ZOMATO RESTAURANT

Today’s world is of Digitalization. It has become easy now to get the analyzed information, which helps us to come to conclusion in bits of seconds. That’s the reason digitalization is booming as it is helping us with the concluded results.

Gone are the days of relying solely on word-of-mouth recommendations or random exploration to find a suitable place to dine. With Zomato, users can effortlessly access a wealth of information about restaurants, ranging from menus and pricing to reviews and ratings from fellow diners. This democratization of information empowers consumers to make informed decisions swiftly, based on the collective experiences and insights shared by others.

The rise of Zomato exemplifies how digitalization is not just a trend but a transformative force in the restaurant industry. It underscores the importance of leveraging technology to enhance customer experience, streamline operations, and drive business growth. Restaurants that embrace these digital platforms stand to benefit from increased visibility, improved customer engagement, and operational efficiencies.

The Objective of this project is to get an idea of following :

1. Average Cost For Two
2. Price Range

Building machine learning models involves a structured process of data preparation, algorithm selection, and model training to derive meaningful insights and predictions from data. Initially, raw data is collected and cleaned to ensure accuracy and consistency, followed by feature engineering to extract relevant patterns and characteristics. The next step involves selecting the appropriate machine learning algorithm based on the problem at hand, whether it's classification, regression, clustering, or other tasks. Model training commences by feeding the prepared data into the chosen algorithm, where parameters are adjusted iteratively to optimize performance. Evaluation of the model's efficacy involves testing against unseen data to gauge its predictive power and generalizability. Iterative refinement is often necessary to enhance model accuracy and address any overfitting or underfitting issues. Finally, deploying the model into production allows it to make real-time predictions, thereby completing the lifecycle of building a machine learning model.

Variables Provided in the Dataset:

Country Name- The name of the Country.

Restaurant Id: Unique id of every restaurant across various cities of the world

Restaurant Name: Name of the restaurant

Country Code: Country in which restaurant is located

City: City in which restaurant is located

Address: Address of the restaurant

Locality: Location in the city

Locality Verbose: Detailed description of the locality

Longitude: Longitude coordinate of the restaurant&#39;s location

Latitude: Latitude coordinate of the restaurant&#39;s location

Cuisines: Cuisines offered by the restaurant

Average Cost for two: Cost for two people in different currencies.

Currency: Currency of the country

Has Table booking: yes/no

Has Online delivery: yes/ no

Is delivering: yes/ no

Switch to order menu: yes/no

Price range: range of price of food

Aggregate Rating: Average rating out of 5

Rating color: depending upon the average rating color

Rating text: text on the basis of rating of rating

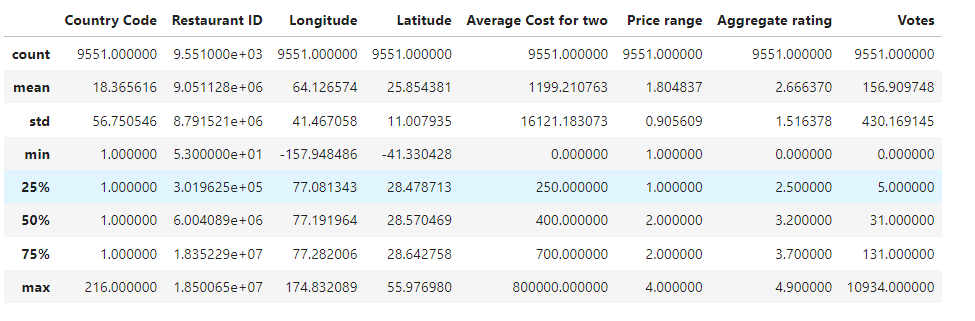
Votes: Number of ratings casted by people

Data Analysis:

Dataset, which encompassed a wide array of attributes crucial for understanding the restaurant landscape. This included restaurant names, locations, cuisines offered, average cost for two, user ratings, reviews, has online delivery,Price range,Currency and more. Each attribute played a pivotal role in shaping the analysis and subsequent modeling efforts. Each Variable plays an important role in framing the understanding of the data. Each attribute played a crucial role in shaping the analytical approach and driving informed decision-making processes. This wealth of data not only facilitated a nuanced understanding of the restaurant industry but also empowered effective modeling efforts aimed at various aspects of restaurant operations and customer preferences.

EDA Concluding Remarks:

Exploratory Data Analysis was pivotal in uncovering patterns and trends within the dataset. Key insights derived from EDA included:



std is very high for Average cost for 2, Longitude and Votes. Outliers would be there. Mean is greater than median except for Longitude, Latitude, Price range anf Aggregate rating. There is a large difference in 75% and max value in Country Code,Longitude,Latitude,Average Cost for two,Votes.

