**DPR**

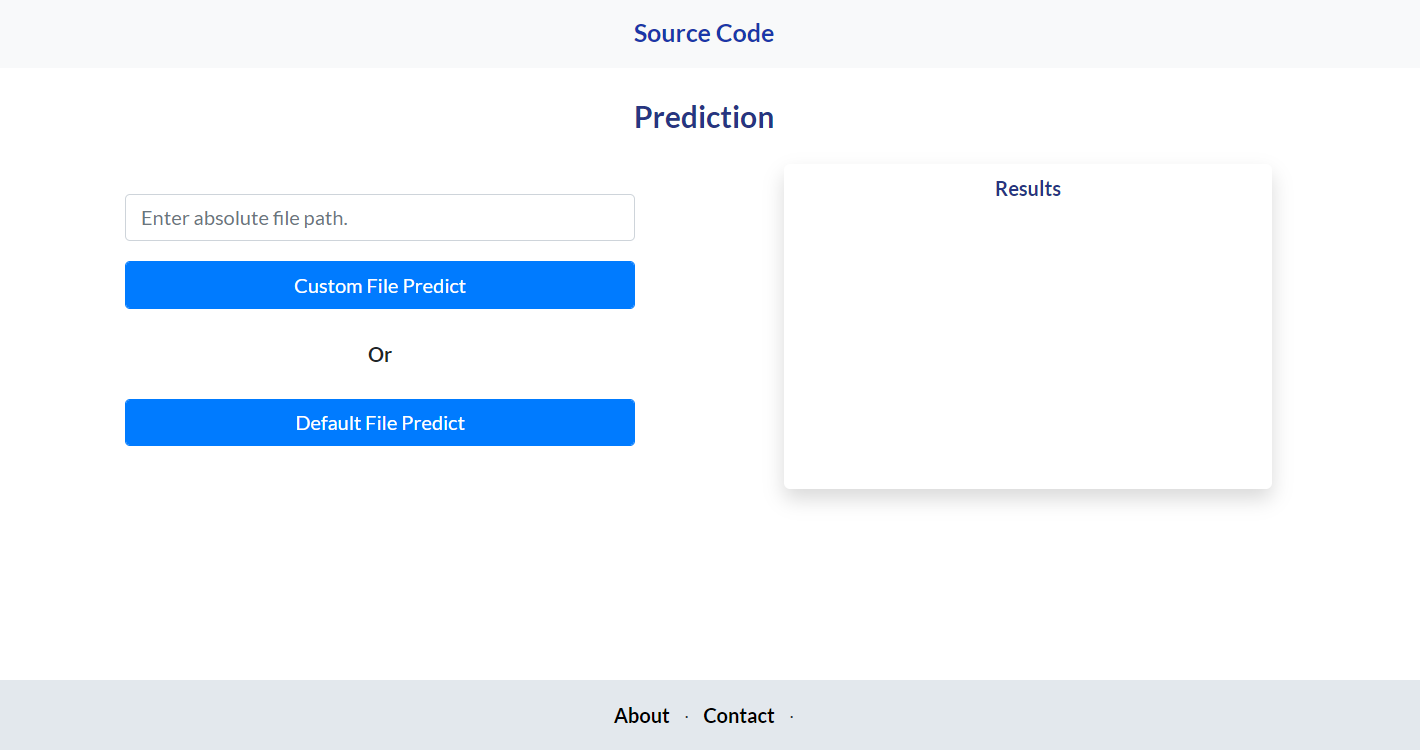
**Project Name:** [Zomato-Restaurant-Price-Prediction](https://github.com/Rishabh1501/Zomato-Restaurant-Price-Prediction)

