SYRACUSE UNIVERSITY · FOUNDED 1870 ·

SUOS
CULTORES
SCIENTIA
CORONAT

THE
iSCHOOL
Syracuse University

IST 718
Big Data Analytics

PROJECT REPORT

Kartheek Sunkara | Rohit Menon
Neel Samant | Shraddha Sawant
Rashmitha Varma

BIG DATA ANALYTICS FEATURING FOOD DELIVERY APP - ZOMATO

Introduction

Food. Everyone loves it, everyone has it. Everyone even talks about it. Food is something we can talk about for hours and hours. The restaurant industry in India has witnessed an unprecedented transformation with the entry of a variety of national and international players. Thanks to the technological revolution, Indian restaurant setups have now gone online to gain more customers and serve them better. It is not surprising that the higher frequency of eating out has also evolved the market for the food services sector. The Indian food service market has come a long way from the early Nineties when it was dominated by unorganized players and few brands. The revolution began in 1996 with McDonalds, Pizza Hut, Dominos Pizza, Subway and Yo!China, among others, setting up shop in the country. Since then, the food services market has been continuously growing. The good news is that the food services industry is set to grow for many years to come, given the rising disposable incomes, a greater population of younger people, the growth of consumers in smaller towns and the widening exposure to new cultures and cuisines besides an increased propensity of eating outside the home. The analysis will mainly help new restaurants in examining the factors affecting their restaurant location.

Purpose of Study

The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the aggregate rating of each restaurant, establishment of different types of restaurant at different places, Bengaluru being one such city has more than 50,000 restaurants with restaurants serving dishes from all over the world. With each day new restaurants opening the industry has'nt been saturated yet and the demand is increasing day by day. Inspite of increasing demand it however has become difficult for new restaurants to compete with established restaurants. Most of them serving the same food. Bengaluru being an IT capital of India. Most of the people here are dependent mainly on the restaurant food as most people don't have time to cook for themselves. With such an overwhelming demand of restaurants it has therefore become important to study the demography of a location. What kind of a food is more popular in a locality. Do the entire locality loves vegetarian food. If yes then is that locality populated by a particular sect of people for eg. Jain, Marwaris, Gujaratis who are mostly vegetarian. These kind of analysis can be done using the data, by studying the factors such as

- Approx Price of food
- Location of the restaurant
- Theme based restaurant or not
- Which locality of that city serves that cuisines with maximum number of restaurants
- The needs of people who are striving to get the best cuisine of the neighborhood
- Is a particular neighborhood famous for its own kind of food

Just so that you have a good meal the next time you step out.

Size

The dataset contains 17 variables all of which were scraped from the zomato website. The dataset contains details of more than 50,000 restaurants in Bengaluru in each of its neighborhood. The data is correct to the best of my knowledge, to that available on the Zomato website until 15 March 2019. The total size of the dataset is approximately 547MB. The dataset examined has the following dimensions:

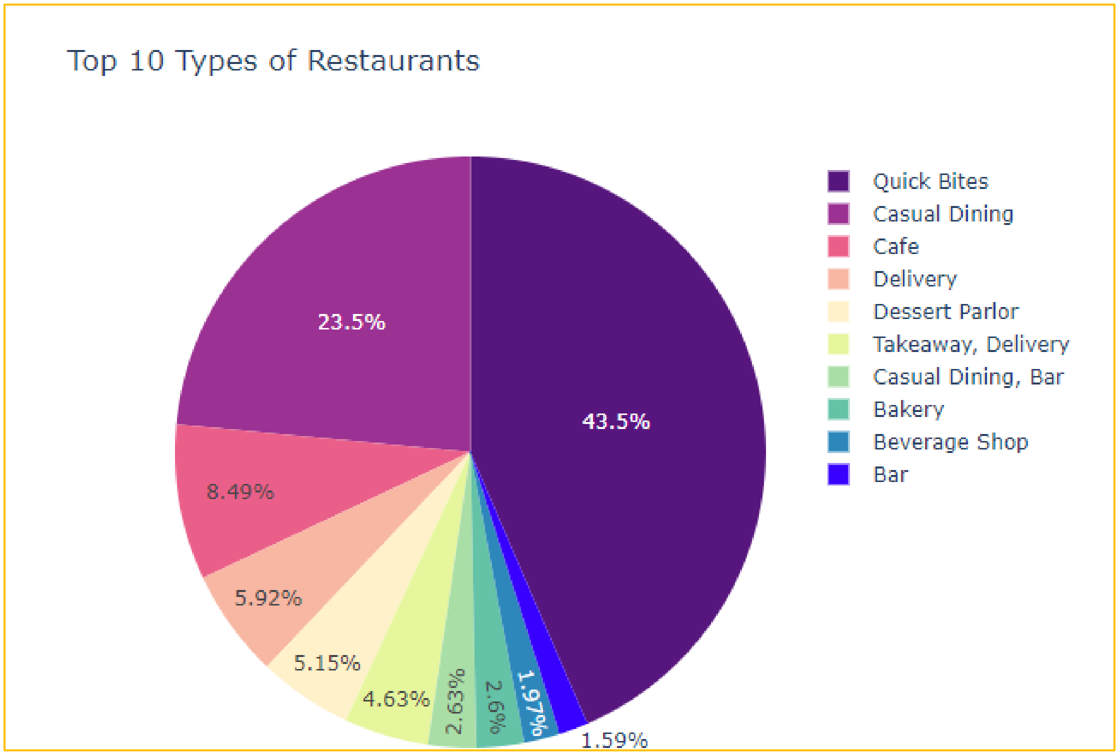
Feature	Result
Number of Observations	51,717
Number of Variables	17

Variable names and description

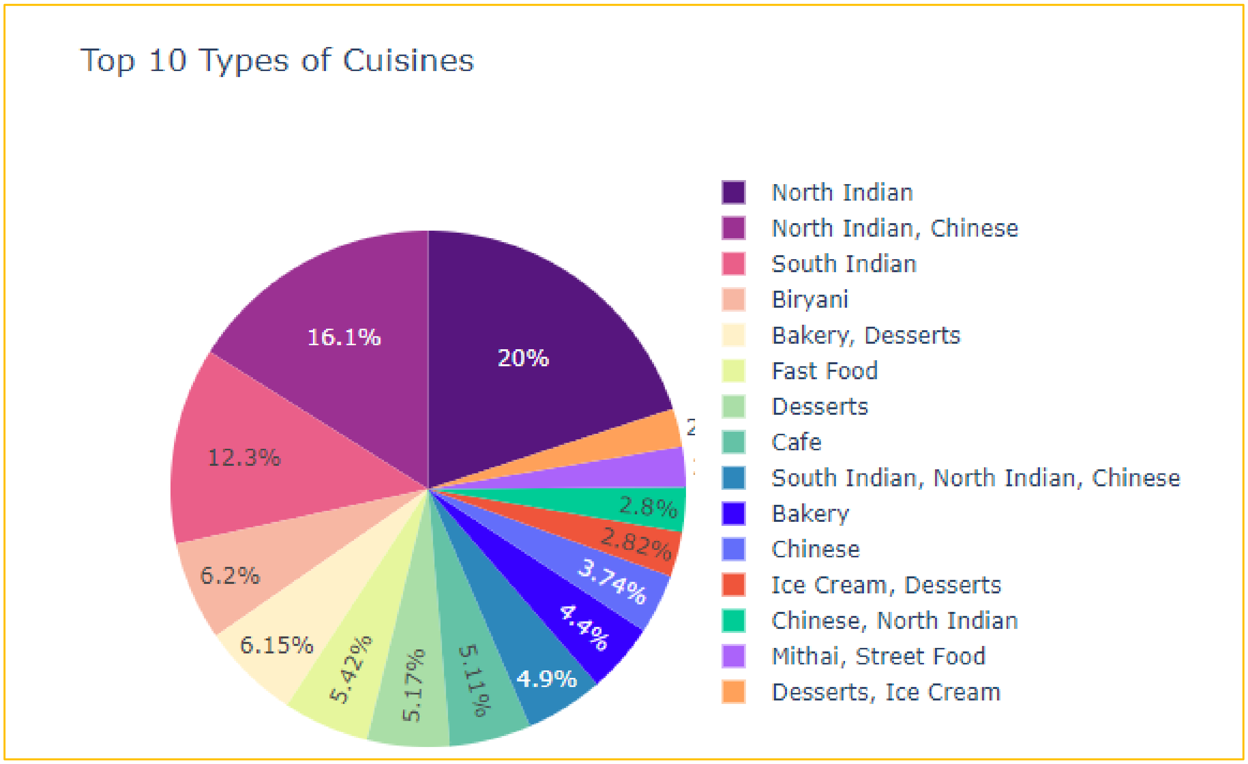
Variable	Type	Unique Values	Description
url	object	51,717	contains the url of the restaurant in the zomato website
address	object	11,495	contains the address of the restaurant in Bengaluru
name	object	8,792	contains the name of the restaurant online
Order	category	2	whether online ordering is available in the restaurant or not
book table	category	2	table book option available or not
Rate	Object	64	contains the overall rating of the restaurant out of 5
Votes	Int	2328	contains total number of rating for the restaurant as of the above-mentioned date
Phone	Object	64	contains the phone number of the restaurant
Location	Category	93	contains the neighborhood in which the restaurant is located
Rest type	Category	93	restaurant type
Dish liked	Object	5271	dishes people liked in the restaurant
Cuisine	Object	2723	food styles
Approx. cost (for two people)	Float	70	contains the approximate cost for meal for two people reviews
List	Object	22513	list of tuples containing reviews for the restaurant, each tuple consists of two values, rating and review by the customer
Menu item	Object	9098	contains list of menus available in the restaurant
listed in(type)	category	7	type of meal
Listed in (city)	Category	30	contains the neighborhood in which the restaurant is listed

