Enhance the guest experience for Indian Hotel Industry using a predictive guest rating model

For Cousera Capstone Project

Anurag Nirmal

March 12, 2020

1. Introduction

1.1. Background

The Indian tourism and hospitality industry have emerged as one of the key drivers of growth among the services sector in India. Tourism in India has significant potential considering the rich cultural and historical heritage, variety in ecology, terrains and places of natural beauty spread across the country.

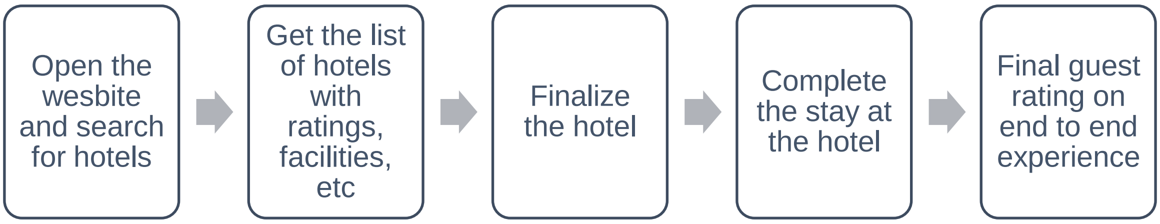
The Indian hotel industry is one of most digital advanced industry where a variety of digital tools are used for planning, booking and providing online feedback. The industry consists of international hotel chains, start-ups like Airbnb, Oyo, etc and fragmented small & medium hotels. It is important to understand the factors contributing to a guest (user or customer who stay at the hotel) experience. Efforts taken to enhance the experience will directly lead to higher business revenue, repeat business opportunity, network effect on other travellers and growth of the overall growth of the tourism industry.

1.2. The Problem Statement

The Indian hotel industry is big and dynamic comprising of big, medium and small players as per their star ratings. While the hotel star rating is important, the experience the guest or customer is visible in the hotel user rating and number of reviews left by the users. A good stay at the hotel is reflected in 5-star rating and positive review. A bad stay at the hotel is reflected in low start rating and negative review.

The project aims to understand the factors responsible (for example, hotel brand, location, point of interest, room area, etc) and their relative importance to determine the user rating. Also, comparison will be drawn against the rating provided by the website (Goibibo.com).

The guest lifecycle can be explained by this flow chart


Open the wesbite and search for hotels
Get the list of hotels with ratings, facilities, etc
Finalize the hotel
Complete the stay at the hotel
Final guest rating on end to end experience

1.3. Interest in the Project

The customer for this project who will have the maximum interest are the following

a) Hotel owners may use the model to understand the feature set and their relative importance in predicting the hotel user rating. They can focus on specific factors to improve their hotel and enhance their revenue in the process

b) Hotel website may use the model to position the hotels on the webpage which have more model rating and hence likelihood to provide better experience to prospective customers

2. Data Acquisition and Preparation

2.1. Data Sources

A pre-crawled dataset (of 4000 Indian hotels) which was created by extracting data from goibibo.com, a leading travel site from India. This dataset is available at **Kaggle.com.**

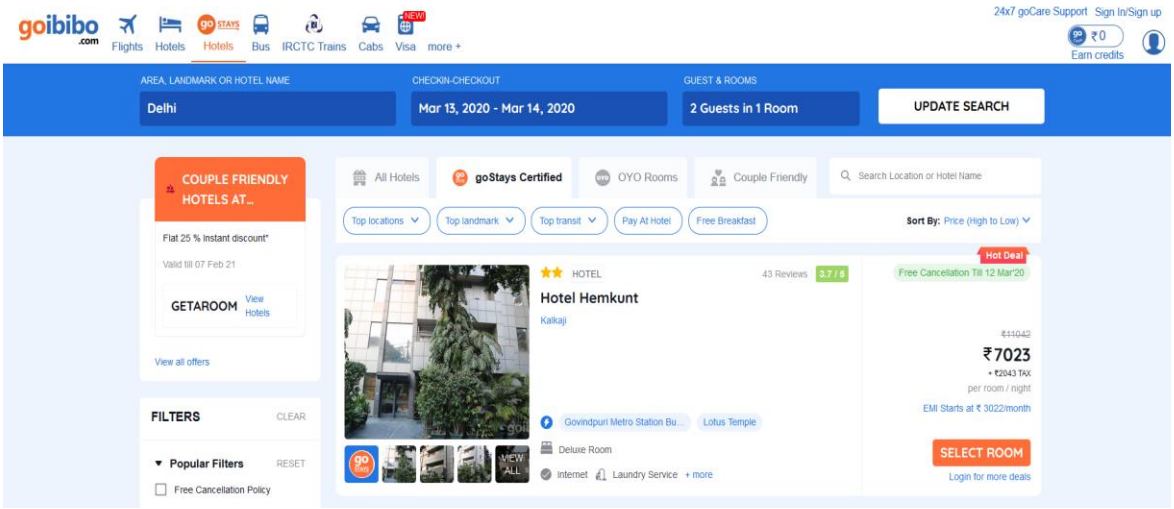
The dataset columns are noted below and explanation provided: -

