

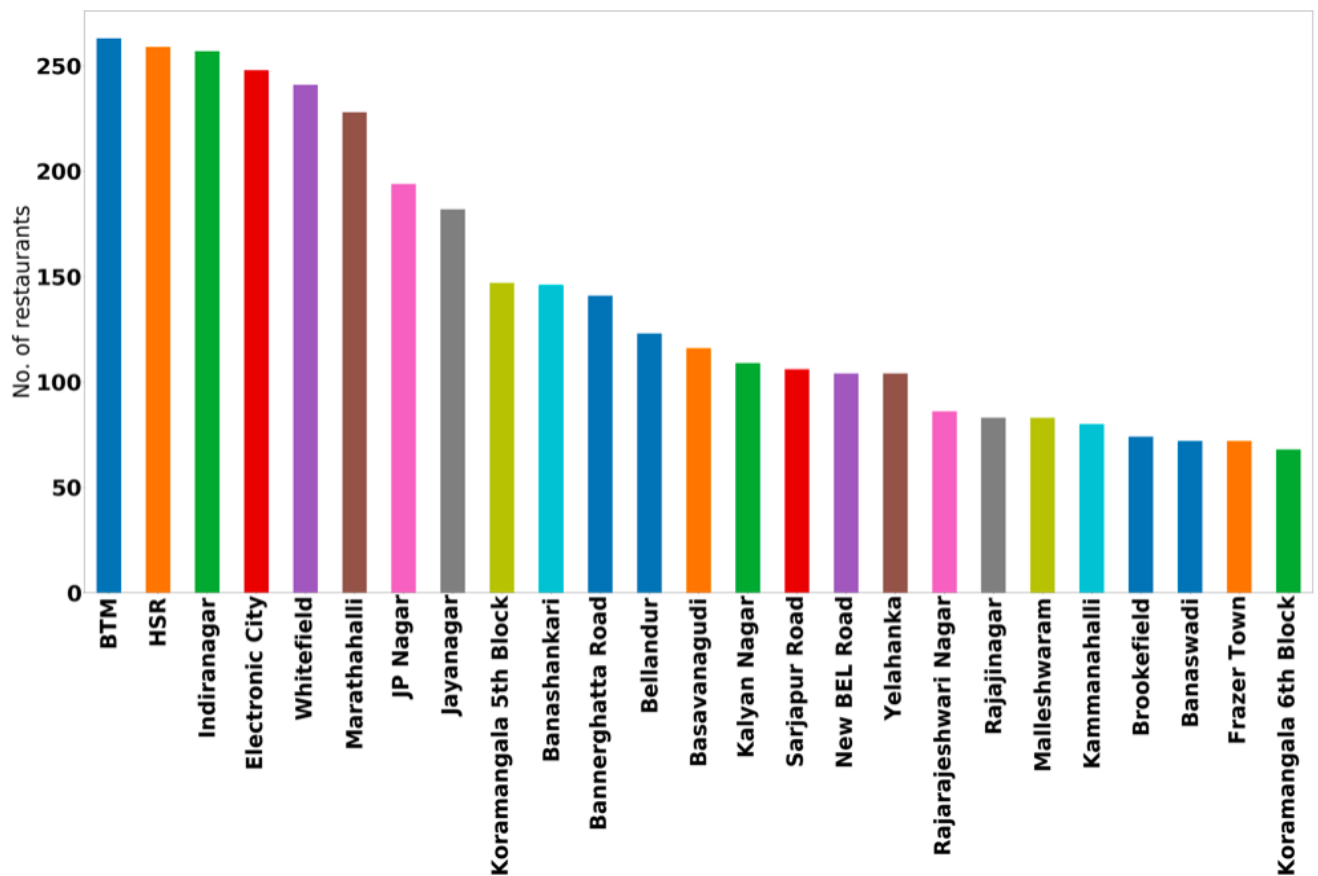
Zomato Restaurants Rating Predictor

IMT2016006

IMT2016089

IMT2016114

Introduction



Problem Statement: Predicting the ratings of restaurants available on Zomato's Bangalore page and find correlation between different features and ratings.

