Machine learning in Predicting Restaurant Food Cost

Problem Definition:

All of us must have craving for at least a few favourite food items, we may also have a few places where we like to get them, a restaurant which serves our favourite food the way we want it to be. But there is one factor that will make us reconsider having our favourite food from our favourite restaurant, the **cost**.

In this article we will try to predict the cost of the food for two-person meal served by the restaurants across different cities in India. We will use Data Science skills to investigate the factors that really affect the cost.

Data Analysis:

We have separate data for training and testing purpose in this example.

Size of training set: **12,690** records

Size of test set: **4,231** records

Following are the columns in our dataset and its description:

**TITLE:**The feature of the restaurant which can help identify what and for whom it is suitable for. This column is of object data type.

**RESTAURANT\_ID:**A unique ID for each restaurant. This column is of integer data type.

**CUISINES:**The variety of cuisines that the restaurant offers. This column is of object data type.

**TIME:**The open hours of the restaurant. This column is of object data type.

**CITY:**The city in which the restaurant is located. This column is of object data type.

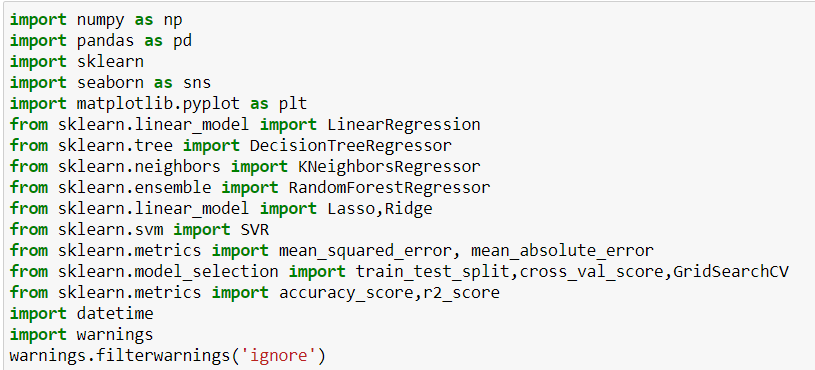
**LOCALITY:**The locality of the restaurant. This column is of object data type.

**RATING:** The average rating of the restaurant by customers. This column is of object data type.

**VOTES:**The overall votes received by the restaurant. This column is of object data type.

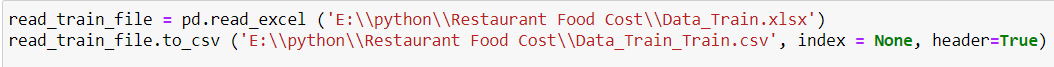
**COST:** The average cost of a two-person meal. This column is of integer data type.

Here ‘**Cost’** is our target variable which we will try to determine in our machine learning model. Now since our target variable can have continuous values this is a clear case wherein we will have to use Linear Regression type of algorithms.



Here we are importing all the necessary libraries and classes which we will be using in our model building process.

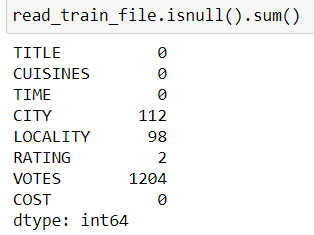
Since this is a regression type of problem we are importing various regression libraries and all the matrices libraries which are used for this type.



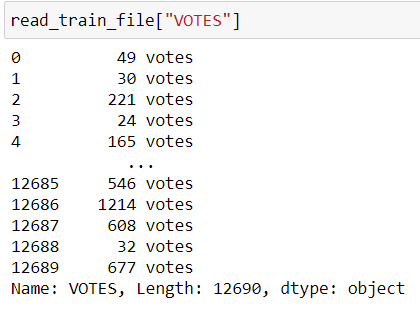
Over here our input file/data is in xlsx file. We first load data in a variable and then use .to\_csv method of pandas to convert it into csv format.