Field	Description
additional_info	Additional facilities provided by the hotel like room service
address	Address of the hotel
area	Which is the area where the hotel is located
city	City where the hotel is located; multiple areas comprise of one city
country	India as the country for all hotels
guest_recommendation	Score between 0 to 100 on recommendation to future travellers
hotel_brand	Is the hotel part of a large chain or single establishment
hotel_category	Two categories Gostays or Regular, Gostays are exclusive platform properties
hotel_description	The check-in and check-out time information
hotel_star_rating	The hotel star rating: 1 to 5
image_count	The number of property or room images uploaded by the hotel
latitude	The Latitude of the property
longitude	The longitude of the property
locality	The locality of the hotel; it is subset of area which is subset of city
page_url	Direct link to the hotel on platform (website)
point_of_interest	Nearby POI where guests visit and have their photos clicked
property_id	Unique ID of the hotel
property_name	Name of the hotel
property_type	Category of the hotel: Guest House, Resort, Homestay, etc
province	The province of the hotel, a subset of city
review_count_by_category	Number of positive, negative and review with images
room_area	Area of the room where the guest stayed in
room_count	The number of rooms in the property
room_facilities	List of facilities provided by the hotel
room_type	Category of the room for which rating is provided by the user
site_review_count	The count of user ratings available on the platform (website)
site_review_rating	Average hotel rating provided by the platform (website)
site_stay_review_rating	Average guest rating on five parameters: Service Quality, Amenities, Food and Drinks, Value for Money, Location and Cleaniness
sitename	Name of the platform or website
state	Location of the property
uniq_id	Unique ID of the property

Table 1 : List of all the columns from the dataset (with explanation)

2.2. Data Cleaning and Preparation

The data cleaning and preparation is the most time-consuming process, and this requires proper understanding of the dataset. We started with visiting the Goibibo webpage to understand the process of searching & finalizing the hotel.



AREA, LANDMARK OR HOTEL NAME: Delhi

CHECKIN-CHECKOUT: Mar 13, 2020 - Mar 14, 2020

GUEST & ROOMS: 2 Guests in 1 Room

Figure 1 : Snapshot of Goibibo.com webpage (after the search query)

From Goibibo webpage, variables like images, number of reviews, hotel rating, location, type of room, room facilities like internet, laundry service and price are completely visible. These visible factors drive the customer purchase decision. The guest rating is given after completing the stay at the hotel and the experience at the hotel has more weightage in the final guest rating.

The following data preparation steps are taken

- Converting the categorical data to numerical data
- Correcting the dataframe datatypes to integer
- Leverage unstructured data in review_count_by_category, room_facilities and site_stay_review_rating

Review_count_by_category	• The column has count of reviews categorized as positive, critical and image reviews • The column was split and dataframe added back to master dataframe • New columns are – ct_pos_review, ct_crit_review and ct_img_review
Room_facilities	• The column has the highest unstructured data where room facilities like room service, telephone, hot water, kitchen, etc are clubbed together in one cell • Splitting didn't work for this column and hence we had to use a for loop to search for keywords and then create dummy variables to define the keywords • Following 15 room facilities variables are created as part of the process : Room Service, Restaurant / Breakfast / Coffee / Tea, Cable / Satellite / Pay TV, Internet, Heating, Hot Water, Shower Facility, Air Conditioning, Bathroom, Telephone, Newspaper, Refrigerator, Housekeeping, Kitchen, Extra Bed
Site_stay_review_rating	• The column has the final guest rating after completing the stay at the property. The rating is categorized into 6 rating scales ○ Service Quality ○ Amenities ○ Food and Drinks

