FIND THE TOP-RATED RESTAURANT USING CLASSIFICATION TECHNIQUES

SUBMITTED TO: SAVITA SEHARAWAT

SUBMITTED BY: GURCHARAN KAUR

STUDENT ID: 0775710

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to our professor Mrs. SAVITA SEHARAWAT for her valuable guidance and support in the completion of my capstone project. Your inspiration and suggestions were so valuable in carrying out this project. I think I will not be able to complete this project without your knowledge and cooperation within limited time frame. I wish to express my deep gratefulness for your consideration and supervision.

OVERVIEW

Bangalore is the capital and biggest city of the Indian territory of Karnataka. With a populace of more than 15 million, Bangalore is the third-biggest city in India and 27th biggest city in the world. Bangalore has an interesting food culture. Eateries from everywhere the world can be found here in Bengaluru, with different sorts of foods. Overall, it is possible that Bangalore is the best spot for foodies.

The food business is always at a rise in Bangalore, with 12,000 or more eateries presently active in the city, the number is yet expanding. The developing number of cafés and dishes in Bangalore draws in me to assess the information to get a few experiences, some interesting facts, and figures. So, there is a big challenge for a new business to find out location which would be always crowded and in demand.

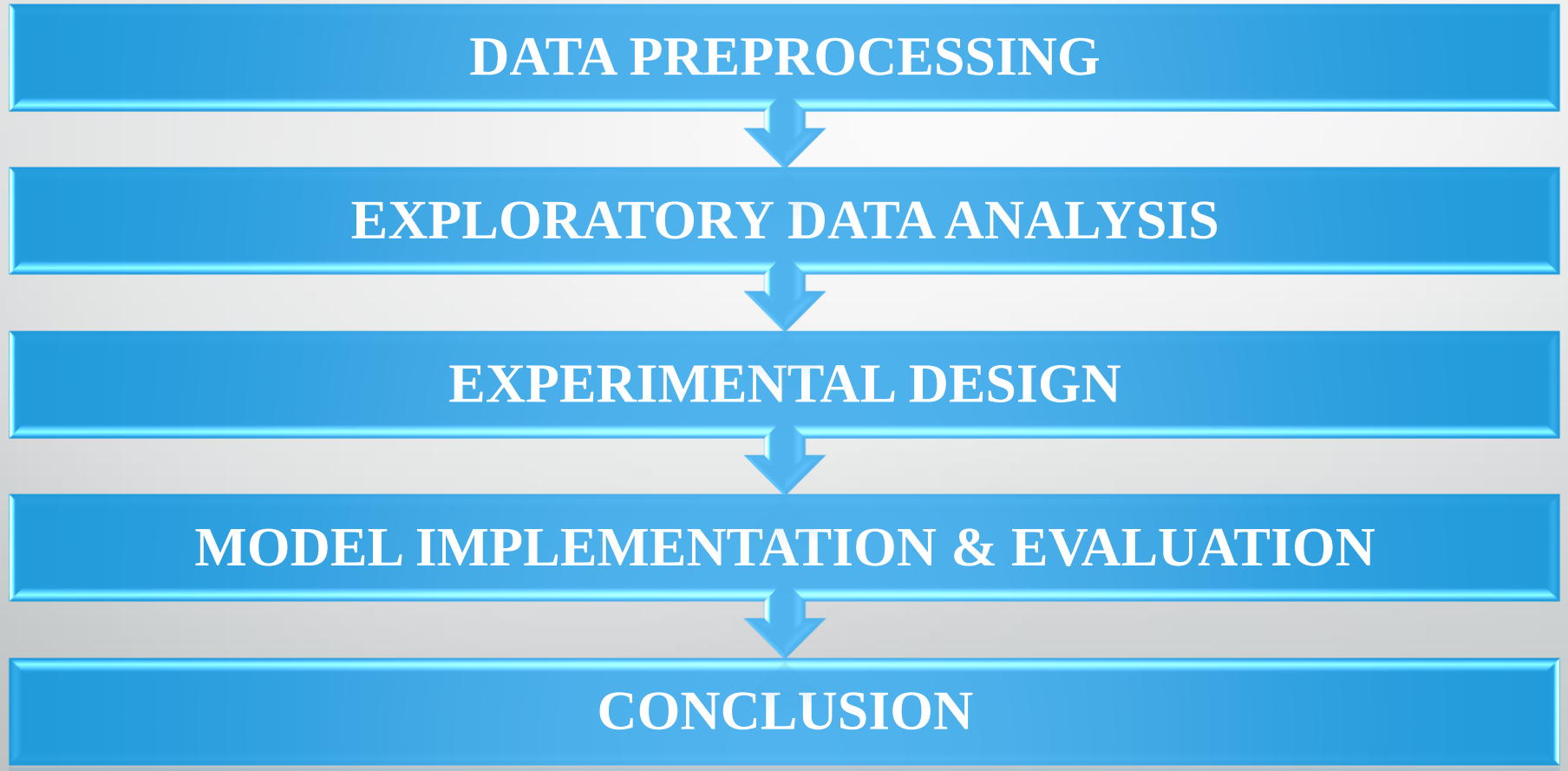
INTRODUCTION

My dataset is based on the restaurants in Bengaluru city. This dataset contains 51717 records and 17 attributes. This dataset basically contains the information regarding the restaurants dine in, takeout and online order options, reviews, and type of restaurants like casual dining, pubs, bars and café, type of cuisine and all.

The main goal of this project to find best top-rated restaurants in Bengaluru city and does the cost of food affect the rating of restaurants. I will try to figure out which cuisines are famous, about dine-in and takeout option in Bengaluru. These days, mostly people do not have time to cook food at home, so they are preferring restaurant food. With such an overwhelming demand of restaurants, it has become important to study the demography of a location. As, its is a big challenge for a new business to find out location which would be always crowded and in demand.

