Team ID: Sc_19

زينب اسعد محمد اسعد	20201700324
زياد ناجي عباس اسماعيل	20201700319
رؤي علاء سعد	20201701079
بدر احمد بدر احمد	20201700183
مهند محمود عبد المعبود رجب	20201701168
محمد جمال إبراهيم شحاته	20201700678

Hotel Rating Prediction

The goal of this project is to build a machine learning model that can predict Reviewer score based on various features such as Average Score, Positive Review and Negative Review.

The dataset used in the project contains information about hotels in different cities around the world.

And here is the columns name and description that we have in dataset:

Dataset Description:

Feature	Description
Hotel Address	
Review Date	
Average Score	Average Score of the hotel, calculated based on the latest comment in the last year.
Hotel Name	
Reviewer Nationality	
Negative Review	Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
Review Total Negative Word Counts	Total number of words in the negative review.
Positive Review	Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
Review Total Positive Word Counts	Total number of words in the positive review.
Reviewer Score	Score the reviewer has given to the hotel, based on his/her experience.
Total Number of Reviews Reviewer Has Given	Number of Reviews the reviewers has given in the past.
Total Number of Reviews	Total number of valid reviews the hotel has.
Tags	Tags reviewer gave the hotel.
Days since review	Duration between the review date and scrape date.
Additional Number of Scoring	There are also some guests who made a scoring on the service rather than a review. This number indicates how many valid scores without review are in there.

Data Preprocessing

The first step in the project was to preprocess the data. This involved filling any missing with mode if it is categorical and median if it is numerical then removing duplicated values, encoding categorical variables like (Hotel Name, Reviewer Nationality, Positive Reviews and Negative Reviews) , get from the address column the country so that we can make encoding easily ,get the number of days from days_since_review column using regular expression and splitting tags for 3 columns trip, Room ,Nights then scaling the numerical variables.

We also split the data into training and testing sets to evaluate the performance of the model.

Feature Engineering

Make new feature from column Review_Date and Tags.

We extract day, year and month from Review_Date column and trip type, room type and stayed nights from Tags column.

Model Building

We evaluated several machine learning algorithms to determine the best model for predicting Reviewer Score. These included linear regression, Polynomial Regression, Lasso ,Decision tree and Ridge.

We try Polynomial with Ridge or lasso and get large MSE.

We compared the performance of these models using test data and selected …. As our final model due to its mean square error and accuracy

Model Evaluation

We evaluated the performance of our model using mean squared error,

mean square error value of Linear Regression is 1.8142205985604327

mean square error value of Polynomial Regression is 1.6649181355291793

mean square error value of Lasso Regression is 2.667086564652559

mean square error value of Ridge Regression is 1.814082174203968

mean square error value of Decision tree is 3.0361643316272895.

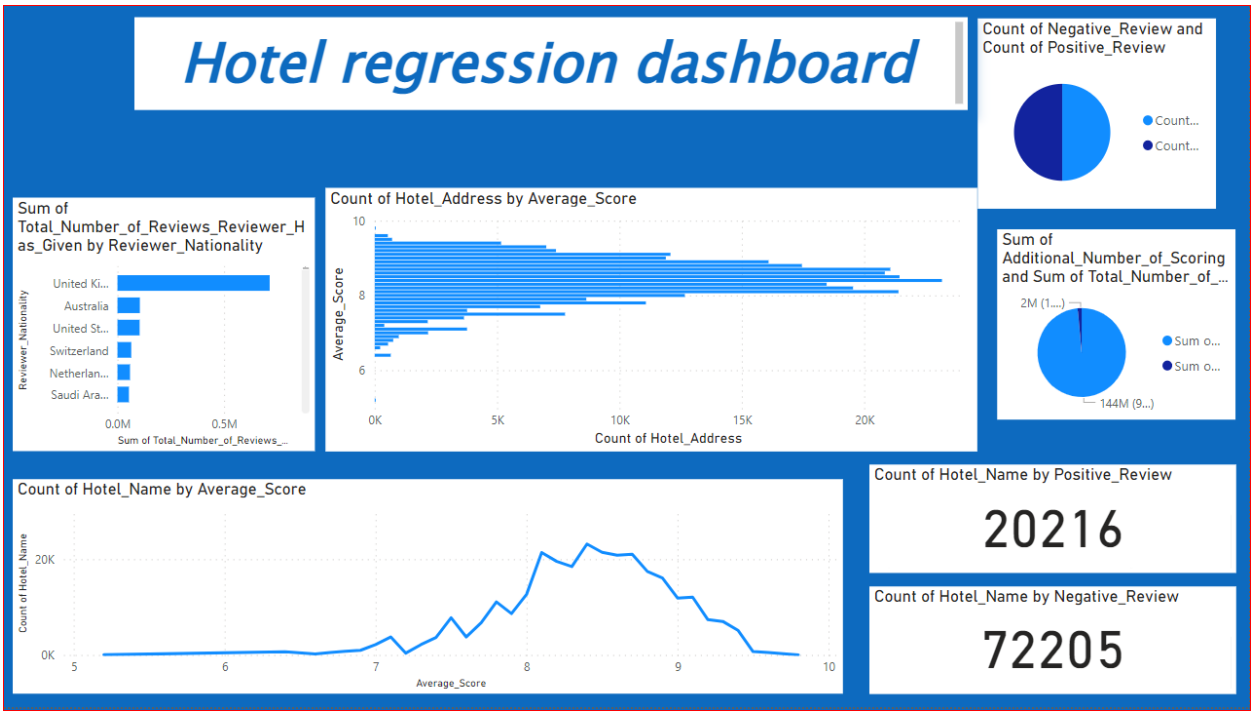
and the best model is the polynomial Regression with degree = 2

Conclusion

When we drop the most effective columns such as Average Score, Review_Total_Positive_Word_Counts and Review_Total_Negative_Word_Counts

And choose different columns to evaluate the models and we get that all columns except the address effect on the models so make the data take all columns.

We know that from power pi Total_Number_of_Reviews_Reviewer_Has_Given counts we get from the people which they are from United Kingdom and there is a file of the power pi attached with the folder of the project.



Bonus Task

We apply sentiment analysis on column of Review (Positive and Negative) and get the score of each review.