DATA EXPLORATORY ANALYSIS

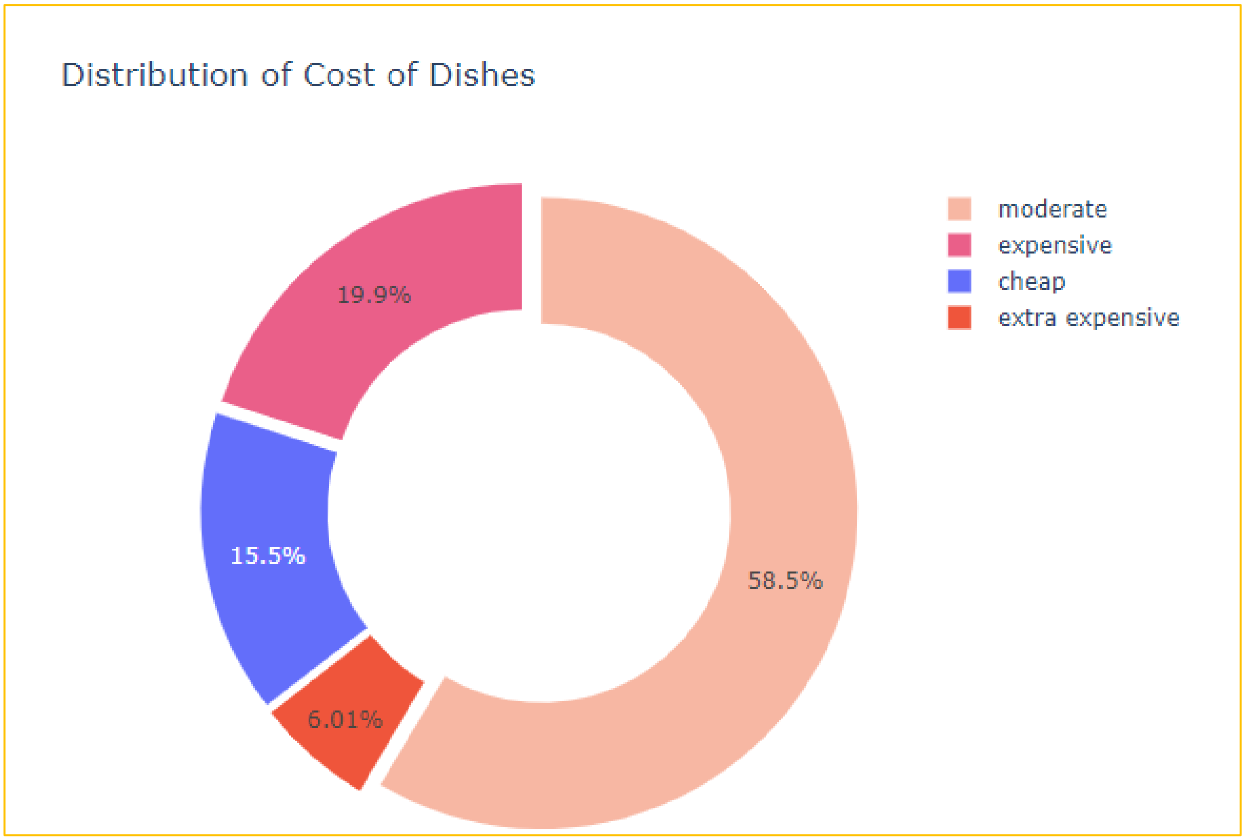
The following visualization shows the top 10 types of restaurants famous in Banglore