This dataset also contains the reviews for each of the restaurant and cost, from which I will find out the overall rating for the area that would be helpful to find out the top-rated location for setting up a new business. Moreover, I can also find which cuisine is popular in the area and many more.

METHODOLOGY



DATA DETAIL

- This dataset contains 51717 records and 17 attributes from which we have 16 object attributes and 1 numeric attribute.
- In this dataset, we have all attributes of object datatype; we should assign the fitting datatype to the attributes. After assigning appropriate data type, now we are having 2 objects, 5 categorical attributes, and 5 numeric datatypes
- Further, we dropped 5 attributes, since that attributes were not giving useful information. The dropped attributes are URL, Phone, Address, Dish liked and Menu item.

DATA PREPROCESSING

Attribute	Missing Values
Rate	7757
Location	21
Restaurant Type	227
Cuisines	45
Cost	345

In this, we dealt with each attribute in our dataset. To begin with, we converted the attribute to suitable data type. From that point forward, we really look at five number summary. Then, we checked the levels of the all-categorical attributes. We additionally look at the missing values in our dataset. Further, I filled the missing values.

Rate, Cost: In these attributes, I replaced missing value with mean.

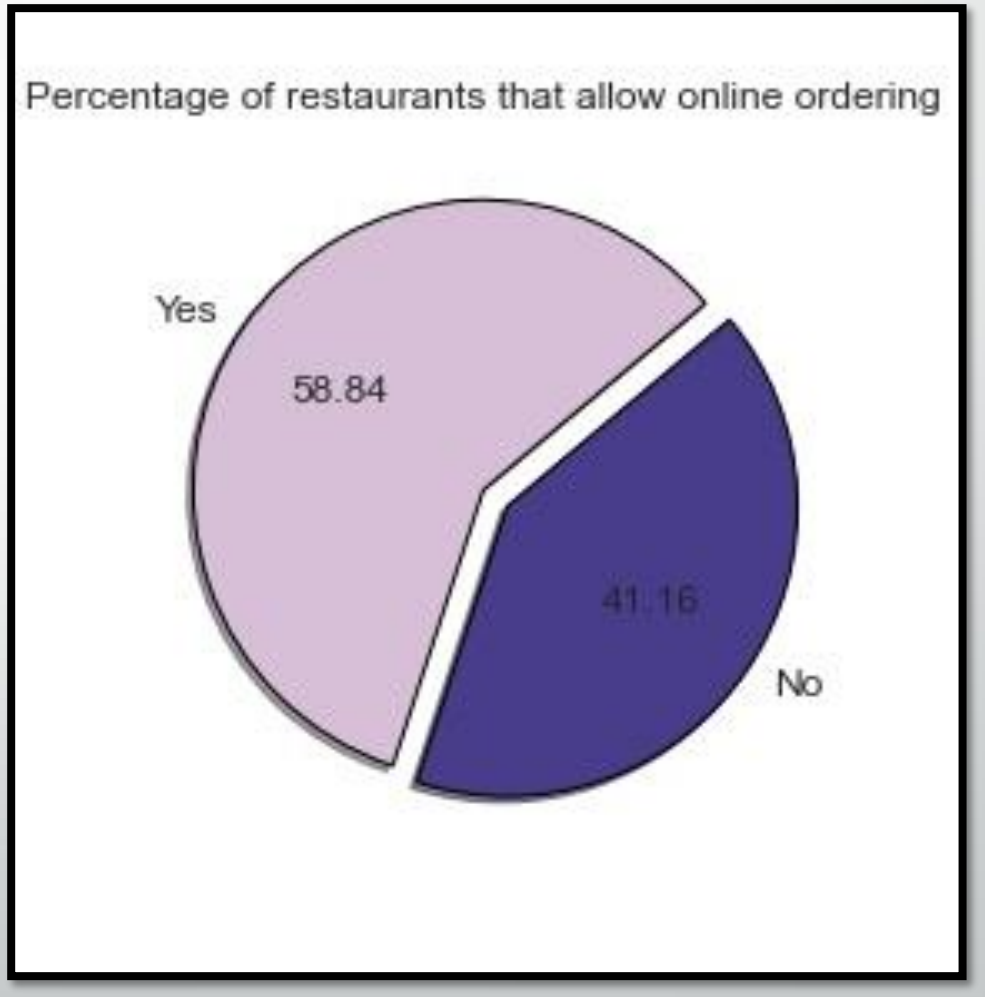
Location, Restaurant type, Cuisines: For these attributes, I fill the missing value with mode.

