DSA211 Statistical Learning with R

Project: Hotel Bookings

G3 Group 9:

KU SEOUNGLIM

LIM JIA YANG, JAMES

MIN HTET HEIN

MUHAMMAD IRFAN BIN MOHD TAHA

TEO GIGI

# Executive Summary

This report provides recommendations to common issues faced by the hotel industry, such as misallocation and inefficient allocation of resources and inaccurate room pricing. This report studies the different factors involved in each booking and produces recommendations that are catered towards improving productivity and increasing profits for City Hotel.

This report is divided into four main sections:

1. Exploratory Data Analysis
2. Room Price Forecasting
3. Booking Cancellation Prediction
4. Reserved Room Type Prediction

Exploratory data analysis of the data allowed us to discover patterns, spot anomalies and test hypotheses with the help of summary statistics and graphical representations. To forecast room prices, a multiple linear regression was conducted to determine the factors and their respective weightages that affect ADR (average daily rate), a metric used to determine how much each customer was spending at the hotel on an average day of their stay. ADR prediction was done using simple and multiple linear regression models to understand more on how each variable can impact ADR. Predicting booking cancellations was done with a decision tree model, and predicting room type that different demographic of customers wanted to reserve was carried out with a multinomial logistic regression.

Each section provides detailed statistical analysis of results and relevant graphs for visualisation. Recommendations for the hotel management team are also provided at the end of each section, along with our overall conclusion and accuracy for each model at the end of the report.

# Background

Hotel booking might seem like a simple operation on the surface, but it can be every hotel manager’s worst nightmare. When planning the perfect stay for their guests, there are many variables that can lead to unexpected problems. Even the most respected hotel chain cannot guarantee the perfect allocation of rooms to guests, proper resource allocation, and right price prediction for their rooms. However, with thorough analysis and machine learning algorithms such as logistic regression models and decision trees, they can no doubt improve their current prediction accuracy.

There are two different hotels contained in this dataset, Resort hotel and City hotel. This project focuses only on City Hotel as the majority of data (66.4%) pertains to this hotel. The dataset contains 31 variables describing the 79,330 observations of this hotel- each observation representing a hotel booking. Information regarding the variables can be found in Appendix A.

The bookings in this dataset are for scheduled stays between 1st July 2015 and 31st August 2017, this includes both bookings that were successfully carried out and bookings that were cancelled. All data elements pertaining to hotel or customers’ confidential information was deleted from this dataset.

The customers that have made bookings come from 168 different countries. As this is too wide a range to analyse, we have simplified it to two groups - PRT and OTH. PRT is the acronym for Portugal, while any other country that is not Portugal will fall into others (OTH). 39.03% of customers are from Portugal and they make up a significantly larger proportion of customers, as compared to customers from the other 167 countries.

We had to first clean the data as some data contained NA and null values. Some of the data had zero Adults, zero Children, and zero Babies, which is illogical. We have removed these rows before performing our analysis.

We then conduct Exploratory Data Analysis on the hotel bookings to better understand the data we were handling. We then wanted to uncover the three areas that we felt every hotel manager would benefit from knowing the answers to. These areas are as follows:

1. Predicting average spending rate of each guest
2. Predicting booking cancellations
3. Predicting type of room guests would reserve

Armed with good predictive models, hotel managers can then better allocate their resources to ensure that the hotel serves its guests to the best of its abilities.

For this project, we assumed that there were no unique adhoc events that occurred between 1st July 2015 and 31st August 2017 that could have caused a large increase/decrease in hotel bookings. Examples of these events are contagious diseases, natural disasters, large scale once-off events such as F1 Grand Prix etc.

# Exploratory Data Analysis

EDA refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations (Patil, 2018). The following are areas of interest that we wanted to explore from the data:

* How long are guests staying for?
* Which countries do guests come from?
* Which distribution channels are used more?
* What party sizes do guests come in?
* Which month has the highest number of cancellations/bookings?

