's from a single hotel, a chain of hotels, or from a website that features several different brands of hotels. Provided were 20 variables, each describing a certain aspect of the reservation. The first, and most important variable being "IsCanceled". Given that our entire project is based around limiting the cancellations, this is the response variable in our models. It, as it sounds, notes whether or not a booked reservation was canceled. The next variable included is "LeadTime". This one clarifies how far in advance the reservation was made. It spans from 0 days, to 737 days, with a median of 57, and a mean of almost 93 days. The following two variables relate to the duration of stay, one denoting weekend nights, and the other week nights. These span from 0 nights, to 19 and 50, respectively. The following three variables count the number of people staying in the room, and classify them by age. There are categories for adults, children, and babies, although there is no description to clarify at what age each group is cut off at. The following columns detail the meal plan that the customer ordered with the room, in addition to their country of origin. Market segment is the next column, denoting whether or not a travel agent or tour operator booked the room for a client. Whether or not the customer is a repeated guest (1 for repeat, 0 for new customer), and the number of prior bookings that they have canceled or not canceled are the following three variables. The room reserved and room assigned are also provided, followed by the number of times the reservation was changed. The deposit type whether that be refundable, non-refundable, or no deposit at all is next. Finally, the customer type, described as whether or not they booked through transient contracts, is provided, along with the required number of parking spaces, and number of special requests. The provided table below shows the descriptive stats for each non-Boolean numeric variable in our data set. If the variable is not shown below, it is either a categorical variable, or simply a 1 or a 0. We also added five more

variables i.e. mutate the data to produce useful columns and subsetting. Below is the snapshot of the variables that we created to get better insights about the model:

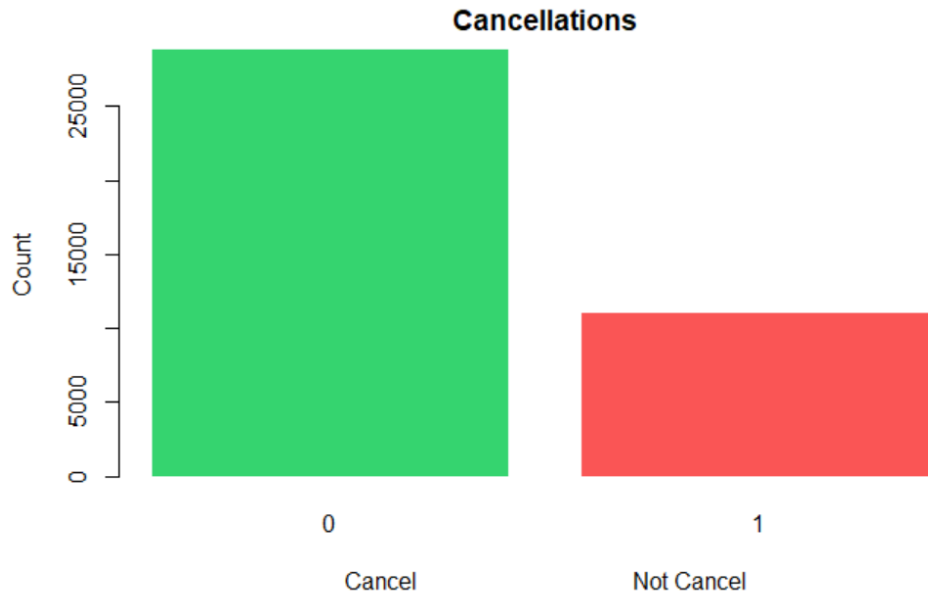
```
#Mutating data to produce useful columns and subsetting
hotel_data$totalpeople <- (hotel_data$Adults + hotel_data$Children +
hotel_data$Babies)
hotel_data$roomassigndiff <- ifelse(hotel_data$ReservedRoomType==
hotel_data$AssignedRoomType,0,1)
hotel_data$kids <- (hotel_data$Babies + hotel_data$Children)
hotel_data$family <- ifelse(hotel_data$kids>0,1,0)

hotel_data$totalstaylength <- (hotel_data$StaysInWeekendNights +
hotel_data$StaysInWeekNights)
```

Variable Name	Min	1Q	Median	Mean	3Q	Max
Lead Time	0	10	57	92.68	155	737
Stay - Weekend Nights	0	1	1	1.19	2	19
Stay - Week Nights	0	2	3	3.129	5	50
Adults	0	0	2	1.867	2	55
Children	0	0	0	0.1287	0	10
Babies	0	0	0	0.0139	0	2
Previous Cancelations	0	0	0	0.1017	0	26
Previous Bookings Not Cancelled	0	0	0	0.1465	0	30
Booking Changes	0	0	0	0.288	0	17
Parking Spaces Required	0	0	0	0.1381	0	8
Number of Special Requests	0	0	0	0.6198	1	5

In order to better understand the goal of our project, we created several visualizations to show any clear tendencies, right off the bat. We wanted to see if there were any trends that a simple graph could show that lead to an increase or decrease in the likelihood of a cancellation. The following graphs break the data set into several different sections in which we analyzed separately.

