October 25, 2021

Hotel Room Price Predictor

Abstract

In order to regulate hotel room pricing in the major cities of Saudi arabia this project went through the process of building a tool for the ministry of tourism to predict the accommodation prices in riyals given the number of beds, the rating score and number of reviews.

Design

A regression model analysis was conducted that encompasses many features and among them the room price.

To gather hotel data we scraped Booking.com , one of the top and most visited online travel agencies.

We utilized several regression models and tested for the best fit; to ensure the accuracy of the predictor tool.

Data

The data scraped is described by 10 features as follows:

1. Hotel Name
2. Location (eg: Riyadh, Jeddah .. etc)
3. Room Type (eg: Suite, Romm .. etc)
4. Price
5. Price for (eg: per 1 night .. etc)
6. Number of Beds
7. Hotel Rating (1 to 10 scale)
8. Rating Title
9. Number of Reviews
10. Room Size in m^2

Algorithms

1. Fetch:

To ensure the highest accuracy and to avoid incorporating seasonality in our data, we scraped the website for one day continuously.

2. Clean:

The dataset had several unwanted columns, duplicate observations, NaN values and spaces in between the categorical features, so we used pandas library to prepare the data for the regression model.

3. Preprocessing:

The following transformation methods were applied, to standardize the values at an equivalent scale and to linearize some of the features that are not linear in nature against the room price:

1. **Feature Scaling:**
    a. Robust scaling
    b. Standard scaling
2. **Guassian Transformations:**
    a. Log
    b. Box-Cox
    c. Polynomial

4. Model Building:

Building the regression model required understanding of how the dependent variable (Room Price) and independent variables are correlated with

each other and how they are distributed, we utilized the pair plot and normal distribution plot to visualize the data initially then we proceeded to work on the below regression models:

1. Multiple linear Regression
2. Polynomial Regression
3. Random Forest
4. Gradient Booster and others

5. Model Testing:

Regularization

1. Ridge Regularization
2. Lasso Regularization
3. Elastic net Regularization

Parameter Tuning

1. Train test split
2. Cross validation
3. Kfold
4. Grid search for Hyper Parameters
5. Randomized search for Hyper Parameters

** Then we checked the outlier influence and multicollinearity using Variance Inflation Factor

Tools

Data Scraping libraries: Extractor & Selenium

EDA libraries: Pandas, numpy, seaborn, matplotlib

Preprocessing libraries: sklearn, Statsmodel, scipy, pylab

Model Building libraries: sklearn

Model Testing libraries: sklearn

Others:

SweetVIS for reporting data.

Plotly Express plotting.

Missingno Plotting.

Yellowbrick for errors prediction.