How I create my Zomato prediction Machine Learning model?



We all Know that Zomato is online food delivery company. Hance Zomato has large amount of customer data that’s why Zomato Dataset is popular dataset for machine learning to understand how ml project works in real world. In this blog I am sharing that how I create Zomato predication Machine learning model. I am sharing my all method, problems, and solutions.

First, we have to understand what client need and what is problem statement and how we attend the problem.

In Zomato dataset prediction problem statement is

1. Predict Average price for two
2. Predict Price Range

Here First problem is regression problem and second is classification problem.

For creating ML model first requirement is data that is the first step Data collection. We have three method of data collection.

1. Data given by client
2. Web scraping
3. Manually data collection from web site

Here data is given by company name Fliprobo. They give two csv file first is Zomato country code file and second is Zomato data file.

Now second step is data reading. We use pandas to read csv file and also marge both file on country code which is same feature in both data set.

After this we have to handle missing value which is third step. We have three methods for handling missing or null value

* Drop null value
* Impute by mean, median or mode
* Impute by KNNiputer

For our dataset we have missing value in Cuisines which is 0.094% of zomato

dataset that’s why we drop that null value.

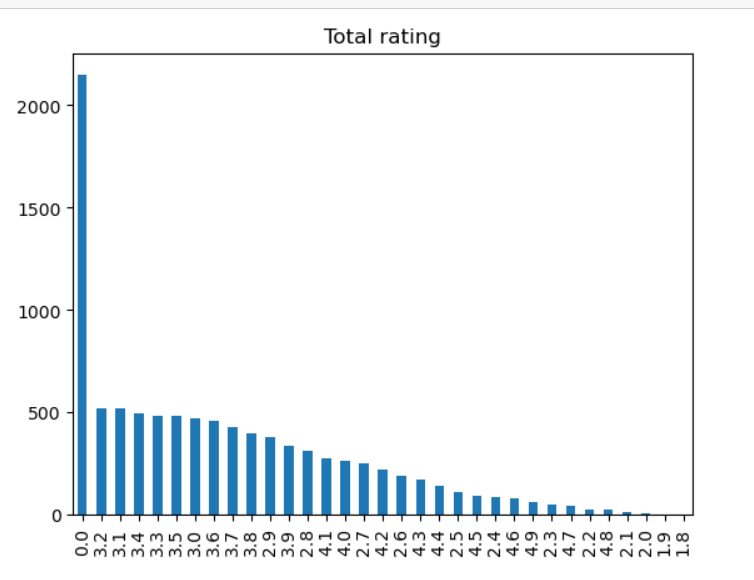
Now we remove that columns or feature that are not import for our prediction

* Address, Cuisines, Locality Verbose and Locality have more than 1000 unique values that is not use in prediction than we drop it.
* Switch to order menu has only one unique value that is no use in prediction we also drop it.
* Also, Restaurant ID and Restaurant Name not use in prediction than we drop it.

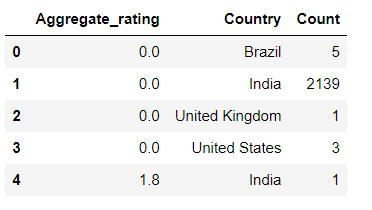
Now we drop duplicate value and remove white space from dataset.

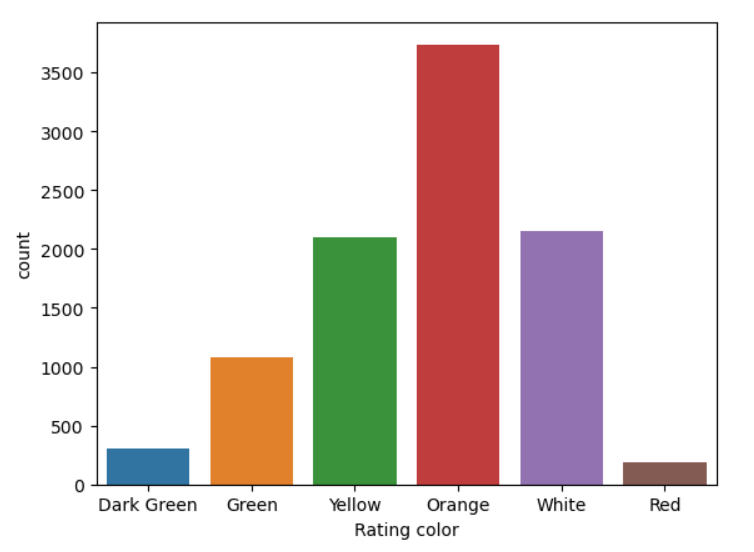
Third step is data analysis.