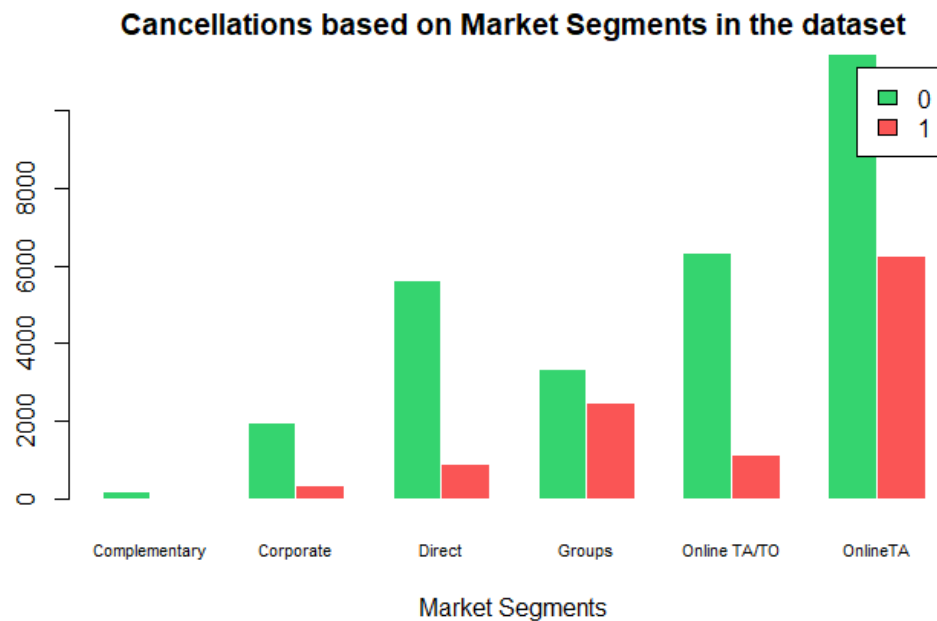
By creating graphs, it allows us to visualize the data and get a better understanding of trends. For example the bar chart below depicts the total number of reservations that were and were not cancelled. Rather than looking at numbers, the graph shows that roughly a quarter of all reservations were cancelled. We're hoping our suggestions can lower that number.



Range <chr>	NotCanceled <dbl>	Canceled <dbl>
0-51	13042	2997
51-101	3850	2112
101-151	2685	1790
151-201	2677	1408
201-251	1712	1250
251-301	954	686
301-351	571	475
351-401	309	160
401-451	12	31
451-501	21	56
501-551	24	0
551-601	0	0
601-651	0	0
651-701	0	0
701-Inf	2	0

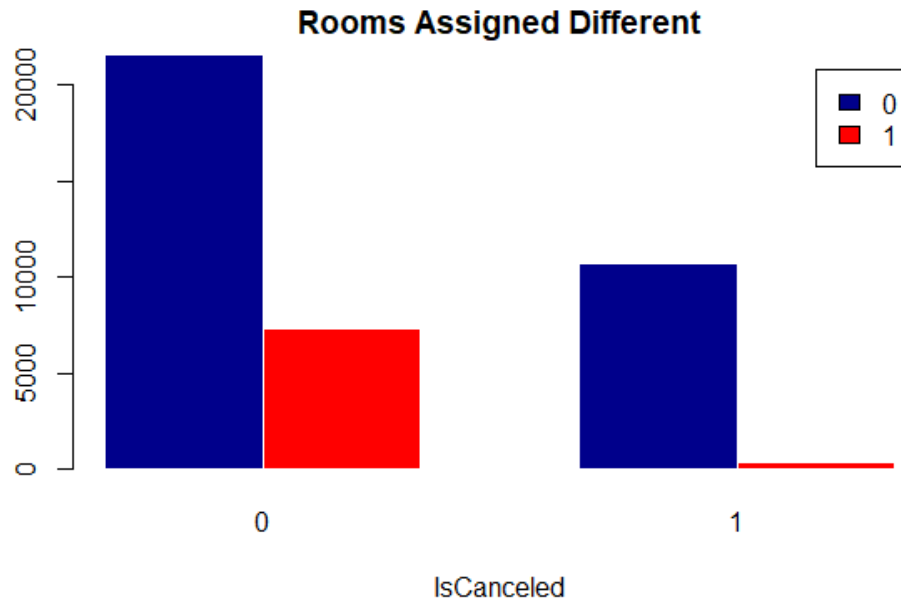
By comparing the values of reservations canceled and not canceled, we can observe that reservations having higher Lead time have almost the same number of cancellations to the

