Final Report

Trend and Sentiment Analysis - Zomato Bangalore

Submitted By

Sriram.K

Richardson Jebasundar

Jagatheesh Pulavan

Rubina Petrishya D

Batch : DSE June 2019
Mentor : Ms.Ann Grace

Table of Contents

Abstract

Food delivery apps have become a rage, especially among millennials, owing to the convenience and ease involved in their usage.In 2018, the food delivery industry was estimated to value $82 billion in terms of gross revenue bookings and is set to more than double by 2025, backed by a cumulative growth rate of 14%. Online food delivery services rely on urban transportation to alleviate customers' burden of traveling in highly dense cities. As new business models, these services exploit user-generated content to promote collaborative consumption among its members. All businesses in the food industry generate data in the form of customer orders, delivery location, GPS, tweets, images, reviews, blogs, updates, etc. The data generated relates to average wait time, experience with the delivery, taste of food and availability As mobile trend grows rapidly, food delivery businesses are now combining unstructured big data with transactional and sales data for tendency and sentiment analysis – in order to leverage mobile app analytics that can help them build brand image and affinity towards customers, thereby increasing sales.

The project aims to gather insights over the operations of Zomato across the restaurants in Bangalore through the analysis of demographic preferences and its variance across locales through trend and sentiment analysis. It attempts to predict the rating of restaurants based on integral factors owing to its success . This would inadvertently help upcoming restaurants and existing franchises to optimize their operations in the future and fix existing issues to the fullest extent possible.

Acknowledgements

We are indebted to our Mentor Ms.Ann Grace for her time, valuable input and guidance. Her experience, support and structured thought process guided us to be on the right track towards completion of this project.Her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We are thankful to Ms. Meena Vardhini, Jr.Data Scientist, Acad-Ops , DSE Program for her unflinching and unabated help extended to us always.

We also thank all the course faculty of the DSE program for providing us with a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Industry Review

Worldwide, the market for food delivery stands at €83 billion, or 1 percent of the total food market and 4 percent of food sold through restaurants and fast-food chains. It has already matured in most countries, with an overall annual growth rate estimated at just 3.5 percent for the next five years. As new business models, these services exploit user-generated content to promote collaborative consumption among its members. .The fact, however, remains that despite the massive revenues and investments, food delivery companies still struggle with profitability, primarily due to their high rates of cash burn.

Predatory pricing is a widely used strategy to beat the competition where companies swallow a loss on the transaction by significantly subsidizing the cost of the meal. Logistical reliability and product quality are beyond their control as these services are contracted out to other parties. If a customer is unhappy with either of elements, the online food company has to bear the monetary penalty. Lastly, fraud across the value chain—whether through restaurant manipulated deliveries or cybersecurity loopholes related to digital payments—is emerging as an increasingly important factor affecting the business. As a result of these trends, several companies have been unable to withstand the heat and have exited the business. What has been disconcerting has been the pattern of failure.

Innovate or perish is a credo that will determine the future of online food delivery companies. One such innovative concept that has exploded onto the scene is that of a ‘dark’ or ‘cloud’ kitchens, essentially commercial facilities dedicated to serving online takeaway orders. The significantly lower cost of capital investment required for this setup, as compared to a full blown restaurant facility, enables food to be offered at cheaper rates. Online food delivery companies are partnering with third party businesses that build and/or run these establishments and are even investing in it themselves to prepare food under different brand names. For example, Swiggy in India runs its own cloud kitchen service called The Bowl Company that prepares meals for Swiggy’s delivery business. Another interesting growth strategy that online food delivery companies are exploring is to take a step backwards on the food supply chain. They are integrating themselves into the delivery chain at the raw material stage, delivering these goods from farmers or producers to warehouses, while also distributing supplies from these warehouses to restaurants and other food preparation businesses. Zomato, for instance, is aggressively setting up gigantic warehouses all over India to store fresh produce, thereby ensuring that not only are quality standards maintained but also that their food sourcing costs are reduced. The end result is a low cost, high quality meal for the customer since Zomato is the intermediary in all aspects of the meal, from sourcing the ingredients right through to cooking the meal in a cloud kitchen. Autonomous Technology and Big Data have become key factors in the race to differentiate and optimize one’s operation over the competition.

Literature Review

1) Evaluation of collaborative consumption of food delivery services through web mining techniques, Juan Correa et.al, Jan 2019

The study aimed to evaluate the impact of traffic conditions (through the use of Google Maps API) on key performance indicators of online food delivery services (through the use of web scraping techniques to retrieve customer's ratings and the physical location of restaurants as provided by Facebook). From a collection of 19,934 possible routes between the physical location of 787 online providers and 4296 customers in Bogotá city, it was found that traffic conditions exerted no practical effects on transactions volume and delivery time fulfillment, even though early deliveries showed a mild association with the number of comments provided by customers after receiving their orders at home.

