STAT 425: Applied Regression and Design

Final Project Report

Exploratory Data Analysis and Statistical Modelling for Hotel Booking Demand

Alena Sorokina

Spring 2020

University of Illinois at Urbana-Champaign

Section 1: Introduction

Nowadays many businesses are becoming data-driven, which means that they tend to make their decisions and solve the problems based on insights learned from the data. They hire Data Scientists, Analysts, and Statisticians, who could implement the comprehensive analysis of the data, apply mathematical and statistical models, and create their own algorithms. The hotel business is not an exception, and one of the problems in the hotel business is a high number of reservation cancellations. Last-minute cancelations lead to the loss of revenue and a potential negative impact on the business. Therefore, the prediction of the possibility of reservation cancellation could save revenue and optimize the use of the hotel's resources.

My project aims to predict the possibility of reservation cancelation based on the statistical analysis of the Hotel Booking Demand dataset. The dataset contains information about City Hotel and Resort Hotel, such as the date of the booking, average daily rate, length of stay, number of adults and children guests, information about the special requests, etc. The data came from the self-titled article written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The dataset was posted on Kaggle in February 2020 and is being used by people who want to practice exploratory data analysis and predictive modeling.

Section 2: Exploratory data analysis

Initial dataset includes 18 variables (description of variables is taken from the Tidy Tuesday GitHub)[1]

Variable	Class	Description
hotel	categorical	Resort Hotel or City Hotel)
is_canceled	categorical	Value indicating if the booking was canceled (1) or not (0)
lead_time	numerical	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	numerical	Year of arrival date
arrival_date_month	categorical	Month of arrival date
arrival_date_week_number	numerical	Week number of year for arrival date
arrival_date_day_of_month	numerical	Day of arrival date
stays_in_weekend_nights	numerical	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	numerical	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	numerical	Number of adults
children	numerical	Number of children
babies	numerical	Number of babies
meal	categorical	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)
market_segment	categorical	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
reserved_room_type	categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
customer_type	categorical	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	numerical	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
total_of_special_requests	numerical	Number of special requests made by the customer (e.g. twin bed or high floor)

I will analyze the City Hotel, so after selecting data corresponding to it, I will drop the variable *hotel*. *Is_canceled* is going to be my response variable, so I will analyze my data with regard to it to understand the relationships between predictors and response and recognize patterns in my data.

Firstly, I want to see the proportion of reservation cancellations out of the total number of reservations. From Figure 1, it could be seen that approximately 27% of all reservations were canceled.

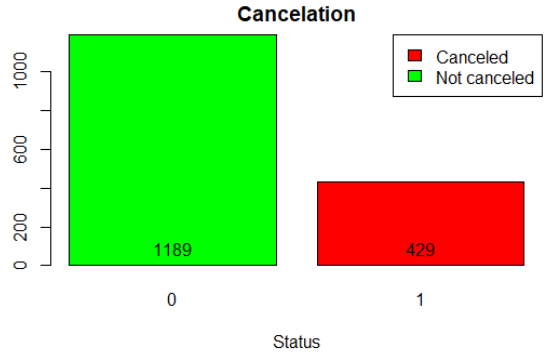


Figure 1. Proportion of reservation cancelations

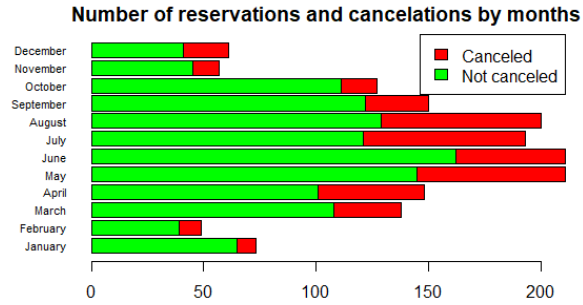


Figure 2. Number of reservations and cancellations by months

Secondly, I want to understand how the hotel demand depends on time. Since *arrival_date_year* contains information for only three years (2015, 2016, 2017), it will be hard to do a time series analysis based on years. I will stick Monthly analysis of the demand, so I will drop the variable *arrival_date_year*. From Figure2, we can see how the demand is changing by months. As expected, the highest demand is demonstrated during the summer months, while the lowest demand is during the coldest months. From Figure 2, we can see that the proportion of cancelations is fluctuating, lower for October, June and March, and higher for July and August.