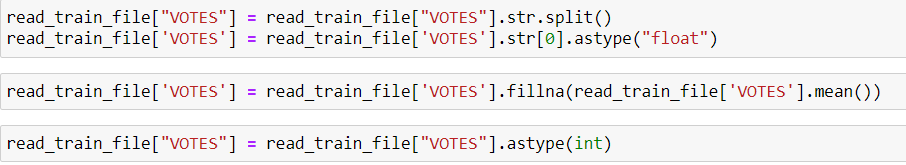
Since ‘RESTAURANT\_ID’ column has unique value for each data entry, we are going to drop this column.



Here we can see that Columns CITY, LOCALITY, RATING and VOTES have null values in it which we will need to work on it.

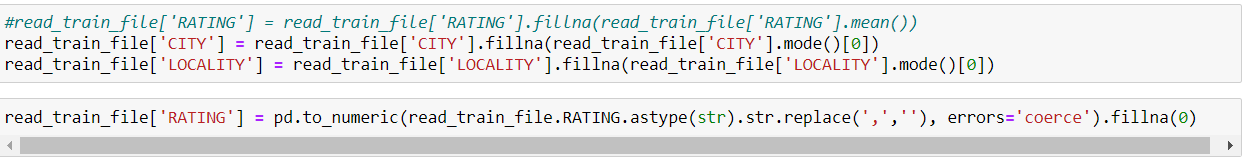


From column ‘VOTES’ we will have to extract only the integer part of data, as it a very useful column in our model building process.



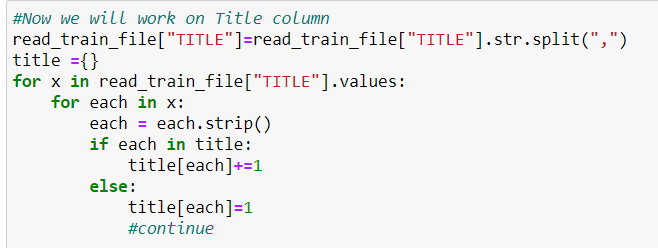
Here we are splitting data in ‘VOTES’ column by using split() method and then extracting first (numerical) part of data.

We can see that this column has 1204 null values and hence we are replacing these null values with mean value of this column. Since this column is of integer data type, we are then converting data in this column into integer.

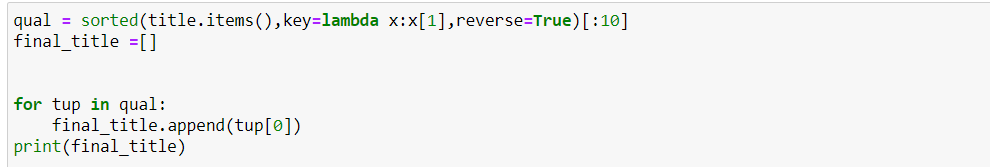


We can see that Columns City and LOCALITY has null values. Since these 2 columns are object data type, we are replacing null values with most used value.

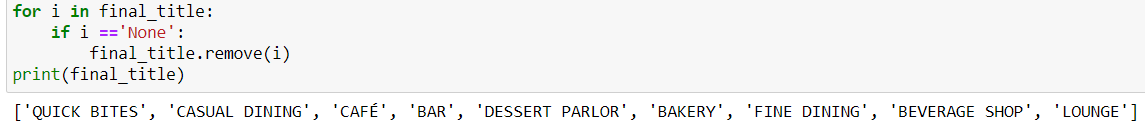
Column RATING has 2 null values, we are replacing null values with ‘0’ value.



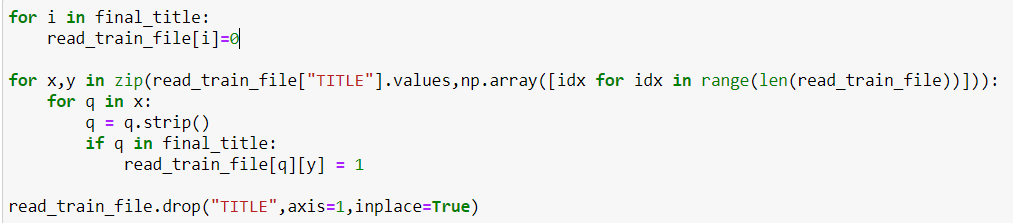
Here we have separated various titles from all rows of column ‘TITLE’ and then grouped them to take a count of number of instances for each title.



