A Data-Driven Approach To Enhance The Indian Hotel Industry: Sentiment Analysis & Recommendations For Improved Customer Experience

Presented By:-

Akash Dey
Arka Dhali
Ipsheetha Nath
Pallavi Mazumdar
Soumya Roy
Subhrajyoti Basak
Tanushree Mandal

Praxis Business School
CELEBRATE YOUR WORTH

BUSINESS PROBLEM

- Hotel Industry is one of the growing sectors in today's market
- Owners can approach us to understand the reasons behind declining customer experience in terms of different service parameters
- Our analysis can help identify which service parameters are critical in improving customer experience
- By addressing these parameters, owners can potentially improve customer satisfaction and loyalty

Technical Overview

Analysis Of Hotel Reviews

Using LDA Topic Modelling To Extract Topics From The Reviews

Building Dashboard Based On Analysis

CONTENT

01 Methodology

02 Model Implementation

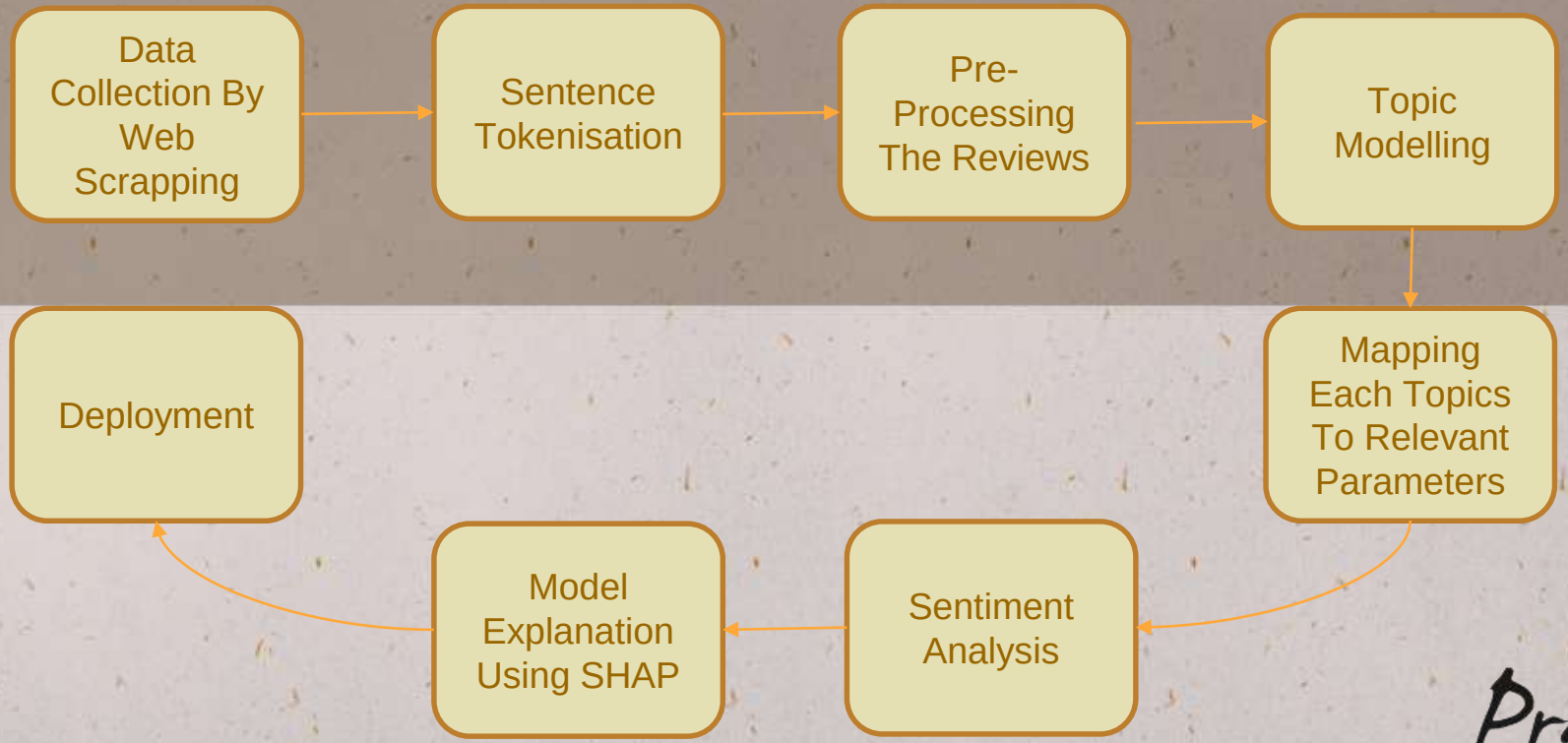
03 Observation

04 Suggested Impact

Praxis Business School
CELEBRATE YOUR WORTH

METHODOLOGY

WORKFLOW



Data Collection By Web Scrapping

- Scraping hotel reviews from Trip Advisor Website
- Extracting reviews of 185 hotels across almost all states in India
- Total reviews scraped: 44,362
- Features extracted: Reviews and Ratings

```
import pandas as pd
data = pd.read_csv('/content/drive/MyDrive/capstone project/Indian_hotel_reviews_v2.csv')
```

```
data
```

	Reviews	ratings
0	Extraordinary hospitality at the hotel. We wen...	5
1	Good place with great scenery. Stayed only for...	5
2	Stayed here for two days. Beautiful property w...	5
3	Very pleasant stay. Friendly staff and excelle...	5
4	We stayed here for 4 nights for business meeti...	5
...	...	...
44358	Excellent hospitality and dining. Supper...	5
44359	The ambience, hospitality and staff dealings a...	5
44360	The wonderful memories in chandys windys woods...	5
44361	It awesome atmosphere and excellent service pr...	5

PRE-PROCESSING

1. Sentence Tokenization
2. Removing Punctuations
3. Lowering The Texts
4. Removing Stopwords
5. Lemmatizing

TOPIC MODELLING

- We have used LDA (Latent Dirichlet Allocation) for our purpose
- Unsupervised Machine Learning Technique for text analysis
