# Modern Data Mining Final Project - Hotel Cancellations

Jacob Canelo-Garcia Kateryna Suprun Kyle Liao

4/30, 2023

## Contents

**Executive Summary** **2**

Background . . . 2
Data Summary . . . 2
Methods and Main Findings . . . 2
Issues and concerns . . . 2

**Detailed Workflow and Analysis** **3**

Goal of the study . . . 3
The Data . . . 3

**Research approach** **5**

Getting Familiar with Data and EDA . . . 5
Data Cleaning and Pre-processing . . . 5
Percentage of Cancelled and Not Cancelled Bookings . . . 5
Top Countries of Origin by Frequency of Bookings . . . 5
Distribution of Average Daily Rates Across Hotel and Room Type . . . 6
Guest Count by Month and Hotel Type . . . 6
Modeling . . . 8
Preparing Data for Model . . . 8
Lasso Regularization . . . 8
Multiple Logistic Regression . . . 11
LASSO + Multiple Logistic Regression . . . 13
Regression Trees . . . 13
Random Forest . . . 15
XGBoost Tree . . . 15
Conclusions . . . 16
Further Extensions . . . 16

# Executive Summary

## Background

Analyzing hotel bookings is important because it helps hotel management make informed decisions that can positively impact the bottom line. By studying patterns in booking data, hotel operators can determine the best times to offer promotions, adjust prices, allocate resources, and improve guest experiences. Additionally, analyzing cancellation data can help hotels identify areas where they may need to improve their booking policies, such as offering more flexible cancellation options or better managing overbooking situations. Predicting cancellations can also help hotels avoid revenue losses by allowing them to adjust staffing levels and resource allocation based on expected occupancy rates. In short, analyzing hotel bookings is crucial for ensuring the success and profitability of any hotel business.

## Data Summary

We have a large data set with each row corresponding to a unique hotel booking. Each booking has information attached to it such as arrival date, how many nights are spent, the number of guests, etc. See the data card below for a breakdown of each feature of the data and the information it represents.

## Methods and Main Findings

In this study, we look to conduct exploratory data analysis to answer a number of questions such as how guest counts vary across months and hotel types, what percentage of bookings are ultimately cancelled, what the top 20 countries in terms of booking frequencies are, etc.

We follow these findings up with modeling to predict whether or not a booking will be cancelled based on other features of a row (e.g. guest count, average daily rate, if a guest is a repeated guest, etc.). This essentially boils down to a classification question, and so we look to tackle this problem with classification models. We first run a logistic regression with all features included. We then use LASSO for regularization and run a logistic regression with select features. Next we create a random forest model, followed by a gradient boosted tree.

Accuracy scores for each model:

Logistic Regression: 0.81

LASSO + Logistic Regression: 0.78

Random Forest: 0.86

XG Boost: 0.82

## Issues and concerns

The data set used here has records for only two years: 2015-2017, which might not be fully representative of all other years. This data set was also recorded pre-COVID. Therefore, the hotel industry might have changed to an extent after COVID, ie. updating their safety and cleaning protocols, reduce common area capacity, and the people's preferences for travel might have also changed. The data set also has a pretty limited attributes set. Having any indicators for weather or news events would add to the accuracy of our models.

# Detailed Workflow and Analysis

## Goal of the study

In this project, we explore what factors influence the demand of different hotel listings. We explore different customer profiles and hotel types and seek to understand what motivates booking decisions. After getting a better understanding of hotel data and factors relevant to cancellations, we look to predict whether a booking will be cancelled or not based on other features of the booking.

In this project, we analyze a data set that contains booking information for different types of hotels in order to deduce what factors influence the demand for listings. We seek to answer questions like how to get the best daily rate when booking dates, what the best time of year to book hotel rooms is, or how to predict if a hotel will have disproportionately high demand compared to other options.

## The Data

This is a data card taking from the original data set that explains variables present in our data.

| Variable | Description |
|---|---|
| hotel | Hotel (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 PMS 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 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 |
| meal | Type of meal booked. Categories are presented in standard hospitality meal packages:<br>Undefined/SC – no meal package<br>BB – Bed & Breakfast;<br>HB – Half board (breakfast and one other meal – usually dinner)<br>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 |

| Variable | Description |
|---|---|
| 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:<br>No Deposit – no deposit was made<br>Non Refund – a deposit was made in the value of the total stay cost<br>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:<br>Contract - when the booking has an allotment or other type of contract associated to it<br>Group – when the booking is associated to a group<br>Transient – when the booking is not part of a group or contract, and is not associated to other transient booking<br>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:<br>Canceled – booking was canceled by the customer;<br>Check-Out – customer has checked in but already departed;<br>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 Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel |

Data files:

- **`hotel_bookings.csv`**

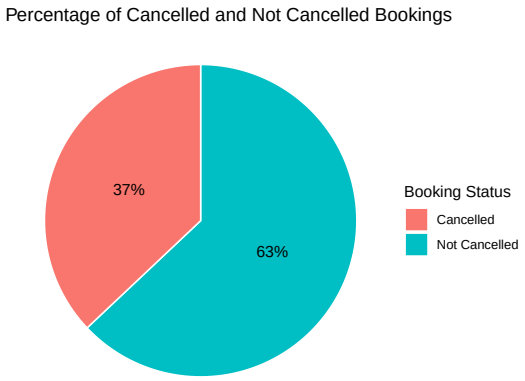
# Research approach

## Getting Familiar with Data and EDA

### Data Cleaning and Pre-processing

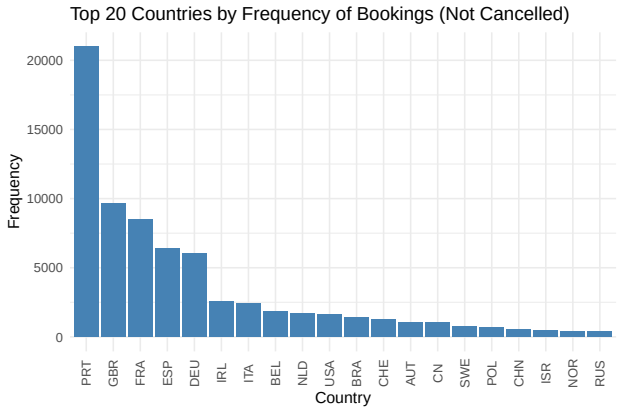
To clean data, we begin by removing bookings where there are no guests at all as this does not make sense. We also look to do some initial visualizations by generating graphs on where guests originate from, percentage of cancelled bookings, distribution of average daily rates across room and hotel types, and a time series of the total number of guests per month by hotel type.

### Percentage of Cancelled and Not Cancelled Bookings




We see that 63% of bookings are not cancelled, and 37% of bookings are cancelled. This is a significant number of cancellations. If we are able to predict cancellations based on other information belonging to a booking, we may be able to save hotels a lot of money.

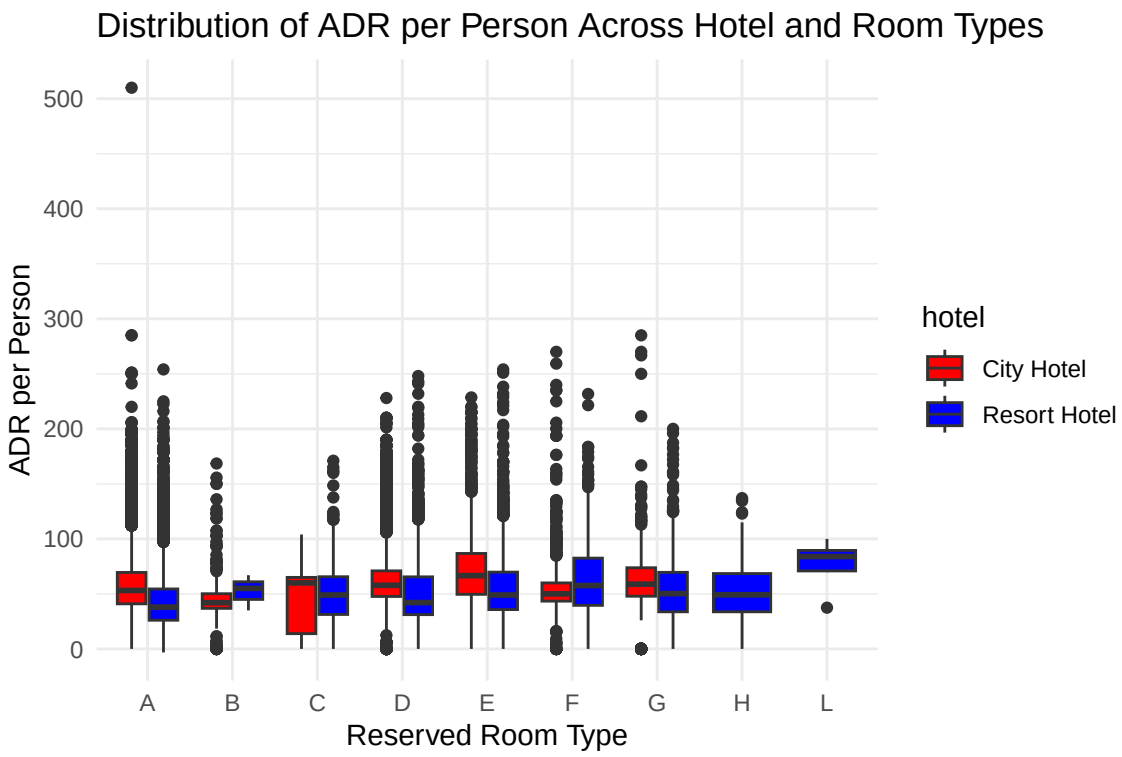
### Top Countries of Origin by Frequency of Bookings




We see that the top 5 countries by frequency of bookings are PRT (Portugal), GBR (Great Britain), FRA (France), ESP (Spain), and DEU (Germany). Intuitively, this makes sense as the hotel industry in these countries is very large, and thus we expect a large number of bookings by people in these countries.