number of no cancellations. This naturally suggests that reservations made quite early on are more likely to be canceled than the reservations made near to the date of travelling.

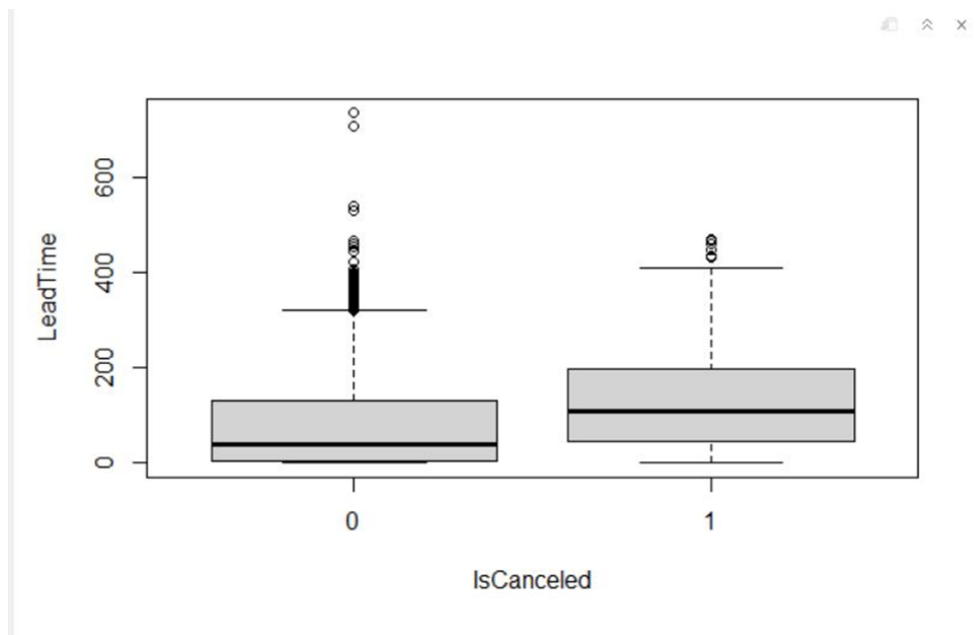


From the graph we can infer that the customers who reserve a booking through an online tourist agent are likely to cancel.