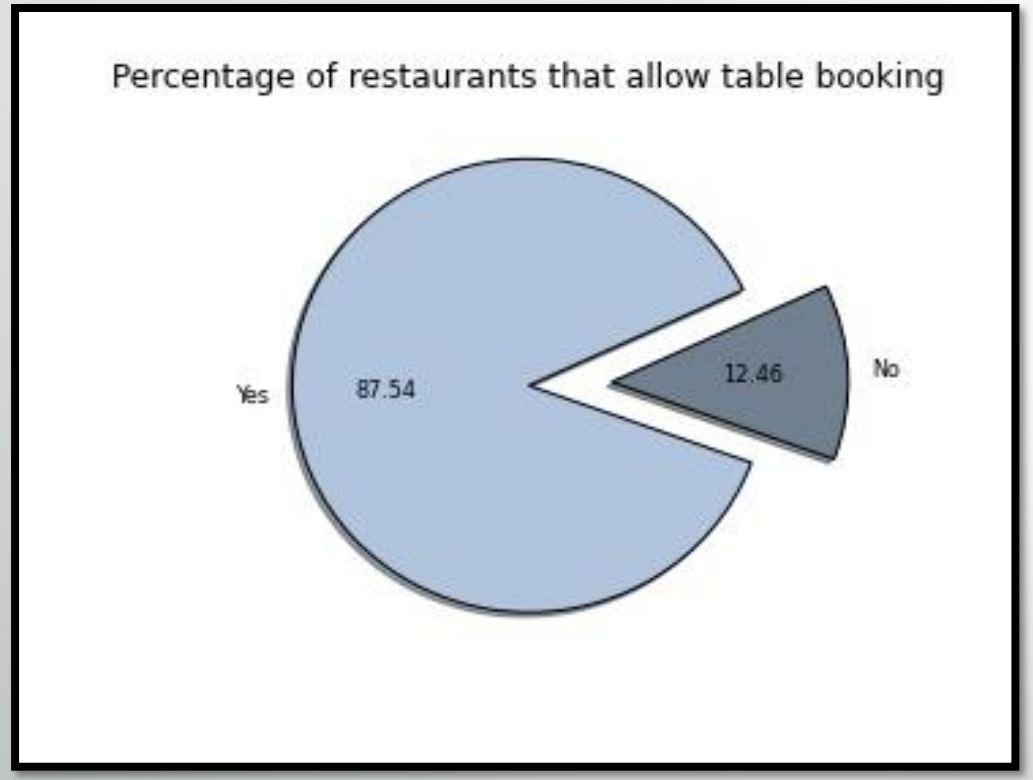
Then I performed EDA, One-hot encoding, Splitting train test, Sampling strategy, Classification Models.

EXPLORATORY DATA ANALYSIS

Q1: HOW MANY RESTAURANTS HAVE ONLINE ORDER OPTIONS?

This pie charts shows the information about the restaurants having option for online order or not. The greater part (58.84%) of the restaurants in the Bangalore city having choice for online order. However, 41.16% restaurants are not giving web-based request choice.



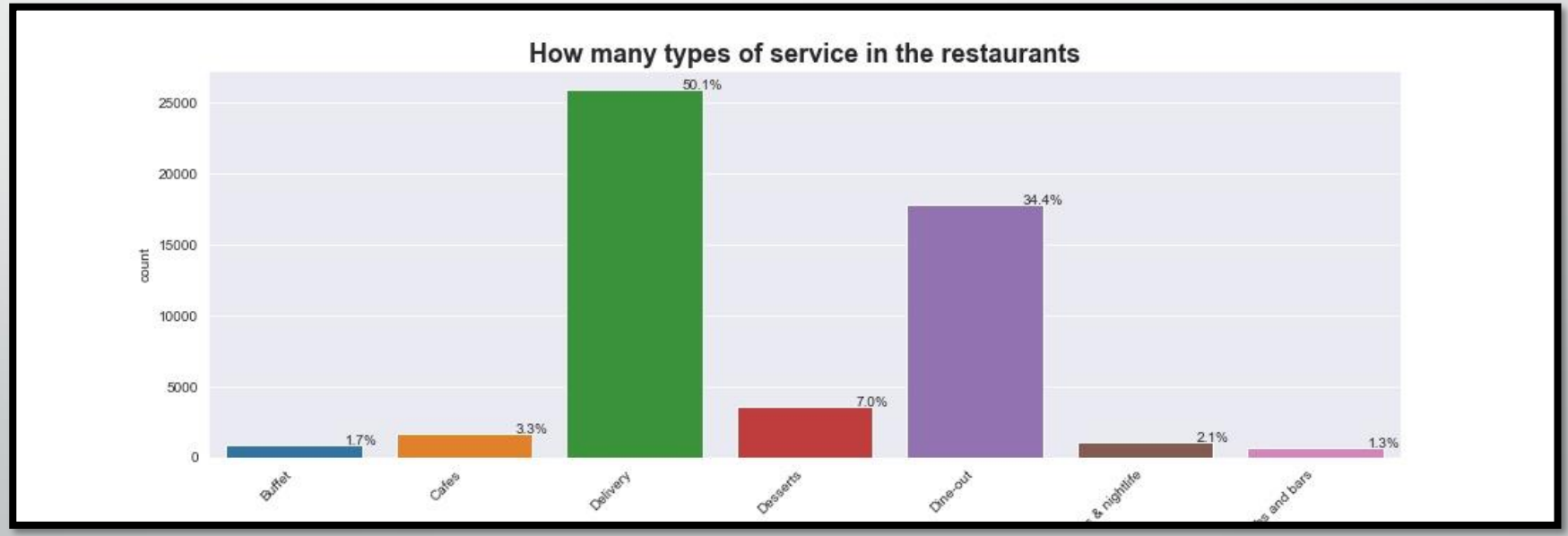


Q2: HOW MANY RESTAURANTS HAVE TABLE-BOOKING OPTION?

The pie chart demonstrates that how much restaurants are having option of pre booking of the table or not. The greater part (87.54%) of the eateries giving table booking choice and just 12.46% of the eateries are not having table-booking choice.

