# Unleashing the Power of Machine Learning on Booking.com Hotel Data

452 Final Project

Master of Science in Business Analytics

University of California, Davis

Priya Iddalgi

Yifan Zhu

Kangjian(James) Gao

# Table of Contents

# Executive Summary

In the competitive hospitality industry, data-driven insights are vital for business growth and informed decision-making. In this project, we utilized data scraped from Booking.com, one of the largest online travel agencies, to analyze hotels in New York, Chicago, and Los Angeles. The extracted data was used to address three business problems using machine learning techniques: 1) Determining the factors that contribute to a hotel's overall rating, 2) Identifying potential competitors for hotels in the three cities, and 3) Predicting the appropriate price for a hotel based on its features.

Our analyses revealed the following key findings:

1. Facilities rating, staff rating, location rating, and the number of reviews significantly impact a hotel's overall rating.
2. The k-Means clustering algorithm was effective in grouping hotels with similar characteristics, allowing businesses to identify potential competitors and benchmark their performance.
3. Random Forest outperformed other machine learning models in predicting hotel prices, with overall rating and the number of reviews as the most critical features in determining price.

# Background：

In today's competitive industry, data is playing a crucial role in driving business growth and making informed decisions. Booking.com is one of the world's largest online travel agencies that offer a wide range of accommodation options for travelers, including hotels, apartments, villas, and hostels, in over 220 countries worldwide. The platform has more than 28 million listings and attracts millions of visitors daily. This makes it a valuable source of data for businesses in the hospitality industry.

We specifically scraped hotels in three main big cities in the US: New York, Chicago, and Los Angeles. The date time range is between May-30-2023 to May-31-2023, and we scraped the lowest price available in each hotel and selected the number of sleepers to 1. By web scraping booking.com and extracting information such as hotel name, hotel price, room type, miles to center, address, facilities rating, free wifi rating, number of reviews, overall rating, and staff rating, we can gain valuable insights into the industry trends, customer preferences, and competition.

The data extracted from booking.com can be used for various purposes to impact the hotel services in the hospitality industry or traveling industry. Here are three of the business problems in which the data can be answered by using machine learning 1. How do different features/ aspects contribute to the overall rating of that hotel 2. How do identify potential competitors of the hotel and make adjustments based on competition? 3. How to set the appropriate price for that hotel based on certain features?

# Business Context :

1. How do different features/ aspects contribute to the overall rating of that hotel?

Traditionally, hotels attempt to solve this problem by relying on customer feedback and internal evaluations. The hotel staff may ask customers for feedback directly, or provide a comment card in the room for customers to fill out. The hotel management team may also evaluate the hotel's facilities and services on a regular basis to identify areas for improvement. It relies on the hotel staff's expertise and knowledge of the hotel's facilities and services. By focusing on improving the areas that customers consistently rate as being important, the hotel can improve its overall rating and attract more customers.

2. How do identify potential competitors of the hotel and make adjustments based on competition?

Traditionally, the hotel management team visits competitors' hotels to evaluate their facilities and services. Based on this research, the hotel can make adjustments to its own pricing, features, and services to remain competitive in the market. By keeping a close eye on the competition and making adjustments as needed, the hotel can remain competitive and attract more customers.

3. How to set the appropriate price for that hotel based on certain features?

Traditionally, hotels attempt to solve this problem by conducting market research. The hotel management team may look at competitors' pricing and compare it to their own, or analyze the local market to identify pricing trends. They may also consider factors such as the hotel's location, amenities, and services when setting the price. It relies on the hotel management team's knowledge and experience of the local market and the hotel's facilities and services. By setting the appropriate price based on these factors, the hotel can attract more customers and increase its revenue.

# EDA:

For exploring the dataset, we tried to see the statistical measures, such as mean, median, mode, standard deviation, min and max for each of the variables. Further, we also tried to visualize the distributions of all the variables along with boxplots to identify any outliers, in order to understand the nature of the features in our dataset. (entire results are in the appendix).

