Zomato restaurant

Zomato is an Indian multinational restaurant aggregator and food delivery company. It was founded by Deepinder Goyal and Pankaj Chaddah in 2008.

1. Problem Definition

The problem presented here involves data analysis to predict (i) Average Cost for two, and (ii) Price range. The dataset provided here has two files.

Dataset 1:

The first file contains the information related to the price of a specific cuisine of a restaurant based on location, ratings, votes and other services like table booking, online delivery etc. This file does not contain the name of countries but has instead a unique numerical code called “Country code”.

Dataset 2:

The second file contains two variables, namely “Country code” and “Country. This dataset can be used to map country code presented in the first file to the name of respective countries. There is a list of total 15 country codes present in it to be mapped to their respective countries.

Features:

There are total 21 variables and 9551 rows present in Dataset 1. There are both categorical and numerical variables present in the dataset.

2. Data Analysis

Step 1: To read and understand the two datasets

The “Country code” variable present in “Dataset 1” was replaced by the “Country” variable, after mapping “Country code” against respective “Country” from “Dataset 2”.

Result: No change in number of variables

Step 2: To identify the variables present

Both categorical and numerical variables are present in the dataset.

Categorical variables: Total 13 categorical variables were present. These were namely:
'Restaurant Name', 'City', 'Address', 'Locality', 'Locality Verbose', **'Cuisines'**, 'Currency', **'Has Table** booking', **'Has Online delivery'**, **'Is delivering now'**, 'Switch to order menu', 'Rating color', 'Rating text'

Numerical variables: There were 8 numerical variables present. These were namely:

'Restaurant ID', 'Country Code', **'Longitude', 'Latitude'**, **'Average Cost for two', 'Price range', 'Aggregate rating', 'Votes'**

Step 3: To analyses variables

There were two variables titled: “Average Cost for two” and “Currency”, which needed to be analysed. This is because “Average Cost for two” contains cost in different currencies. So, in order to make the value in “Average Cost for two” consistent in terms of same currency, a currency exchange rate as on 14th May 2024, 2 PM IST was used. So, all the values got converted into USD. After this, the “Currency” and “Average Cost for two” variable was replaced by a single variable ““Average Cost for two_USD”.

Step 4: Dropping irrelevant variables

Following variables were dropped:

'Average Cost for two', 'Currency', 'Restaurant ID', 'Restaurant Name', 'Rating color','City', 'Address', 'Locality', 'Locality Verbose', 'Rating text', 'Switch to order menu', 'Country Code'

“Currency”, “Average Cost for two”: These variables were replaced by a single variable ““Average Cost for two_USD” as explained in Step 3.

'Restaurant ID', 'Restaurant Name': These variables just signify the identity of the restaurant and had no physical influence on the price of the cuisine.

'Rating color', 'Rating text': There were three variables namely: 'Rating color', 'Rating text' and 'Aggregate rating' giving the same information. This would have led to multicollinearity. So out of these 'Rating color', and 'Rating text' were dropped.

'City', 'Address', 'Locality', 'Locality Verbose’: There were six variables present in the dataframe pertaining to the location of the restaurant. These were 4 categorical variables 'City', 'Address', 'Locality', 'Locality Verbose', and 2 numerical variables ‘latitude’, and ‘longitude’. So, in order to avoid multicollinearity, the 4 categorical variables were dropped.

'Switch to order menu': This categorical variable had only one unique value. So, this variable was fixed all across the different rows and had no variation. So, this variable was not at all influencing the price of the cuisine and hence was dropped.

'Country Code': The “Country code” variable present in “Dataset 1” was replaced by the “Country” variable, after mapping “Country code” against respective “Country” from “Dataset 2”. Hence there was a new variable “Country” instead of “Country code”.

So, total number of variables dropped = 11.

Result: No. of variables remaining = 10.

Step 5: To explore missing values

Missing values in variable **"Cuisines"** = 9

There were 9 missing values present in the categorical variable **"Cuisines".** These missing values were imputed by the mode of the remaining values in the variable.

Step 6: To explore distribution of all the variables

The numerical variables were segregated as continuous and discrete. These were

Discrete variables: 'Price range', 'Votes'

Continuous variables: 'Longitude', 'Latitude', 'Aggregate rating', 'Average Cost for two_USD'

The continuous variables were plotted using a histogram. All these variables were observed to be highly skewed. The skewness of the variables was:

Longitude: **2.81**, Latitude: **3.08**, Aggregate rating: **0.95** and Average Cost for two_USD: **11.46.**

So, the variable “Average Cost for two_USD” had the maximum skewness.