1. Geographical Insights: Distribution of restaurants across different cities and regions.
2. Cuisine Preferences: Popular cuisines based on user reviews and ratings.
3. Price Sensitivity: Relationship between average cost for two and restaurant ratings.
4. Relationship between average cost for two and Has Table booking.
5. Relationship between average cost for two and 'Has Online delivery'.
6. Relationship between average cost for two and 'Is delivering now'.
7. Relationship between average cost for two and Country.
8. Relationship between Votes and ‘Rating Text’.
9. Relationship between Votes and ‘'Has Table booking'
10. Relationship between Votes and ‘'Has Table booking'.
11. Relationship between Votes and ‘Is delivering now'.
12. Relationship between Votes and ‘Country'.
13. Relationship between Votes and ‘Average Cost for Two According to Rating text'.
14. Relationship between Votes and ‘Average Cost for Two According to Price Range'.
15. Relationship between Rating and ‘Average Cost for Two According to Rating text'.
16. Relationship between Rating and ‘Average Cost for Two According to Price Range'.
17. Presented Bar Plot for Average Cost for two according to the Currency and Price Range.
18. Presented Bar Plot for Aggregate rating according to Country and Rating.
19. Presented Scatter Plot for Restaurant Location Categorized by Price range.
20. Seasonal Trends: Variations in dining preferences across different times of the year.

We observe a moderate positive correlation between the 'Average Cost for Two' variable and the 'Votes,' 'Price Range,' and 'Country Code' variables. Conversely, the 'Price Range' variable shows a weak positive correlation with the 'Votes,' 'Average Cost for Two,' and 'Country Code' variables.

Pre-processing Pipeline:

1. Plotted Heatmap for Correlation Matrix.

Observed:

The heatmap displays both positive and negative correlations. The variable 'Average Cost for Two' exhibits a weak correlation with all other variables. The 'Price Range' variable shows a moderate positive correlation with 'Aggregate Rating' (+0.44), 'Votes' (+0.31), and 'Country Code' (+0.24). Additionally, 'Country Code' demonstrates a moderate positive correlation with 'Longitude'.

'Restaurant Name' Column: Although 'Restaurant Name' exhibits high cardinality, it potentially influences both target variables. For instance, it might indicate whether the restaurant is associated with a well-known brand. Frequency-Based Encoding will be applied to this column.

Locality' and 'Locality Verbose' Columns: From the previous analysis, it is evident that both features provide very similar information. Therefore, 'Locality Verbose' will be dropped, and Frequency-Based Encoding will be applied to 'Locality' to handle its high cardinality.

Cuisines' Variable: This variable has a unique characteristic where each row can contain multiple cuisine categories. To address this, binary columns will be created for each cuisine category, and the categories will be aggregated to reduce the number of unique values. This approach aims to make the variable more manageable.

22-Encoded the Unique Values Of each Category to get the result and the model is able to understand and extract valuable information.

23-Encoded all the Categorical values and dropped the columns which are not needed in the predictions.

1. Treated the outliers as they can negatively affect the statistical analysis and the training process of a machine learning algorithm resulting in lower accuracy.

Use Cube Root transform method to deal with the skewness.

Visualized the distribution of the columns after removing skewness.

Observed:

All the distributions got right skewed.

Separated the independent and target variables into x and y.

Separated both the Variables “Average Cost for Two” and the “price Range” into x and y.

Scaled the Data for both Average Cost for Two and Price Rnage. Scaling the Data increase efficiency, reduce manual errors.

After Scaling Fined the VIF Values(varience inflation factor) in each scaled column.

* VIF values greater than 10 suggest a high correlation among these features. To address this issue, we will begin by dropping one of the columns. If the multicollinearity persists, we will then consider removing the column with the highest VIF.

We drop the 'Malay' and 'Indian' columns to solve the multicollinearity problem since the first one was the second with higher VIF value, then the feature remaining highest VIF value.

Counted the Frequency of the target Price Range. But the dataset was imbalanced.

The dataset was imbalance. So, we applied SMOTE for oversampling the data.

Started Finding the Best Random Sate:

Splitting the Dataset into Training and Testing:

The features selected in the preceding step were approved to develop classification models. Initially the dataset was randomized to obtain an arbitrary permutated sample. It was followed by splitting of the dataset into training (70% of the dataset) and test (30%) sets.

Building Models:

CLASSIFICATION ALGORITHMS

Linear Regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable[5]. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0)

Random Forest:

Random forests or random decision forests are an ensemble learning technique for classification [6].Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, “Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.” Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

Decision Tree:

Decision Tree calculation has a place with the supervised learning algorithms. Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

Used other Methods as well such as:

KNN

GradientBoostingRegressor

Lasso Model

Ridge Model

To calculate:

Mean squared error

Root Mean squared error

Cross-validation results (R2

Average R2

R2\_Score(test

Calculated the difference for all the models between R2 Score and Cross Validation Score.

In order to apply Hyper Parameter tuning we are going to select Random Forest Regressor as our final model.

We got the Maximum R2 Score – 98.143%

Calculated Accuracy using Different Models:

RandomForestClassifier

ExtraTreesClassifier

LogisticRegression

SVC

GradientBoostingClassifier

AdaBoostClassifier

BaggingClassifier

Calculated the Training Accuracy and Model Accuracy Score for all the models.

Evaluate cross-validation for each model.

The best Model we got as RandomForestClassifier .

The Accuracy Score got as= 97.784 %

Conclusion:

We have two target variables: 'Average Cost for Two', which is continuous, prompting the development of a regression model, and 'Price Range,' a categorical variable with four possible values, requiring the creation of a classification model."

The chosen regression model is the RandomForestRegressor, achieving a R2 score of 98.143 % for predicting 'Average Cost for Two'.

For the classification task, the RandomForestClassifier achieved an accuracy score of 97.784 % in predicting 'Price Range'.