Predicting booking cancellations for hotels

Group 2

Introduction

Our project aims to provide valuable insights on hotel cancellations using analytics. In order to systematically address the needs of the project, we have decided to adopt the "Cross-Industries Standard Process for Data Mining" (CRISP-DM), which will form the structure of our report. We found this useful since this process emphasizes the reiterative process of improving and updating each section based on new insights made as we progress and make new discoveries about the data and business problem.

Business Understanding

Understanding hotel cancellations

Booking cancellation is a critical challenge faced by the hospitality industry because it has a direct impact on their revenue generating potential. When customers cancel their bookings, there are serious implications on the hotel because it not only affects their occupancy rates, but also any demand forecasting activities and budgeting that accounted for the canceled bookings.

Revenue management strategies such as dynamic pricing, overbooking and strict cancellation policies are employed in an effort to address booking cancellations and maximize occupancy rates. However, when done based on intuition only, these strategies might potentially backfire and result in negative consequences, such as loss sales, deteriorated business reputation and fall in customer loyalty. Therefore, it is of the industry's interest to more accurately predict booking cancellations and employ appropriate strategies based on data analysis.

Goal of our analysis

Using the four types of analytics (descriptive, diagnostic, prescriptive, predictive), we aim to provide insights on what affects booking cancellations and provide recommendations on how to improve cancellation rates. Following which, we will build a model to predict canceled bookings. The goal of our analysis is to answer 2 main questions:

- Is booking cancellations related to factors like lead time and market segment?
- Can we predict with reasonable accuracy if a booking will be canceled based on its attributes?

Data Understanding and Preparation

The dataset was extracted from Kaggle and can be found here

The dataset is accompanied by documentation of the data, which can be found here

```
## Loading required package: ggplot2
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: reticulate

## Loading required package: readr

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: ggcorrplot

## Loading required package: rpart

## Loading required package: rpart.plot

## Loading required package: caret

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Loading required package: e1071

## Loading required package: ROCR

## Loading required package: plyr

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following object is masked from 'package:purrr':
##
##     compact

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## Loading required package: randomForest

## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## Loading required package: caTools
```

Missing data

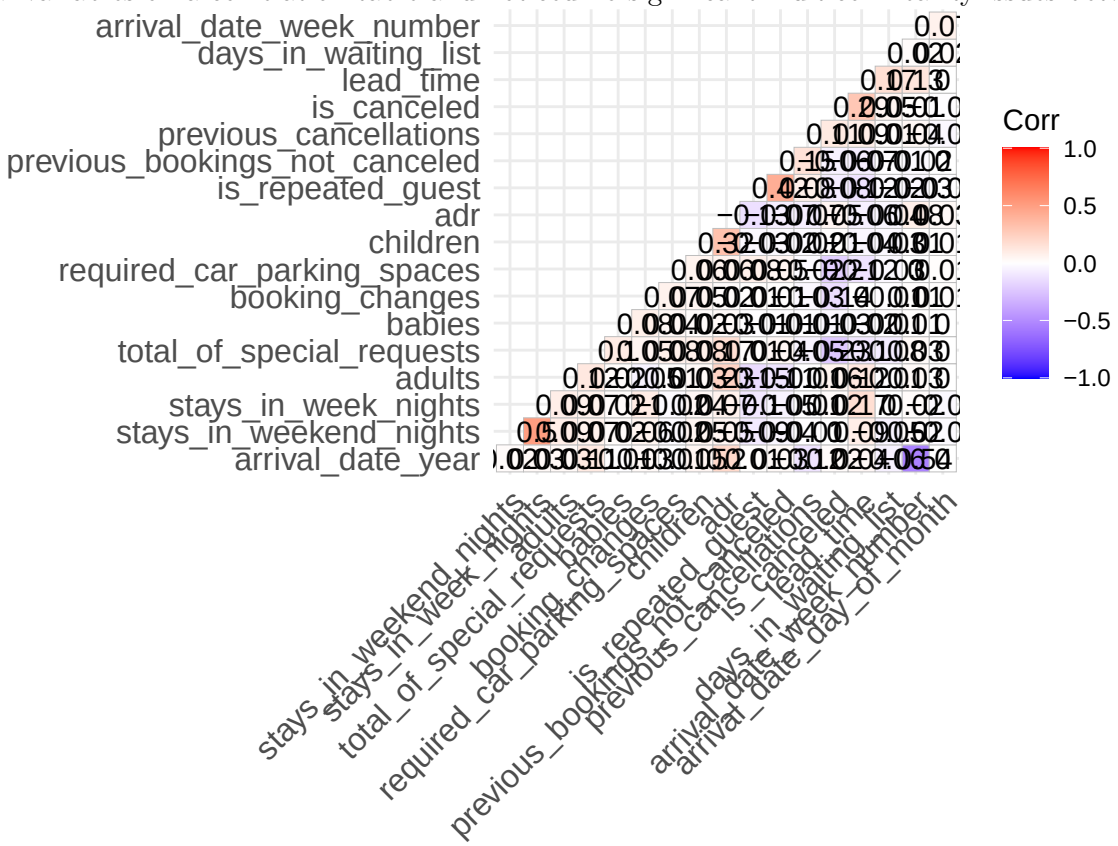
We checked the dataset for empty columns and found that only the children column has 4 empty cells that are "N.A" cells.

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

To address this, we replaced the missing values in the "Children" column from the corresponding Babies column.

Correlation table

We plot all the numerical variables on a correlation table and noticed no significant multicollinearity issues between



the numerical variables.

Number of booking cancellations in City vs. Resort Hotel

From the table above, we know that majority of the data comes from the city hotel (66% of data), with about twice the amount of data as compared to the resort hotel (33%). Knowing this, we might consider performing a separate analysis for city hotel and resort hotel or our analysis might be skewed towards the city hotel.

From this plot we see that we have more data points for cancellations in the City Hotel as compared to the Resort hotel as well. Again, it suggests that performing a separate analysis for both hotels might be appropriate.

