What features of a hotel increases customer spend for a single booking?

Lab 2 - Carol Sun, Celina Wu, Patrick Yim

Introduction

Travel and hospitality industries have been booming since travel restrictions have been lifted around the world. According to statistics collected by United Nations, international travel is back at 80% of the pre-pandemic level and 235 million tourists traveled in the first 3 months of 2023. As hotels and others hospitality services are recovering from weak demands during the pandemic, we want to find out what are the factors that lead individuals to spend more money at a hotel in a single booking? Is it offering additional features, or the meals offered? Or is it the type of deposit, time on wait list that has an impact on the overall amount spent? This study tries to derive what are the functional characteristics that result in increased hotel revenue. We are hoping that our study can be used by hotel owners to navigate and potentially align on the specific financial benefit of certain characteristics and relationships.

This study estimates the value of implementing certain features within hotels, utilizing a data set from real hotel databases located in Portugal. The data set is broken up into two primary hotels, a resort hotel in the city of Algarve and a city hotel in Lisbon. By applying a set of regression models, we hope to estimate the value that results from certain features and characteristics.

Data and Methodology

The data in this study comes from the Hotel booking demand datasets, sciencedirect.com. It was compiled and made publicly available by Nuno Antonio, Ana de Almeida, and Luis Nunes. This data article describes hotel booking data of a city hotel and resort hotel located in Portugal. Dataset for both resort and city hotel share the same structure, with 31 variables describing the 40,060 observations of Hotel 1 and 79,330 observations of Hotel 2. Each observation represents a hotel booking. Datasets for Hotel 1 and Hotel 2 comprehend bookings with planned arrival date between the 1st of July of 2015 and the 31st of August 2017. The dataset including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. [1] The dataset we are using originated from a dataset that was provisioned by the research on revenue management. Due to the size and scope of the dataset, we narrowed down our primary focus to booking related to aviation and online Travel Agency and using data from the aviation data known as Passenger Name Record (PNR). To address the limitation of the unique requirements of industry-specific data, datasets were curated to determine the likelihood of hotel bookings being cancelled. Additionally, the leakage of future information to influence the dataset was prevented by including timestamps of the target variables after the timestamps of the input variables.

To determine the economic value, we analyze the factors that influence the amount of money spent on a single hotel booking. The dependent variable, Y, represents the amount of money spent with one booking (average daily rate * number of nights stayed). We examine the impact of various independent variables, X,

[1] Nuno Antonio, Ana de Almeida, Luis Nunes, Hotel booking demand datasets, Data in Brief, Volume 22, 2019, Pages 41-49, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2018.11.126. (https://www.sciencedirect.com/science/article/pii/S2352340918315191)

on customer spending, including the features listed above. This takes the assumption that both hotels offer some types of options for these variables.

After screening out the data that was canceled or did not have to spend money, there were 73419 rows remind in the dataset. To operationalize the economic value, we analyze the factors that influence the amount of money spent on a single hotel booking. The dependent variable, Y, represents the amount of money spent with one booking (average daily rate * number of nights stayed). We examine the impact of various independent variables, X, on customer spending, including features like hotelType, assigned_room_type, days_in_waiting_list, deposit_type, distribution_channel, market_segment, and meal. This takes the assumption that both hotels offer some types of options for these variables.Market_segment and distribution_channel functions the same as features. Therefore, distribution_channel is also eliminated. Then the columns with categorical data were split up to become binary columns to be used with regression test.