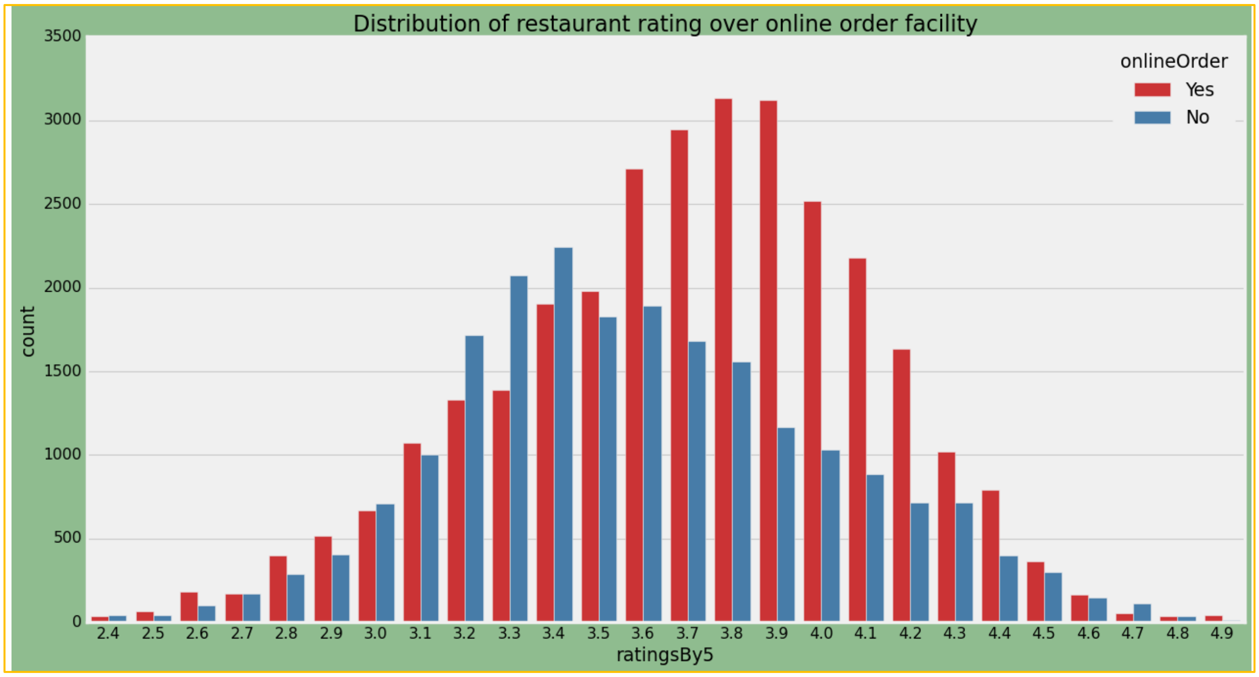
The following visualization shows the top 10 types of cuisines famous in Banglore