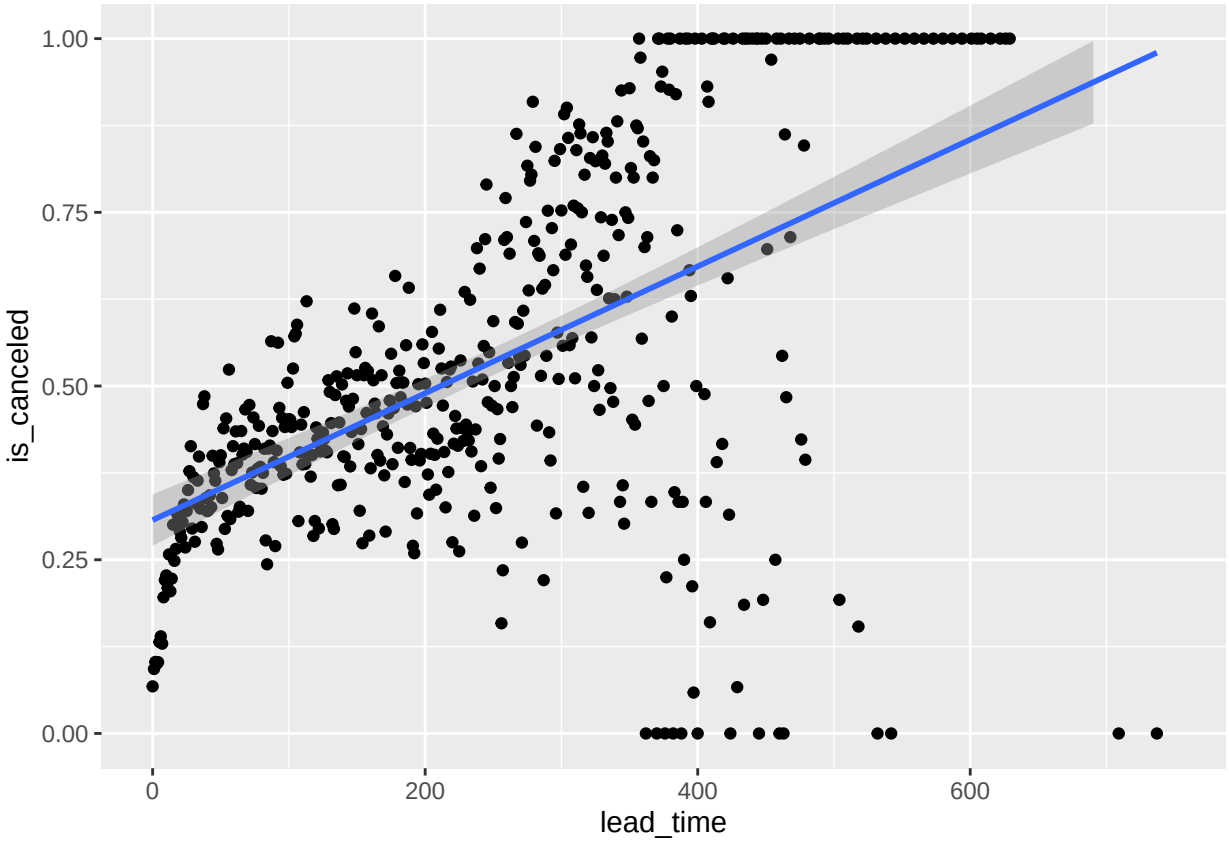
Descriptive Analytics

We dive deeper to identify trends, patterns or anomalies in booking cancellations. This section aims to highlight **two key discoveries** from our data exploration and their implications which guide the way we proceed with data analysis and building of the models.

Trends in booking cancellations

Discovery 1: Potential correlation between lead time and cancellations

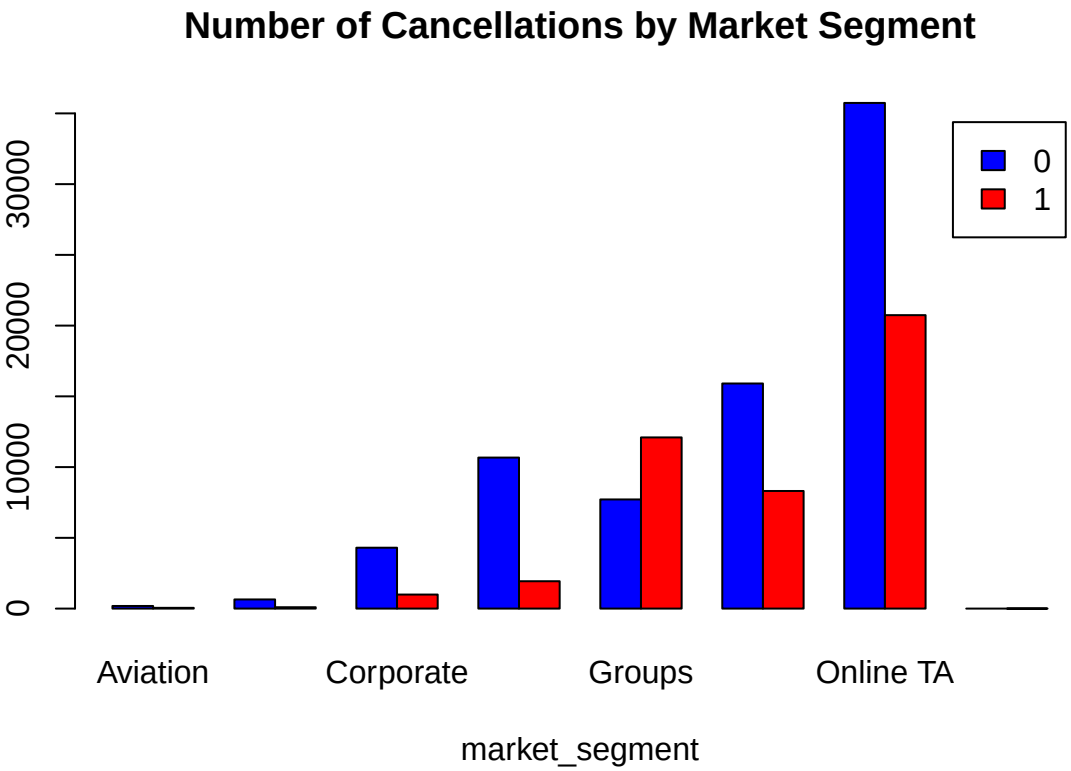
```
## `geom_smooth()` using formula 'y ~ x'
```



Apart from the anomalies where lead time is more than 750 days, the scatter plot revealed that as lead time increases, there is a higher probability of bookings being canceled.