First, I analyze Aggregate rating, Rating color and I found that

* 2148 people not give any rating
* Dark Green have rating between 4.5-4.9 Rating
* Green has rating between 4.0-4.4 Rating
* Yellow has rating between 3.5-3.9 Rating
* Red has rating between 1.8-2.4 Rating
* White has rating between No Rating
* Orange has rating between 2.5-3.4 Rating

Then I plot Aggregate rating bar graph which is given here

Then I plot count plot to visualize rating color and observed that most of ratting is in orange color also white color which is o rating. Also analyze which country has 0 rating.



* Also, 90% order are from India which is obvious because Zomato is Indian company
* Also, 65% order from India has no online delivery

After this we have 9551 rows and 15 columns, 8 string columns and 7 numeric columns.

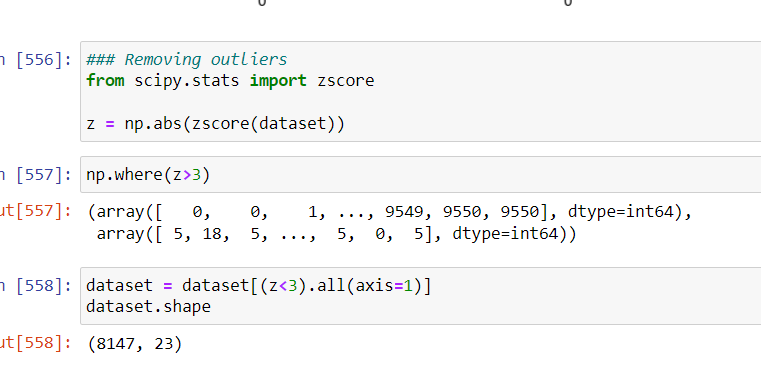
Now we convert categorical columns into numerical columns. for this we use LabelEncoder and get dummies method.

Currency, City, Price range and Country have more than 2 unique value that’s why we use LableEncoder and for other categorical columns we use get dummies method.

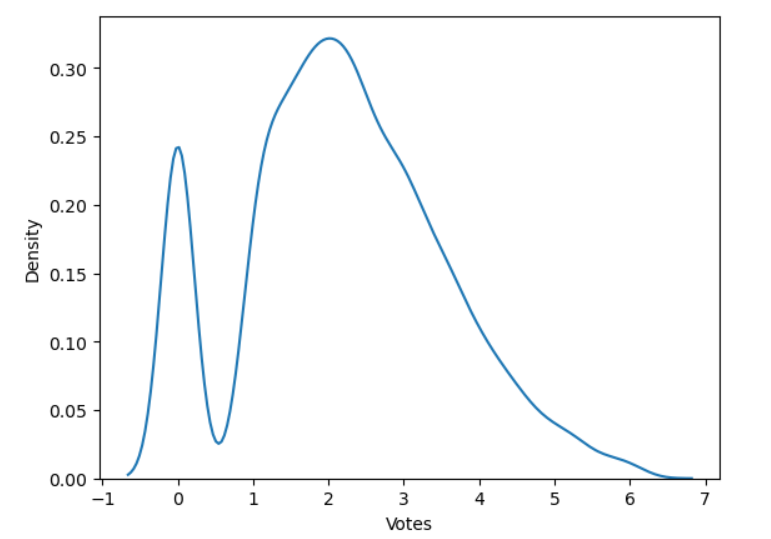
Now we perform Exploratory data analysis (EDA)

