Predicting Hotel Ratings: The Data-Driven Approach

Lokesh Kumar
Data Science and Business Analytics
Great Learning
Chennai, India
mlokesh.2911@gmail.com

Krithika Natarajan
Data Science and Business Analytics
Great Learning
Chennai, India
krithika.natarajan11@gmail.com

Parthiban P
Data Science and Business Analytics
Great Learning
Chennai, India
parthi6391@gmail.com

Vigneshwaran A
Data Science and Business Analytics
Great Learning
Chennai, India
vigneshwaranarul278 @gmail.com

Anand Mohan
Data Science and Business Analytics
Great Learning
Chennai, India
manand2992@gmail.com

Nimesh Marfatia
Data Science and Business Analytics
Great Learning
Chennai, India
nmarfatia@gmail.com

***Abstract*— The Indian Tourism and Travel industry is poised for significant growth that will result in boosting the business of Indian hotels. In the age of the internet, the primary source of information about the hotels are online booking websites that carry user reviews and ratings. Usually, the rating is a numerical figure that the user enters on the website that may not describe the entire information about the hotel. This project aims at arriving at an overall rating for the hotel from the text reviews, numerical user ratings and the presence of significant amenities offered by the hotels. The project employs web scraping and data analysis techniques to assess customer ratings and feedback from booking.com for Indian hotels. The analysis identifies high-rated amenities, areas for improvement, and factors affecting hotel success. The study aims to empower hotels with actionable insights for service enhancement, thus increasing customer satisfaction, loyalty, and business success.**

Keywords - Customer rating, Reviews, Web scraping, Data analysis

I. Introduction

In the backdrop of a remarkable surge in online hotel reservations in recent years and the flourishing Indian Tourism and Travel industry, projected to reach USD 125 billion by FY27 with international arrivals expected to hit 30.5 million by 2028, understanding the pivotal hotel attributes influencing choices becomes paramount. While extensive attention has been devoted to this topic in the realm of tourism and hospitality research, the examination of these attributes within the context of online booking remains largely uncharted as discussed by Dickinger & Mazanec, 2008[1].

This paper's overarching objective is to equip hotels with the tools to identify priority amenities and services that not only enhance guest satisfaction but also foster loyalty and stimulate positive recommendations. To achieve this, the project conducts a thorough analysis of customer feedback and ratings, offering hotels invaluable insights and recommendations for service enhancement. In doing so, it positions them for a competitive advantage, business growth, and an overall improved guest experience.

II. Literature Review

The Indian hotel industry has seen remarkable growth due to increased domestic and international travel [2]. The online booking platforms have simplified the process of search for a good hotel and booking a suitable accommodation before arrival at a travel destination. However not all hotels live up to the promise and many a times this leads to customer dissatisfaction. There could be scenarios where customers are happy about few amenities and recommend the hotel improves the other amenities. These sentiments of customers are captured in the feedback and rating shared by them on booking.com. Good reviews automatically drive business and average reviews can hamper the brand image. It becomes imperative on the part of the hotels to take action addressing customer concerns and clarifying the same on the booking portal. Also, hotels need to keep in mind to continuously monitor and keep their rating at a higher scale. Customer loyalty and word of mouth advocacy can help the hotel business to boom. Hence understanding customer preferences and feedback is vital for success.

When considering hotel ratings, customers commonly place a premium on factors such as cleanliness, comfort, security, location, value for money, and personalized services [2]. Additionally, amenities such as Wi-Fi availability, dining quality, and fitness facilities hold significant value. Customers actively engage in hotel reviews on online platforms, with a particular focus on aspects like personalized service, problem resolution, and staff conduct [2]. In this study, we have segmented hotels into three categories based on room tariffs and amenities, aiming to discern whether guests at specific hotel categories prioritize particular amenities and rate their overall experience accordingly.

In their study, Maleerat Sodanil [3] conducted a comparative analysis of three classifier methods: Naive Bayes, Decision Tree, and Support Vector Machine. Their goal was to determine which of these methods yielded the highest accuracy when applied to sentiment analysis in hotel reviews. M. Geetha, Pratap Singha and Sumedha Sinha [4] aims to establish a relationship between customer sentiments

in online reviews and customer. In order to understand the effect of word-of-mouth reviews and hotels sales Qiang Ye, Rob Law and Nib Gu [5] have used log-linear regression model and found significant relationship between the consumer reviews and business performance of the hotel.
In this paper we aim to present analysis of the customer's ratings and reviews to figure out the amenities that hold importance to the customers for various class of hotels ranging from budget to luxury.

