Utilizing Restaurant Data to Predict Rating and Create Recommendations for Customers

Yash Nema

Section Overview

- End-to-End Machine learning Project Overview
- Problem Ideation
- Data Capturing
- Data Cleaning and Exploration tasks
- Modeling Approach
- Data Modeling and Results
- Deployment
- Future Work

End-to-End Machine learning Project Overview

End-to-End Overview

Data Silos

Centralized Version control & Model

Deployment Silos

Collection of Data Sets

Exploratory data analysis

Data modelling

Deployment

Problem Ideation

Project Setting

- The restaurant industry has tripled in last 25 years with the advent of Online ordering and changing of geographical conditions

- Rating will play a crucial role as it will be a measure of their customer approval as well as can be parameter for

- Restaurant owners coming into the business need to understand the data so that they can perfect in term of some their offerings like
    - Cuisine's
    - Location
    - Rate/Cost etc.

Project Definition

- Goal 1: Understand what factors impact Rating of a restaurant
- Goal 2: Identify and cluster restaurants based on common attributes and see trends
- Goal 3: Predicting new restaurant rating depending on input parameters
- Goal 4: Building recommendation system to suggest restaurant based on customer preference

Data Capturing

Data Sources & Description

Zomato API(Kaggle Dataset)

Sn	Column Name	Description
1	URL	Restaurant URL
2	Address	Address of the restaurant
3	Restaurant Name	Name
4	Has Online Delivery	Yes/No
5	Has Table Booking	Yes/No
6	Aggregate Rating	Average rating out of 5
7	Votes	Ratings casted by people
8	Phone	City
9	Locality	Location in the city
10	Restaurant Type	Type of Restaurant
11	Dish Liked	Dish liked
12	Cuisines	Cuisines offered
13	Average cost of two	Cost for two people
14	Review	Reviews
15	Menu Items	Yes/No
16	Listed type	Type of Serving
17	Listed City	City Name

Data Cleaning & Exploration

Data Cleaning

- Dropped Columns Like URL, Phone and dish liked as they are personal and has no influence on decision parameter

- We started with **17** columns and after cleaning we came down to **14** columns

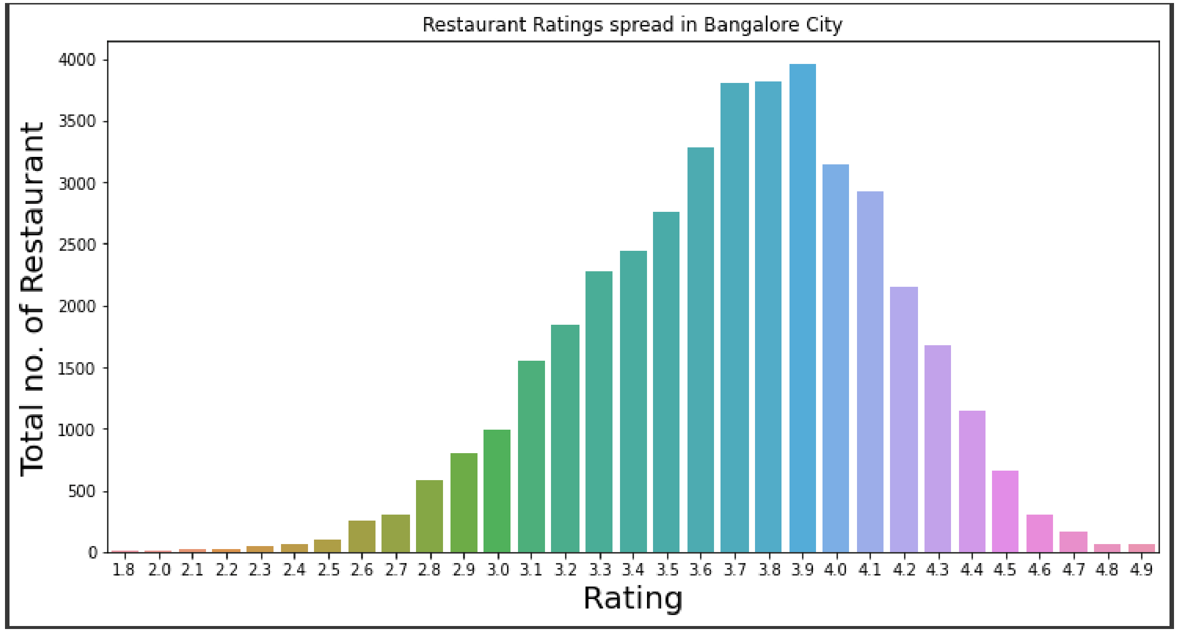
- Removing NaN values resulted in rows from **51717** to **41237**