This suggests there might be a **potential correlation** between lead time and canceled bookings. We might be able to verify this when doing further analysis later.

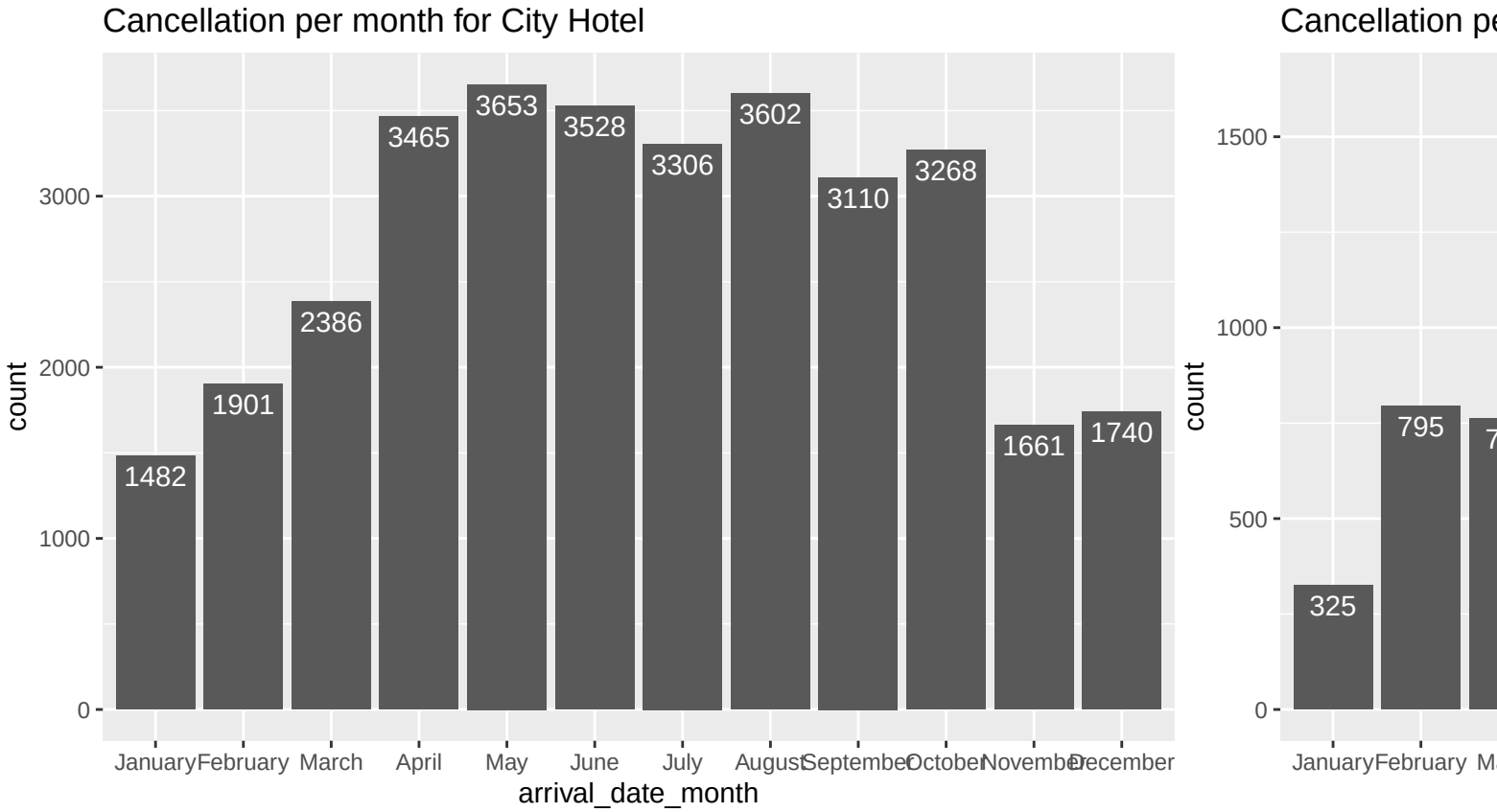
Discovery 2: The "Groups" segment is the segment with the highest % of cancellations



Other insightful discoveries to contextualise data

As mentioned earlier, we found that there is a significant portion of our hotel bookings that are from City Hotel as compared to Resort Hotel. To confirm if we should just focus our analysis on one of these segments as the models might differ, we checked if there was any varying trends between these two sub-groups. One such example we identified is **seasonality of cancellations**.