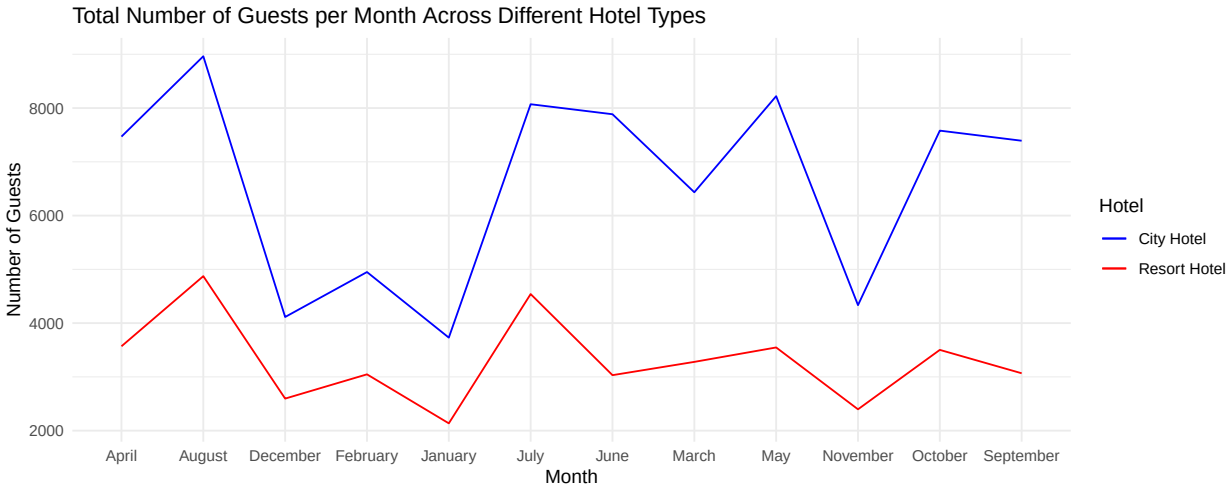
### Distribution of Average Daily Rates Across Hotel and Room Type

```
## [1] "Resort Hotel" "City Hotel"
```

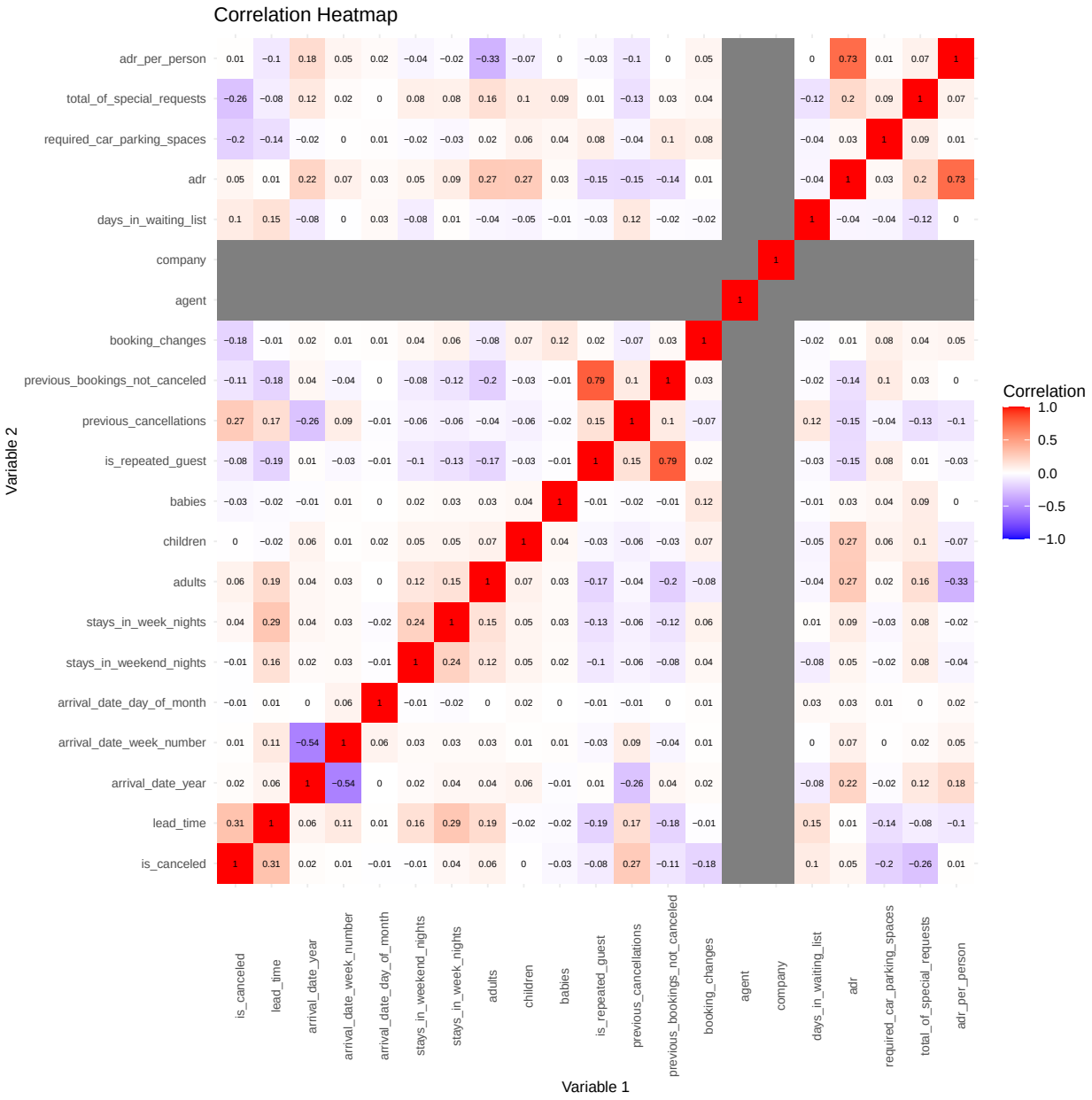



We see here across room types there is not a clear trend between ADR per Person and the hotel type. For room types A, C, D, E, G, we see that city hotels are on average slightly higher in terms of ADR/pp. For the other room types, we see that resort hotels have comparable median prices or exceed the median ADR/pp of city hotels.