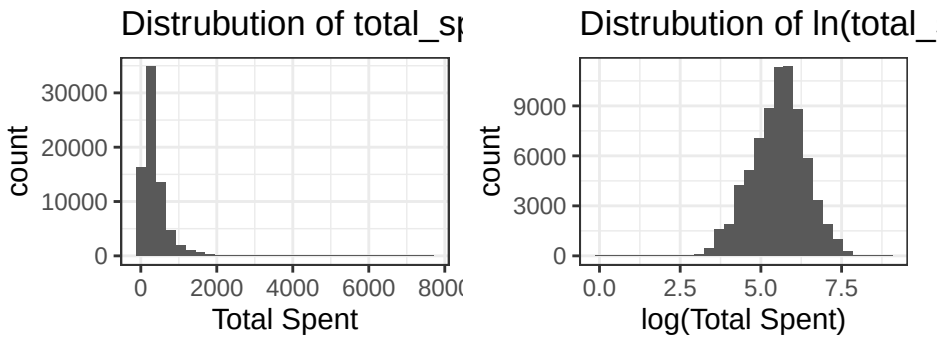


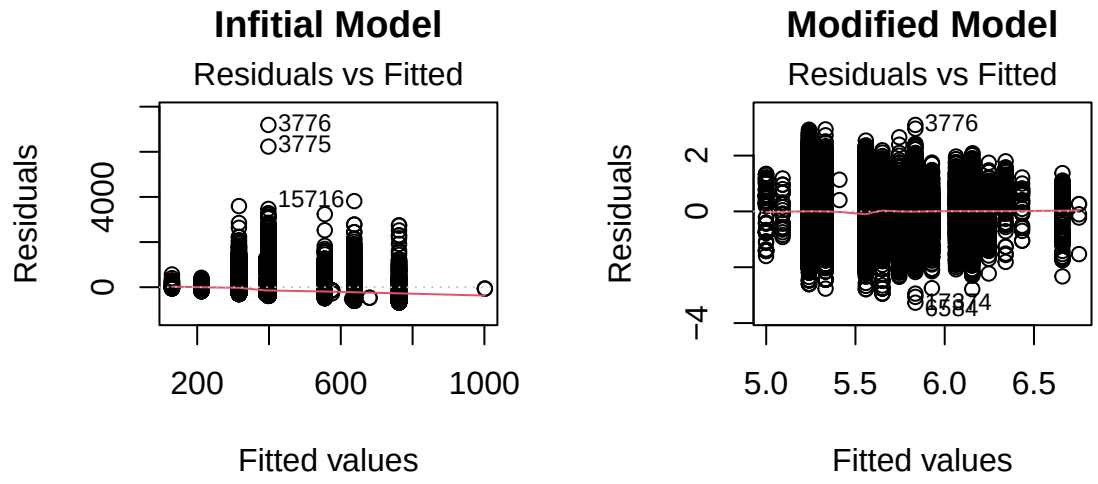
Figure 1: Distribution of total money spent

With the log of distribution of the amount money spent per booking, we found that the distribution is in a normal form but skewed toward left.Therefore, the we eliminate the date with the log of total money spent that is less or equals to 2.5 and left with 73321

The first model came up with the combination of the first category of the categorical features and all metric feature. This model includes city hotel, assigned room type A, Aviation market, bed and breakfast meal, days in waiting list and no deposit. Then the model was modified with the other features in the categorical data and eliminating each feature to find the balance between positive coefficient that would increase the amount of spending and the amount of residual gaining. The result of the trial and error gave us an aggressive model with possibly highest amount of coefficient and moderate residual increase. This model includes resort hotel, assigned room G, aviation market, full board meal and no deposit.

With this model, we did assumption tests to see if the model has any problem. We found that there is a problem with Linear Conditional Expectation with the residual plots were off to have high values. - IID - all row of data representing the result of single booking with all the categorical and metric data being independent - No Perfect Collinearity - by running correlation function, the biggest index is 0.13.

Therefore, we came up with another model using log(total_spent) instead and modified variables to get a better fit.



With the not so aggressive model and the log transformation of total spent, the residual seems to be more balanced. The rest of the assumption tests were conduct with the the modified model and seems to satisfy the conditions.

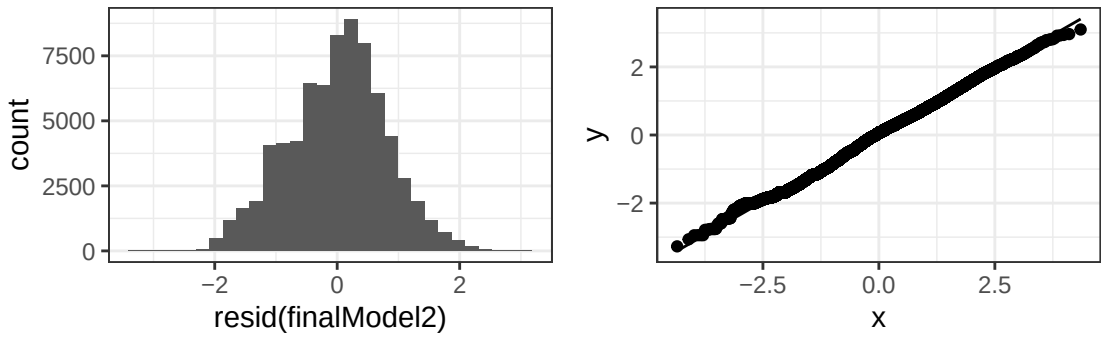


Figure 2: Residual Distribution

The linear regression model assumption:

- IID - all row of data representing the result of single booking with all the categorical and metric data being independent
- No Perfect Collinearity - by running correlation function, the biggest index is 0.13 which means there is no perfect colinearity beteen variables. The variables of VIF are less then 1.5 means there is low levels of multicollinearity.
- Linear Conditional Expectation - the problem was resolved through using log transformation of total spent and the change to not so aggressive features.
- Constant Error Variance - Examine the scale-location plot. Homoskedasticity show up on this plot as a flat smoothing curve.
- Normal Distribution of Errors - The residual histogram and relationship plot suggest that some effects exist and that they may be roughly linear.

