Ömer Yiğit – r0767950

STATISTICAL CONSULTING – HOTEL RESERVATIONS REPORT

1. Introduction

The reservations dataset contains booking information of the customers from a city hotel and a resort hotel, from 2014 October to 2017 September, spanning across three years. The raw data has information about more than 100.000 reservations and 32 variables like date, number of days, daily rate, special requests (twin bed, high floor…) etc. The dataset is analyzed from two perspectives: Customer perspective and hotel management perspective. Naturally, customers and hotel management are interested in answering different questions. Using this dataset, those questions are tried to be answered and certain conclusions are reached.

2. Data Cleaning and Preprocessing

A data analysis should always start with cleaning the data to reach more powerful conclusions. Missing and inaccurate data points are detected. Raw data contains some missing points, erroneous values, and outliers. Since the contributions of these data points to the entire dataset in percentages are negligibly small (less than 2%), they are deleted. Then, in order to have a better understanding, new variables are created from the existing variables, like total price, if a stay includes weekend or not etc. Also, certain variables that have identical information are deleted. In the end, the dataset ready for the analyses. The complete list of variables is given in the Appendix.

3. Customer Perspective

People spend so much time to arrange the best possible holiday and handle their reservations accordingly. A hotel should provide a satisfying service with affordable prices. Also, people may have different preferences in terms of holiday period. In the light of such information, there are two main questions of interest that should be answered:

1. When is the best period of the year to book a hotel?
2. What is the optimal length of stay in order to get the best daily rate?

3.1. The best period of the year

The number of reservations across the year fluctuates monthly. *Figure 1* below shows the relationship.

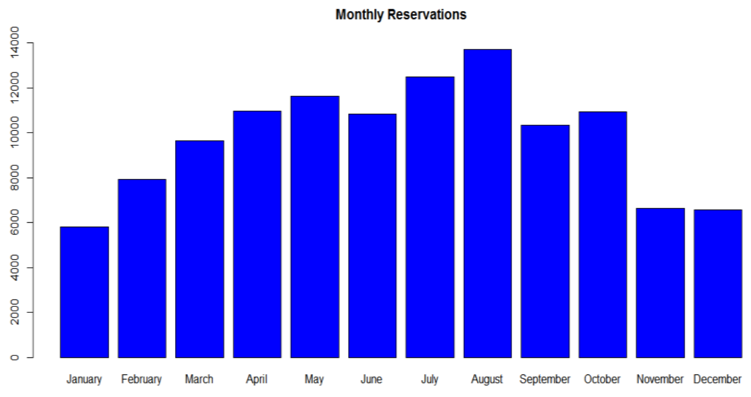

Figure 1: Monthly reservations made by the customers

Most of the reservations are done in summer season, whereas the winter season is the least preferred. Also, last week of the year has a hidden spike, indicating the Christmas holiday. Another observation is that the city hotel, on average, is 40% more expensive during the year, but the resort hotel, on average, is 50% more expensive in the summer season.

High number of reservations also indicate higher daily rates. Therefore, an off-season holiday at the resort would be the cheapest option. If crowd and liveliness is preferred, a late-summer holiday would be a good choice, considering the number of reservations (hence the rates) decreases in September.

3.2. Optimal length of stay

Intuitively, longer stays indicate cheaper daily rates. Also, longer stays cost more in total. Therefore, an optimum point should be achieved to have a satisfyingly long holiday with reasonable prices. The relationship between the length of stay and price is shown below in *Figure 2* and *Figure 3*.