### Guest Count by Month and Hotel Type



Based on this graph, number of guests in City Hotels is always greater than number of guests in Resort Hotels. August looks to be the busiest month for both types of hotels, and January is the lowest.




This correlation map captures both linear and nonlinear relationships as it depends on spearman coefficients. We see associations that intuitively make sense to us like a positive association between the number of adults and children in a listing and the adr. Other feature assocations that are particularly notable for our classification question is that we see high positive associations between lead times and adr as well as previous cancellations and adr. Strong negative associations include the presence of booking changes or special requests, and required car parking spaces.

## Modeling

**Preparing Data for Model** We notice that a lot of entries in 'company' attribute are nulls:

```
## [1] 0.9443771
```

Over 94% of 'company' column is null. Therefore, the best strategy will be to drop the column altogether. If we were to remove the null rows, most of the other data would be gone.

We will turn categorical entries: hotel type, arrival_date_month, meal, country, market_segment, agent, distribution_channel, reserved_room_type, customer_type, reservation_status, reserved_room_type, assigned_room_type.

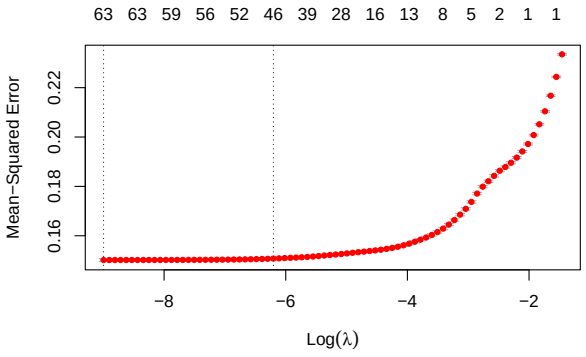
We identifying categorical columns, as those will be turned into factors:

```
##  [1] "hotel"                 "arrival_date_month"
##  [3] "meal"                  "country"
##  [5] "market_segment"        "distribution_channel"
##  [7] "reserved_room_type"    "assigned_room_type"
##  [9] "deposit_type"          "customer_type"
## [11] "reservation_status"    "reservation_status_date"
```

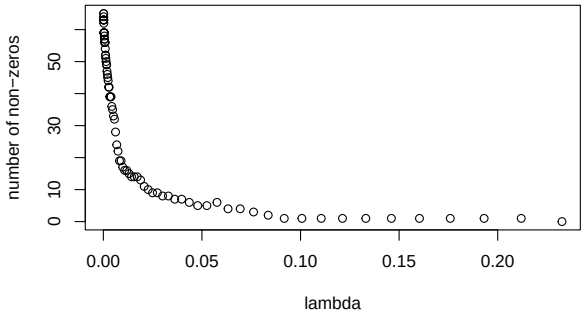
Here, we convert categorical features to factor variables.

We drop the `agent` and `company` features as they are predominately null, and we believe that they will not contribute to the model. We also drop reservation_status_date and reservation_status as those are directly correlated with is_canceled. Therefore, these would overpower any effects other attributes would have. We also drop the column country. We recognize that country's climate as a certain date might effect whether customers cancel the booking. In this study, we are interested in such effect. Rather, we look directly at booking statistics available to hotels to notice any correlations, internationally.

### Lasso Regularization




From the above plot we see that the best value for $\log(\lambda)$ is between about -9 and -6.3.




We extract the coefficients from the cross-validated Lasso model fit stored in the **cv.fit** that have the coefficients corresponding to the value of lambda that gives the smallest value of the mean cross-validated error plus one standard error.

```
##  [1] "lead_time"                      "arrival_date_monthDecember"
##  [3] "arrival_date_monthFebruary"     "arrival_date_monthJune"
##  [5] "arrival_date_monthMarch"        "arrival_date_monthSeptember"
##  [7] "stays_in_weekend_nights"        "stays_in_week_nights"
##  [9] "adults"                         "children"
## [11] "babies"                         "mealFB"
## [13] "mealHB"                         "mealUndefined"
## [15] "market_segmentComplementary"    "market_segmentOffline TA/TO"
## [17] "market_segmentOnline TA"        "distribution_channelDirect"
## [19] "distribution_channelGDS"        "distribution_channelTA/TO"
## [21] "previous_cancellations"         "previous_bookings_not_canceled"
## [23] "reserved_room_typeB"            "reserved_room_typeC"
## [25] "reserved_room_typeD"            "reserved_room_typeE"
## [27] "reserved_room_typeF"            "reserved_room_typeG"
## [29] "assigned_room_typeB"            "assigned_room_typeC"
## [31] "assigned_room_typeD"            "assigned_room_typeE"
## [33] "assigned_room_typeF"            "assigned_room_typeG"
## [35] "assigned_room_typeH"            "assigned_room_typeI"
## [37] "assigned_room_typeK"            "assigned_room_typeL"
## [39] "booking_changes"                "deposit_typeNon Refund"
## [41] "days_in_waiting_list"           "customer_typeGroup"
## [43] "customer_typeTransient"         "adr"
## [45] "required_car_parking_spaces"    "total_of_special_requests"
```