2) Consumer Preference and Attitude Regarding Online Food Products in Hanoi, Vietnam, Anh Kim Dang et.al (May 14, 2018)

This study aimed to examine: 1) how the Internet has changed consumers food-buying behavior and identify its associated factors; 2) consumers' concern about food safety information of online food products. A cross-sectional study was performed from October to December 2015 in Hanoi—a Vietnamese epicenter of food service. 1736 customers were randomly chosen from food establishments of 176 communes. Data was collected through face-to-face interviews using structured questionnaires. The majority of participants reported using the Internet to search for food products (81.3%). The most crucial factors influencing food purchases through the Internet were convenience (69.1%) and price (59.3%). Only one-third of participants selected products based on accurate evidence about food safety certification or food origin. The majority of participants were concerned about the expiration date (51.6%), while brand (9.8%) and food licensing information (11.3%) were often neglected. People who were either female or highly influenced by online relationships and those having difficulty in doing usual activities were more likely to look for online food products.

Data Dictionary

url	URL of the restaurant in the zomato site.
address	Address of the restaurant in Bengaluru
name	Name of the restaurant
online_order	States the availability of online ordering in the restaurant
book_table	States the availability of table booking in the restaurant
rate	The overall rating of the restaurant out of 5
votes	Total number of ratings for the restaurant as of 15th March 2019
phone	Phone number of the restaurant
location	Neighborhood in which the restaurant is located
rest_type	Restaurant type
dish_liked	Dishes liked by the customers in the restaurant
cuisines	Cuisines offered by the restaurant
approx_cost(for two people)	Approximate cost for meal for two people
reviews_list	List of tuples containing reviews for the restaurant, each tuple consists of two values, rating and review by the customer
menu_item	List of menus available in the restaurant
listed_in(type)	Type of dining the restaurant specializes in
listed_in(city)	The cluster of neighbourhoods to which the restaurants belongs to

Variable Categorization

Dataset info

Number of variables	17
Number of observations	51717
Missing cells	37700 (4.3%)
Duplicate rows	0 (0.0%)
Total size in memory	6.7 MiB
Average record size in memory	136.0 B

Variables types

Numeric	1
Categorical	13
Boolean	2
Date	0
URL	1
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

`address` has a high cardinality: 11495 distinct values	Warning
`approx_cost(for_two_people)` has a high cardinality: 71 distinct values	Warning
`cuisines` has a high cardinality: 2724 distinct values	Warning
`dish_liked` has a high cardinality: 5272 distinct values	Warning
`dish_liked` has 28078 (54.3%) missing values	Missing
`location` has a high cardinality: 94 distinct values	Warning
`menu_item` has a high cardinality: 9098 distinct values	Warning
`name` has a high cardinality: 8792 distinct values	Warning
`phone` has a high cardinality: 14927 distinct values	Warning
`phone` has 1208 (2.3%) missing values	Missing
`rate` has a high cardinality: 65 distinct values	Warning
`rate` has 7775 (15.0%) missing values	Missing
`rest_type` has a high cardinality: 94 distinct values	Warning
`reviews_list` has a high cardinality: 22513 distinct values	Warning
`votes` has 10027 (19.4%) zeros	Zeros

Data Preprocessing : Data Cleaning

Features/stages	Stage-1	Stage-2	Stage-3	Stage-4	Stage-5	Stage-6	Stage-7
Rate	7775	7760	10027	8522	0	0	0
Phone	1208	0	0	0	0	0	0
Location	21	21	21	21	0	0	0
Rest_type	227	227	227	227	147	0	0
Dish_liked	28078	28027	28027	28027	19394	19307	0
Cuisines	45	45	45	45	8	0	0
Cost	346	345	345	345	0	0	0

Explanation of Stages:

Stage-1 – The values represent the total number of null values in each column.

Stage-2 – drop_duplicates code is used to drop the duplicates rows in the dataset.

Stage-3 – Rate column has values like '-' and 'New' which needs to be cleaned. So in this stage they are converted to np.nan. that is why above above rate column has increased null values.

Stage-4 – Rate column null value is handled by using reviews_list column. Extracting information from the tuples containing individual rating by the customers.