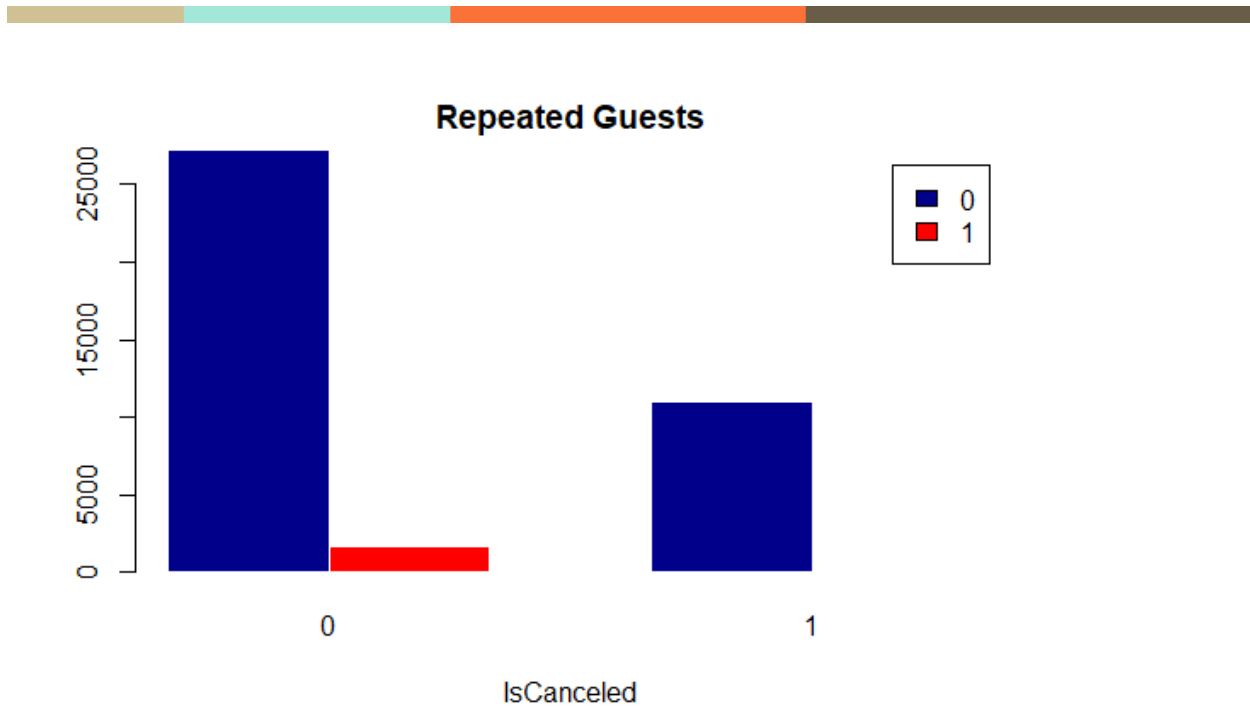
This next chart looks at a subset of the “groups” option of the market segment variable. By subsetting variables, we can see which trends are especially strong in certain groups. We decided to look at groups, because as the chart at the bottom shows, that was the market segment that had the highest proportion of canceled reservations, comparatively. In this case, the chart below shows that there were surprisingly more cancellations with the non-refundable deposit type compared to the others. Although this goes against common logic, the trend remains the same, albeit to a lesser magnitude, amongst other subsets.



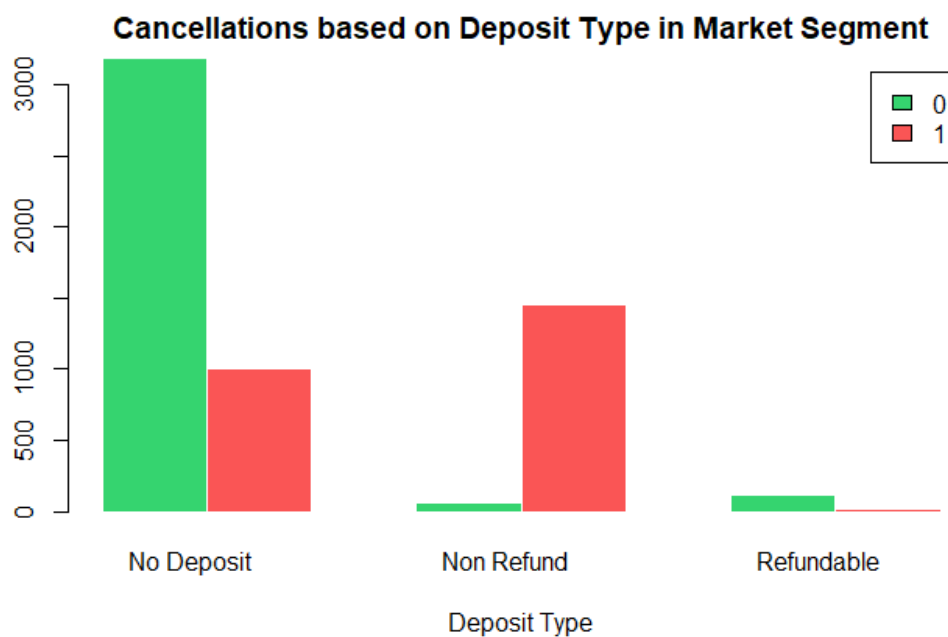
It can be found that customers tend to cancel when they are assigned to different room compared to the room assigned.



The canceled median is higher than the lead time median, that is, people who have greater lead time have canceled more than the people having lesser lead time.