Steps of EDA

1. Remove outliers
2. Remove skewness

To remove outliers, we use zscore method which is given below

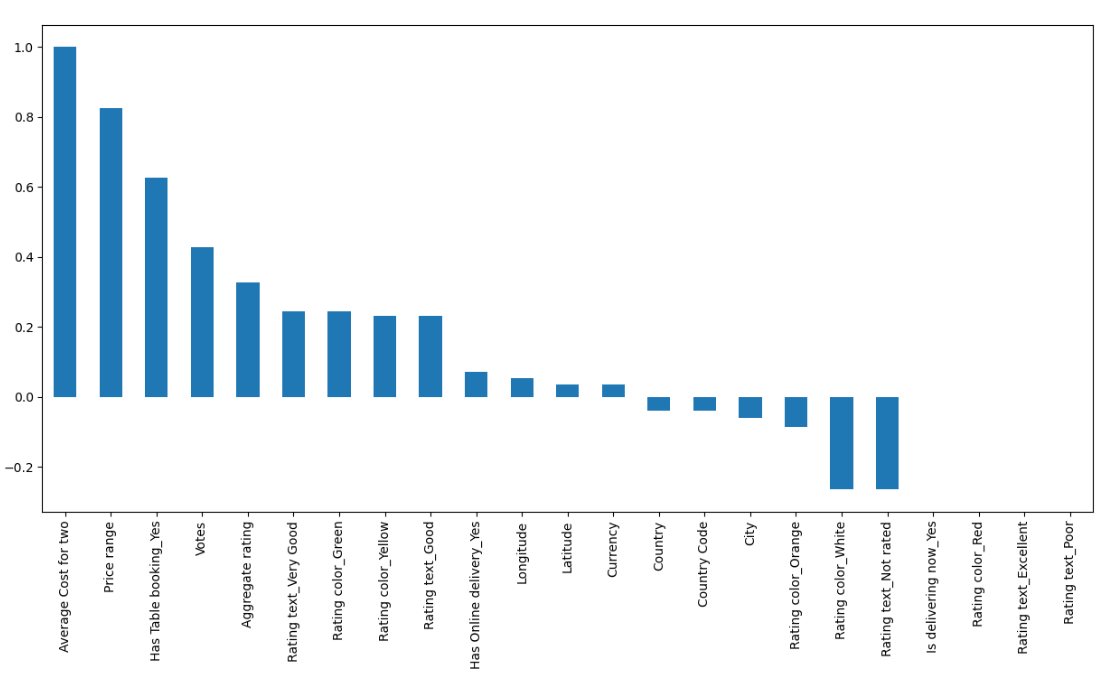
Now for skewnesswe plot kdeplot and we got votes have skewness and to remove skewness we use square root method. We use square root method twice because in one time this is not get bell shape. And last plot is given below.



After this Data is prepare for model building

Now splitting data into target variable and input variable In First task we perform regression that’s why first target variable is Average Cost for two and other column be an input variable.

Now we plot correlation plot to observed how feature correlated with target variable. The graph is given below.



Here we can see that Price range is highly correlated with target variable.

Now next task is to select feature which means remove unnecessary feature from input data.

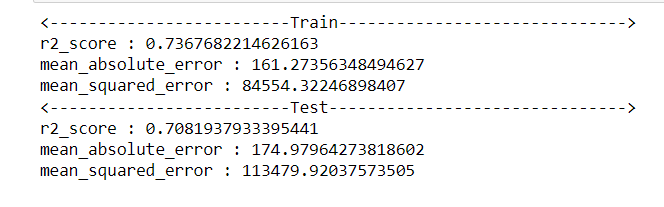
For selecting feature, we use Ridge regression because I used various method but in ridge regression, I got good accuracy and generalized model.

After using ridge regression method, we have now got 5 important feature which is Currency, Price range, Country, Has Table booking and Has Online delivery\_yes.