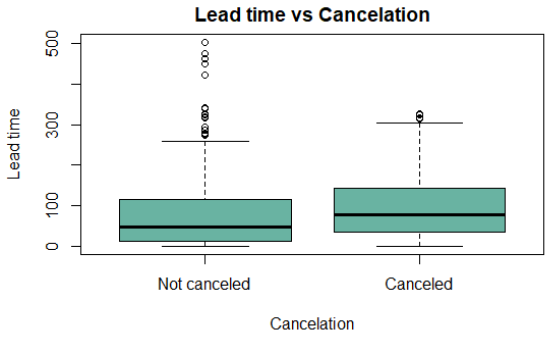


Figure 3. Lead-time vs Reservation cancelation

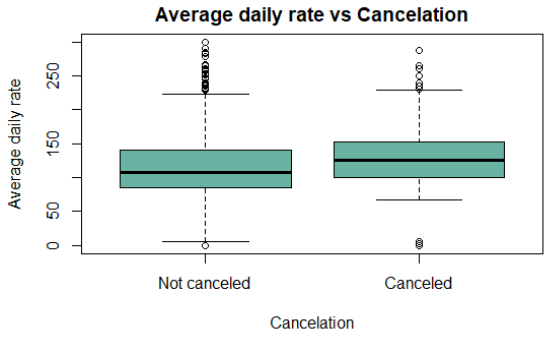


Figure 4. Average daily rate vs Reservation cancelation

Then, I would like to determine how the lead-time (number of days that elapsed between the entering date of the booking and the arrival date) correlates with the reservation cancelations. From the boxplot (Figure

3), we could see that the earlier reservations are more likely to be canceled. From Figure 4, it could be seen that bookings that are more expensive have higher probability of being canceled, and this seems reasonable because people may get a better deal and call off old bookings.

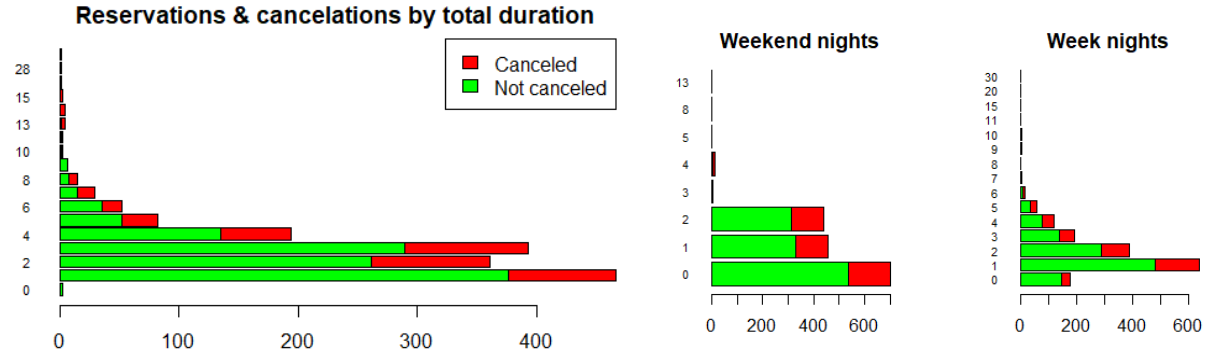


Figure 5. Reservations and cancelations by the duration

Moving to the next variables, I want to look at the relationship between reservation cancelation and the length of the stay. We could see that the percentage of cancelations is generally similar for different durations, with minor changes in the tails of the data: shorter durations are less likely to be canceled, while longer length demonstrates the opposite.

