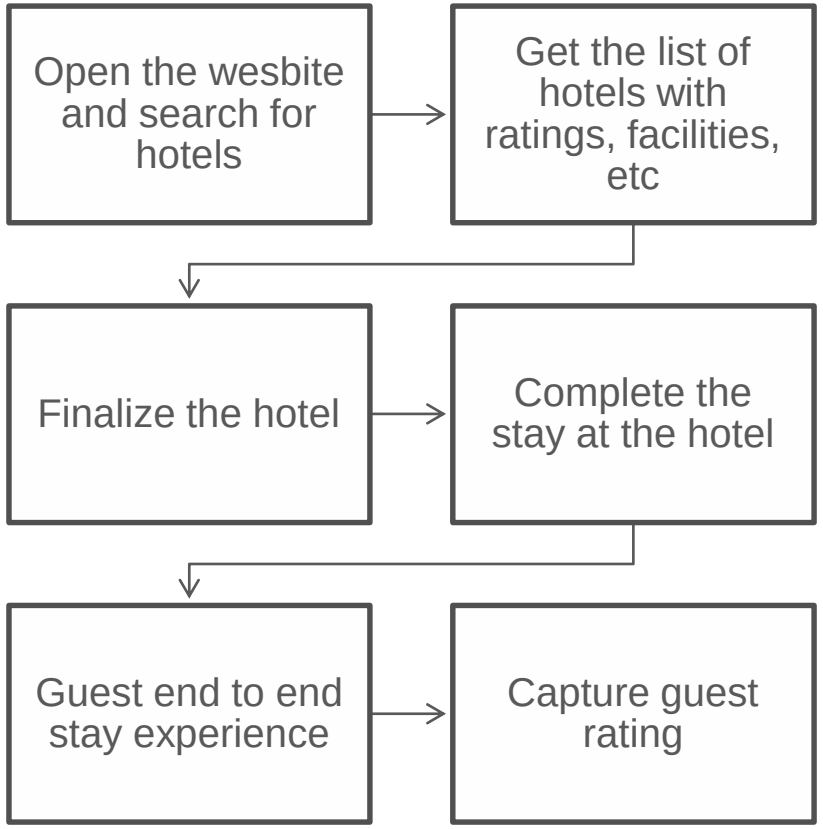
Enhance the guest experience for Indian Hotel Industry using a predictive guest rating model

For Cousera Capstone Project

Anurag Nirmal
March 12, 2020

Predicting the guest experience at Indian hotels is important for hotel owners and travel booking platforms

Guest lifecycle from booking to stay



Key objectives and outcome of the study

- The project aims to understand the factors responsible to determine the user rating after staying at the property. A predictive machine learning model will be suited for this purpose. Also, comparison will be drawn against the rating provided by the website (Goibibo.com)
- The study is aimed at hotel owners and travel booking platforms who will have maximum interest
- 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