The proposed methodology uses sentiment polarization concept of VADER sentiment analysis tool [2]. Vader sentiment analysis assigns a polarity either positive, negative or neutral and assesses text reviews, it also effectively measures the intensity of the sentiment. This is done by assigning polarity score to each text. The VADER sentiment analysis tool has been particularly picked for text review analysis as it is effective in analyzing text data with informal language and even non textual emoticons.

III. Data Preparation

A. Data Collection

The study used various web scraping techniques to collect data from www.booking.com. The project complies with ethical and legal standards for data scraping and analysis.

Initially, Instant Data Scraper, a browser extension used for web scraping was used to extract raw data from the website. The data was cleaned and significant data like the name of the hotel, city, URL of the bookings page were obtained using this method.

Using the URLs obtained for each hotel, data was further extracted from the website. An automated python script was created using the **BeautifulSoup4** library for this purpose. The script identifies the values of the desired variables using XPATH. The library was used to extract salient features of the hotel such as ratings, amenities offered, price, etc., A total of **7072** records and **23** variables were collected.

To collect the user text reviews for each hotel, another automated python script was created using **Selenium** library. A total of **98304** reviews were collected across various hotels.

B. Data Dictionary

The model adopted for predicting the hotel rating is based on Linear Regression. The data preprocessing involves the dropping of unwanted variables, variable transformation, missing value treatment, duplicate value treatment and outlier treatment to suit the modelling approach selected.

Following is the data dictionary of the data collected using the web scraping technique from the website.

TABLE I – DATA DICTIONARY

S. No	Variable	Data Type
1	ID	Integer
2	Hotel Name	Object
3	Address	Object
4	URL	Object
5	Date of Data collection	DateTime
6	City	Object
7	Number of Ratings	Integer
8	Features	Object
9	Distance from City center	Float
10	Remarks	Object
11	Metro	Boolean
12	Beach	Boolean
13	Staff	Float
14	Facilities	Float
15	Cleanliness	Float
16	Value for Money	Float
17	Location	Float
18	Free Wi-Fi	Float
19	Comfort	Float
20	Overall Rating	Float
21	Travel Sustainable Property	Object
22	Price	Float
23	No of star	Float

C. Variable Elimination

The variables such as "ID", "hotel name", "address", "URL of the hotel", "date of data collection" and "city" have been dropped because they are categorical variables that are used to identify the hotels uniquely.

"Number of ratings" has been dropped because it is just the measure of how many users have reviewed the hotel. It neither clearly explains nor influences the overall rating.

"Remarks", "Metro" and "Beach" have been dropped because there were a large number of missing values in these variables.

D. Variable Transformation

A categorical feature "Accessibility" was created to denote how accessible a hotel is from the centre of the city. This was created from an existing feature "Distance from City Centre" and it is ordinally encoded as 1,2 and 3. All hotels that are within a 6 KM radius are encoded as 3. Hotels within a 15 KM radius are encoded as 2 while the rest are encoded as 1. Furthermore, dummy encoding was done for the feature "Accessibility" while the "Distance from City Centre" was dropped.

A new feature "Type of City" was created to classify the cities into purely tourism friendly cities and a mix of business-oriented as well as tourism friendly cities. Tourism friendly cities were named as "Destinatia" while the rest were named as "Excursa". The feature was label encoded and a new Boolean variable was created as "Type of City Destinatia".

The "Travel sustainable property" feature is a badge given to hotels by the website based on the level of Travel

Sustainability that they exhibit. There are 4 levels namely ranging from 0 to 3. We have ordinally encoded the values as 0, 1, 2 and 3. Further the feature was dummy encoded and split into three Boolean columns.

The hotels offer rooms on a per night basis and the value ranges from as low as Rs. 600/- to as high as Rs 40000/-. The hotels are classified by the website by means of star ratings ranging from 1 to 5. As the customer base and the target audience of the variety of hotels selected are starkly contrasting, we have categorized the hotels into three types. A new feature "Hotel Category" was created using the "Price" and "No. of star" variables. The three categories were named as "Thrift Haven", "Tranquil Retreat" and "Grandiose Manor". After segregating the hotels into three categories, the "No. of star" and "Price" features were dropped from the dataset. The categorization logic is described in the table below.