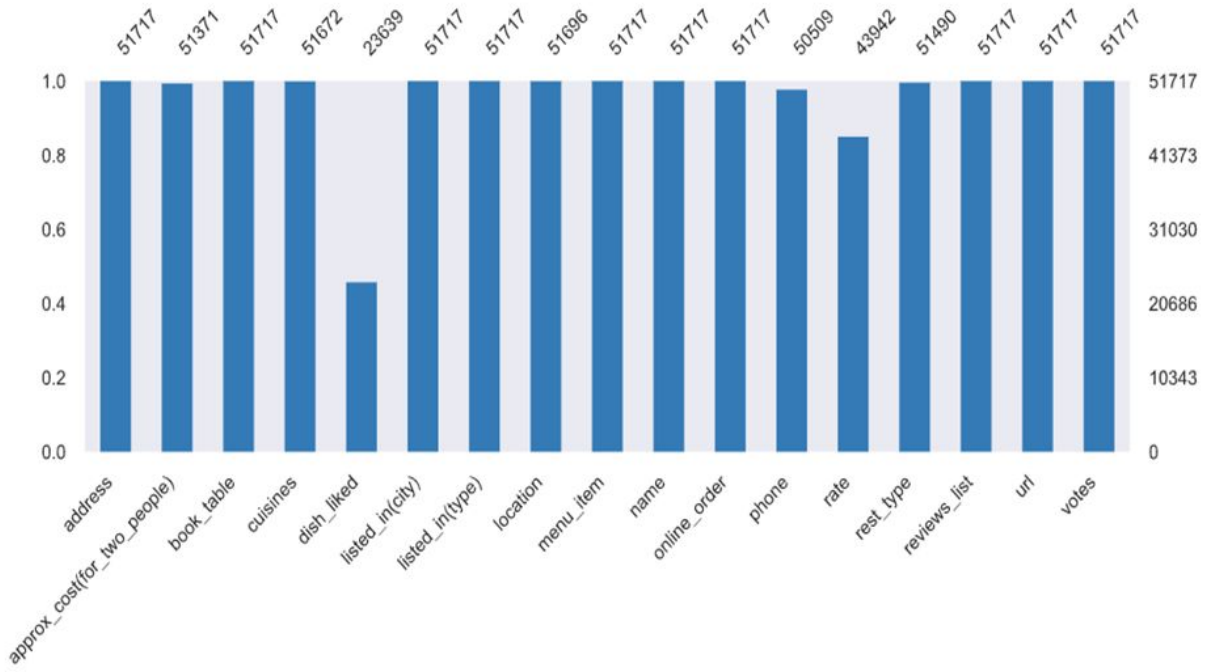
Stage-5 – The remaining null values in rate column is dropped. Technically we are dropping 8522 records. This is a huge reduction which is not appreciable, but due to this the location column is cleaned, rest_type column is reduced. The main reason behind dropping is the dish_liked column had significant reduction.

Stage-6 – The remaining null value in rest_type and cuisines is also dropped. Other imputation technique like groupby, bfill, ffill we tried but it saved only 22 records. So comparing to the whole data this is negligible.

Stage-7 – The huge null value in dish_liked column is handled by extraction information from the reviews_list column and intersecting it with the disk_liked list created using the available records.

Finally all the null values are handled the total data loss is 17.33% from the original data size of 51,654 to 42,699.

Null value checks :



Irrelevant Features :

>By domain knowledge,we deem the following features as irrelevant.

1. Url
2. Phone number

Data Transformation

The rate was initially a string variable of the sort ‘3.4/5’.By slicing and type casting the rate variable,we made it ready for further analysis.For restaurants having no rating,we mined the reviews of all customers and extracted each numerical score assigned by each customer,thus imputing the average of all scores for each restaurant.

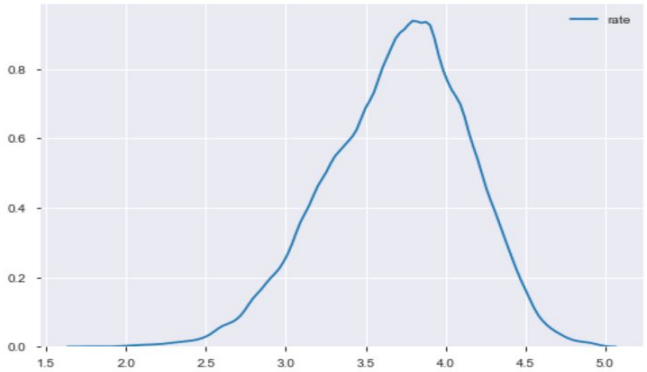
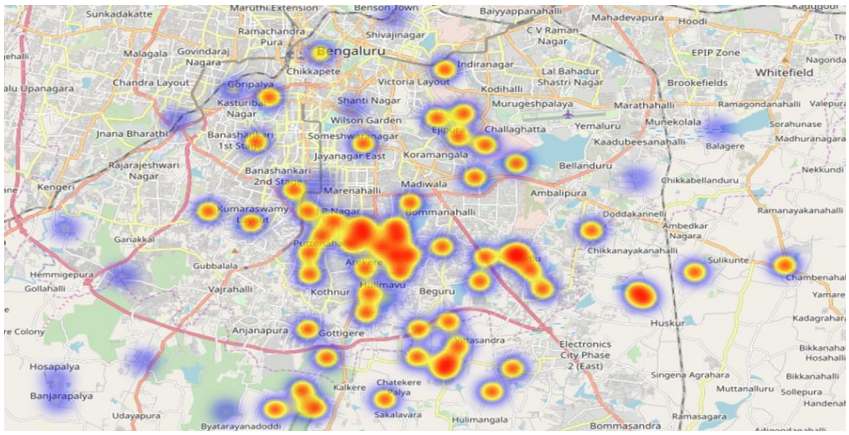