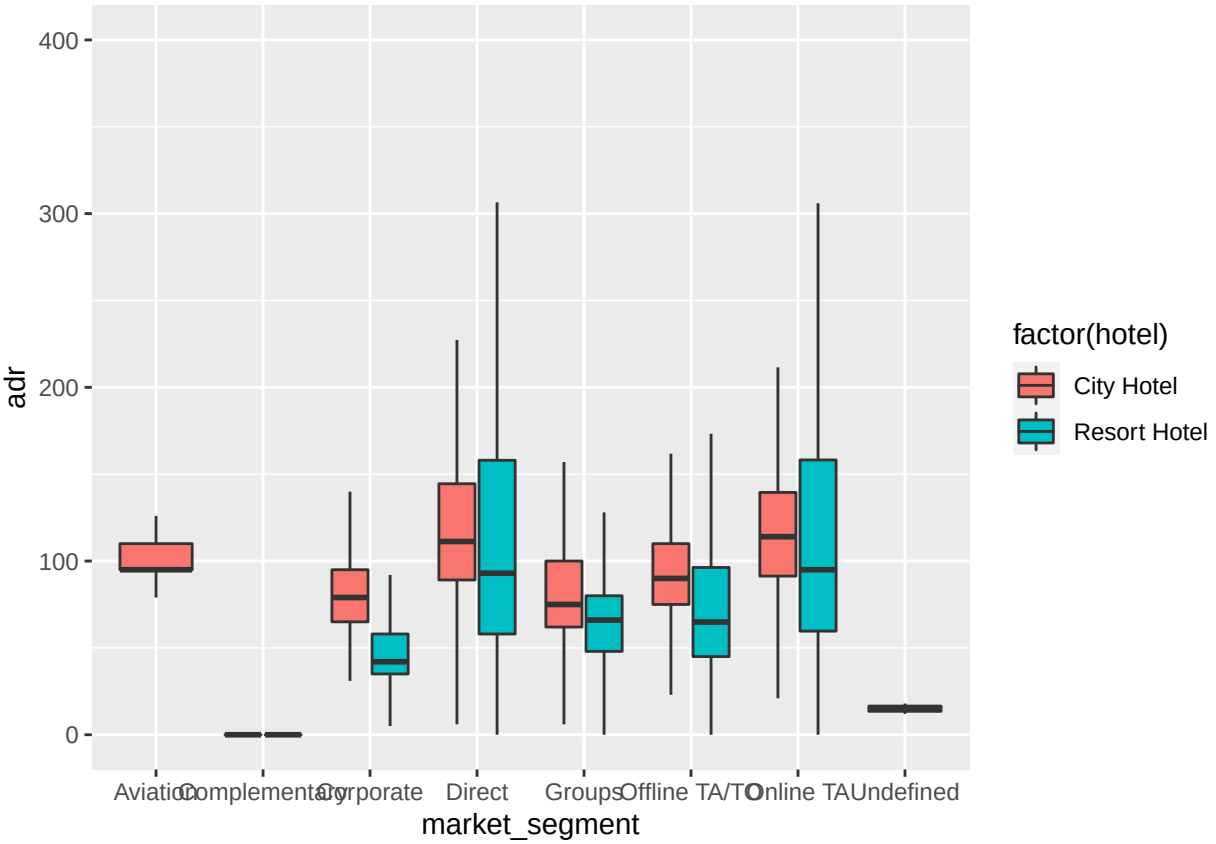
Seasonality of cancellations



We know that for **City Hotel**, customers tend to cancel more from the period of **April to October**. For **Resort Hotel**, **July and August* seem to have the most significant number of cancellations.

In light of this, we know that there might be some form of **seasonality** when it comes to booking cancellations because of the seasonality of hotel bookings, where customers tend to book holidays during the holiday seasons. It might be more effective to perform a separate analysis on City hotels and build a predictive model solely on City hotel data.

Hotel room pricing variation by market segment



Average Daily Rate (ADR) is defined by the average revenue generated from a single night's stay. We can compare the average ADR of each market segment to see which segments have the most price variation.

There is an anomaly from offline TA/TO where the ADR is >4000 for city hotels. We know this could potentially be a one-off high price that was charged, or perhaps an error in data entry. Because prices are clustered mostly around the $400 range, we used that as an upper bound.

Based on the vertical box plot, it can be inferred that the **Direct** and **Online TA** segments have the highest price variation. A possible explanation is that dynamic pricing is more common in these segments.

On the other hand, we know that **Aviation** might be the least flexible in negotiating prices because of its low price variation. A possible explanation for this is that this segment, possibly airline companies, usually have fixed arrangements with the hotels to ensure predictable prices.

#Predictive Analytics We found CART and Random Forest most relevant in this context and also produces the best results for predicting booking cancellations.

Data Preparation for models

We need to first process and remove some of the variables before building our prediction model. The reason are as follows.

Firstly, as mentioned in the section above, we decided to experiment with building a predictive model based on **City Hotel** data only, since there might be significant variation for Resort Hotel. Secondly, variables with more than 53 levels (country, company and agent) are removed due to the restrictions of R, also the models take too long to run when these variables are included. Thirdly, we removed variables (reservation_status and reservation_status_date). They cannot be used for prediction because they are directly related to booking cancellations. Including these variables will result in an AUC = 1.0, and the model won't be useful.

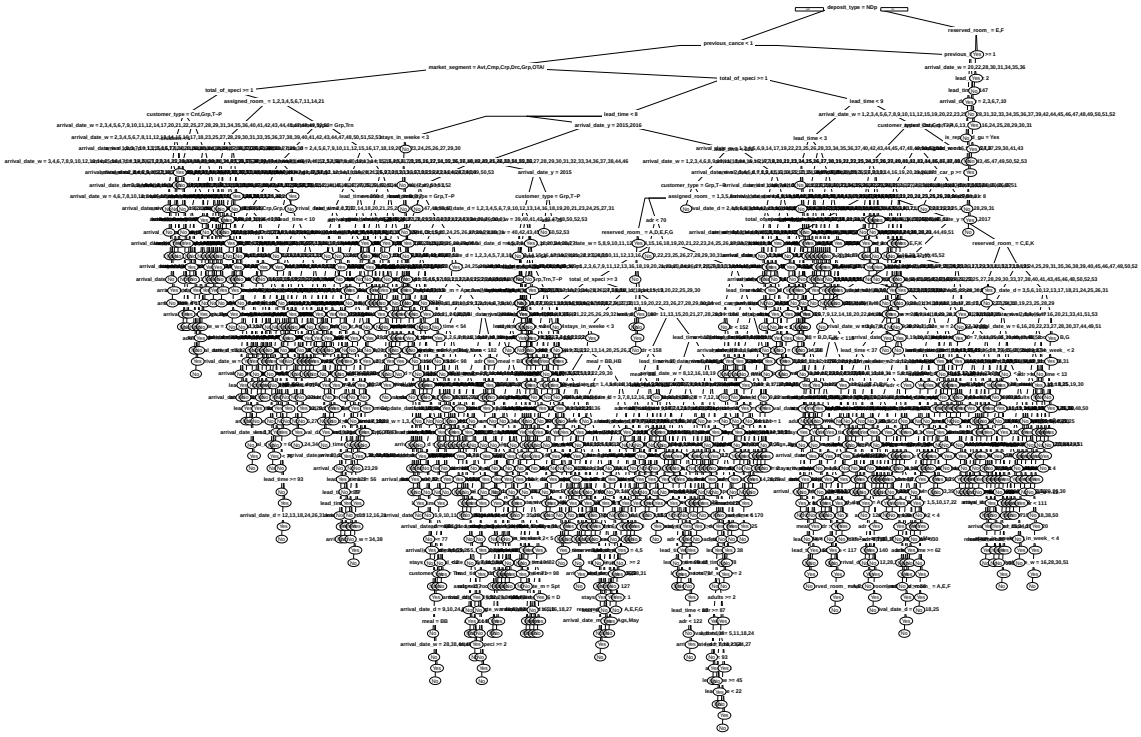
We split this new data set into a training set and a test set, so we can have a test set to check the AUC of the model and evaluate its performance.