A boxplot as shown in Fig. 6.1 was used to identify if there were outliers in the above-mentioned variables.

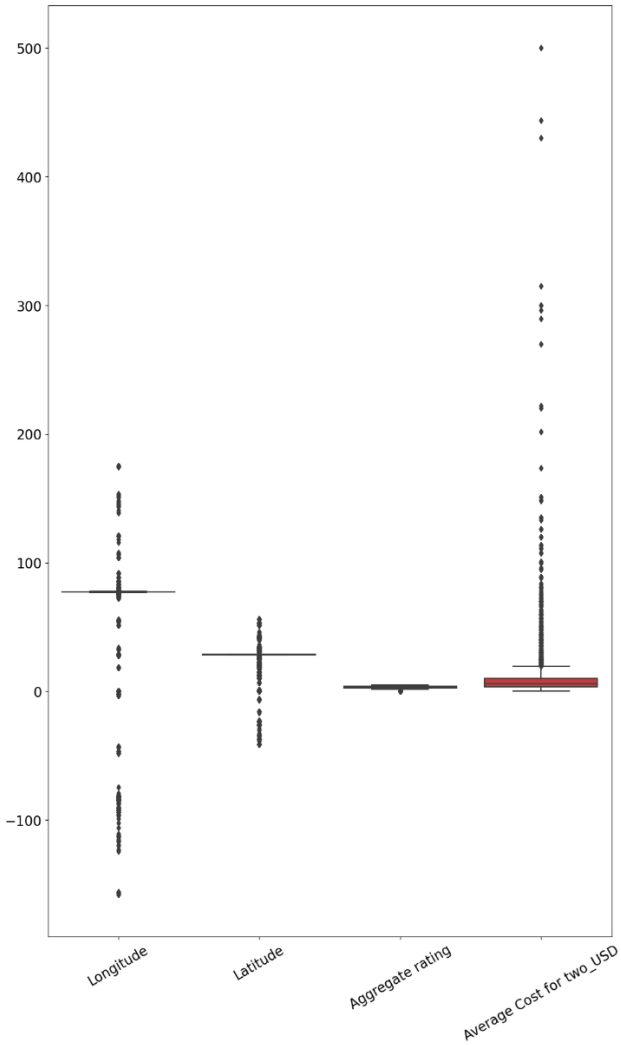

Fig. 6.1: Box plot of continuous variables

It was clearly visible in Fig. 6.1, that there were a significant number of outliers present.

Then, a count plot was plotted using the seaborn library for the categorical variables and also for the discrete type numerical variables 'Price range', and 'Votes'. The categorical variable “cuisine” had a very large number of values. This can be observed in Fig. 6.2.