## How long are guests staying for?

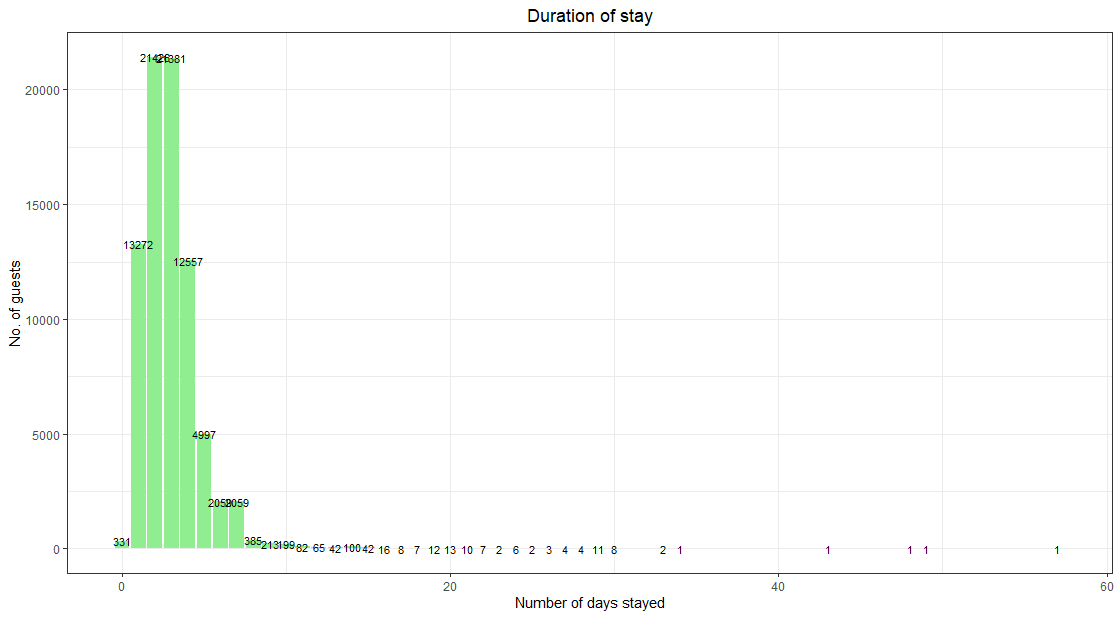


Fig 1.a. Graph of Duration of Stay



Fig 1.b. Summary of Duration of Stay

From the graph, we can deduce that guests’ duration of stay largely falls between 1-4 days; with the majority being 2 and 3 days, which is also reflected in the mean and median scores of 2.978 days and 3 days respectively. The summary statistics echo this as well as the third quantile (75%) of all guests’ stay is 4 days, meaning that 75% of all guests stay at the hotel for 4 or fewer days.

Understanding the duration of stay of guests is useful information for the hotel as they can better schedule their guests to ensure maximum utilization of their rooms. Having knowledge of guests’ duration of stay can also aid the hotel in other aspects such as recommending better suited packages and plans for guests. For example, the hotel can advise its partnered attraction companies to offer packages that are more suited to guests’ duration of stay.

## Which countries do guests come from?

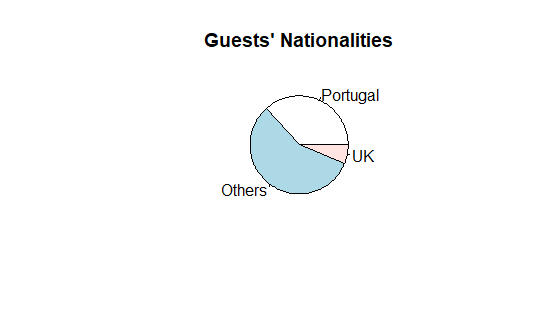


Fig 2. Pie Chart of Guests’ Nationalities

The hotel's guests mainly came from Portugal (about 40%) whereas the rest of the guests’ nationalities is a minority in the data set. The second largest group would be from the United Kingdom (about 10%) which means that “Portugal” is much more significant as compared to all the other nationalities. As such, we would like to focus more on Portugese guests due to the large disparity between “Portugal” and all the other nationalities.

## Which distribution channels are used more?

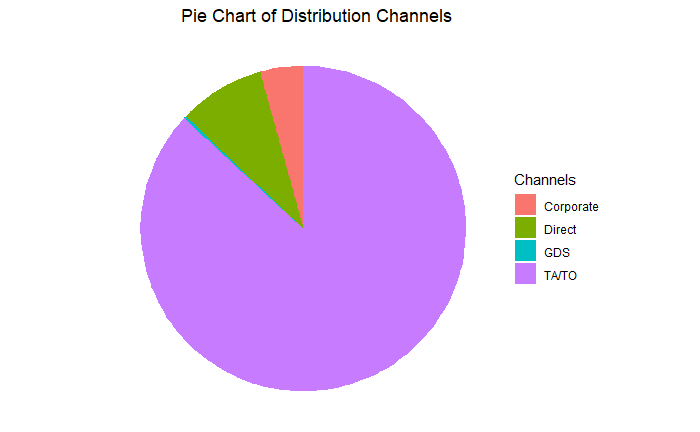


Fig 3. Pie Chart of Distribution Channels

Distribution channels refer to channels where hotel bookings were made from. The hotel classifies its channels through 4 categories: Corporate, Direct, GDS and TA/TO. Corporate refers to bookings made by companies for their staff, Direct refers to bookings made directly on the hotel’s website, GDS refers to a Global Distribution System and lastly, TA/TO refers to travel agencies or tour operators.

From this pie chart, we can see that a vast majority of all bookings are made through TA/TO. This is useful information to the hotel as it advises the marketing team to prioritize their marketing efforts towards TA/TO to impact the largest group of the hotel’s guests. In contrast, bookings made through GDS are few, indicating that this distribution channel is not effective in bringing in guests.

