أكاديمية سدايا
SDAIA Academy

Data Science Bootcamp

Project Proposal 2

AI Solution for Booking.com

Presented by

Samah Alrefaei

Ashjan Basri

16 Sep. 21

Introduction

Recently online hotels reservations become an important platform in all the hotels industry. customers can view and make their choice before arrival, check the location and read the reviews and hotels ratings before doing the reservation. Booking.com is the world's leading digital travel companies and their website offers more than 28 million reported accommodation listings. In line with the company's mission Save time, save money!, we proposing a solution using Machine Learning (ML) model, precisely linear regression to predict hotels ratings to improve the company's revenue during Haj and Omrah seasons. Since Saudi Arabia is the Middle East's second most popular tourist destination, by 2030 the ministry of al Hajj will host more than 30 million pilgrim per year. We are planning to develop our model to assist booking.com to invest more on hotels reservation in Makkah. Based on the hotels rating the most rated hotels will appear on the top listing for clients to book. The model will be trained and tested on real data scraped from booking.com website. In this project will present the insights extracted from the data whether linear regression model is appropriate to solve the problem implementation and results. The rest of the report is organized as the following: Section two present data description. Section three presents project prepressing. Section four present data analysis tools. Section five model results, and report is concluded with conclusion in section five.

Data Description

Our data scraped from booking website, the total number of data is 47 rows and five futures the target variable that our model will predict is the hotels rating and the input features that our model will learn are the hotels reviews and number of reviews. The below table shows the data description.

Variables	Data type	Description
Hotel_name	String values	Presents hotels names
hotel_rating	Numeric values	Present hotels rating out of 8
num_reviews	Numeric values	Present total number of reviews
review	Categorical values	Present clients review fabulous good, very good or bad
dist_centre	Numeric values	Present the distance from the center of Makkah

We apply a Data Expletory Analysis on our data and find correlation using heatmap between the valuables as the figure (1) shows hotels rating and number of reviews has a high correlation

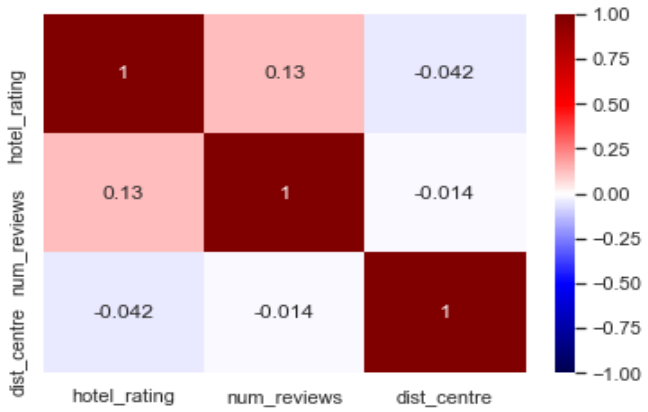

Figure 1 features correlation

The below figure (2) shows the distribution and the relation between features using pair plot of all numeric features such as the closest distance of a hotel to the center the higher rating that the hotel will get reviews. This give us indication our model will perform well.