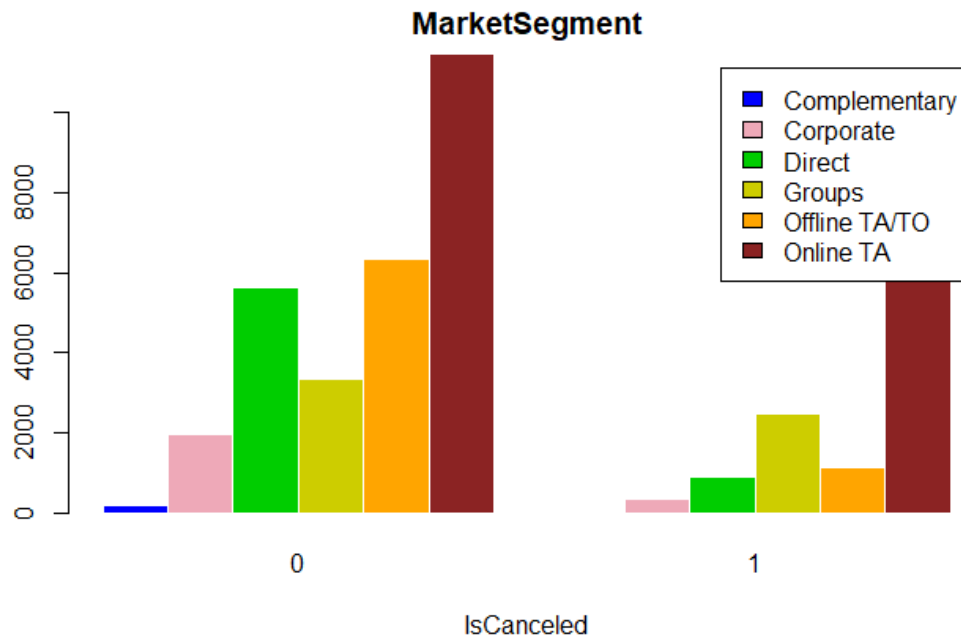


We can infer from the above graph, the repeated customers are likely to cancel less as compared to new customers.



The final chart we looked at also included the market segment variable. For this one we broke the dataset into canceled and not canceled. From there, we chose to look at the count of market

segments in each category. As can be seen below, there was a stark difference in both direct, and offline TA/TO. This suggests that those two options tend to have a lower cancellation rate.



4. Workflow

The dataset is imported in RStudio and loaded. The packages and libraries are installed to perform visualizations, modeling, subset, etc. We performed the data manipulation methods to find the missing values(using is.na function), subset data(subset function), and summary/structure(summary function) the data. This helps to discern the data in order to convert the datatype to numeric and thus, the quality of data is measured. Once the data is cleaned, data visualization is performed. As we know, "IsCanceled" is a key variable here, data is divided into segments to glean the insights with other associated variables to find the limitations of cancellations. Using boxplot, we discovered that the canceled median is higher than the lead time median, that is, people who have greater lead time have canceled more than the people having lesser lead time.

4.1 Data Cleaning:

- Examining the entire dataset based on the columns that it has.
- Eliminating the missing value in the country table (based on the context) because there are many outliers in the dataset.
- Converting all the character values to numeric ones using as.factor().

4.2 Exploratory Data Analysis:

- Creating different types of inferential statistics like histogram, bar plots, box plots, pie charts to get a clear idea on which basis the hotel cancelation is being made
- Creating a subset of IsCanceled columns into two parts.
- Working particularly on the data which is based on hotel cancellations.
- Comparing IsCanceled with various important columns to find out cancelations based on data context.

4.3 Feature Selection for Machine Learning Process:

- Creating data partition into train set and testset using the original data
- Applying all the different models such as linear modeling, supervised model, support vector machine, ksvm model, association rule mining techniques. to compare which model gives the more accurate data.
- Creating a confusion matrix to check the accuracy of the particular model and to compare the models with one another.

4.4 Applying Rules

- After finding out the better model, applying rules on the modeled data to infer some of the insights as to why the hotel cancellations are being made.
- Providing some suggestions which will reduce the number of hotel cancellations soon.

5. Coding Methods


5.1 Visualizations

Visualization is a quintessential strategy to derive insights from the data which engenders decision-making. It may considerably increase the quality and artistic view of the graphs while also increasing the desired efficiency. For this purpose, the “ggplot2” package is installed to perform any type of charts/ graphics. Some of the arguments involved are customizing a title, highlighting, adding annotations or using faceting.

The following are the visualizations used in the datasets -

1. Histogram

The histogram is a graph that divides data into bins (or breaks) and displays the frequency distribution of these bins. You may also alter the breaks to examine how they affect data display in terms of readability. The `geom histogram()` function in `ggplot2` can be used to create



histograms. It only accepts one numeric variable as input. This program divides the variable into bins and counts the number of data points in each bin automatically.

