Machine Learning to Predict Restaurant Ratings in Bangalore

Shridevi Angadi
20122029
MSc. DS(I)

Shubham Chowdhury
20122040
MSc. DS(I)

Rakshith
20122059
MSc. DS(I)

Kumar Gaurav
20122065
MSc. DS(I)

Akshay K
20122044
MSc. DS(I)

I. INTRODUCTION (*HEADING 1*)

In addition to being a technological hub, in the recent years, Bangalore has become a cultural hub as well. Intermingling of people of varying cultures and background has led to Bangalore becoming a vibrant and dynamic playground for traditions, lifestyles and genres, and a major part of that is food. Bangalore has restaurants that serve food from around the world, with almost every type of cuisine available at some restaurant or other. Even then, the number of restaurants is rising as even with more than 12,000 restaurants currently, the industry isn't saturated yet. However, it has become difficult for newer restaurants to compete with already established restaurants. The main issues that continue to pose a challenge to them are high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain and over-licensing.

The data from Zomato aims to analyze the demography of the location and predict the success of a restaurant based on the features, theme, cuisine, cost and other factors. This will provide an insight to help analyze whether the restaurant is capable of competing with existing restaurants and predict its chances of success.

II. PROBLEM STATEMENT

Prediction of restaurant ratings is a good precursor to the planning & establishment of a food establishment, as it helps to understand if the setup is capable of generating business effectively. Altering the cuisines, reservation process, convenience of order placement, location within the city, and pricing play a major part in this decision process, which this model helps to perform.

III. PROBLEM OBJECTIVES

All of the features of a restaurant contribute, in different degrees, to its success. Some of the hypothetical scenarios are listed:

- Online orders: Since these ratings are tracked on Zomato, which is an online food delivery service, the provision of online orders plays a direct role on the ratings, as it makes its food more accessible to working classes & corporates in that locality
- Cuisine: Since the cultural demographic changes from one location to another, a Jain restaurant, for example, may have better chances of success in locations where the population is predominantly vegetarian
- Cost: Cost is one of the important factors that people take into considering before going out to eat. It is therefore important for a restaurant to price it's menu items competitively to stay in business
- Number of votes: In addition to the online rating, the number of votes plays an important part in the decision process, as it decides the reliability of the rating

IV. DATA COLLECTION

The dataset was collected from Kaggle, where it was independently scraped for educational purposes. The data was accurate to that available on the Zomato website until 15 March,2019.

Data was extracted in two phases:

- Phase 1: URL, name and address of the restaurant were extracted which were visible on the Zomato page.
- Phase 2: The recorded data for each restaurant and each category was read and data for each restaurant was scraped individually.
- The data contains the following information:

Column Name	Description
url	URL of restaurant details (Zomato)
address	Address of the restaurant
name	Name of the restaurant
online_order	Whether restaurant accepts online orders
book_table	Whether restaurant has provision of booking a table
rate	Current rating (Out of 5)
votes	Total votes
phone	Phone contact details
location	Location
rest_type	Type (casual dining, quick bites etc.)
dish_liked	Liked dishes
cuisines	Cuisines served
approx_cost(for two people)	Approximate cost of a meal for two
reviews_list	List of reviews provided
listed_in(type)	Type of listing in Zomato
listed_in(city)	Location of listing in Zomato

V. DATA PREPARATION

The raw data was prepared to make it compatible to the methods of analysis it was planned to be subjected to, the steps used for which are listed below.

Firstly, null values were found for all the columns.

```
In [10]: data.isna().sum()

Out[10]: url                              0
         address                          0
         name                             0
         online_order                     0
         book_table                       0
         rate                          7775
         votes                            0
         phone                         1208
         location                        21
         rest_type                      227
         dish_liked                   28078
         cuisines                        45
         approx_cost(for two people)    346
         reviews_list                     0
         menu_item                        0
         listed_in(type)                  0
         listed_in(city)                  0
         dtype: int64
```

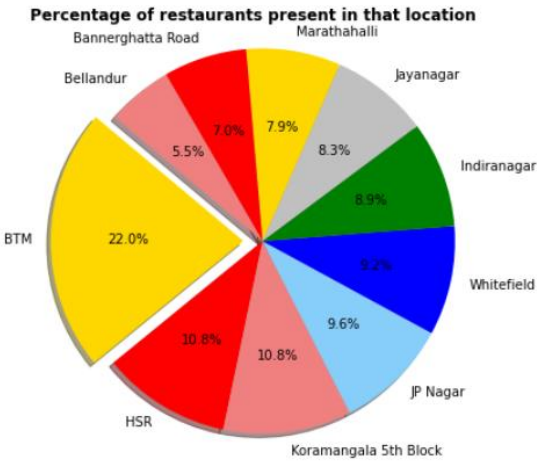
Different strategies were used for dealing with these null values. Irrelevant columns (e.g. phone) were dropped. Values 'New' were replaced with 0 for rate.

