Analysis of Hotel Room Pricing In the Indian Market

By :
Chirag Gupta
Chirag.gupta364@gmail.com
IIT (ISM) Dhanbad

1. Introduction

The aim of this report is to analyse the various factors involved in the pricing strategy of hotels in the Indian Market. The data set we are working on is a sequential data set consisting of details of various hotels taken over a range of 42 cities in 8 different states. This report addresses the issues concerning the "prices of Hotel rooms" with respect to the Indian market and various cities of India.

In this report, we have to investigate whether the internal factors like swimmingpool, freebreakfast, freewifi, starrating , hotelcapacity have any effects on room rents? Whether external factors like population, metrocity, weekend and touristdestination also have any effect compared to non metro city, non-weekend and no tourist spot? The second issue concerns that what are the most important factors and least important factors which affect hotel room rents?

This field study empirically investigates the pricing of hotel rooms located in different cities of India. We estimate a regression of hotel room prices in a mixed-model framework. Our model accounts for both fixed-effects and random-effects, controlled for unobserved heterogeneity. Our analysis reveals a significant factors embedded in hotel room rent among different cities of India.

2. Review of Dataset

The data set we are working on here is a sequential data set with records for consecutive days for 42 different cities all over the country. The dataset is compiled from www.hotels.in and aggregates the hotel prices on 8 different dates at different hotels across these different cities.

- **RoomRent:** To set a common room type for correct comparative analysis, we study the Room Rents for the cheapest room of each hotel with double occupancy for a night. Roomrent values are specified in Indian rupees (INR).

The dataset captures some of these **external factors**, as explained below**-**

- **Date** - We have hotel room rent data for the following 8 dates for each hotel: { Dec 18, Dec 21, Dec 24, Dec 25, Dec 28, Dec 31, Jan 4, Jan 8}.For Simplicity, Dates are converted into numeric values using $dummy\ varaibles \in \{1,2,3,4 \dots ,8\}$ respectively.
- **IsWeekend** - Whether the day indicated by the date falls on a weekend or not. We use '0' to indicate week days, '1' to indicate weekend dates (Sat / Sun)
- **IsNewYearEve** - Whether the day indicated by the date is on New Year's Eve. We use '1' for Dec 31, '0' otherwise
- **CityName** - Name of the City where the Hotel is located
- **Population** - Population of the City in 2011
- **IsMetroCity** - Whether it is a Metro City or not. We use '1' if CityName is {Mumbai, Delhi, Kolkata, Chennai}, '0' otherwise
- **IsTouristDestination** - Whether it is a tourist destination or not. We use '1' if the city is primarily a tourist destination, '0' otherwise

- **CityRank:** Specifically, we used a dummy variable CityRank to index the 42 cities, where $index \in \{0,1,2,3,4 \dots\}$ are given to 42 cities.

  *** CityRank will coverup the effect of CityName.*

The dataset captures some of these **internal factors**, as explained below

- **Name of the Hotel**
- **Hotel Address**
- **Hotel Pincode**
- **Hotel Descriptions**
- **Star Rating -** In India, the Ministry of Tourism has formulated a scheme for classification of operational hotels using a "Star" rating. Hotels are rated as either 5 Star, 4 Star, 3 Star, 2 Star or 1 Star. Accordingly, we classified the hotels in our dataset using their star rating.
- **Airport** - Distance of hotels from the Airport(in Kms.).It is possible that hotels located close to the airport are able to charge a price premium for the greater convenience and easy access. In order to control for this alternate explanation, we recorded the distance between a given hotel and the closest airport.

- **Hotel Capacity -** Ultimately, the number of rooms in a hotel denotes the available supply and it is expected that this will keenly influence the price that a hotel will set. Accordingly, we used this as a control variable to account for the possibility that the room price set by a hotel may depend upon the supply of available rooms.
- **Free Wifi** – we use 1' if the hotel offers Free Wifi, '0' otherwise.
- **Free Breakfast** – we use 1' if the hotel offers Free Breakfast, '0' otherwise.
- **Has Swimming Pool** – we use 1' if they have a swimming pool, '0' otherwise.

- **HotelRank:** Specifically, HotelRank is a dummy variable created to index the 1670 hotels so that it can be used for analysis as HotelName, Address and description , where $index \in \{1,2,3,4 \ldots 1670\}$ are given to 1670 different hotels randomly.