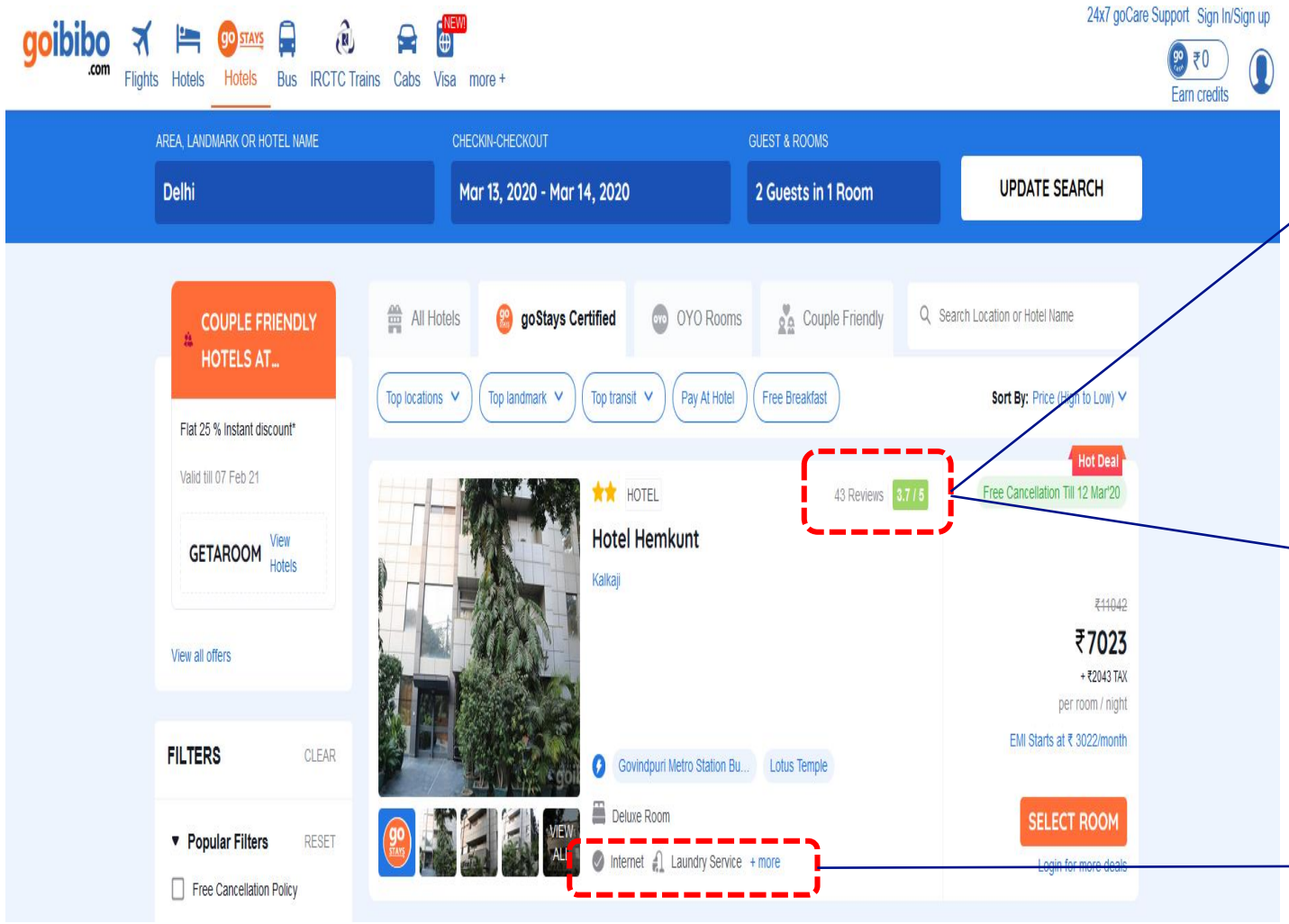
A word about the data

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

This is how the dataset looks like ->

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

Data preparation started with visiting goibibo.com to understand the booking process. Next, standard data cleaning and preparation tasks were carried out



Site 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 and Cleanliness
- *Action* - The column data was split and new columns added back to master dataframe

Count of Reviews

- The column has count of reviews categorized as positive, critical and image reviews
- *Action* - The column was split and dataframe added back to master dataframe. New columns are – ct_pos_review, ct_crit_review and ct_img_review

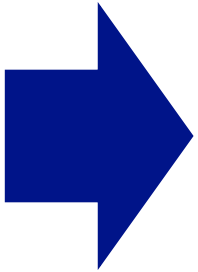
Room Facilities

- The column has the highest unstructured data where room facilities like room service, telephone, hot water, kitchen, etc are clubbed together in one cell
- *Action* - for loop to search for keywords and then create dummy variables to define the keywords

After data understanding, next step is to finalize the feature set and target set

Feature Set

hotel_brand_label
hotel_category_label
hotel_star_rating
image_count
property_type_label
ct_pos_review
ct_crit_review
ct_img_review
room_count
room_type_label
site_review_count
room_service
restaurant
cable
internet
heating
hot_water
shower
air_cond
bathroom
telephone
newspaper
refrigerator
housekeeping
kitchen
extra_bed
poi_length



