Section C (Strategy) Team 61:

Zixing Chen, Yaoyao Nai, Jocelyn Ramm, Sanj Saxena, Ankit Chandra Shekar

Business Understanding

Goal: Maximizing Revenues generated by predicting the probability of cancelation. As of the data we have, around 37% of all bookings are canceled.

We are pulling out data to classify customers into those who are likely to cancel and those who are likely to go through with the booking. The high rate of cancelations will result in lower revenue for the hotel. A logistic regression helps us identify customers likely to cancel, we can target them to reduce the likelihood of cancelations, and consequently increase the expected revenue.

Data Understanding

There are 32 variables in the hotel data set, all of which are described in the appendix (Table 1). The key areas of data are: demographic information (eg, number of adults, children, country of origin, etc), booking information, previous booking information, booking methods, booking amenities, cancelations, and average daily price of the booking, etc. (To see preliminary visualizations, see Appendix images 1 - 5)

Using k-means to explore the 32 variables, it is apparent that there are three clear clusters in the data. This is apparent due to the large overlap in the diagrams of the four clusters identified by the k-means model, whereas the two or three k-means clusters are still relatively distinct between the two dimensions. There were key attributes which split the clusters when there were two clusters, these included: cancelation rate, average lead time, and average previous bookings not canceled. It is apparent that the clusters were grouped into bookings with loyal customers, and customers with little or no previous experience with the hotel. This is inferred from the stark

difference between the values of is_cancelled and previous_bookings_not_cancelled. When there are three clusters, the bookings are grouped by those getting discounts/sales on their stays, individuals which are likely to cancel, and bookings with a high average daily rate. When there are four clusters of data, the bookings are grouped into those with high costs (high adr and many special requests), bookings with small lead time (suggesting that the guests booked the trip on a whim), bookings with high or low cancelation rates. As the three clusters remain distinct with little overlap between them, this is the best way to interpret the groups of bookings in the data. (For more details, refer to Appendix for Tables 2 - 4 and images 6 - 8)

Data Preparation

The business problem aims to maximize the expected revenue, to do this, we need to know the revenue and probability of cancelation for the booking. To do this, we need to predict the probability of cancelation by making the is_canceled variable the target variable for the modeling. Building on this, another model will be built to see how the expected revenue varies as the average daily rate is modified. We can proceed with this modeling once our dataset is clean and ready for use.

We identified many areas needing cleaning in the hotel data set. This included, but was not limited to: replacing null entries in various columns, removing columns irrelevant for the data prediction (or compounded from other columns), creating data types into factors, etc. The variables company, arrival_date_year, reservation_status_date, and reservation_status were all removed from the data set. The reservation status contained information which would only be known at the time of check-in (meaning it would hold more information than we would have at the desired time for prediction), so would not be beneficial to have in our prediction model. The company variable was removed due to its high number of missing data (roughly 94% of the

company variable was null). Finally, the arrival_date_year was removed because not much seasonality is captured in the variable. Further, having the arrival_date_year variable being an integer could produce unexpected results, and transforming the variable into a factor would make predictions for future years impossible to determine.

The remaining missing values for each variable were all replaced with 0 regardless of their variable. This made sense for the 4 missing values in the children column, as they were interpreted as scenarios where no children were present. Similarly, both country and agent will be converted into factors instead of characters or integers, meaning the null values having a factor labeled null or 0 is irrelevant.

Factors were made of all remaining variables of the character type, the agent variable (as there is no inherent logic that a higher agent_id necessarily improves or hinders the agent's performance in any way) and all the integers for year, month, week and day of the year. The reason why we are making factors of the dates is because there will be seasonal changes which will have varying impact on the booking and cancelation amount. These changes will be captured by making the variables factors instead of leaving them as integers. Upon further inspection, it is clear that some variables have very small numbers of observations for some factors. These variables include: country, agent, assigned_room_type, and reserved_room_type. In assigned_room_type and reserved_room_type, there are only 1 and 6 instances of room 'L' (this affects 6 rows of data total). In order to prevent overfitting, these 6 rows were removed from the data set (there is no information on the type of room which room 'L' is, so there is no way to group this data with a larger existing room type without potentially compounding effects). Further, 50 agents in the agent variable only booked one booking in the hotel data set, so the column was removed to prevent overfitting for these levels and because there would be no intuitive way to group the

agents into larger levels. For the country variable, there were many factors with only one booking originating in that country. To tackle this issue, we created a new variable called continent. This variable contained the information of the continent of origin rather than the specific country. Consequently, the data was grouped into bins of minimum size (Antarctica and French Southern Territories are in the continent of Antarctica, these were left as "null" values alongside the null country values, this is due to their unique geographical location with few related data points). After this variable was created, the country variable was dropped from the data set.