None of the numerical variables seem to be normally distributed. The price and number of reviews seem skewed towards the left, while all the ratings seem skewed towards the right. Some features are highly correlated, such as overall ratings with facilities ratings and staff rating, which is expected from this kind of data.Additionally, most room types are basic, and data is for three different cities of NY,LA and Chicago.

# Analyses:

Analysis 1:

The first business problem that we tried to address in this project is to run regressions, and set overall ratings as dependent variables. For regression models, we checked four assumptions for linear/multiple models. The interpretation for the final model Figure J is the following: the Multiple R-squared values is 0.9687, which indicates that approximately 96.87% of the variation in the overall rating can be explained by the included variables. This means that the model has a strong fit for the data. The F-statistic is 2800, with a p-value less than 2.2e-16, which suggests that the overall model is statistically significant. The individual variables' significance can be interpreted based on their p-values: Price: p-value = 0.619834, not significant. Miles to center: p-value = 0.599024, not significant. Facilities rating: p-value < 2e-16, highly significant. Free Wi-Fi rating: p-value = 0.979646, not significant. Location rating: p-value < 2e-16, highly significant. The number of reviews: p-value = 0.000402, significant. Staff rating: p-value < 2e-16, highly significant.

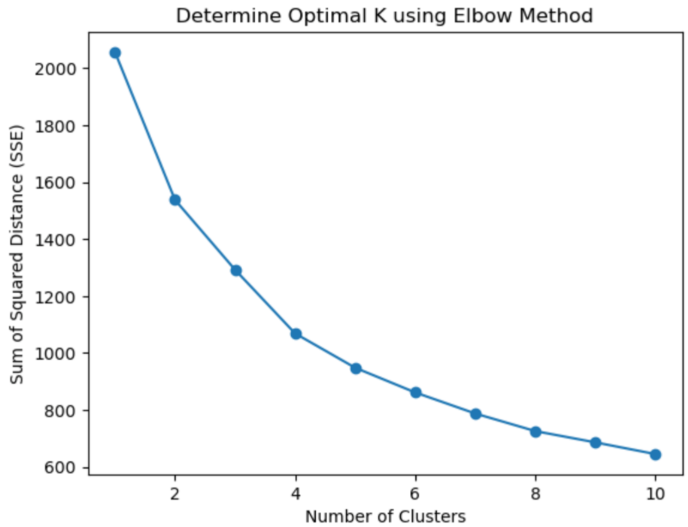
Analysis 2:

The second business problem that we tried to address in this project is to identify potential competitors for each hotel in the city of New York, Chicago, and Los Angeles. Our approach is to group hotels that are most similar to each other in terms of relative distance using the K-Means algorithm. The first step of our analysis is to do some data preprocessing. If the input variables are categorical, we know that the algorithm would consider the relative distance between two data points to be equal to 1. Thus, we decide to remove all categorical variables in the dataset. Since the algorithm is distance based, we also applied robust scaling to make sure all the data are measured in the same level while less sensitive to outliers. After having the standardized data, we then shuffled the data to make the first 90% of rows be the training set and the remaining 10% of rows be the testing set.

After done with data preprocessing, we then moved to the next step of deciding the optimal K for the K-Means algorithm. In this step, we applied the elbow method to find the biggest drop in sum of squared distance (SSD) between two consecutive points. In the case of New York, we found that the optimal K is equal to 2 (Figure 1). We then took K = 2 as the input variable and fit the model on the training set, which gives us the initial sum of squared distance equal to 1537.842 and the initial sample size equal to 261. Before fitting the model using all the data, we also did some validation to prove the model is effective. To achieve this, we compare the percentage increase in SSD with the percentage increase in sample size and found the percentage increase in SSD is consistently lower than the percentage increase in sample size in the case of New York. Therefore, we are confident to conclude that the model is effective, and we then run the model using the entire data set to get the detailed membership information for each cluster. We did the same analysis for all three cities, and the membership list can be found in Figure 2, Figure 3, and Figure 4 in the appendix.