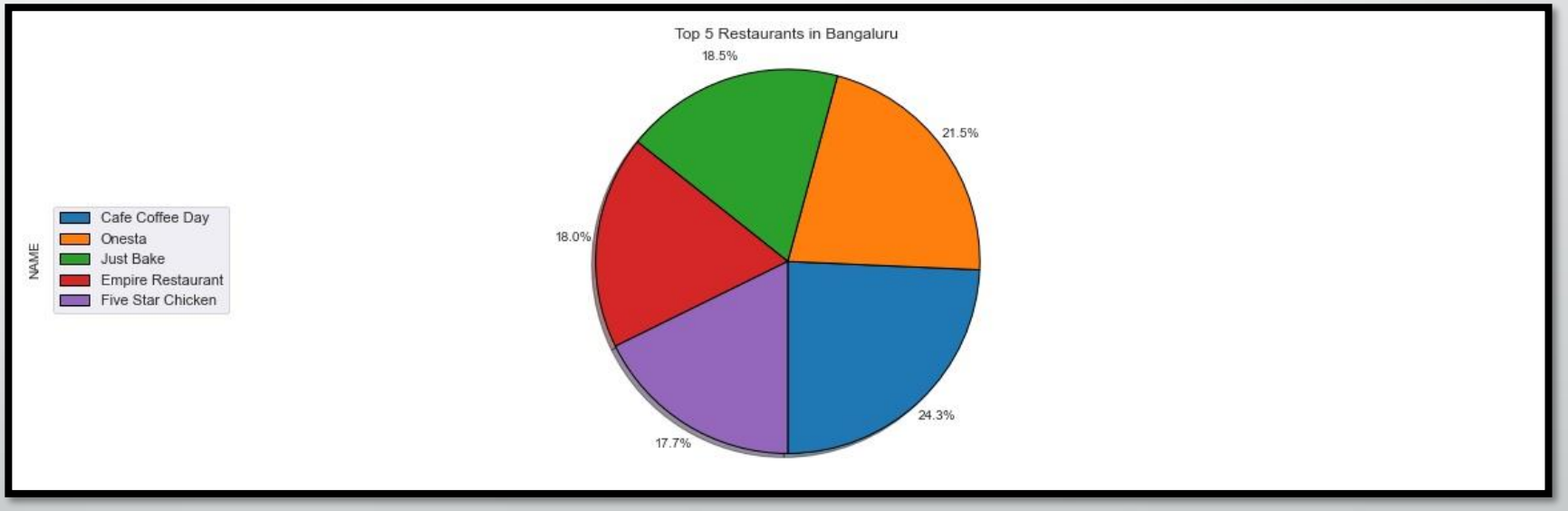
Q3: HOW MANY TYPES OF SERVICE ARE PRESENT IN THE RESTAURANTS?

There are all out 7 kind of services in the eateries. Top-notch services is the delivery option. From these services, clients like to Dine-out and do delivery service, as contrast with different sorts like buffet, cafes and bars and so on.



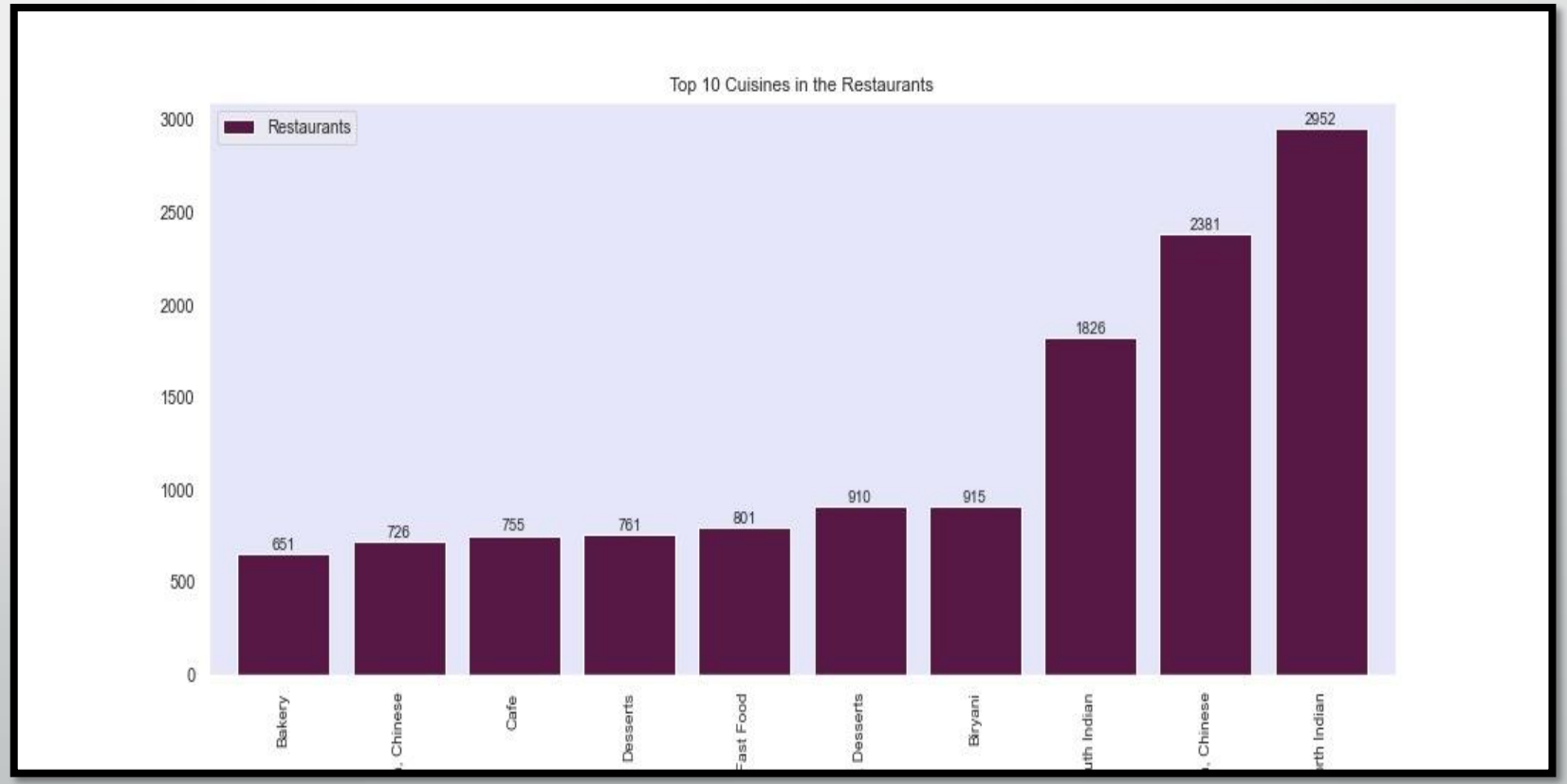
Q4: FIND THE TOP 5 RESTAURANTS NAME IN BENGALURU

This pie graph gives the data about the best five restaurants in Bangalore. Top first eatery is Café Coffee Day (24.3%); second one Onesta (21.5%), third one is Just Bake (18.5%)