  ***HotelRank will cover up the effect of HotelName , Address, Pincode, Descriptions.*

3.Tools Used

The entire analysis of this dataset was done on RStudio Version 1.0.143 which has the underlying R Language version 3.4.1 . RStudio provides a very useful tool for working with R and is the most widely used enterprise ready tool used in the industry for comprehensive analysis. The dataset was obtained as csv files segregated by cities. They were then consolidated in R into a single data frame to run analysis on it collectively.

4. Hypothesis

We study how the price of a room at a hotel is affected by external and internal factors. We will frame our hypothesis based on these factors as well as with other binary functions which may affect the room rent of a hotel. Therefore, we make the following hypothesis.

Hypothesis Testing:-

1. **H1:** *The RoomRent of Hotels which having swimmingpools are higher.*
   - Performing one tailed t test, Since p-value is 2.2e-16 which is much less than 0.05 suggests there is significant difference between the means of our sample populations and we would reject our null hypothesis. This means hotels with swimming pools have higher room prices than hotels without swimmingpools.
2. **H1:** *The RoomRent of hotels in city having tourist destination are higher.*
   - Performing one tailed t test, Since p-value is 2.2e-16 which is much less than 0.05 suggests there is significant difference between the means of our sample

populations and we would reject our null hypothesis. This means hotels with in city having tourist destination have higher room prices.

3. **H1:** *The RoomRent of Hotels having free breakfast are higher.*
   - Performing one tailed t test, Since p-value is 0.8367 which is much greater than 0.05 suggests there is no significant difference between the means of our sample populations and we can not reject our null hypothesis. This means hotels having free breakfast have not much effect on room prices.
4. **H1:** *The RoomRent of Hotels in non Metro city are higher than in metro city.*
   - Performing one tailed t test, Since p-value is 2.2e-16 which is much less than 0.05 suggests there is significant difference between the means of our sample populations and we would reject our null hypothesis. This means hotels with in non metro city have higher room prices than metro city.
5. **H1:** *The RoomRent of Hotels having free wifi are higher than paid or no wifi.*
   - Performing one tailed t test, Since p-value is 0.2212 which is much greater than 0.05 suggests there is no significant difference between the means of our sample populations and we can not reject our null hypothesis. This means hotels having free wifi have not much effect on room prices.
6. **H1:** *The RoomRent of Hotels on newyeareve are higher than other days.*
   - Performing one tailed t test, Since p-value is 1.5e-05 which is much less than 0.05 suggests there is significant difference between the means of our sample populations and we would reject our null hypothesis. This means hotels room prices are higher on newyeareve.
7. **H1:** *The RoomRent of Hotels on weekends are higher than weekdays.*
   - Performing one tailed t test, Since p-value is 0.302 which is much greater than 0.05 suggests there is no significant difference between the means of our sample populations and we can not reject our null hypothesis. This means hotels room prices have not much effect on weekends than weekdays.

5. BORUTA ANALYSIS

Boruta is a package in R which can select variables and features and is capable of working with any classification method that outputs variable importance measure (VIM). Boruta performs a top-down search for relevant features and variables by comparing their attributes' importance with importance achievable at random w.r.t. a fixed variable. It eliminates variables one by one and lists out the most important variables that correlate to a certain variable.

As per the Boruta Test run, 11 attributes confirmed important: Airport, CityRank, FreeBreakfast, FreeWifi, HasSwimmingPool and 6 more;

Importance of features: StarRating> Airport > Hotel Capacity > HasSwimmingpool > HotelRank > CityRank > Population > TouristDestination > FreeBreakfast> Metrocity > Freewifi

3 attributes confirmed unimportant: Date, IsNewYearEve, IsWeekend

6. Correlation

For Finding relations of each variable with Room rents of hotel.The Corrgram was plotted for all the variables in the data set. We found these significant relationships/inferences from corrgram:

- Strong positive correlation between RoomRent and Star Ratings.
- Strong positive correlation between RoomRent and HasSwimmingPool.
- Positive correlations are seen in case of RoomRent with Airport and Hotel Capacity.
- Positive correlations are seen in case of RoomRent with Hotel Rank and CityRank,.
- Negative correlations are seen in case of RoomRent with Population and TouristDestination.