Figure 1 : Distribution of Rating

Alternate sources of data

In order to visualize the location of various restaurants in a map, we have scraped two features from the web, namely longitude and latitude.Using these new features added,we were able to successfully distinguish restaurants on a geographical scale.

Figure 2 : Basemap of restaurants across Bengaluru



Statistical tests performed

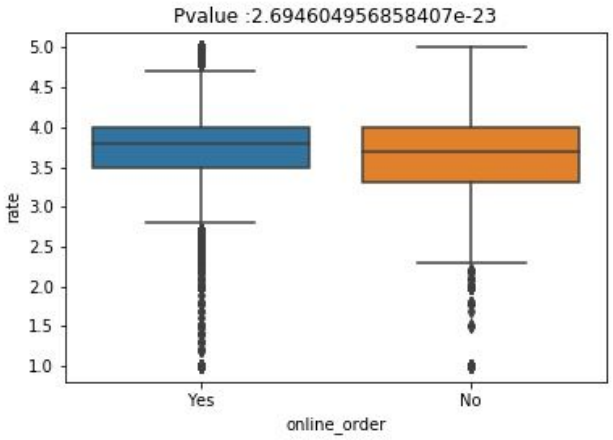
Here, we need to decide which out of the two categorical variables, online_order and book_table plays a significant role towards our target variable. The null and alternative hypothesis for these tests are given below. In both cases, we assume a 95% confidence interval.

H0 : The variable online_order/book_table is not significant

Ha : The variable online_order/book_table is significant

Online_order :

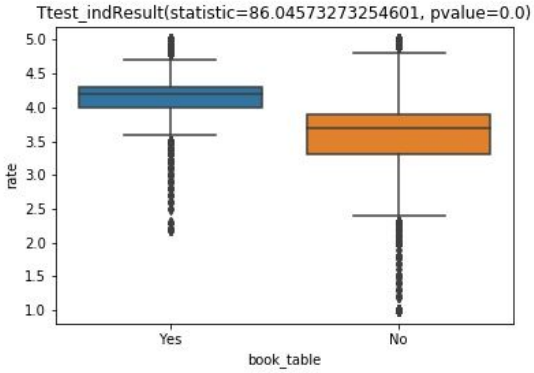
Figure 2 : Boxplot of online_order vs rate



Here, the p-value obtained is greater than 0.05, and hence we cannot reject the null hypothesis, and thus we deem that online_order does not have a significant effect on rating.

Book_table :

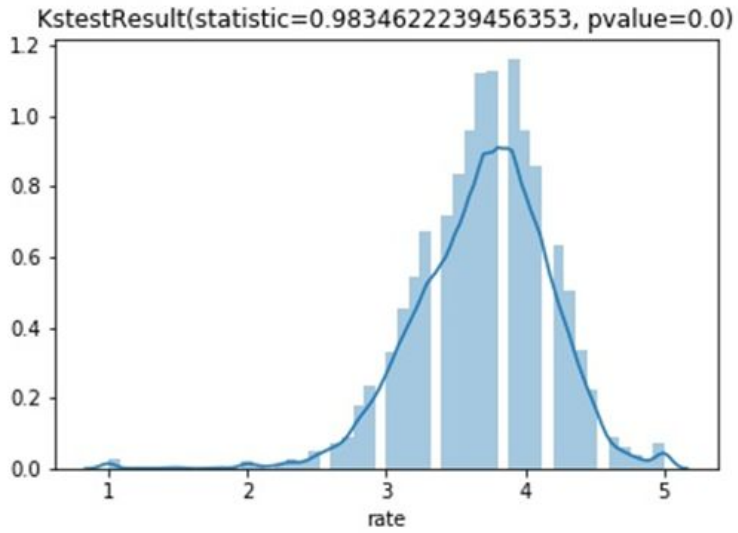
Figure 3: Boxplot of book_table vs rate



Here, the p-value obtained is less than 0.05, and hence we can reject the null hypothesis, and thus we deem that book_table has a significant effect on rating

