**Zomato restaurant project**

Author- Rebanta Dutta

1. **Problem Definition-**

Zomato is an Indian company that helps foodies and users to find reviews and in depth of the restaurants and cuisines. The company collects information of large number of restaurants and helps foodies to search and find for the best ones that suit their budget and taste. Zomato is an Indian company, which was founded in 2008 and presently it is one of the leading platforms for research, and discovery of food and thousands of restaurants across the world. It guides the users and the foodies to search for restaurants.

In this problem, a dataset containing 9551 data has been used for analysis and prediction of the average cost for two and the price range. I have used the dataset, to analyse the independent columns and the features with libraries like seaborn and matplotlib to gain useful insights from the data and build the model with higher level of accuracy. The average cost for two would determine the amount of money charged for the visit of two people in a particular restaurant. While the price range indicates the range in which the price of the food falls. It would become much more easier for the foodies to identify the budget and the cost of the food before visiting any particular cuisine or restaurant in any part of the world.

**Predictions**-

* To predict the average cost for two people
* To predict the price range of the food in the restaurant

1. **Data Analysis-**

The dataset contains 9551 rows and 21 columns

**Features-**

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. Currency: Currency of the country
12. Has Table booking: yes/no
13. Has Online delivery: yes/ no
14. Is delivering: yes/ no
15. Switch to order menu: yes/no
16. Aggregate Rating: Average rating out of 5
17. Rating color: depending upon the average rating color
18. Rating text: text on the basis of rating of rating
19. Votes: Number of ratings casted by people

**Label**-

* Average Cost for two: Cost for two people in different currencies
* Price range: range of price of food

For analysis of the data and building the machine learning models, I have imported the necessary libraries like numpy, pandas, seaborn, matplotlib and others. After that, I have loaded the dataset and checked for presence of any null values in the dataset. I have found that there are no null values and data contains a mixture of int64, object and float64 datatypes.

1. **EDA Concluding Remark-**

The machine learning model is to be built for prediction of price range and the average price of two people. On analysis of data, it has been found that price range of the restaurants depends on multiple factors like if the restaurant is delivering now, country code, table booking, currency it is accepting, online delivery, rating color and rating text. I have found that price range is high for restaurants that accepts Indian currency, no online delivery option, and rating color orange and white. Moreover, from the distribution plot it can be witnessed that skewness is present in majority of the columns. From the exploratory data analysis of the above data, it can be concluded that the restaurants having no online delivery option, accepting indian currency have higher price range.

1. **Pre-Processing Pipeline-**

In order to improve the accuracy of the model, data pre processing is one of the most important steps in machine learning.

Data pre-processing is a predominant step in machine learning to yield highly accurate and insightful results. It is generally a concept of transforming raw data into a clean set of data. Further, data pre-processing minimises noisy data, checks for missing values and other inconsistencies in data before the execution in order to ensure data is in an appropriate format.

Firstly, I have used the label encoding technique to transform categorical or object type into integer data for analysis. I have also used heatmap to check for correlation between the independent variables. From the heatmap, I have found that multicollinearity exists between the columns locality and locality verbose. So we have dropped the columns locality verbose because it has lesser correlation with the target variable. Further, from the descriptive statistics I have seen that outliers are present from the difference between the 75th quartile value and max value. The boxplot and the zscore method or technique has been used to identify the outliers and the total number of outliers in the dataset. After removal of the outliers, I have found that there is 11% loss of data.

Secondly, I have also found that there is imbalance in the dependent variable or the target variable. Imbalance in the data reduces accuracy of the model and so to balance the data, I have used the SMOTE technique. It is a technique that helps to generate synthetic samples for the minority class and it is a part of the Scikit learn.

After application of the SMOTE technique to balance the data, I have used the distribution plot and the skew() to check for the skewness in the data. It has been found that skewness is present in certain columns. I have used the log1p() transformation technique to remove skewness of the data, while keeping in mind that the target variable does not get transformed.

Lastly, I have used StandardScaler as a part of the pre processing technique for scaling of the data. This technique is used to normalise a large range of independent variables or high varying values of data.

1. **Building Machine Learning Models-**

The goal of building a machine learning is to train a model on historical data for prediction of some value. In this problem, we have to build 2 models for prediction of average cost of two and the price range. One of the problem is a classification problem and the another one is a regression problem. For prediction of:

* Average cost for two:

I have used different regression model after splitting of the data using the train\_test\_split function. The model has been built using KNeighbor regressor model, RandomForestRegressor, and Linear Regressor model. Using different metrics like mean squared error, r2 score and mean absolute error, I have found that random forest regressor model performs the best with higher level of accuracy score and has the minimum difference with the CV score. Further, in order to fit the model, I have used the hyper parameter tuning and found that the best parameters for the randomforestregressor model is 'criterion'= 'poisson', 'max\_depth'= 15, 'max\_features'= 'log2', and 'n\_estimators'= 200. After building the final model with these parameters, the model performs with an accuracy of 91%.

* Price range:

For prediction of the price range, I have used the classification model. I have used different classification models like logistic regression, decision tree classifier, random forest classifier, and KNeighbor classifier. Using metrics like classification report, confusion\_matrix and accuracy\_score, I have found that the random forest classifier model performs the best with an accuracy of 99%. I have selected this model and used hyperparameter tuning to best fit the model and to identify the most appropriate parameters.

1. **Concluding Remarks-**

It can be concluded that train\_test\_split function is an important part of the machine learning learning model. The train and test data needs to be split accurately in order to prevent the model from overfitting. Lastly, both the models performs with a good level of accuracy and the performance of the model has been viewed from the difference between the predicted output and the expected output.