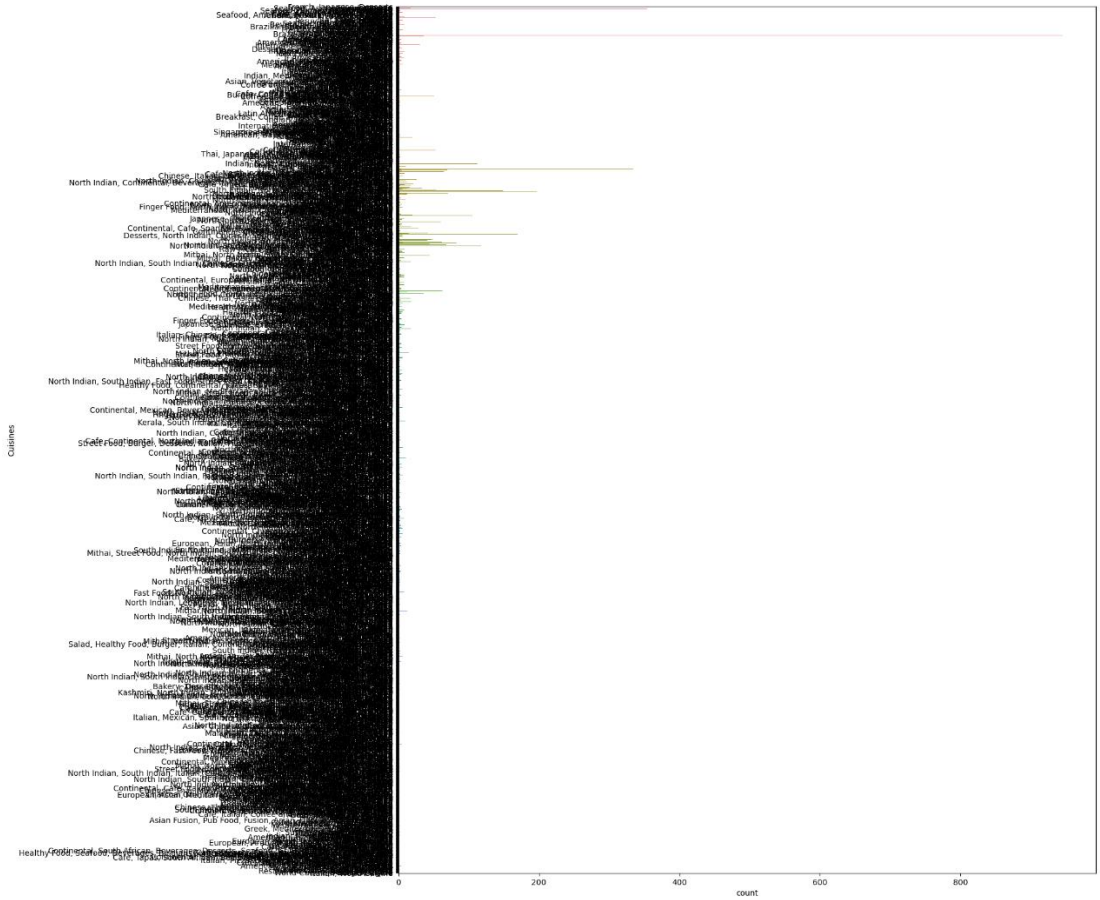

Fig. 6.2: Count plot- Cuisines

The discrete variable ‘Votes’ had also a large number of values. Therefore, the data values were grouped into 8 bins, which can be seen in Fig. 6.3. The plot shows that a large number of votes were in the bin 0 to 1366.75. There were very few votes in the remaining bins.

