Hotel Data Analysis

BADM 557

Goal of the analysis : To predict the cancellation of the booking based on the sample data

Background of the dataset : This data set contains booking information for a city hotel and a resort hotel and includes 32 variables representing 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. For a detailed description of column data, please refer to Appendix section 1.

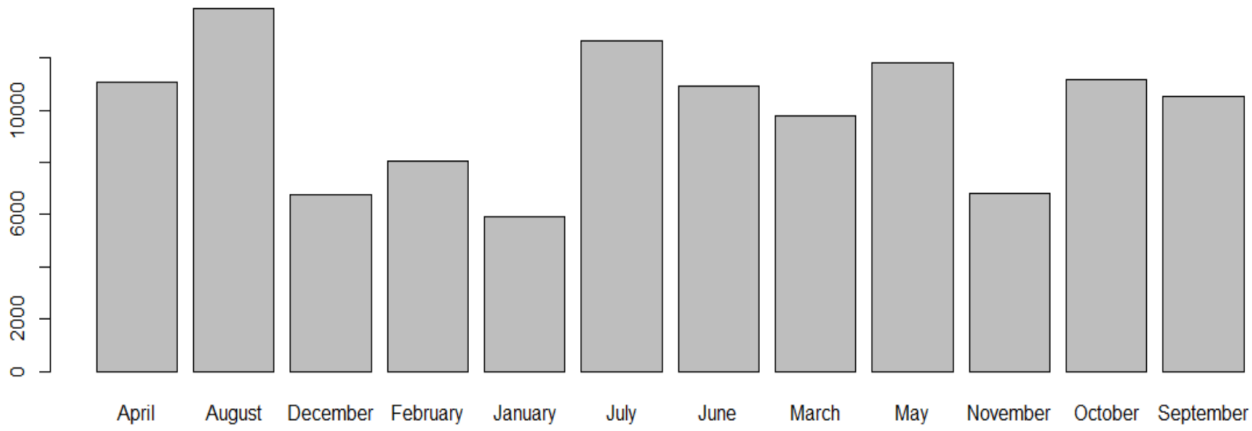
Source URL: https://www.kaggle.com/jessemostipak/hotel-booking-demand

We have further updated the dataset for this project by taking a 1000 rows subset and all personally identifying information has been removed from the original data.

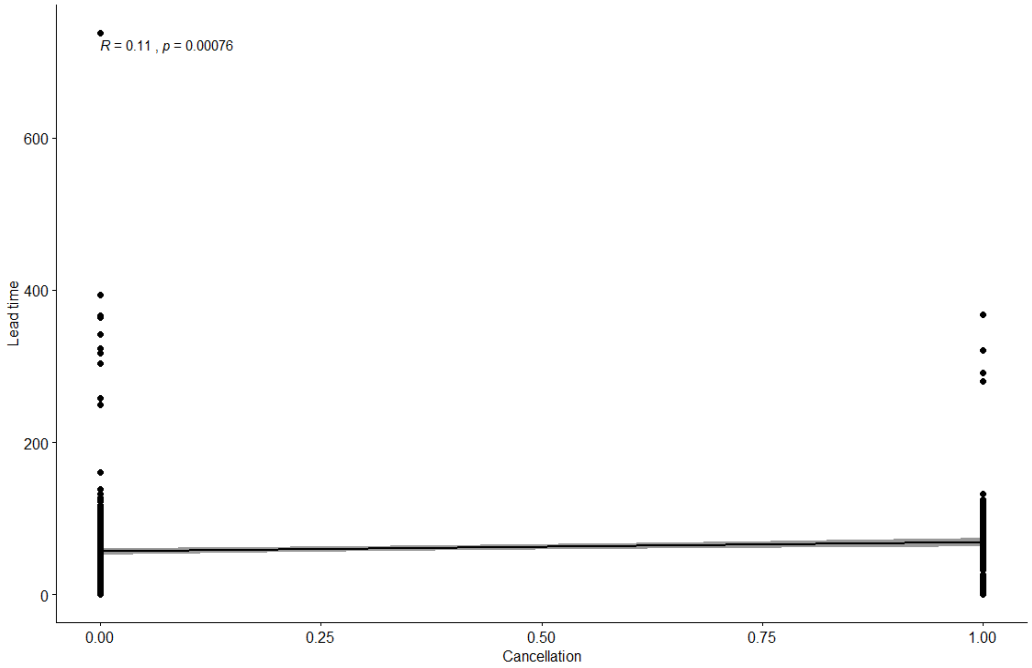
Descriptive statistics and pre-processing :

Average (means) values of variables such as Booking Cancellations, Stays on Weekends and on Weekdays to be 0.351, 1.235 and 2.763 respectively.

- **Month wise bookings :** The plot below shows months on the X-axis and number of bookings on the Y-axis. As shown, August, July, May and June are the months with the highest number of bookings. This trend could possibly indicate the preference of travelers to travel during Summer as compared to the other times of the year.



- **<u>Correlation Analysis :</u>** The plot below shows that there is some correlation between the booking cancellations(X-axis) and lead time(Y-axis). From the plot, it can be inferred that the customers are more likely to cancel bookings if the lead time is more. Similar observations were also seen in the case of Parking Spots indicating the customer may have more of a tendency to cancel a booking if there isn't a parking spot available.