- Takes input on the number of topics and helps to cluster similar words under one topic
- Assumes Each Document is a Mixture of Various Topics
- Each Topic is a Mixture of Various Words

IMPLEMENTATION OF TOPIC MODEL

- We used Bayesian Optimisation and Coherence Score to fine-tune our topic model
- We got our best Coherence Score of 0.5 and Optimal number of topics as 6
- Then each of our sentence tokens was assigned the topic with the highest probability
- At the end we went through the words under each topic to map the topics to relevant service parameters of the hotel industry
- We identified the following service parameters from the topics identified by our LDA model: **Hotel Amenities, Room Experience, Staff Responsiveness, Location, Dining Experience, Transportation Service**

SENTIMENT ANALYSIS USING BERT MODEL

- BERT – Bidirectional Encoder Representations from Transformers
- Jacob Devlin and his colleagues developed BERT at Google in 2018
- BERT allows the architecture or model to learn the heterogeneity in data patterns and perform effectively across a variety of NLP applications which gathers knowledge from both the left and the right sides

Implementation Of BERT Model

- BERT model gives the output as POSITIVE or NEGATIVE sentiments along with their respective probability scores
- Using the BERT model, we got the sentiment labels of each of our sentence tokens
- Finally we created a dataframe to store the sentence tokens, sentiment labels and probability scores

POST SENTIMENT ANALYSIS STEPS

- We converted the probability scores of each of our sentence tokens to regression scores
- Regression scores with negative labels were multiplied with -1
- If a particular review has multiple sentence tokens with the same topics then the respective regression scores were averaged
- Then we created a pivot table containing the review indexes in the rows and the topics in the columns

NEED OF IMPUTATION

- After creating the pivot table we saw that for some of the reviews, some service parameters are coming as zero
- We assumed that a customer who wrote a review must have experienced all the identified service parameters of the hotel
- So we imputed the zero values with the help of an iterative imputer