The following visualization shows the distribution of the cost of dishes in restaurants in Banglore

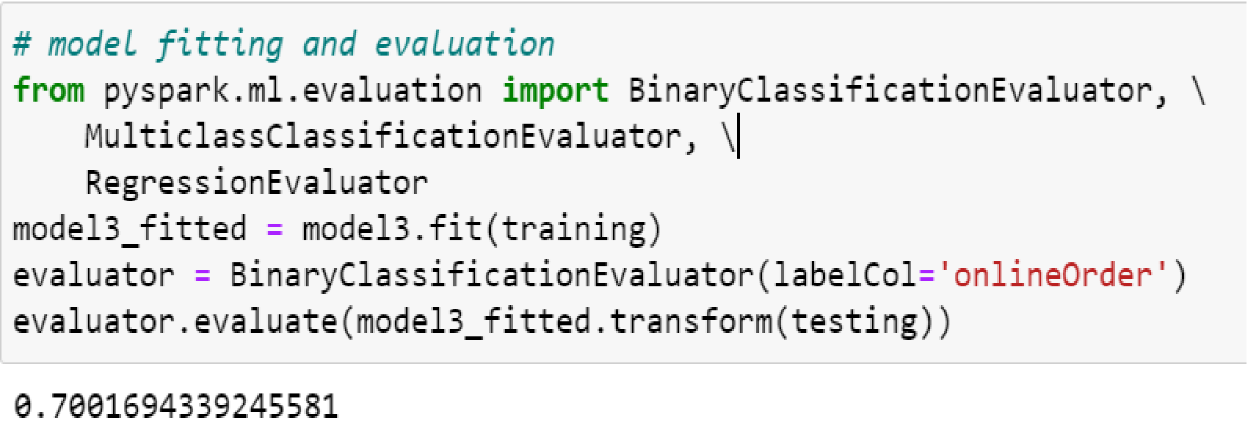


The following visualization shows the distribution of the rating of a restaurant based on its online delivery facility in Banglore



MODELS

Random Forest Classifier

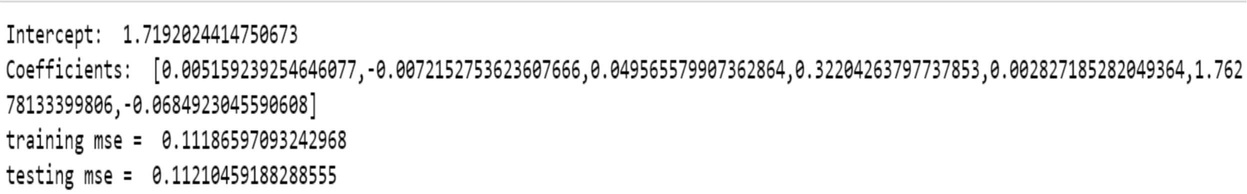


```
# model fitting and evaluation
from pyspark.ml.evaluation import BinaryClassificationEvaluator, \
    MulticlassClassificationEvaluator, \
    RegressionEvaluator
model3_fitted = model3.fit(training)
evaluator = BinaryClassificationEvaluator(labelCol='onlineOrder')
evaluator.evaluate(model3_fitted.transform(testing))
```

```
0.7001694339245581
```

The above random forest classifier has number of tress as 20, and the maximum tree depth is 5 and it is approximately 70.017% accurate.