Model 1: CART Model

We began by finding the optimal cp value using cross validation where k=5. 5 Seems to be an appropriate value since our dataset has around 100K rows, and a 1:5 ratio is around 80% to 20% training-test split. The optimal cp value was found to be 3eˆ-05.

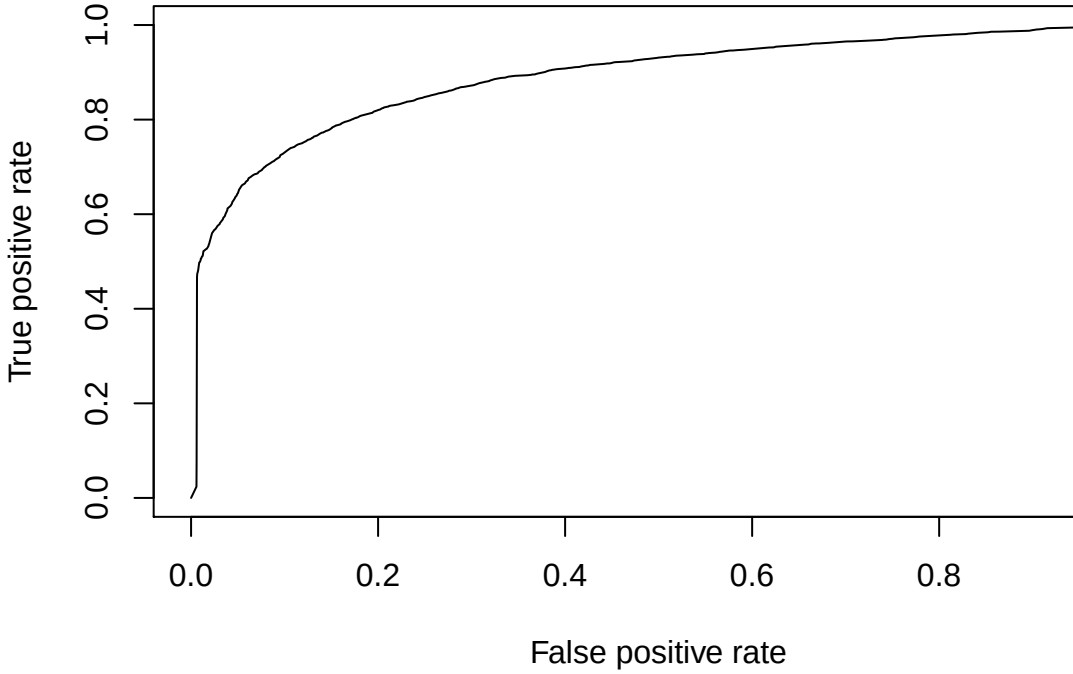
We assigned the optimal cp value of 3eˆ-05 to train the classification tree.

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



Testing the CART model with test data

Predicting Probability outcomes of CART Model



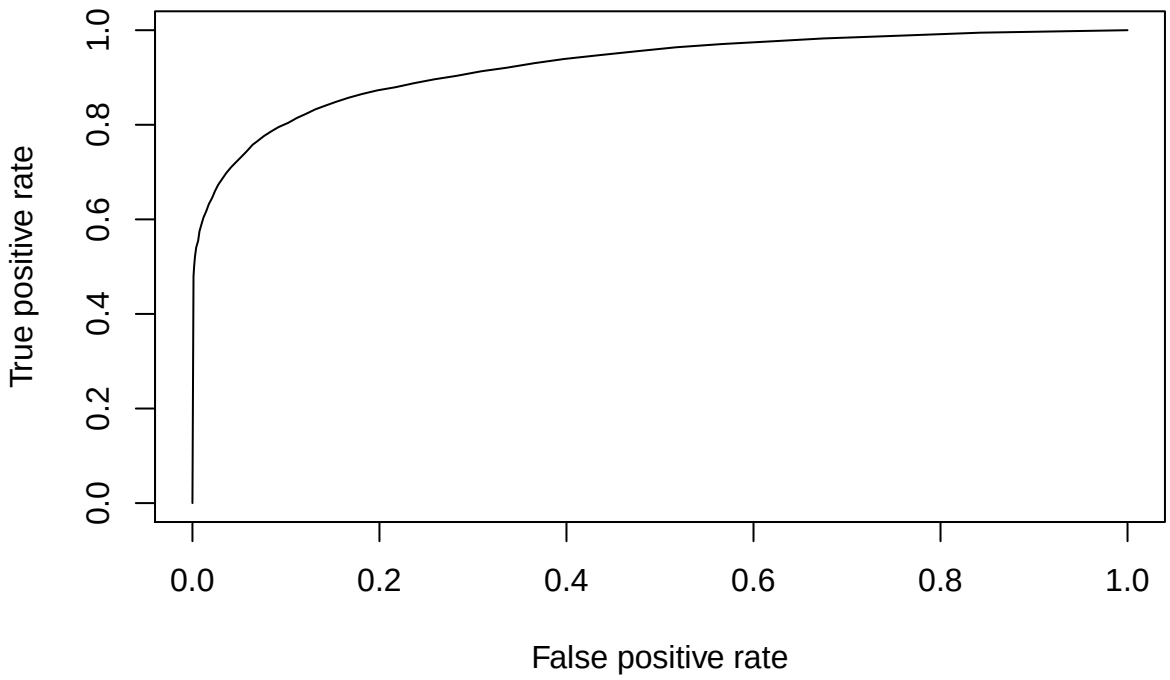
Plotting the ROC for the CART model

Deriving the AUC for the CART model

Our CART model returned an AUC of **88.8%**, based on just City hotel data alone.