- Renamed some columns name

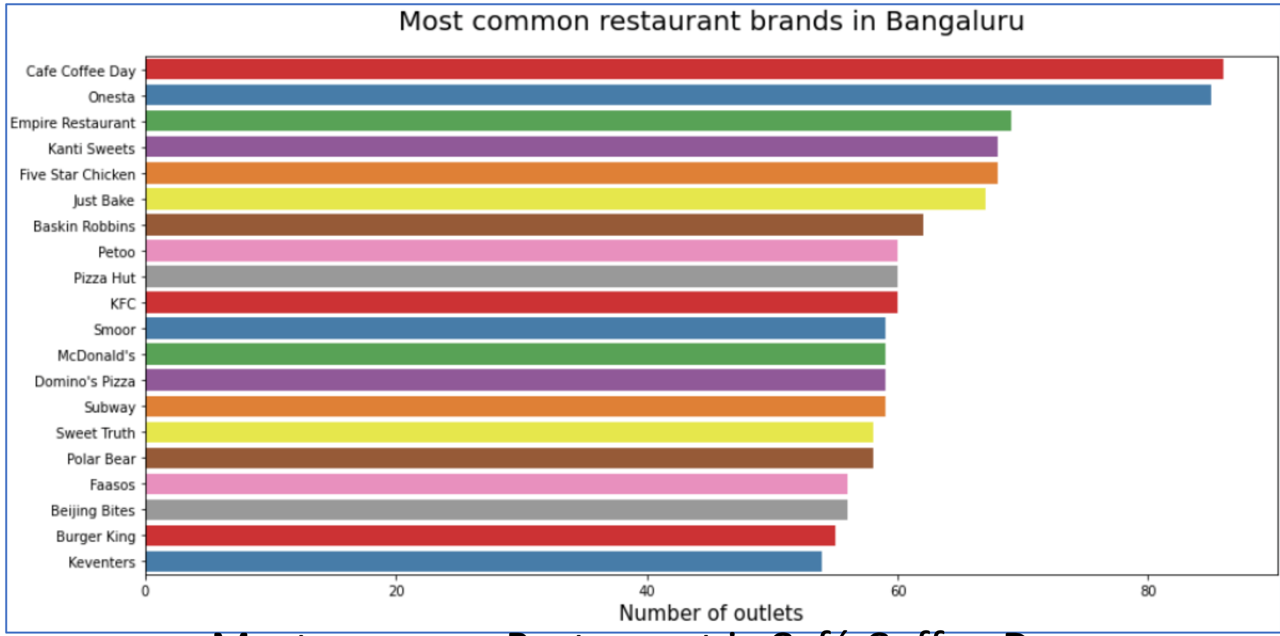
Data Exploration

- How many restaurants are in each area within Bangalore?
- How does ratings vary with restaurants?
- How many restaurants offer Online ordering option and who offers offline only options
- What percentage of restaurants have “book a table” option?
- Which areas in Bangalore vote the most, and what's the average number of votes for all areas
- What are the most popular cuisines offered by restaurants?
- What are the most popular restaurant types?
- How is type of service distributed among restaurants in Bangalore?
- Does offering “Table booking” impact the Restaurant ratings
- How much do restaurants charge for 2 people?
- Which restaurants have the most branches in Bangalore
- Do top cuisines change depending on the cost for two?
- How is correlation between rate column and votes cost.