	○ Value for Money ○ Location ○ Cleanliness • The column data was split and new columns added back to master dataframe

Table 2 : Understand of the key variables and data preparation actions on them

3. Methodology and Exploratory Data Analysis

3.1. Target Variable and Feature Selection

The target set is the final guest rating which is computed as average of 6 rating scale Service Quality, Amenities, Food and Drinks, Value for Money, Location, Cleanliness

We selected 28 variables as part of feature set, these are

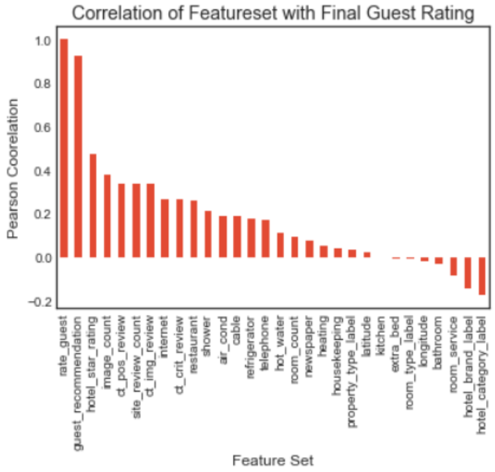
SNo	Field	Description
1	guest_recommendation	Score between 0 to 100 on recommendation to future travellers
2	hotel_brand_label	Is the hotel part of a large chain or single establishment
3	hotel_category_label	Two categories Gostays or Regular, Gostays are exclusive platform properties
4	hotel_star_rating	The hotel star rating: 1 to 5
5	image_count	The number of property or room images uploaded by the hotel
6	property_type_label	Category of the hotel : Guest House, Resort, Homestay, etc
7	ct_pos_review	Count of number of positive, negative and review with images
8	ct_crit_review	Count of number of positive, negative and review with images
9	ct_img_review	Count of number of positive, negative and review with images
10	room_count	The number of rooms in the property
11	room_type_label	Category of the room for which rating is provided by the user
12	room_service	Captured from hotel_facilities column
13	restaurant	Captured from hotel_facilities column
14	cable	Captured from hotel_facilities column
15	internet	Captured from hotel_facilities column
16	heating	Captured from hotel_facilities column
17	hot_water	Captured from hotel_facilities column
18	shower	Captured from hotel_facilities column
19	air_cond	Captured from hotel_facilities column
20	bathroom	Captured from hotel_facilities column
21	telephone	Captured from hotel_facilities column
22	newspaper	Captured from hotel_facilities column
23	refrigerator	Captured from hotel_facilities column
24	housekeeping	Captured from hotel_facilities column
25	kitchen	Captured from hotel_facilities column
26	extra_bed	Captured from hotel_facilities column
27	poi_length	Length of point of interest column string

Table 3 : List of 27 variables which are part of Featureset

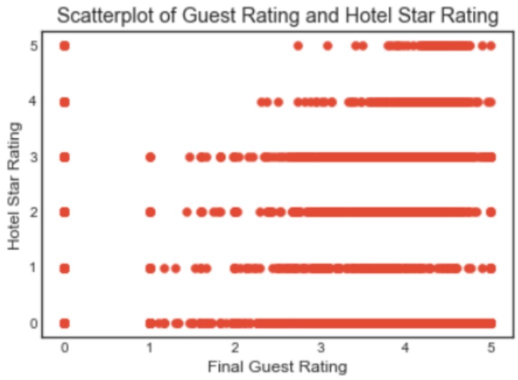
3.2. Inferences from the Data

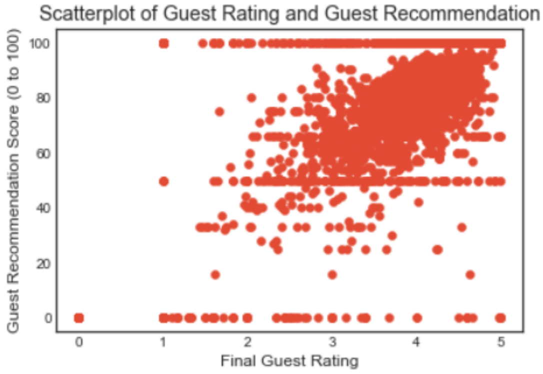
We have drawn multiple inferences between target and feature set from the dataset. The key highlights are listed below