Columns irrelevant to analysis (e.g. url, address, reviews_list) were dropped.

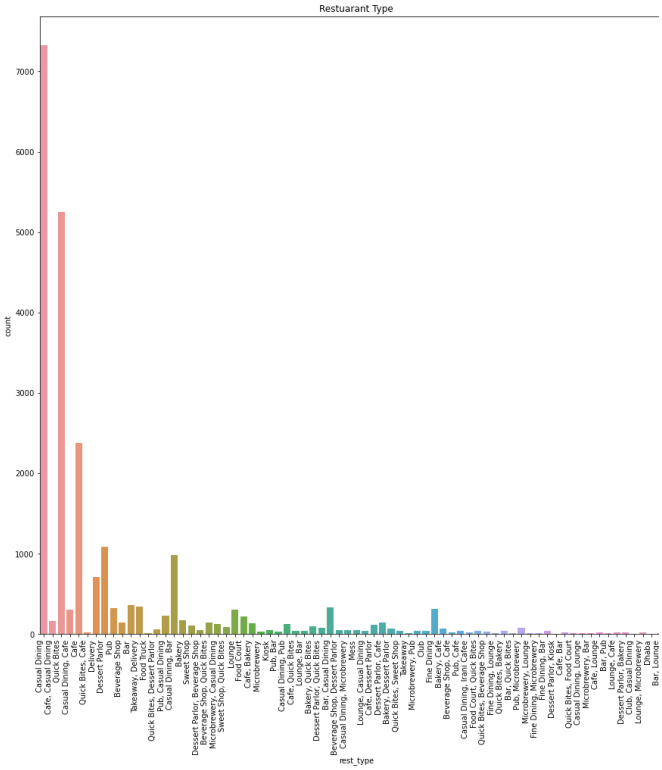
Since datatype of only 'votes' was int64, the remaining numeric values were cleaned & formatted to float or int datatypes.

Duplicates were removed to avoid redundancy.

