A Hotel Cancellation Prediction Tool

TEAM #64 PROJECT PROPOSAL

Team: **Patil, Rutuja** Liu, Yuxin Hui, Fei van Keulen, Alessio

Contents

Introduction

Overview

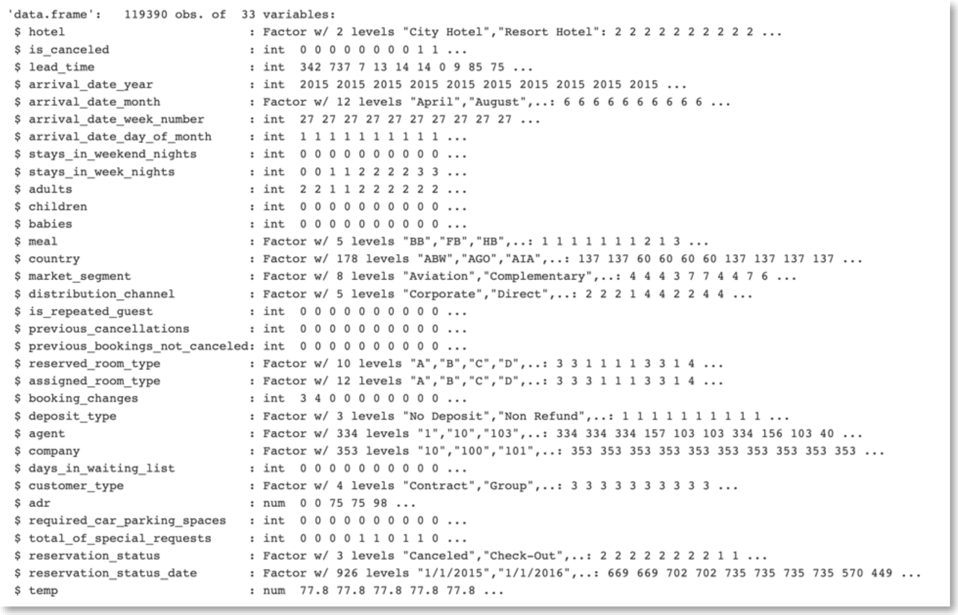
With the advent and outreach of cellular web-technologies including ecommerce marketplaces, hotel booking platforms, and travel websites, the hospitality industry has seen major revolutions in terms of market exposure and capitalization. In the last decade, this revolution is especially noticeable in the lodging sector, where existing hotels, bed & breakfasts, inns, and other similar businesses have been witnessing an expanding portfolio of alternative and more affordable solutions for travelers in seek for a temporary accommodation. Other than the inherent increase in the competition, hotel managers and concierge, must deal with a more concrete problem: increasing reservation cancellations patterns.

Objective & Problem Statement

This study attempts to describe and understand the underlying forces that drive reservation cancellation decisions by customers toward their original lodging choices. Leveraging data analytics tools and practices, this analysis distillated those factors that accompany a customer throughout their travel experience, into an essential set of variables that will predict future reservation and cancellation patterns with similar characteristics. Finally, based on these observations, the project serves as an effective prescription tool for optimizing vacancies and profits.

Data Overview

Data preprocessing was a crucial step in our study. It helped enhance the quality of data to promote the extraction of meaningful insights. This technique involved preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. The raw data included 33 variables, both numerical and categorical variables shown below.



```
'data.frame':   119390 obs. of  33 variables:
 $ hotel                         : Factor w/ 2 levels "City Hotel","Resort Hotel": 2 2 2 2 2 2 2 2 2 2 ...
 $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
 $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
 $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ arrival_date_month            : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27 27 ...
 $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
 $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
 $ children                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ meal                          : Factor w/ 5 levels "BB","FB","HB",..: 1 1 1 1 1 1 1 2 1 3 ...
 $ country                       : Factor w/ 178 levels "ABW","AGO","AIA",..: 137 137 60 60 60 60 137 137 137 137 ...
 $ market_segment                : Factor w/ 8 levels "Aviation","Complementary",..: 4 4 4 3 7 7 4 4 7 6 ...
 $ distribution_channel          : Factor w/ 5 levels "Corporate","Direct",..: 2 2 2 1 4 4 2 2 4 4 ...
 $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
 $ reserved_room_type            : Factor w/ 10 levels "A","B","C","D",..: 3 3 1 1 1 1 3 3 1 4 ...
 $ assigned_room_type            : Factor w/ 12 levels "A","B","C","D",..: 3 3 3 1 1 1 3 3 1 4 ...
 $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
 $ deposit_type                  : Factor w/ 3 levels "No Deposit","Non Refund",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ agent                         : Factor w/ 334 levels "1","10","103",..: 334 334 334 157 103 103 334 156 103 40 ...
 $ company                       : Factor w/ 353 levels "10","100","101",..: 353 353 353 353 353 353 353 353 353 353 ...
 $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ customer_type                 : Factor w/ 4 levels "Contract","Group",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ adr                           : num  0 0 75 75 98 ...
 $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
 $ reservation_status            : Factor w/ 3 levels "Canceled","Check-Out",..: 2 2 2 2 2 2 2 2 1 1 ...
 $ reservation_status_date       : Factor w/ 926 levels "1/1/2015","1/1/2016",..: 669 669 702 702 735 735 735 735 570 449 ...
 $ temp                          : num  77.8 77.8 77.8 77.8 77.8 ...
```

Figure 1

Data preprocessing

Distribution Analysis

Distribution analysis started with the visualization of the distribution for each variable in our data set. This was a good technique to get an immediate understanding of how data flows through the sample period (2015-2017), and it could hint at trends and possible cyclical patterns.