After LASSO regularization, we pick out columns that are significant and run multiple. logistic regression

```
##
## Call:
## lm(formula = is_canceled ~ ., data = data.fl.sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1937 -0.3041 -0.1313  0.3328  1.9342
##
## Coefficients:
```

```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    8.038e-02  9.786e-03   8.214  < 2e-16 ***
## lead_time                      5.792e-04  1.350e-05  42.913  < 2e-16 ***
## arrival_date_monthAugust      -1.354e-02  5.202e-03  -2.603 0.009239 **
## arrival_date_monthDecember     2.719e-02  6.149e-03   4.421 9.82e-06 ***
## arrival_date_monthFebruary     2.034e-02  5.873e-03   3.464 0.000533 ***
## arrival_date_monthJanuary      1.347e-02  6.465e-03   2.084 0.037186 *
## arrival_date_monthJuly        -2.297e-02  5.266e-03  -4.362 1.29e-05 ***
## arrival_date_monthJune        -3.837e-02  5.374e-03  -7.140 9.38e-13 ***
## arrival_date_monthMarch       -2.395e-02  5.524e-03  -4.336 1.45e-05 ***
## arrival_date_monthMay         -2.251e-02  5.249e-03  -4.289 1.80e-05 ***
## arrival_date_monthNovember    -8.179e-03  6.161e-03  -1.327 0.184377
## arrival_date_monthOctober     -8.422e-03  5.341e-03  -1.577 0.114816
## arrival_date_monthSeptember   -3.653e-02  5.438e-03  -6.717 1.87e-11 ***
## stays_in_week_nights           6.342e-03  7.222e-04   8.781  < 2e-16 ***
## stays_in_weekend_nights        5.093e-03  1.342e-03   3.793 0.000149 ***
## adults                         2.857e-02  2.172e-03  13.157  < 2e-16 ***
## children                       4.778e-02  3.840e-03  12.444  < 2e-16 ***
## babies                         3.204e-02  1.186e-02   2.701 0.006906 **
## mealFB                         5.835e-02  1.410e-02   4.137 3.52e-05 ***
## mealHB                        -4.851e-02  3.704e-03 -13.095  < 2e-16 ***
## mealSC                         5.157e-02  4.337e-03  11.892  < 2e-16 ***
## mealUndefined                 -1.008e-01  1.184e-02  -8.510  < 2e-16 ***
## distribution_channelDirect    -7.187e-02  6.104e-03 -11.775  < 2e-16 ***
## distribution_channelGDS       -1.485e-01  2.887e-02  -5.144 2.69e-07 ***
## distribution_channelTA/TO      4.395e-02  5.386e-03   8.160 3.38e-16 ***
## distribution_channelUndefined  1.283e-02  3.935e-01   0.033 0.973987
## reserved_room_typeB            1.461e-01  1.548e-02   9.435  < 2e-16 ***
## reserved_room_typeC            1.462e-01  1.680e-02   8.702  < 2e-16 ***
## reserved_room_typeD            1.544e-01  5.200e-03  29.686  < 2e-16 ***
## reserved_room_typeE            2.162e-01  9.599e-03  22.526  < 2e-16 ***
## reserved_room_typeF            1.944e-01  1.425e-02  13.641  < 2e-16 ***
## reserved_room_typeG            2.711e-01  1.964e-02  13.805  < 2e-16 ***
## reserved_room_typeH            1.761e-01  3.649e-02   4.825 1.40e-06 ***
## reserved_room_typeL            4.126e-02  1.762e-01   0.234 0.814838
## assigned_room_typeB           -1.230e-01  1.113e-02 -11.047  < 2e-16 ***
## assigned_room_typeC           -1.629e-01  1.033e-02 -15.766  < 2e-16 ***
## assigned_room_typeD           -1.720e-01  4.651e-03 -36.990  < 2e-16 ***
## assigned_room_typeE           -2.203e-01  8.757e-03 -25.152  < 2e-16 ***
## assigned_room_typeF           -2.640e-01  1.229e-02 -21.477  < 2e-16 ***
## assigned_room_typeG           -3.091e-01  1.749e-02 -17.677  < 2e-16 ***
## assigned_room_typeH           -1.885e-01  3.341e-02  -5.641 1.69e-08 ***
## assigned_room_typeI           -2.527e-01  2.129e-02 -11.873  < 2e-16 ***
## assigned_room_typeK           -2.472e-01  2.888e-02  -8.561  < 2e-16 ***
## assigned_room_typeL            7.771e-01  4.311e-01   1.803 0.071433 .
## booking_changes               -4.642e-02  1.854e-03 -25.043  < 2e-16 ***
## deposit_typeNon Refund         5.216e-01  4.348e-03 119.955  < 2e-16 ***
## deposit_typeRefundable         5.350e-02  3.114e-02   1.718 0.085846 .
## days_in_waiting_list          -4.136e-04  6.666e-05  -6.205 5.48e-10 ***
## customer_typeGroup            -4.579e-02  1.775e-02  -2.580 0.009892 **
## customer_typeTransient         1.261e-01  6.497e-03  19.404  < 2e-16 ***
## customer_typeTransient-Party   1.354e-02  6.912e-03   1.960 0.050037 .
## adr                            8.163e-04  2.994e-05  27.270  < 2e-16 ***
## required_car_parking_spaces   -2.390e-01  4.836e-03 -49.424  < 2e-16 ***
```

```
## total_of_special_requests     -9.730e-02  1.574e-03 -61.817  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3934 on 118674 degrees of freedom
## Multiple R-squared:  0.3376, Adjusted R-squared:  0.3373
## F-statistic:  1141 on 53 and 118674 DF,  p-value: < 2.2e-16
```

Notice that is_canceled is binary, and 0 means booking was not canceled while 1 means booking was canceled. Therefore, a positive coefficient in front of a variable means that variable increases the number of canceled bookings while a negative coefficient decreases the number of cancelled bookings.