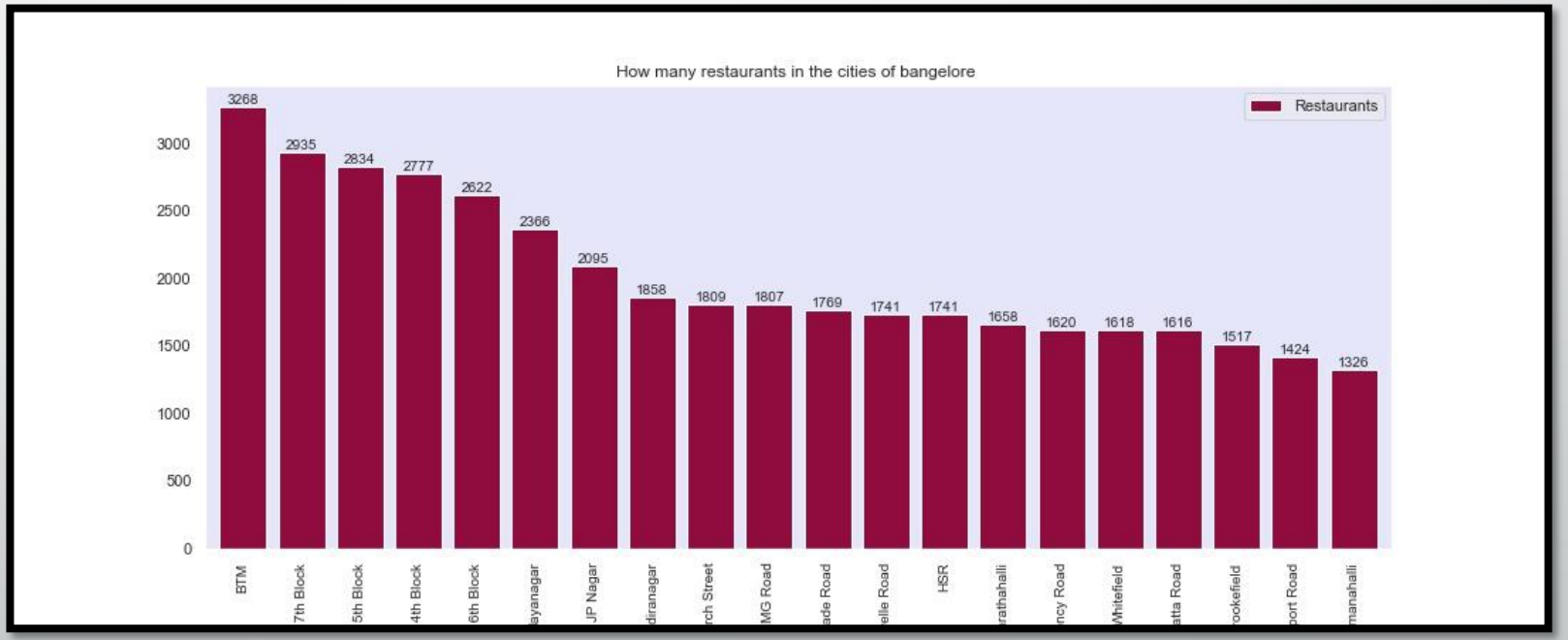
Q5: WHAT KIND OF CUISINE IS MOST POPULAR IN THE LOCALITY?

The bar graph shows the main 10 cuisine which are mostly preferred by customer. The absolute first food, which is famous, is North Indian cuisine. Second one is Chinese. Notwithstanding, among top 10 Bakery is toward the end.