	Hotel_amenoties	Room_experience	Staff_responsiveness	Location	Dining experience	Transportation service
44343	4.929043	6.082725	8.936465	5.413013	7.046925	3.813577
44344	6.949794	8.820985	9.345813	7.104495	8.973406	5.420721
44345	5.461898	6.864208	8.513015	7.449536	7.196591	4.424485
44346	6.789341	9.003978	7.807496	8.798617	8.152650	5.446214
44347	7.937997	7.750592	8.983475	6.604187	8.247464	5.287475
44348	1.474260	4.200196	7.273865	5.987189	7.636171	2.533178
44349	-8.221877	2.273540	3.821853	8.502917	1.143449	-5.034103
44350	4.999220	6.173070	8.955460	5.576564	7.113666	3.086618
44351	3.893825	6.687356	5.953942	4.613787	0.690907	6.060969
44352	5.377567	6.662818	8.089914	7.437486	7.306660	4.334485

Final Score Matrix

- Snap of Score Matrix

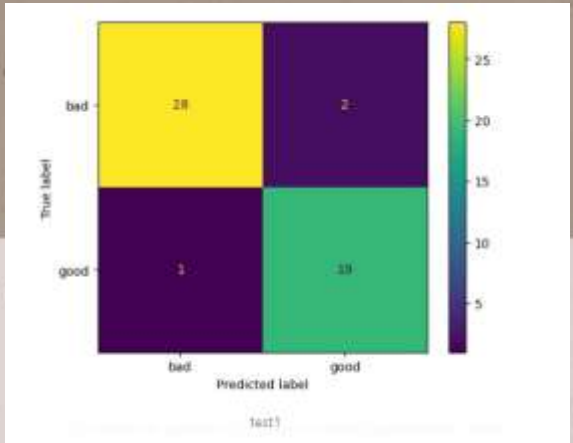
	Hotel_ameneties	Room_experience	Staff_responsiveness	Location	Dining experience	Transportation service	Rating	Review_label
44353	5.727069	6.334018	8.231076	6.126762	9.035511	4.367103	5	good
44354	8.935558	8.283552	9.028354	6.866587	8.278125	5.708846	5	good
44355	4.927237	6.090047	8.857212	5.520796	7.038473	3.623786	5	good
44356	6.748763	8.501155	9.119997	8.854669	8.458502	5.518166	5	good
44357	5.649501	6.956865	9.079513	6.919033	7.675368	4.512644	5	good
44358	5.720938	6.605692	8.607698	6.127910	8.596513	4.399339	5	good
44359	4.806338	5.950607	8.692202	5.427132	6.912186	3.718257	5	good
44360	4.674302	5.806882	8.263286	5.587883	6.733817	3.619621	5	good
44361	-5.285596	9.052677	9.072217	5.063494	3.643362	1.117887	5	good
44362	4.760017	6.465596	8.960575	6.674676	7.302448	4.082948	5	good

- Reviews with a rating >= 4 are labelled as Good
- Ratings below 4 are labelled as Bad

Train For Classification

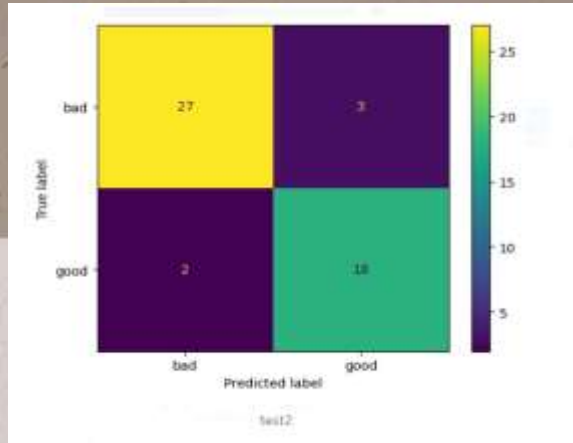
- We used a random forest classifier to train a model to predict good reviews and bad reviews
- Here are the snapshots of our model performance on unseen data