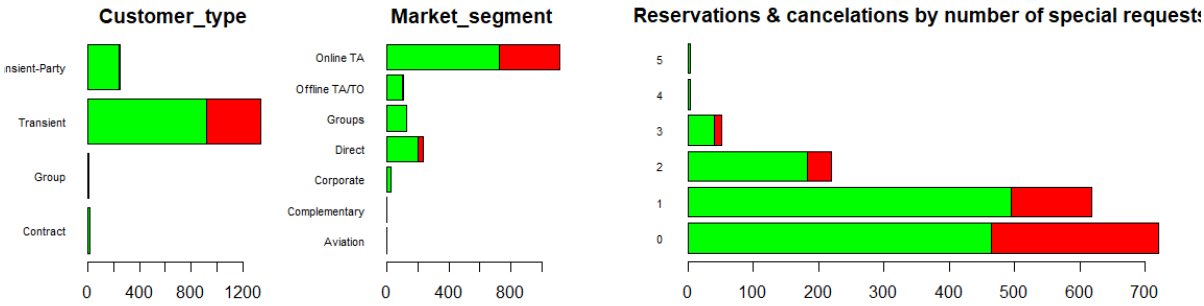


Figure 7. Customer type Figure 8. Market segment Figure 9. Number of special requests

From Figure 7, we could see that the Transient customer type is prevalent in our dataset, and it is more likely to be called off. From Figure 8, we can see that Online Travel Agents are more probable to cancel their reservation, and from Figure 9, we could not define the group with the largest percentage of cancelations.

From the variables descriptions, it could be clearly seen that some of the variable repeat the same information, for example, *arrival_date_month* and *arrival_date_week_number*. To avoid the correlation problems in my model, I will left the *arrival_date_month* variable in my model, because it is informative and easier to use and interpret, and I will exclude *arrival_ date_week_number*, *arrival_date_year* (explained earlier), and *arrival_date_day_of_month* from my final model.

From the correlation matrix (Figure 10), we could see that our numerical variables are not correlated.

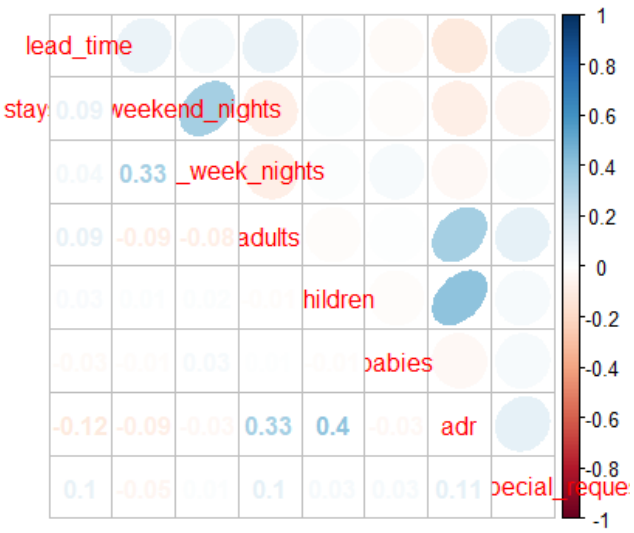


Figure 10. Correlation matrix

So far, I have only five categorical variables in my dataset: *arrival_date_month*, *meal*, *market_segment*, *reserved_room_type*, and *customer_type*. From the exploratory analysis, we have seen that in the *customer type* there is a prevalent type that makes up to 83% of total values, so I will not include the interactions with this variable. I would like not to include the interaction, as there will be many NA values and it will make the interpretation of my model harder.

Section 3. Method

3.1. Logistic Regression

I would like to start with the simple model. Since my output is binary, I would not be able to use a linear regression model for my analysis because it violates the assumption that the response variable is normally distributed. I would start my analysis from the Logistic Regression Model.