Data Exploration



Most Restaurant's has a Rating of 3.9



Most common Restaurant is Café Coffee Day

Data Exploration

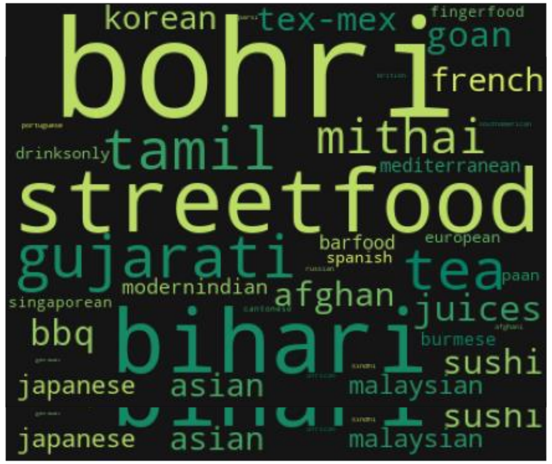
	Rest_type	Frequency
0	Quick Bites	14193
1	Casual Dining	11221
2	Cafe	3960
3	Dessert Parlor	2268
4	Delivery	1666
5	Takeaway	1357
6	Delivery	1278
7	Bar	1228
8	Bakery	1131
9	Bar	1045
10	Beverage Shop	981
11	Casual Dining	961
12	Quick Bites	937
13	Pub	731
14	Cafe	643

Most Restaurant's are Quick Bites

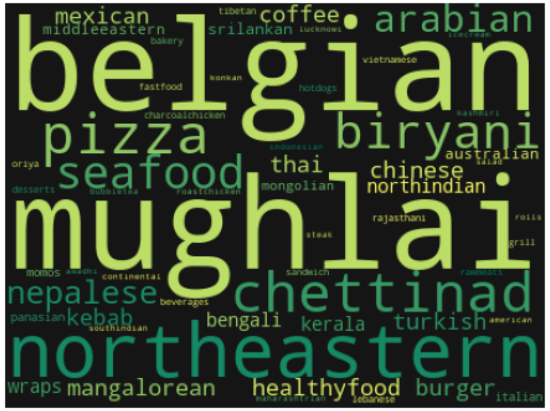
	Cuisines	Frequency
0	Chinese	10321
1	North Indian	9695
2	North Indian	7503
3	Fast Food	4446
4	Cafe	3951
...	...	...
179	German	3
180	North Eastern	2
181	Paan	2
182	Pan Asian	1
183	Singaporean	1

Most common Cuisine is Chinese

Low Budget



Mid Budget



High Budget

Modeling Approach

Modeling Approach

We started exploring data focusing on the Goal

Goal 1

- Encoded categorical variables using imputation or binary encoding
- Build Linear Regression, Decision Tree and Extra Tree Regression Model to understand factors impacting rating

Goal 2

- Used Elbow Curve to identify cluster
- Cluster will help us understand how each restaurant relates to each other

Goal 3

- As most columns are numerical will use Multiple Linear Regression, Decision tree, Random Forest to predict new data points columns
- Will use R2 score to see model accuracy on Train and Test Data