Linear Regression

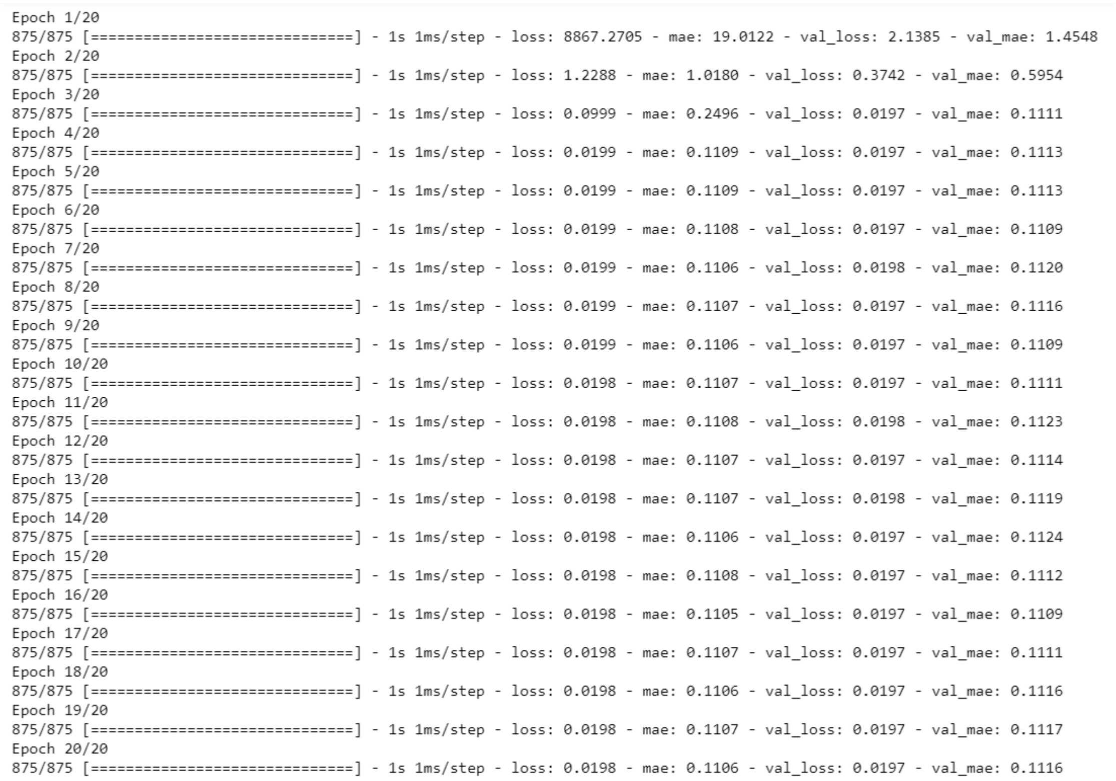


```
Intercept:  1.7192024414750673
Coefficients:  [0.005159239254646077,-0.007215275362360766б,0.049565579907362864,0.32204263797737853,0.002827185282049364,1.762
78133399806,-0.0684923045590608]
training mse =  0.11186597093242968
testing mse =  0.11210459188288555
```

For the linear regression model, the dependent variable is the overall rating of the restaurant and the independent variables are votes, location, online order, book table option availability, restaurant type, dishes liked and approximate costs for 2 people respectively. The above image displays the coefficients of the independent variables and the mean squared error for the training and testing datasets.