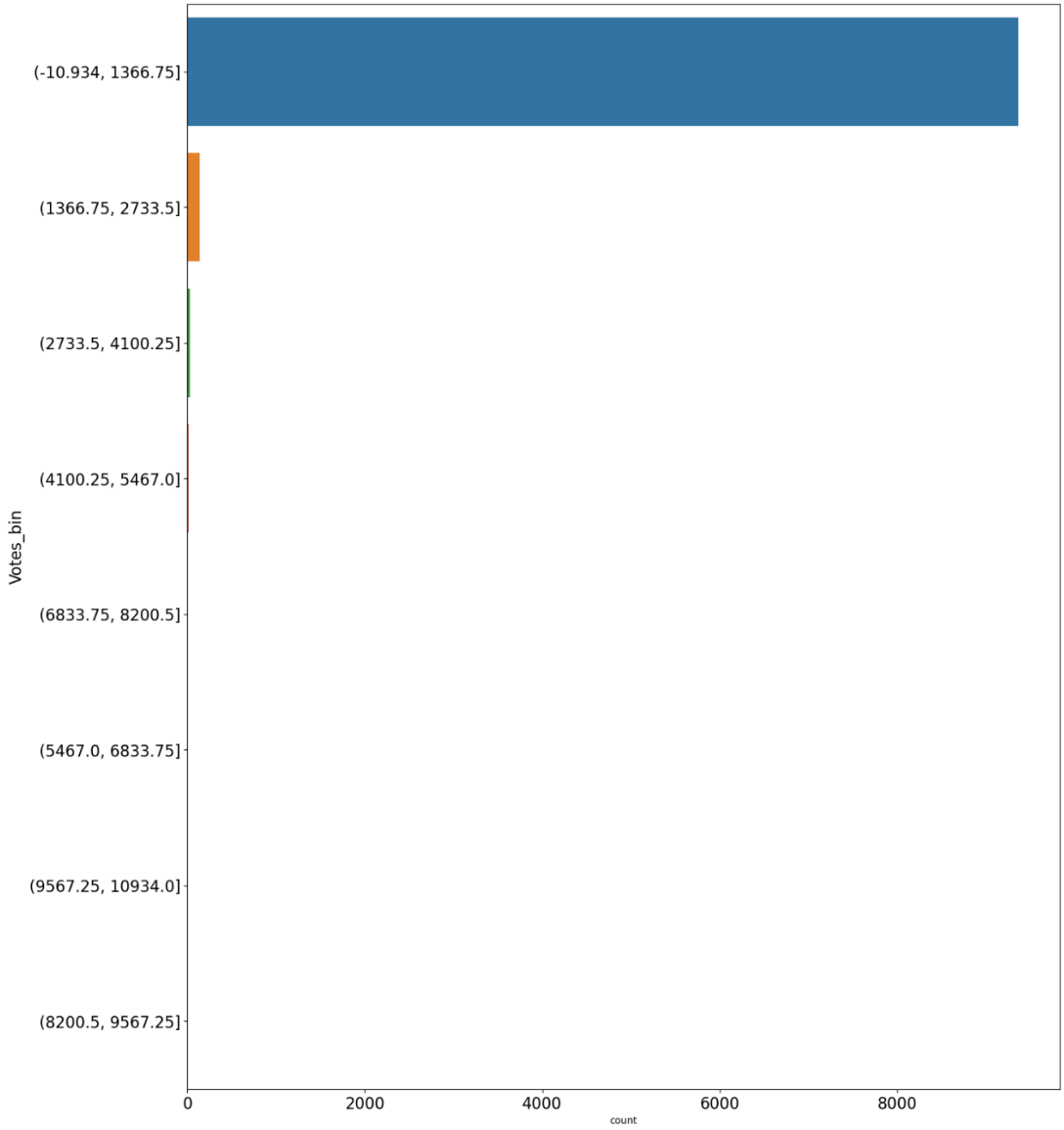

Fig. 6.3: Count plot- Votes

Step 7: To check for multicollinearity

A correlation analysis was performed among the numerical variables using heatmap plot of seaborn library. As can be seen in the heatmap in Fig. 6.4, there was no significant correlation found between variables.