After viewing the correlogram plots of all the variables in the table we observed some variables like Star ratings, Hotel capacity, Airport and swimming pool have a strong relation between each other and might be crucial in creating a model for prediction of room rents.

7. Regression Model

We analyzed the research question using three nested models.

Model 1: We first established the effect of most important features StarRating , HotelCapacity , Airport on the price of a room in a hotel with the simplest model we could come up with. We regressed the room rent of hotels on the specified variables, as follows.

$$RoomRent = \alpha_0 + \alpha_1 * StarRating + \alpha_2 * HotelCapacity + \alpha_3 * Airport + \alpha_4 * HasSwimmingPool + \in \quad \text{--- (1)}$$

Model 2: Next, as a robustness check, we defined a detailed model accounting for all the independent variables present for analysis, which may also influence the variation in hotel room prices. Our revised regression model was as follows:

$$RoomRent = \alpha_0 + \alpha_1 * StarRating + \alpha_2 * HotelCapacity + \alpha_3 * Airport \\ \alpha_4 * HasSwimmingPool + \alpha_5 * HotelRank + \alpha_6 * Population + \alpha_7 * IsMetroCity + \\ \alpha_8 * IsTouristDestination + \alpha_9 * Date + \alpha_{10} * Freewifi + \alpha_{11} * IsNewYearEve + \\ \alpha_{12} * CityRank + \alpha_{13} * IsWeekend + \alpha_{14} * FreeBreakfast + \in$$

--- (2)

We estimated Model 2 using linear least squares and decided the important factors which have significant p- values and irrelevant factors which have p-value greater than 0.05.

Model 3: Next, for final model ,We have removed irrelevant variables like CityRank, IsWeekend, FreeBreakfast.So, Our revised regression model was as follows:

$$RoomRent = \alpha_0 + \alpha_1 * StarRating + \alpha_2 * HotelCapacity + \alpha_3 * Airport \\ \alpha_4 * HasSwimmingPool + \alpha_5 * HotelRank + \alpha_6 * Population + \alpha_7 * IsMetroCity + \\ \alpha_8 * IsTouristDestination + \alpha_9 * Date + \alpha_{10} * Freewifi + \alpha_{11} * IsNewYearEve + \in$$

--- (3)

We estimated Model 3 using linear least squares.We expected that rerunning the regression with the three less independent variables would fit the data better. Recall that the Akaike Information Criterion (AIC) developed by Akaike, (1974) and the Bayesian Information Criterion (BIC) developed by Schwarz (1978), represent the trade-off between the goodness of fit of the model and the complexity of the model. If Model 3 indeed fits the data better than Model 2, we expected the AIC and BIC of Model 3 to be less than Model 2.

Results

From the regression analysis of Model 2 and 3(Table 1) which also yielded statistical support for our hypothesis. Recall that Model 3 extended Model 2, by including three less independent variables, as shown in equation (3). This regression analysis for model 3 also yielded α_j>0, with p <0.05, as shown in Table 1 where $j \in \{1,2,3,4 ... 11\}$. As expected, we observed a positive relationship between the hotel room prices and the hotel star ratings,

$\alpha_1 > 0$, with $p < 0.0001$. Model 3 fit the data better than Model 2, as indicated by the AIC. The AIC of Model 3 was less than the AIC of Model 2. Overall, we found Model 3 to be better than Model 2 in explaining the relationship between roomrents and different factors.

7. Conclusion

The most significant variables in the analysis of Hotel Pricing of 42 Cities was found to be the following:-

- StarRating
- Airport
- Hotelcapacity.
- Has SwimmingPool

The least significant variables which are not used in the regression analysis of Hotel Pricing of 42 Cities was found to be the following:-

- IsWeekend
- FreeBreakfast
- CityRank

Final Variables used in Regression Analysis are StarRating, HotelRank, HotelCapacity, Airport, HasSwimmingPool, Populatio, IsMetroCity, IsTouristDestination, Date, FreeWifi, IsNewYearEve. These variables have significant t values and the P-value of the proposed model was found to be 2.2e-16 (<0.05).So, this model is acceptable and the Adjusted R-Squared value of best model is 0.1987 and Multiple R-Squared value is 0.1994.