Shapiro-Wilk Test for Normality :

Figure 4 : Distribution plot for rating



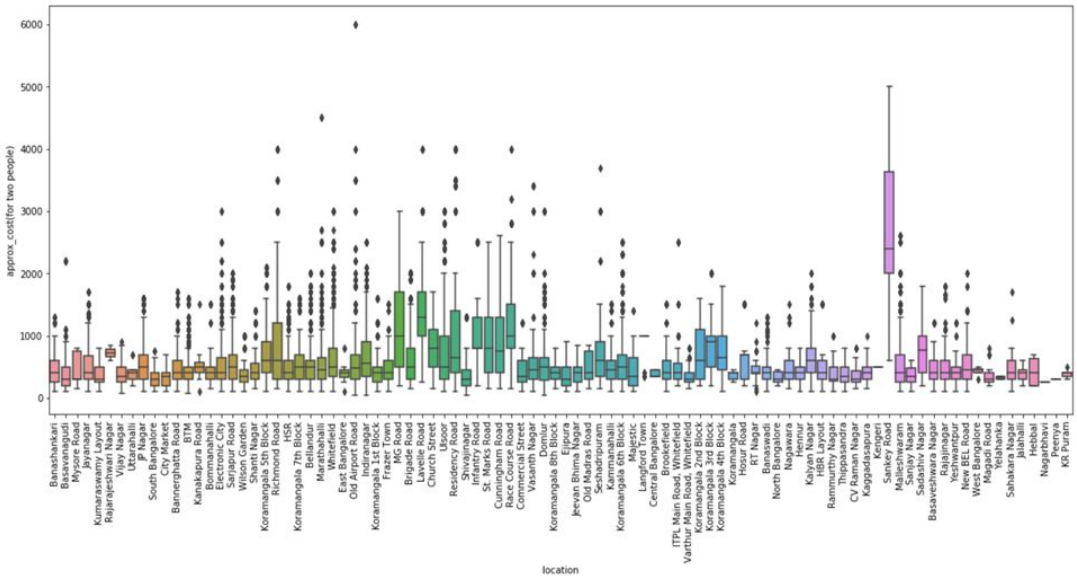
H0 : The rate is normally distributed

Ha : The rate is not normally distributed

The p-value obtained is less than 0.05, therefore we reject the null hypothesis.
The rate is not normally distributed.