Overall, arrival date month is a significant variable, with June being the most negatively influential month, and the number of cancellations decreases in June. December increases the number of bookings being canceled the most out of all months. From all coefficients, getting assigned room type L has a positive largest coefficient relation to whether the booking is cancelled (number of cancellations increases).

Total of special requests has a negative effect on is_canceled, so an increase in one unit of total of special requests decreases is_canceled by 1.3168 units, so number of cancellations becomes smaller.

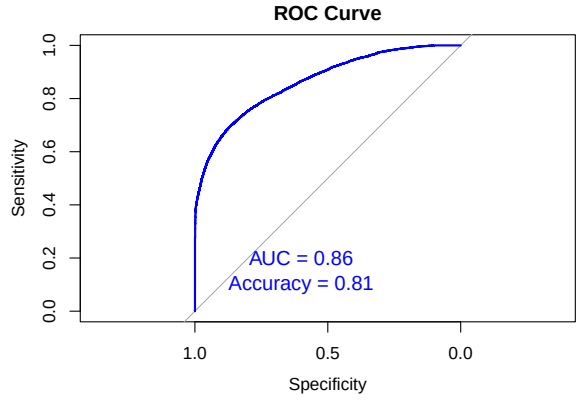
Number of adults, children, and babies increases the number of canceled bookings. This makes sense since the more people there are in a group, the higher the chance that there might be some circumstance at which the family/group of people cannot make it to the hotel anymore.

### Multiple Logistic Regression

In this section, we begin by establishing a baseline for classification models. This logistic regression will serve as a benchmark for comparing the performance of future models. We choose to begin with a logistic regression as it is interpretable and offers reasonable performance when modeling relationships between a binary response variable and one or more predictor variables.

We first create a train test split of 70/30. Next, we run a logistic regression while including most of the features of the original data.

We then plot the accuracy and AUC to assess the model and find that the model has a 0.81 accuracy and 0.84 AUC, which are decently high.