Goal 4

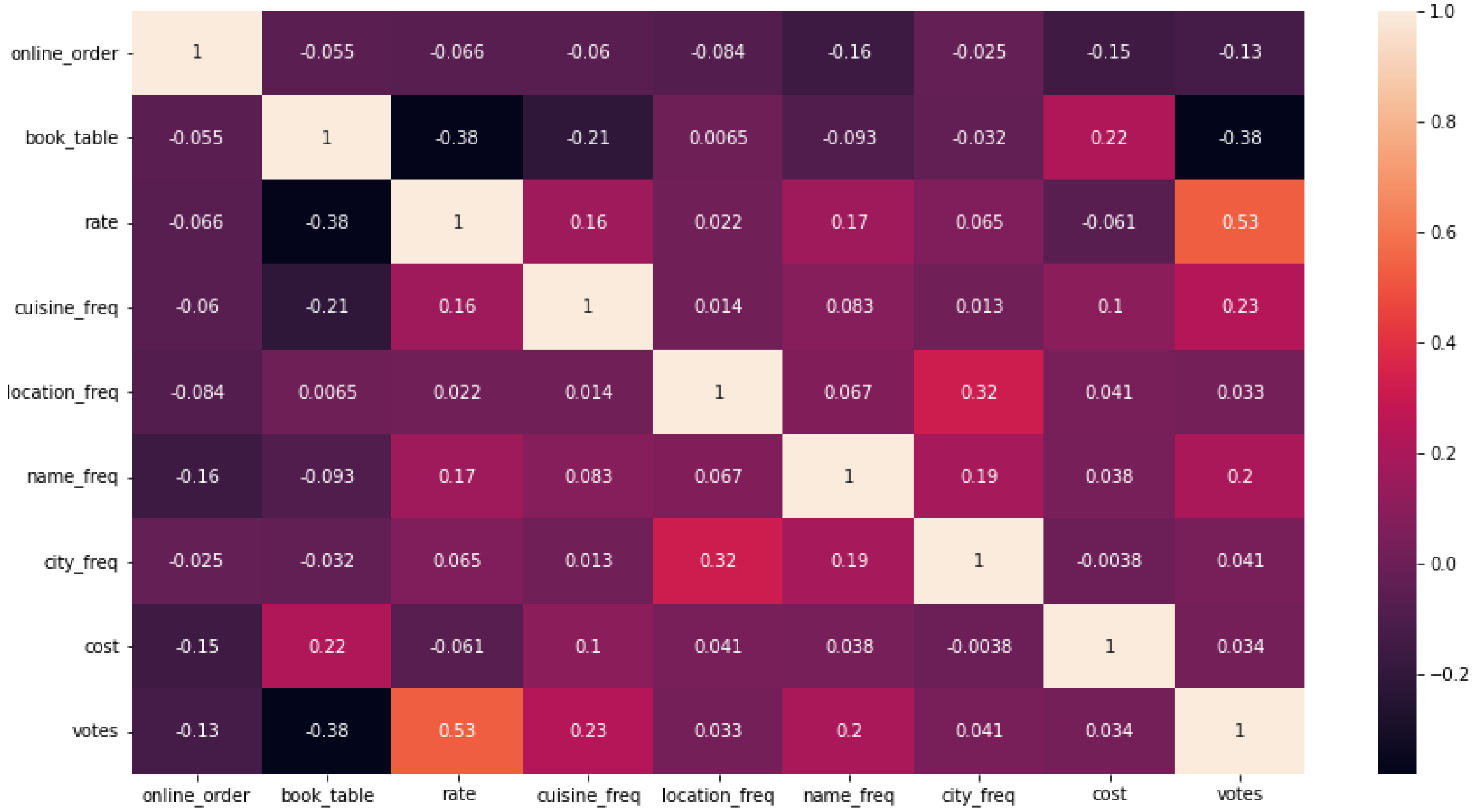
- For recommendation model will use TF-IDF statistical method to understand cuisines
- implementing a content-based recommendation using the rating information

Data Modeling and Results

Variable Importance

Goal 1

- After analyzing we got to know that factors like
  - Restaurant Votes,
  - Location and
  - Booking table option, and
  - cost for two-person meal are important
- Factors like
  - Online Order
  - Cuisine range offered
  - Menu items are not important factors