TABLE 2 – HOTEL CATEGORIES

Category	**Star**	**Price range (in Rs)**	**Hotels in the category**
Thrift Haven	1,2	Less than 2500	3020
Tranquil Retreat	3	2500 to 5000	1618
Grandiose Manor	4,5	More than 5000	738

The "Features" column consists of the salient amenities that are offered by the hotels such as "Air Conditioning", "Free Breakfast", "Free Wi-Fi", "Cab Services", "Television", "Special Treatments" such as Spa, Salon etc., The "Features" column was extracted as a colon separated field so as to make it easy for further data extraction. The various features were converted into separate Boolean variables using python script and the top 21 features were alone retained for the analysis. The 21 features were selected after thorough consideration based on the commonality among hotels, domain knowledge, essentiality, and significance. Among the 21 features, there are some features that have been clubbed together into a single feature. For example, "Security" feature consists of "CCTV", "24-hour Security Monitoring", "Fire Alarm", etc., The features selected are predominantly the amenities that the customers commonly look for in Indian hotels. After the transformation of the amenities into multiple columns, the "Features" column was dropped.

The "Free Wi-Fi" numerical feature has a lot of missing values and thus is not a good option to retain the feature. However, we obtained a categorical variable "Free Wi-Fi" from the "Features" that describes whether the hotel offers Wi-Fi for free or not. Both the features were combined into a single categorical variable and the numerical feature was dropped.

The following table explains the clubbing of various amenities into a single variable formed from the "Features" variables.

TABLE 3 - FEATURES

Variable	**Combined Features**
Air conditioning	Air conditioning[a]
Customer Support	24-hour front desk, Tour desk, Wake-up service
Free Wi-Fi	Free Wi-Fi
Smoking Zone	Non-smoking rooms, Designated smoking area
English	English
Room service	Room service
Hindi	Hindi
Laundry	Laundry, Ironing service, Dry cleaning, Linen
Security	CCTV in common areas, 24-hour security, CCTV outside property, Safety deposit box, Security alarm, Smoke alarms
Luggage storage	Luggage storage
TV	Flat-screen TV,TV
Family rooms	Family rooms
Free toiletries	Free toiletries, Toilet paper
Lift	Lift
Cab Service	Car hire, Airport shuttle
Furnished	Wardrobe or closet, Clothes rack, Desk, Telephone, Electric kettle, Trouser press
Special Treatment	Breakfast in the room, Express check-in/check-out
Restaurant	Restaurant
Free Breakfast	Breakfast
Currency exchange	Currency exchange
View	View[a]

[a] The features on the right represent a single feature in this case, thus making parent and derived features identical

In order to include the text user reviews extracted from individual hotel webpages into the rating prediction analysis, we have performed Natural Language Processing. The NLP was performed on the text reviews using the Natural Language Processing Tool Kit Library (nltk) using Python. After performing the NLP on the texts, we have performed a sentiment analysis using Valence Aware Dictionary and Sentiment Reasoner (VADER). VADER is a lexicon and rule-based sentiment analyzer that is sensitive to the internet slang. VADER has been extensively trained on social media and internet-based text datasets to understand the sentiment and emotions of texts of online users. Hence, this approach is more suitable for our study as the reviews are found online.

The reviews were initially checked for nulls and duplicates and then a Bag of Words approach was used. The review text was converted into lower case and all the punctuations were removed. Lemmatization was performed on the review dataset using WordNetLemmatizer. Furthermore, Part of Speech tagging was performed to ensure the words are tagged as nouns, pronouns, adjectives, verbs etc. Subsequently, the stop words and trivial words were removed from the text reviews.

Upon completion of text processing, the sentiment analysis was performed using VADER. VADER uses

Sentiment Intensity Analyzer to perform the analysis and gives the Polarity of the reviews as its output. The Polarity score is a numerical value that explains how positive or how negative the sentiment of a text is.

The "Polarity" score was subsequently included to the original dataset against each hotel and was included in the prediction analysis of overall hotel ratings.

E. Exploratory Data Analysis

We identified 139 duplicate values in the dataset and have dropped them. Additionally, we found 53 duplicate hotels identified by grouping the Hotel name and City. These records were dropped as well.

TABLE 4 – MISSING VALUES

S.No	Variables	Null Values
1	Staff	18
2	Value for Money	23
3	Cleanliness	25
4	Facilities	25
5	Location	25
6	Comfort	27
7	Polarity	78
8	Price	226

We dropped rows for the columns Staff, Facilities, Cleanliness, Value for Money, Location, Polarity and Comfort as the number of nulls are negligible when compared to size of the data identified.

The distribution of hotels among the Indian cities considered is depicted in the figure below.