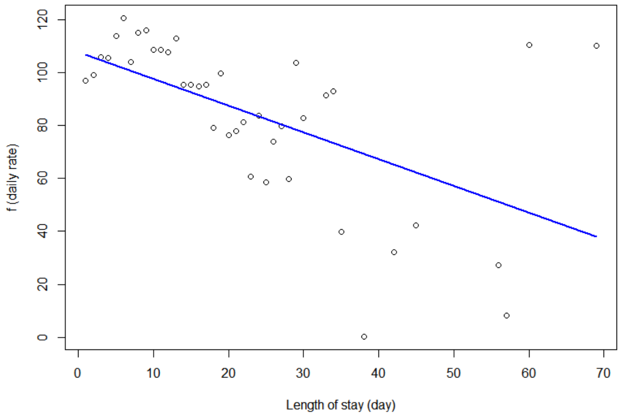

Figure 2: Daily rate is lower on longer stays

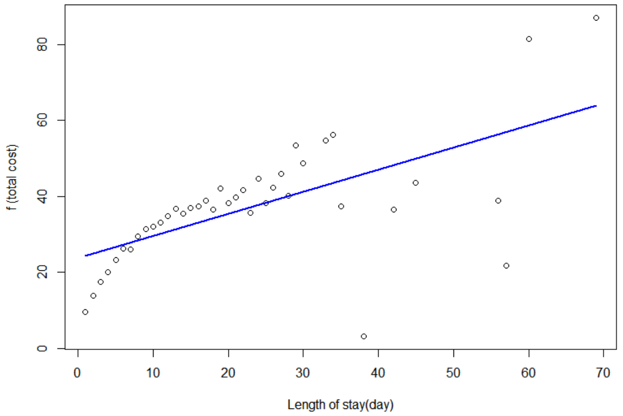

Figure 3: Total cost is higher on longer stays

Although it is difficult to find a precise, optimum number on the length of stay, statistical tests show that there is a significant difference in daily rates between the reservations that include weekends and the reservations that do not include weekends. Omitting the extreme values, average daily rates are 99.06$ and 95.39$ for the reservations that include weekends and do not include weekends, respectively. Hence, it can be said that short stays (around 5 days) which do not include weekends are the cheapest option. Also, since the longer stays provide cheaper daily rates, a 12-day reservation which does not include the second weekend can be another satisfying alternative.

4. Hotel Management Perspective

As a hotel management, there are two main variables of interest that needs to be predicted to provide a better service to customers: Cancellation and special requests. Hotel management needs to answer the following questions:

1. Can we predict whether a customer would cancel the reservation or not?
2. Can predict how many special requests a customer would have?

Both variables have different type of information. Cancellation is a binary variable, whereas special requests is a count variable. Therefore, different approaches are needed to predict these outcomes.

4.1.Cancellations

The cancellation is a binary variable, it takes only the values of 0 (not cancelled) and 1 (cancelled). Predicting the possibility of a cancellation can provide flexibility to the hotel management. There are certain binary classification methods to apply for predicting whether a customer would cancel the reservation or not. Before applying the models, the dataset is randomly divided into two: Training set and test set. Training set and test set consist of the 80% and 20% of the actual data, respectively. Training set is used to apply the model and test set is used to validate if the model is accurate or not. This is a necessary step, because predicting the future outcomes with the same data on which the model is applied can yield unrealistically high prediction accuracy. Also, it should be noted that due to randomness, the prediction accuracies might change each time. However, these changes would not be significantly different than each other. Two methods, namely, logistic regression and random forest, are applied and compared in the following sections.

4.1.a. Logistic Regression

Logistic regression method assigns probabilities to customers, according to given values of the variables included in the model. Then, these assigned probabilities are predicted. Based on a pre-specified cut-off probability point (mostly 50%), it is predicted whether a customer would cancel the reservation or not. In order to apply a valid model with high prediction accuracy, it is important to decide on which variables should be included.

Variables are selected with backward selection method. First, a model is created by including all variables. In each step, the variable which improves the model the most when omitted is deleted from the model. This process continues until no further improvement is observed. The aim of this process is to find and eliminate the unimportant variables that have no significant effect in predictions. Then, the ultimate model is used to predict the outcomes on the test set. The actual and predicted outcomes given below in *Table 1*.