**About the Project:**

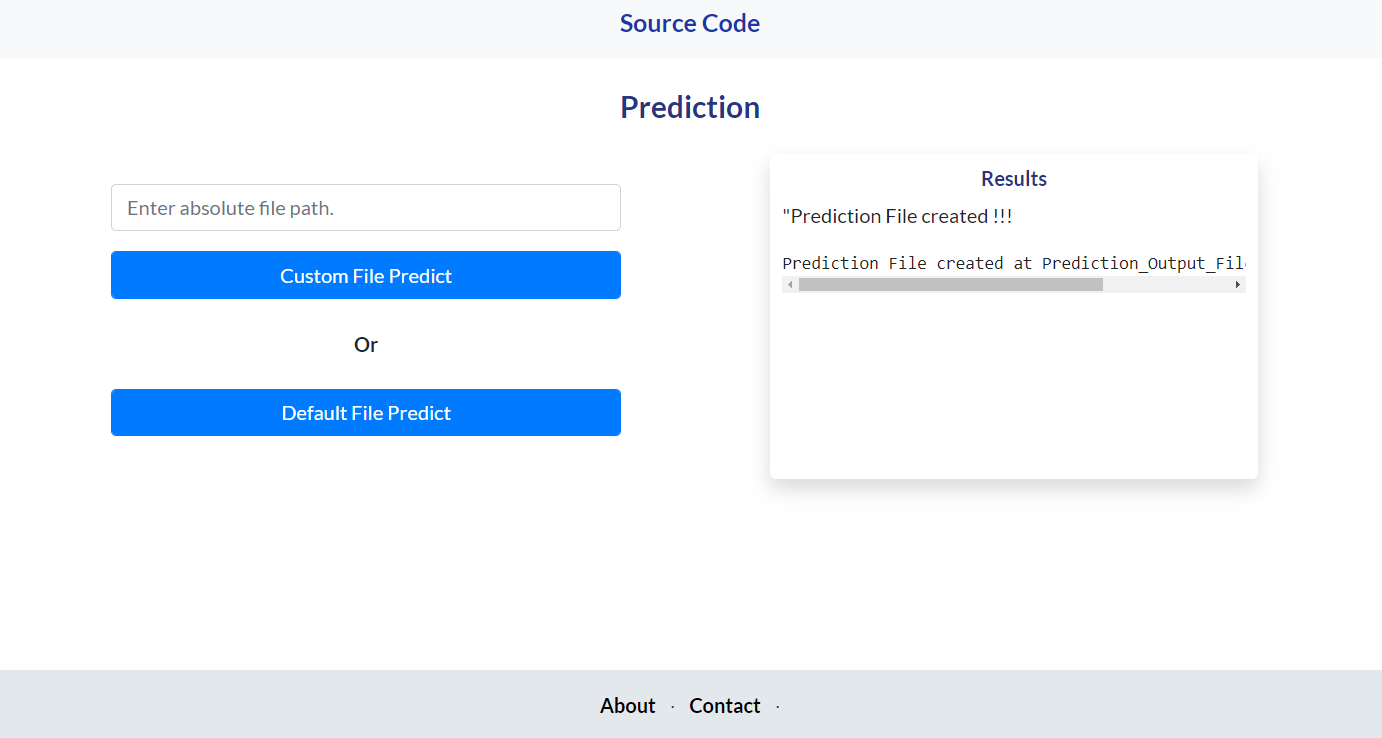
The goal of this project is to build a regression model to predict the cost of restaurant for two people based on the given indicators in the training data. The project follows a very structural approach and uses Flask Framework that provides a very intuitive UI. The Inputs are taken in the form of a .csv file and the Prediction file is also created in a .csv format.

Some Snaps of the project:

* **The index page of the project**



* **Page after Prediction**

****

**Features**:

* Works with every Operating System.
* Can be used with both Post and Get Requests.
* Very Simple to use and understand.
* Properly maintained logs.

**Data Description**

Data Description: This dataset predicts the visibility distance based on the different indicators as below:

1. **serial:** The index/serial number
2. [**URL:** The](URL:The) URL for the restaurant
3. **address:** The address of the restaurant.
4. **name:** The name of the restaurant.
5. **online\_order:** Does the restaurant allow online order (Yes/No)
6. **book\_table:** Does the restaurant allow table booking (Yes/No)
7. **rate:** Restaurant rating out of 5.
8. **votes:** The number of votes for the restaurant.
9. **Phone:** Resyaurant contact number
10. **Location:** Location of the restaurant
11. **rest\_type:** Type of the restaurant (Casual Dining, Café etc.).
12. **dish\_liked:** The Most liked dishes in the restaurant (Pasta, Lunch Buffet, Ma-sala Papad etc.)
13. **cuisines:** North Indian, Mughlai, Chinese etc.
14. **approx\_cost** (for two people)**:** In Rupees. (target Column)
15. **reviews\_list:** The List of reviews
16. **menu\_item:** Open Dosa', 'Benne Set Dosa' etc.
17. **listed\_in (type):** Buffet, café etc
18. **listed\_in (city):** The part of city where restaurant is listed.

Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

**Data Validation**

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

1. **Name Validation-** We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the name, we check for the length of date in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder."
2. **Number of Columns-** We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then the file is moved to "Bad\_Data\_Folder."
3. **Name of Columns-** The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".
4. **The datatype of columns-** The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad\_Data\_Folder".
5. **Null values in columns-** If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad\_Data\_Folder".

**Data Insertion in Database**

1) **Database Creation and connection-** Create a database with the given name passed. If the database is already created, open the connection to the database.

2) **Table creation in the database-** Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" based on given column names and datatype in the schema file. If the table is already present, then the new table is not created and new files are inserted in the already present table as we want training to be done on new as well as old training files.

3) **Insertion of files in the table-** All the files in the "Good\_Data\_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".

**Model Training**

1) **Data Export from Db-** The data in a stored database is exported as a CSV file to be used for model training.

2) **Data Preprocessing**

a) Drop columns not useful for training the model. Such columns were selected while doing the EDA.

b) Replace the invalid values with numpy “nan” so we can use imputer on such values.

d) Check for null values in the columns. If present, remove them.

e) Scale the training and test data separately

f) Encode the categorical values.

3) **Clustering-** KMeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using "KneeLocator" function. The idea behind clustering is to implement different algorithms.

To train data in different clusters. The Kmeans model is trained over preprocessed data and the model is saved for further use in prediction.

4) **Model Selection-** After clusters are created, we find the best model for each cluster. We are using two algorithms, "Decision Tree Regressor" and "XGBoost regressor". For each cluster, both the algorithms are passed with the best parameters derived from GridSearch. We calculate the Rsquared scores for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

**Prediction Data Description**

Client will send the data in multiple set of files in batches at a given location. Data will contain climate indicators in 10 columns.

Apart from prediction files, we also require a "schema" file from client which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns and their datatype.

**Data Validation**

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

1) **Name Validation-** We validate the name of the files on the basis of given Name in the schema file. We have created a regex pattern as per the name given in schema file, to use for validation. After validating the pattern in the name, we check for length of date in the file name as well as length of time in the file name. If all the values are as per requirement, we move such files to "Good\_Data\_Folder" else we move such files to "Bad\_Data\_Folder".

2) **Number of Columns-** We validate the number of columns present in the files, if it doesn't match with the value given in the schema file then the file is moved to "Bad\_Data\_Folder".

3) **Name of Columns-** The name of the columns is validated and should be same as given in the schema file. If not, then the file is moved to "Bad\_Data\_Folder".

4) **Datatype of columns-** The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If dataype is wrong then the file is moved to "Bad\_Data\_Folder".

5) **Null values in columns-** If any of the columns in a file has all the values as NULL or missing, we discard such file and move it to "Bad\_Data\_Folder".

**Data Insertion in Database**

1) **Database Creation and connection-** Create database with the given name passed. If the database is already created, open the connection to the database.

2) **Table creation in the database-** Table with name - "Good\_Data", is created in the database for inserting the files in the "Good\_Data\_Folder" on the basis of given column names and datatype in the schema file. If table is already present then new table is not created, and new files are inserted the already present table as we want training to be done on new as well old training files.

3) **Insertion of files in the table-** All the files in the "Good\_Data\_Folder" are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is moved to "Bad\_Data\_Folder".

**Prediction**

1) **Data Export from Db-** The data in the stored database is exported as a CSV file to be used for prediction.

2) **Data Preprocessing**

a) Drop columns not useful for training the model. Such columns were selected while doing the EDA.

b) Replace the invalid values with numpy “nan” so remove them.

d) Scale the training data.

e) convert categorical to numerical

3) **Clustering-** KMeans model created during training is loaded, and clusters for the preprocessed prediction data is predicted.

4) **Prediction-** Based on the cluster number, the respective model is loaded and is used to predict the data for that cluster.

5) Once the prediction is made for all the clusters, the predictions along with the original names before label encoder are saved in a CSV file at a given location and the location is returned to the client.