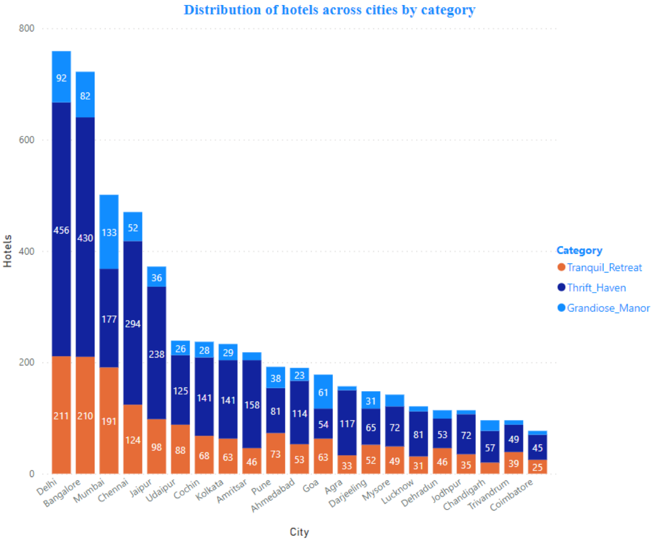


Figure 1- Hotel distribution across cities

"Price" was imputed with mean values of hotel prices. This was done because all the hotels with price as null were having either 1 or 2 stars only. Hence, they all would fall into the same category. The hotel categorization was done after the value imputation of prices.

The box plots were plotted for the dataset for all the numerical data before the hotel categorization was performed. Initially the categorization was done purely based on the "No. of Star" awarded to the hotel. However, as it can be evidently seen in the Figure 2, the "Price" values had outliers and needed outlier treatment. Hence the logic was applied as explained in TABLE 1 to ensure the outliers were removed from the dataset as visible in Figure 3. This resulted in meaningful segregation of hotels into three categories.

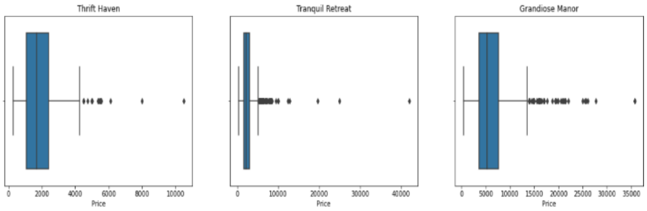


Figure 2 - Outliers in Price

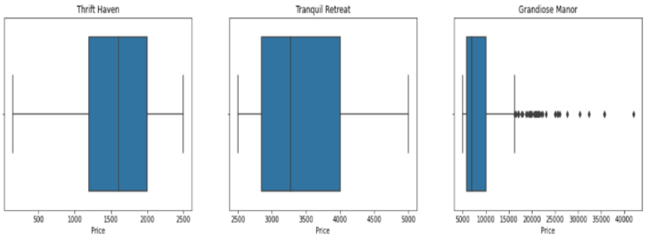


Figure 3 - Outliers treated in Price

The correlation between the numerical variables can be seen in the figure below across the three categories. The numerical variables individually were generally having a high correlation with the target variable.

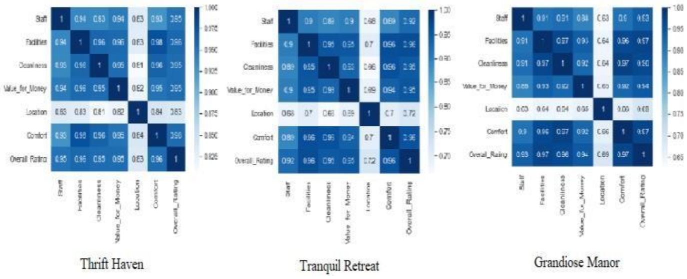
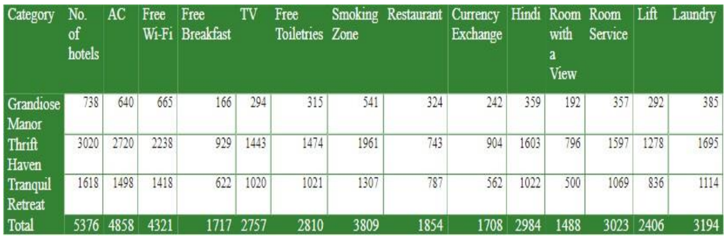


Figure 4 - Heat map for numerical variables

The figure below shows the offerings of the prominent amenities across the hotel categories that form a part of the regression results.