Target Set

rate_guest =

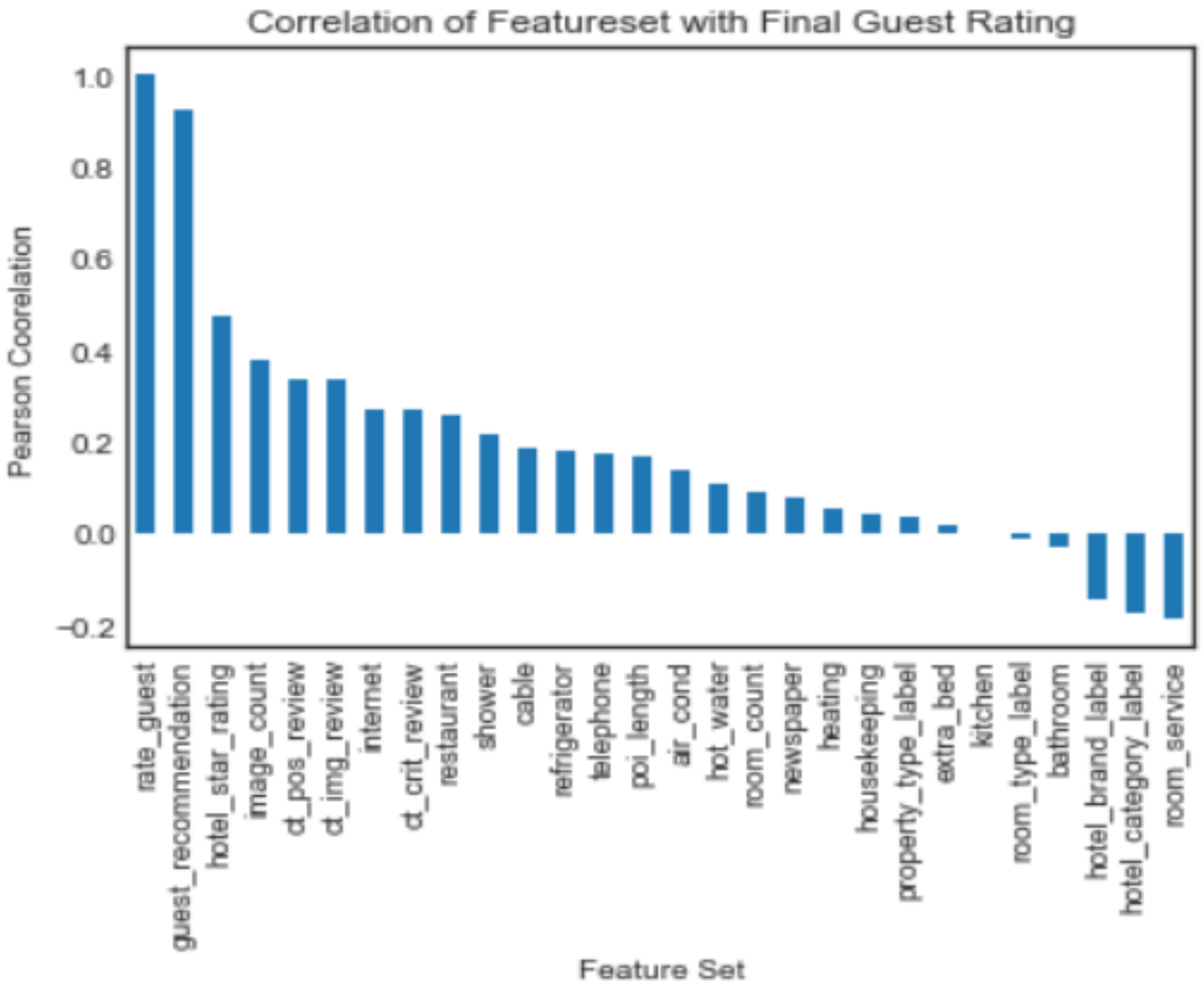
The target set is the final guest rating which is computed as average of 6 rating scale

- Service Quality
- Amenities
- Food and Drinks
- Value for Money
- Location
- Cleanliness

Exploratory Data Analysis to draw inferences from the dataset …(1/2)

Pearson Correlation – We calculated the Pearson correlation coefficient for the feature set and target variable (rate_guest)