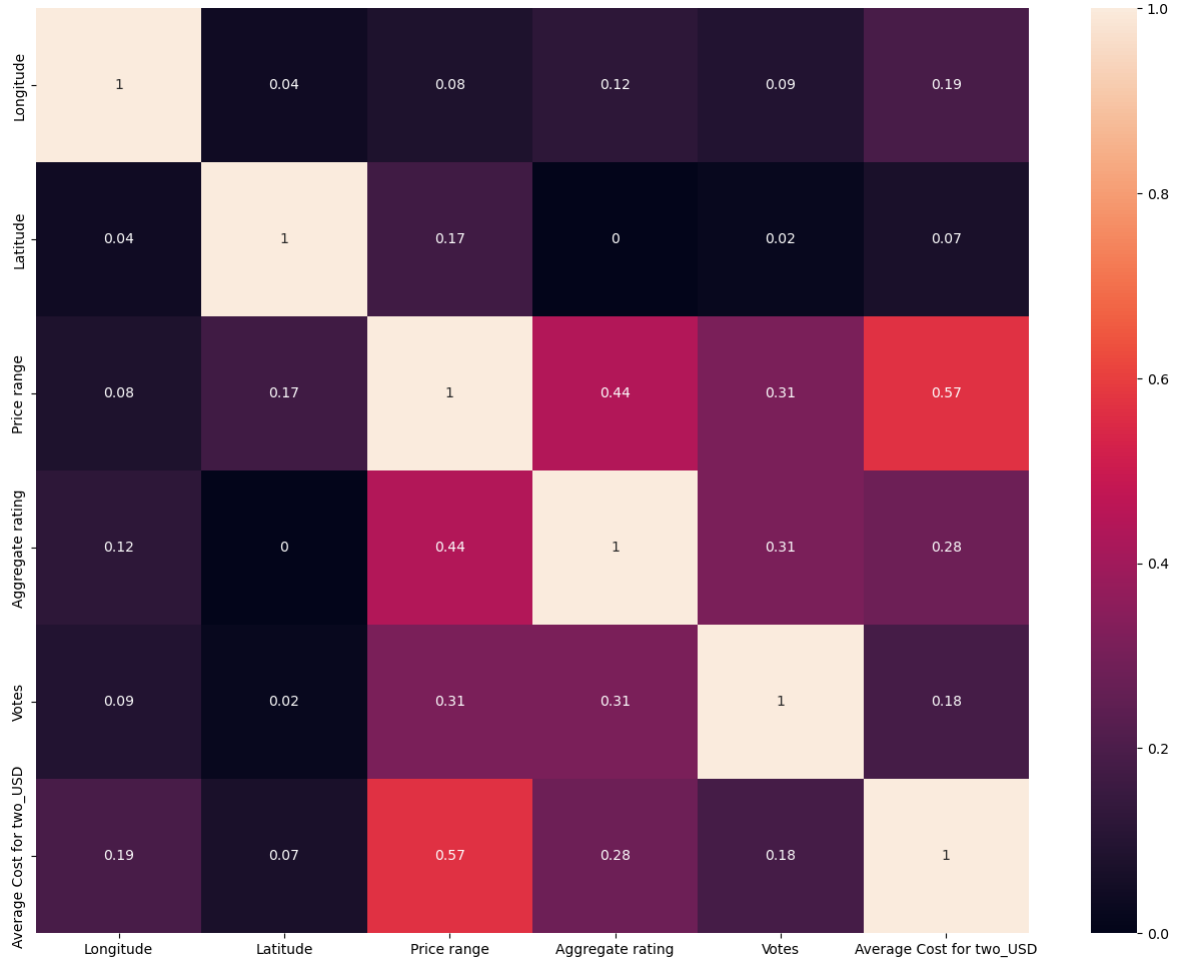

Fig. 6.4: Correlation heatmap

Step 8: Feature engineering

Once all the variables were visualized and their skewness, outliers and multicollinearity were assessed, the next task was to work on these variables to make them suitable for model building. The various steps performed to make the variables suitable for model building has been presented in the steps listed below:

(i) **To identify outlers:** The ouliers present in the variables which was shown in Fig. 6.1, had to be identified so that they can be removed. In order to identify these variables, z score was to be calculated. There was a threshold limit of 3 set to filter out the outliers from the remaining data in each of the variables with significant skewness. A threshold limit of 3 in z-score means a standard deviation of -3 to 3 from the mean of a normal distribution. Within this range of the normal distribution, there are 99.7 % of the data points. Any data point lying beyond this range was considered outlier. Z- score is calculated using the following formula:

$$z = \frac{x - \mu}{\sigma}$$

where, x = data point, μ = mean of the data, σ = standard deviation

There were no outliers found in the variable “Aggregate rating". The outliers found in other variables were dropped from the dataset. After dropping the outliers, number of rows reduced to 8820 and the skewness of the variables also reduced significantly particularly of variable "Average Cost for two_USD" as mentioned below:
Longitude: **3.37**, Latitude: **2.51**, Aggregate rating: **0.89** and Average Cost for two_USD: **2.91.**

However, there is still considerable skewness present. A tree-based model like Decision tree, Random forest, Gradient boosting and Adaboost algorithms is expected to perform well with these data.

(ii) **Categorical variables to numerical transformation:**

(a) Variable: 'Cuisines'

There were 1489 cuisines present in this variable. To perform numerical encoding of this variable, a clustering-based encoding was used. This encoding helped to create cluster of similar cuisines before encoding them. A K-means clustering algorithms was used. The cuisines were grouped into 250 clusters. The number of clusters was obtained using elbow method. In this method, the number of clusters was selected based on the observation where the rate of decrease of inertia became reasonably small. After clustering the cuisines, a new variable titled “Cuisines_Label” was created which was a column that mapped each of the cuisines by a number decided by the cluster to which they belonged. Following creation of this new variable, the “Cuisines” variable was dropped. So the number of variables effectively remained unchanged.

(b) 'Has Table booking', 'Has Online delivery', 'Is delivering now'

The remaining categorical variables 'Has Table booking', 'Has Online delivery', and ‘Is delivering now' were encoded using dummy encoding. Dummy encoding is basically one-hot encoding with drop first. Drop first helped to avoid duplicacy of information in terms of creation of additional variables, and hence multicollinearity was reduced.

Thus, all the variables were now in numerical form without any missing values.

(iii) **Scale of variables:**

After encoding all the variables, the scale of different variables was analysed. It was observed that there were variables like “Votes” which had a maximum value of 10934, whereas there were categorical variables encoded having maximum value of 1.
Therefore, these variables had to be scaled. All the variables were scaled using “Min max scaler” to convert them in the range 0 to 1.

3. Exploratory data analysis (EDA) concluding remarks:

Initially, there were total 21 variables and 9551 rows present. There were 9 missing values found in the variable **“Cuisines”**. As it was a categorical variable, therefore the missing values were imputed by the mode of the remaining data in the variable.