The purpose of this analysis is to help hotels in all three cities to find their potential competitors. For those hotels in the same cluster, they are most similar to each other in terms of price, miles to center, address, facilities rating, free WIFI rating, location rating, number of reviews, overall rating, and staff rating. All this information can be used for Competitive Analysis, so that the business can have a better understanding of where to improve by comparing to potential competitors' performance.

Figure 1




Analysis 3 :

Building models for hotel room price prediction:

In this section, we explored and compared various machine learning models such as CART, Boosted Trees, Random Forests and Neural networks to predict the prices of the various hotels we scraped. We removed the extra columns not required for prediction, split the data into train and test sets, and evaluated the model performance on the test set with two evaluation metrics : Mean Square Error and Rsq.

We first fit a CART Tree to our regression model. Since the categorical variables of Location and Room Type were important for our prediction, we used One-Hot encoding to convert these variables into numbers and feed it as input to grow our CART Tree. The plotted tree can be seen in Fig. 1.

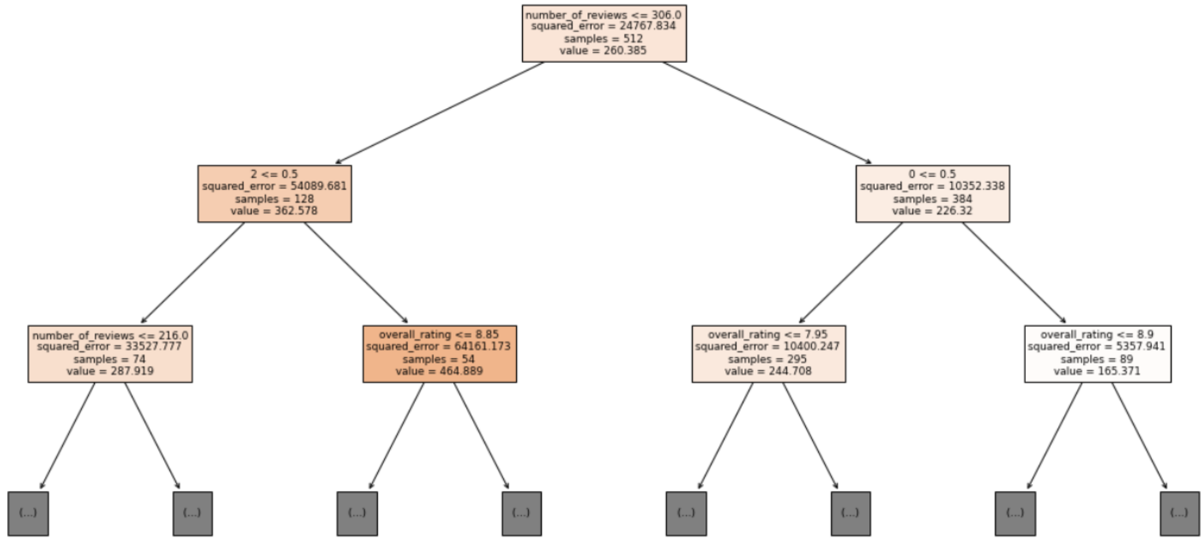



Figure 5

We proceeded to use XGBoost library to make a boosted model, and then used an ensemble method - Random Forest. We then fit a simple neural network with 64 layers and a ReLU function as activation, and ran it for 100 epochs. To finetune it, we increased the number of layers to 128 and 256 and ran it for 300 cycles. Here is a summary of all the results for comparison.

| Evaluation Metric | CART | XGBoost | Random Forest | NN (64 layers, 100 epochs) | NN (256 layers, 300 epochs) |
|---|---|---|---|---|---|
| Mean Sq Error | 10487.18 | 12458.02 | 6341.75 | 10013.68 | 8047.09 |
| Rsq | 0.179 | 0.025 | 0.504 | 0.216 | 0.37 |

Figure 6

Random forest outperforms all these other methods. Many features in our dataset are correlated (overall rating is correlated with wifi,location rating etc). Random Forest’s ability to reduce overfitting, handle non-linear relationships, handle correlations  and simple tuning parameters make it a great choice for our dataset and business needs. We can see, through the CART, tree that the most important features in determining the Price are ‘overall rating’ and ‘number of reviews’ (because they have splits early on in the tree growing process) while ‘miles to center’ and ‘wifi rating’ are relatively less important.