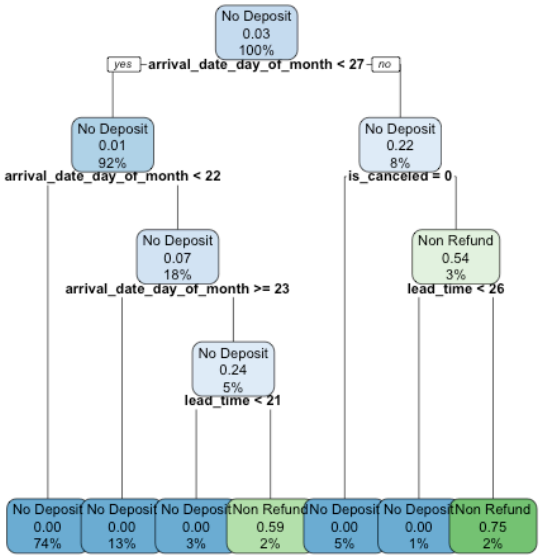


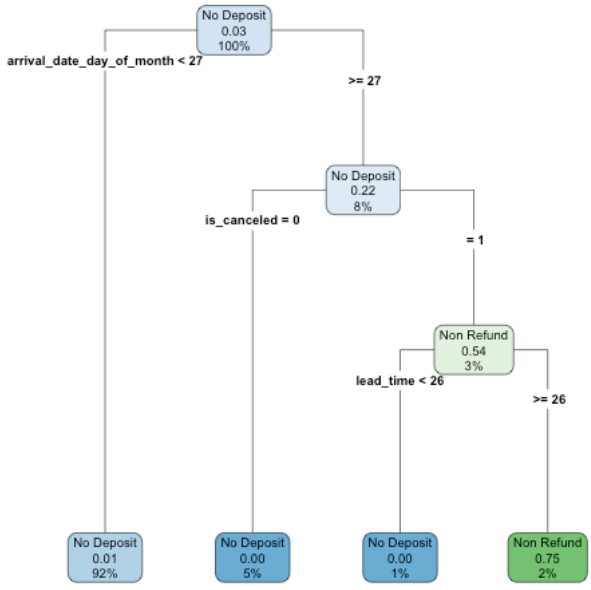
The descriptive statistics was therefore indicative of the underlying trends in the data and gave a view of possible correlations. Further analysis was done for definitive results and prediction.

<u>Analysis Method</u>:

We have used **Supervised learning** method using **Decision Tree** to perform the analysis of the data.
To predict customer's behavior especially during the peak holiday season, we took "Deposit Type" as the analyzing parameter to predict the patterns. They were of two types : No Deposit and Non-Refund.
No Deposit: The customer has made the booking but has not paid any deposit to confirm the booking.
Non-Refund: A deposit was made in the value of the total stay cost.

Below are the non-pruned and pruned decision trees :





It can be inferred from the trees that 92% have not made any deposit when their check in date was less than 27days. The other 8% of the people did not make any deposit because their booking got cancelled.

Confusion Matrix

	No Deposit	Non-Refund
No Deposit	192	1
Non-Refund	0	6

Accuracy=198/199
Sensitivity= 6/7
Specificity=192/192

<u>Implications from the results :</u>

As we can see from the decision tree, the cancellation rate for bookings in the online industry is quite high. Once the reservation has been cancelled, there is almost nothing to be done. This creates discomfort for many institutions and creates a desire to take precautions. Therefore, predicting reservations that can be cancelled and preventing these cancellations will create a surplus value for the institutions.

These prediction models enable hotel managers to mitigate revenue loss derived from booking cancellations and to mitigate the risks associated with overbooking. Booking cancellations models also allow hotel managers to implement less rigid cancellation policies, without increasing uncertainty. This has the potential to translate into more sales, since less rigid cancellation policies generate more bookings. Concurrently, development of these models should contribute to improve hotel revenue management.

Appendix :

1. Columns Description:

hotel : Hotel type (H1 = Resort Hotel or H2 = City Hotel)
is_canceled : Value indicating if the booking was canceled (1) or not (0)
lead_time : Number of days that elapsed between the entering date of the booking into the system 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 weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults : Number of adult guests
children : Number of children
babies : Number of babies
meal : Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country : Country of origin. Categories are represented in the ISO 3155–3:2013 format
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_date : Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when the booking was canceled or when did the customer checked-out of the hotel

2. References:

https://core.ac.uk/download/pdf/233833078.pdf
https://www.researchgate.net/publication/310504011_Predicting_Hotel_Booking_Cancellation_to_Decrease_Uncertainty_and_Increase_Revenue

Code Snippet

```r
View(hotel)
summary(hotel)
library(caret)
colnames(hotel)
newdata = newdata[, c(1:500)]
colnames(newdata)
col= as.factor(newdata$is_canceled)
set.seed(80)
trainIndex <- createDataPartition (newdata$is_canceled, p=0.8, list=FALSE,times=1 )
trainIndex
trainset <- newdata[trainIndex, ]
testset <- newdata[-trainIndex, ]
library(rpart)
library(rpart.plot)
tree <- rpart (is_canceled ~ . , trainset, method="class")
plot(tree)
text(tree)
tree
rpart.plot (tree,type=4,cex=.4)
rpart.plot(tree,type=2,cex=.4)
predict_tree <- predict (tree, testset, type="class")
predict_tree
table (predict_tree)
table (predict_tree,testset$is_canceled)
table (testset$is_canceled,predict_tree)
predict_new<- predict(tree, newdata, type="class")
predict_new
result<- cbind (newdata,predict_new)
View(result)
printcp (tree)
plotcp (tree)
ptree <- prune (tree, cp=tree$cptable[which.min(tree$cptable[,"xerror"]), "CP"])
rpart.plot(ptree,type=4,cex=.6)
predict_ptree<-predict(ptree,testset,type="class")
table (predict_ptree,testset$is_canceled)
predict(ptree,testset[1,],type="class")

```