Now, we split data into tarin and test data with test ration with 30%.

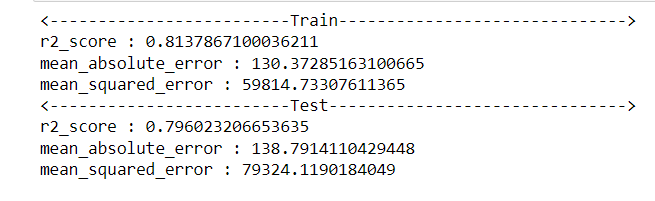
I used StandardScaler, minmax scale but I got good result with out using scaling method. Now we can make model.

First, Algorithm we use Linear regression and to calculate performance we use r2 score, mean absolute error and mean squared error.

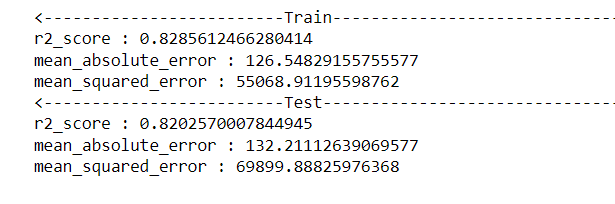
From Linear regression we got:

We can see that tarin accuracy is 73% and test accuracy is 70% that means data is genrelized and not overfitted.

Second, Algorithm we use KNeighbors Regressor:

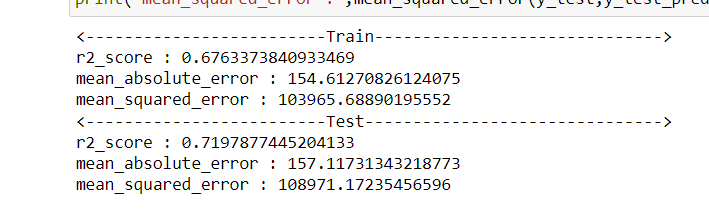


We can see that tarin accuracy is 81% and test accuracy is 79% that means data is genrelized and not overfitted.

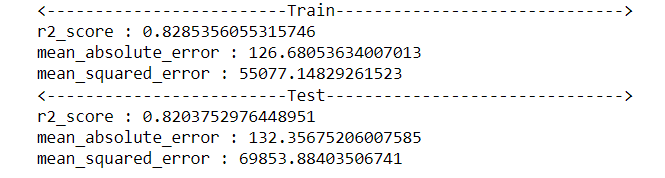
Third, Algorithm we use Gradient Boosting Regressor..

We can see that tarin accuracy is 83% and test accuracy is 82% that means data is genrelized and not overfitted.

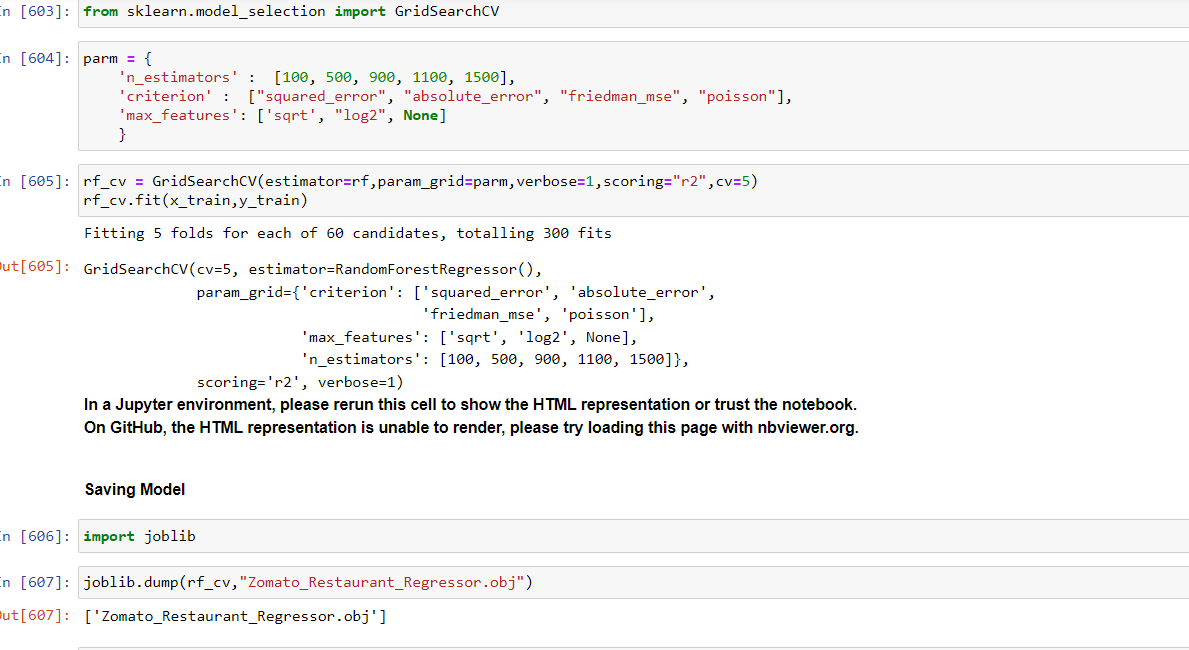
Forth, Algorithm we use Ada Boost Regressor.