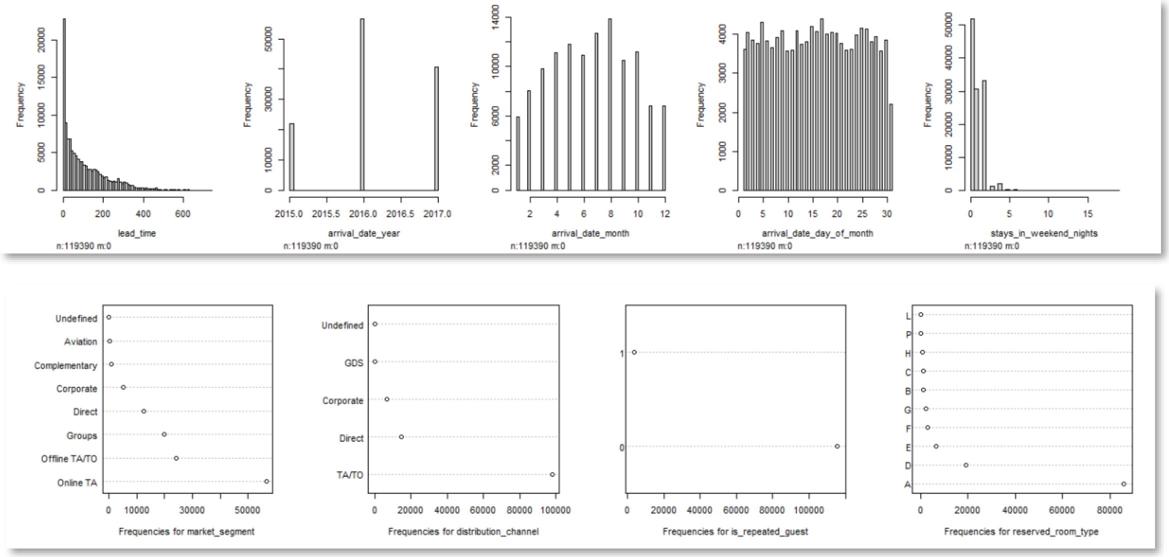


Figure 2

Outlier Analysis (Univariate)

Outlier Analysis helped us remove observation that can potentially skew or bias the results in our study. The easiest way to detect outliers was by using visualization techniques. Note that outlier detection required a qualitative approach before deciding whether the data set needs to be discarded. Univariate Outlier Analysis involved addressing one variable at a time, free from any relationship they may have with other variables. This relationship would become apparent only after a statistical model was built, and for this reason this outlier analysis step was considered At Priori: before the actual study takes place.

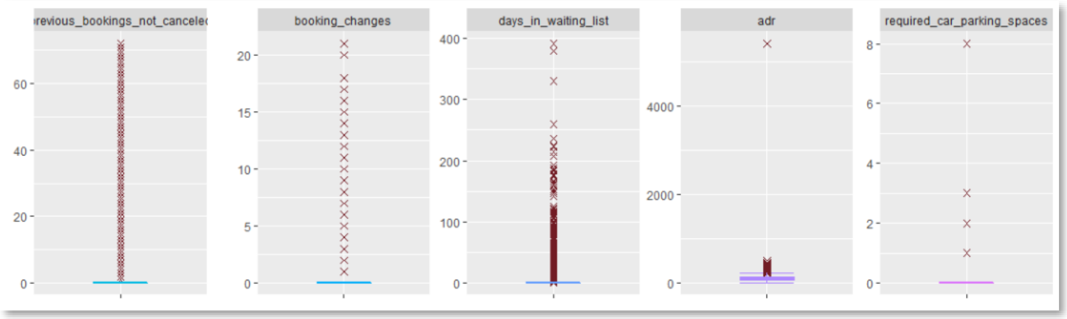
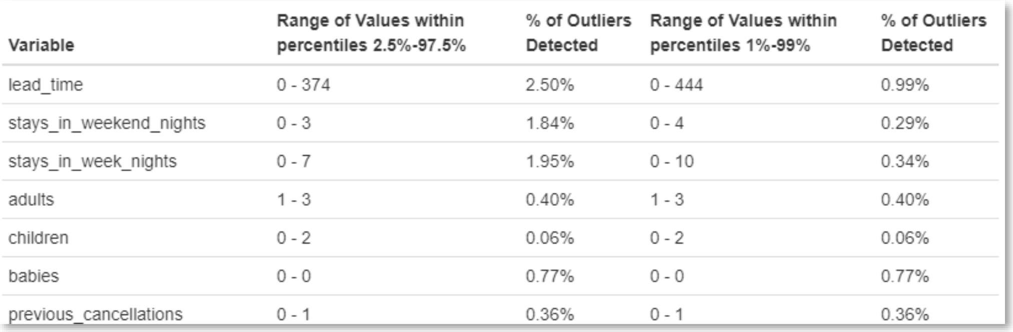


Figure 3

After a first glance, it looked like many of these variables have observations that are found outside the whiskers for each boxplot. This would suggest that the data point is an outlier. However, a more qualitative approach was required when interpreting outliers. In the next step addressed outliers using a combination of the Percentile method and some educated guessing based on the underlying significance of each variable.



Variable	Range of Values within percentiles 2.5%-97.5%	% of Outliers Detected	Range of Values within percentiles 1%-99%	% of Outliers Detected
lead_time	0 - 374	2.50%	0 - 444	0.99%
stays_in_weekend_nights	0 - 3	1.84%	0 - 4	0.29%
stays_in_week_nights	0 - 7	1.95%	0 - 10	0.34%
adults	1 - 3	0.40%	1 - 3	0.40%
children	0 - 2	0.06%	0 - 2	0.06%
babies	0 - 0	0.77%	0 - 0	0.77%
previous_cancellations	0 - 1	0.36%	0 - 1	0.36%

Figure 4

The table above outlines how we constructed two different percentile ranges, each yielding different outlier results. We chose to adopt the second, stricter range: anything within 1 and 99 % of the distribution is deemed valid, whereas anything beyond that is considered an outlier.

Note that the percentile method only really applied to numerical variables with a large range. This wouldn't be affective with small categorical variables that have a reduced number of levels. In that case we adjusted the percentiles range manually or we have adopted a qualitative approach with the aid of the visual plots in the previous step.

