Prediction Of Hotel Guests' Meal Choices Based On Hotel Booking Demand

Jing Zhang

2020/5/22

Contents

1 Introduction

When booking a hotel, people consider many factors, such as the best time to book a hotel, the location of the hotel and so on. For the hotel manager, how to better attract customers to book the hotel? Now I want to explore this issue from a different perspective. I want to know how likely it is for people to choose to eat in hotels so as to provide some constructive suggestions for hotels to attract more customers. My final goal is to make predictions about the possibility that people will choose to eat in hotels.

2 Data preparation and description

This dataset comes from the Kaggle(https://www.kaggle.com/jessemostipak/hotel-booking-demand#hotel_bookings.csv) and contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

Additionally,the data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The paper claimed that the data comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. All personally identifying information has been removed from the data.

There are a total of 119,390 observations and 32 variables in this data set, among which 14 are factor variables and the rest are all numerical variables.

3 Exploratory data analysis

After preliminary cleaning of the data and processing of missing values, let's investigate the following questions:

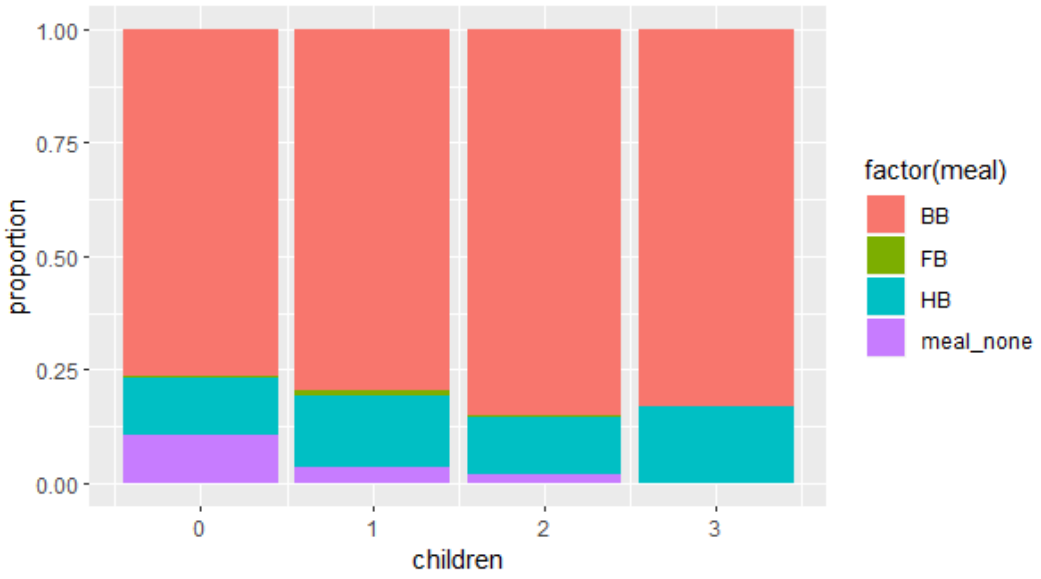

Figure 1: Analysis about Children

1. Do elders with children choose to eat in hotels?

From figure 1 we can know that whether the guests take children or not, a large proportion of them choose to eat in hotels. Moreover, no matter the number of children is large or small, the proportions of the four dining choices have slight difference. Therefore, the variable representing children will be converted into a binary variable when modeling.

2. Do elders with babies choose to eat in hotels?

From figure 2 we can know that a large proportion of guests with babies choose to eat in hotels. What' s more, the guests with babies will be more likely to eat in hotels. Therefore, the variable representing babies will be chosen to the model and be converted into a binary variable when modeling.

3. How about availability of parking?

From figure 3 we can know that hotels without parking lots have a large proportion of guests. Moreover, whether hotels with parking lots or not, there is slight difference among the four dining choices. In addtion, the proportions of city hotel and resort hotel are different. Then the variables representing parking lots and the type of hotel will be chosen when modeling.

4. Does the average daily rate affect people's choice to eat in hotels?

From figure 4 we can know that the boxplot of the four dining choices are different. And the length of the box who choose to eat in hotels is longer than those who don't eat in hotels. In this case, the variables representing average daily rate will be chosen when modeling.