Variable Importance

Goal 1

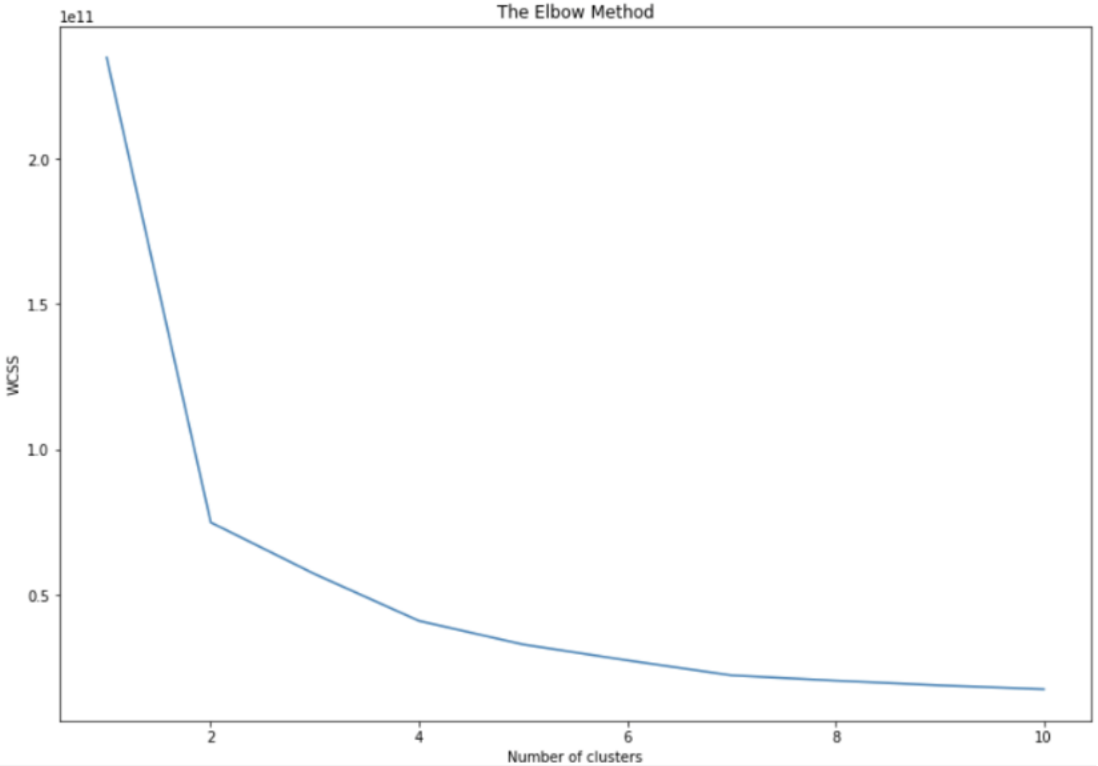
- After analyzing we got to know that factors like
  - Restaurant Votes,
  - Location and
  - Booking table option, and
  - cost for two-person meal are important
- Factors like
  - Online Order
  - Cuisine range offered
  - Menu items are not important factors

	Model Implemented		
Features	Linear Regression	Decision Tree	E Tree Regression
online_order	-0.23	0.02	0.02
book_table	-0.67	0.02	0.18
cost	-0.06	0.07	0.09
votes	0.27	0.56	0.30
cuisine_freq	0.04	0.04	0.06
location_freq	-0.03	0.08	0.08
name_freq	0.03	0.09	0.10
city_freq	0.07	0.03	0.03
type_Buffet	-0.06	0.00	0.00
type_Cafes	-0.01	0.00	0.00
type_Delivery	-0.05	0.00	0.00
type_Desserts	0.10	0.00	0.00
type_Dine-out	-0.04	0.00	0.00
type_Drinks & Nighlife	0.02	0.00	0.00
type_Pubs and bars	0.04	0.00	0.00
bakery	-0.02	0.00	0.01
bar	0.14	0.00	0.01
beverageshop	0.11	0.00	0.00
bhojanalya	-0.68	0.00	0.00
cafe	0.38	0.02	0.02
casualdining	0.10	0.01	0.01
club	0.08	0.00	0.00
confectionery	-0.20	0.00	0.00
delivery	-0.04	0.01	0.01
dessertparlor	0.53	0.02	0.02
dhaba	-0.63	0.00	0.00
finedining	0.71	0.00	0.00
foodcourt	-0.22	0.00	0.00
foodtruck	-0.04	0.00	0.00
iranicafee	-0.16	0.00	0.00
kiosk	0.09	0.00	0.00
lounge	0.08	0.00	0.00
meatshop	0.57	0.00	0.00
mess	0.11	0.00	0.00
microbrewery	0.08	0.00	0.00
pub	0.11	0.00	0.00
quickbites	-0.08	0.01	0.01
sweetshop	0.05	0.00	0.00
takeaway	-0.13	0.00	0.00