## What party sizes do guests come in?

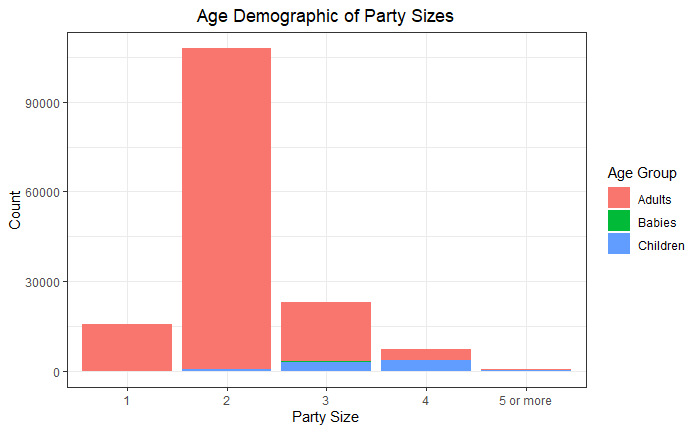


Fig 4. Bar Chart of Age Demographic of Party Sizes

The hotel categorized the age of their guests into three groups, adults, children and babies. From the bar chart, we can see that guests’ bookings of party size two are the majority, this is likely made up of couples. We can also see that party sizes of four are made up of half adults and half children; these are likely to be young families with two children. Lastly, there are very few bookings made for party sizes five and above, this implies that the hotel might not be well suited for bigger families, hence deterring bigger families and larger groups from staying at the hotel.

## Which month has the highest number of cancellations & bookings?

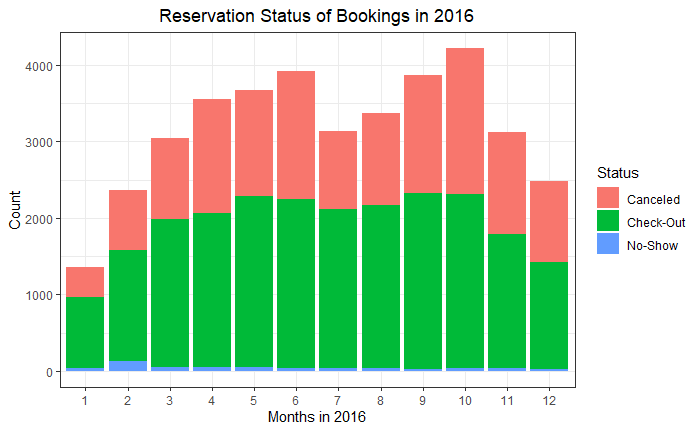


Fig 5. Bar Chart of Reservation Status of Bookings in 2016

Another area we wanted to study was the number of bookings made per month and the breakdown of the status of these bookings, which is reflected in this stacked bar chart . Every booking made can result in one of three outcomes: Canceled, Check-Out or No-Show. Canceled refers to bookings canceled by the guest before staying at the hotel, Check-Out refers to guests “successfully” completing their stay at the hotel, and No-Show refers to guests that did not stay at the hotel but did not cancel their booking either.

From the graph, we can see that business peaks from May to October, then begins to taper off from November to January, before picking up again in February. From this information, the hotel can better prepare itself for the volume of guests coming in. Knowing its lull periods is useful as well as the hotel can schedule its potentially business-disruptive activities then; such as renovation/repair works, sending staff for training or encouraging staff to clear their leave then etc.

# Price Forecasting Models

## **Introduction**

To help City Hotel manage their resources better (in terms of logistics and staff, etc), we analyse the city hotel data to understand how different independent variables affect the ADR.

We also found out about the price forecasting for the rooms based on the month of visit and the demographic of the visitors and the characteristics of their stay.

The average daily rate (ADR) is a metric used in the hospitality industry to indicate the average revenue earned for an occupied room on a given day. The average daily rate is one of the key performance indicators of the City Hotel. Thus, we will find out which factors can significantly impact the ADR of the room in City Hotel so that it can focus its resources on the most important areas to keep generating high revenues.

## **Analysis**

For our study of how ADR is determined, we focus on four main areas of the independent variables for the hotel to pay attention to – the visitors’ demographics, hotel services used and type of days they stayed on and the months of visit. We will then generate the models to predict ADR for each part.

Demographics of the visitors include the citizenship of visitors in Portugal, the family composition of each stay and the customer type. The citizenship of the visitors is separated into two simpler categories - Local (Portugese) and foreigner (all other countries) as the individual countries of the visitors. The stay of the foreign visitors can improve the ADR by as much as 21.87% per visitor. For each stay, the number of adults and children (mainly due to the adults whom they stay together with as there is a significant interaction effect between them) makes a significant contribution to the ADR. The number of babies has no significant impact on the ADR. The ADR rate is impacted positively by the customer types - Contract, Transient, Group and Transient-Party in decreasing order.

|  |  |
| --- | --- |
|  |  |
| Fig 6. Boxplot of ADR against Country | Fig 7. Boxplot of ADR against No. of Adults |
|  |  |
| Fig 8. Boxplot of ADR against No. of Children | Fig 9. Boxplot of ADR against No. of Babies |

Overall Model to predict ADR using demographics:

adr=67.24-15.09citizenshipPRT+17.26adults+3.36children-3.20customer\_typeGroup+11.54customer\_typeTransient+0.88customer\_typeTransient-Party+16.77adults\*children

Type of days the visitors stay on (whether it is weeknight or weekend) also has a significant effect on the ADR- weekend night stay leads to greater increase in ADR than the week night stays. There is a negative interaction effect between these two variables so the visitors who stay on week day will not have their hotel stay reached to weekend night so it drops the ADR predicted.