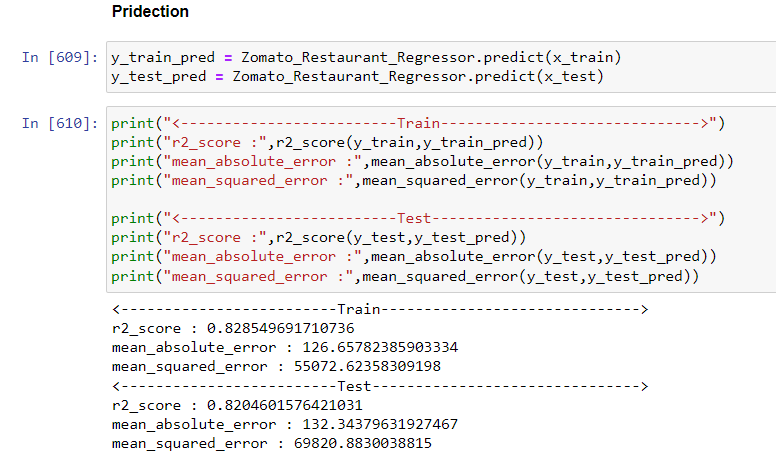


Fifth, Algorithm we use Random Forest Regressor



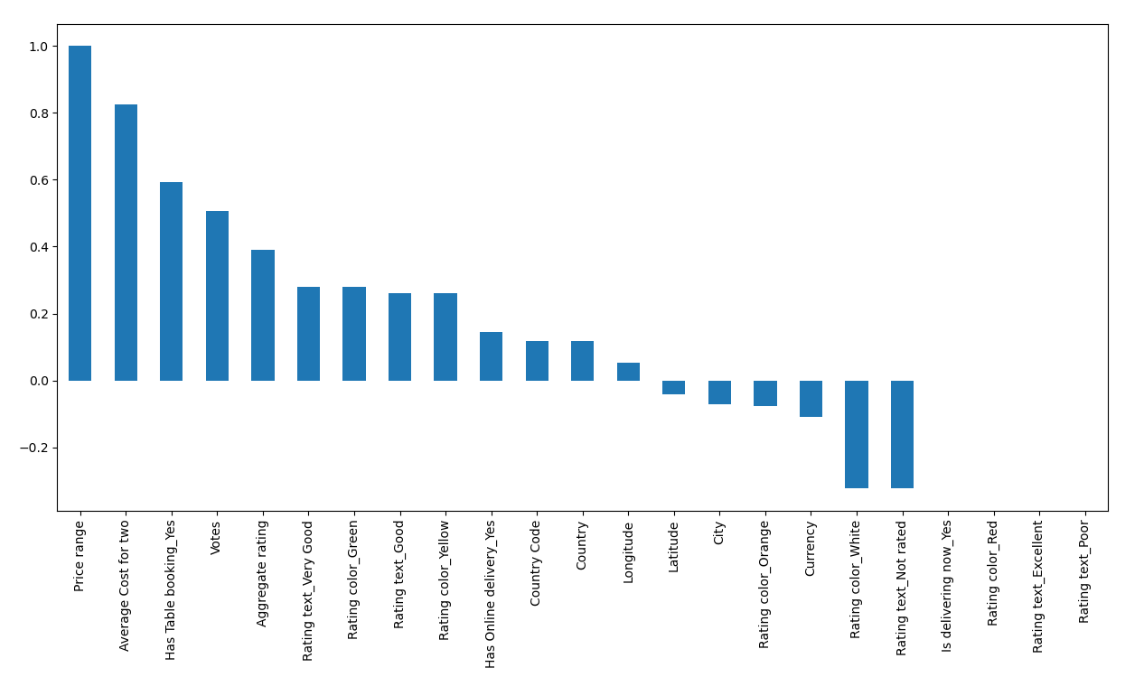
After using various model, I got that Random Forest Regressor is working good and we perform hyperparametric tuning on it.