Model Selection and Implementation

Goal 2

- Using Elbow curve, we found that three clusters are optimal



Goal 3

- After summarizing the model performance on the validation test. We can now predict new ratings using the Extra tree

	Model Training Data Performance			
Metric	**Naïve Model**	**Multi Regression**	**Decision Tree**	**Extra Tree Regressor**
R2 Score	0	0.29	0.94	0.99
Mean Absolute Error	0.1	0.08	0.014	0.0003
RMSE	0.44	0.37	0.105	0.0097

Model Validation Data Performance			
Metric	**Multi Regression**	**Decision Tree**	**Extra Tree Regressor**
R2 Score	0.27	0.86	0.94
Mean Absolute Error	0.08	0.023	0.011
RMSE	0.38	0.168	0.108

Recommendation System

- Recommendation Systems are a type of information filtering systems to improve the quality of search results
- They are active information filtering systems which personalize the information coming to a user based on his interests
- **Recommender system will look at the reviews of other restaurants, and System will recommend us other restaurants with similar reviews and sort them from the highest rated**.
- After creation of the dataset in the required format we chose **Term Frequency-Inverse Document Frequency** while creating the Recommendation model vectors. The models priorities review similarities and shows the respective **cuisines, Mean rating and cost for dining for the** recommend restaurants.

TOP 8 RESTAURANTS LIKE Woodee Pizza WITH SIMILAR REVIEWS:

	cuisines	Mean Rating	cost
Mojo Pizza - 2X Toppings	Pizza	4.13	600.0
Pizza Stop	Pizza, Italian	3.27	500.0
Pizza Hut	Pizza, Fast Food	3.03	750.0
Pizza Hut	Pizza	3.03	750.0
Deshi Fusion Pizza	Pizza, Italian, Chinese, Rolls, Biryani	2.94	750.0
Deshi Fusion Pizza	Pizza, Chinese, Rolls	2.94	750.0
The Tower Of Pizza	Pizza, Italian	2.40	500.0
Crunch Pizzas	Italian, Pizza	2.11	600.0

Example of finding restaurants similar to “Onesta” a Pizza chain in Bangalore, India