A result of this combined procedure can be seen here:

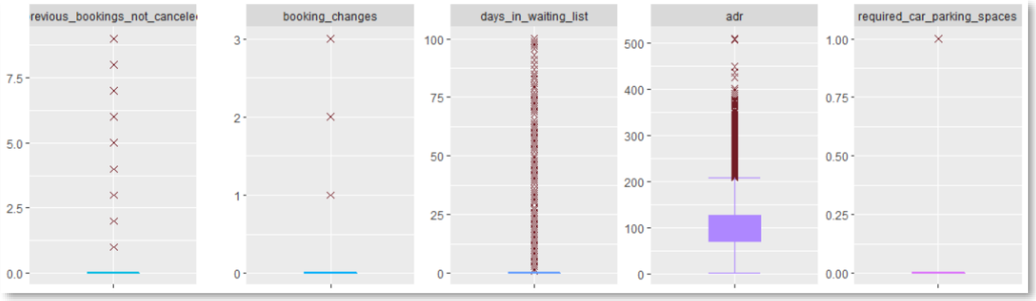


Figure 5

Missing value imputation

After addressing obvious outliers, our next step was to fill in missing values in any in our variables. The outcome of this was going to be more precise as outliers could have skewed our biased our data imputation techniques. As usual we will begin with some visual aid:

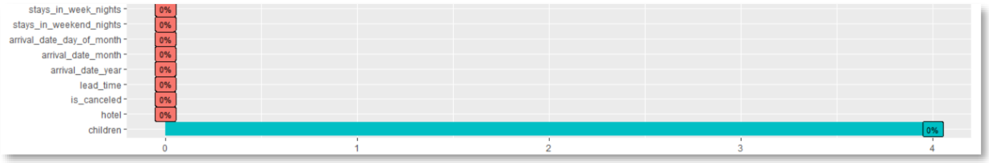


Figure 6

It emerged that the only variable which contained missing values is the *children* variable. *Children* was of numerical data type INT and it was discrete. Therefore, the proper imputation technique in this case would be the mode of the other values in the distribution.

Multicollinearity

Our last step in the data-preparation phase involves the detection of possible multicollinearity between predictor variables, that is a relationship between independent variables. As usual we will deploy some visual tools to get an immediate understanding of the problem at hand. In this case we will draw a correlation matrix plot on a pair of variables. We will start by using an educated guess as to which variables could be correlated (Figure 7), and integrate this will a more in-depth approach toward multicollinearity detection using Cramer's V statistical approach (Figure 8):

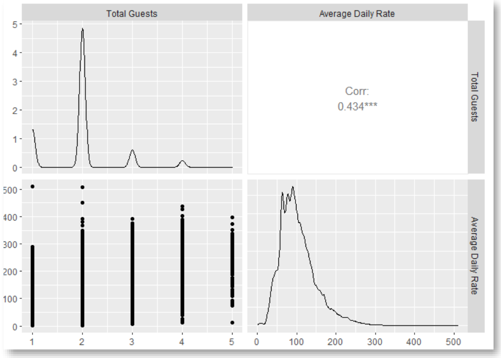


Figure 7

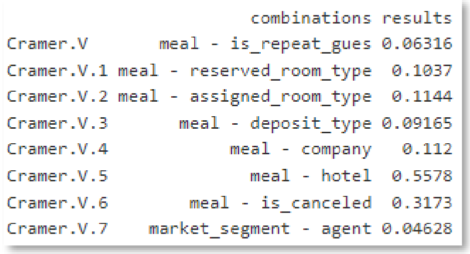

```
                              combinations results
Cramer.V       meal - is_repeat_gues 0.06316
Cramer.V.1 meal - reserved_room_type  0.1037
Cramer.V.2 meal - assigned_room_type  0.1144
Cramer.V.3        meal - deposit_type 0.09165
Cramer.V.4             meal - company   0.112
Cramer.V.5               meal - hotel  0.5578
Cramer.V.6         meal - is_canceled  0.3173
Cramer.V.7    market_segment - agent 0.04628
```

Figure 8

From the study it emerged that no pair of variables displayed a coefficient significant enough to suggest multicollinearity.

Data modeling

Data Approach & Methodology

Our study attempts to describe reservation cancellation patterns, via descriptive and predictive analytics. Since the outcome of our response variable was binary "Canceled: 1" or "Not Canceled: 0", we leveraged logistic regression.

As previously observed, our datasets included 32 variables, and many categorical variables includes 10+ levels. Due to the complexity of our dataset, simply running automated variable selection functions, such as stepwise or random forest, would not generate correct outputs. Having too many predictors in the model not only makes it difficult to interpret the result, but also easily overfits the model.

Therefore, our approach for this analysis included 5 steps:

1. Observe the dataset and remove any unnecessary variables
2. Run forward stepwise AIC regression function to further eliminate predictors for resort hotels and city hotels separately
3. Observe the coefficients, ANOVA tables, and further reduce levels of categorical variables
4. Fit the desired variables with logistic regression on training data
5. Validate the accuracy of the final model on testing data

At the end of the analysis, we hoped to find a set of essential factors that affect the hotel reservation cancellation for future reference.

Variable Selection

This step was to perform some qualitative analysis and understand how each variable serves our purpose of study. Our methodology on variable selection included the following steps:

Firstly, we removed time series column *reservation_status_date* as this could be done separately in the future. We also removed *company* columns because it contained 94% Null and 6% Other which would not make any use recommendations.

Secondly, we removed some variables have repeated information with others:

reservation_status	is identical to the response variable *is_canceled* after observation, so we removed it
distribution_channel	Is identical to *market_segment,* so we removed it
arrival_date_year *arrival_week_number* *arrival_date_day_of_month*	have an intrinsic seasonality meaning, which can be explained just through the *arrival_date_month* variable, so we have removed them
previous_bookings_not_canceled	has very similar information as *previous_bookings_cancelled,* and for the purpose of predicting cancellation patterns, only cancelled reservations are relevant to our analysis

Thirdly, we combined some variables into a more representable variable as below:

childrens *babies*	will be combined into a new variable: *total_children*
stays_in_week_nights *stays_in_weekend_nights*	will be combined into a new variable: *total_nights_stayed*
arrival_date_year *arrival_week_number* *arrival_date_day_of_month*	have an intrinsic seasonality meaning, which can be explained just through the *arrival_date_month* variable, so we have removed them
assigned_room_type *reserved_room_type*	will be combined into a new variable: *mismatched_room* indicating whether the room had been changed (1) or not (0)

Finally, we grouped some levels of categorical variables by investigating their distributions, and further reduced the dimensions in data:

Agent_group	contains 4 groups: agent #9, #240, null, others
Special_request_group	2 groups: no special request (0); One or more special requests (1+)
Booking_change	2 groups: no booking changes (0); One or more changes (1+)
Children_group	2 groups: no children (0); one or more children (1+)
Meal_group	3 groups: Bed and Breadfast (BB); Half Board (HB); Full Board (FB)

Thus far, we grouped most of the categorical variables within 4 levels and reduced the number of predictors from 32 to 21.