2. Box Plot

Boxplots are one way of determining how evenly distributed the data in a data collection is. The data set is divided into three quartiles. This graph depicts the data set's minimum, maximum, median, first quartile, and third quartile. Drawing boxplots for each data set allows you to compare the distribution of data across data sets. The minimum, the 25th percentile, the median, the 75th percentile, and the maximum are all statistically significant numbers in a box plot. As a result, it's beneficial for visualizing the data's spread and drawing inference from it.


3. Bar Plot

Bar graphs are useful for illustrating variations of cumulative totals across multiple groups. For bar plots in several categories, stacked plots are performed. The `barplot(height)` function creates barplots, where height is a vector or matrix. The values determine the heights of the bars in the plot if height is a vector. If `beside=FALSE` and height is a matrix, each plot bar corresponds to a column of height, with the values in the column indicating the heights of stacked "sub-bars." When `beside=TRUE` and height is a matrix, the values in each column are juxtaposed rather than stacked. Option names should be included. To label the bars, we use `arg=(character vector)`. To make a horizontal barplot, we use the option `horiz=TRUE`.

4. Linear Model

Based on one or more input predictor variables X , linear regression is used to predict the value of an outcome variable Y . The goal is to build a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable so that we may use this formula to estimate the value of the answer Y when only the predictor variable(s) values are known. The mathematical equation can be generalized as follows:

$$Y = \beta_1 + \beta_2 X + \epsilon$$



where, β_1 is the intercept and β_2 is the slope. Collectively, they are called *regression coefficients*. ϵ is the error term, the part of Y the regression model is unable to explain.

5. Support Vector Machine

SVM (Support Vector Machine) is a supervised machine learning algorithm for classifying data into different categories. SVM, unlike most algorithms, employs a hyperplane that serves as a decision boundary between the various classes.

SVM can be used to create numerous separating hyperplanes, dividing the data into segments with just one type of data in each segment.

SVM's characteristics include:

1. A supervised learning algorithm is SVM. SVM is trained on a collection of labeled data in this way. SVM examines the labeled training data before classifying any new input data based on what it has learned during the training phase.
2. SVM has the advantage of being able to solve both classification and regression issues. The SVR (Support Vector Regressor) is utilized for regression problems, despite the fact that SVM is well recognized for classification.
3. Using the kernel approach, SVM may be utilized to classify non-linear data. The kernel trick entails translating data into a new dimension with a distinct dividing line between data classes. Following that, you can easily build a hyperplane between the various data classes.

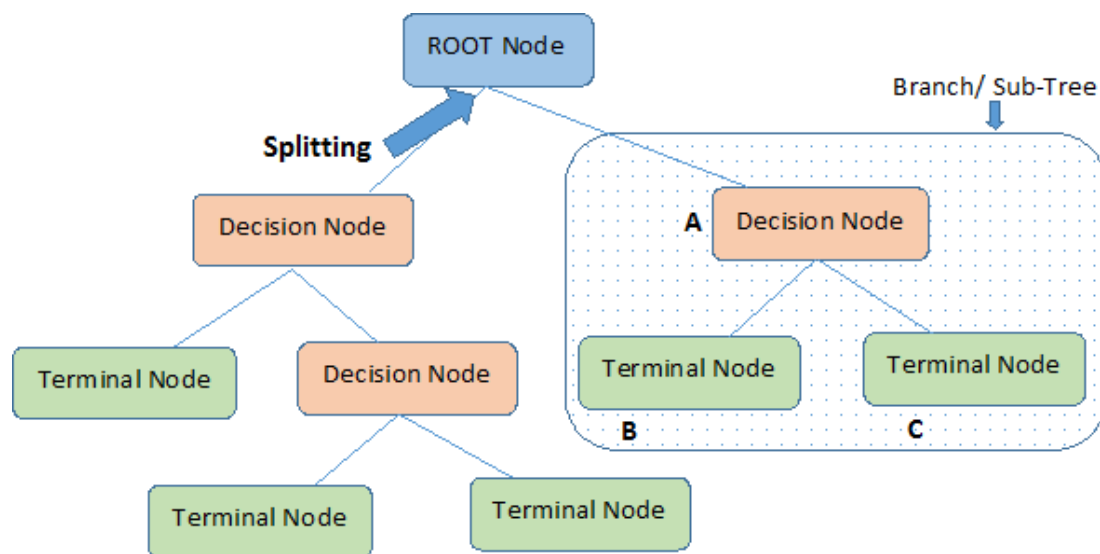
6. Maps

Given data with location variables, it allows us to create a world map with the color of the country signifying the magnitude of the variable. This allows users to have a better understanding of an actual map compared to a column of just abbreviations. The map below shows our dataset, with the color of the country corresponding to the average in which the country reserves in advance.