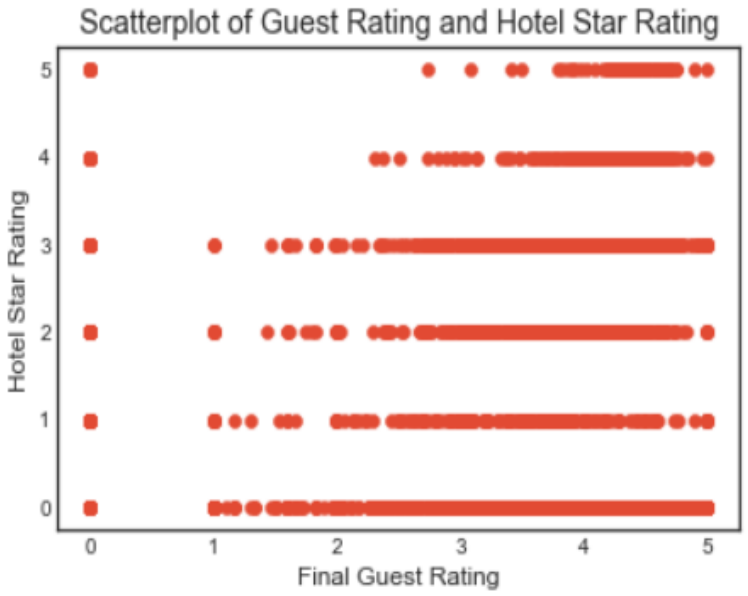
The highest correlated features are Guest recommendation, Hotel star rating, Image count, Count of positive review, Site review count and Count of image reviews



Exploratory Data Analysis to draw inferences from the dataset …(2/2)

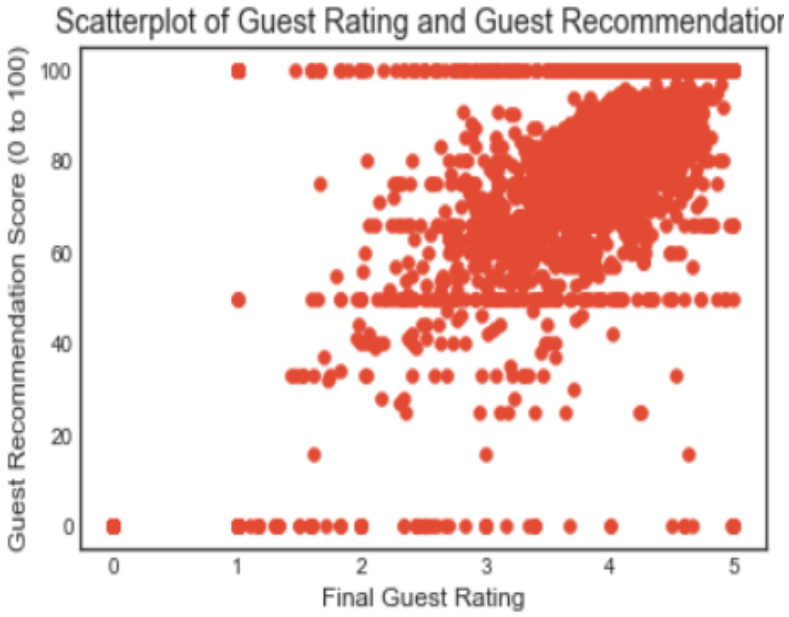
Hotel Star Rating – this is the industry established rating scale for hotels

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

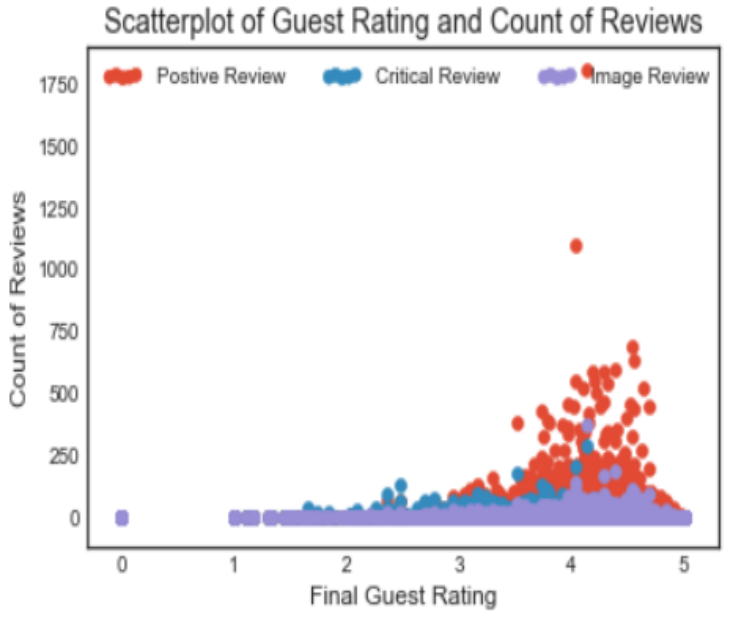


Guest Recommendation – this is the recommendation score given by guest to future travellers

There is clear link between the two which means that a guest will recommend the hotel when he rates it higher also



Count of Reviews – the count is divided into positive, critical and image reviews. We see no clear correlation as suggested by Pearson correlation.