Model to predict ADR from whether the visitors stay weeknight or weekend night:

adr=97.315+2.969\*stays\_in\_week\_nights+3.298\*stays\_in\_weekend\_nights–0.583\*stays\_in\_week\_nights\*stays\_in\_weekend\_nights

Hotel Services used by the visitors include Meal Packages, Required Car Parking Spaces, Reserved Room Type and Total Special Requests. For Meal Packages, Half Board (HB) which includes breakfast and one other meal, led to increase in predicted ADR while Bed & Breakfast(BB) and Full board(FB) which includes all three meals, results in decrease in predicted ADR. Positive impact on the number of required car parking spaces will lead to increase in predicted ADR. The increase in the predicted ADR is in the decreasing order of room types – G,F,E,D,A,B,C,P. Reserved room type is the most significant contributor to ADR as it has the lowest variance from the prediction model using only it as the independent variables. The increase in the total of special requests also leads to increase in predicted ADR.

Model to predict ADR from all types of services used:

adr=83.76-29.91\*mealFB+19.33\*mealHB+2.74\*mealSC+7.17\*required\_car\_parking\_spaces-6.01\*reserved\_room\_typeB-9.63reserved\_room\_typeC+36.21\*reserved\_room\_typeD+58.56\*reserved\_room\_typeE+91.94\*reserved\_room\_typeF+ 103.93\* reserved\_room\_typeG-93.67\* reserved\_room\_typeP+6.20\* total\_of\_special\_requests

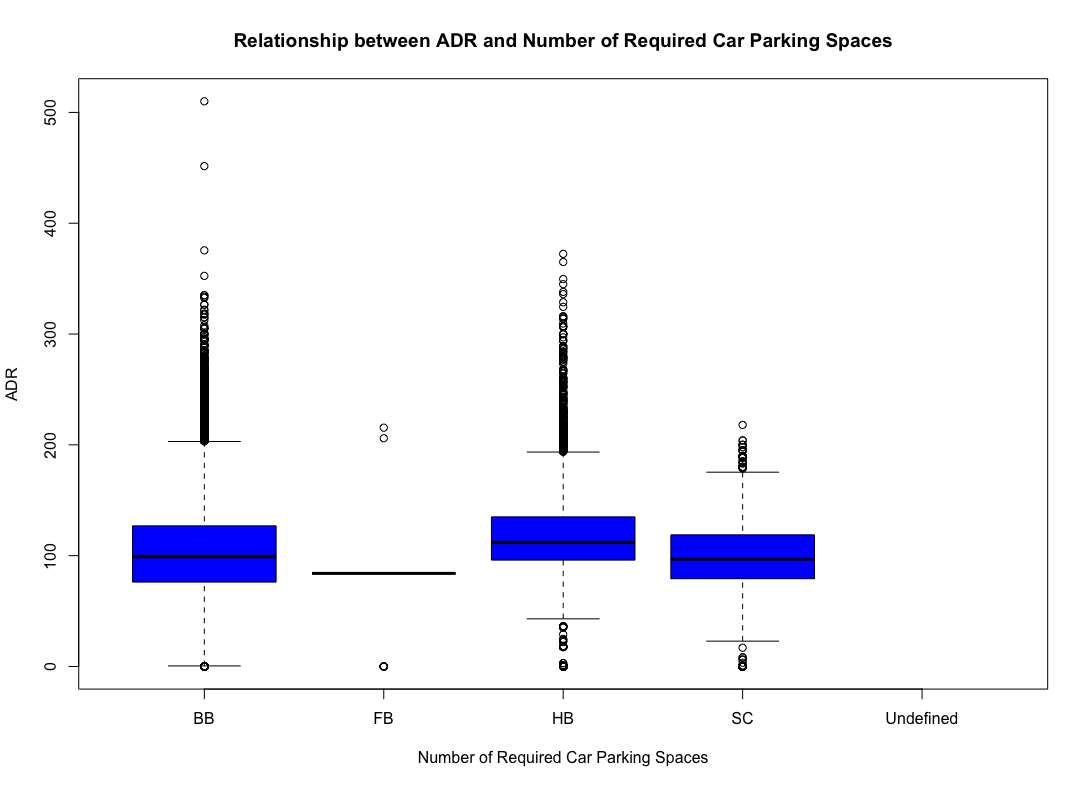
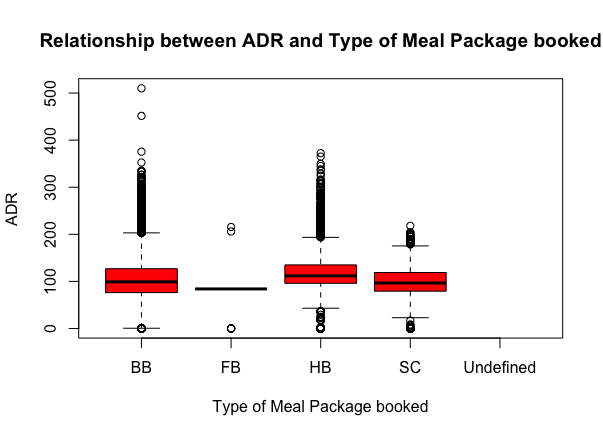