Model Fitting

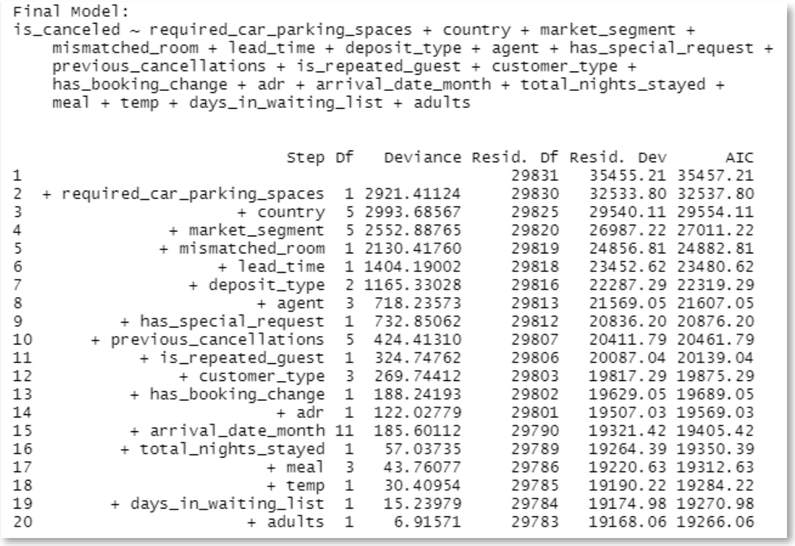
Resort Hotel

After we explored variable selection from a qualitative point of view, we then deployed some statistical machine learning algorithms to help us further identify strong vs weak predictors. We separated the data into Resort vs. City hotel dataset since we want to know if cancellation behavior varies by different types of hotels. And for each subset, we used 80% for model training, and 20% for model testing purposes.

Logistic Regression was used since it is a commonly used model for classification problems.

Then we used Forward Stepwise AIC Regression. Stepwise AIC regression will automatically select one variable at a time to add to the model for AIC improvement. It stopped adding variables until there were no more AIC improvements.

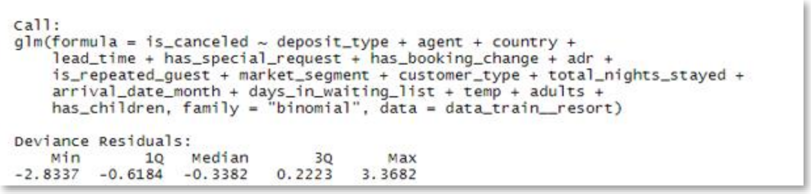
Note that one constraint of forward method is that data cannot have multicollinearity, or else it will remove the relationships. In our prior finding, we did not find any multicollinearity in our prior analysis. Stepwise AIC suggests the following variables to be included in the final mode below in the ANOVA table.



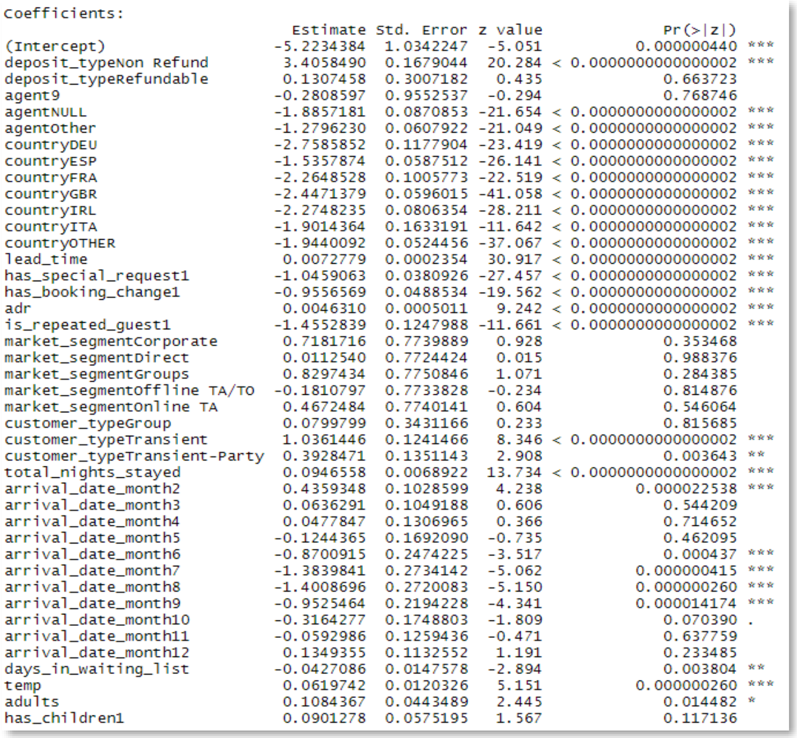
```
Final Model:
is_canceled ~ required_car_parking_spaces + country + market_segment +
    mismatched_room + lead_time + deposit_type + agent + has_special_request +
    previous_cancellations + is_repeated_guest + customer_type +
    has_booking_change + adr + arrival_date_month + total_nights_stayed +
    meal + temp + days_in_waiting_list + adults


                          Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                                29831   35455.21 35457.21
2  + required_car_parking_spaces  1 2921.41124     29830   32533.80 32537.80
3                    + country  5 2993.68567     29825   29540.11 29554.11
4             + market_segment  5 2552.88765     29820   26987.22 27011.22
5            + mismatched_room  1 2130.41760     29819   24856.81 24882.81
6                  + lead_time  1 1404.19002     29818   23452.62 23480.62
7               + deposit_type  2 1165.33028     29816   22287.29 22319.29
8                      + agent  3  718.23573     29813   21569.05 21607.05
9        + has_special_request  1  732.85062     29812   20836.20 20876.20
10    + previous_cancellations  5  424.41310     29807   20411.79 20461.79
11         + is_repeated_guest  1  324.74762     29806   20087.04 20139.04
12             + customer_type  3  269.74412     29803   19817.29 19875.29
13        + has_booking_change  1  188.24193     29802   19629.05 19689.05
14                       + adr  1  122.02779     29801   19507.03 19569.03
15        + arrival_date_month 11  185.60112     29790   19321.42 19405.42
16       + total_nights_stayed  1   57.03735     29789   19264.39 19350.39
17                      + meal  3   43.76077     29786   19220.63 19312.63
18                      + temp  1   30.40954     29785   19190.22 19284.22
19      + days_in_waiting_list  1   15.23979     29784   19174.98 19270.98
20                    + adults  1    6.91571     29783   19168.06 19266.06
```

