****

**Detailed Project Report**

***Restaurant Rating PREDICTION***

***January 23, 2023***

***Written by Sofana Benoutiq & Othmane Zoubairi***

**Problem Statement**

The basic idea of analyzing the dataset is to get a fair idea about the factors affecting the establishment of different types of the restaurant at different places in different Cities, aggregate rating of each restaurant, With each day new restaurants opening the industry hasn't been saturated yet and the demand is increasing day by day.

Inspite of increasing demand it, however, has become difficult for new restaurants to compete with established restaurants.

Most of the people here are dependent mainly on the restaurant food as they don’t have time to cook for themselves so that it is very Important to predict witch Restaurants are good and witch Restaurants are not.

**Architecture**

** Start**

** Data Ingestion**

** Data Validation**

** Data Tranformation**

**Model Trainer**

** Hyperparameter Tuning**

**Model Evaluation**

**Prediction**

** Model Pusher**

**Save Model**

**Aplication interface**

**Pushing App to Cloud**

**Deployement**

**END**

**Data Description**

1. Restaurant Id: Unique id of every restaurant across various cities of the world
2. Restaurant Name: Name of the restaurant
3. Country Code: Country in which restaurant is located
4. City: City in which restaurant is located
5. Address: Address of the restaurant
6. Locality: Location in the city
7. Locality Verbose: Detailed description of the locality
8. Longitude: Longitude coordinate of the restaurant’s location
9. Latitude: Latitude coordinate of the restaurant’s location
10. Cuisines: Cuisines offered by the restaurant
11. Average Cost for two: Cost for two people in different currencies (local currency)
12. Currency: Currency of the country
13. Has Table booking: yes/no
14. Has Online delivery: yes/ no
15. Is delivering: yes/ no
16. Switch to order menu: yes/no
17. Price range: range of price of food
18. Aggregate Rating: Average rating out of 5
19. Rating color: depending upon the average rating color
20. Rating text: text on the basis of rating of rating
21. Votes: Number of ratings given

**Data Ingestion:**

The project is constructed with folder and files within them.

The first step we create data ingestion folder to insert url data and to split data in train and test.

**Data Validation**

In this step, we perform different sets of validation on the given set of training files.

1. We create folder Validation, then we validate the name of the files based on the given name in the schema file
2. Number of Columns - We validate the number of columns present in the files,
3. Name of Columns - The name of the columns is validated and should be the same as given in the schema file.
4. The datatype of columns - The datatype of columns is given in the schema file.
5. Null values in columns - If any of the columns in a file ,we use datadrift and report html file to remove it.
6. Report file give us all Graphic about data and deep analysis about correlation of Data.

**Data Transformation**

We split data in train and test data path and transform it base on

preprocessing . we save the preprocessing Model and we pickle it

if we have another data, we can directly use the saved

preprocessing object

Transformation Folder will also be exported as a Train and Test

file to be used for model training.

**Model Trainer**

- After clusters and Model Yaml are created, we find the best model for each cluster. We are using two Regression algorithms, "Random Forest Regressor" and “Linear Regression". For each cluster, both the algorithms are passed with the best parameters derived from GridSearchCV.

We calculate the RMSE scores for both models and select the model with the best ACCURACY score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

**Model Evaluation**

Model can make two kinds of wrong predictions:

Predicting that the customer will give a good Restaurant Rating when they don"t. - False Positive.

Predicting that the customer will not give a good Restaurant Rating when they do. - False Negative. The Restaurant’s objectives are:

Make Marketing Expenditure more efficient and focused on the customers that would actually give a good rating to the restaurant. Predict and Identify all potential customers who will give a good rating to the restaurant. Metric for Optimization:

For the above objectives, its important that both False positive and False negative values are low. Hence we would want the RMSE score to be maximized. The greater the F1-Score, greater the chances of predicting both classes correctly.

We will build following models, tune them and compare the outcome of all the models:

Linear Regression

Stacking Regressor

Bagging Regressor

Decision Tree model.

Extra tree Regressor

Random Forest Regressor

ADABoost.

GradientBoost.

XG Boost.

**Model Pusher**

We save best model in pickle file and we use it by prediction in new dataset

**Deployment**

We will be deploying the best model into the cloud Platform Hiroku.

This is a workflow diagram for the prediction of using the trained model.

**Conclusion**

We collect data from CSV file, we tried to fill estimate values using related column. Firstly we design a random model, a **random model** is something which randomly chooses values from 1.0 to 5.0, such model gives **MSE 2.12**.

We tried only 5 one-hot encoded features and try different models Random Forest Regressor was most learning model, so we tune model using **grid-search** technic, **minimal MSE** = 0.03485. Then we tried with 7 one-hot encoded features and try on different models. Again Random Forest regressor was winning the race. we achieved **MSE = 0.01404**.

Then we done some **Feature Engineering**, used response coded feature, but this time “Linear Regression” perform well than previous model, Random Forest Regressor is winning the race as usual. we achieved **MSE =0.00353**.

Both model have hight accuracy between 96.00% and 98% without underfitting the Data.