Fig 10. Boxplot of ADR against Type of Meal package Fig 11. Boxplot of ADR against No. of car parking spaces

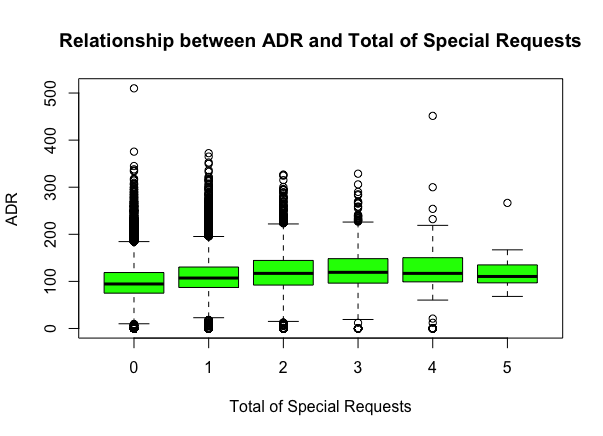
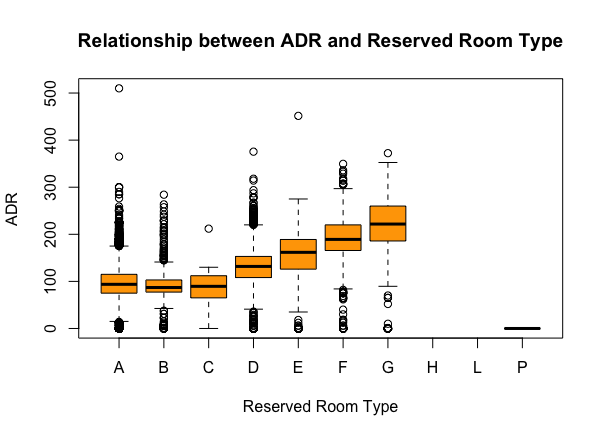


Fig 12. Boxplot of ADR against Reserved Room Type Fig 13. Boxplot of ADR against Total Special Requests

Month of visit also has impact on the ADR. With the month of April as the baseline, our model shows that there is an increase in predicted ADR in May, June and August with only a small decrease in July (which is not statistically significant). Thus, ADR is usually up from the late spring to the end of Summer (from April to August) and the hotel earn increased revenue in the summer period as it will be more visitors coming due to summer holidays.

Model to predict ADR from month of visit:

adr=111.25+3.47\*August-22.43\*December-26.16\*February-28.62\*January-0.52\*July-7.82\*June-19.43\*March-10.39\*May-23.18\*November-11.28\*October-1.25\*September

## **Recommendation**

City Hotel’s management could use the results derived from the model to decide which area to best allocate the hotel’s resources in which period of time so that it can focus on earning maximum revenue from the visitors they received. This will also help the city hotel to plan resource allocation better so that there will be less wastage of resources as they will be mainly channeled towards those areas which makes a significant impact on the ADR.

# 

# Predict Cancellations

## **Introduction**

Of all the bookings, 58.28% of the bookings were eventually cancelled. This can potentially lead to resource wastages if City Hotel is unable to foresee the potential cancellations of bookings, which eventually will affect their potential revenue and profit. To aid City Hotel in better managing their resources, we analysed the booking data to understand what attributes of a booking can affect the probability of a booking cancellation. This can allow the hotel management to have a better understanding of more significant factors and take precautionary measures if deemed necessary.

## **Analysis**

The correlation matrix in Fig highlights the correlation between the variables. Specifically of interest, it shows which independent variables are highly correlated to whether a booking is cancelled (or not).

As such, the variables *Deposit Type* (deposit\_type\_Non\_Refund), *Countr*y (countryPRT), *Lead Time* (lead\_time), *Number of Previous Cancellations* (previous\_cancellations), *Distribution Channel* (distribution\_channelTA.TO), *Customer Type* (customer\_type\_Transient.Party) are more correlated with a *Cancelled Booking* (is\_canceled\_Y).

Boxplots and scatterplots were plotted for the highly correlated independent variables identified against the booking status to further analyse, as seen in Fig 15.

As seen from Fig 14, in general, customers tend to cancel their bookings as lead time increases. The difference is more significant for customers from PRT as seen in Fig 16, as the lead time increases, customers from PRT are more likely to cancel their bookings, as compared to customers from OTH.

Customers of customer type Transient tend to cancel their bookings more with higher lead time, as seen in Fig 17, as compared to the other three customer types.