We can see that last final model has 82% accuracy and generalized model.

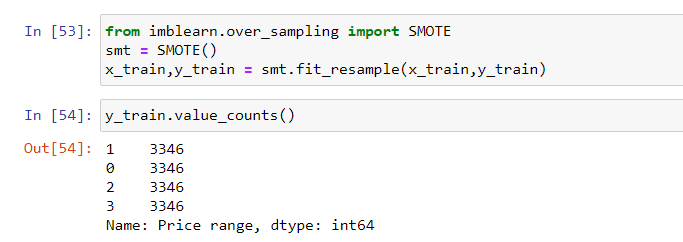
Now we perform classification task.

Again, splitting data into input variables and target variable. Now our target variable is Price range.Now we plot correlation with target variable.

Here we use feature selection as Lasso regression. And by this method I got 8 feature which is Country Code, City, Longitude, Latitude, Average Cost for two,

Votes, Has Table booking\_Yes and Has Online delivery\_Yes.

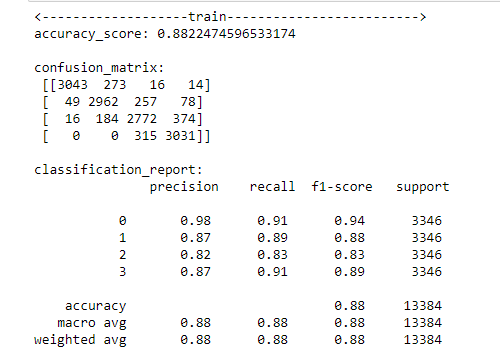
Now splitting data into tarin and test data. Also handling imbalance data and convert into balance data by SMOTE method.

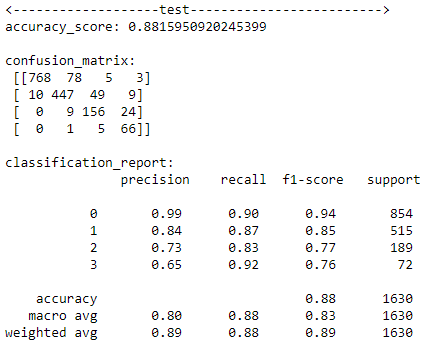


Here also I used StandardScaler, minmax\_scale but I got good result with out using scaling method. Now we can make model.

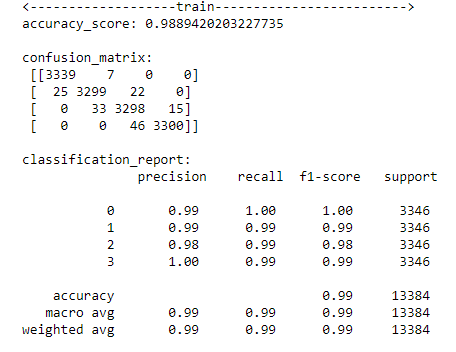
Now we billed model.