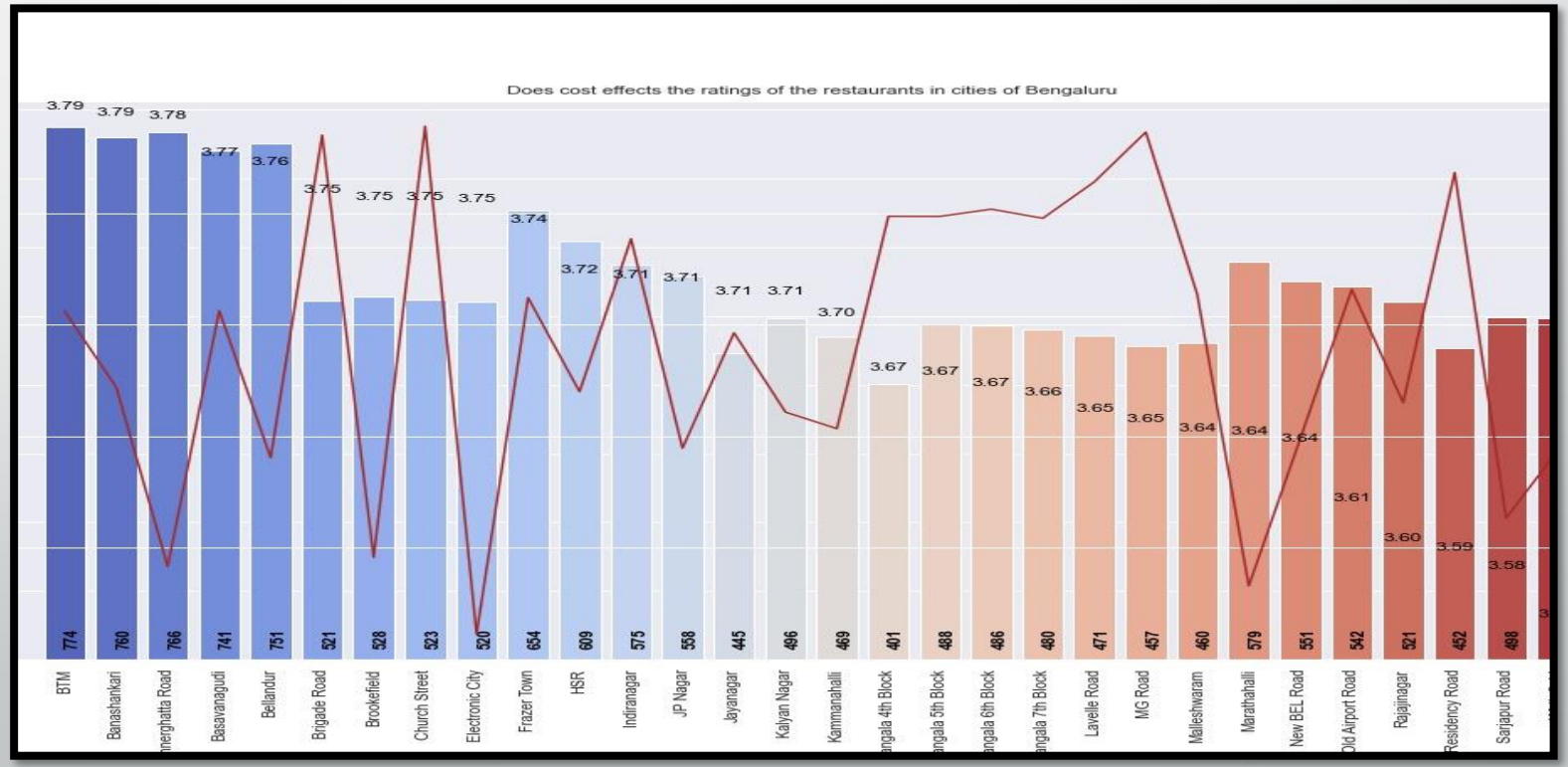
Q6: HOW MANY NUMBERS OF RESTAURANTS IN EACH NEIGHBORHOOD?

This bar diagram shows the count of eateries in every neighborhood of the Bangalore city. BTM (3268) is having biggest number of restaurants. In contrast, Old airport road and Kammanahalli areas are having most minimal number of restaurants.



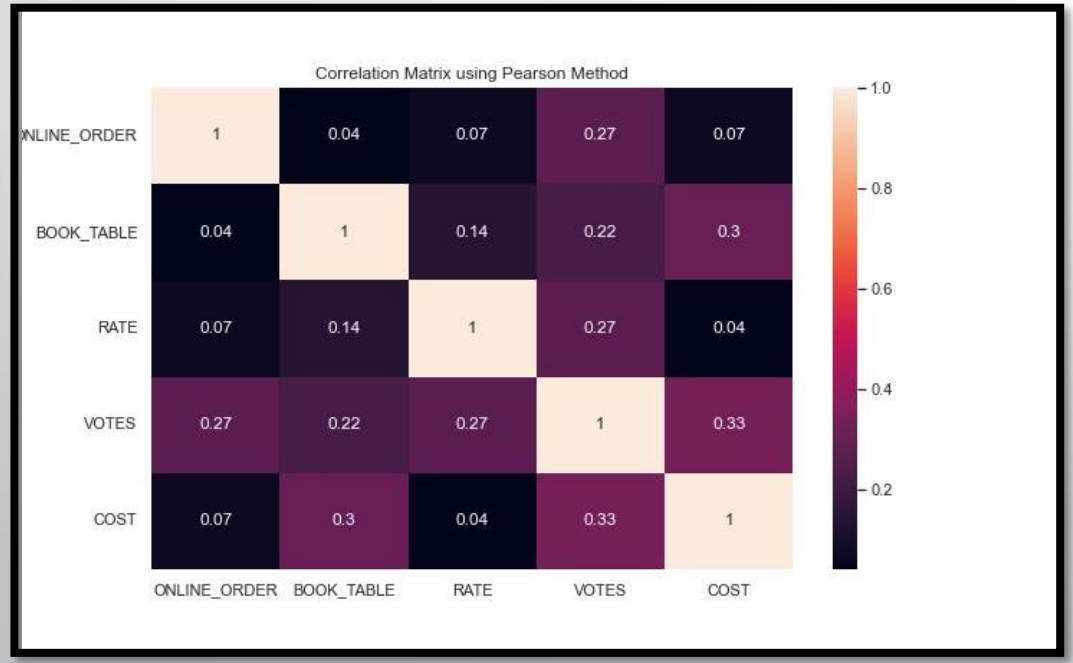
Q7: DOES COST EFFECTS THE RATINGS OF THE RESTAURANTS IN CITIES OF BENGALURU?