Firstly, I will divide my dataset to the Training (80%) and Test (20%) parts, and for each model, I will report training and testing error. My first model does not include interactions, and it is formulated as:

is_canceled ~ lead_time + arrival_date_month + stays_in_weekend_nights + stays_in_week_nights + adults + children + babies + meal + market_segment + reserved_room_type + customer_type + adr + total_of_special_requests.

Test error: 76.2%
Train error: 80.6%
AIC: 1232.7

This is a good result for the simple model. In order to reduce my model, I will conduct the Stepwise Mode Selection with AIC and BIC criteria.

Model selected according to AIC criteria:

is_canceled ~ lead_time + stays_in_weekend_nights + stays_in_week_nights + meal + market_segment + reserved_room_type + customer_type + adr + total_of_special_requests

Test error: 77.5%
Train error: 79.5%
AIC: 1215.8

Model selected according to BIC criteria:

is_canceled ~ lead_time + stays_in_week_nights + market_segment + adr + total_of_special_requests

Test error: 75.6%
Train error: 79.1%
AIC: 1251.6

From the results, it could be seen that the first simple model demonstrates better results than the reduced models in terms of test and train errors. However, the lowest AIC score belongs to second model, therefore, it should be considered as the best out of given threes.

Interpretation:

I am going to interpret the chosen model.

```
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -0.6474174  0.4048532  -1.599  0.11004
## lead_time                      0.0010530  0.0001359   7.748 1.90e-14 ***
## stays_in_weekend_nights        0.0330315  0.0120868   2.733  0.00637 **
## stays_in_week_nights           0.0195733  0.0066048   2.963  0.00310 **
## mealHB                        -0.1136286  0.0832642  -1.365  0.17260
## mealSC                         0.0825528  0.0301417   2.739  0.00625 **
## market_segmentComplementary    0.3803881  0.5615250   0.677  0.49826
## market_segmentCorporate        0.2226992  0.3982992   0.559  0.57617
## market_segmentDirect           0.2018799  0.3891297   0.519  0.60399
## market_segmentGroups           0.2510352  0.3929063   0.639  0.52299
## market_segmentOffline TA/TO    0.1560545  0.3910901   0.399  0.68994
## market_segmentOnline TA        0.4199998  0.3885234   1.081  0.27990
## reserved_room_typeB            0.0659305  0.0841005   0.784  0.43322
## reserved_room_typeC           -0.0756212  0.3983276  -0.190  0.84946
## reserved_room_typeD            0.0628861  0.0318982   1.971  0.04889 *
## reserved_room_typeE           -0.0210135  0.0611709  -0.344  0.73126
## reserved_room_typeF            0.0444317  0.0658622   0.675  0.50004
## reserved_room_typeG           -0.4486334  0.0829075  -5.411 7.47e-08 ***
## customer_typeGroup             0.0079426  0.1623920   0.049  0.96100
## customer_typeTransient         0.2251150  0.1056151   2.131  0.03324 *
## customer_typeTransient-Party   0.0545032  0.1167258   0.467  0.64063
## adr                            0.0023393  0.0003516   6.652 4.28e-11 ***
## total_of_special_requests     -0.1245769  0.0131426  -9.479  < 2e-16 ***
## ---
```

For example,

The probability of reservation cancelation per day in lead time increases by a factor of exp(0.001053)=1.

The probability of reservation cancelation per day in stays_in_weekend_night increases by a factor of exp(0.0330315)=1.395.

The probability of reservation cancelation per day in stays_in_week_night increases by a factor of exp(0.0195733)=1.02.

The probability of reservation cancelation per euro in adr increases by a factor of exp(0.0023393)=1.

3.2. Regression Tree

The next model I would try to use in my analysis is the regression tree model.

