**Analysis of Hotel Pricing in Indian Market**

**Purbasha Chatterjee**

[**pchatterjee10@gmail.com**](mailto:pchatterjee10@gmail.com)

**Oregon State University**

**Introduction**:

Whenever Indian customers look for booking a room in a hotel, they tend to research over different prices of a room and other offerings provided by the hotel. Certain metrics influences the price of a hotel. Hotels tend to offer pricing policies that influences a guest perception (Anna S. Mattila, 2005).

This paper captures the analysis of the hotel pricing. Hotel pricing refers to the comparisons of prices of different hotels across different parts with respect to certain other metrics such as star rating, locations, event such as weekend or New Year / Christmas, Free WiFi, complimentary breakfasts, etc. Nowadays, we find several websites such as trivago, expedia which compares the prices of hotel with each other and also with other websites to provide the customer with their best suited results. This strategy improves the hospitality sector attracting highest Foreign Direct Investment (IBEF, 2017)

In this project, hotels across 42 different cities has been analyzed to compare the prices. It analyzes the strategies and metrics which directly influences the price of a room in a hotel.

**An empirical field of study of hotel pricing strategy in India:**

It is studied that how the price of a room of hotels are influenced by certain number of factors.

It was observed that hotel capacity, presence of swimming pool and star rating influenced the pricing of the room of a hotel. Some of the hypothesis were considered to test their statistical significance. They are :

* *Hypothesis is average room rent is with higher star rating hotels is more than the ones with low star rating*
* *Hypothesis is tourist destination has more population*
* *Hypothesis is higher city rank cities have more number of star rating hotels*
* *Hypothesis average hotel capacity is always more for star rating hotels*
* *Hotels closer to airport has higher star ratings*
* *Hotels closer to airport has higher room rent*

**Data:**

The suggested dataset has been collected from [www.hotel.in](http://www.hotel.in). The dataset includes features like room rent, star rating, city name, city rank, population, isweekend, hotel names, closeness to airport, hotel description, etc. It includes several hotels across 42 different cities of India. The collected data contains the features of hotel with respect to 8 different dates. Features of the datset were analyzed deeply and their relations were identified. The focus on the features were made depending upon its correlation with other features and its level of influence to hotel’s room tariff.

**Star rating** – Several factors influence the rating of a hotel such as customer service, quality of rooms, availability of WiFi, complimentary breakfast, infrastructure, etc. In this dataset rating from 1-5 was considered and analyzed that maximum count of hotels lies in the star rating of 3.

**Hotel Capacity** – Most of the high rating hotel has higher hotel capacity but some of the 3 rating hotels were found to have occasionally higher capacity than the hotel with star rating 4.

**Presence of swimming pool –** Hotels with rating 4 and 5 are all supported with a swimming pool. Some of the hotels with star rating above 3 that is 3.5 have swimming pool.

**Approach:**

Initially, my approach towards the problem was to analyze the data, hence I started with describing dataset which would reflect the mean, standard deviation, variance and median. This was followed by analysis of the corrogram graph. It was observed that room rent, star pricing, presence of swimming pool and hotel capacity serves strong correlation among each other. IsWeekend and IsNewYear also seemed to share strong correlation with each other and to a certain level of positive correlation with room rent, star renting, presence of swimming pool and hotel capacity but the former shared the strongest correlation. Hence, room rent was considered as the response variable and star rating, hotel capacity and presence of swimming pool as the predictor variable.

Once, the variables were figured out, linear regression model was applied. The model was studied carefully to note the observations. Certain hypothesis were also considered and t-test was applied to verify those hypothesis. Please refer the section below for the checked hypothesis. Followed by this Random Forest regression (Biau, 2012) was applied to further analyze the data.