Category	No. of hotels	AC	Free Wi-Fi	Free Breakfast	TV	Free Toiletries	Smoking Zone	Restaurant	Currency Exchange	Hindi	Room with a View	Room Service	Lift	Laundry
Grandiose Manor	738	640	665	166	294	315	541	324	242	359	192	357	292	385
Thrift Haven	3020	2720	2238	929	1443	1474	1961	743	904	1603	796	1597	1278	1695
Tranquil Retreat	1618	1498	1418	622	1020	1021	1307	787	562	1022	500	1069	836	1114
Total	5376	4858	4321	1717	2757	2810	3809	1854	1708	2984	1488	3023	2406	3194

Figure 5 - Amenities offerings across categories of hotels

IV. Modelling

The dependent variable or the target variable for the model will be “Overall Rating”. The independent variables selected for the modelling are as follows:

Staff, Facilities, Cleanliness, Value for Money, Location, Comfort, Polarity, Type of City Destinatia, Air conditioning, Customer Support, Free Wi-Fi, Smoking Zone, English, Room service, Luggage storage, Hindi, Laundry, Free Breakfast, Restaurant, Security, TV, Special Treatment, Free toiletries, Cab Service, Family rooms, Lift, Furnished, Currency exchange, View, Travel Sustainable Property 1, Travel Sustainable Property 2, Travel Sustainable Property 3, Accessibility 2, Accessibility 3.

Separate models were created for the three hotel categories and the results were studied separately. Commonly, we have used Linear Regression, Ridge, Lasso, Elastic NET regression techniques, Random Forest, and Decision Tree Regressor.

Before proceeding with the modelling, we had to check the modelling assumptions for the regression.

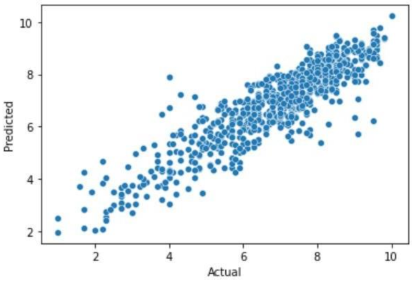


Figure 6 - Linearity Check - Predicted vs Actual

- Linearity - The relationship between the predictor variables and the target variable is linear, which is met across all the variables. Linearity check was carried out using pair plots.

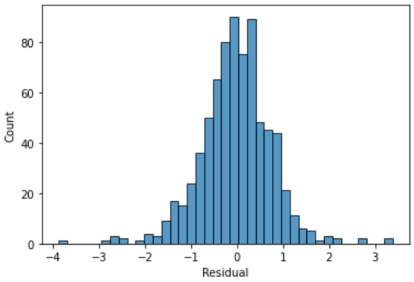


Figure 7 - Histogram of Residuals

- Normality - The residuals follow a normal distribution. The histograms of the numerical variables confirm the normality of the data distribution.

- Homoscedasticity - The spread of the residuals (the differences between the observed and predicted values) is consistent across the range of the predictor variables.

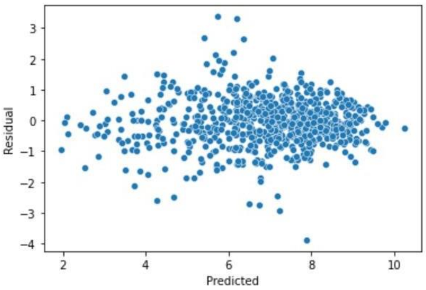


Figure 8 - Homoscedasticity check

- No multicollinearity - High multicollinearity can lead to unstable and unreliable estimates of the model coefficients. In such cases wherever the VIF value is greater than 5, we have addressed the multicollinearity issue by removing one or more correlated variables.

For all the models, the first step was model preparation and feature selection. After creating the data frame with all the essential variables, the model selection process started off with multi collinearity check. We calculated the Variance Inflation Factor to identify variables that are highly collinear. VIF was calculated only for the following numerical variables - 'Staff', 'Facilities', 'Cleanliness', 'Value for money', 'Location', 'Comfort', 'Polarity'. Step by step the variables were dropped until all remaining variables had VIF value less that 5.

Subsequently, for all the datasets, Ordinary Least Squares regression was applied to identify the R Squared and Adjusted R Squared of the variables considered at 95% confidence level. The variables were dropped one after the other based on the p-value by backward elimination technique. The process was repeated until all remaining variables had a p-value less than 0.05. The model was locked as the final model. All regression techniques were preformed on the respective datasets for the respective categories and the models were compared. All the models have been built with a train-test split ratio of 75:25. We have performed modelling using the following techniques – Linear Regression, Lasso, Ridge, Elastic Net, Decision Tree regressor and Random Forest regressor. Based on the output of all models and the explanation of independent variables we have chosen Linear Regression as the best model in all categories. The model output comparisons can be found below.