My full tree is overloaded with nodes, so I will use tree pruning in order to select the best sub-tree. Tree is defined by the number of parameters that correspond to the number of leave nodes. The model selection criteria is:

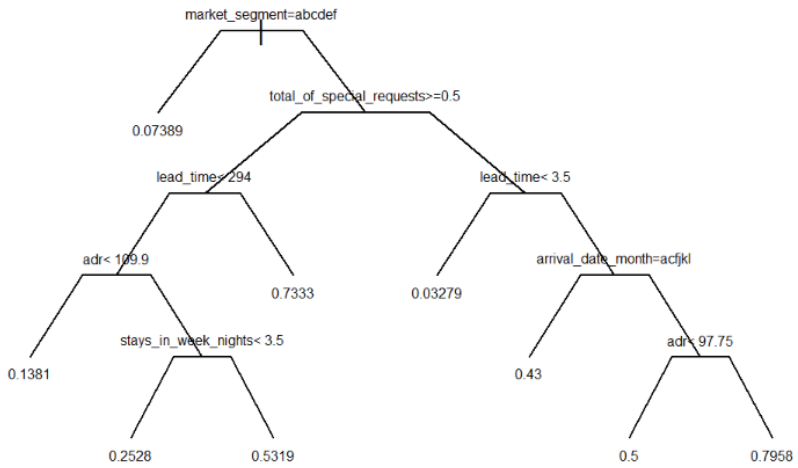
$$RRS + cp * tree\_size$$

I have chosen CP (complexity parameter) to be equal to 0.001, in order to construct the big enough tree for consequent pruning. For each tree size, I will compute the corresponding CV error, and select the subtree with the smallest CV error. The rpart method also uses cross-validation on the data.

There are two methods of selecting the sub-tree:

1. **size.min**: Pick the optimal tree size which has the smallest CV error.

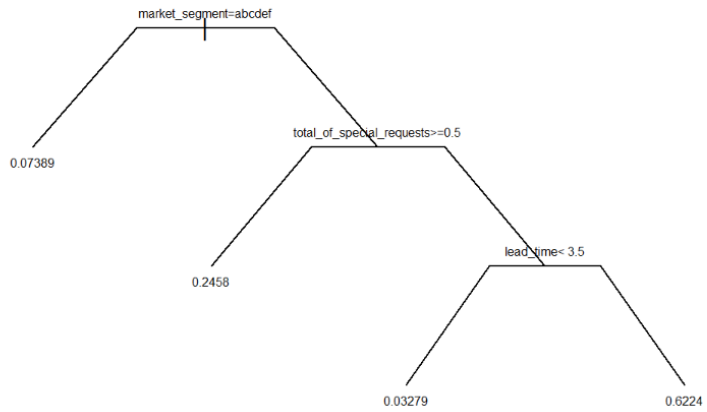
Model selected by **size.min** method:



Test error: 79.6%
Train error: 80.8%

2. **size.1se**: Pick the smallest tree size whose CV error is within one-standard-error of the smallest CV error.

Model selected by **size.1se** method:



Test error: 80.9%
Train error: 78.9%

From the results, it could be seen that the model selected by size.1se criteria gives us better test error, and it has smaller number of nodes, therefore, the second model is better.

3.3 Random Forest

Random forest is a Supervised Learning algorithm which uses multiple regression/classification trees for classification/regression problem. It is called ensemble learning – a method that combines outputs from multiple learning models to provide more accurate predictions than an individual model.

Random forest model

Test error: 98.8%
Train error: 84%

Random forest gives us the best results in terms of train & test errors, so it empirically proves that the ensemble learning outperforms any individual model.

Section 4. Discussion of the results

My project aimed to predict the probability of hotel reservation cancellation using the Hotel Booking Demand dataset. As my response variable was binary, I decided to use logistic regression, regression tree, and random forest models. The best model out of all aforementioned is the random forest, since it is an ensemble model, and it outperforms any of the individual models. However, the Random Forest Model is harder to interpret in comparison with the individual models.

I believe that the knowledge which I get from this project could be applied to real industrial problem-solving. It helped me to look into the course material more generally and better understand the practical application of the statistical analysis.