**Article on Zomato Restaurant Project**

**Problem Definition:**

**There were two kind of analysis that was required from the dataset provided**

1. **Average cost for two**
2. **Price range**

**The dataset provided included information like Restaurant Name, City, Address, Locality, Cuisines, Currency accepted, Number of times the table was booked which indicates the count of times people have visited the restaurant, Number of times Online Delivery has happened which indicates how many times online orders are placed, also a column to indicate whether Zomato delivers from restaurant and Aggregate rating from the customers and other columns as well.**

**Data Analysis:**

**PFB the steps carried out to analyze and clean the data**

**1) Data Exploration:**

**a) In this step we explored that there are two datasets namely Zomato.csv and Country.xlsx.**

**b) Both the datasets were added as DataFrame and joint with the help of merge command.**

**c) Identified that the dataset combined had 9551 rows and 2 columns**

**d) Data had a mixture of text data and numerical data**

**2) Data Cleaning**

**a) While doing a check on the missing values for the Feature columns identified that the Cuisines column had 9 missing values**

**b) Got the missing values updated by using the fillna and mode function**

**3) Feature Engineering**

**a) Distributed columns as per numerical and categorical**

**b) Went through a statistical summary of the features to understand if the skewness is towards right or left**

**c) Identify columns which has outliers so that the normal distribution can be created**

**d) Created a regression plot graph to identify relationship between the features and label.**

**e) Made Scatter plot and pairplot to have a stronger case in identifying which features have a higher relationship with the label**

**f) Used the boxplot graphs to identify which features have outliers and needs to be fixed**

**g) Then implemented the correlation feature to identify which feature has higher skewness**

**h) Found skewness in Country code, Votes and Price Range and used the cuberoot method to get it normally distributed.**

**i) Used the Ordinal Encoder function to convert the text columns into numerical values.**

**j) Used the heatmap function to check the correlation between features and labels, and features to features.**

**k) Identified that Votes has a high correlation with Price range and Aggregate rating**

**l) As none of the columns had a high correlation with the label kept the columns and did not delete any of them**

**EDA Concluding remarks:**

**a) Missing values have been updated, Column which will not impact the outcome has been removed. Did a statistical summary of numerical columns and did the first run on identifying columns which has right and left skewness. Identified relationships between features and label by using regression plot and removed the skewness with the help of boxplot and cuberoot method. Text columns converted to numerical values by using Ordinal encoders**

**Pre-Processing Pipeline:**

**a) Separating features and label by dropping the label column.**

**b) Used the Standard Scalar function to normalize the feature columns**

**c) Used the Variance Inflation factor to check how much the features are related to other features and is that resulting in biasness**

**d) Locality Verbose, Locality, Country and Country Code have the highest relationship with each other**

**e) Removed the Locality column as it was causing the highest Multicollinearity**

**f) As there was still high Multicollinearity between Country Code and Country remove the Country code column**

**g) As the Multicollinearity challenge is solved split the data into training and testing datasets in order to predict Average price for two.**

**Building Machine Learning Models for scenario 1:**

**a) Imported Machine Learning models like Linear0Regression, RandomForest Regressor, DecisionTree Regressor, KNNNeighbors Regressor, Lasso and Ridge.**

**b) Tested the models wit Train and Test data and identified that the best results on R2 score on training data came for Random Forest Regressor and DecisionTree Regressor.**

**c) Got the same result after trying cross validation score**

**d) As the accuracy score was too high needed the model to go through Hyperparameter tunning to check if the accuracy can be increased and done without any biasness**

**e) Got the final results and compared the predicted data with actual data.**

**Building Machine Learning Models for scenario 2:**

**a) In the second scenario the label column considered is Price Rating**

**b) Split the data into training and testing dataset**

**c) Ran the Classification models RandomForest Classification, ExtraTrees Classifier, Logistic Regression, SVC, GradientBoosting Classifier, AdaBoostClassifier, BaggingClassifier**

**d) Got the best accuracy on RandomForest Classification and ExtraTrees Classifier**

**e) Ran the Cross Validation Score across the parameters and got the same results**

**f) Applied Hyperparameter tuning on RandomForest Classifier and got 100% accuracy on the predicted value as compared to actual test values.**

**Concluding Remarks on Zomato project:**

**a) In conclusion, the application of Machine Learning models on the Zomato restaurant dataset has yielded insightful results, with the RandomForest Regressor and RandomForest Classifier emerging as the top performers. These models have demonstrated their robustness and predictive power, effectively capturing the complexities within the data. The success of these algorithms underscores the potential of Machine Learning in transforming data into actionable insights, paving the way for more informed decision-making in the culinary business landscape. As we continue to refine these models, we can anticipate even more sophisticated analyses that will further enhance the dining experience and operational efficiency for restaurateurs and customers alike. This endeavour not only reflects the strides made in data science but also the endless possibilities that lie at the intersection of technology and gastronomy.**