A. *Thrift Haven*

Based on VIF, the variables that were retained among all the numerical variables - 'Location', 'Comfort', ‘Polarity’.

Furthermore, based on the backward elimination technique, the following variables were retained – 'Polarity', 'Location', 'Type of City Destinatia', 'Accessibility 2', 'Free Wi-Fi', 'Air conditioning', 'Smoking Zone', 'Free Breakfast', 'TV', 'Travel Sustainable Property 1','Free toiletries'.

The values of R Squared, Adjusted R Squared, RMSE and MAPE can be found in the following tables.

TABLE 5 – THRIFT HAVEN – PERFORMANCE - TRAIN

	Train		
	R²	MAPE	RMSE
Linear Regression	**0.791**	**0.100**	**0.764**
Ridge	0.791	0.100	0.763
Lasso	0.700	0.123	0.915
Elastic Net	0.788	0.117	0.855
Decision Tree	0.973	0.030	0.273
Random Forest	0.976	0.037	0.264

TABLE 6 - THRIFT HAVEN – PERFORMANCE - TEST

	Test		
	R²	MAPE	RMSE
Linear Regression	**0.797**	**0.102**	**0.765**
Ridge	0.797	0.104	0.771
Lasso	0.694	0.130	0.948
Elastic Net	0.794	0.124	0.855
Decision Tree	0.972	0.031	0.273
Random Forest	0.880	0.070	0.570

For Thrift Haven, the R Squared and Adjusted R Squared are high for Linear regression. The RMSE and MAPE values are the lowest in Linear regression again. Although the numbers were significant in Decision Tree and Random Forest regressors, the explanation of individual independent variables was more evident in Linear Regression than the other regressors. Hence, we decided to go ahead with the Linear Regression model.

Linear Regression: The regression equation for the Multiple Linear Regression model executed for Thrift Haven is as follows:

*- 0.431 + 2.33 * Polarity + 0.722 * Location + 0.065 * Type of City Destinatia + 0.115 * Accessibility_2 + 0.159 * Free Wi-Fi + 0.599 * Air conditioning - 0.182 * Smoking Zone + 0.197 * Free Breakfast + 0.147 * TV + 0.093 * Travel Sustainable Property 1 + 0.271 * Free toiletries*

The coefficients effectively explain the importance of the selected variables. The "Polarity" of the user text reviews significantly contributes to the Overall rating of the hotel.

While establishing new hotels, consideration must be given to the Location of the hotel and its accessibility from the center of the city as the customers in this segment of hotels give importance to both these aspects.

Furthermore, focusing on establishing / improving air conditioning, free wi-fi, television will have a positive impact on the overall rating of the hotel. These are some essential features that any customer would prefer in their stay.

A sumptuous complimentary breakfast offered to the customers checking in to the hotel will help ensure favorable impression about the hotel.

Travel sustainability is a priority for many customers. Especially in an age where the world is ready to move adopt eco-friendly products and are mindful about energy usage, travel sustainability label and the corresponding practices will add more value to the hotels.

The negative coefficient of Smoking Zones explain that this section of customers does not give much preference to private smoke area in the room.

B. Tranquil Retreat

Based on VIF, the following variables were retained among all the numerical variables - 'Location', 'Cleanliness', 'Polarity'.

Furthermore, based on the backward elimination technique, the following variables were retained – 'Location', 'Polarity', 'Restaurant', 'Travel Sustainable Property 3', 'Free Wi-Fi', 'TV', 'Type of City Destinatia', 'Smoking Zone', 'Free Breakfast', 'Currency exchange', 'Hindi', 'Air conditioning'.

The values of R Squared, Adjusted R Squared, RMSE and MAPE can be found in the following tables.

TABLE 7 – TRANQUIL RETREAT – PERFORMANCE - TRAIN

	Train		
	R²	MAPE	RMSE
Linear Regression	**0.718**	**0.070**	**0.631**
Ridge	0.718	0.070	0.630
Lasso	0.543	0.092	0.802
Elastic Net	0.709	0.079	0.693
Decision Tree	0.947	0.029	0.276
Random Forest	0.968	0.025	0.217

TABLE 8 - TRANQUIL RETREAT – PERFORMANCE - TEST

	Test		
	R²	MAPE	RMSE
Linear Regression	**0.753**	**0.069**	**0.606**
Ridge	0.754	0.070	0.612
Lasso	0.587	0.093	0.792
Elastic Net	0.752	0.080	0.693
Decision Tree	0.951	0.028	0.259
Random Forest	0.798	0.058	0.529

For Tranquil Retreat, the best values of R Squared, Adjusted R Squared, RMSE and MAPE can be observed in Linear Regression. Also, the explanation of individual independent variables was evident in Linear Regression than the other regressors. Hence, we decided to choose the Linear Regression model.