**Observations:**

For the above mentioned hypothesis following results were observed:

*Hypothesis is average room rent is with higher star rating hotels is more than the ones with low star rating –*

The p-value was observed to be less than 0.05. Hence, we could say that they are statistically significant and we can reject the null hypothesis. 95 percent confidence interval:  
5345.575 5595.491

*Hypothesis is tourist destination has more population –*

We also observe p-value to be less than 0.05 in this case and hence we can reject the null hypothesis. 95 percent confidence interval: -4489400 -4344272

*Hypothesis is higher city rank cities have more number of star rating hotels –*

The p-value was observed to be less than 0.05 and hence the hypothesis could be rejected. 95 percent confidence interval: -11.60542 -11.14419

*Hypothesis average hotel capacity is always more for star rating hotels –*

The hypothesis was rejected because of p-value less than 0.05.95 percent confidence interval: 57.74628 60.35913

*Hotels closer to airport has higher star ratings –*

The hypothesis was rejected because of p-value less than 0.05.95 percent confidence interval: 17.31176 18.08786

*Hotels closer to airport has higher room rent –*

The hypothesis was rejected because of p-value less than 0.05.95 percent confidence interval: -5577.792 -5327.875

**Experimental Setup:**

Initially linear regression model was applied with response variable as room rent and predictor variable as star rating, hotel capacity and presence of swimming pool. The relation was observed as:

RoomRent = -6896.154 + 3597.32177\*StarRating + (-15.55769 )\*HotelCapacity

+ 2528.88529 \* HasSwimmingPool + **ϵ**

The residual standard error was 6710 and F-statistic as 857.5 on 3

This implies that with every 1 increment in star rating there is increment in room rent by Rs 3597.32177, whereas with 1 increment in hotel capacity there is decrement by 15.55769. With every increment in swimming pool, there is increment in room rent by 2528.88529

F- statistics denotes the level of difference between the mean of the samples. Taking the probability of F could gives the probability that the null hypothesis for a model could be true.

Summary of the model:

Residuals:

Min 1Q Median 3Q Max

-10804 -2295 -946 1002 310110

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6896.154 340.549 -20.25 <2e-16 \*\*\*

hotels.df$StarRating 3597.322 111.670 32.21 <2e-16 \*\*\*

hotels.df$HotelCapacity -15.558 1.006 -15.47 <2e-16 \*\*\*

hotels.df$HasSwimmingPool 2528.885 157.894 16.02 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6710 on 13228 degrees of freedom

Multiple R-squared: 0.1628, Adjusted R-squared: 0.1626

F-statistic: 857.5 on 3 and 13228 DF, p-value: < 2.2e-16

Another linear regression model was applied to weigh the relation between star rating, hotl capacity, presence of swimming pool and distance from the airport. Following relation was observed :

RoomRent = -7288.048 + 3522.99\*StarRating + (-14.77562)\*HotelCapacity

+ 2708.40013 \* HasSwimmingPool + 25.34377 \* Airport + **ϵ**

The residual standard error was 6687 and F-statistic as 671.7 on 4

This implies that with every 1 increment in star rating there is increment in room rent by Rs 3522.99, whereas with 1 increment in hotel capacity there is decrement by 14.77562. With every increment in swimming pool, there is increment in room rent by 2708.40013 and with 1km closeness to airport there is increment by 25.34377

Summary of the model:

Residuals:

Min 1Q Median 3Q Max

-10785 -2265 -876 982 310437

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7288.048 341.691 -21.329 <2e-16 \*\*\*

hotels.df$StarRating 3522.990 111.531 31.588 <2e-16 \*\*\*

hotels.df$HotelCapacity -14.776 1.006 -14.695 <2e-16 \*\*\*

hotels.df$HasSwimmingPool 2708.400 158.397 17.099 <2e-16 \*\*\*

hotels.df$Airport 25.344 2.590 9.786 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6687 on 13227 degrees of freedom

Multiple R-squared: 0.1688, Adjusted R-squared: 0.1686

F-statistic: 671.7 on 4 and 13227 DF, p-value: < 2.2e-16

Finally, few more predictor variables were added to explore relation among each other and it was observed to be as follows:

RoomRent = -7455.073 + 3519.28483\*StarRating + (-14.78609)\*HotelCapacity

+ 2714.14551 \* HasSwimmingPool + 25.62737 \* Airport + 227.84314 \* FreeWiFi + -59.31294 \* FreeBreakfast + **ϵ**

The residual standard error was 6687 and F-statistic as 447.9 on 6

This implies that with every 1 increment in star rating there is increment in room rent by Rs 3519.28483, whereas with 1 increment in hotel capacity there is decrement by 14.78609. With every increment in swimming pool, there is increment in room rent by 2714.14551 and with 1km closeness to airport there is increment by 25.62737. Additionally for every increment in FreeWifi, thereis increment in room rent by 227.843 and for every increment in FreeBreakfast, there is decrement by 59.313.