Here we have created list of title of restaurant wherein top 10 titles which has most number of count are selected.

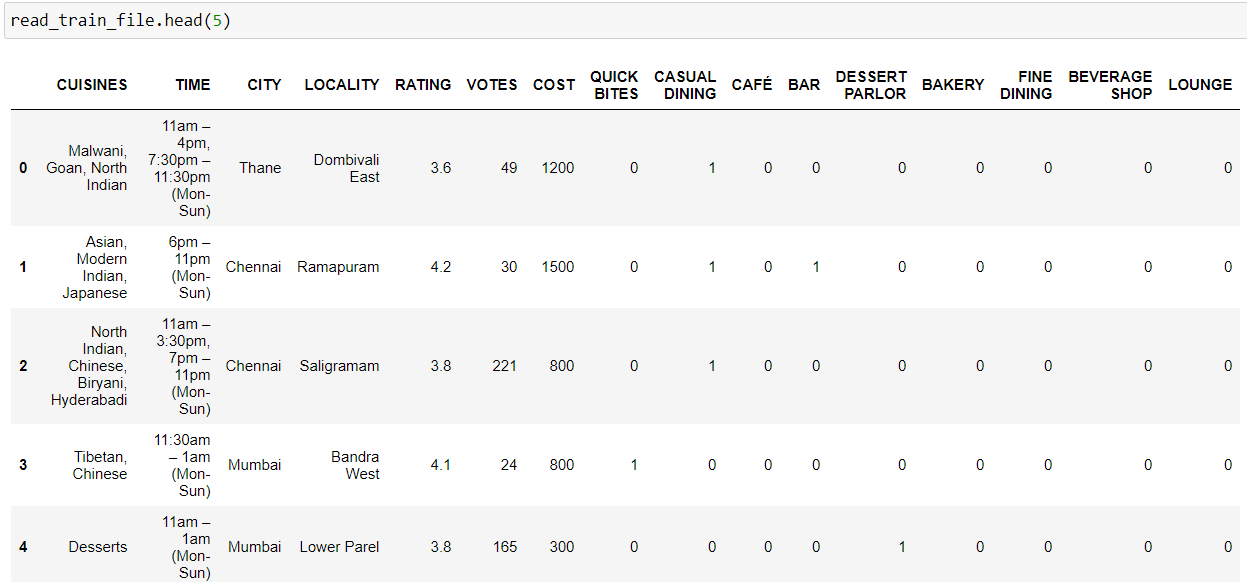


We had found that one of title has value 'None' and hence we have removed it from our list.