An additional dummy variable has been added: matching_room. This variable will have value 1 if reserved_room_type = assigned_room_type, else 0. This is to help capture the impact of an individual not receiving the room which they booked. Through further investigations, we realized this interaction is significant as the probability an individual cancels after their assigned room is different to their reserved room (roughly 5.4%) is lower than when an individual is assigned the same room as their reservation (roughly 41.6%).

As a result of preparing the hotel data set for the modeling, we have 28 predictor variables and one target variable. There are also no null values in the table. All required variables are transformed into factors ready for the modeling.

Modeling

To prevent overfitting and to analyze the out-of-sample accuracy of the models, the data set was split into training data ("train") and testing data ("test"). The training data contained 80% (roughly 96000 bookings) of the total data and the testing data contained the remaining 20% (roughly 24000 bookings). These subsets of data are chosen at random, to minimize any bias within the data across all variables of data. This was cross-checked by ensuring that the average

cancellation rate of both sections were roughly equal. Consequently, when the models are run, the training set ensures the model has a good fit, and the testing set is "unknown" data to the model, so we are able to accurately judge the out-of-sample accuracy of the model. A variety of models will be analyzed and compared by their out-of-sample accuracies.

The business problem requires a prediction of a classification variable (is_canceled), which restricts the types of models that can be used. The possible models include: logistic regression, CART, and random forest. Due to the large number of variables and our limited computation power, random forest was eliminated as an option.

Firstly, both CART and a full logistic regression model with all variables were run to explore the data and analyze the out-of-sample performance of the models. It was apparent, but not statistically significant at the 5% significance level, that the full logistic regression model had a better performance when faced with the "unknown" test data set (all metrics to judge performance can be found in the next section). Consequently, stepwise elimination was conducted with different information criterion (AIC and BIC) to further refine the logistic regression model to prevent overfitting and ensure simplicity.

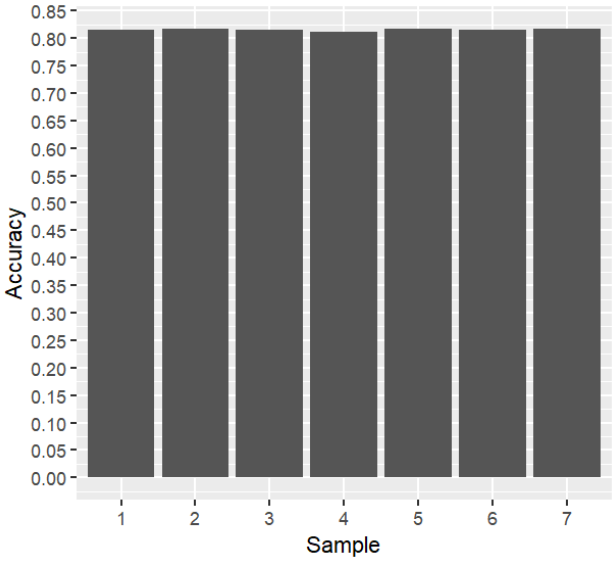
Evaluation

In order to evaluate the models and choose the best one, we need to remember our initial business objectives that we had in mind while formulating these models: To accurately predict cancellation rates across customers. While we can keep an eye on In-Sample Accuracy, it is essential to look at the Out-of-Sample Accuracy to determine which model is the best fit. First, we need to look at a baseline null model to compare the other models against. Using an initial logistic regression model, this null model was fitted to the training set to decipher the null out-of-sample accuracy of predicting the cancellation of bookings. Following this, the other

models were used to analyze the out-of-sample accuracy in order to ensure the best fit. The accuracy of various models are summarized below:

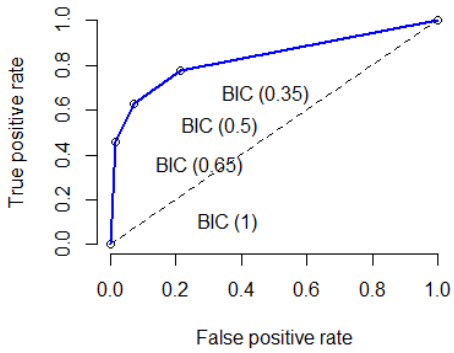
Model	In-Sample Accuracy	Out-of-Sample Accuracy	OOS Accuracy Confidence Interval	Number of Coefficients
Null	0.6294	0.6287	(0.6227, 0.6346)	1
CART (Image 11)	0.8076	0.8059	(0.8010, 0.8107)	6 branches
Full Logistic Regression Model	0.8147	0.8123	(0.8074, 0.8170)	152
AIC Selected Logistic Regression Model (Formula 1)	0.8147	0.8122	(0.8074, 0.8170)	151
BIC Selected Logistic Regression Model (Formula 2)	0.8145	0.8148	(0.8100, 0.8194)	60

As seen in the table, it is clear that the full logistic regression model and the BIC selected logistic regression model obtained the highest out-of-sample accuracy. In order to minimize overfitting, the BIC selected logistic regression model will be used due to its balance of simplicity and high out-of-sample accuracy. This graph depicts the out-of sample accuracy across various random samples of the hotel data:



It is clear that the variance between the samples is small, which implies that the model is not overfitting the data. The out-of-sample accuracies range from 0.8122 to 0.8164. Further, all predictions and tests for

accuracy have been made under the assumption that individuals who have a probability of canceling greater than 0.5 will cancel. Through analyzing the true positive and false positive rates for different cut-off values, a depiction was made to assist in choosing the ideal cut-off for the cancelation predictions. It is clear that the threshold of 0.5 balances the true negative rate and false negative rates best. The firm should decide whether they wish to penalize type I errors (predicting the booking will cancel when, in reality it doesn't) or type II errors (predicting the booking will not cancel when, in reality it does) more. After this decision is made, various other threshold levels can be chosen to adapt to their judgements (a couple example cut-off points are provided).



Deployment

Through analysis of this logistic regression model, five key variables were identified to have the largest impact on the probability of cancelation. These variables were: lead_time, matching_room, previous_cancellations, total_of_special_requests, and deposit_type.

Variable	Coefficient in the Model	Relative Importance
total_of_special_requests	-7.169e-01	54.61
deposit_typeNon Refund	5.423e+00	41.76
previous_cancellations	2.799e+00	39.70
matching_room	1.761e+00	35.34

lead_time	3.970e-03	35.11

Through this analysis, hotels will be able to modify the cost of the booking or change various features of the booking (eg. increasing or requiring a deposit, providing a discount, etc.).

Ways to influence key metrics to reduce cancelations:

There should be a focus on increasing the number of special requests for the booking. This could be done by focusing more on bookings which center around a special occasion. It would be expected that individuals celebrating would create more special requests, such as asking for a bottle of champagne to celebrate a birthday, or rose petals to greet a newly-married couple. Hotels can target these guests through advertising their hotel on specific chat rooms online. It is likely that word-of-mouth would also be highly beneficial in spreading these recommendations to other groups.

The data suggests that bookings with non-refundable deposits increase the probability a booking is canceled. Using intuition, it would be assumed that introducing a non-refundable deposit would increase an individual's barrier to cancel the booking, and consequently decrease the probability of canceling. It could be possible that this variable has compounding effects with variables outside of those in the model or data set. Alternatively, it is possible the deposit amount is not large enough to create a sense of hesitation while canceling the booking, and this amount needs to be increased.

It is clear that an increase in previous cancellations would make the current booking more likely to be canceled. To reduce this, suggestions would include increasing their deposit amount and creating a larger cancelation charge. Again, this will increase the barriers to cancel the booking, because the hotel guest will view the booking as more valuable.

In order to reduce cancelations further, hotels can increase the number of bookings which have the reserved room different from the assigned room. This would likely be due to unexpected upgrades to the booking, which can increase customer satisfaction and clearly reduces the cancelations. In order for hotels to maximize this quality, they should upgrade bookings where possible, and it is profitable.