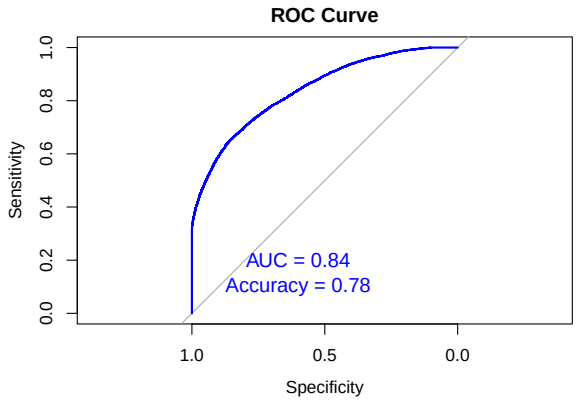

Next, we use LASSO to regularize the model and limit the number of features we select for the logistic regression model. The idea is to develop a parsimonious logistic regression model that can predict cancellations based on a smaller set of features that still explain a large proportion of variability in the data.

```
##
## Call:
## glm(formula = is_canceled ~ ., family = binomial, data = data.fl.sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1545  -0.7998  -0.3853   0.3073   3.5398
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.529e+00  7.048e-02 -35.888  < 2e-16 ***
## lead_time                       4.346e-03  9.511e-05  45.697  < 2e-16 ***
## arrival_date_monthAugust       -1.838e-01  3.350e-02  -5.487 4.10e-08 ***
## arrival_date_monthDecember      2.722e-01  4.086e-02   6.661 2.72e-11 ***
## arrival_date_monthFebruary      2.176e-01  3.889e-02   5.594 2.22e-08 ***
## arrival_date_monthJanuary       1.398e-01  4.429e-02   3.157 0.001595 **
## arrival_date_monthJuly         -2.164e-01  3.332e-02  -6.494 8.38e-11 ***
## arrival_date_monthJune         -2.543e-01  3.526e-02  -7.213 5.49e-13 ***
## arrival_date_monthMarch        -8.794e-02  3.691e-02  -2.383 0.017194 *
## arrival_date_monthMay          -1.554e-01  3.425e-02  -4.538 5.69e-06 ***
## arrival_date_monthNovember      1.079e-01  4.203e-02   2.567 0.010264 *
## arrival_date_monthOctober      -1.452e-02  3.592e-02  -0.404 0.685996
## arrival_date_monthSeptember    -2.171e-01  3.685e-02  -5.893 3.80e-09 ***
## stays_in_week_nights            2.829e-02  4.578e-03   6.180 6.39e-10 ***
## stays_in_weekend_nights         2.469e-02  8.673e-03   2.847 0.004414 **
## adults                          1.670e-01  1.682e-02   9.929  < 2e-16 ***
## children                        2.172e-01  2.450e-02   8.864  < 2e-16 ***
## babies                          1.884e-01  8.257e-02   2.282 0.022513 *
## mealFB                          3.832e-01  1.062e-01   3.608 0.000309 ***
## mealHB                         -3.823e-01  2.513e-02 -15.213  < 2e-16 ***
## mealSC                          2.098e-01  2.486e-02   8.439  < 2e-16 ***
## mealUndefined                  -8.458e-01  9.689e-02  -8.729  < 2e-16 ***
## distribution_channelDirect     -4.647e-01  4.676e-02  -9.938  < 2e-16 ***
## distribution_channelGDS        -7.560e-01  1.900e-01  -3.979 6.92e-05 ***
## distribution_channelTA/TO       3.344e-01  4.060e-02   8.236  < 2e-16 ***
## distribution_channelUndefined   5.099e-01  3.956e+03   0.000 0.999897
## reserved_room_typeB             9.407e-01  1.001e-01   9.399  < 2e-16 ***
## reserved_room_typeC             1.367e+00  1.305e-01  10.477  < 2e-16 ***
## reserved_room_typeD             1.339e+00  4.482e-02  29.869  < 2e-16 ***
## reserved_room_typeE             2.188e+00  9.346e-02  23.413  < 2e-16 ***
## reserved_room_typeF             2.272e+00  1.366e-01  16.629  < 2e-16 ***
## reserved_room_typeG             3.205e+00  2.022e-01  15.848  < 2e-16 ***
## reserved_room_typeH             2.162e+00  4.526e-01   4.777 1.78e-06 ***
## reserved_room_typeL             9.131e-01  1.162e+00   0.786 0.432074
## assigned_room_typeB            -7.833e-01  7.983e-02  -9.812  < 2e-16 ***
## assigned_room_typeC            -1.403e+00  9.990e-02 -14.041  < 2e-16 ***
## assigned_room_typeD            -1.509e+00  4.272e-02 -35.313  < 2e-16 ***
## assigned_room_typeE            -2.255e+00  9.008e-02 -25.033  < 2e-16 ***
## assigned_room_typeF            -2.844e+00  1.291e-01 -22.023  < 2e-16 ***
## assigned_room_typeG            -3.539e+00  1.960e-01 -18.053  < 2e-16 ***
## assigned_room_typeH            -2.284e+00  4.412e-01  -5.177 2.26e-07 ***
## assigned_room_typeI            -4.416e+00  4.942e-01  -8.936  < 2e-16 ***
## assigned_room_typeK            -2.648e+00  3.492e-01  -7.584 3.34e-14 ***
## assigned_room_typeL             1.865e+01  3.956e+03   0.005 0.996239
```

```
## booking_changes               -3.732e-01  1.536e-02 -24.299  < 2e-16 ***
## deposit_typeNon Refund         5.080e+00  1.066e-01  47.646  < 2e-16 ***
## deposit_typeRefundable         3.393e-01  2.121e-01   1.600 0.109601
## days_in_waiting_list          -3.074e-03  4.850e-04  -6.339 2.32e-10 ***
## customer_typeGroup            -5.980e-01  1.634e-01  -3.660 0.000253 ***
## customer_typeTransient         8.009e-01  4.631e-02  17.295  < 2e-16 ***
## customer_typeTransient-Party   7.591e-02  4.902e-02   1.549 0.121496
## adr                            6.780e-03  2.244e-04  30.216  < 2e-16 ***
## required_car_parking_spaces   -1.640e+01  4.166e+01  -0.394 0.693806
## total_of_special_requests     -6.216e-01  1.090e-02 -57.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 156693  on 118727  degrees of freedom
## Residual deviance: 105790  on 118674  degrees of freedom
## AIC: 105898
##
## Number of Fisher Scoring iterations: 16
```

### LASSO + Multiple Logistic Regression

We find that after regularization and multiple logistic regression, we can predict booking cancellations with an accuracy of 0.78 and an AUC of 0.84. This is not too much lower than our original model that included all features. We may assume that the predictor variables are all highly relevant for predicting the response variable, which may explain why Lasso regularization did not improve model performance significantly.

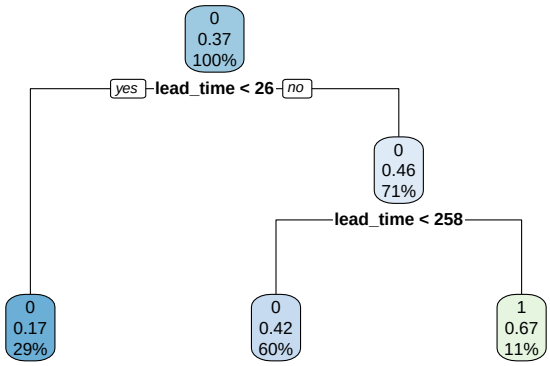



### Regression Trees

We begin by generating a train test split.

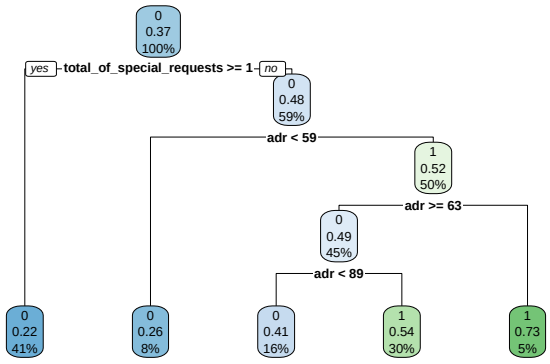
### Single predictor (lead_time)

```
## Accuracy: 0.6647762
```




We begin by fitting a regression tree on a single factor, `lead_time`, so to see how well a decision tree can predict hotel cancellations with limited information. We find that the accuracy of this model is around 0.66, which is low as expected.