Figure 9

We used the variables emerged from the StepAIC procedure and fed them into our new logistic regression model.



```
Call:
glm(formula = is_canceled ~ deposit_type + agent + country +
    lead_time + has_special_request + has_booking_change + adr +
    is_repeated_guest + market_segment + customer_type + total_nights_stayed +
    arrival_date_month + days_in_waiting_list + temp + adults +
    has_children, family = "binomial", data = data_train__resort)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8337  -0.6184  -0.3382   0.2223   3.3682
```



```
Coefficients:
                               Estimate Std. Error z value             Pr(>|z|)    
(Intercept)                  -5.2234384  1.0342247  -5.051          0.000000440 ***
deposit_typeNon Refund        3.4058490  0.1679044  20.284 < 0.0000000000000002 ***
deposit_typeRefundable        0.1307458  0.3007182   0.435             0.663723    
agent9                       -0.2808597  0.9552537  -0.294             0.768746    
agentNULL                    -1.8857181  0.0870853 -21.654 < 0.0000000000000002 ***
agentOther                   -1.2796230  0.0607922 -21.049 < 0.0000000000000002 ***
countryDEU                   -2.7585852  0.1177904 -23.419 < 0.0000000000000002 ***
countryESP                   -1.5357874  0.0587512 -26.141 < 0.0000000000000002 ***
countryFRA                   -2.2648528  0.1005773 -22.519 < 0.0000000000000002 ***
countryGBR                   -2.4471379  0.0596015 -41.058 < 0.0000000000000002 ***
countryIRL                   -2.2748235  0.0806354 -28.211 < 0.0000000000000002 ***
countryITA                   -1.9014364  0.1633191 -11.642 < 0.0000000000000002 ***
countryOTHER                 -1.9440092  0.0524456 -37.067 < 0.0000000000000002 ***
lead_time                     0.0072779  0.0002354  30.917 < 0.0000000000000002 ***
has_special_request1         -1.0459063  0.0380926 -27.457 < 0.0000000000000002 ***
has_booking_change1          -0.9556569  0.0488534 -19.562 < 0.0000000000000002 ***
adr                           0.0046310  0.0005011   9.242 < 0.0000000000000002 ***
is_repeated_guest1           -1.4552839  0.1247988 -11.661 < 0.0000000000000002 ***
market_segmentCorporate       0.7181716  0.7739889   0.928             0.353468    
market_segmentDirect          0.0112540  0.7724424   0.015             0.988376    
market_segmentGroups          0.8297434  0.7750846   1.071             0.284385    
market_segmentOffline TA/TO  -0.1810797  0.7733828  -0.234             0.814876    
market_segmentOnline TA       0.4672484  0.7740141   0.604             0.546064    
customer_typeGroup            0.0799799  0.3431166   0.233             0.815685    
customer_typeTransient        1.0361446  0.1241466   8.346 < 0.0000000000000002 ***
customer_typeTransient-Party  0.3928471  0.1351143   2.908             0.003643 ** 
total_nights_stayed           0.0946558  0.0068922  13.734 < 0.0000000000000002 ***
arrival_date_month2           0.4359348  0.1028599   4.238          0.000022538 ***
arrival_date_month3           0.0636291  0.1049188   0.606             0.544209    
arrival_date_month4           0.0477847  0.1306965   0.366             0.714652    
arrival_date_month5          -0.1244365  0.1692090  -0.735             0.462095    
arrival_date_month6          -0.8700915  0.2474225  -3.517             0.000437 ***
arrival_date_month7          -1.3839841  0.2734142  -5.062          0.000000415 ***
arrival_date_month8          -1.4008696  0.2720083  -5.150          0.000000260 ***
arrival_date_month9          -0.9525464  0.2194228  -4.341          0.000014174 ***
arrival_date_month10         -0.3164277  0.1748803  -1.809             0.070390 .  
arrival_date_month11         -0.0592986  0.1259436  -0.471             0.637759    
arrival_date_month12          0.1349355  0.1132552   1.191             0.233485    
days_in_waiting_list         -0.0427086  0.0147578  -2.894             0.003804 ** 
temp                          0.0619742  0.0120326   5.151          0.000000260 ***
adults                        0.1084367  0.0443489   2.445             0.014482 *  
has_children1                 0.0901278  0.0575195   1.567             0.117136    
```

Figure 10

market_segment	shows a p-value that is not significant enough to reject the Null-Hypothesis
customer_typeTransient	has higher odds of cancellation compared to the other levels in the variable
agent9	is not statically different from 240 (which is the baseline) so they can be grouped.
Country	has 8 level, and the level can be reduced further by looking at the 95% Confidence Interval (CI) chart (see below) • CIs that are far apart from each other indicated that the estimated coefficients are different from • Cis that overlap indicates the coefficients are not statistically different from each other
arrival_date_month3 *arrival_date_month4* *arrival_date_month5* *arrival_date_month10* *arrival_date_month11* *arrival_date_month1*	corresponding to the months of March, April, May, October, November, and December, do not show statistical significance, and they can be grouped with the baseline (January)

For what concerns the country variable, we constructed the 95% confidence interval chart to see if the estimated means are uniquely different from each other.