TEST 1



- Model predicted 95% of all GOOD reviews correctly
- Model predicted 93.33% of all BAD reviews correctly
- Overall Accuracy 94%

TEST 2



- Model predicted 90% of all GOOD reviews correctly
- Model predicted 90% of all BAD reviews correctly
- Overall Accuracy 90%

Praxis Business School
CELEBRATE YOUR WORTH

WHAT IS **SHAP**

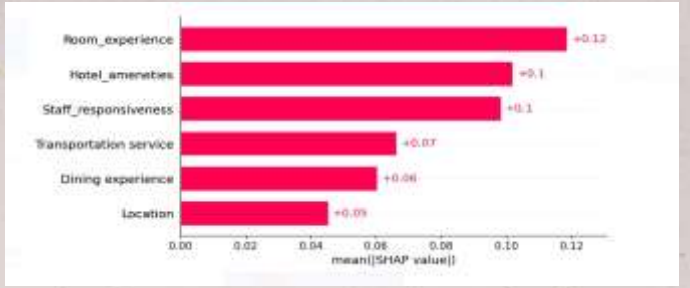
- **SHAP** stands for **SH**apley **A**dditive ex**P**lanations
- It is a framework for explaining the output of machine learning models.
- It is based on the concept of Shapley values from
- **SHAP** assigns an importance value to each feature in a prediction by calculating the difference between the prediction with and without that feature cooperative game theory

HOW **SHAP** CAN BE USED FOR OUR PURPOSE

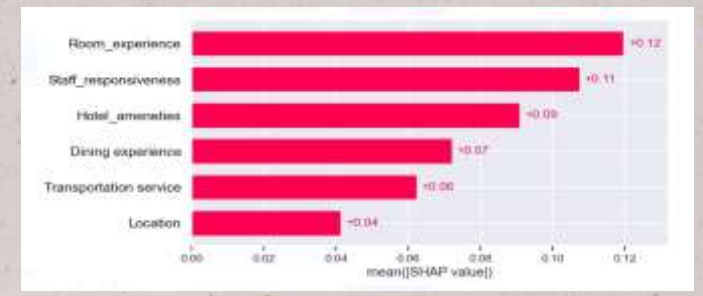
- Since SHAP is using shapely values to calculate feature importance
- We use the same to quantify the importance of each of our service parameters to understand why the customers of a hotel are giving bad reviews

OBSERVATION

BAR PLOTS USING SHAP



West Bengal

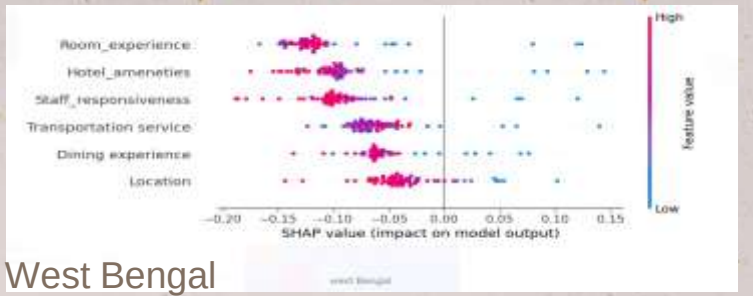


Rajasthan

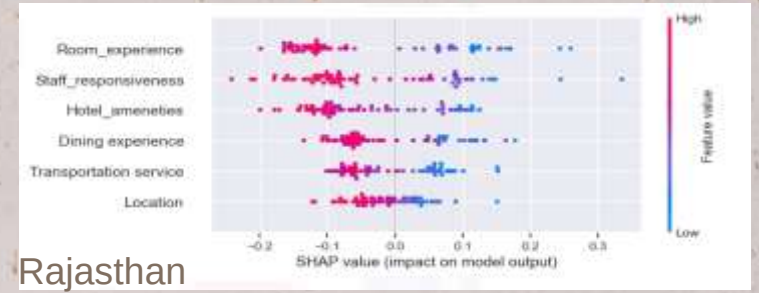
- As we can see in the above graphs for both the above states Room experience is coming out as the most important feature
- The second most important feature for West Bengal is Hotel Amenities but for Rajasthan is Staff Responsiveness
- So based on our analysis we can say that the hotels in Rajasthan must give more importance to staff responsiveness and dining experience to improve their customer's level of satisfaction