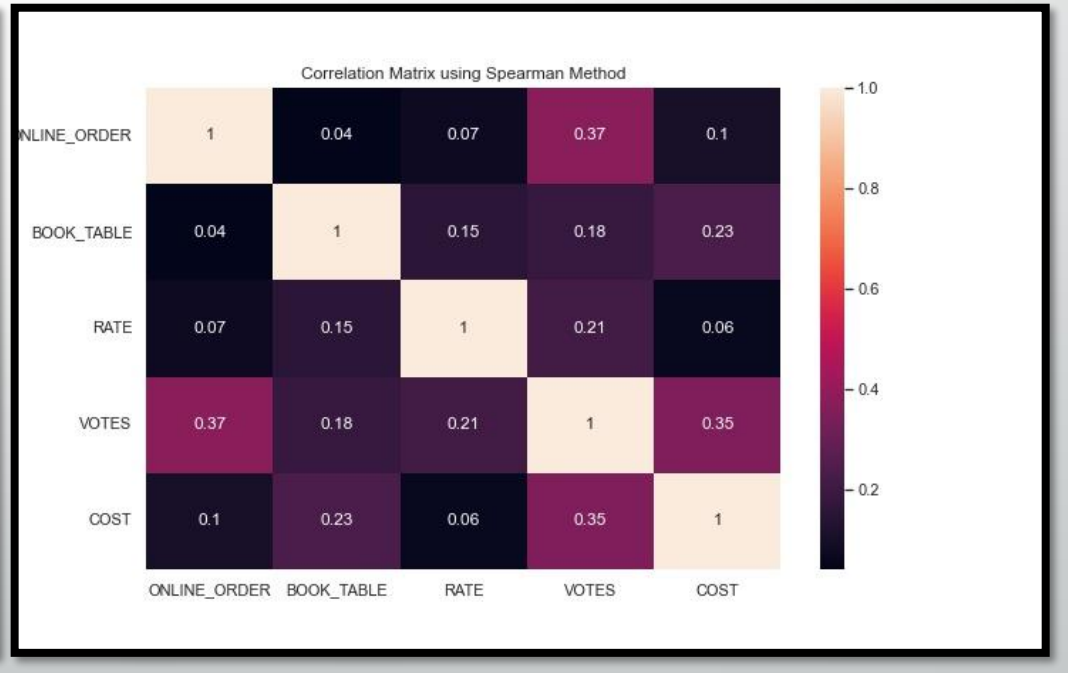
There are the best 30 urban communities wherein clients likes to eat in and takeout. In the referenced urban communities, the expense of 2 man lies in the middle 367 to 448 and rating is change between 3.64 to 3.52. By examining the above diagram, we can say that cost doesn't impact the rating of the eateries in Bengaluru.



CORRELATION MATRIX USING PEARSON AND SPEARMAN METHOD

The charts illustrates the connection between two factors utilizing two unique strategies for correlation matrix, which shows what the adjustment of one variable means for other variable. The worth changes between - 1 to 1. There is no solid connection has been seen between these the variables as a whole.





ONE-HOT ENCODING & MIN-MAX SCALING

ONE-HOT ENCODING

- One hot encoding technique represents the categorical data into binary vectors. It is a common process prior to performing classification techniques. I performed one hot encoding procedure on categorical attributes- Online request, Book table, Location, Restaurant type, Cuisines and Service type.

MIN-MAX SCALING

- It essentially shrinks the range, now somewhere in the range between 0 and 1 (or - 1 to 1 assuming there are negative qualities). It is method to normalise the data utilizing Python's min-max functions.
- Min Max Scaler does not reduce the significance of outliers. The default range for the feature returned by Min Max Scaler is 0 to 1.

EXPERIMENTAL DESIGN

TRAIN-TEST-SPLIT

The train_test_split work is for parting a single dataset for two distinct purposes: training and testing. The testing subset is for building your model. The testing subset is for utilizing the model on unknown information to assess the performance of the model. In this 70% of the information is used in training set and 30% by testing.

UNDER SAMPLING STRATEGY

This system refers to gathering of procedures to adjust our dataset. It eliminates the example from the training dataset, which has a place with the majority class. I performed under sampling method on my dataset, in light of the fact that before that my information was not stable. After performing this technique, I made my information stable.

MODEL IMPLEMENTATION & EVALUATION

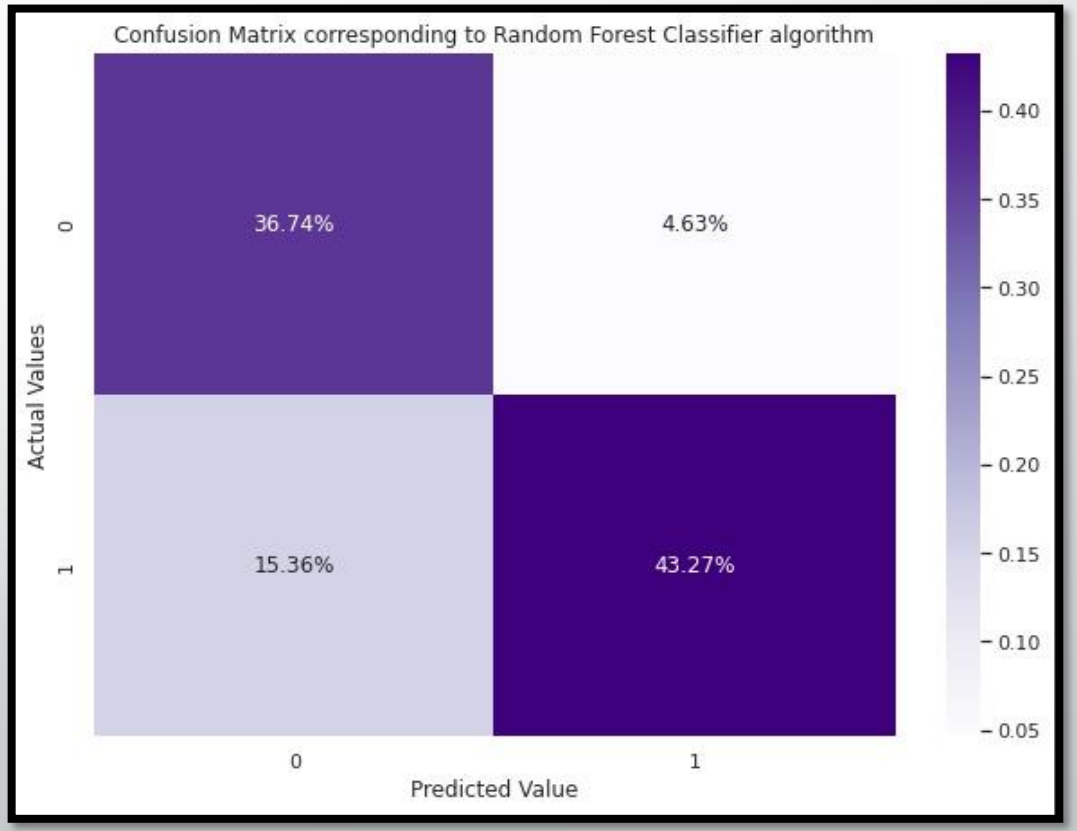
Model	Train-Test-Split
Naive Bayes	55.012322
Logistic Regression	60.847444
KNN	72.986724
Decision Tree	79.569123
Random Forest	80.411797

In the above table is the comparison of accuracies. I used linear regression, naïve bayes, knn, decision tree and random forest classifier for our dataset. Among all models, Random Forest Classifier is the best-fit model on our dataset which gives the 80% accuracy for our dataset.