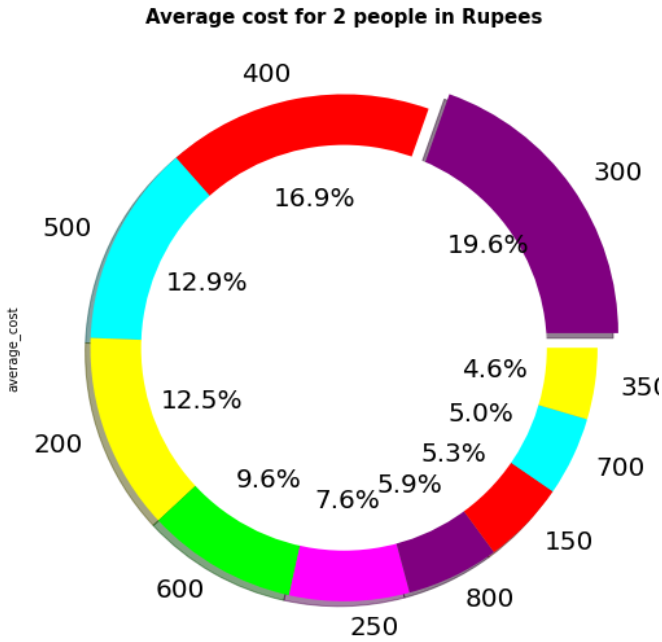
VI. EXPLORATORY DATA ANALYSIS



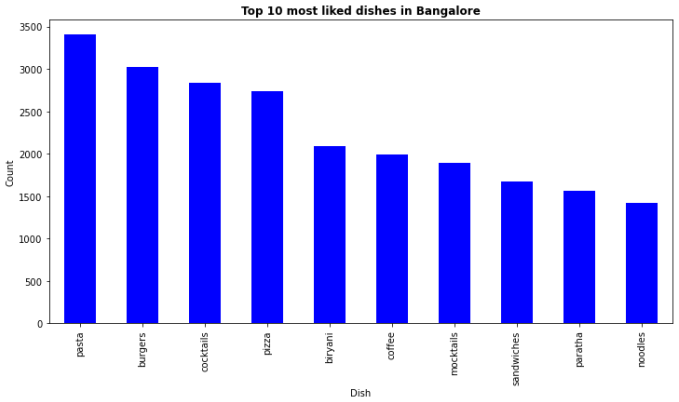
Highest number of restaurants were found to be in BTM, followed by HSR and Koramangala 5th Block.



Most frequently visited restaurant type was found to be Casual Dining, followed by quick bites and Cafes.

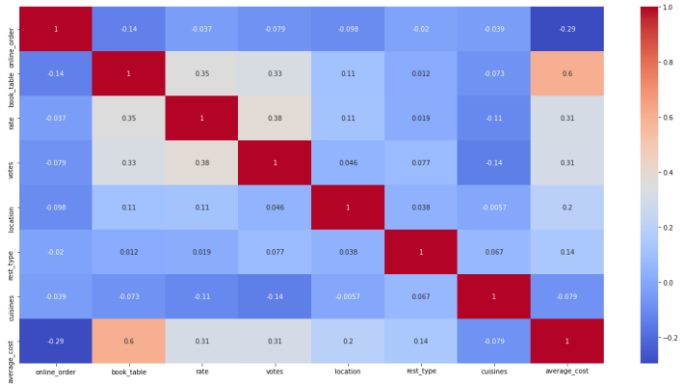


Most common average cost of dining for two people was found to be ₹300.



Pasta & burgers were found to be the top two liked dishes in Bangalore.

- 30428 restaurants are accepting online orders
- 45268 restaurants do not have the book table feature.
- The most liked cuisine in Bangalore is -North Indian



Correlation matrix was found to establish how the variables were correlated.

It was found that average cost, location, votes, and table booking were positively correlated with the restaurant's ratings, with highest correlation between votes and rating.

VII. MODEL BUILDING

Linear Regression, Decision Tree Regressor, Random Forest Regressor, K Nearest Neighbors Regressor, XGBoost, and Gradient Boosting Regressor were used for the model building part.

Maximum accuracy for predictions was found to be yielded by Random Forest Regression.

In [37]:
```
#selecting best models

model_selc = [LinearRegression(),
              DecisionTreeRegressor(),
              RandomForestRegressor(n_estimators=10),
              KNeighborsRegressor(),
              GradientBoostingRegressor()]

kfold = RepeatedKFold(n_splits=5, n_repeats=10, random_state= None)
cv_results = []
cv_results_mean =[]
for ele in model_selc:
    cross_results = cross_val_score(ele, X_train, Y_train, cv=kfold, scoring =
'r2')

    cv_results.append(cross_results)

    cv_results_mean.append(cross_results.mean())
    print("\n MODEL: ",ele,"\nMEAN R2:",cross_results.mean() )
```

```
 MODEL:  LinearRegression()
MEAN R2: 0.21787340503251784

 MODEL:  DecisionTreeRegressor()
MEAN R2: 0.8929557928261734

 MODEL:  RandomForestRegressor(n_estimators=10)
MEAN R2: 0.9261732530203263

 MODEL:  KNeighborsRegressor()
MEAN R2: 0.6210055140924114

 MODEL:  GradientBoostingRegressor()
MEAN R2: 0.752416342860573
```

In [38]:
```
#Let's try xgboost now
my_xgb = xgb.XGBRegressor(objective='reg:linear',learning_rate = 0.1, n_estima
tors = 100,verbosity = 0,silent=True)
xgb_results = cross_val_score(my_xgb, X_train, Y_train, cv=kfold, scoring ='r
2')
print("\n MODEL: XGBOOST","\nMEAN R2:",xgb_results.mean() )
```

```
 MODEL: XGBOOST
MEAN R2: 0.8074725983036702
```

VIII. CONCLUSION

Exploratory data analysis yielded some important insights to consider for restaurants, including restaurant density by location, most preferred cuisines and dishes, and most frequent average cost.

Based on these inferences, parameters were selected for training the model to predict the rating for a restaurant.