Similarly, by visualizing some variables, children, babies, type of hotel, month of arrival date, meal, average daily rate, adults, number of car parking spaces required by the customer, number of special requests made by the customer (e.g. twin bed or high floor), number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel, and number of weekend nights (Saturday or Sunday) the guest stayed

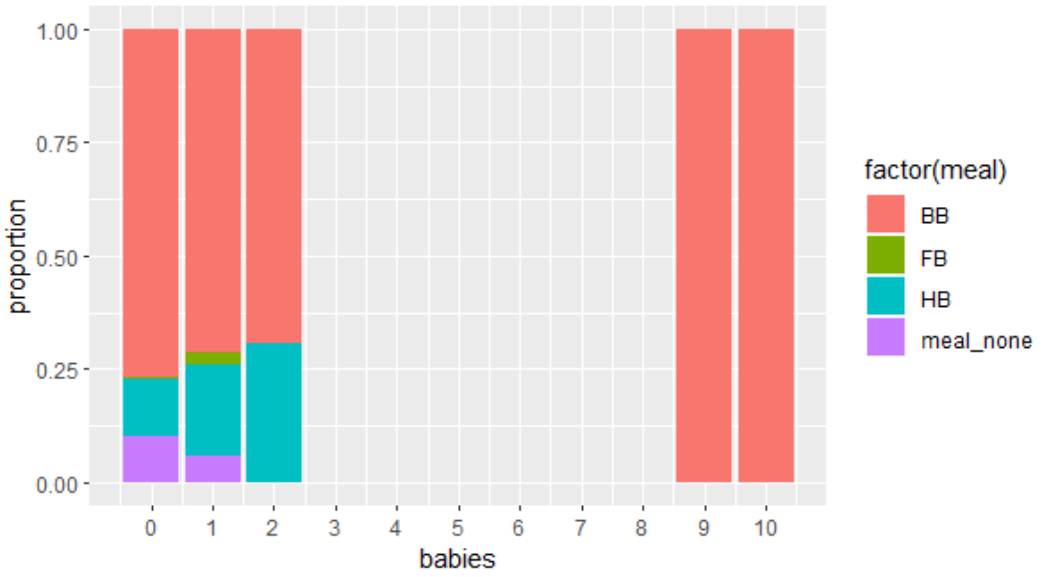

Figure 2: Analysis about Babies

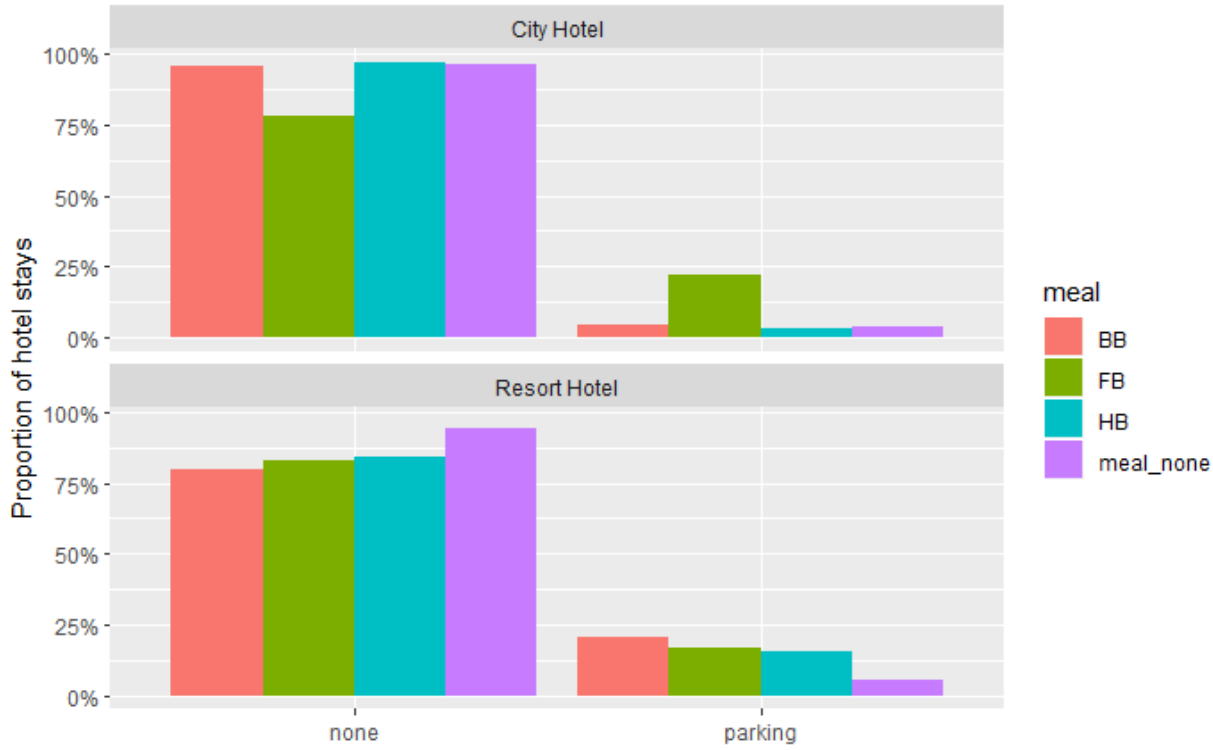

Figure 3: Analysis about Availability of Parking

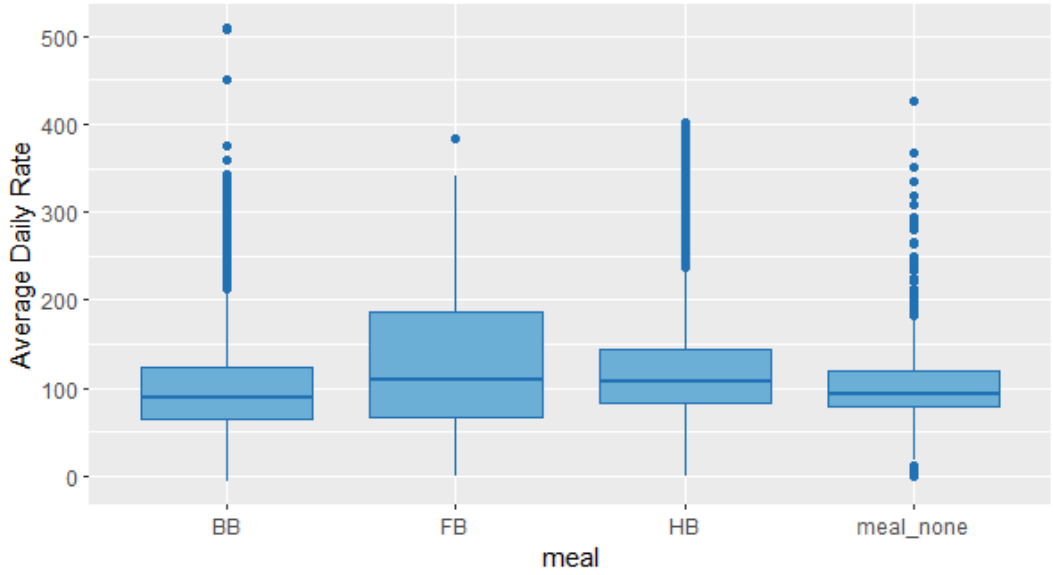

Figure 4: Analysis about Average Daily Rate

or booked to stay at the hotel these eleven features are eventually extracted to form a new data set for formal modeling, where the meal feature represents the dependent variable.

4 Formal modeling and results

To make predictions about the possibility that people will choose to eat in hotels, I choose three different methods to measure the effect of the prediction. These three methods are logistic regresson, k-nearest-neighbor and decision tree, each of them has its own pros and cons.

By using **tidymodels**, we can make formal modeling and train the model on the training dataset. we know that ROC Curves shows the tradeoff between true positive rate and false positive rate of classification algorithms. That's to say, ROC shows how many correct positive classifications can be gained as we allow for more and more false positives. Then I plot the ROC Curves of these three methods to visualize their results.

As shown in figure 5, we know that the k-nearest-neighbor model performs better than the decision tree and the decision tree model performs better than the logistic regression, since the ROC curve of knn is farthest from the diagonal of the plot, the decision tree's is the second and the logistic regression's is the nearest.

To make the conclusions more persuasive, I have calculated several common metrics of each method – accuracy, AUC, true positive rate and true negative rate. The results are displayed in table 1.