Out of 21variables, 13 were categorical and remaining were numerical. After the EDA, during which 11 variables were dropped, the number of variables reduced to 10. These

were dropped because either they were irrelevant like '**Restaurant ID', 'Restaurant Name'** or there were multiple columns presenting the same or similar information like **'City', 'Address', 'Locality', 'Locality Verbose'**. Some variables like "**Currency",** and **"Average Cost for two"** were replaced by a single variable **"Average Cost for two_USD"** after converting the cost present in various currencies to a more consistent form of all the currencies in USD.

There was significant skewness present in variables Longitude: **2.81**, Latitude: **3.08**, Aggregate rating: **0.95** and Average Cost for two_USD: **11.46.** So, these variables were analysed for the presence of outliers. After removing the outliers from these variables, their skewness reduced, particularly of variable **Average Cost for two_USD**. The reduced values of skewness were: Latitude: **2.51**, Aggregate rating: **0.89** and Average Cost for two_USD: **2.91.** After removing outlers, the number of rows reduced to 8820, which was a reduction of **7.6 %** of rows. This was considered not very significant to affect the performance of the model. There were no significant correlations observed among the variables.

The categorical variable "Cuisines" was converted to numerical variable using "**cluster-based encoding**". The reason to use this encoding was that, there were 1489 cuisines. So, to encode them, clusters of similar cuisines were created using K-means clustering. A total of 250 clusters were found to be the most optimum number based on the elbow method. Therefore, there was 250 clusters created for 1489 cuisines to represent them numerically.

Other categorical variables **'Has Table booking', 'Has Online delivery', 'Is delivering now'** were encoded using dummy encoding method.

Finally, after encoding all the variables, the scale of different variables was analysed. It was observed that there were variables like **"Votes"** which had a maximum value of 10934, whereas there were categorical variables encoded like **'Has Table booking'** having maximum value of 1. Therefore, these variables had to be scaled. All the variables were scaled using "Min max scaler" to convert them in the range 0 to 1.

Thus, the data was ready for model building.

4. Pre-processing Pipeline

Before training the model, the dataset was segregated as independent variables and target variable. There were two predictions to be made in this problem. These were:

(i) Average Cost for two:
In this part, the target variable was "Average Cost for two". This is a continuous variable. So, it was basically a regression problem. The other variables, which were independent, were segregated from the target variable "Average Cost for two".

(ii) Price range
In this part, the target variable was "Price range". This is a continuous variable. So, it was also a regression problem. The other variables, which were independent, were segregated from the target variable "Price range".

In both predictions, the whole dataset was then split into train and test data set using sklearn library's "train_test_split function" in the ratio 70:30. The random state was estimated based on maximum r2 score in "Random forest model".

5. Building Machine Learning Models

There were total 8 algorithms used to build the model and assess the best performing model. A 10-fold cross validation was also performed to improve the performance of the model. The regression algorithms used to build the model and their respective **r2 score** and **cross validation score** have been presented in the table below:

S. No	Regression algorithms		(i) Average Cost for two		(ii) Price range	
			r2 score	Cross validation score	r2 score	Cross validation score
1	Ensemble models	Random Forest	0.912	0.877	0.988	0.982
2		Gradient Boosting Regression	0.902	0.879	0.982	0.976
3		AdaBoost Regression	0.792	0.707	0.892	0.872
4		Bagging Regression	0.896	0.871	0.987	0.979
5	Linear models	Linear Regression	0.716	0.711	0.731	0.712
6		Ridge Regression	0.715	0.711	0.73	0.712
7		Lasso Regression	0.417	0.436	-0.001	-0.001
8	K Neighbors	K Neighbors Regression	0.835	0.811	0.828	0.811

Among all the regression models “Random Forest” model has the best r2 score. Therefore, this model was used for hyperparameter tuning to explore the possibilities to increase its accuracy further. However, hyperparameter tuning did not do well. The hypertuned model had poorer accuracy than the default “Random forest” model. Therefore, the default “Random forest model” was finally selected to be the final regression model and was saved using pickle library.

6. Concluding Remarks

As evident from the table, in both the prediction problems, r2 score for ensemble models is better than K Neighbors model, which in turn is better than linear models. Among the ensemble models, “Random forest model” has the highest r2 score and cross validation score. The reason for this is that there was some skewness present in the variables even after dropping the outliers. As we know, that ensemble models work well in the presence of skewness and outliers.