Praxis Business School
CELEBRATE YOUR WORTH

BEESWARM PLOT USING SHAP



West Bengal



Rajasthan

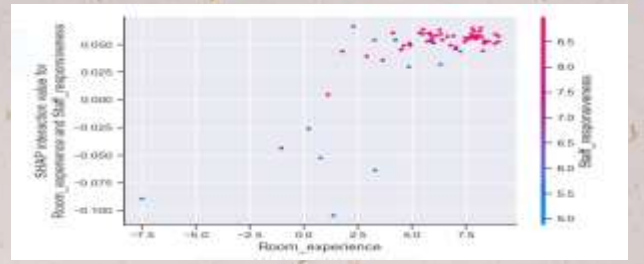
In the X-axis, we are plotting the model output(log odds) w.r.t. the reviews being bad.
Positive values on the X-axis mean a higher probability of the reviews being bad
Negative values on the X-axis mean a lower probability of the reviews being bad

From the above plots we can conclude the following:

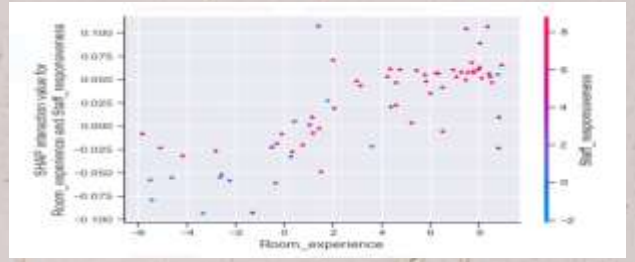
- In the case of both West Bengal & Rajasthan, we can see, in most of the reviews, the scores of the service parameters are on the higher side and higher scores are decreasing the probability of the review being bad
- Additionally in the case of West Bengal, we observe that the scores of the “Transportation Service” is on the lower side and should be worked upon to increase the same

Praxis Business School
CELEBRATE YOUR WORTH

DEPENDENCE PLOT FROM SHAP



West Bengal

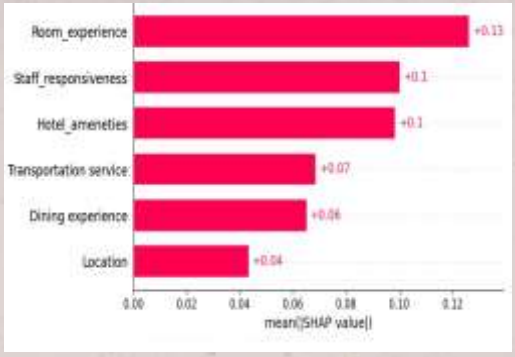


Rajasthan

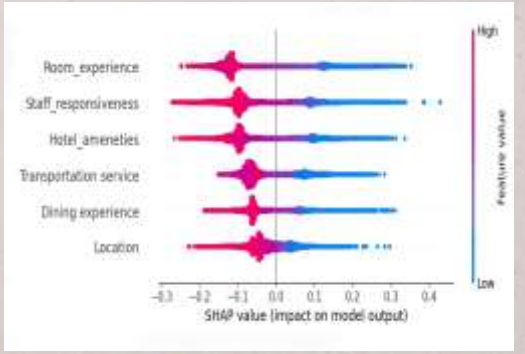
In the X-axis we have the actual feature scores of one of the interacting variables, while, the colors are representing the feature scores of the other interacting variable.
The Y-axis represents the SHAP interaction values(log-odds), positive values of the log-odds mean a higher probability of the review being bad and vice-versa