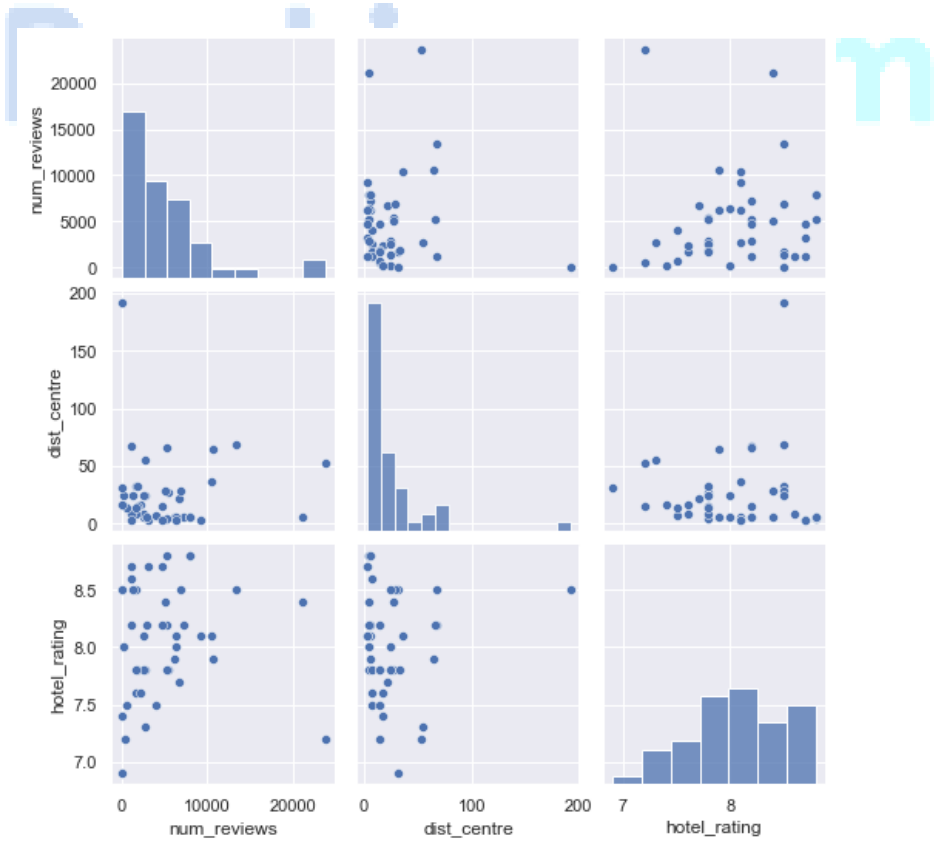

Figure 2 futures distribution

Prepossessing Methos

We apples cleaning techniques from the data such as:

- Remove the string values from numeric
- Drop columns are not need it such as hotels name.
- Convert values type to float such as hotel rating and distance columns.
- Covert categorical variables into dummy variables.

Data Analysis Tools

- BeautifulSoup
- Selenium
- Python (Pandas, numpy, pickle)
- Matplotlib
- seaborn
- scikitlearn

Discussion and results

we split the data into training set and validation set and trained our model in two phases. The first time the model has 0.884 $\mathbf{R^2}$ error in the second time the model got 0,019 $\mathbf{R^2}$. The two models trained on the whole data set with the dummy features -1. We selected the second model and test our model the below table shows the results of the final model.

Accuracy on Testing model (FINAL MODEL)	
$\mathbf{R^2}$	0.890
MAE	0.461-2
MSE	0.243
RMSE	0.493

We plot the model performance using scoter plot in figure (3) the actual data and the prediction data, we can say the model perform well. Most of the predicted data are close to the actual data.

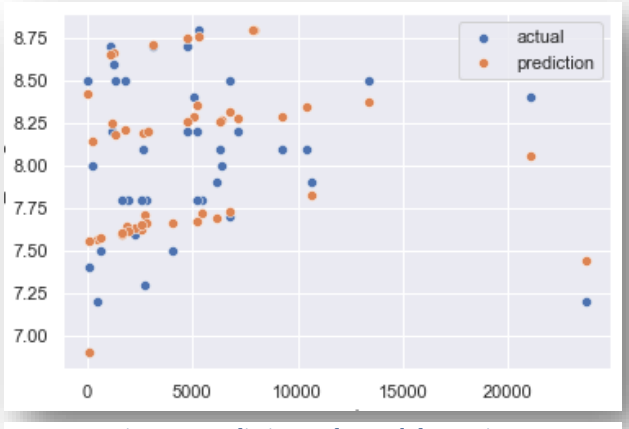

Figure 3 prediction and actual data points

We plot our error using a histogram plot in figure (4) to show the performance of our model.

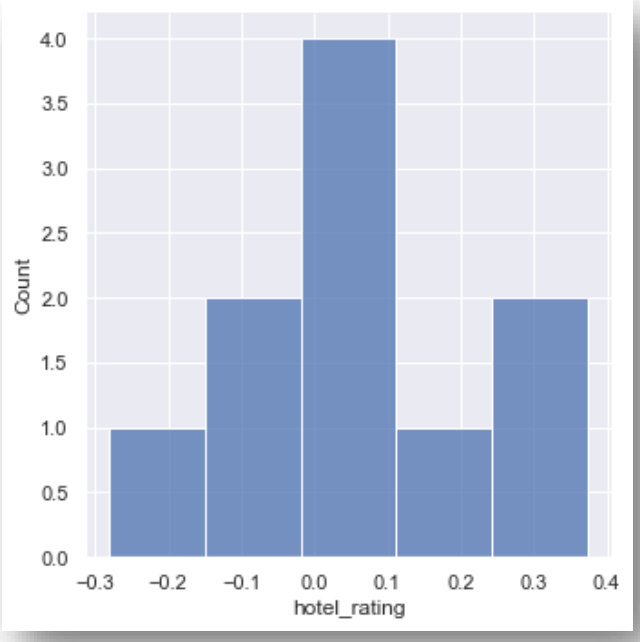

Figure 4 error rate testing model

Conclusion

In this project, we will propose an AI solution for booking.com to predict the hotels rating that has high number of reviews and closest location. Our model has a good potential for booking.com for investment since hotels in Makkah by 2030 will host more than 30 million pilgrim per year. It will enhance booking.com revenue. In the future we are planning to expand our dataset, add more features and tested with more assumptions. Also, we are planning to use our model to predicted the reviews using sentiment analysis methods.