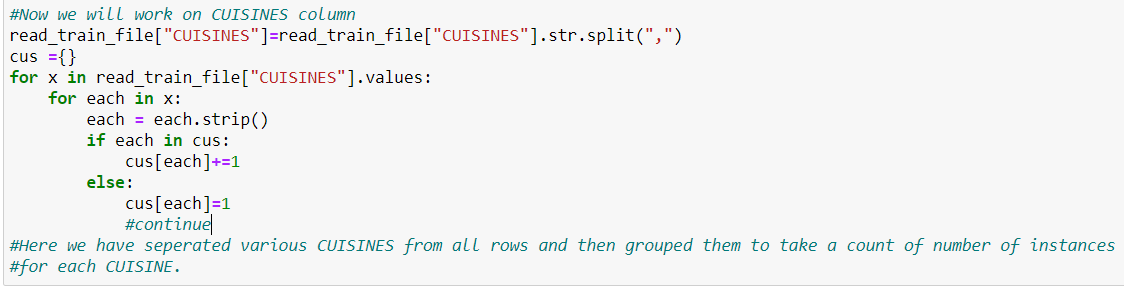


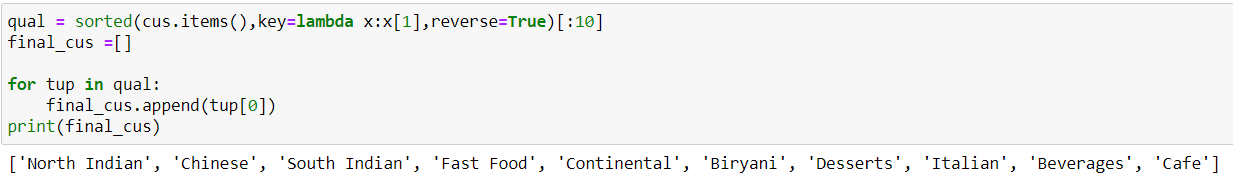
Now we are setting each value in our list as a new column in our dataset and then giving value as ‘1’ to entries where this new column is found else ‘0’ is set. Finally, we are then dropping column ‘TITLE’ from our dataset.

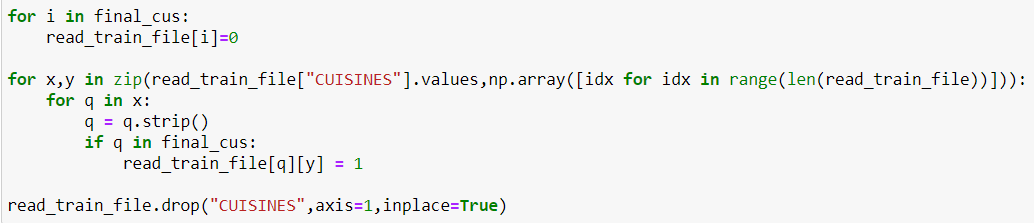
We can see below output after this operation is performed. Notice the new columns which are now added in our dataset.

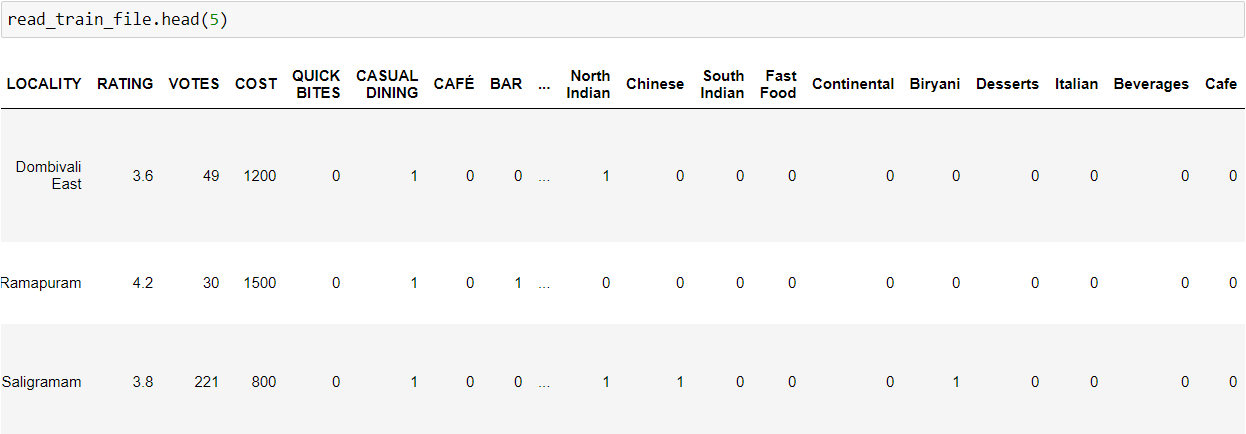


We are then performing similar operation for column CUISINES also as below:

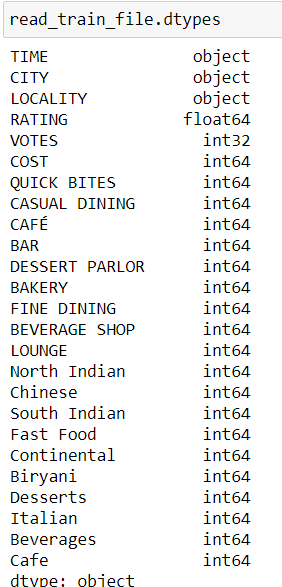




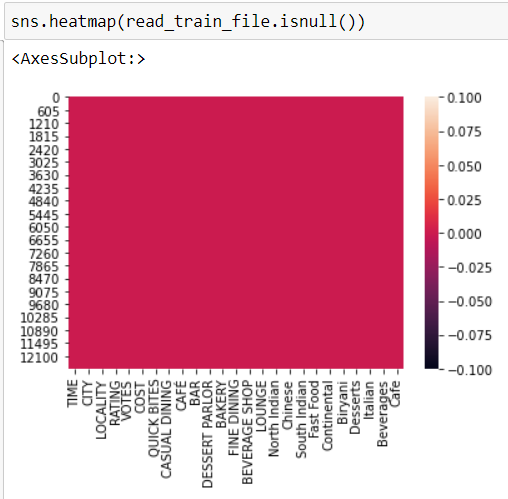




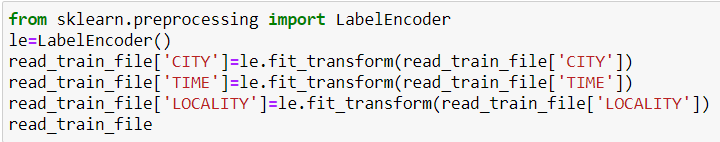
Please notice the new columns like ‘Café’, ’Desserts’, etc which are now added to our dataset.



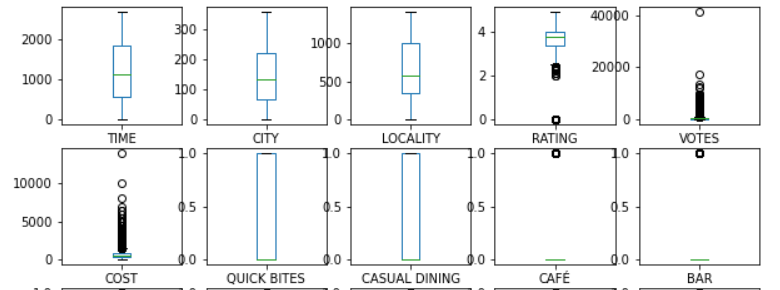
We can see list of all columns which are now part of our dataset along with their data type.



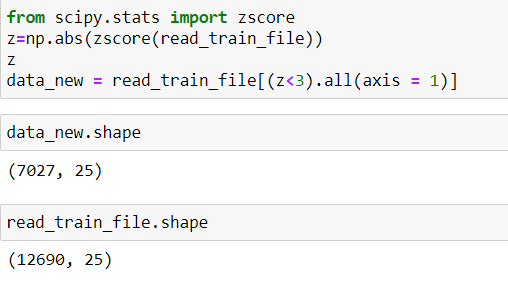
We can see heat map shows that there are no null values present on our dataset now.



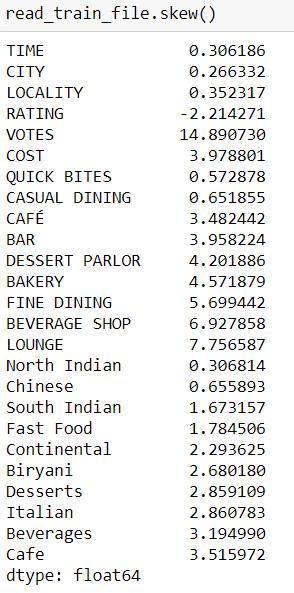
We can see that columns Time, Locality and City are of object data type and hence we are encoding them into numerical data type.



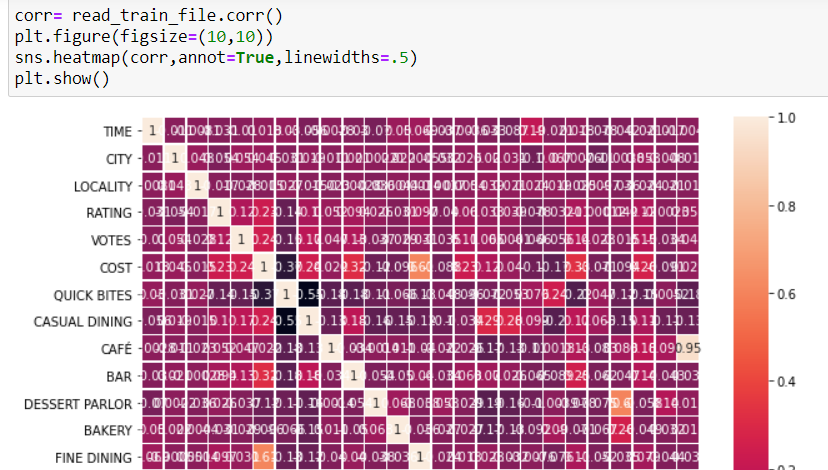
Now we can see that columns Rating, Votes and Cost have outliers present in them.