Finally, the data suggests that a larger lead time (that is a larger number of days between the booking date and the check-in date) results in a higher probability of cancelation. Reasons for this could include: individuals have forgotten about their booking and do not show up, the guests do not have their PTO approved, or there was a change in plans or an emergency. The hotel has little control over most of these reasons for cancelations, however the hotel can minimize the probability that the individuals forgot about their booking and are no-shows. This can be done by sending reminders to the hotel guests at intervals between the booking date and the check-in date. A consequence of this could be that the booking is canceled earlier than without the reminder messages, as the individuals may cancel before the check-in date where they otherwise would have been a no-show. These messages could be sent by email, text, mail, or phone call. Further, we can limit this possibility by opening bookings to a maximum of 2-3 months in advance. This would help open up slots to individuals who are indeed certain about their plans and are less likely to cancel their reservations.

<u>Using overbooking to counteract loss of revenue by cancelations:</u>

In order for overbooking to be effective, the hotel would need to evaluate its utilities for various scenarios ([canceled, predicted canceled], [not canceled, predicted canceled], [not canceled, predicted not canceled], [canceled, predicted not canceled]). Through the information gained by these determined utilities, the expected utility ($E(\text{utility}) = \sum(\text{utility}_i * P(i))$, where i is false

positive, false negative, true positive, and true negative) would be able to be maximized using the logistic regression model with BIC selection described in the modeling section. This shows the revenue (higher revenue due to greater effective bookings) - risk(lower customer satisfaction due to cancellation of their booking from the hotel's end) tradeoff that comes with overbooking. The extent of overbooking and the required guard rails we must establish is something that the firm needs to decide in order to proceed with the strategy.

Appendix

Contributions of Each Team Member

Business		
	Understanding	Everyone
	Key Issue	Everyone
	Solution	Everyone
	Deployment of Solution	Yaoyao Nai, Sanj Saxena
Dataset		
	Finding	Ankit Chandra Shekar
	Cleaning	Everyone
	Visualizations	Zixing Chen, Jocelyn Ramm, Ankit Chandra Shekar
	Exploration (PCA, K-means)	Zixing Chen, Sanj Saxena
	Modeling	Everyone
	Testing (overfit)	Jocelyn Ramm, Yaoyao Nai
Presentation		
	Content + Design	Ankit Chandra Shekar, Sanj Saxena
Document	Content + Completion	Jocelyn Ramm, Zixing Chen, Yaoyao Nai

Table 1: Description of Variables and their missing values

Variable Name	Description	Number of Missing Values
hotel	Character - type of hotel	0
is_canceled	Integer - whether booking is canceled (1) or not (0)	0

lead_time	Integer - number of days elapsed between the entering of the booking and the arrival date	0
arrival_date_year	Integer - year of arrival	0
arrival_date_month	Integer - month of arrival date	0
arrival_date_week_number	Integer - week number of year for arrival date	0
arrival_date_day_of_month	Integer - day of arrival date	0
stays_in_weekend_nights	Integer - number of weekend nights that guests stayed or booked to stay	0
stays_in_week_nights	Integer - number of week nights (Mon-Fri) that guests stayed or booked to stay	0
adults	Integer - number of adults	0
children	Integer - number of children	4
babies	Integer - number of babies	0
meal	Character - meal packages (BB - bed and breakfast, HB - half board, FB - full board)	0
country	Character - country of origin	488
market_segment	Character - market segment designation (TA - travel agent, TO - tour operators)	0
distribution_channel	Character - booking distribution channel (TA - travel agent, TO - tour operators)	0
is_repeated_guest	Integer - value indicating if the booking name is a repeated guest (1) or not (0)	0
previous_cancellations	Integer - number of previous canceled bookings by the customer prior to current booking	0
previous_bookings_not_canceled	Integer - number of previous bookings not canceled by the customer prior to current booking	0
reserved_room_type	Character - code for type of room reserved	0
assigned_room_type	Character - code for type of room assigned	0
booking_changes	Integer - number of changes or amendments made to the booking from the moment the	0

	booking was entered	
deposit_type	Character - indication if the customer made a deposit to guarantee booking	0
agent	Integer - ID of the travel agency that made the booking	16.3k
company	Integer - ID of the company/entity that made the booking or responsible for paying the booking.	113k
days_in_waiting_list	Integer - number of days the booking was in the waiting list before it was confirmed to the customer	0
customer_type	Character - type of booking (contract, group, transient, transient parking)	0
adr	Integer - Average Daily Rate	0
required_car_parking_spaces	Integer - number of car parking spaces required by the customer	0
total_of_special_requests	Integer - number of special requests made by the customer	0
reservation_status	Character - 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	0
reservation_status_date	Date - Date at which the last status was set	0