		Actual	
		0	1
Predicted	0	13462	3401
	1	1142	5481

Table 1: Cancellation predictions according to logistic regression (0: no cancellation, 1: cancellation)

4.1.b Random Forest

Random forest is a classification method that constructs numerous decision trees and outputs the majority decision made by those trees. This method performs better on balanced data. However, the hotel dataset regarding cancellation is not balanced; there is only 37% cancellation. Therefore, a procedure is needed to obtain a balanced data. The minority part is randomly sampled with replacement until the number of observations from both levels (cancelled and not cancelled) are the same. This procedure is called bootstrapping.

Then, on this balanced dataset, training and test sets are created. The model is trained with different number of trees, varying from 50 to 500 with the increments of 50. When these models are tested on the test set, the lowest prediction error occurred on the model with 200 trees. This procedure is called cross-validation. The aim of this procedure is to prevent yielding unrealistically high prediction accuracy. Backward selection method is applied here as well. The future analyses are done on this model with the test set. The actual and predicted outcomes are given below in *Table 2*. The sum of the values in *Table 2* is higher due to bootstrapping procedure.

		Actual	
		0	1
Predicted	0	13051	1626
	1	1249	13442

Table 2: Cancellation predictions according to random forest (0: no cancellation, 1: cancellation)

4.1.c. Comparison of the Models

When comparing two models, the most common measures are accuracy and AUC. Accuracy is simply the ratio between the sum of diagonals (accurate predictions) and the sum of all numbers in the table. AUC value is also important, because it prevents obtaining unrealistically high prediction accuracies by testing different cut-off values. For example, if 99% of the customers do not cancel their reservations, the management can have 99% accuracy by simply doing nothing and assuming that nobody will cancel their reservations. This obviously would be a wrong decision.

The closer these values are to 1, the better the model is. The accuracy and AUC values for the logistic regression model is 0.81 and 0.85, respectively, whereas the accuracy and AUC values for the random forest model are both 0.90. Hence, we can conclude that the random forest model is better at predicting the cancellations with 90% success.

4.1.d. Cancellation Profile

Customers who, in general, have reservations for longer durations and with more people (hence, more costly) are more prone to cancel their reservations, because these parameters create more uncertainty. Also, customers who cancelled a reservation before are more likely to cancel a reservation. On the contrary, returning who have not cancelled any reservations before are the least likely to cancel a reservation. Lastly, some customers prefer to change bookings instead of cancelling.

4.2.Special Requests

The graphical information about the special requests is given below in *Figure 4*. According to the numbers, majority of the customers do not have any special requests. The average special request made by a customer is 0.57.

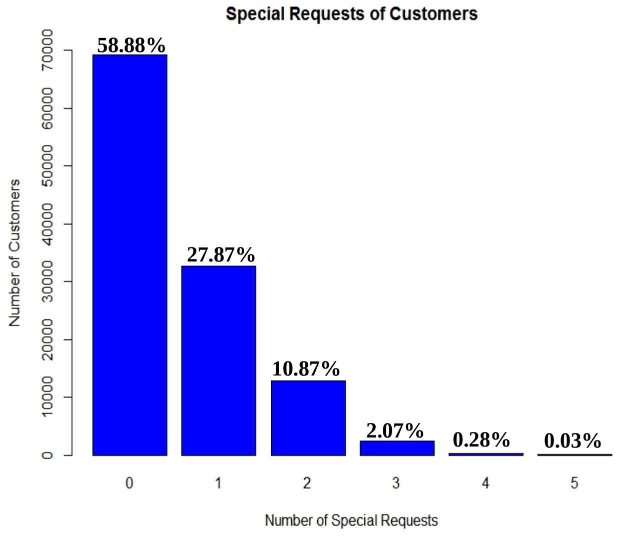

Figure 4: Customer information about special requests

This variable is a count variable, ranging from 0 to 5. Variables that have count values are generally assumed to follow a poisson distribution. Number of phone calls per day, number of patients arrive at a hospital etc. are some examples of data that has poisson distribution. Hence, poisson regression method is applied, and the results are analyzed in the following section.