Despite the deposit type as non-refundable, customers with deposit type non-refundable tend to cancel their bookings more than customers of deposit type refundable and no deposit, as seen in Fig 18 and Fig 19.

|  |  |
| --- | --- |
|  |  |
| Fig 14. Boxplot of Lead Time against Booking Status | Fig 15. Correlation Matrix |
|  |  |
| Fig 16. Boxplot of Lead Time against Booking Status and Country | Fig 17. Scatterplot of Lead Time and Previous Cancellations as a function of  Customer Type and Booking Status |
|  |  |
| Fig 18. Scatterplot of Lead Time and Previous Cancellations as a function of  Deposit Type and Booking Status | Fig 19. Boxplot of Lead Time against Distribution Channel and Booking Status |

## **Decision Tree**

We decided to use decision trees for easy illustration of possible customer categories for each possibility - booking is cancelled (noted in leaf nodes as Y outcome), booking is not cancelled (noted in leaf nodes as N outcome).

Instead of using the tree library, we used the rpart library as rpart is a superset of the tree object (Carbon, 2005) and both categorical and continuous data can be used in rpart. This is in our favour as our data set contains a handful of categorical data. The rpart library includes pruning and cross validation as well.

We separated the data into training and test sets for the decision tree.

From the decision tree in Fig 20, the following customer categories would most likely result in a cancelled booking:

1. Deposit type is Non-Refundable and Refundable
2. Deposit type is No Deposit, Number of Previous Cancellations is greater than 0.5 (logically speaking greater than 0 since Number of Previous Cancellations is a whole number)
3. Deposit type is No Deposit, Number of Previous Cancellations is smaller than 0.5 (logically equals to 0 since Number of Previous Cancellations is a whole number), Lead Time more than 8.5 days, Total Number of Special Requests more than 0.5 (logically speaking greater than 0 since Total Number of Special Requests is a whole number), Customer Type is Transient, Distribution Channel is TA/TO, Average Daily Rate more than $91.00.