7. Decision Tree

A decision tree is a form of supervised learning algorithm that can be used to solve issues in both regression and classification. Both categorical and continuous input and output variables are supported. The terminologies of decision trees are -

- The complete population or sample is represented by the Root Node. It's then separated into two or more homogeneous groups.
- The process of splitting a node into two or more sub-nodes is known as splitting.
- A Decision Node is formed when a sub-node splits into more sub-nodes.
- A Terminal Node, also known as a Leaf, is a node that does not split.
- Pruning is the process of removing sub-nodes from a decision node. Splitting is the polar opposite of pruning.
- The term "branch" refers to a segment of a tree.
- The parent node of the sub-nodes is referred to as the parent node, while the sub-nodes are referred to as the child of the parent node.



Note:- A is parent node of B and C.

6. Model Results

From the results of our linear models, the most glaring variables that affected cancellation rate were: lead time, returning customer, deposit type, room type differentiation, and market segment. Lead time was positive, meaning that for every one increase in lead time unit (in this case day), there's a .007 percent increase in the cancellation rate. Although it's a small increase, it is still significant. The other numeric variables show negative estimates, meaning there is a decreased chance in cancellation with an increase in the variable. The output of our model is shown below:

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.847e-01	3.266e-03	87.185	< 2e-16	***
DepositTypeNon Refund	5.883e-01	1.005e-02	58.549	< 2e-16	***
DepositTypeRefundable	-1.220e-01	3.351e-02	-3.641	0.000272	***
LeadTime	5.101e-04	2.148e-05	23.748	< 2e-16	***
roomassigndiff	-2.152e-01	5.144e-03	-41.843	< 2e-16	***
IsRepeatedGuest	-9.551e-02	9.813e-03	-9.733	< 2e-16	***
RequiredCarParkingSpaces	-2.422e-01	5.750e-03	-42.123	< 2e-16	***

It's important to note that the adjusted r-squared value is not as high as we would like, but this is just showing which variables are important and in which way does it affect the cancellation rate. The most surprising outcome is the difference in deposit types. Common sense would lead us to believe that a refundable deposit would lead to more cancellations, but it's actually the opposite.

Our next model we used was an SVM model, we used IsCanceled as the dependent variable, and all of the remaining variables as independent variables. Using this method, our model accuracy came out to be 87%. This means that given a data entry, the model can correctly predict whether or not that entry will end up cancelling. We are able to use this model

to see which variables are most important in deciding whether or not the reservation will be canceled. Featured below is the output of that model, showcasing an 87.5% accuracy rate.

	Reference	
Prediction	0	1
0	15988	1630
1	1374	5043


Accuracy : 0.875
95% CI : (0.8708, 0.8792)
No Information Rate : 0.7224
P-Value [Acc > NIR] : < 2.2e-16



7. Suggestions to Ownership

Given the results of our models and looking at the graphs, it is clear to us that we can make several suggestions.

- **Don't allow reservations more than two months in advance.** Now this is a slippery slope, as some customers may be deterred from booking with your hotel if they're trying to book far in advance. However, our data shows that there is a positive relationship between lead time and cancellation rate. If we were to limit the maximum lead time to 60, that would give customers less opportunity to have things come up forcing them to cancel their trip.
- **To create more parking spaces.** There is a negative correlation between the number of parking spaces requested and the cancellation rate. In other words, the more parking spaces required, the lower the cancellation rate. If ownership were to build a parking garage or buy land around the hotel to create more parking spaces and guarantee availability, it is less likely customers will cancel their trip.
- **To offer more refundable deposits.** This strategy may change based on some details of the hotel we're working for, more on that later. Possibly the most surprising find in the data set was that cancellation rates were higher with non-refundable deposits. And while that's a great insurance option, the hotel would rather have the cost of the room for several nights rather than just a one-time smaller deposit. If our hotel is consistently sold out, then more non-refundable deposits would be better. That way the cancellation rate would go up, giving us the deposit money, but also opening up the room to be sold again. It would be a win-win. However, in a more likely scenario, our hotel has vacancies it's trying to fill, the refundable deposit will hopefully prevent cancellations. The next variable we found to be important was when the assigned room type was different from the reserved room. Although we are unclear what the quality of



the room is based on our data set, it can be assumed that guests are more likely to keep their reservation booked if they are getting a deal. In the following section however, we go more in depth about this variable.