### Two predictors (total_of_special_requests and adr)

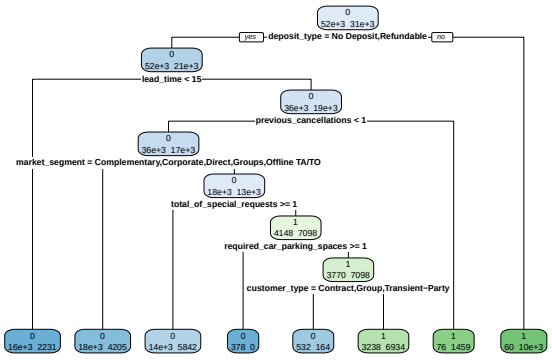



```
## Accuracy: 0.6760346
```

Arbitrarily we selected two factors, `total_of_special_requests` and `adr`, to see if a regression tree can make accurate predictions with just two factors and the interactions between them. The accuracy of this model is 0.67, which is a slight improvement over our previous single factor model.

### All avaiable predictors

```
## Accuracy: 0.8101241
```




Finally, we feet a decision tree over all available predictors. This final tree had the highest accuracy among the regression trees so far, with an accuracy of 0.81. Within this chart we can get a sense of the most important predictors (e.g. `no refunds`, `lead_time`, `previous cancellations`), as well as interactions between factors.

### Random Forest

```
## Accuracy: 0.8695603
```

We also decide to fit a random forest model on our data. Because we do not specify `ntree` or `mtry`, the randomForest function's default values of 500 trees to grow and the square root of the number of predictors as the number of variables randomly sampled at each split of the tree are used. We find that this model has an accuracy of almost 0.87!

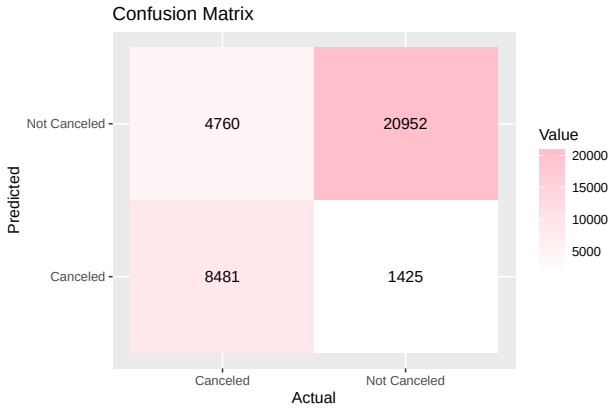
### XGBoost Tree

In this final section, we look to use XG boosting, which is a form of gradient boosting, in order to generate a high accuracy predictive model. XGBoosting iteratively builds decision trees to minimize the loss function, using gradient boosting to correct errors made by previous models. In the following model, we use XGBoost with logistic regression through the "binary:logistic" objective function. This allows us to use boosting for our hotel cancellation binary classification question.

```
## Confusion Matrix and Statistics
##
##
## pred_factor     0     1
##           0 20952  4760
##           1  1425  8481
##
##                Accuracy : 0.8264
##                  95% CI : (0.8224, 0.8303)
##     No Information Rate : 0.6282
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6081
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9363
```

```
##             Specificity : 0.6405
##          Pos Pred Value : 0.8149
##          Neg Pred Value : 0.8561
##              Prevalence : 0.6282
##          Detection Rate : 0.5882
##    Detection Prevalence : 0.7219
##       Balanced Accuracy : 0.7884
##
##        'Positive' Class : 0
##
```

We see that with XGBoosting, we can create a model that has high sensitivity but lower specificity. The overall accuracy of our XGBoost model is 0.825, which is a slight increase from our initial multiple logistic regression. Graphed below is the confusion matrix, where we can see the true positive, true negative, and false positive/negative frequencies/rates.




## Conclusions

Given the accuracy scores for each model, we find that Random Forest yields the highest accuracy score at 0.86 when compared to other models. In addition to having the highest accuracy score, we believe this model is favorable when compared to the other models we used. Unlike some other models (ex. LASSO), Random Forest does not overfit data as a result of using various trees and random sampling. Furthermore, this model was able to accommodate our data set with categorical and numerical data. However, Random Forest requires hyperparameter tuning and little control over results. XGBoost computed the second highest accuracy score. This technique is also less prone to overfitting and is useful when determining variables significant in making predictions. Still, XGBoost is more difficult to interpret/visualize than Random Forest and can lead to overfitting if not tuned adequately. When compared to other models, LASSO may have calculated a lower accuracy score because of small coefficients that were excluded from the model and correlation between certain variables.

## Further Extensions

In finding the best parameters in Random Forest, we were mindful of the impact of hyperparameter tuning, which is both computationally intensive and can lead to overfitting. We were careful in reviewing variables in this data set to limit overfitting by eliminating irrelevant/redundant variables. However, we are aware that this may be insufficient in mitigating negative impacts and our results may be subject to overfitting. As for other models that could have been considered, neural nets would have been a promising avenue for our analysis. Because neural networks have the ability to handle various data types and handle complex/incomplete data, neural nets would have been helpful if not for the potential challenges in manipulating categorical data.