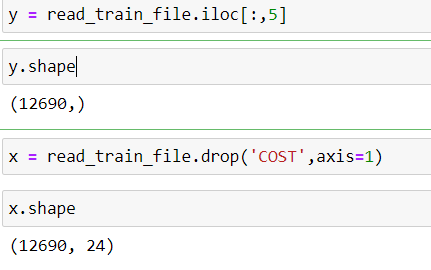
By removing outliers using z-score technique we can see that there is 44% data loss and hence we will go ahead with outliers for our model building process.



We can also see that there is lot of skewness present in our data (especially in columns RATING, VOTES and COST) on which we will have to work.



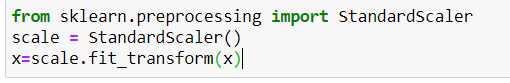
Above heatmap shows that column ‘Casual Dining’ and ‘Quick Bites’ are having correlation.



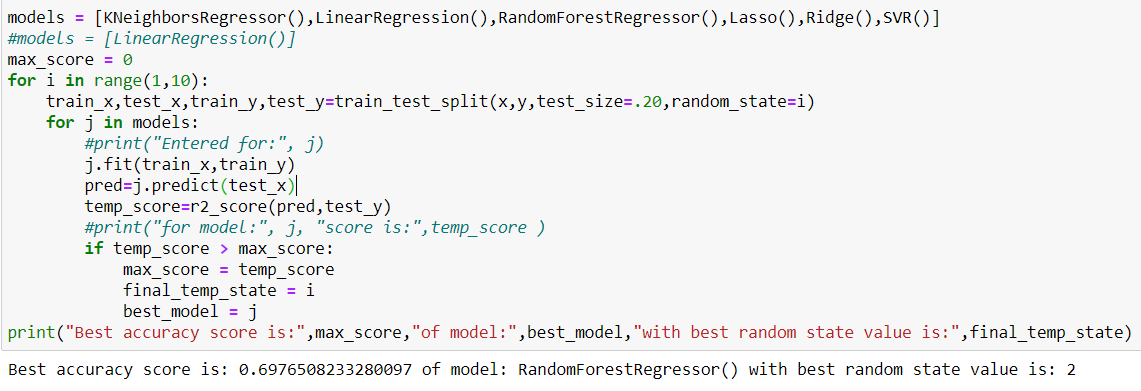
Now we have separated our target variable and our data variables accordingly.



We are treating skewness by using ‘yeo-johnson’ method.



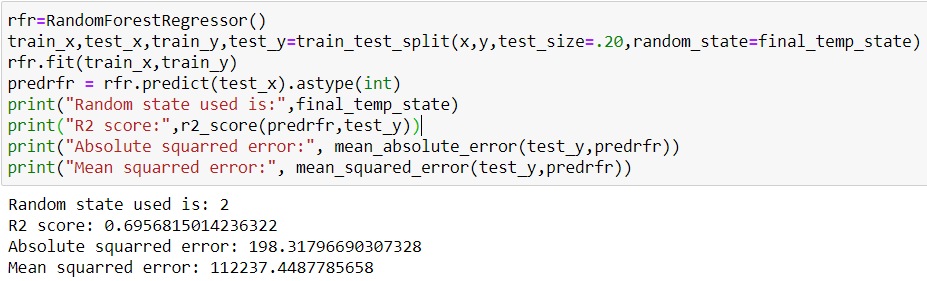
By using StandardScaler class we are transforming our data variables which will be helpful in our model building process.