##Model 2: Random Forest We built a Random Forest with our training data with all variables. We experimented with tuning the parameters ntree, nodesize and mtry, and derived these values to return the optimal results.

Predicting using HotelForest on the test data to check ROC and AUC



AUC of our Random Forest model

Our Random Forest model returned an AUC of **92.5%**, based on just City hotel data alone.

Discussion of Model Results

From the CART diagram, we can infer that there are some variables that are more important than others, such as deposit type, previous cancellations and reserved room. This is inferred based on the CART tree, which returns a decision tree with the aforementioned variables seen in the top 2 levels.

Our random forest model returns a relatively high AUC of **92.5%**, so we know our model can predict city hotel cancellations relatively well, and will be the model that will be used by the hotel to predict cancellations in the future.

Also, in proving our initial hypothesis of having different models for City Hotel and Resort hotel due to its difference in demand, we ran the same model to include both City and Resort data, which returned similar AUC of ~92%. Therefore, the hotel could potentially use all rows of data to build a complete model and the segregation is not necessary.

We will further discuss the implications of the model's results in the conclusion section later in the report.

#Diagnostic Analytics

Based on the prior data analysis performed, we narrowed down the top 2 factors we believe lead to the most cancellations based on data for City hotels. These factors are lead_time and market_segment, based on our descriptive analysis. We also want to test the deposit_type variable in our diagnostic analysis, since the decision tree in our CART model suggests that this is a significant variable.

In this section, we test these factors for causality between the aforementioned variables and cancellation rates.

##Lead Time

As lead time is a continuous variable, we created a new column lead_bin which is a binary variable. This column is created by assigning a value of 1 if lead_time is more than 100. This way, we can see if we keep all control variables constant, does lead time have an effect on the cancellation rate. Based on the analysis above, we found that at lead time = 100 is when the rate of cancellations start to decrease which is why we use lead_time > 100 as the binary constraint as a starting point.

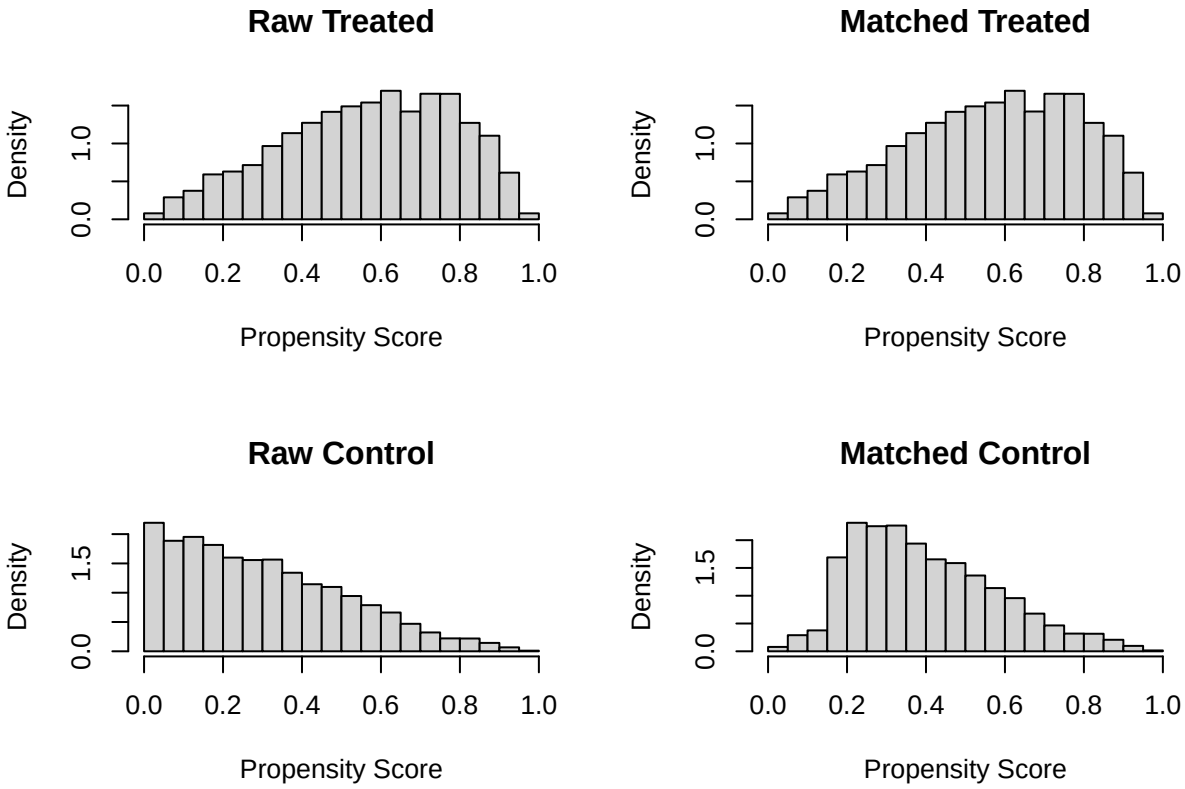
To ensure that the number of data points above 100 is significant enough, we check for how many bookings take place with a lead time of more than 100.

Indeed it is significant enough, since there is about a 50:50 split. We further analyze the cancellation rate for bookings with lead time less and more than 100 for the unmatched data.

We can see from this that the mean cancellation rate (probability of cancellation) is 0.315 for rooms booked earlier than 100 days from its arrival date, and 0.564 for rooms with more than 100 days in lead time to arrival date.

Matching based on lead time > 100 for selected variables. We selected the control variables based on what we feel will be possibly affected by the lead time. Due to the interest of space, we have left out the explanations for each variable.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```



The matching of data is not ideal, but given the large number of data points it is hard to match very accurately so we continued with our analysis with the given match results

We will now analyze to see if there is a significant difference in the cancellation rates once the data has been matched, with having only the lead time as the differing factor.

We can see from this that the mean cancellation rate (probability of cancellation) for the matched data is 0.345 for smaller lead times and 0.564 for larger lead times.