4.2.a. Poisson Regression

In order to predict the outcome of a variable that has poisson distribution, the most common approach is to fit a poisson regression model to the data. Variables are chosen with forward selection method, which is a similar method to backward selection. The forward selection method starts from a null model which has no variables. In each step, the variable which contributes to the model the most is found and added. This process continues until no further improvement is observed. The validity of the model is satisfied with statistical tests. Then, the training and the test sets are created with a systematic approach. The proportions given above in *Figure 4* are preserved in both datasets. Three alternative family parameters are tested and compared. Technical properties aside, the estimates are similar in all these methods, with the only significant difference being the variability of the estimates. Parameter estimates with smaller variability indicates more precise predictions.

According to the model outputs, average prediction error in absolute values (mean absolute error – MAE) is 0.45 and in squared values (mean square error – MSE) is 1.64. Compared to possible alternative models, these numbers are quite small, indicating the model is successful. Hence, this poisson regression model can successfully be used for future predictions for the number of special requests a customer would have.

4.2.b. Special Requests Profile

Customers who, in general, have more special requests made their reservations online, because special requests are easier to specify with online booking. They can also edit their preferences and add special requests before the arrival. Also, regular customers who repeatedly had reservations before are more demanding and may have more special requests.

Conclusions

After the data is described and made ready for the analyses, it is analyzed both from the customer and from the hotel management perspective.

Customers can take the following observations into consideration before booking a hotel:

- A month when the number of reservations is high also has higher daily rate.
- City hotels are more expensive than resort hotels, except summer. Also, weekends are more expensive than weekdays.
- An off-season holiday at a resort is the cheapest option. Early September can also be a good alternative for a late-summer holiday.
- A 5 day-long holiday without a weekend is the cheapest option. A 12 day-long holiday without the second weekend can also be a good alternative.

Hotel management can take the following observations into consideration to be more flexible for uncertainties and to provide more satisfying customer experience:

- Cancellations can be predicted beforehand with 90% precision by using a random forest model, and number of special requests can also be predicted beforehand with very little deviation by using a poisson regression model.
- Longer reservations, reservations with more people, and reservations made by one who has cancelled a reservation before are indicators of a possible cancellation. Also, loyal customers who have had reservations before would most likely not to cancel their reservations.
- Online reservations include more special requests, in general, due to easiness. Also, regular customers can be more demanding in terms of special requests.

APPENDIX

Variables in the Dataset

Variables from 1 to 32 are in the raw data. Variables from 33 to 38 are created using the raw data.

	Variable
1	hotel
2	is_canceled
3	lead_time
4	arrival_date_year
5	arrival_date_month
6	arrival_date_week_number
7	arrival_date_day_of_month
8	stays_in_weekend_nights
9	stays_in_week_nights
10	adults
11	children
12	babies
13	meal
14	country
15	market_segment
16	distribution_channel
17	is_repeated_guest
18	previous_cancellations
19	previous_bookings_not_canceled
20	reserved_room_type
21	assigned_room_type
22	booking_changes
23	deposit_type
24	agent
25	company
26	days_in_waiting
27	customer_type
28	adr (average daily rate)
29	required_car_parking_spaces
30	total_of_special_requests
31	reservation_status_date
32	is_weekend_included
33	stays_in_total_nights
34	people
35	reserved_assigned_same
36	is_special_request
37	total_cost

Variables Used in the Models

Numbers are given as coded above.

Section	Model	Variables Used
4.1.a	Logistic regression	1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16,17, 18, 19, 22, 23, 27, 28, 29, 30, 37, 38
4.1.b	Random forest	1, 3, 5, 6, 13, 15, 16, 18, 22, 23, 27, 28, 29, 30, 33, 34, 35, 36, 37, 38
4.2.a	Poisson regression	1, 3, 10, 11, 12, 13, 15, 16, 17, 18, 19, 22, 23, 26, 27, 28, 29, 33, 34, 36, 38