Linear Regression: The regression equation for the Multiple Linear Regression model executed for Tranquil Retreat is as follows:

> *0.22 + 0.496 * Location + 2.092 * Polarity + 0.195 * Restaurant + 0.18 * Travel Sustainable Property 3 + 0.842 * Free Wi-Fi + 0.35 * TV + 0.254 * Type of City Destinatia - 0.266 * Smoking Zone + 0.307 * Free Breakfast + 0.408 * Currency exchange + 0.44 * Hindi + 0.24 * Air conditioning*

Like Thrift Haven, the "Polarity" of the user text reviews significantly contributes to the Overall rating of the hotel.

Location, aesthetics, sustainability of the property plays vital role in the rating calculation. The hotels can look to score higher in these areas. In fact, in contrast to Thrift Haven, the preference for Travel Sustainable property is higher for Tranquil Retreat.

Establishing / improving air conditioning, free wi-fi, television, provision of a quality complimentary breakfast will have a positive impact on the overall rating of the hotel.

Customers opting for Tranquil Retreat hotels are looking for additional features like Currency Exchange. This shows that the Tranquil Retreat hotels in India not only attracts Indians, but also attract foreign tourists.

In a nation where Hindi is spoken by over 70% of the population, there is a strong customer preference for Hindi-speaking staff within this hotel segment.

C. Grandiose Manor

Based on VIF, the following variables were retained among all the numerical variables - 'Location', 'Cleanliness', 'Polarity'.

Furthermore, based on the backward elimination technique, the following variables were retained – 'Location', 'Polarity', 'Restaurant', 'Travel Sustainable Property 3', 'Free Wi-Fi', 'TV', 'Type of City Destinatia', 'Smoking Zone', 'Free Breakfast', 'Currency exchange', 'Hindi', 'Air conditioning'.

TABLE 9 – GRANDIOSE MANOR – PERFORMANCE - TRAIN

	Train		
	R^2	MAPE	RMSE
Linear Regression	**0.909**	**0.028**	**0.280**
Ridge	0.909	0.028	0.279
Lasso	0.870	0.034	0.334
Elastic Net	0.902	0.050	0.580
Decision Tree	0.984	0.009	0.110
Random Forest	0.954	0.021	0.203

TABLE 10 - GRANDIOSE MANOR – PERFORMANCE - TEST

	Test		
	R^2	MAPE	RMSE
Linear Regression	**0.908**	**0.031**	**0.308**
Ridge	0.908	0.032	0.312
Lasso	0.858	0.041	0.387
Elastic Net	0.898	0.059	0.580
Decision Tree	0.980	0.011	0.130
Random Forest	0.785	0.042	0.443

For Grandiose Manor, based on the results from various models as mentioned in the tables above we have decided to go with Linear Regression model.

Linear Regression: The regression equation for the Multiple Linear Regression model executed for Grandiose Manor is as follows:

> *-1.417 + 0.876 * Value for Money + 0.108 * Location + 0.416 * Polarity + 0.079 * View + 0.176 * Room service + 0.192 * Free Wi-Fi + 0.075 * Lift + 0.843 * Laundry + 0.087 * Restaurant + 0.068 * Travel Sustainable Property 3 + 0.091 * Smoking Zone*

With higher hotel rates, customers anticipate receiving commensurate value for their expenditure. They typically seek rooms with scenic views and private smoking areas to enhance their stay.

Amenities such as prompt room service, laundry facilities, and a diverse culinary offering contribute positively to the hotel's rating. Therefore, it is advisable for hotels to prioritize the provision and upkeep of these services.