<table>
<tr>
<td><b>Pearson Correlation –</b> We calculated the Pearson correlation coefficient for the feature set and target variable (rate_guest)

The highest correlated features are
• Guest recommendation
• Hotel star rating
• Image count
• Count of positive review
• Site review count
• Count of image reviews</td>
<td>
</td>
</tr>
<tr>
<td><b>Hotel Star Rating</b> – this is the industry established rating scale for hotels

We can see from scatter plot that there is no correlation between the final rating and star rating. Example, for hotel star rating = 1, the final rating ranges from 1 to 5

<i>Note: disregard the hotel star rating = 0 as this is case of missing data</i></td>
<td>
</td>
</tr>
<tr>
<td><b>Guest Recommendation</b> – this is the recommendation score given by guest to future travellers

There is clear link between the two which means that a guest will recommend the hotel when he rates it higher also</td>
<td>
</td>
</tr>
<tr>
<td><b>Count of Review</b> – the count is divided into positive, critical and image reviews

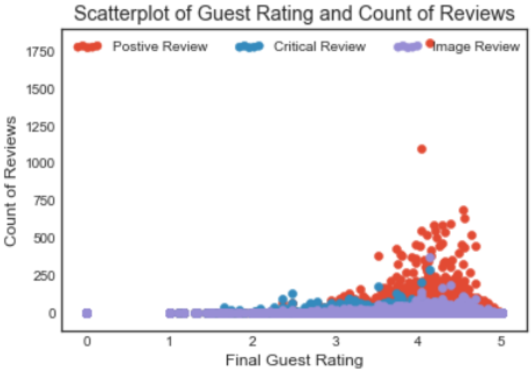
We see no clear correlation as suggested by Pearson correlation. All the positive reviews tend to be located at rating of 4 and 5 while image reviews are across 1 to 5 ratings</td>
<td>
</td>
</tr>
</table>

Table 4 : Exploratory data analysis and inferences from the data

4. Result of the Study

There are two types of Machine Learning models namely, supervised which works on labelled data and unsupervised learning which work with unlabelled data. Since the data we are dealing here is more labelled and structured, we will use supervised ML techniques for creating the predictive model.

There are two types of supervised learning

- Classification which works with category or class of the data
- Regression which are more suitable for continuous values

4.1. Classification – Supervised Learning Model

I have used classification models on our hotel dataset to determine the target variable

- K Nearest Neighbours (KNN)
- Decision Trees
- Logistics Regression
- Support Vector Machines (SVM)

4.2. Comparison of Different Models

The Decision Tree classification model has shown the best results for our data with 67% accuracy on Jaccard Score and 75% accuracy on F1 Score.

	K Nearest Neighbours (KNN)	Decision Tree	Logistics Regression	Support Vector Machine (SVM)
Jaccard Accuracy Score	0.50	**0.67**	0.57	0.63
F1 Score	0.62	**0.75**	0.68	0.73
Log Loss	-	-	0.92	-

Table 5 : Performance of classification models (best performance is red bold)

5. Discussion and Recommendations

The following decision tree is build using the Decision Tree model. The model has used **ct_pos_review** as the attribute to classify the guest rating = 0 and >0. (Note that there are 6 rating possibilities in our model 0 to 5).