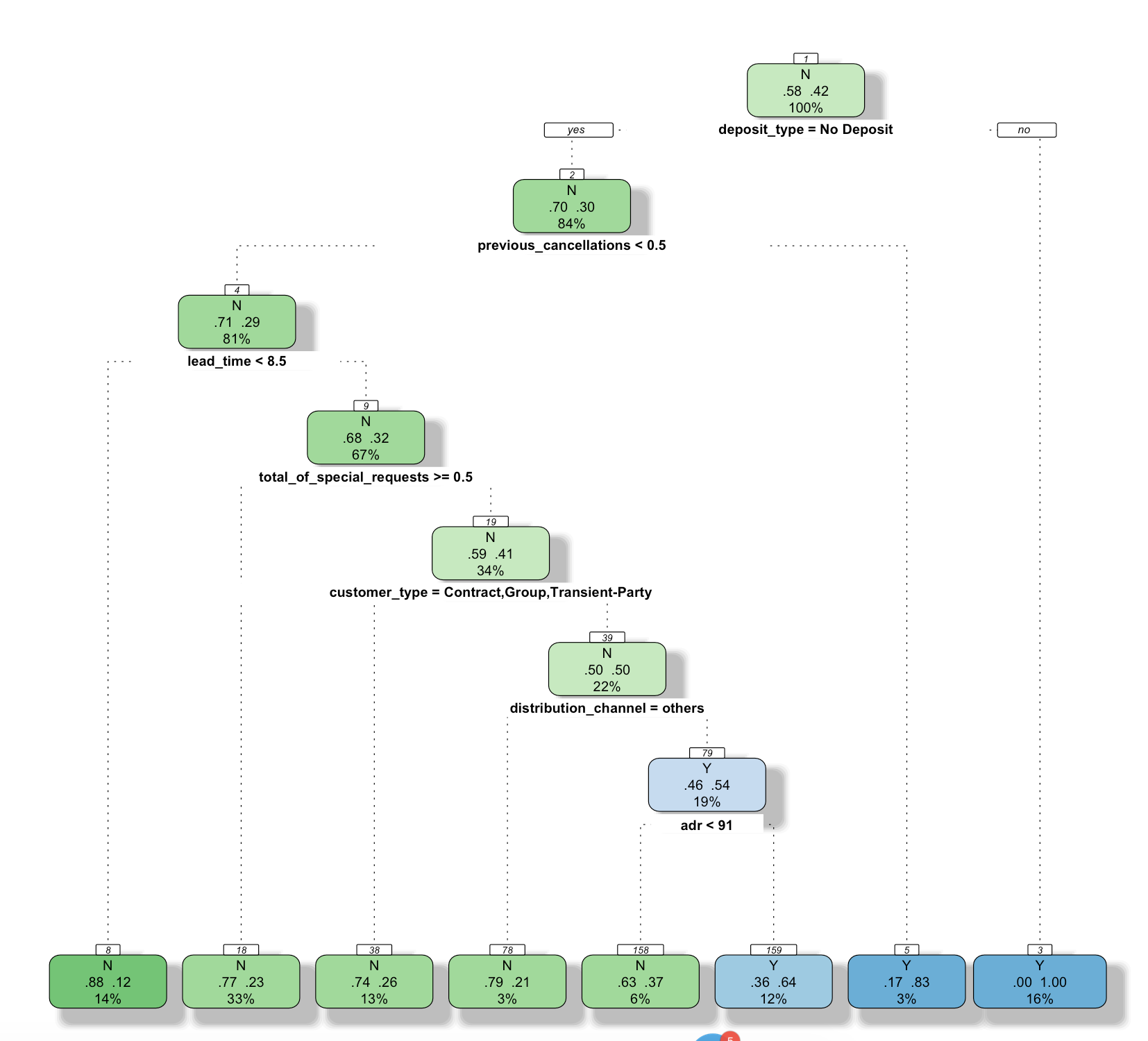


Fig 20. Decision Tree

## **Recommendations**

City Hotel Management should keep a lookout of customers that fall into these three broad categories as they are more likely to cancel their bookings, as seen from the analysis done. As such, they account for the possibility of cancellations from these customer groups and allocate resources accordingly.

Moreover, although a customer’s country is not a sub-root in the decision tree, as seen from the analysis earlier, City Hotel Management should also keep a lookout of customers from Portugal with prolonged lead time.

# Predict Type of Reserved Room

**Introduction**

In the data we have gathered for City Hotel, we found that 9.97% of the customers from 2015-2017 were not assigned to the rooms that they had previously reserved. This means that they were either offered an upgrade or a downgrade based on the hotel’s availability. Another possible reason is that the room was not up to the customer’s expectations. This poses the issue of efficient allocation of rooms in the hotel management. As such, we analyse the data to see what affects a consumer’s decision in choosing a type of room. In this analysis, we have used historical data to predict a customer’s choice of room based on their respective demographics and needs. This will help the hotel’s management to understand the significant factors behind a customer’s choice of room and assist customers in selecting the suitable room based on their needs.

**Analysis**

We gathered the data from City Hotel and eliminated irrelevant variables before running a regression. For this particular regression, we used the following independent variables: *total\_stay, adults, children, babies, meal, PRT, market, distribution, is\_repeated\_guest, deposit\_type, adr, required\_parking\_spaces* and *total\_special\_requests.* In this model, we combined the 2 variables, *stays\_in\_week\_nights* and *stay\_in\_weekend\_nights* into 1 variable, *total\_stay* as the difference between a weeknight or weekend night stay is not important for this model. As we are interested in finding out the probabilities of a customer choosing each type of room based on their demographic, we have decided to use a Multinomial Logistic Regression to run this model instead of a Binomial Logistic Regression. After running the regression, checking the Deviance Statistic and p-values, our results are as follows:

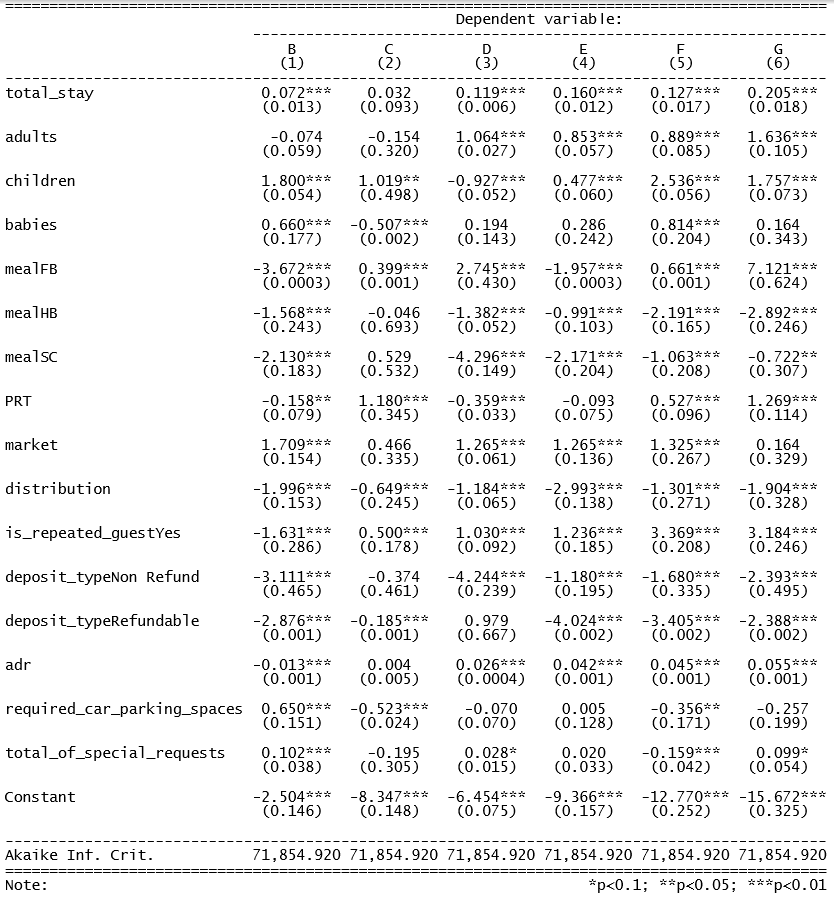


Fig 21. Result of Multinomial Logistic Regression

From the results, we could see that most of the variable coefficients are significant. In this model, we have chosen room type A as our reference category for ease of inference as room type A is the most common in the dataset. Some important conclusions we could derive from this result include:

* Repeated customers are more likely to reserve the more premium and expensive rooms as compared to first-time customers *(room type F & G)*
* Families travelling with Children and/or Babies are likely to choose *room type F*
* Customers who would opt for the Full-Board meal option are likely to choose *room type D & F*
* Family with Babies and/or require parking spaces tend to avoid *room type C*
* Customers who intend to stay longer tend to choose the more premium and expensive rooms

**Recommendations**

City Hotel’s management could use the results and conclusions derived to strategise their marketing campaigns to target the right audience. The management could also use the results to predict the suitable room type and provide recommendations to customers. We could automatically generate the suitable type of rooms for the different customers based on their requirements and budget as seen in the following table:

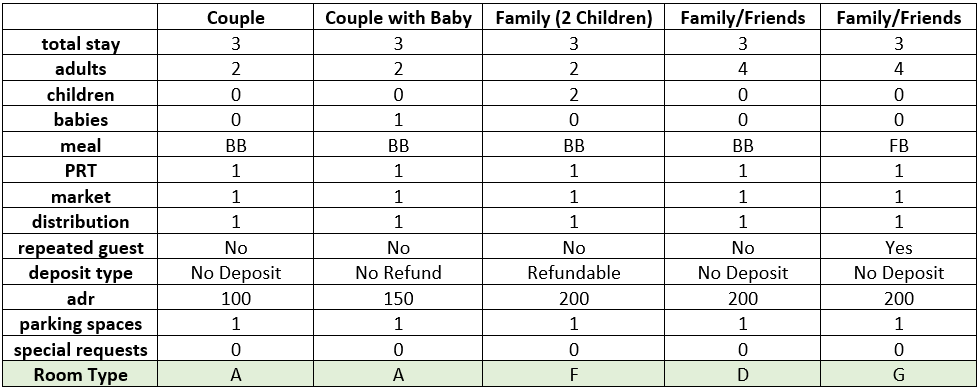


Fig 22. Recommendation Table

With such the following predictions, the Hotel’s management can efficiently allocate their marketing resources to target the specific audience effectively, catering to their needs. Also, customers will be able to get accurate recommendations on a suitable room type that satisfy their needs and requirements for their stay at City Hotel, which helps to prevent disappointment upon arrival.

# Conclusion

With an accuracy of 79.70%, the model to predict which customer types are more likely to cancel a booking is reliable and useful for City Hotel’s management. To better manage and allocate resources, City Hotel should be on the lookout of these three customer categories:

1. Deposit type is Non-Refundable and Refundable
2. Deposit type is No Deposit, Number of Previous Cancellations is greater than 0.5 (logically speaking greater than 0 since Number of Previous Cancellations is a whole number)
3. Deposit type is No Deposit, Number of Previous Cancellations is smaller than 0.5 (logically equals to 0 since Number of Previous Cancellations is a whole number), Lead Time more than 8.5 days, Total Number of Special Requests more than 0.5 (logically speaking greater than 0 since Total Number of Special Requests is a whole number), Customer Type is Transient, Distribution Channel is TA/TO, Average Daily Rate more than $91.00.

City Hotel’s management can also use the third model to predict the possible room types for each demographic of customers to better strategise their marketing campaigns in order to maximise their sales potential. This model has an overall accuracy rate of 84.70% and can also accurately provide recommendations for customers on the suitable room type for their stay.

We are sure the above recommendations will be able to adequately address the inherent problems faced by City Hotel and in the hotel industry.

# References

Carbon. (2005). [R] Difference between "tree" and "rpart". Retrieved March 20, 2020, from <https://stat.ethz.ch/pipermail/r-help/2005-May/070905.html>

Mostipak. J. (2020). Hotel Booking Demand. Retrieved March 5, 2020, from https://www.kaggle.com/jessemostipak/hotel-booking-demand

Patil. (2018). What is Exploratory Data Analysis?. Retrieved March 24, 2018, from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

# Appendix

## Appendix A - Variables & Description

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Ǡ hotel | Resort Hotel or City Hotel |
| # is\_canceled | Value indicating if the booking was cancelled (1) or not (0) |
| # lead\_time | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| # arrival\_date\_year | Year of arrival date |
| Ǡ arrival\_date\_month | Month of arrival date |
| # arrival\_date\_week\_ number | Week number of year of 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 week nights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| # adults | Number of adults |
| # children | Number of children |
| # babies | Number of babies |
| Ǡ 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) |
| Ǡ country | Country of origin. Categories are represented in the ISO 3155–3:2013 format |
| Ǡ market\_segment | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| Ǡ 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 |
| Ǡ reserved\_room\_ type | Code of room type reserved. Code is presented instead of designation for anonymity reasons. |
| Ǡ 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 |
| Ǡ 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. |
| Ǡ agent | ID of the travel agency that made the booking |
| Ǡ 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 |
| Ǡ 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) |
| Ǡ 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 |
| Ɗ reservation\_status \_date | Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel |

|  |  |
| --- | --- |
| Legend | |
| Ǡ | String |
| # | Float |
| Ɗ | Date |