We aim to find correlation between different features and ratings. This will help restaurants as they will have better understanding of features that are important for customers. This will also give an insight on geolocation benefits for a restaurant.

Data Analysis and Prediction

For customers like us, this will give an insight as to where are most of the restaurants located and which restaurant has the features that matters the most to customers.

Scraping Dataset from Zomato and preprocessing it.

We scraped Zomato's Bangalore website to extract desired features that are available on a restaurant card on Zomato



We got all the features for total of 6000 restaurants and created a robust dataset containing all the information available on the card.

We processed the data by removing all the restaurants which are void of ratings as they are new with very less votes.

```
In [24]: df1 = pd.read_csv('zomato_res_final.csv', delimiter=',')
df1 = df1[df1.Rating != 0]
df1 = df1[df1.Cost != 0]
df1.head()
```

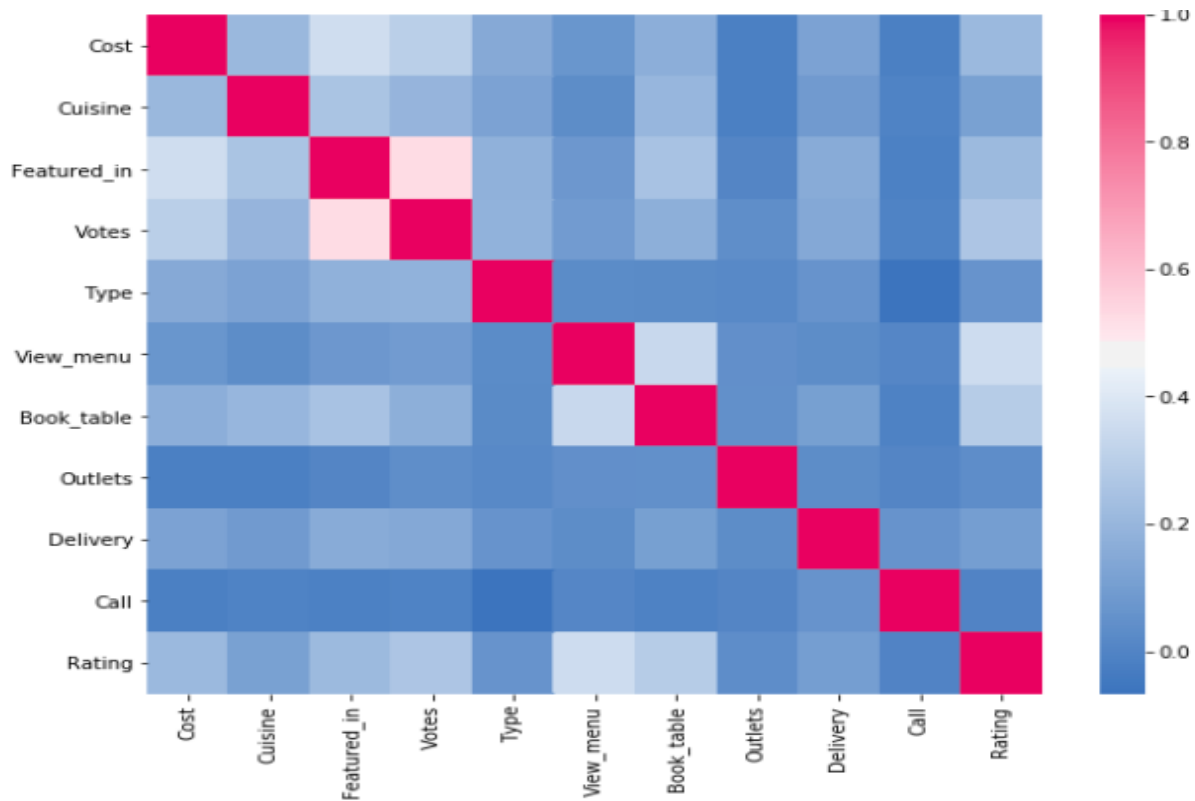
Out[24]:

	Name	Cost	Cuisine	Featured_in	Hours	Votes	Location	Type	View_menu	Book_table	Outlets	Delivery	Call	Rating
0	The Black Pearl	1500	4	2	12noon - 4pm, 6pm - 12midnight (Mon-Sun)	5882	Marathahalli	2	1	1	1	0	0	4.9
1	The 13th Floor	1700	3	3	4pm - 11pm (Mon-Thur), 4pm - 1am (Fri-Sat), 12noon - 1am (Sun)	4419	Barton Centre	1	1	1	1	0	1	4.3
2	Byg Brewski Brewing Company	1600	3	2	12:30pm - 1am (Mon-Sun)	3382	Hennur	1	1	1	1	0	1	4.6
3	Fly Dining	14000	1		5pm - 11pm (Mon-Sun)		Nagawara	1	0	0		0	1	
4	AB's - Absolute Barbecues	1600	4	2	12noon - 4pm, 6:30pm - 12midnight (Mon-Sun)	11160	Marathahalli	1	1	1	4	0	1	4.8

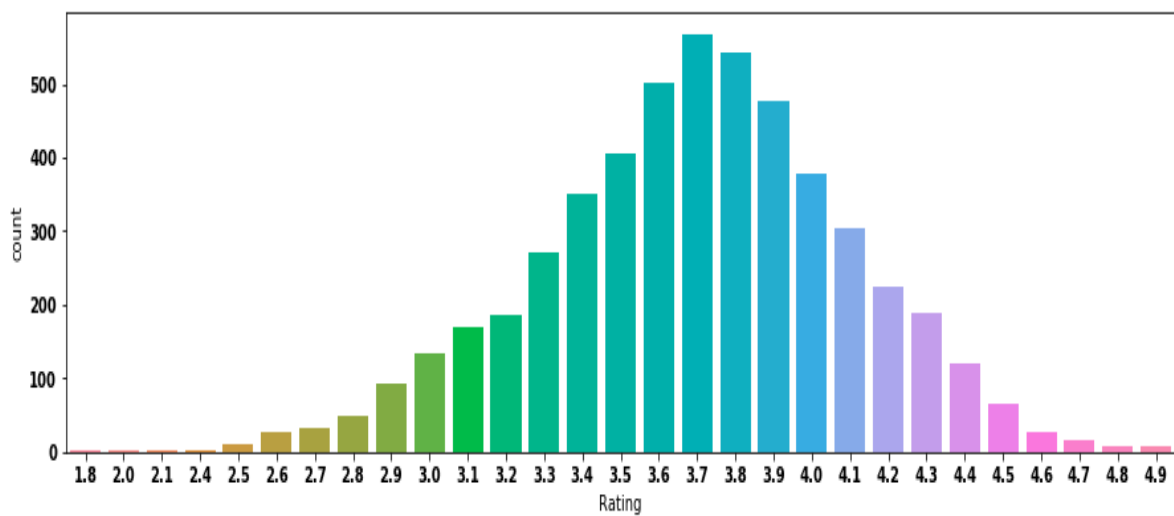
Data Analysis and Prediction

Data Analysis and Visualization:

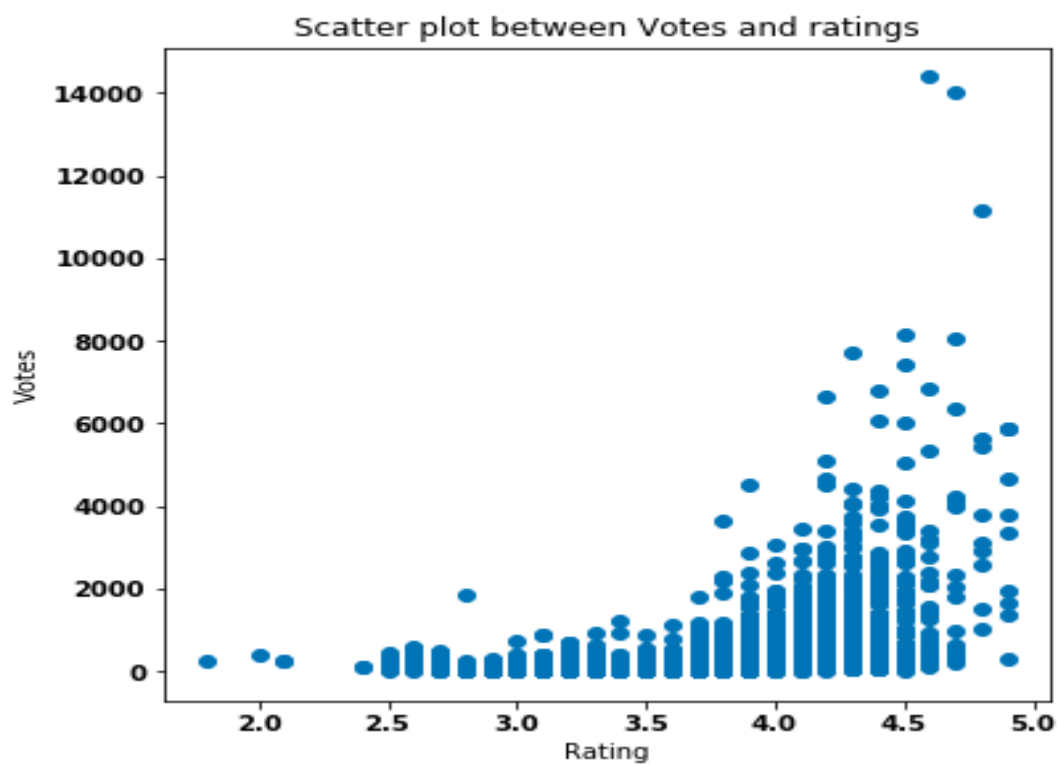
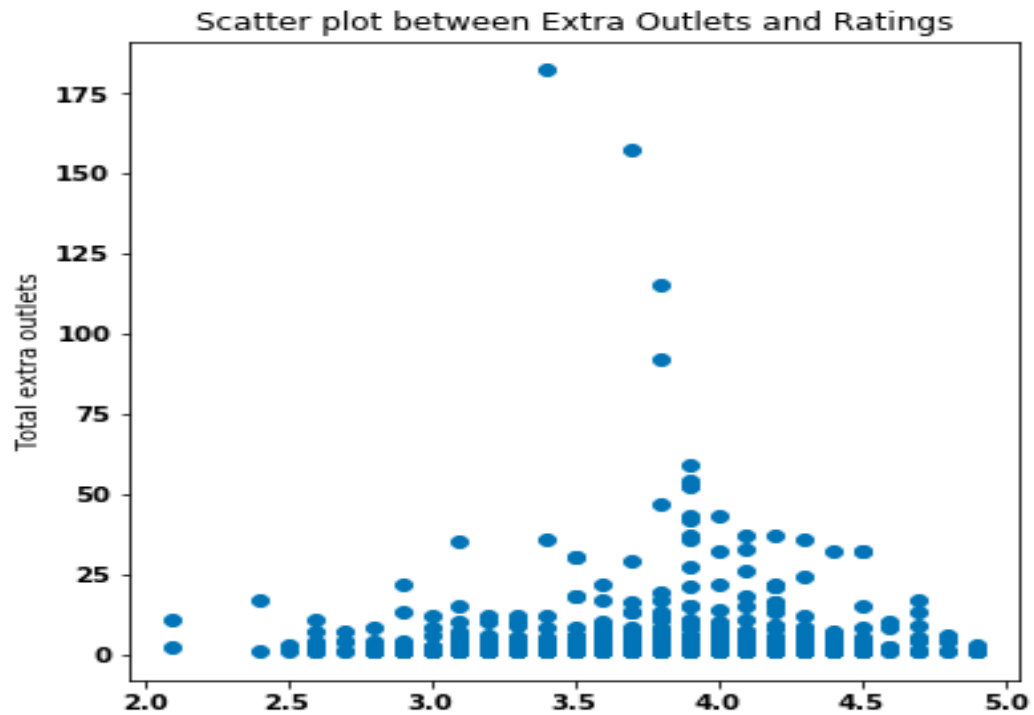
Correlation matrix:



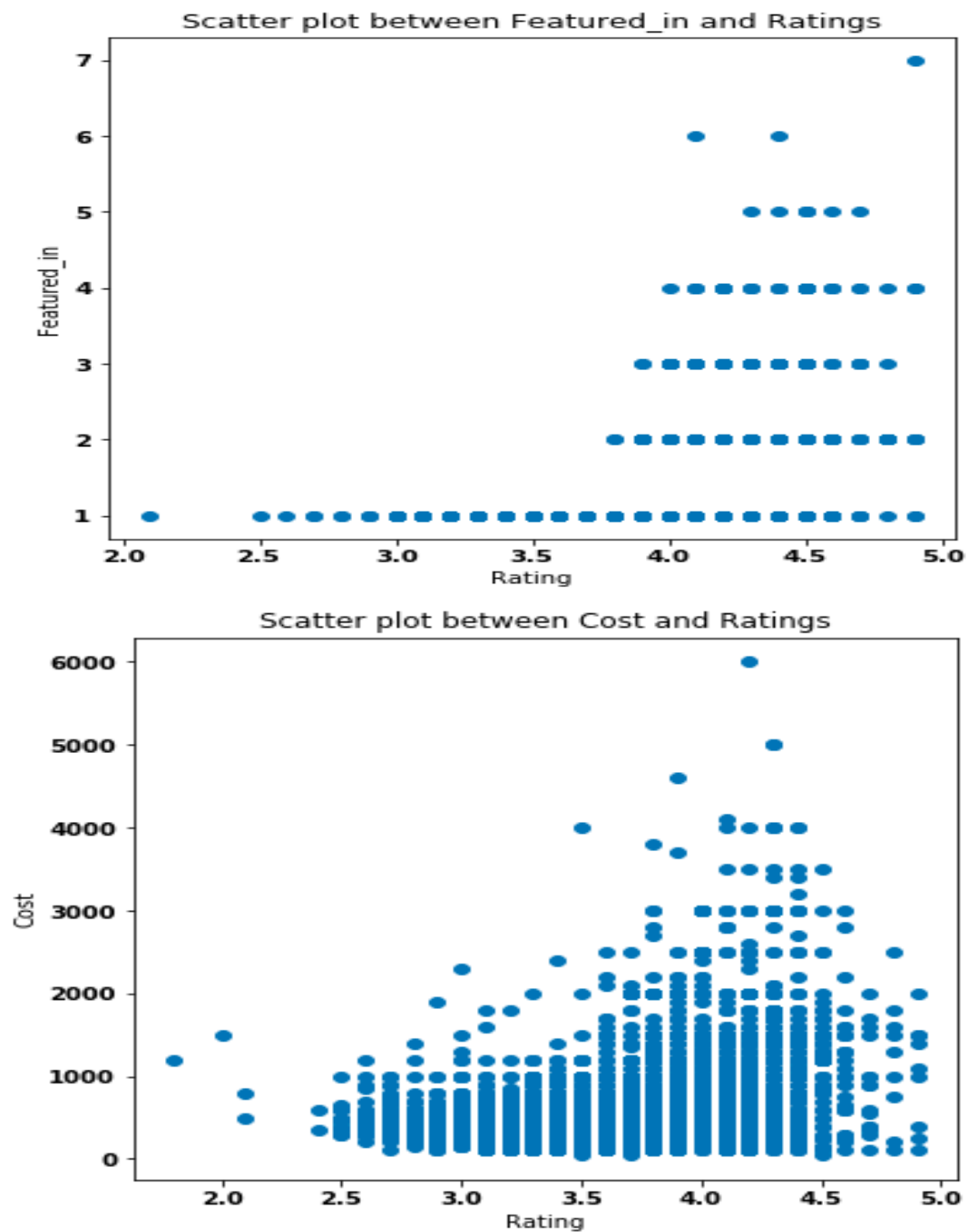
Most frequent ratings among restaurants:



Data Analysis and Prediction



Data Analysis and Prediction



We can see there is a strong correlation between Featured in categories and ratings, we can also see a strong correlation between votes and ratings as expected.