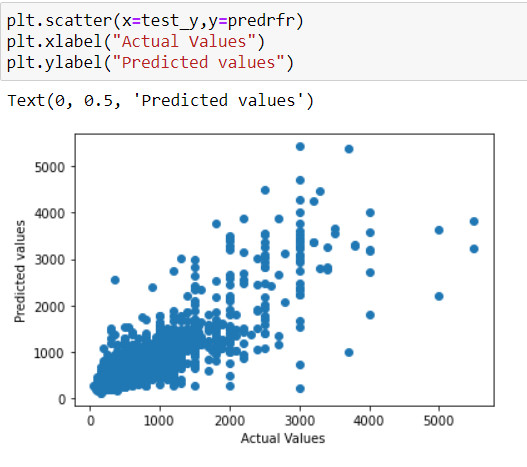
Here we are using ‘train\_test\_split’ method to split data into train and test parts by using test size of 20%. So here 80% of data is used for training and then remaining 20% is used for testing.

We are testing our model with KNeighborsRegressor(), LinearRegression(), RandomForestRegressor(), Lasso(), Ridge() and SVR() methods to check which algorithm gives us the best performance and for which best random state value between 1 and 100.

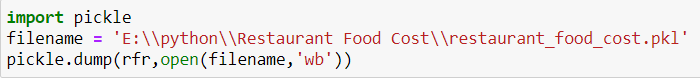
We can see that RandomForestRegressor algorithm gives us the best r2\_score of 69.76% for random state value of 2.



Here we are predicting the output by using ‘predict’ method and then using this predicted values for finding out the r2\_score, mean\_absolute\_error and mean\_squared\_error values.



We can see that average cost of restaurant is highly concentrated between 500 to 1500 Rupees for two-person meal.

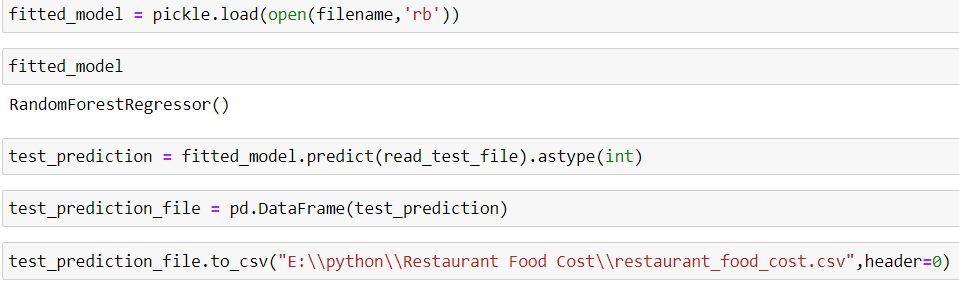


Now we are saving our model in a pickle object so that it can be used for our test dataset.

From our training data set we have:

1. Dropped column ‘RESTAURANT\_ID’.
2. Splitted and extracted numerical value part from column ‘VOTES’.
3. Replaced null value from columns ‘CITY’ and ‘LOCALITY’ with most used values in these columns.
4. Replaced null value with 0 in column ‘RATING’.
5. Separated various values from each row from columns ‘TITLE’ and ‘CUISINES’, grouped them, used 10 most used values from these columns and then created new columns for these values.
6. Label Encoded columns ‘CITY’, ‘TIME’ and ‘LOCALITY’.
7. Treated skewness by using ‘yeo-johnson’ method.

We have to perform same operations on our test dataset also before using our model on it.



After performing data cleaning process on test data also we are now using our saved model on test data.

Here we are using load method of pickle library to do this.

We have loaded our model in ‘fitted\_model’ variable and then used predict method to find value of output variable ‘COST’ for our test data and then saving our output in a csv file.

Complete source code of my model can be found at path:

<https://github.com/japitale/DT_Evaluation_projects/blob/main/ProjectsDynamicsM20_project14_Restaurant_Food_Cost%20.ipynb>

**We can see how machine learning technology can be used to predict the COST for two-person meal for different restaurants in various cities.**