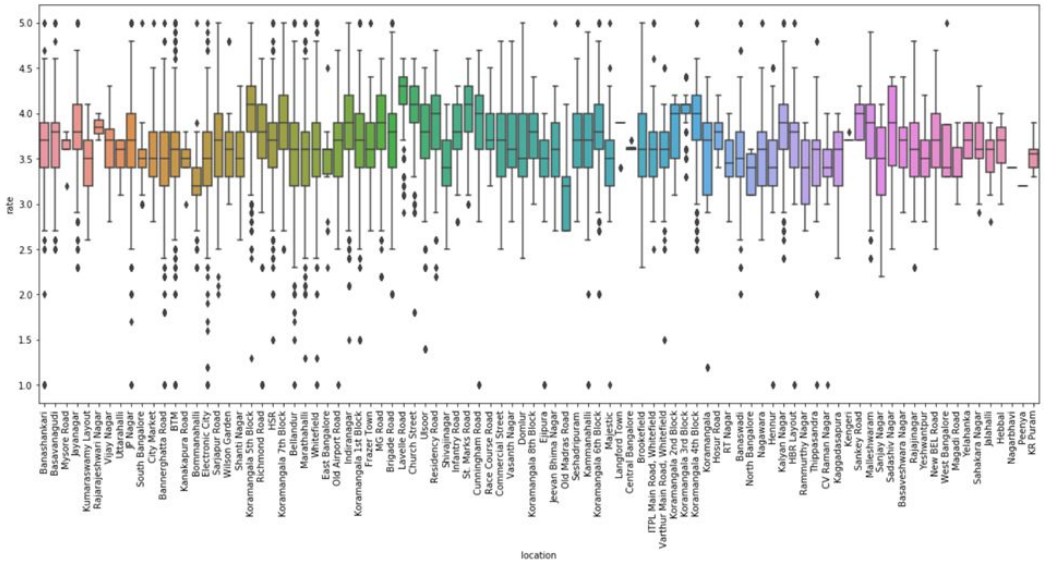
Exploratory Data Analysis

Figure 5: Location Versus Approximate cost for two people



#

Figure 6 : Location vs rating



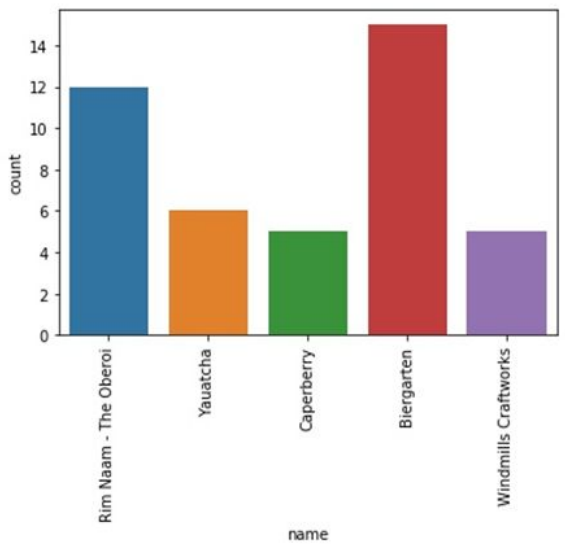
By plotting location versus cost and rating, one can understand if a location's overall average cost and rating is influenced by select restaurants or if all restaurants contribute equally towards the target variable.

Figure 7 : Good value for money

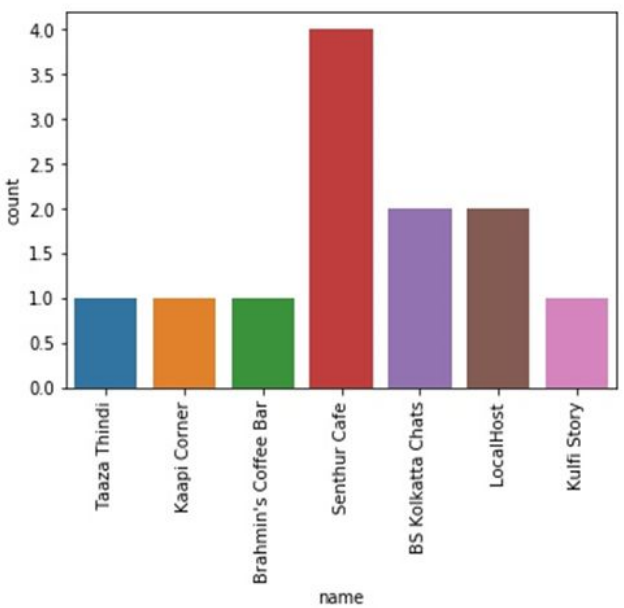
Most expensive dishes

Most economic dishes

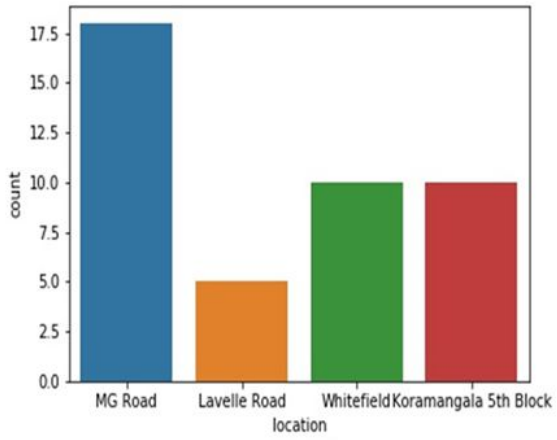
Most expensive restaurants



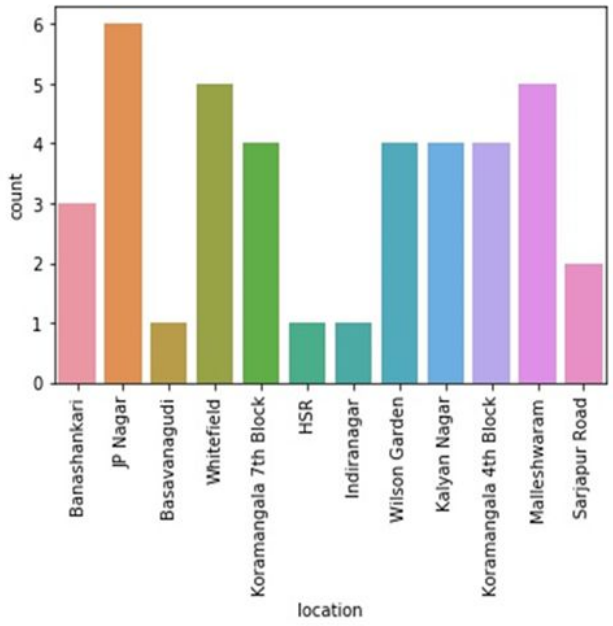
Most economic restaurants



Most expensive locations to dine at



Most economic locations to dine at



The above set of graphs showcase the expensive versus economic choices of dishes, restaurants, and location progressively. From this, one can understand how to get a good bang for the buck