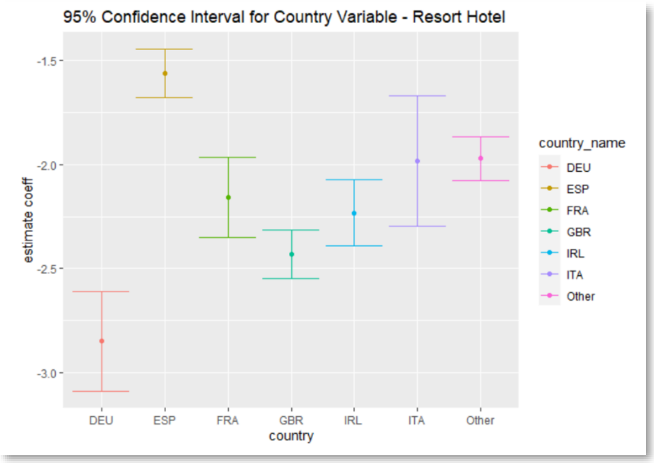
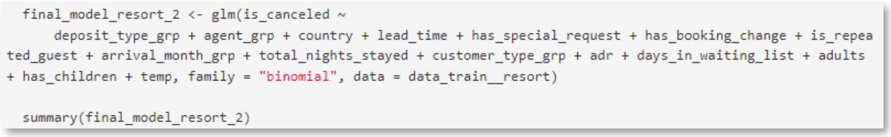


Figure 11

This plot showed that:

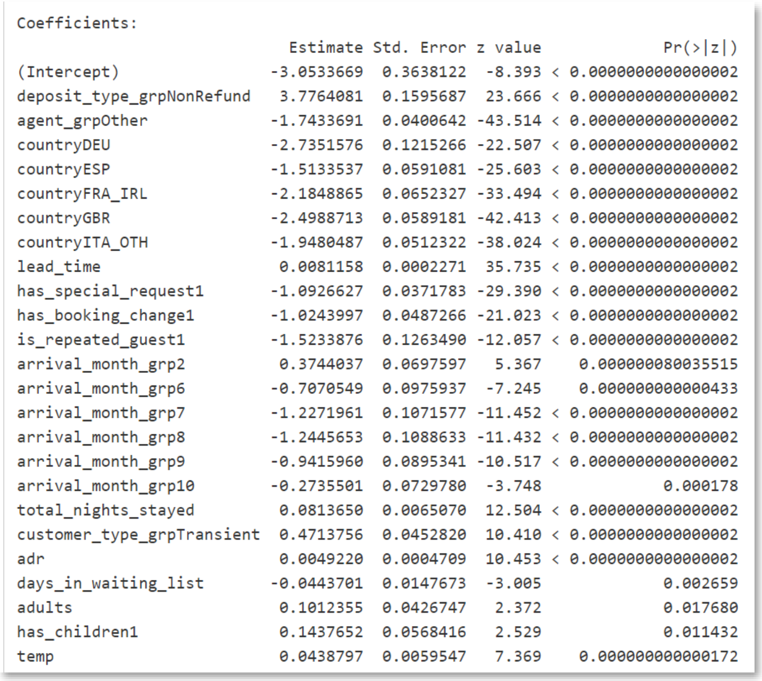
- DEU, ESP do not overlap with others, so their coefficients are statistically
- FRA and IRL overlap with each other, so their coefficients are not statistically different
- GRB will be considered by itself
- ITA and Other overlapped, so their coefficients are not statistically different

After further grouping the variables and the removal of insignificant variables, we trained the model again, and repeated the same process until we identified no more grouping and removal needs.

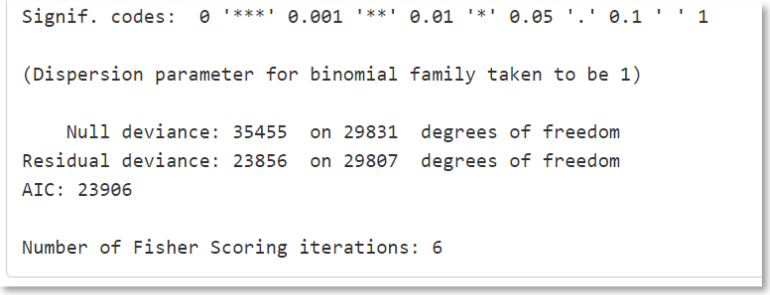


```
  final_model_resort_2 <- glm(is_canceled ~
      deposit_type_grp + agent_grp + country + lead_time + has_special_request + has_booking_change + is_repea
ted_guest + arrival_month_grp + total_nights_stayed + customer_type_grp + adr + days_in_waiting_list + adults
+ has_children + temp, family = "binomial", data = data_train__resort)

  summary(final_model_resort_2)
```

Figure 12



```
Coefficients:
                          Estimate Std. Error z value           Pr(>|z|)
(Intercept)             -3.0533669  0.3638122  -8.393 < 0.0000000000000002
deposit_type_grpNonRefund  3.7764081  0.1595687  23.666 < 0.0000000000000002
agent_grpOther          -1.7433691  0.0400642 -43.514 < 0.0000000000000002
countryDEU              -2.7351576  0.1215266 -22.507 < 0.0000000000000002
countryESP              -1.5133537  0.0591081 -25.603 < 0.0000000000000002
countryFRA_IRL          -2.1848865  0.0652327 -33.494 < 0.0000000000000002
countryGBR              -2.4988713  0.0589181 -42.413 < 0.0000000000000002
countryITA_OTH          -1.9480487  0.0512322 -38.024 < 0.0000000000000002
lead_time                0.0081158  0.0002271  35.735 < 0.0000000000000002
has_special_request1    -1.0926627  0.0371783 -29.390 < 0.0000000000000002
has_booking_change1     -1.0243997  0.0487266 -21.023 < 0.0000000000000002
is_repeated_guest1      -1.5233876  0.1263490 -12.057 < 0.0000000000000002
arrival_month_grp2       0.3744037  0.0697597   5.367   0.000000080035515
arrival_month_grp6      -0.7070549  0.0975937  -7.245   0.000000000000433
arrival_month_grp7      -1.2271961  0.1071577 -11.452 < 0.0000000000000002
arrival_month_grp8      -1.2445653  0.1088633 -11.432 < 0.0000000000000002
arrival_month_grp9      -0.9415960  0.0895341 -10.517 < 0.0000000000000002
arrival_month_grp10     -0.2735501  0.0729780  -3.748            0.000178
total_nights_stayed      0.0813650  0.0065070  12.504 < 0.0000000000000002
customer_type_grpTransient  0.4713756  0.0452820  10.410 < 0.0000000000000002
adr                      0.0049220  0.0004709  10.453 < 0.0000000000000002
days_in_waiting_list    -0.0443701  0.0147673  -3.005            0.002659
adults                   0.1012355  0.0426747   2.372            0.017680
has_children1            0.1437652  0.0568416   2.529            0.011432
temp                     0.0438797  0.0059547   7.369   0.000000000000172
```