- **To offer a small discount to returning customers.** The final solution we can suggest to ownership is to offer a small discount to returning customers. We found there to be a negative correlation between returning customers and their resulting cancellation rate. In order to entice the customers who are less likely to cancel back to the hotel, we suggest offering them a slightly discounted rate.


8. Further Research and Future Uses

Although our models were able to provide useful insight to management about potential strategies that can be used to lessen cancellations, the models can be improved. If our dataset had more information available to us, our models would likely be more accurate. Some useful variables that would be helpful for future research include:

- Total expected price of stay / Price of refund.
- A more detailed description of room type.
- The coordinates of the customer's hometown.
- The dates of their stay.

These added variables, just by using common sense would likely have an effect on the chances of cancellation rate. Looking at the price of the stay and amount refunded/deposited, customers would be far less likely to cancel if their price of stay is thousands of dollars and the amount refunded is only a small fraction of that. Although our dataset surprisingly shows that refundable deposits lead to fewer cancellations, rather than non-refundable, we may be able to suggest a change based on this. Depending on the popularity of our hotel, it may be beneficial to take in the non-refundable deposit, therefore opening up the room to be sold again. If it's a down season for tourism though and there is an abundance of excess rooms, management would rather have the price of the room rather than just the deposit.


The next added variable which could be helpful is the detailed room description. In our provided dataset, the room type was given a letter A-P. We came up with a variable showing if the room that the customer reserved is different from the room that they actually received. The variable proved to be statistically significant and with a negative coefficient, leads us to believe that a change in room leads to fewer cancellations. However, there is no index to describe the quality of a room, for example room type A is the most basic room, while P is the most luxurious room. If this were to be provided, we could split the created variable into two parts. One



showing if the actual room given to the customer is an upgrade to what they reserved, or vice versa, a downgrade. It would make sense that an upgraded room (for the price of the cheaper room) would lead to fewer cancellations, but we can't know that unless we are given the provided data. If that were the case, we could make a suggestion to offer an upgraded room for the price of a lower quality room, if availability allows. Although it may not be a business-savvy decision to offer better rooms for lower prices, our goal was to lower cancellations, and that would do the trick.

The final two variables that could be useful are the coordinates of the customer's hometown, and the date of their stay. In order for these to be useful, we would also need more information about the hotel, especially its location. In a real-life scenario we would be working for the hotel, so that wouldn't be an issue to obtain. With these added variables, we would be able to create a new variable calculating the distance between a customer's hometown and the hotel. Logically, the variable could go both ways. The further one travels, the less likely they are to cancel because it's a rare occurrence to get the opportunity to travel. On the other hand, they would be more likely to cancel thanks to the added stress and unknowns that come with traveling a long distance. Similarly, the date of the stay would be useful to understand whether or not it's a tourist season. If our hotel is located in Miami and customers are booking for spring break, our model would benefit by taking situations like these into account. We could also build variables about season, and therefore weather around the date variable. For example, a hotel in Syracuse during the dead of winter may get more cancellations than in summer due to the weather. As mentioned before, in a real-world situation in which we know more specifics about our hotel, we could create a model that would have a better understanding of which variables play a role in impacting the cancellation rate.

In the given scenario, we were under the impression that we were only working for one single hotel or small hotel chain. If we were to approach the booking websites such as Travelocity, Booking.com, Hotel Trivago etc. they could use this information just as well. Even



looking at a grander scale, we could use similar approaches with the entire tourism industry. Given the proper data, we could run similar machine learning models on car rental cancellations, restaurant reservations, or airline tickets. Every company is trying to get a competitive advantage compared to its competitors, and would jump at the opportunity to find out how they can bring in more money.



9. Conclusion

At the current levels, 27.8% of all reservations get canceled. By looking at the data and coming up with solutions to fix this problem, we believe we can lower that number significantly. If ownership were to, when possible, upgrade rooms, not allow for reservations to be booked further than two months in advance, and allow more refundable deposits, we believe the cancellation rate will fall. However, if we were to be provided a more complete data set, it would allow us to better fit our model and improve our suggestions.