Residuals:

Min 1Q Median 3Q Max

-10783 -2286 -875 967 310387

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7455.073 396.215 -18.816 <2e-16 \*\*\*

hotels.df$StarRating 3519.285 111.754 31.491 <2e-16 \*\*\*

hotels.df$HotelCapacity -14.786 1.009 -14.660 <2e-16 \*\*\*

hotels.df$HasSwimmingPool 2714.146 158.681 17.104 <2e-16 \*\*\*

hotels.df$Airport 25.627 2.604 9.840 <2e-16 \*\*\*

hotels.df$FreeWifi 227.843 226.177 1.007 0.314

hotels.df$FreeBreakfast -59.313 123.964 -0.478 0.632

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6687 on 13225 degrees of freedom

Multiple R-squared: 0.1689, Adjusted R-squared: 0.1685

F-statistic: 447.9 on 6 and 13225 DF, p-value: < 2.2e-16

Finally random forest regression was applied for further exploration and analysis of relation among the metrics.

1 decision tree was created for each ensemble. The 3 of the independent variable can assessed as shown by the importance. Proximity measure could not be calculated. Mean squared error was observed to be 500. Probably predicting the model with the test cases could enhance the result.

Summary of the model:

Length Class Mode

call 3 -none- call

type 1 -none- character

predicted 13232 -none- numeric

mse 500 -none- numeric

rsq 500 -none- numeric

oob.times 13232 -none- numeric

importance 3 -none- numeric

importanceSD 0 -none- NULL

localImportance 0 -none- NULL

proximity 0 -none- NULL

ntree 1 -none- numeric

mtry 1 -none- numeric

forest 11 -none- list

coefs 0 -none- NULL

y 13232 -none- numeric

test 0 -none- NULL

inbag 0 -none- NULL

terms 3 terms call

**Conclusion:**

All the features provided in the dataset were analyzed in detailed with the help of different graphs. Further analysis was made by applying regression model. The room rent of a hotel varies strongly with respect to the star rating, hotel capacity and presence of swimming pool. Additionally, closeness to airport along with FreeWifi and complimentary breakfast do add value to this.

describe(hotels.df$RoomRent)

vars n mean sd median trimmed mad min max range skew

X1 1 13232 5473.99 7333.12 4000 4383.33 2653.85 299 322500 322201 16.75

kurtosis se

X1 582.06 63.75

describe(hotels.df$StarRating)

vars n mean sd median trimmed mad min max range skew kurtosis se

X1 1 13232 3.46 0.76 3 3.4 0.74 0 5 5 0.48 0.25 0.01

describe(hotels.df$HotelCapacity)

vars n mean sd median trimmed mad min max range skew kurtosis se

X1 1 13232 62.51 76.66 34 46.03 28.17 0 600 600 2.95 11.39 0.67

describe(hotels.df$HasSwimmingPool)

vars n mean sd median trimmed mad min max range skew kurtosis se

X1 1 13232 0.36 0.48 0 0.32 0 0 1 1 0.6 -1.64 0

# References

Anna S. Mattila, S. C. (2005). The Impact of Hotel Pricing Policies on Perceived Fairness and Satisfaction with the Reservation Process. *Journal of Hospitality and Leisure Marketing*, 15.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research 13 (2012) 1063-1095*.

IBEF. (2017, June). *Tourism & Hospitality Industry in India*. Retrieved from ibef: https://www.ibef.org/industry/tourism-hospitality-india.aspx