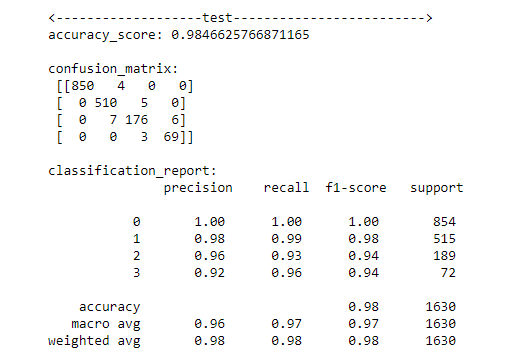
First model is logistic regression and I got accuracy as



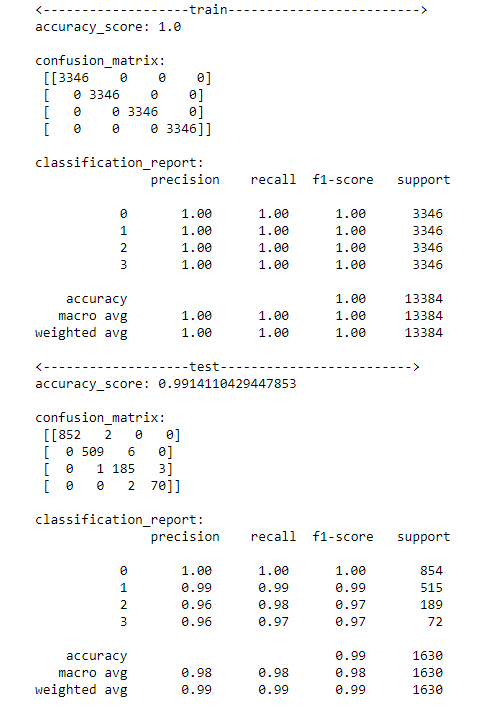
Here we can see that our model is giving 88% accuracy on test data and 88% accuracy on train data, Hance data is generalized.

Similarly, KNeighbors Classifier



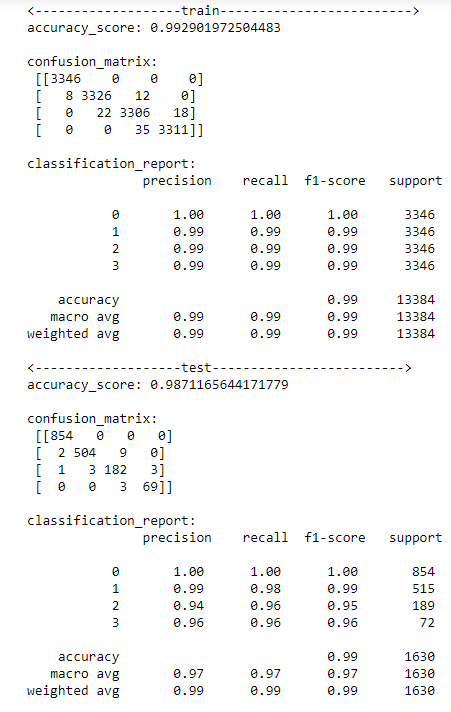


Here also we can see that our model is giving 98% accuracy on test data and 99% accuracy on train data, Hance data is generalized .

Random Forest classifier,

Here also we can see that our model is giving 99% accuracy on test data and 100% accuracy on train data, Hance data is generalized.

Gradient Boosting Classifier,



From above all we get Random Forest classifier is best model and we perform hyperparametric tunning on it.

After using hyperparametric tunning I got 99% accuracy and generalized model.

 Concluding remark:

* Zomato have to focused on online delivery.
* Also have to expand in more country.
* Zomato have to give more facility no online customer because Zomato have more offline delivery than online.
* Try to get more rating to understand how Zomato perform in state wise.