From this we can see that the rate of cancellation is caused by a larger lead time, as the rate of cancellation did not change much with all else being constant. To further confirm our hypothesis, we will do a logistic regression to see if the variable lead_bin is an significant variable by analyzing its p-value.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

As the variable lead_bin has a 3* significance, we should conclude that lead time is a significant factor for hotel booking cancellations.

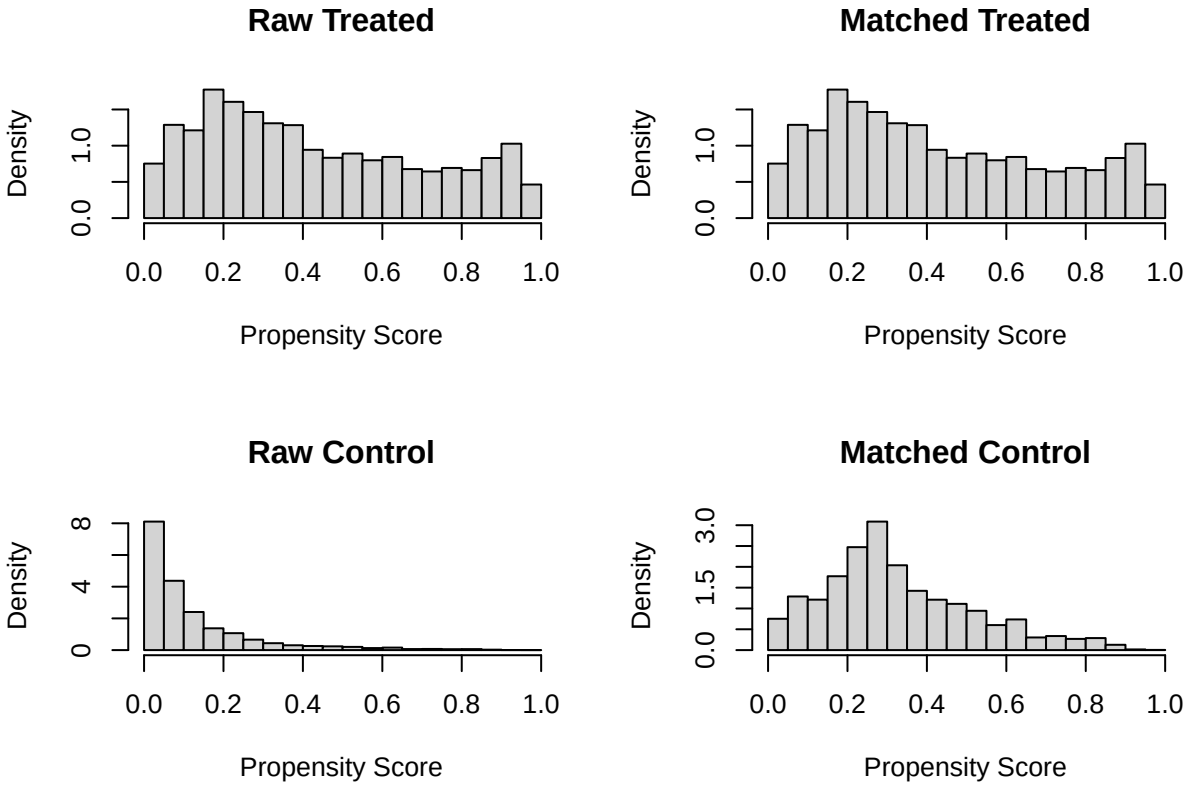
However, this analysis is only to compare between rooms booked within 100 days in advance against rooms booked more than 100 days in advance. As we increased the days from 100 to 200 and 300, we started to see that the number of cancellations were significantly different from the matched and unmatched data. We can therefore conclude that as lead time increases, there are other factors that contribute to the cancellation of rooms, and it is hard to point an

exact number of days which from cancellations are affected by lead time.Regardless, the hotel needs to be cautious **as lead-time increases, there is a higher chance of cancellations.**

##Market Segment The next variable that we would like to check for causation would be market segment. From our earlier analysis, we found that the market segment 'Groups' had the highest cancellation rate of about 68%. We repeated the same process of creating a new binary variable for if market_segment == 'Group' and tested for causality.

We can see from this that the mean cancellation rate (probability of cancellation) is 0.359 for rooms not booked in groups and 0.688 for rooms booked in groups.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```



From the graph, we can see that the matching of values for groups are much better matched than lead time.

We can see from this that the mean cancellation rate (probability of cancellation) increased to 0.409 for rooms not booked in groups and rooms booked in groups remained similar.

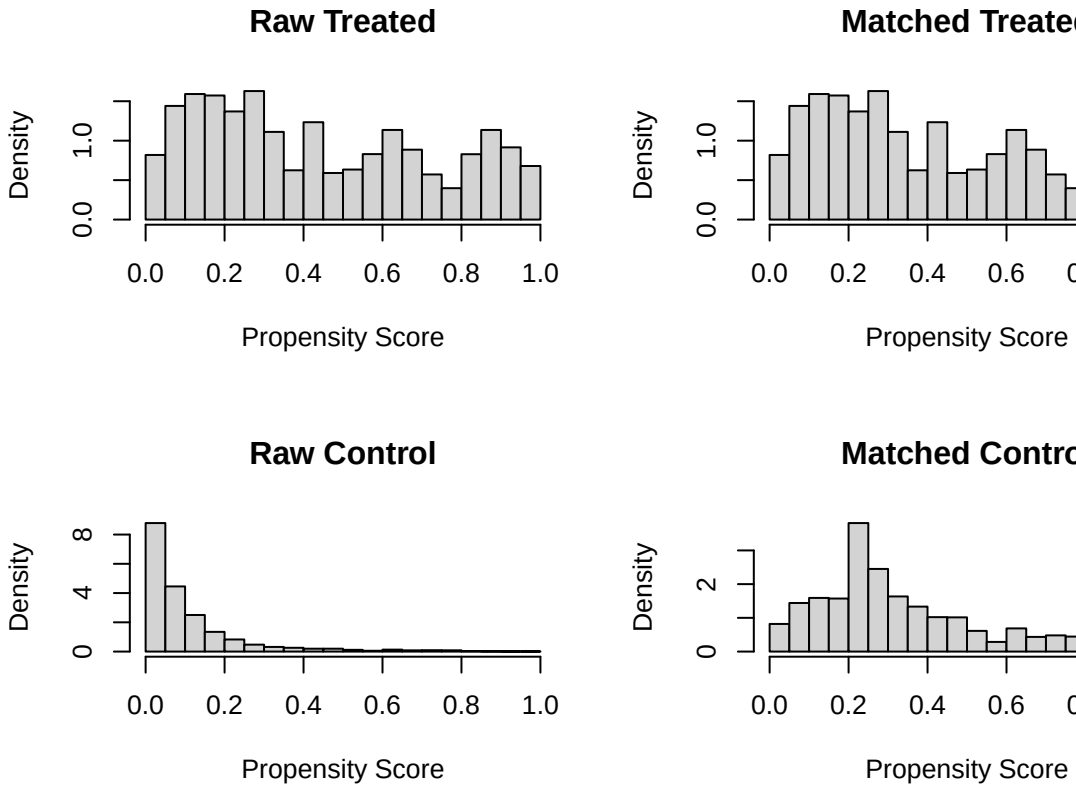
Since the difference is not significant as well, we can say that bookings by groups has a significant relationship with cancellations. We will do a logistic regression again to confirm the hypothesis.