# Recommendation and Business Value:

For analysis 1, focus on improving the facilities rating, as it has a high positive coefficient (0.6676) and is highly significant. Better facilities will likely lead to higher overall ratings. Pay attention to the staff rating, which also has a high positive coefficient (0.3460) and is highly significant. Providing excellent customer service and well-trained staff can contribute to higher overall ratings. The location rating has a positive coefficient (0.1423) and is highly significant. While it may be challenging to change the hotel's location, businesses can improve the perceived location by offering shuttle services, partnering with local attractions or businesses, or highlighting nearby points of interest to guests. Although the number of reviews has a smaller positive coefficient (0.00001416), it is still significant. Encourage guests to leave reviews, as more reviews can contribute to a better overall rating. Price and miles to center have insignificant p-values, indicating that these variables may not significantly impact the overall rating in the current model. However, it's essential to keep these factors competitive in the market. The free Wi-Fi rating is not significant in the current model, but it is still important to provide reliable and fast Wi-Fi services to meet guests' expectations.

For analysis 2, since we applied the K-Means model to the entire dataset and obtained cluster members' info for each hotel in the cities, hotels can analyze the membership lists (Figure 2 for New York, Figure 3 for Chicago, and Figure 4 for Los Angeles) to do more search on hotels within the same cluster. These hotels are potential competitors, as they are most similar in terms of price, miles to center, address, facilities rating, free WIFI rating, location rating, number of reviews, overall rating, and staff rating. Within each competitor pool, the marketing competitive analysis model and SWOT model can be well applied, and by evaluating the performance of possible "rivals", the hotel may have a better idea of where to improve.

For analysis 3, since this is a smaller dataset, it is difficult to comment on which model will overall perform the best for the entire industry, but it can still give us some insights about the important features that determine the fluctuations in price and predict it. Our recommendation would be for the businesses to make use of the various features (present in our data), along with other data sources (such as seasonality, brand, location, amenities, and competition), and use ensemble methods like Random Forest to build a price predictor model. They should focus mostly on getting the highest overall ratings and maximizing the number of reviews in order to quote a price as per their needs,by incentivizing more customers to vote. The price prediction model and identification of such important features that predict the price can provide valuable insights to optimize their pricing strategy, improve operations, and stay competitive in the market.

# Conclusions:

The insights gained from this project can provide valuable guidance for hotels to optimize their operations and pricing strategy. By focusing on improving facilities, staff, and location ratings, and encouraging guests to leave more reviews, hotels can enhance their overall ratings and subsequently charge higher prices.

Utilizing clustering algorithms to identify potential competitors enables hotels to better understand their market position and make strategic adjustments. Lastly, implementing ensemble methods like Random Forest for price prediction models can help hotels stay competitive and maximize their revenue.

**Appendix**

Fig A.1 Histograms of all Numerical Variables:

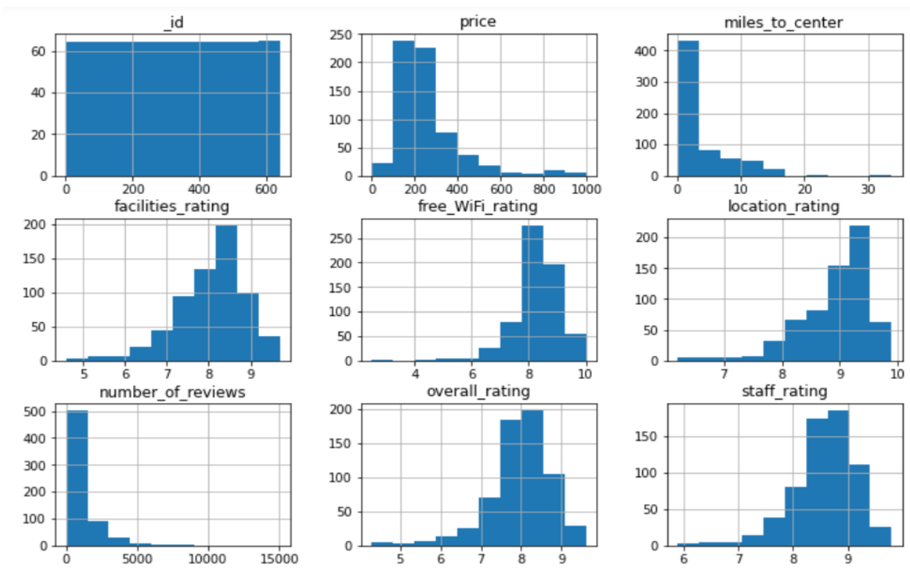



Fig A.2 : Correlation Plots between different numerical variables

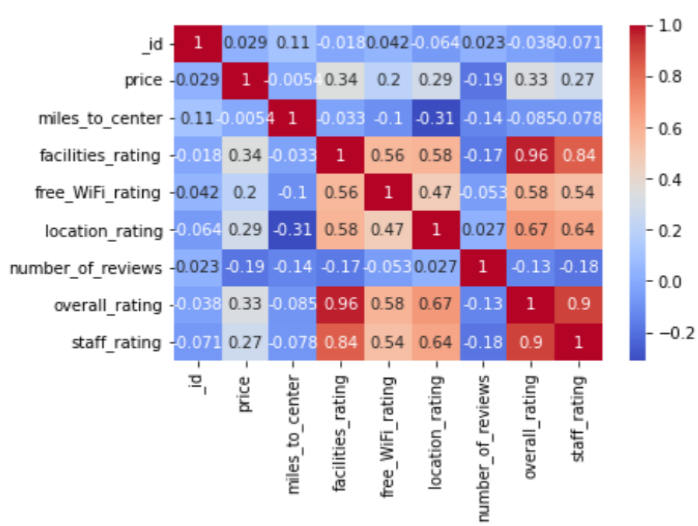


Fig A.3: Regression model on how do different features contribute to the overall rating of that hotel

```
lm(formula = y ~ ., data = x1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98073 -0.08344 -0.00155  0.07401  0.45489 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.653e+00  9.898e-02 -16.695  < 2e-16 ***
price             -1.990e-05  4.009e-05  -0.496 0.619834    
miles_to_center   -7.512e-04  1.428e-03  -0.526 0.599024    
facilities_rating  6.676e-01  1.344e-02  49.665  < 2e-16 ***
free_WiFi_rating   2.101e-04  8.233e-03   0.026 0.979646    
location_rating    1.423e-01  1.314e-02  10.833  < 2e-16 ***
number_of_reviews  1.416e-05  3.981e-06   3.558 0.000402 ***
staff_rating       3.460e-01  1.836e-02  18.850  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1392 on 633 degrees of freedom
Multiple R-squared:  0.9687,    Adjusted R-squared:  0.9684 
F-statistic:  2800 on 7 and 633 DF,  p-value: < 2.2e-16
```

Figure 2 (Partial)

| Competitors Pool 1 | Competitors Pool 2 |
|---|---|
| Kixby | Courtyard New York Downtown Manhattan/Financial District |
| Hilton Club New York | Midtown West Hotel |
| Sanctuary Hotel New York | Four Points by Sheraton Manhattan Chelsea |
| Gild Hall - A Thompson Hotel | TownePlace Suites by Marriott New York Manhattan/Chelsea |
| Cambria Hotel New York - Chelsea | W New York - Union Square |
| Hilton Club The Quin New York | Hampton Inn Manhattan Grand Central |
| San Carlos Hotel New York | Fairfield Inn & Suites by Marriott New York Manhattan/Chelsea |
| Conrad New York Midtown | YOTEL New York Times Square |
| Courtyard New York Manhattan/Midtown West | Casamia 36 Hotel |
| 1 Hotel Central Park | Redford Hotel |
| Pestana CR7 Times Square | Hyatt Place NYC Chelsea |
| Arlo Midtown | Courtyard by Marriott Times Square West |
| New York Marriott Downtown | Hyatt Grand Central New York |
| Courtyard by Marriott New York Manhattan/Central Park | Holiday Inn Express - Times Square South, an IHG Hotel |
| Courtyard New York Manhattan/Midtown East | The Herald 8 by LuxUrban |
| The Historic Blue Moon Hotel - NYC | Bentley Hotel |
| New York Marriott Marquis | Americana Inn |
| Best Western Premier Empire State Hotel | Hilton Garden Inn New York Central Park South-Midtown West |
| Iberostar 70 Park Avenue | The Benjamin Royal Sonesta New York |
| Artezen Hotel | Hilton Garden Inn Times Square |
| ... | ... |