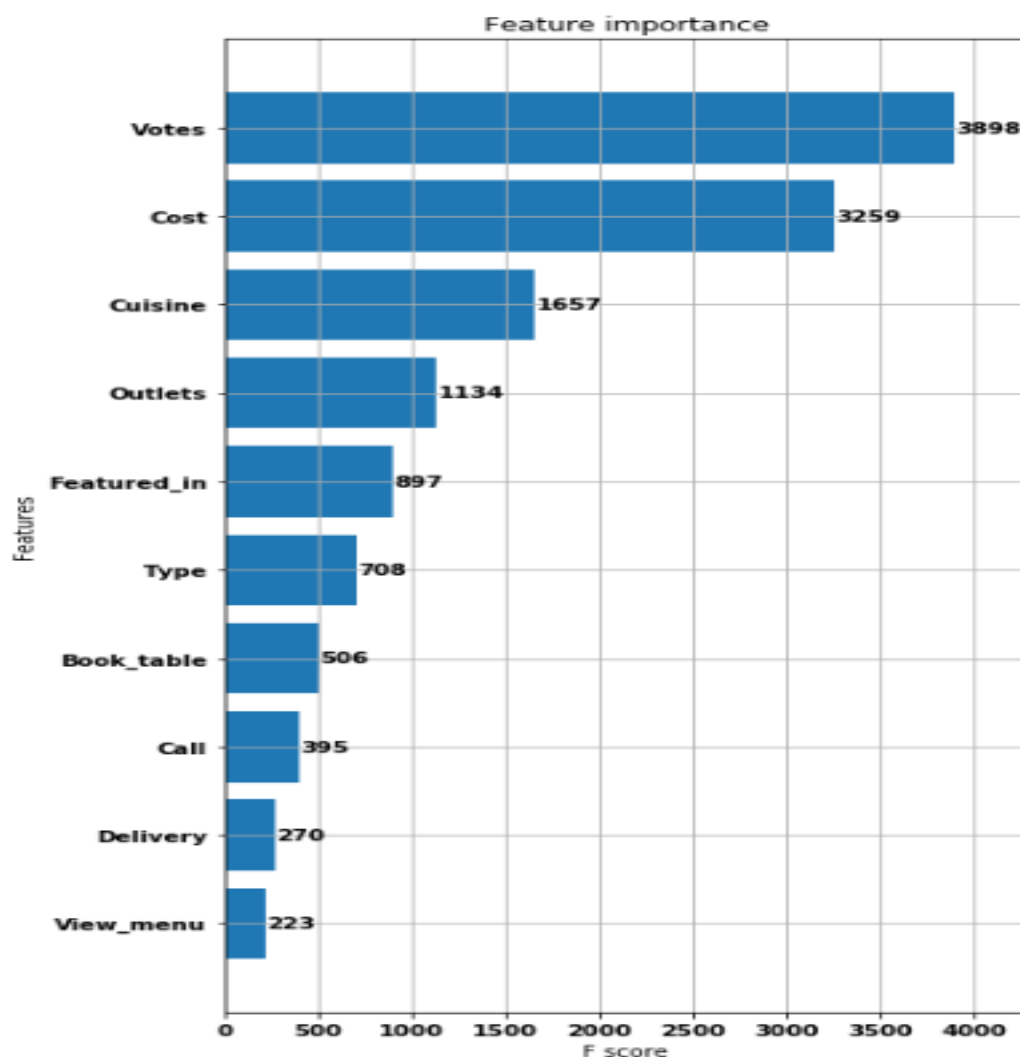
Data Analysis and Prediction

We will now use Xgboost to train on 70% dataset and test on 30%

From Xgboost we got repeated accuracy of 90-91%

```
error = rmspe(train_probs, X_test['Rating'].values)
print('root mean percentage error', error)
print('accuracy', (1 - error)*100)
test_probs = gbm.predict(xgb.DMatrix(test[features]))
submission = pd.DataFrame({"Name": test["Name"], "Rating": test_probs})
submission.to_csv("xgboost_kscrip_submission.csv", index=False)
```

```
Validating
('root mean percentage error', 0.097186100634352429)
('accuracy', 90.281389936564764)
```



We can see that most important features are: ***Votes, Cost, No. of cuisines offered, and no. of categories that a restaurant has been featured in.***