Table 2: Main features of the k-means clusters (with 2 clusters)

Variable	Cluster 1	Cluster 2
is_canceled	41.05%	17.36%
lead_time	115.55 days	47.38 days
previous_bookings_not_canceled	0.0179	0.7277

Table 3: Main features of the k-means clusters (with 3 clusters)

Variable	Cluster 1	Cluster2	Cluster 3
is_canceled	36.23%	47.60%	15.96%
days_in_waiting_list	6 days	0.08 days	0.001 days
lead_time	161.32 days	40.81 days	82.20 days
total_of_special_requests	0.163	0.48	0.906
adr	83.13	97.44	116.66

Table 4: Main features of the k-means clusters (with 4 clusters)

Variable	Cluster 1	Cluster2	Cluster 3	Cluster 4
is_canceled	47.85%	15.99%	30.85%	36.06%
lead_time	161.98 days	82.45 days	40.22 days	69.07 days
adr	83.16	115.08	95.08	171.73
total_of_special_requests	0.159	0.908	0.482	0.7317

Visualizations of the Hotel Data Set:

Image 1: How the average daily rate of the bookings varies by month of year

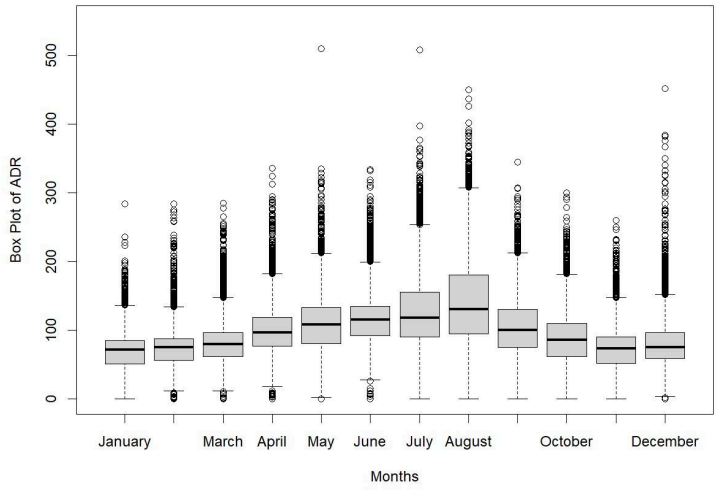


Image 2: How the cancelation rate varies as lead time increases

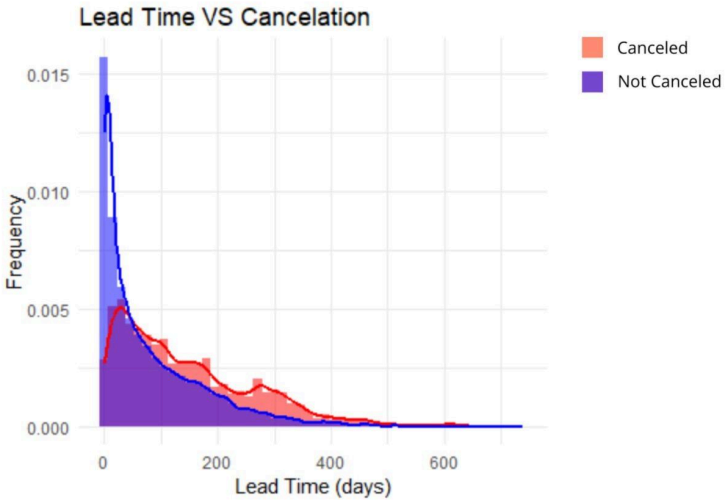
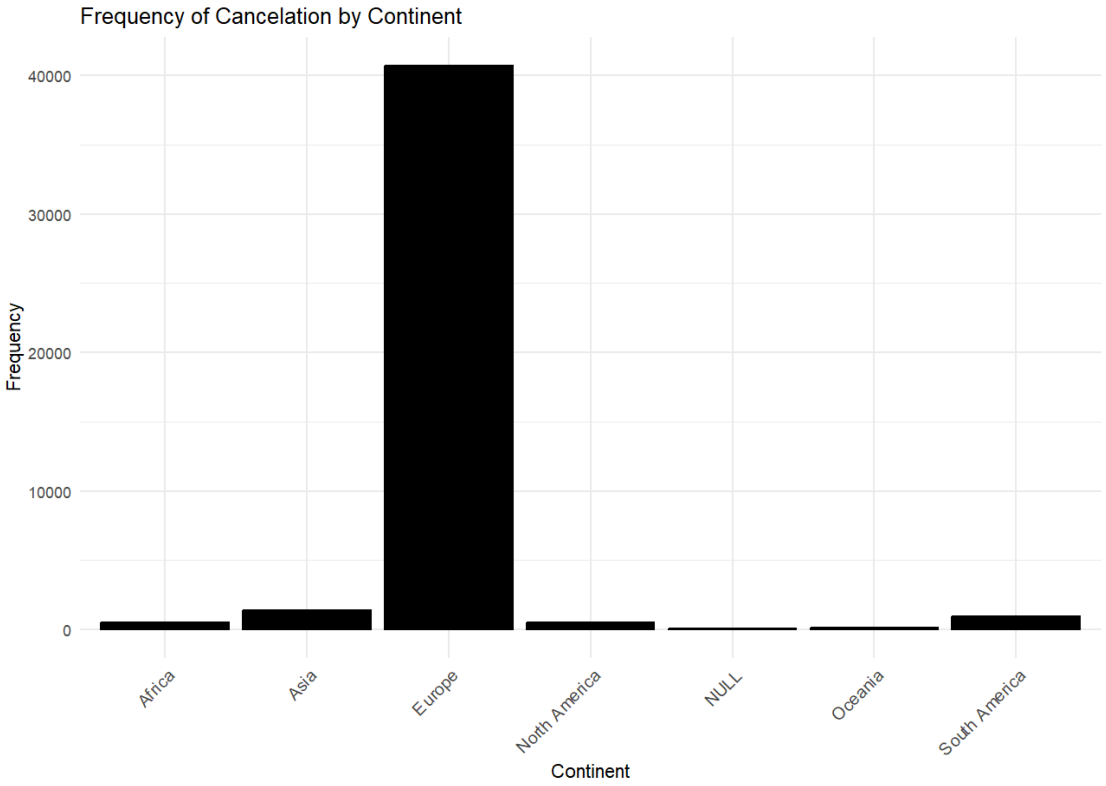


Image 3: How the cancelation rate varies as the total number of special requests change