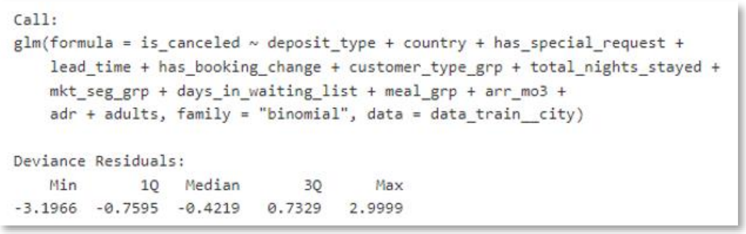


```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35455  on 29831  degrees of freedom
Residual deviance: 23856  on 29807  degrees of freedom
AIC: 23906

Number of Fisher Scoring iterations: 6
```

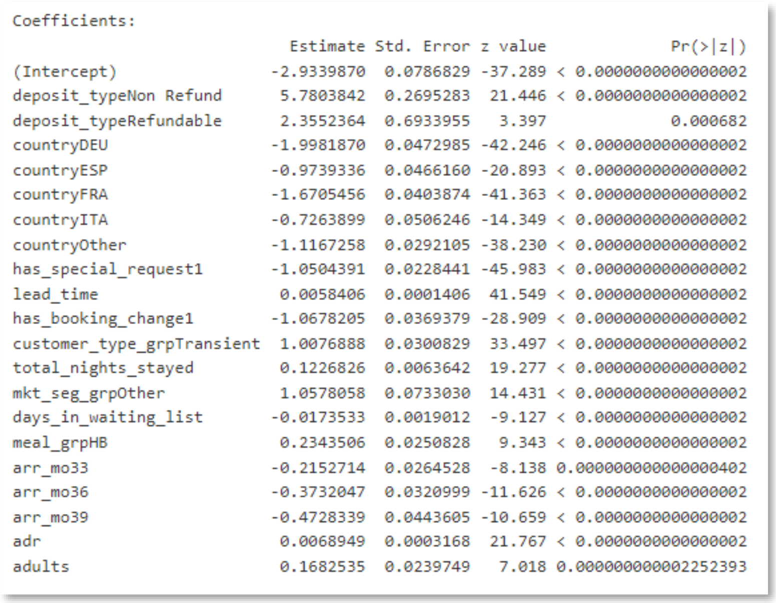
Figure 13

City Hotel

Similar approaches were taken for fitting the City Hotel model. The following variables were selected in the final model. Overall, the same variables used in the City Hotel are also being used in the final Resort Hotel model.



```
Call:
glm(formula = is_canceled ~ deposit_type + country + has_special_request +
    lead_time + has_booking_change + customer_type_grp + total_nights_stayed +
    mkt_seg_grp + days_in_waiting_list + meal_grp + arr_mo3 +
    adr + adults, family = "binomial", data = data_train__city)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1966  -0.7595  -0.4219   0.7329   2.9999
```



```
Coefficients:
                            Estimate Std. Error z value          Pr(>|z|)
(Intercept)               -2.9339870  0.0786829 -37.289 < 0.0000000000000002
deposit_typeNon Refund     5.7803842  0.2695283  21.446 < 0.0000000000000002
deposit_typeRefundable     2.3552364  0.6933955   3.397             0.000682
countryDEU                -1.9981870  0.0472985 -42.246 < 0.0000000000000002
countryESP                -0.9739336  0.0466160 -20.893 < 0.0000000000000002
countryFRA                -1.6705456  0.0403874 -41.363 < 0.0000000000000002
countryITA                -0.7263899  0.0506246 -14.349 < 0.0000000000000002
countryOther              -1.1167258  0.0292105 -38.230 < 0.0000000000000002
has_special_request1      -1.0504391  0.0228441 -45.983 < 0.0000000000000002
lead_time                  0.0058406  0.0001406  41.549 < 0.0000000000000002
has_booking_change1       -1.0678205  0.0369379 -28.909 < 0.0000000000000002
customer_type_grpTransient 1.0076888  0.0300829  33.497 < 0.0000000000000002
total_nights_stayed        0.1226826  0.0063642  19.277 < 0.0000000000000002
mkt_seg_grpOther           1.0578058  0.0733030  14.431 < 0.0000000000000002
days_in_waiting_list      -0.0173533  0.0019012  -9.127 < 0.0000000000000002
meal_grpHB                 0.2343506  0.0250828   9.343 < 0.0000000000000002
arr_mo33                  -0.2152714  0.0264528  -8.138 0.000000000000000402
arr_mo36                  -0.3732047  0.0320999 -11.626 < 0.0000000000000002
arr_mo39                  -0.4728339  0.0443605 -10.659 < 0.0000000000000002
adr                        0.0068949  0.0003168  21.767 < 0.0000000000000002
adults                     0.1682535  0.0239749   7.018 0.000000000002252393
```

Figure 14

The final model revealed the following results for the Resort Hotel:

- The Non-Refundable deposit type was significant and had higher cancellation probability than the other two types (refundable and no deposit). The non-refundable definition was customers made full deposit when making a reservation, refundable was customers that made partial payment. But this field did not say whether the payment can be refunded if a reservation is canceled. It would have been more helpful to have that piece of information.
- Agent ID with value of “NULL” and “Other” have lower cancellation rates than agent ID “9” and “240”. A Null value means the customer made the reservation themself involving no agents.
- Reservations made by Portuguese visitors have a much higher cancellation probability. Germany visitors have the lowest cancellation rate.
- 1 unit (day) increase in lead time, cancellation probability increased by 0.8% .
- Reservations that have special requests have lower cancellation probability than the ones that do not
- Reservations that had booking changes had a lower probability of cancellation than the ones that have not.
- If a reservation was made by a repeated guest, that reservation presented lower probability compared to one reservation made by a new guest.
- February has the highest cancellation rates; July has the lowest cancellation rates.
- For each additional night of stay, the cancellation probability increases by 8.2%.
- A reservation made by a “transient” customer has a high cancellation rate.
- A $1 increase in the Average Daily Rate (ADR) is associated with a 0.5% increase in cancellation probability.
- A 1 day increase in the wait line was associated with 0.4% decrease in the cancellation probability
- A reservation that had children had a high probability of cancellation.

- A 1-degree F temperature increase was associated with a 4.9% increase in the cancellation probability.