Location-wise Preferences :

There are various factors that influence a customer's inclination towards a restaurant. Satisfying customer's palettes is a daunting task, judging from the diverse palette profiles of people, each having their own preferences. In the modern era, delectable food no longer makes the cut. This study highlights the importance of analysing the demographics of regions to zero in on the customers' specific needs.

Table 1 : Common customer preferences across all locations

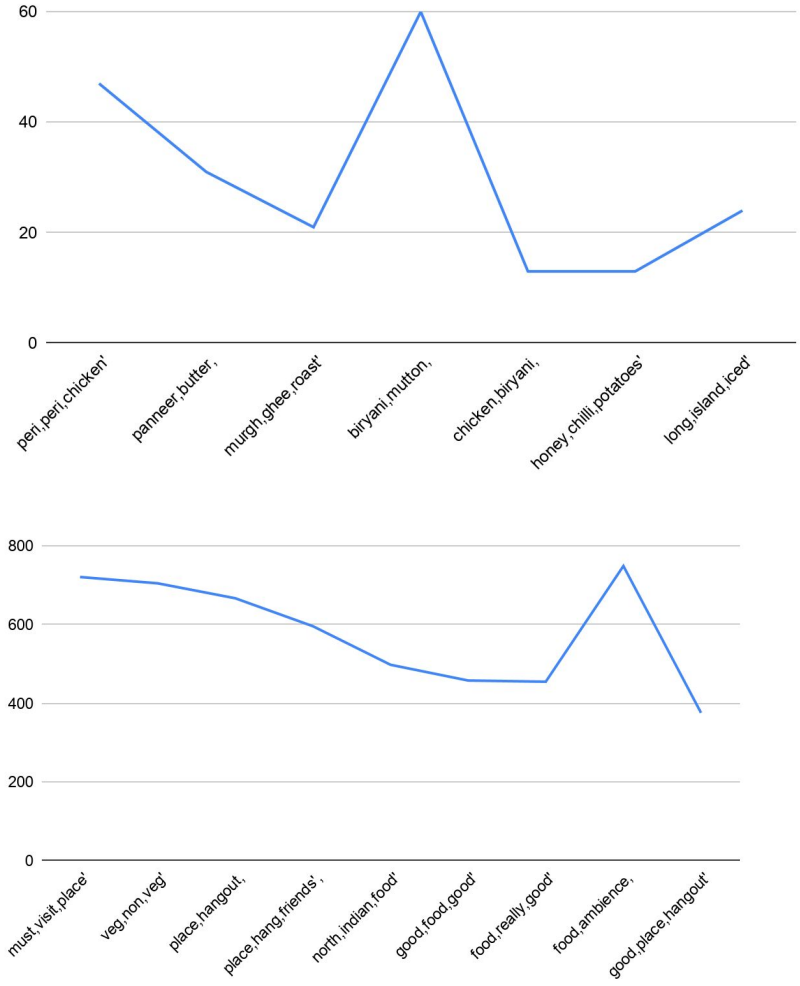
Feature	Preference
Unique restaurant attributes	Ambience Value for money Customer service Distinctive restaurant themes
Prominent cuisines	North Indian Chinese South Indian
Restaurant types	Casual dining Quick bites Food trucks(across central and east bangalore)
Favorite dishes	Biryani Paneer butter masala Peri- peri chicken

The text data of several districts of Bangalore have a rather interesting story embedded within them. Narrating this story would give restaurants some idea on how to enhance their business. BTM(Byrasandra, Tavarekere and Madiwala),Banashankari, Church Street, MG Road, Residency Road and Old Airport Road are some of the districts with contrasting customer preferences which will be analysed in detail.

BTM (Byrasandra, Tavarekere and Madiwala):

Upon inspecting the trigrams from the figure given below, we can understand that almost all restaurants here are flourishing in their trade. Verified online sources indicate that BTM is a highly sought after residential area with many educational institutions being erected in the midst of this town and the fact that BTM is in close proximity to other prominent districts such as Koramangala, HSR layout,Bannerghatta Road, J P Nagar and Jayanagar. Students pursuing courses in the institutions here are the food critics of the 20th century. By posting photos of their meals on various social media platforms such as instagram, they passively promote the restaurants thus increasing their popularity. Hence, setting up shop here that serves good north indian food is definitely a good start.