Four classification machine learning models helped us classify the target variable correctly

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 models

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

Following classification model are used

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

The **Decision Tree classification model** has shown the best results for our data with 66% 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	0.36	**0.66**	0.42	0.46
F1 Score	0.51	**0.75**	0.56	0.61
Log Loss	-	-	1.18	-

The decision tree model tells us about the **five** variables which drive the guest stay experience at the property

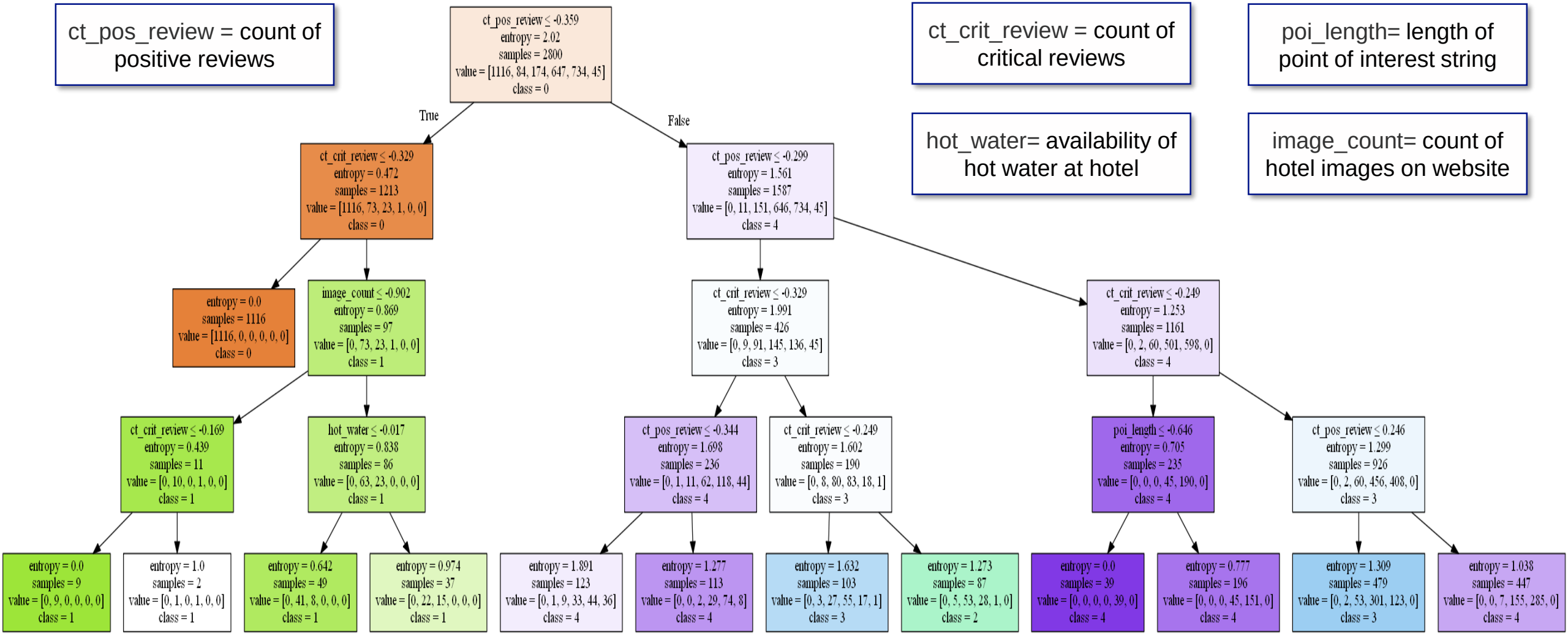
ct_pos_review = count of positive reviews

ct_crit_review = count of critical reviews

poi_length= length of point of interest string

hot_water= availability of hot water at hotel

image_count= count of hotel images on website



Recommendations for hotel owners and travel booking platforms

Count of reviews

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

Point of interest

The **point of interest** which are tourist places close to the property form a significant part of the guest rating. Properties which are closer will attract more guests and hence more business

Count of images

The **count of images** at the hotel drives the hotel selection decision and finally the guest stay experience

Room facilities

From **room facilities** perspective, basic facilities are important to the guest like hot water and internet. Since the model accuracy was not higher a further study on large dataset is recommended

End of Presentation