HOTEL RATING PREDICTION

TEAM ID: SC_30
TEAM EMAIL: ahmedmohsen2492@gmail.com

TEAM MEMBERS

Ahmed Esmail Mohamed	20201700024
Ebrahim Tarek Mohamed	20201700009
Ahmed Mohsen Fathelbab	20201700067
Ebrahim Ahmed Abdelmaboud	20201700007
Ahmed Mohamed Kamal	20201700080
Yasmine Khaled Atta	20201701154

PREPROCESSING

1. **SPLIT THE DATASET**
   20% for testing (58063 rows) 80% for training (232252 rows)

2. **HOTEL_ADDRESS**
   It was divided into two columns **(hotel_country, hotel_city)** using regular expressions, pycountry and GeoText.

3. **REVIEW_DATE**
   It was converted to DateTime first then extracted the year only to column **(Review_year)** then Review_Date column converted to ordinal.

4. **NEGATIVE_REVIEW, POSITIVE_REVIEW**
   Applied sentiment analysis to classify if the review positive or negative using vader.

5. **REVIEW_TOTAL_NEGATIVE_WORD_COUNTS**
   If the review already negative so it filled by the word count but if the review is positive so it filled by zero.

6. **REVIEW_TOTAL_POSITIVE_WORD_COUNTS**
   If the review already positive so it filled by the word count but if the review is negative so it filled by zero.

7. **TAGS**
   It was divided into multiple columns **(type_of_trip, with_a_pet, people, Room_Type, nights, submitted_from_mobile).**

8. **DAYS_SINCE_REVIEW**
   it contains days or day word after the number and both have been removed.

9. **LAT, LNG**
   Both contains null data and filled with average.

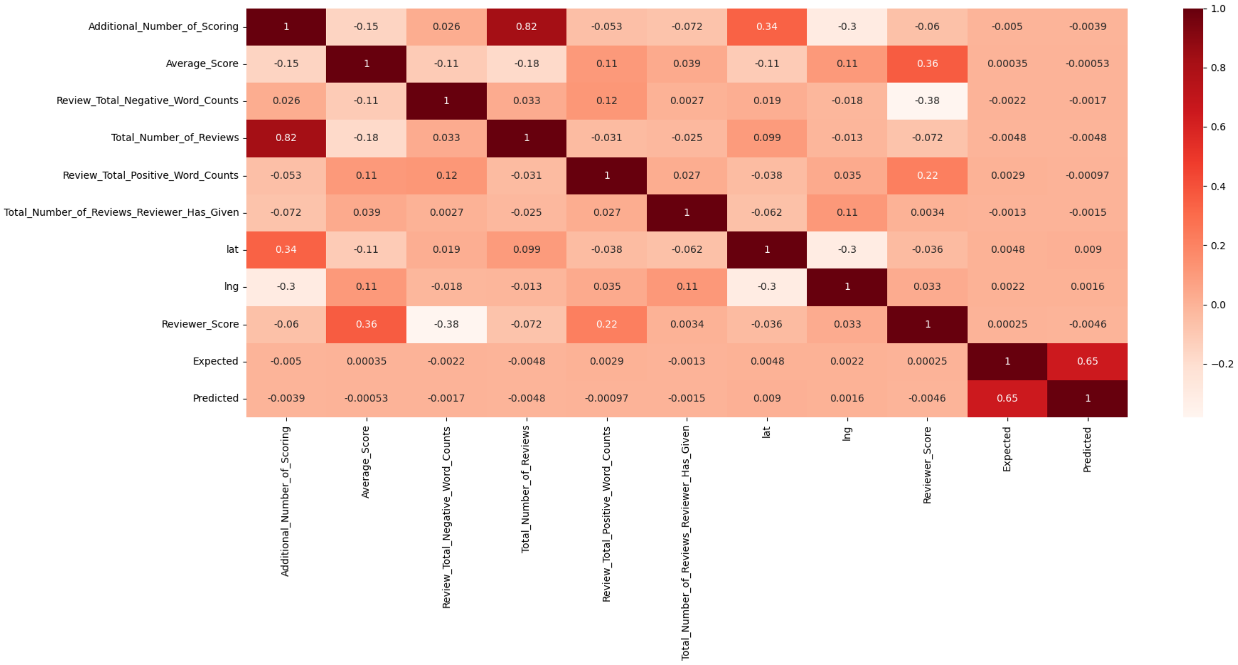
10. **LABEL ENCODING (LabelEncoder model)**
    DEFINITION: is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

    Applied on **(Hotel_Name, Reviewer_Nationality, type_of_trip, people, hotel_country, hotel_city).**

11. FEATURE SELECTION

Applied based on the correlation with **Reviewer_Score** column and all the columns with correlation more than or equal 0.2 have been selected **(Average_Score, Review_Total_Negative_Word_Counts, Review_Total_Positive_Word_Counts).**