We conclude the following from the above plots:
In the case of the above-mentioned states, we can see that although the scores of Staff Responsiveness and Room Experience are on the higher side still their combined effect is slightly increasing the probability of the reviews being bad. Although this is counter-intuitive, this is happening may be due to the higher prices of the rooms. Since higher prices come with higher expectations from the customers

ANALYSIS ON ALL THE REVIEWS

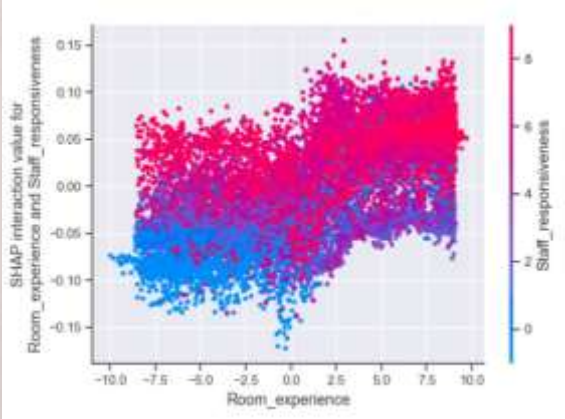


In the bar plot, we observe that “Room Experience” proves to be the most important service parameter for all the reviews which is in the same line as that of our state-wise analysis



In the beeswarm plot, we observe that in some cases lower scores of “Staff Responsiveness” is pushing up the value of log odds beyond 0.3 and even crossing 0.4. So, we can say that for some customers Staff Responsiveness is playing a more important role in determining the quality of their experience

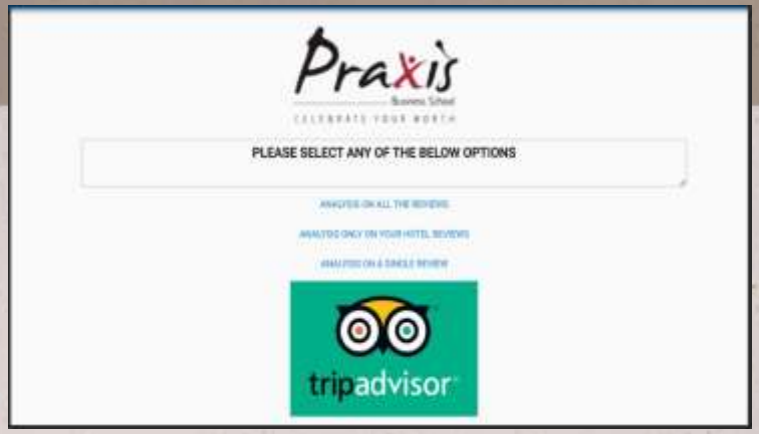
Analysis On All The Reviews



- The dependence plot reveals some data points with high scores for both interacting features, yet they marginally push up the log odds/probability of bad reviews. This could be due to the high price of rooms, indicating unmet expectations for the level of service. Higher prices often come with higher expectations
- Some data points have lower room experience scores but higher staff responsiveness scores. The SHAP interaction value for these data points varies mainly between 0 - 0.05. This suggests that good staff experience has overridden the effect of lower room experience scores
- Some data points have low scores for both interacting features but don't significantly increase the probability of bad reviews. These may be the value-for-money hotels where customer expectations are lower

Praxis Business School
CELEBRATE YOUR WORTH

Web Interface



Video Of The Working Model

LIMITATIONS

- Users should know to analyze the graphs
- The app needs heavy calculations, so proper GPU support is needed for better performance
- The price parameter wasn't included initially, which would have helped to explain the sentiments of the customers connecting with the hotel industry service parameters
- Only TripAdvisor was scrapped. Reviews from other service providers were not included.
- Around 10 hotels on average from 20 states in India, we can include more hotels for better analysis

FUTURE SCOPE

- We can include more reviews so that our outputs can generate better state-wise representations of hotels
- We are planning to give our users an option using which they can get the outputs w.r.t the reviews are bad or good
- We are also planning to add a feature that can analyze our output graphs automatically and present the same in natural language

Thanks!