Image 4: How the cancelation rate varies based on hotel type

Image 5: How the cancelation rate varies across the continents



Images 6 - 8: Depictions of K-means for the data

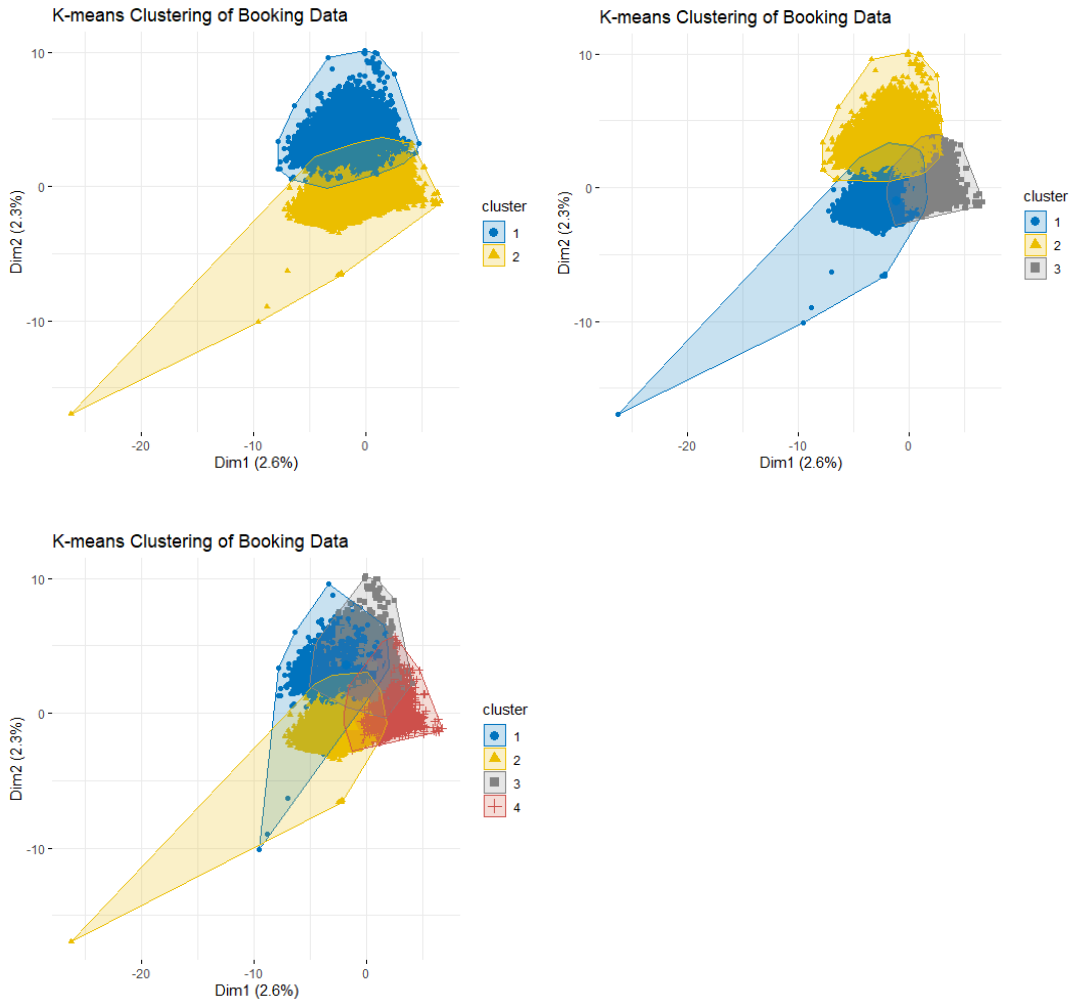


Image 9: PCA Analysis showing variance explained by various factors

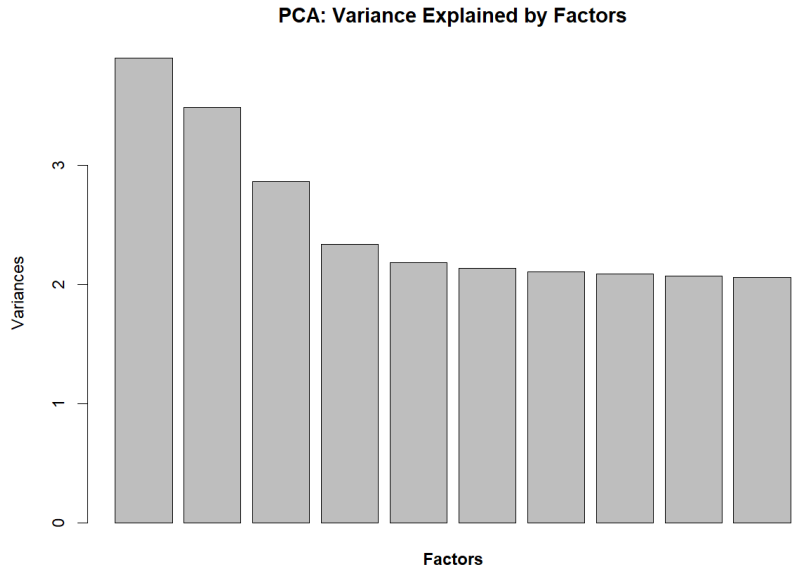


Image 10: How the cancelations vary by high, medium, and low-risk groups (risk thresholds decided by the three k-means values)