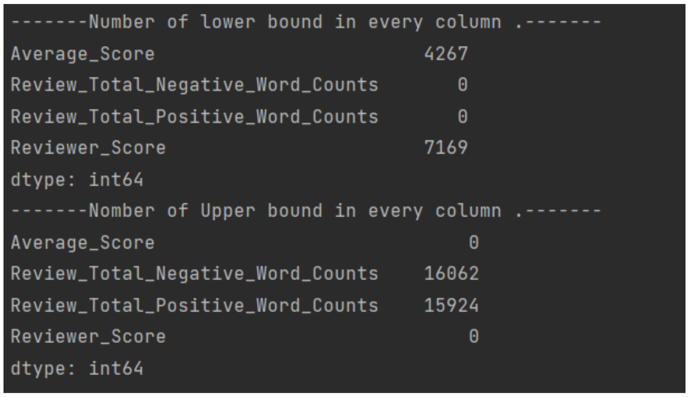
-Then we create a heat map of the correlation matrix using the heatmap function from Seaborn that allows us to visualize the strength and direction of the correlations between different variables in the data:



11. HANDLING OUTLIERS

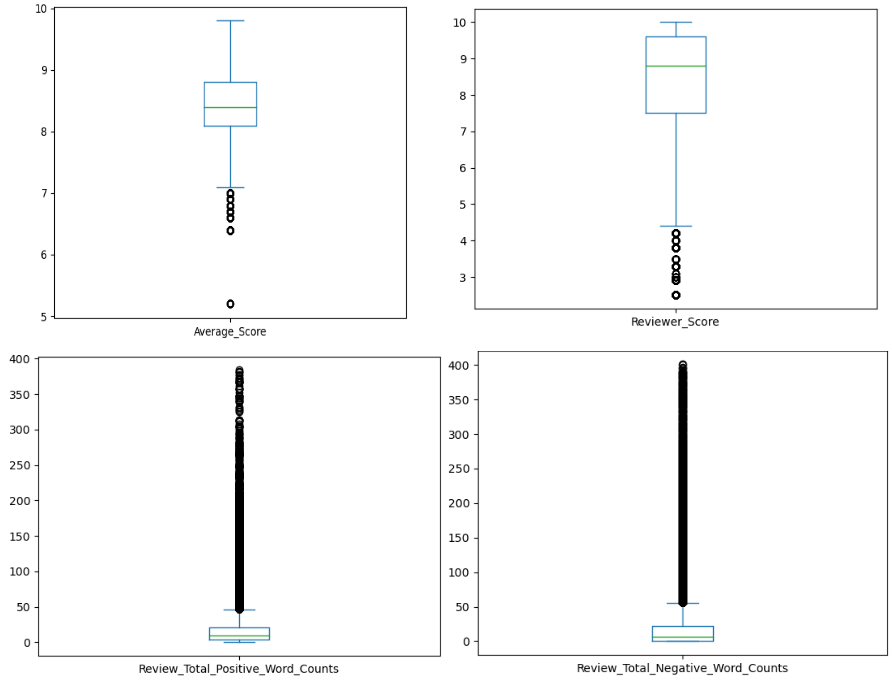
DEFINITION: An outlier is an observation in a dataset that lies an abnormal distance from the other observations in the dataset. Outliers can occur due to measurement error, data entry errors or natural variation in the data. Outliers can affect the results of statistical analysis and machine learning models, so it is important to identify and handle them appropriately.

1- USING STATISTICAL METHOD FOR DETECTING OUTLIERS :



```
-------Number of lower bound in every column .-------
Average_Score                        4267
Review_Total_Negative_Word_Counts       0
Review_Total_Positive_Word_Counts       0
Reviewer_Score                       7169
dtype: int64
-------Nomber of Upper bound in every column .-------
Average_Score                           0
Review_Total_Negative_Word_Counts    16062
Review_Total_Positive_Word_Counts    15924
Reviewer_Score                          0
dtype: int64
```

2- USING BOX PLOT VISUALIZATION FOR DETECTING OUTLIERS :



3- REMOVING OUTLIERS USING THE INTERQUARTILE RANGE (IQR) METHOD.

12. FEATURE SCALING (MinMaxScaler)

DEFINITION: It involves transforming the feature values of a dataset so that they fall within a specified range, typically between 0 and 1.

Applied on selected features.

MODELS

Linear Regression

DEFINITION: is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable

MSE: 0.032908836186091354

SCORE: 0.29922637317226874

BEST PARAMETERS: {'fit_intercept': True, 'normalize': True}

Polynomial Regression

DEFINITION: Polynomial regression is a kind of linear regression in which the relationship shared between the dependent and independent variables Y and X is modeled as the nth degree of the polynomial. This is done to look for the best way of drawing a line using data points.