Figure 3 (Partial)

| Competitors Pool 1 | Competitors Pool 2 |
|---|---|
| Hampton Inn Chicago-Midway Airport | Homewood Suites by Hilton Chicago Downtown |
| EDGEBROOK MOTEL | The Emily Hotel |
| Best Western Plus Hyde Park Chicago Hotel | Hyatt Regency McCormick Place |
| Holiday Inn Chicago Midway Airport S, an IHG hotel | Hilton Garden Inn Chicago Downtown Riverwalk |
| Sleep Inn Midway Airport Bedford Park | Fairmont Chicago Millennium Park |
| Hilton Garden Inn Chicago/Midway Airport | Majestic Hotel |
| Hampton Inn Chicago North-Loyola Station, Il | Hotel Chicago West Loop |
| Skylark Motel | SpringHill Suites Chicago Downtown/River North |
| Red Roof Inn Chicago-Alsip | Best Western Chicago Downtown-River North |
| | Freehand Chicago |
| | Best Western Grant Park Hotel |
| | Hilton Garden Inn Chicago Downtown/Magnificent Mile |
| | Fairfield Inn and Suites Chicago Downtown-River North |
| | Moxy Chicago Downtown |
| | La Quinta by Wyndham Chicago Downtown |
| | The Langham Chicago |
| | Crowne Plaza - Chicago West Loop, an IHG Hotel |
| | Godfrey Hotel Chicago |
| | Waldorf Astoria Chicago |
| | The Westin Chicago River North |
| | ... |

Figure 4 (Partial)

| Competitors Pool 1 | Competitors Pool 2 |
|---|---|
| Hotel 850 SVB West Hollywood at Beverly Hills | H by H Hospitality |
| Palihouse West Hollywood | Travelodge by Wyndham Culver City |
| The Hoxton, Downtown LA | Courtyard by Marriott Los Angeles LAX / Century Boulevard |
| The London West Hollywood at Beverly Hills | Mr C Beverly Hills |
| STAY OPEN Venice Beach | Lexen Hotel - Hollywood |
| The Garland | Hotel Angeleno |
| The Prospect Hollywood | Hampton Inn & Suites Santa Monica |
| Best Western Plus LA Mid-Town Hotel | Tuscan Garden Inn |
| El Royale Hotel - Near Universal Studios Hollywood | Cal Mar Hotel Suites |
| La Mirage Inn - Hollywood | Ramada by Wyndham Los Angeles/Wilshire Center |
| Hollywood Celebrity Hotel | W Hollywood |
| The Godfrey Hotel Hollywood | Hometel Suites Hotel |
| The Charlie West Hollywood | Best Western Plus Commerce Hotel |
| Hyatt House LAX Century Blvd | Best Western Royal Palace Inn & Suites |
| Ocean View Hotel | Little Tokyo Hotel |
| Dream Hollywood | Hollywood Le Bon Hotel |
| Petit Ermitage | Sheraton Gateway Los Angeles Hotel |
| Comfort Inn Near Old Town Pasadena in Eagle Rock CA | GOLDSTAR INN MOTEL |
| Palihotel Westwood Village | Hilton Garden Inn Los Angeles Marina Del Rey |
| Knights Inn Los Angeles Central / Convention Center Area | Carlyle Inn |
| ... | ... |