The final regressions of the form:

$$\widehat{log(TotalSpent)} = \beta_0 + \beta_1 \cdot HotelCity + \beta_2 \cdot HBMeal + \beta_3 \cdot onlineTA + \beta_4 \cdot roomG + \beta_5 \cdot NoDeposit + \mathbf{Z}\gamma$$

where $TotalSpent$ is the amount of money spent with one booking (average daily rate * number of nights stayed), β_1 represents the bookings with city hotel, β_2 represents the bookings with halfboard meal, β_3 represents the bookings went through online travel agency, β_4 represents the bookings with room type G assigned, β_5 represents the bookings required no deposit, β_6

$\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

Results

As seen in our stargazer table, we were left with City Hotel, Half Board Meal, Online Travel Agent Market, Tour Operators/Travel Agents Distribution, Assigned Room Type G, and No Deposit. Interpreting the coefficients, we can look at the City Hotel variable, which holds a coefficient of 0.075. This means that opting for a city hotel experience is linked to an incremental increase in the log of money spent, which could likely be attributed to the allure of city-based accommodations and proximity to different sightseeing locations and amenities, as well as higher cost per square feet due to population density. The half board meal option also correlates to 0.557 increase to log of money spent, meaning this could reflect an increased experience of inclusive culinary experience of half board meals. The online travel agent variable could align to the convenience, transparency, and specific targeted matches of hotel stays with the direct customer, increasing amount spent due to alignment. Lastly, the no deposit variable helps customers spend more on hotel bookings because of the perceived financial flexibility and reduced commitment that is found with a deposit requirement, encouring customers to spend more during their stay.

Limitations

Consistent regression estimates require an assumption of independent and identically distributed (iid) observations. IID in the context of our research study is that each booking's spending is not influenced by previous or bookings in the future, and the distribution of spending is consistent between bookings. Collinearity was examined previously between variables, which occurs when predictors are correlated, which was satisfied through the VIF of 1.5. There are various external factors that could significantly impact the validity of our study. First and foremost, because the dataset is limited to two cities and hotel types within Portugal, there could be a generalization bias around location. Global location is likely a primary factor in a consumer spending on hotels, with additional complexities resulting from the limited scope of our data set. Date of collection of data is also a factor that we need to consider as having an impact on our study. As the data was collected between July 2015 and Aguust 2017, the post COVID world that is today has had a significant impact on the hotel industry. Thirdly, some other external factors that can influence hotel bookings is likely around economic trends (e.g., Global Recession), events (e.g, Corona Virus Pandemic), and even potential weather and natural disasters. These external factors are likely significant drivers for an individuals spend in relation to hotel bookings.

Conclusion

This study estimated the economic value of certain features in hotels that impact customer spending. The regression models provide insights into which features contribute to increased customer spend for a single booking. Hotel owners may want to know the financial benefit of offering certain room types or deposit types and which market segment is likely to bring in the top spender. The ultimate goal of this line of work is to provide consistent, accurate takeaways and analysis for hotel investors and owners to make the best decisions to maximize revenue potential.

Table 1: Estimated Regressions

	Output Variable: Dollar Amount/ log(Dollar Amount)		
	total_spent		log(total_spent)
	(1)	(2)	(3)
City Hotel	−62.23***		0.09***
	(2.46)		(0.01)
Assigned Room Type A	−109.28***		
	(2.42)		
Resort Hotel		82.06***	
		(2.42)	
Half Board Meal			0.60***
			(0.01)
Online Travel Agency			0.32***
			(0.01)
Assigned Room Type G		239.09***	0.51***
		(7.84)	(0.02)
Market in Aviation	87.01***	76.06**	
	(23.28)	(23.60)	
No Deposit	159.88***	186.63***	0.24***
	(21.13)	(21.42)	(0.05)
Bed and Breakfast	−118.09***		
	(2.76)		
days in waiting list	−0.42***		
	(0.08)		
Full Board Meal		363.84***	
		(18.20)	
Constant	384.80***	129.70***	5.00***
	(21.25)	(21.46)	(0.05)
Observations	73,321	73,321	73,321
R^2	0.06	0.04	0.09
Adjusted R^2	0.06	0.04	0.09
Residual Std. Error	311.62 (df = 73314)	316.05 (df = 73315)	0.80 (df = 73315)
F Statistic	839.88*** (df = 6; 73314)	572.20*** (df = 5; 73315)	1,412.03*** (df = 5; 73315)
Note:			*p<0.05; **p<0.01; ***p<0.001