From the table 1, we can know that each metric's value of the k-nearest-neighbor model is the largest, the value of the decision tree is the second and the value of the logistic regression is the lowest. For example, as for the k-nearest-neighbor model, its accuracy is 0.756 higher than the others means that this model has higher ratio of correct predictions, its AUC is 0.823 higher than the rest means that we can quickly compare these learning models and conclude that the k-nearest-neighbor model performs better.

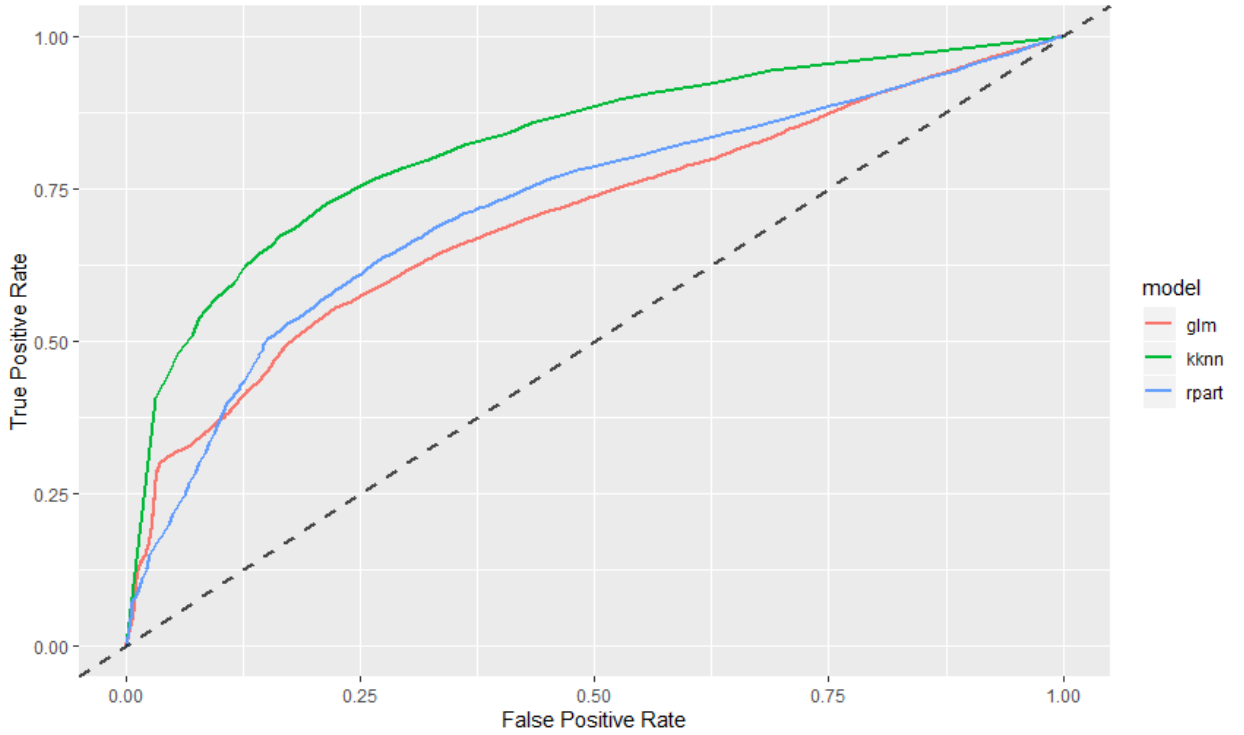

Figure 5: ROC Curve of Different Methods

Table 1 : Metrics of Different Methods

Method	Accuracy	AUC	True Positive Rate
Logistic Regression	0.663	0.700	0.705
K-Nearest-Neighbor	0.756	0.823	0.772
Decision Tree	0.681	0.720	0.704

From above analysis and results, I will use the results of the k-nearest-neighbor model to make predictions since its performance is better. We can know that the rate that people choose to eat in hotels is 0.772 means that the guests are large likely to eat in hotels. Therefore, for the hotel manager, in order to increase the turnover, a series of discount packages can be launched in food and beverage department to attract more guests to stay in the hotel.

5 Further discussion

Since each algorithm has its own advantages and disadvantages, I only studied three algorithms in this report, and readers interested in this report can also try other algorithms to make predictions and see if the results are better.

6 Supplementary materials

Supplementary materials are available at https://github.com/RainySeason2019/hotel-booking-demand, which contain the dataset, codes and so on.