TABLE 11 – T STATISTIC VALUE FOR THRIFT HAVEN

Thrift Haven			
Parameters	**T stat**	**Parameters**	**T stat**
Location	41.9	Accessibility_2	4.4
Polarity	27.7	Free_WiFi	3.8
Air_conditioning	5.2	Type_of_City_ Destinatia	2.9
Free_Breakfast	4.8	Travel_Sustainable _Property_1	2.0
TV	4.6	Smoking_Zone	-4.7
Free_toiletries	4.5		

TABLE 12 – T STATISTIC VALUE FOR TRANQUIL RETREAT

Tranquil retreat			
Parameters	**t stat**	**Parameters**	**t stat**
Location	23.0	Hindi	4.6
Polarity	19.4	Travel_Sustainable _Property_3	3.1
Currency_ exchange	7.9	Restaurant	3.0
Type_of_City_ Destinatia	7.8	Air_conditioning	2.4
Free_WiFi	6.1	TV	2.1
Free_Breakfast	5.9	Smoking_Zone	-6.6

TABLE 13 – T STATISTIC VALUE FOR GRANDIOSE MANOR

Grandiose Manor			
Parameters	**t stat**	**Parameters**	**t stat**
Value_for_ Money	47.7	Lift	3.0
Location	5.2	Travel_Sustainable _Property_3	2.7
Free_WiFi	4.5	Restaurant	2.2
Polarity	4.3	Room_service	2.1
Laundry	4.1	View	2.0
Smoking_Zone	3.0		

The t value explains how much significant the variable is in the model. With the t value we can prioritize the parameters that has more impact on the model.

V. CONCLUSION

From the analysis one can establish that based on the segment of the preferred budget category of the rooms classified in this paper as Thrift Haven(Standard), Tranquil Retreat(Deluxe) and Grandiose Manor(Luxury) customer's expectation with the amenities offered vary. We were able to infer from the modelling and analysis that hotels catering to Thrift Haven segment of customer should consider location and accessibility from the city center, as these matter to customers in this segment. Also focusing on amenities like excellent air conditioning, free Wi-Fi, and quality television positively impacts the hotel's overall rating. In case of the customers preferring rooms in the category of Tranquil Retreat. Location, aesthetics, sustainability, and added amenities drive higher hotel ratings. Tranquil Retreat outperforms Thrift Haven in sustainability and attracts foreign tourists with services like Currency Exchange, while Hindi-speaking staff are preferred in this segment. When it comes to the luxury segment Grandiose Manor With higher hotel rates, customers expect value for their money, including rooms with good views and private smoking areas. Services like express room service, laundry, and a multi-cuisine restaurant enhance the hotel's rating, making them key areas of focus.

In all three categories, a well-located, travel-sustainable property boosts hotel rating. Customers also expect standard amenities like elevators and free Wi-Fi, which can enhance reviews and spread positive word-of-mouth.

Some action points those hotels on booking.com must pickup from analysis of customer rating and reviews –

Highlight Amenities: If certain amenities consistently receive praise in reviews, make sure to emphasize them in your marketing. For example, if guests rave about your spa, promote spa packages.

Online Reputation Management: Monitor and respond to reviews promptly, both positive and negative. Engage with customers to show that you value their feedback.

Highlight Positive Reviews: Showcase positive reviews on your website and marketing materials. Feature them prominently in areas like your homepage and booking pages.

Use Testimonials: Convert positive reviews into testimonials. Include these testimonials in email marketing campaigns and on your social media profiles.

Leverage User-Generated Content: Encourage guests to share photos and stories about their stay on social media using a unique hashtag. Repost these on your own social profiles to increase engagement and trust.

Improve Negative Feedback: Take action on negative feedback by addressing the issues raised in reviews. Show potential guests that you are committed to improving.

Employee Training: Train your staff to provide exceptionalservice, as positive interactions can lead to better reviews andrepeat business.

REFERENCES

[1] Dickinger, A., Mazanec, J. (2008). Consumers' Preferred Criteria for Hotel Online Booking. In: O'Connor, P., Höpken, W., Gretzel, U. (eds)Information and Communication Technologies in Tourism 2008. Springer, Vienna. https://doi.org/10.1007/978-3-211-77280-5_22

[2] Beny, Moha & Barakbah, Ali & Muliawati, Tri. (2020). Data Analyticsfor Hotel Reviews in Multi-Language based on Factor Aggregation of Sentiment Polarization. 324-331. 10.1109/IES50839.2020.9231625.

[3] Maliyaem, Maleerat. (2016). Multi-Language Sentiment Analysis for Hotel Reviews. MATEC Web of Conferences. 75.03002. 10.1051/matecconf/20167503002.

[4] M. Geetha, Pratap Singha, Sumedha Sinha,Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis,Tourism Management,Volume 61,2017,Pages 43-54,ISSN 0261-5177,https://doi.org/10.1016/j.tourman.2016.12.022.

[5] Ye, Qiang & Law, Rob & Gu, Bin. (2009). The Impact of Online User Reviews on Hotel Room Sales. International Journal of Hospitality Management - INT J HOSP MANAG. 28. 180-182. 10.1016/j.ijhm.2008.06.011.