As the variable groups has a 3* significance, we can indeed conclude that **group bookings** might potentially be a **strong causal factor** for hotel booking cancellations.

##Deposit Type The final variable that we would like to check for significance would be deposit type. From the CART model, we found that the deposit type is significant, and confirmed it when we checked that bookings with deposit_type as 'Non Refund' had the highest cancellation rate of about 99%. We then did the same steps above by creating a new binary variable for if deposit_type == 'Non Refund' and repeated the process. Non Refund refers to bookings that do not have a refund for cancellations.

We can see from this that the mean cancellation rate (probability of cancellation) is 0.998 for rooms with no refund and 0.304 for rooms with no deposit and refundable rooms.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```



Analysis of matched Non Refund values

Cancellations post matched analysis

After the data is matched we can see from this that the mean cancellation rate is still similar at 0.375 for rooms with no deposit and refundable and the same 0.998 for rooms with no refund. Since this is counter-intuitive, we would like to further check the significance of this through a logistic regression model.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

From this analysis, we can conclude that **deposit type: no refund** might have a **strong causal relationship** to the cancellations of bookings.

#Prescriptive Analytics 1. Algorithm that adjusts the hotel's level of overbooking based on seasonality

In order to maximize our overbooking strategy, prescriptive analytics can help by automatically adjusting how much to overbook the hotel based on factors that change throughout the year. This ensures the hotel implements the optimal level to overbook, and minimize the possibility of the negative consequences mentioned in the earlier section. One example is as such: When the algorithm identifies that it is currently the season for holiday bookings, it can automatically adjust to increase the level of overbooking, since customers are more likely to cancel during this season. This insight is based on our descriptive analysis in the earlier section.

2. Algorithm which tailors pricing and cancellation policies based on customer profile

Prescriptive analytics can be used to maximize a hotel's profit generating potential by charging the highest price possible to customers based on the type of customer. The hotel can also implement tailored cancellation policies to customers based on their history of cancellations. Our analysis shows that there is a pattern in certain segments which are more profitable, as well as reveals that customers who have had a history of cancellations are more likely to cancel.

Conclusion - Discussion on Findings

In our introduction, it was stated that the goal of our analysis is to attempt to answer two main questions:

- Is booking cancellations related to factors like lead time and customer type?
- Can we predict with reasonable accuracy if a booking will be canceled based on its attributes?

Based our data analysis, we distilled the most critical insights to these questions in the following section.

Significant variables and potentially causal relationships that affect booking cancellations

Lead Time

Based on our analysis, we learned that lead time is one of the most significant factors that affects cancellations. When lead time increases, customers are more likely to cancel their bookings. We first observed this with descriptive analysis in the scatter plot, then verified from our CART tree that lead times is one of the top-level separators in the decision tree, as well as the diagnostic analysis that revealed lead time is indeed a causal factor as well.

Deposit Type

One factor that we found from the CART model that might prove to be significant is deposit_type. We further confirmed it from the diagnostic analysis that there is a causal relationship. This results of this analysis were counter intuitive because it showed that rooms with No Refund had a 99% cancellation rate, as compared to rooms with No Deposit. We now know that customers are not bothered by the monetary penalty when deciding to cancel. For hotels, making it a Non Refundable rooms will not prevent cancellations.

Can we predict with reasonable accuracy if a booking will be canceled based on its attributes?

From our models, we found that with the given data the model can predict booking cancellations at a relatively high confidence as it had an AUC of **88% for CART** and **92% for Random Forest**. This proves that on top of the early warning signs provided by our analysis above, the models can indeed be **useful** in predicting booking cancellations.

Limitations and Further Studies

##Limitations on Dataset

Lack of customer behavior data

From our analysis, we discovered that variables relating to the **customer's behavior** show significance in leading to booking cancellations. Unfortunately, the dataset available to us is limited in providing variables that relate to customer behavior. However, we understand that such data is not as readily available and might potentially be difficult to monitor.

Scope of dataset

Our dataset is limited to City and Resort hotels in Portugal, which may/may not be fully relevant in other contexts. For example, we know that Portugal attracts a certain archetype of tourists who are interested in its rich history and are willing and able to travel to Western Europe. This archetype may differ for a hotel in a city like New York, which is known for other reasons.

Recommendations for future studies

Include more customer behavior data

One recommendation that can improve our study is to source for more customer demographic and customer behavior data. Examples include data which provides information on a customer's: age, gender, browsing sessions. With this additional data, we could potentially understand the **customer's booking journey** in greater detail, and identify alternative signs that are good predictors of booking cancellations.

Cross-reference results and findings from hotels in a different country

In order to improve our model and increase its relevance to hotels outside of Portugal, one recommendation is to perform a similar analysis on data from other hotels in other cities and cross-reference the results. Doing so will allows us to verify if our insights are specific to the context of Portugal City and Resort Hotels, or indeed applicable to other hotels.