Other than **ct_pos_review** which is the number of reviews written about the hotel on the website, following variables are important

- image_count – count of image reviews on the website
- poi_length – the length of point of interest (properties located close to large number of POI will be higher length)
- hot_water – whether the room had hot water in the facilities
- ct_crit_review – count of critical reviews on the website

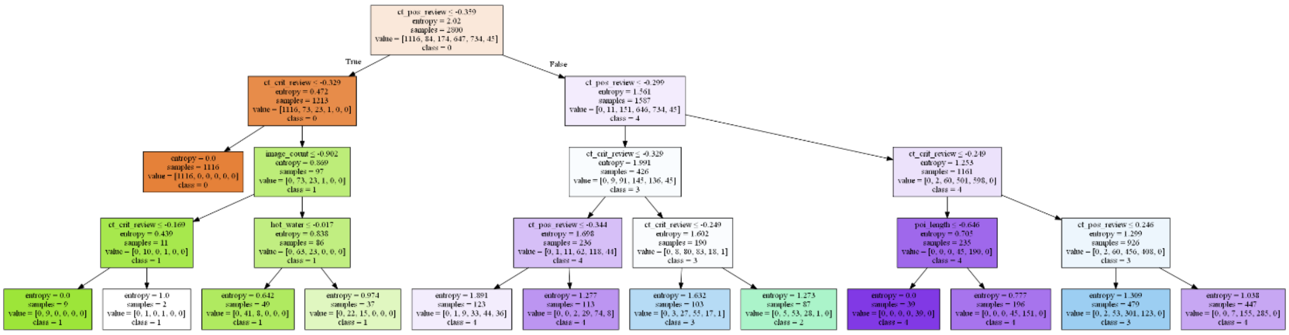


Figure 2 : Decision Tree model using complete set

We also created another model using the room_facilities (the unstructured list of facilities on the website and experienced by the guest at the hotel). The decision tree from the model is

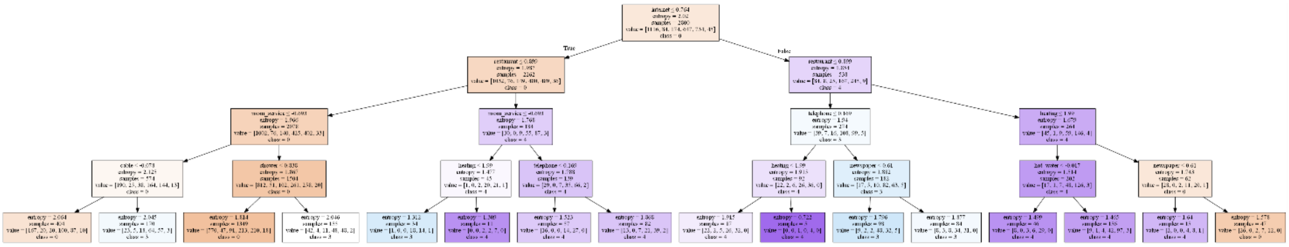


Figure 3 : Decision Tree model using room facilities feature set

Though the model accuracy is low at 30%, the key facilities which are important for guest experience are

Important to guest experience	Not important
• Internet • Restaurant • Room service • Cable • Shower • Heating • Telephone • Newspaper • Hot water	• Air conditioning • Bathroom • Refrigerator • Housekeeping • Kitchen • Extra bed

Table 6 : List of room_facilities variables categorized into important and not important

The recommendation for the hotel owners and Goibibo website is the following

- The count of reviews (positive, critical and image) are the primary drivers of guest experience. During the stay at the hotel, the most positive and negative experience is highlighted by the guest. The regular or okay experience may not be converted into a positive or critical review. Hence by increasing the number of reviews (positive, critical and image) will drive the guest experience at the hotel
- The count of images at the hotel drives the hotel selection decision and finally the guest stay experience
- The point of interest which are tourist places close to the property form a significant part of the guest rating
- From room facilities perspective, basic facilities are important to the guest like hot water and internet. Since the model accuracy was not high, a further study on large dataset is recommended

6. Conclusion

In this study, a predictive classification model was created using Decision Trees to understand the drivers of guest experience of staying at hotels. The study was primarily aimed for hotel owners and Goibibo website to undertake steps to ensure that guest has a happy stay at their hotels.

We understood that count of reviews (positive, critical and image), count of images, point of interest and room facilities are primary drivers of enhancing the guest experience at the hotel. A high accuracy (~75%) is observed from the decision tree model.

For further study, I would recommend looking at factors like point of interest, weather or climatic conditions in the city, duration of the stay and large set of room facilities to understand the guest experience.