For City Hotel, the findings were as follows:

- Similar finding for *deposit_type* variable. Non-refundable and refundable customers have higher cancellation rates than non-deposit customers.
- Germany visitors have the lowest cancellation rate, and Portuguese visitors (base) have the highest cancellation rate. This finding is similar to the Resort Hotel one.
- Variables of *has_special_request, lead_time, has_booking_change , total_nights_stayed, days_in_waiting_list, adr, adults* have similar conclusion like the Resort hotel.
- The Transient customer column has a higher cancellation rate, and the magnitude is higher than that of the resort hotel one.
- Winter months (Jan through Feb) have higher cancellation rates than the other months. September has the lowest cancellation rate
- The Corporate market segment has the lowest cancellation rate.
- The Half Board has a higher cancellation rate than all the other types.

Prediction results on test dataset:

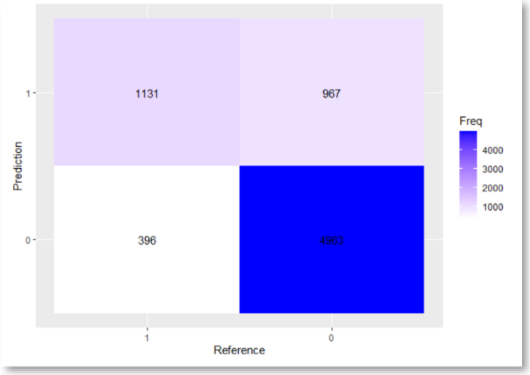


Figure 15

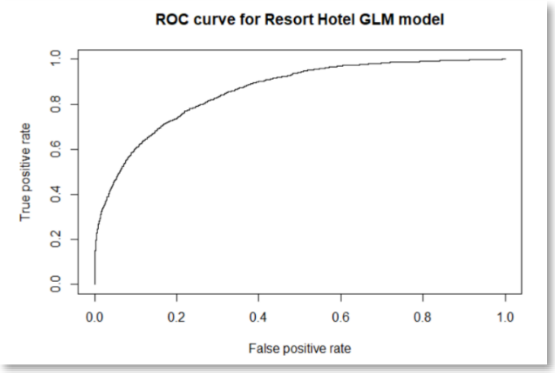


Figure 16

Our predictions revealed the following in the Confusion Matrix. The AUC value for the Resort hotel is 86.13%.

Our analysis of the City Hotel model revealed the following results. The AUC value for the city hotel was 85.18%

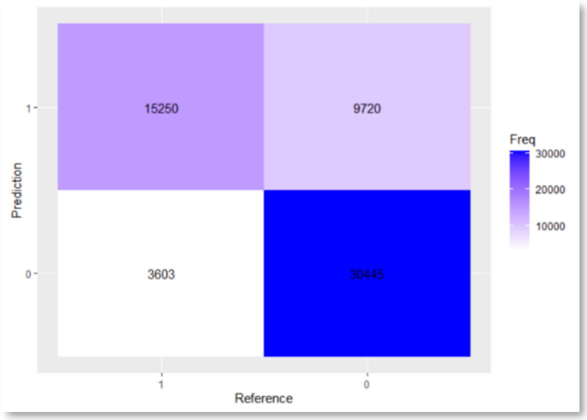


Figure 17

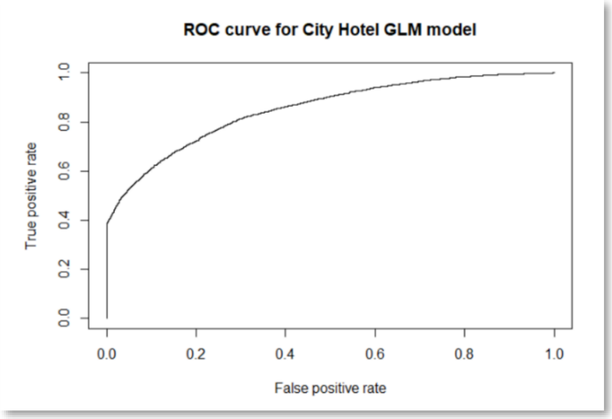


Figure 18

Observations

The result of the model trials, and the accuracy measurements, reveal a relatively truthful model that can predict rather successfully whether a booking will be susceptible to a cancellation in the future. Although our methodology is not perfect, and there are several results we would have preferred to verify through other means, our journey towards development of the model has been fruitful and beneficial.

Because of time constraints and computational challenges, we had to make some assumptions for our study to simplify procedure, nonetheless observing a realistic interpretation and solution of the problem at hand. On the other hand, one of the approaches we would have wanted to explore in more detail is the study of cyclical patterns, fluctuations, and trends, by exploring the data from a time-series based point of view. Because vacations are a seasonal occurrence, it is a mere consequence that vacation patterns throughout the year will affect the number of cancellations for any given hotel. We assume with a certain degree of confidence that this analysis would most likely have revealed that cancellations are more frequent during the non-summer months due to the consistency of summer vacations.

An extra data-preparation step that we would have liked to implement was cross validation. This would have proven very beneficial for our modeling phase, effectively helping the training procedure by discarding random effects in favor of real effects. Despite the objectivity of this technical solution, there are many other variables that go into the estimation of whether a customer is likely to fulfill or cancel their hotel room reservation. As an example, the unforeseen events that took place in recent years, in the context of the pandemic, have harshly impacted the hospitality sector globally.

We believe that within a certain confidence interval, and with exhaustive historical data, any data analyst should be able to build a model that can accurately predict pattern of occurrence and outcomes for future observations.

Conclusions

In this analysis we have attempted to shine light on which factors, common within the Hospitality industry, and to which extent, may affect hotel room booking cancellations. There are several considerations when it comes to developing a model for such a specialized sector, some of which we were able to identify and analyze, and some other we have found hard to gather insight for leading us to make assumptions.

Our analysis has concluded that on top of common factors that influence customer's behavior, there are also some less intuitive components that will affect reservation cancellation. Through the means of a statistical study, we were able to identify those and give a possible reasonable explanation, consolidating our conclusions into a solid tool that could help managers and concierge maximize the vacancy efficiency on their premises.

This project has been as great opportunity to learn real life applications of business data analysis and we can only hope to produce better models and analysis to contribute to a more meaningful application outside of the classroom.

Sources

Hotel Data set: https://www.sciencedirect.com/science/article/pii/S2352340918315191

Find our extensive study in our project folder, namely: GroupProject.html