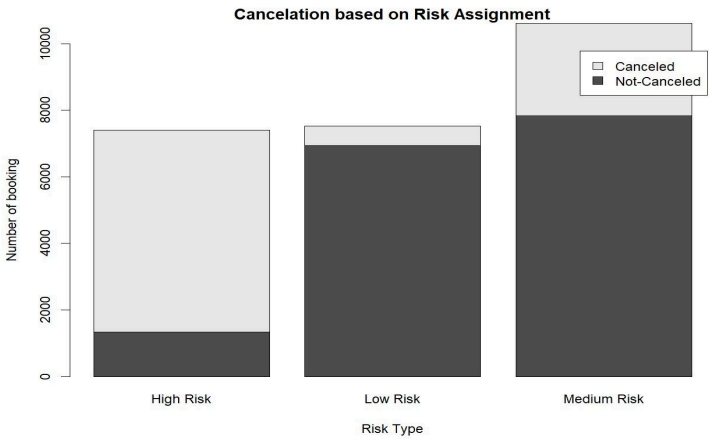
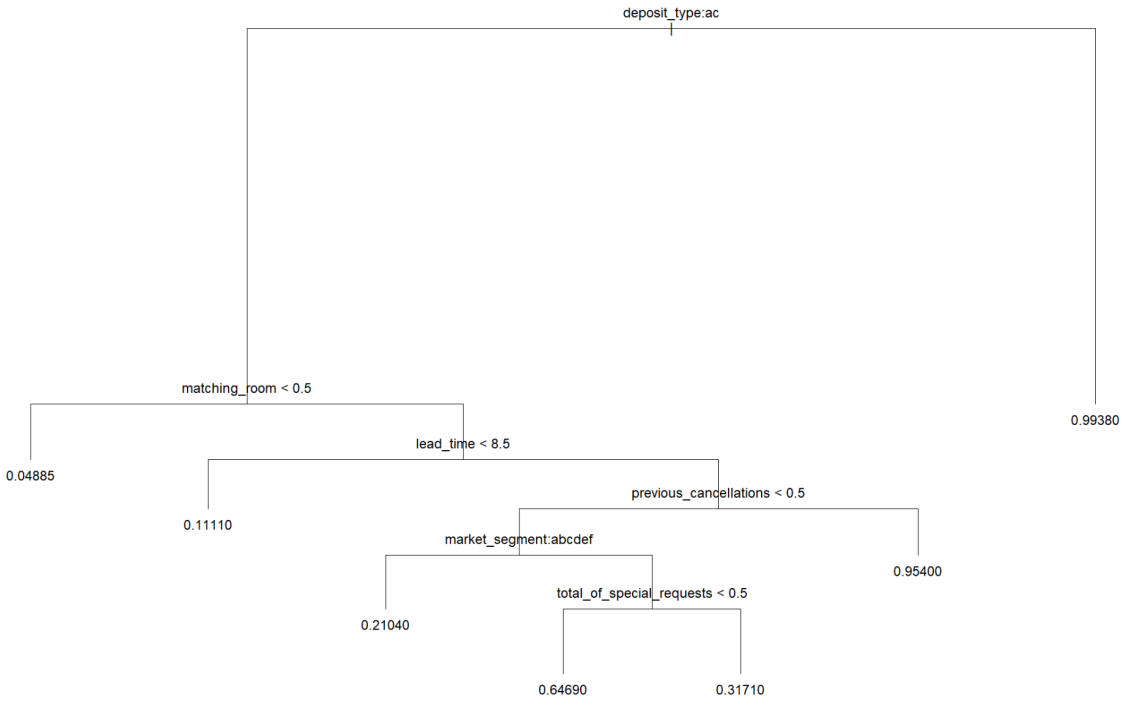


Image 11: CART model



Formula 1: BIC

is_canceled ~ hotel + lead_time + arrival_date_month + arrival_date_week_number + arrival_date_day_of_month + stays_in_weekend_nights + stays_in_week_nights + adults + children + babies + meal + market_segment + distribution_channel + is_repeated_guest + previous_cancellations + previous_bookings_not_canceled + reserved_room_type + assigned_room_type + booking_changes + deposit_type + customer_type + adr + required_car_parking_spaces + total_of_special_requests + matching_room + continent

Formula 2: AIC

is_canceled ~ hotel + lead_time + arrival_date_month + stays_in_week_nights + adults + children + meal + market_segment + distribution_channel + is_repeated_guest + previous_cancellations + previous_bookings_not_canceled + assigned_room_type + booking_changes + deposit_type + customer_type + adr + required_car_parking_spaces + total_of_special_requests + matching_room + continent