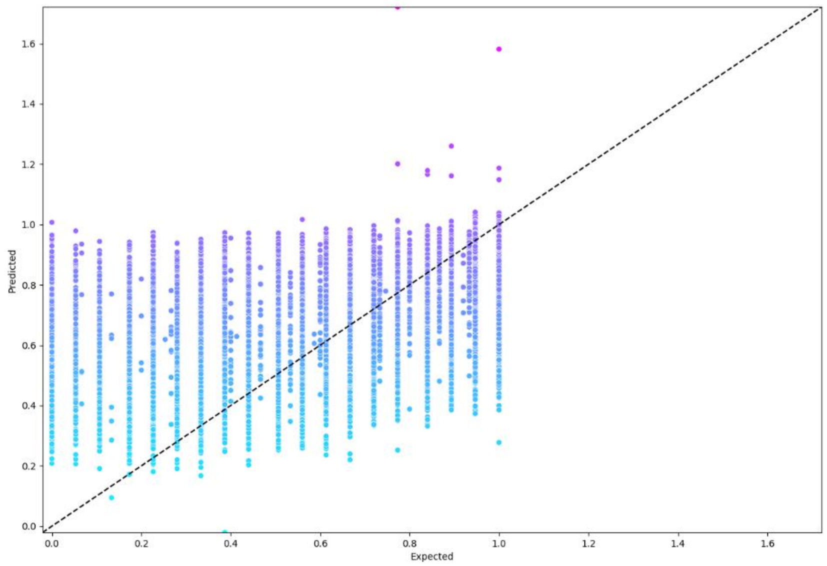
- **Used pipeline to merge linear regression and polynomial features.**

MSE: 0.027699723107936558

SCORE: 0.41997005121604314

BEST PARAMETERS: {'linear__fit_intercept': False, 'linear__normalize': True, 'poly__degree': 7, 'poly__include_bias': True}

VISUALIZE THE EXPECTED VS. PREDICTED VALUES IN POLYNOMIAL REGRESSION MODEL :



Random Forest Regression

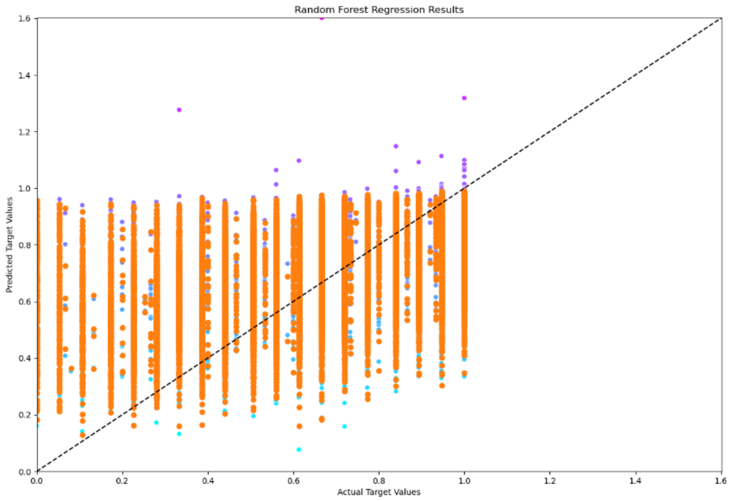
DEFINITION: a supervised learning algorithm that uses ensemble learning method for regression.

MSE: 0.027724169074332266

SCORE: 0.4372377176111383

BEST PARAMETERS: {'max_depth': 10}

VISUALIZE THE EXPECTED VS. PREDICTED VALUES IN RANDOM FOREST REGRESSION MODEL :



Ridge Regression

DEFINITION: Is A Technique Which Is Used For Analyzing Multiple Regression Where The Data Suffers From Multicollinearity.

MSE: 0.032909089824720 24

SCORE: 0.29923119202735904

BEST PARAMETERS: {'alpha': 0.0001, 'fit_intercept': True, 'normalize': True, 'solver': 'sag'}

Gridsearchcv

DEFINITION: a technique for finding the optimal parameter values from a given set of parameters in a grid

Applied to find best parameters for each model.

CONCLUSION

In the Hotel Rating Prediction project , we aim to build a machine learning model that can predict the rating of a hotel based on various features such as address, Average Score, and customer reviews.

To accomplish this task, we first need to gather and preprocess the data. This involves collecting data from various sources such as hotel booking websites and customer review platforms, cleaning the data to remove any missing or incorrect values, and transforming the data into a format suitable for machine learning.

Next, we can use various machine learning algorithms such as linear regression, Ridge Regression, and random forest regression to train and test our models. We can also use techniques such as feature selection, Feature scaling and gridsearchcv to optimize the performance of our models.

Once we have trained and tested our models, we can use them to predict the ratings of new hotels based on their features. We can also evaluate the performance of our models using metrics such as mean squared error (MSE) and score.

Overall, the Hotel Rating Prediction project using Python requires a combination of data preprocessing, machine learning, and data analysis skills to successfully build an accurate and reliable model that can predict hotel ratings.