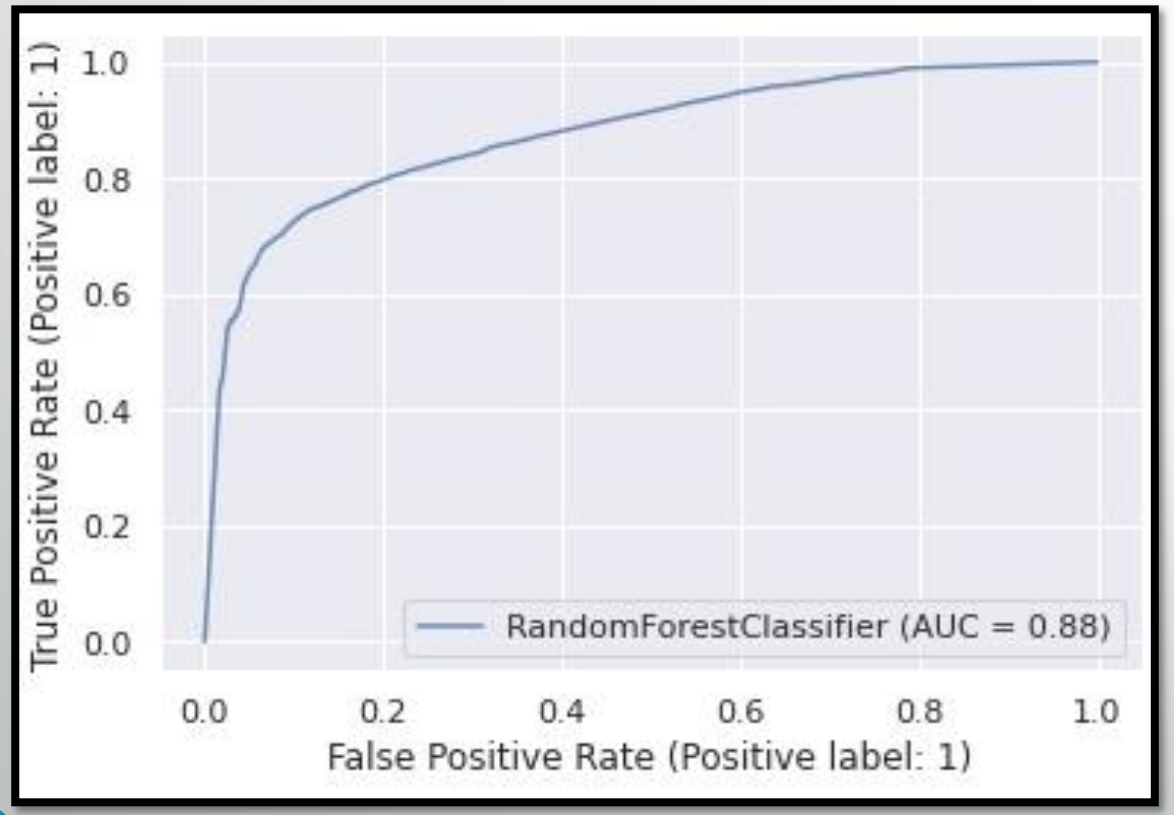
- Logistic Regression: It is use to predict the dependent variable value based on the given independent variable. Therefore, this technique shows the linear relationship between input and output variable.
- K Nearest Neighbors: : It is a supervised machine-learning algorithm. This algorithm represents the k nearest neighbor. It is used for classification and regression both..
- Gaussian Naive Bayes: It is a special type of NB algorithm. It is used when the features having continuous values.
- Random Forest: A supervised Machine Learning Algorithm is used widely in Classification and Regression problems. It makes decision trees on different samples.
- Decision Tree Classifier: It is a non-parametric learning algorithm used for classification and regression. The mail goal is to create a model that predicts the value of a target variable.

CONFUSION MATRIX CORRESPONDING TO RANDOM FOREST CLASSIFIER

The confusion matrix shows the accuracy of the optimal model. It shows that correspond to 0, only 4.63% values are wrong and 36.74% is right. However, corresponds to 1 is 15.36% are wrong values and 43.27% is right. So, it means almost 20% is predicted wrong and 80% is right correspondingly to 1.



ROC CURVE OF THE OPTIMAL MODEL RANDOM FOREST CLASSIFIER ALGORITHM



The AUC value vary between 0.5 to 1, where 0.5-0.6 auc denotes a bad classifier, 0.7 to 0.8 is considered acceptable, and 1 denotes an excellent classifier and our model gives the auc 0.8, which is much closed to 1. That means our roc curve gives the fair auc score.

CONCLUSION

We extracted data from CSV file, in this dataset many values were missing, and we did not throw up all values. Therefore, instead of removing NULL values, we tried to fill these. We have performed exploratory data analysis to answer the research questions. We used one-hot encoded features and performed different models.

Random Forest Regressor is the best-fit model, so we made correlation matrix and ROC curve with our best-fit model.

THANKYOU