Table 1: Regression Analysis of the Hotels room rent study

```
# model 2 - with all features in "hoteldata"
fit2<-lm(RoomRent ~ . , data=hoteldata)
summary(fit2)

##
## Call:
## lm(formula = RoomRent ~ ., data = hoteldata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12273  -2290   -702   1140 309442
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -9.786e+03  4.691e+02 -20.861  < 2e-16 ***
## Population             -1.130e-04  3.573e-05  -3.162 0.001570 **
## CityRank                3.011e-01  1.029e+01   0.029 0.976656
## IsMetroCity            -7.299e+02  2.153e+02  -3.390 0.000702 ***
## IsTouristDestination    2.007e+03  1.474e+02  13.613  < 2e-16 ***
## IsWeekend              -3.740e+01  1.248e+02  -0.300 0.764337
## IsNewYearEve            7.311e+02  1.884e+02   3.880 0.000105 ***
## Date                    7.305e+01  2.603e+01   2.806 0.005023 **
## StarRating              3.531e+03  1.103e+02  32.022  < 2e-16 ***
## Airport                 9.338e+00  3.154e+00   2.961 0.003073 **
## FreeWifi                5.030e+02  2.230e+02   2.255 0.024138 *
## FreeBreakfast           4.462e+01  1.231e+02   0.362 0.717108
## HotelCapacity          -1.018e+01  1.028e+00  -9.907  < 2e-16 ***
## HasSwimmingPool         2.044e+03  1.610e+02  12.699  < 2e-16 ***
## HotelRank               1.389e+00  1.184e-01  11.728  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6565 on 13217 degrees of freedom
## Multiple R-squared:  0.1994, Adjusted R-squared:  0.1985
## F-statistic: 235.1 on 14 and 13217 DF,  p-value: < 2.2e-16


# model 3 - best fit model
fit3 <- lm(RoomRent ~ . - CityRank - FreeBreakfast - IsWeekend,
data=hoteldata)
summary(fit3)

##
## Call:
## lm(formula = RoomRent ~ . - CityRank - FreeBreakfast - IsWeekend,
##     data = hoteldata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12263  -2287   -704   1142 309399
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            -9.793e+03  4.254e+02 -23.020  < 2e-16 ***
## Population             -1.140e-04  2.252e-05  -5.061 4.23e-07 ***
## IsMetroCity            -7.235e+02  2.121e+02  -3.411 0.000650 ***
## IsTouristDestination    2.005e+03  1.368e+02  14.650  < 2e-16 ***
## IsNewYearEve            7.127e+02  1.784e+02   3.995 6.50e-05 ***
## Date                    7.427e+01  2.570e+01   2.890 0.003862 ** 
## StarRating              3.533e+03  1.100e+02  32.122  < 2e-16 ***
## Airport                 9.437e+00  2.701e+00   3.494 0.000478 ***
## FreeWifi                5.148e+02  2.206e+02   2.334 0.019625 *  
## HotelCapacity          -1.021e+01  1.023e+00  -9.985  < 2e-16 ***
## HasSwimmingPool         2.042e+03  1.592e+02  12.832  < 2e-16 ***
## HotelRank               1.393e+00  1.180e-01  11.807  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6564 on 13220 degrees of freedom
## Multiple R-squared:  0.1994, Adjusted R-squared:  0.1987 
## F-statistic: 299.2 on 11 and 13220 DF,  p-value: < 2.2e-16
```

```
# model 3 is equivalent to (RoomRent ~ StarRating + HotelRank + Airport + 
HotelCapacity + HasSwimmingPool + Population+
#              IsMetroCity + IsTouristDestination  + Date+ FreeWifi  + 
IsNewYearEve ,data = hoteldata)


# AIC of model(3)
AIC(fit2)

## [1] 270173.3


# AIC of model(3)
AIC(fit3)

## [1] 270167.5
```

#Coefficents of the best fit model (3)

```
fit3$coefficients

##          (Intercept)           Population          IsMetroCity 
##        -9.793152e+03        -1.139882e-04        -7.235076e+02 
## IsTouristDestination         IsNewYearEve                 Date 
##         2.004711e+03         7.127156e+02         7.426967e+01 
##           StarRating              Airport             FreeWifi 
##         3.532794e+03         9.437204e+00         5.147855e+02 
##        HotelCapacity      HasSwimmingPool            HotelRank 
##        -1.021488e+01         2.042397e+03         1.392768e+00 
```