Deployment

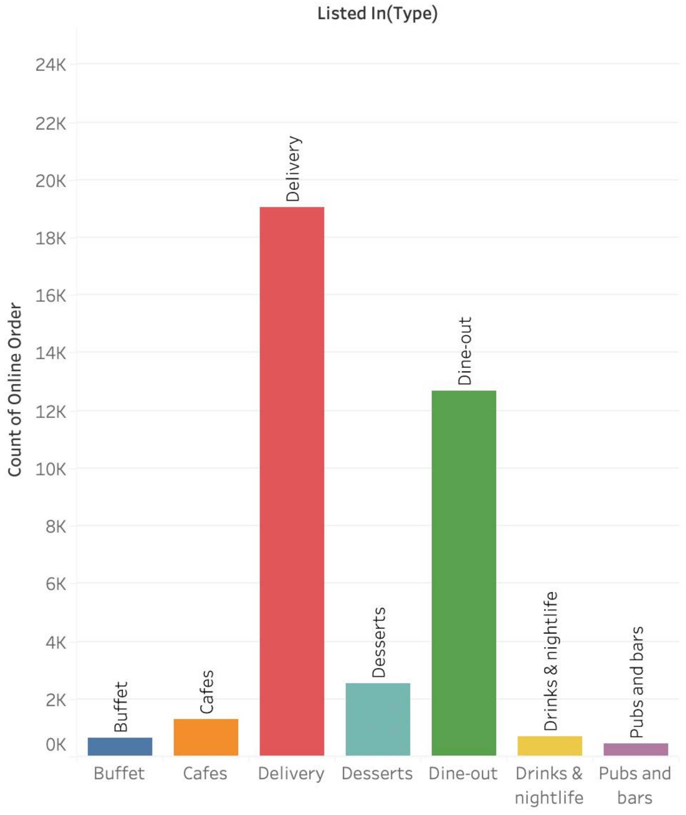
Deployment

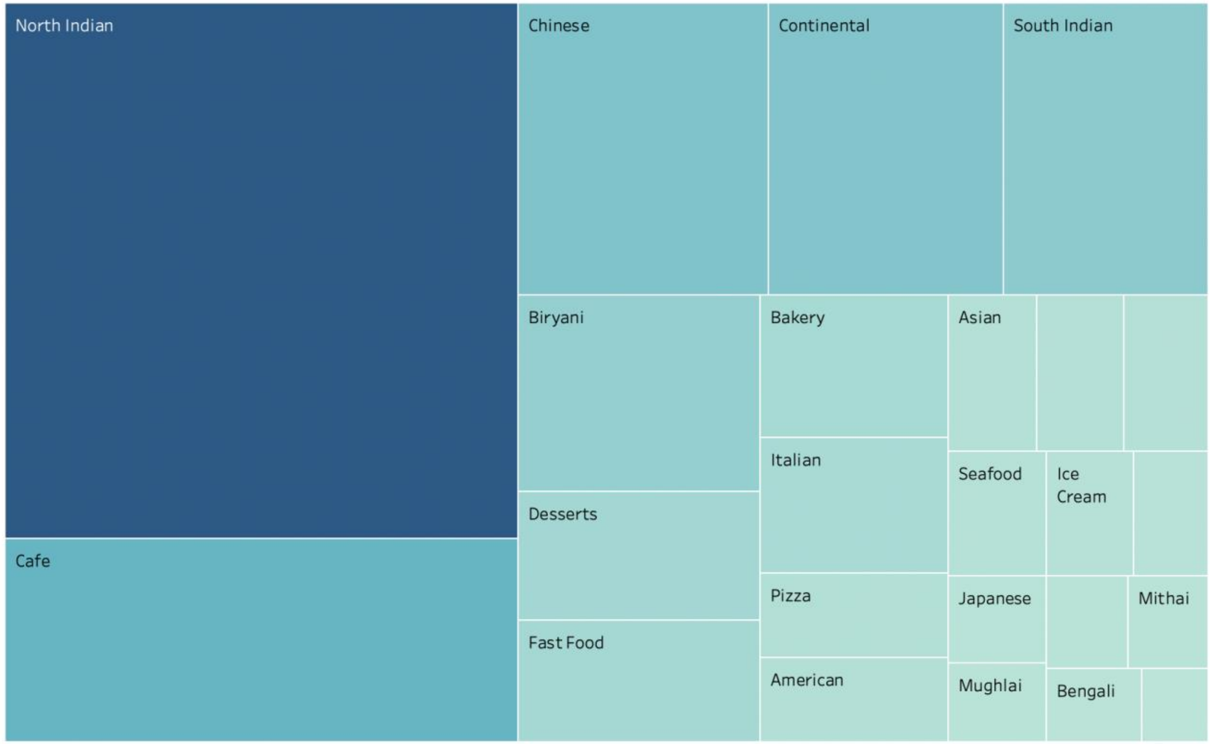
EDA, Variable Importance and Rating Predictor

Recommendation System

Tableau Dashboard Views

Sample Tableau Views





Key Takeaways

Summary of the 4 Goals and recommendations

1

Variable Importance

We saw through our EDA, Data models like Extra tree regression model how predictors such as Votes, Online Booking availability, Cost per meal for 2 people impact rating of a restaurant. This can further be improved by adding new variables like demographics and GDP

2

Clustering Restaurants

We used Elbow curve to determine that restaurants can be clustered into 3 distinct groups.

The cluster wise data can be used to understand trends among the restaurants in each cluster and ratings for new restaurants

3

New Restaurant Prediction

We found that Extra Tree Regression preformed the best on Train and Test dataset. It had the highest model accuracy, low RMSE and MAPE errors

This model can be used to predict ratings for new restaurants

4

Recommendation Model

We computed TF-IDF matrix using the customer Review list for each restaurant and then computed cosine similarity to recommend similar restaurants. This model can be easily deployed using Google Big Query on Tableau and other deployment Silos