Neural Network

We have implemented the neural network using the keras package with 3 layers in total. There are 7 nodes in the input layer, 9 nodes in the output layer and a regressor predictor in the output layer. The loss value is calculated using stochastic gradient descent. Using this neural network, we are predicting the approximate cost for 2 people at a restaurant.



```
Epoch 1/20
875/875 [==============================] - 1s 1ms/step - loss: 8867.2705 - mae: 19.0122 - val_loss: 2.1385 - val_mae: 1.4548
Epoch 2/20
875/875 [==============================] - 1s 1ms/step - loss: 1.2288 - mae: 1.0180 - val_loss: 0.3742 - val_mae: 0.5954
Epoch 3/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0999 - mae: 0.2496 - val_loss: 0.0197 - val_mae: 0.1111
Epoch 4/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0199 - mae: 0.1109 - val_loss: 0.0197 - val_mae: 0.1113
Epoch 5/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0199 - mae: 0.1109 - val_loss: 0.0197 - val_mae: 0.1113
Epoch 6/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0199 - mae: 0.1108 - val_loss: 0.0197 - val_mae: 0.1109
Epoch 7/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0199 - mae: 0.1106 - val_loss: 0.0198 - val_mae: 0.1120
Epoch 8/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0199 - mae: 0.1107 - val_loss: 0.0197 - val_mae: 0.1116
Epoch 9/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0199 - mae: 0.1106 - val_loss: 0.0197 - val_mae: 0.1109
Epoch 10/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1107 - val_loss: 0.0197 - val_mae: 0.1111
Epoch 11/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1108 - val_loss: 0.0198 - val_mae: 0.1123
Epoch 12/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1107 - val_loss: 0.0197 - val_mae: 0.1114
Epoch 13/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1107 - val_loss: 0.0198 - val_mae: 0.1119
Epoch 14/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1106 - val_loss: 0.0197 - val_mae: 0.1124
Epoch 15/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1108 - val_loss: 0.0197 - val_mae: 0.1112
Epoch 16/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1105 - val_loss: 0.0197 - val_mae: 0.1109
Epoch 17/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1107 - val_loss: 0.0197 - val_mae: 0.1111
Epoch 18/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1106 - val_loss: 0.0197 - val_mae: 0.1116
Epoch 19/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1107 - val_loss: 0.0197 - val_mae: 0.1117
Epoch 20/20
875/875 [==============================] - 1s 1ms/step - loss: 0.0198 - mae: 0.1106 - val_loss: 0.0197 - val_mae: 0.1116
```

RECOMMENDATIONS

Following recommendations are deduced from the results generated by the models:

- **Linear Regression Model –**
  Independent Restaurant Owners should take into account the factors- votes, location, online order services, booking table services, restaurant type, liked dishes and approximate cost for 2 people, when developing strategies to improve their business or invest in a new ventures.

- **Random Forest Classifier Model –**
  Zomato should analyze the restaurant type, location, approximate cost, votes, ratings and cuisines to predict and identify potential clients who they can convince to subscribe to Zomato’s online delivery services.

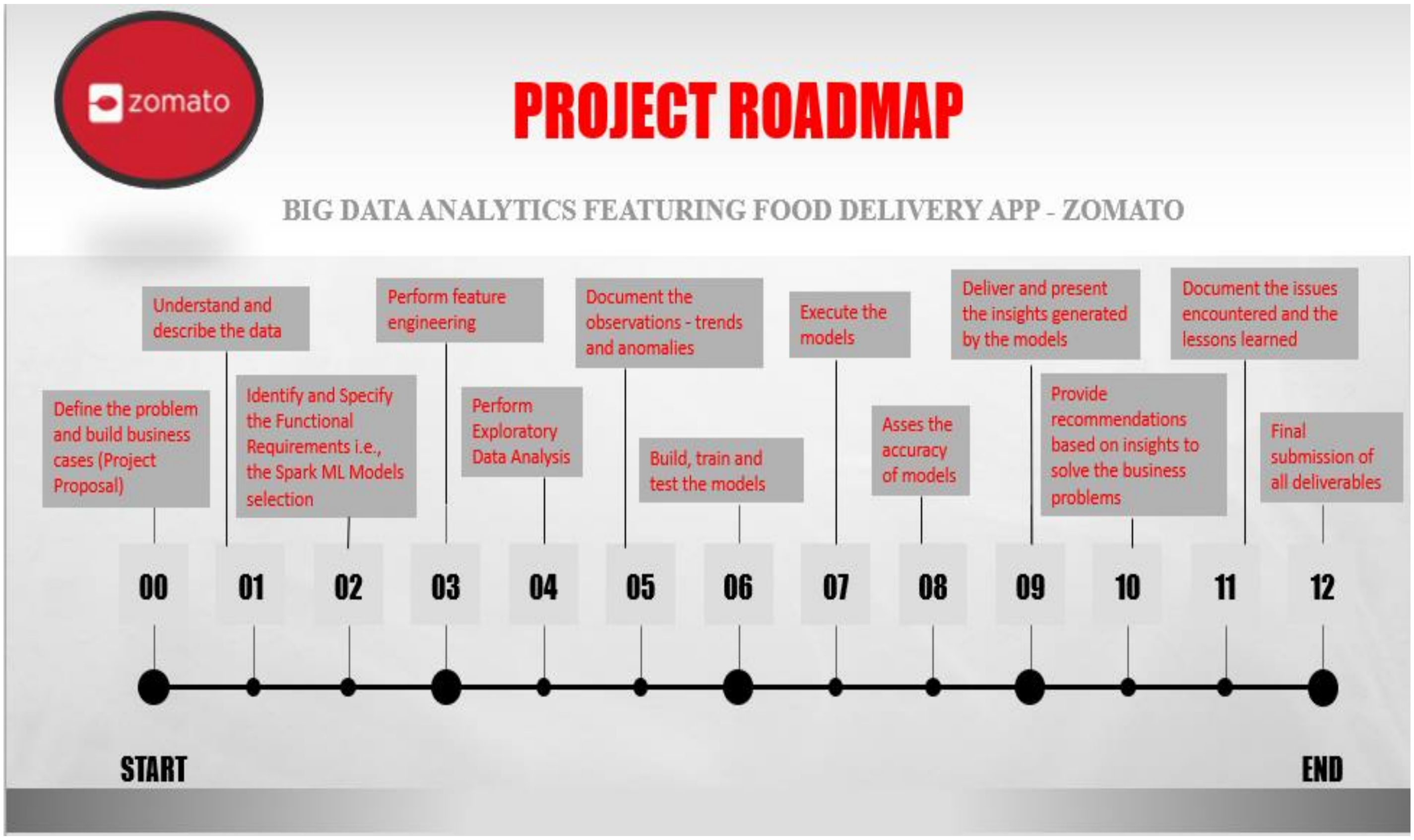
- **Neural Network Model –**
  Zomato can implement the algorithm that predicts the approximate cost for 2 people eating out using votes, cuisine, location, and ratingsby5 preferences to provide highly appreciated assistance to the users on budget.

ISSUES AND CONCERNS

- Data quality challenges due to missing data, inconsistent data, logic conflicts and duplicate data.
- Time-consuming implementation and execution of code due to the limitations pertaining to the lack of computational resources.
- Attaining co-ordination when working remotely.

PROJECT ROADMAP

The following timeline illustrates the workflow followed to achieve the objectives of this project:



PROJECT ROADMAP

BIG DATA ANALYTICS FEATURING FOOD DELIVERY APP - ZOMATO

00 Define the problem and build business cases (Project Proposal)

01 Understand and describe the data

02 Identify and Specify the Functional Requirements i.e., the Spark ML Models selection

03 Perform feature engineering

04 Perform Exploratory Data Analysis

05 Document the observations - trends and anomalies

06 Build, train and test the models

07 Execute the models

08 Asses the accuracy of models

09 Deliver and present the insights generated by the models

10 Provide recommendations based on insights to solve the business problems

11 Document the issues encountered and the lessons learned

12 Final submission of all deliverables

START

END

The following deliverables were produced as the project progressed:

- Project Proposal
- Source Code for Feature Engineering
- Exploratory Data Analysis Visualizations
- Source code for the Spark ML Models
- Project Presentation Slides
- Project Report