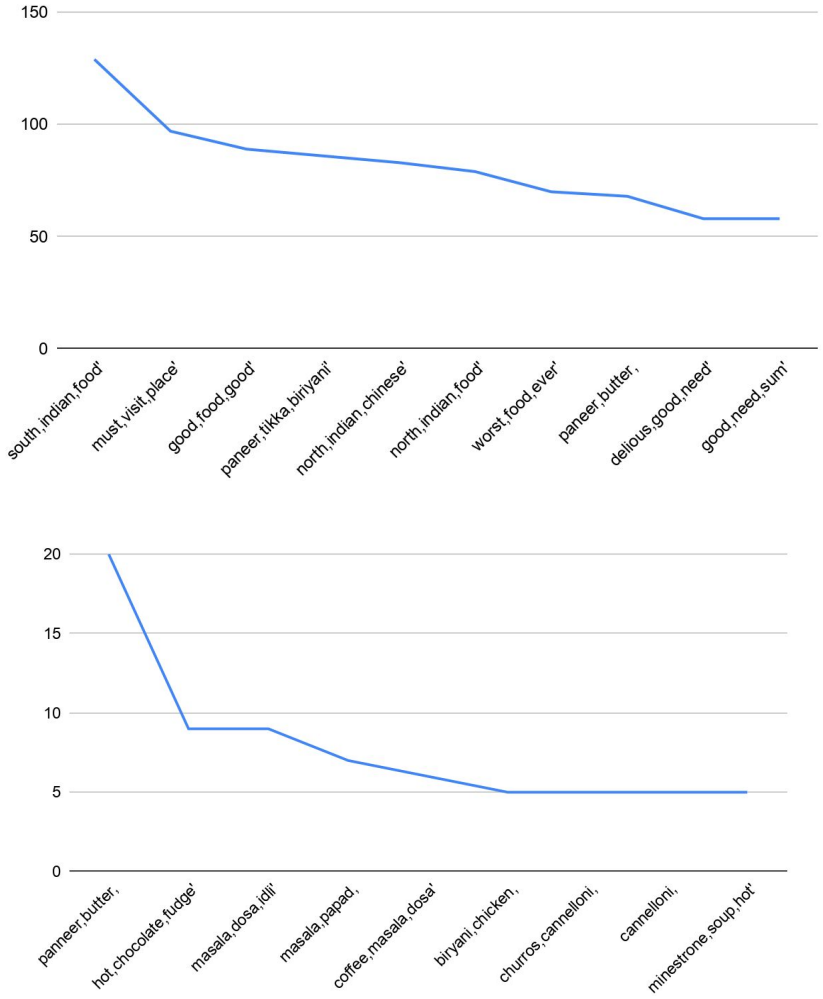
Figure 8 : Trigrams for BTM



Banashankari:

Banashankari has a good number of restaurants with positive reviews. In fact almost all of the restaurants have high ratings. Yet, in the figure below, the phrase ‘worst food ever’ is common in most customer reviews indicating extreme disappointment . On further analysis of some of the scathing reviews, and the total number of reviews for each restaurant, we found the number to be really low. One such restaurant, Udupi Ruchi Grand, has a rating of 4.2 ,yet people have given scathing reviews.Obviously, this is due to the number of reviews for the restaurant being 60. Thus, a restaurant can be considered good only if it is highly rated by a large number of people.

Figure 9 : Trigrams for Banashankari



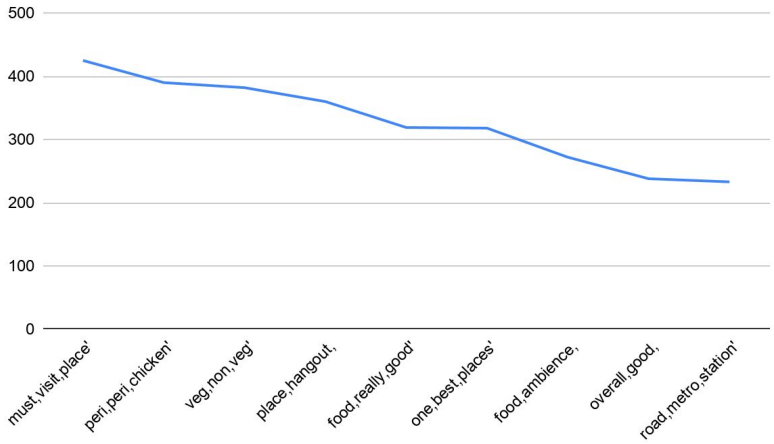
Church Street, MG Road, Residency Road :

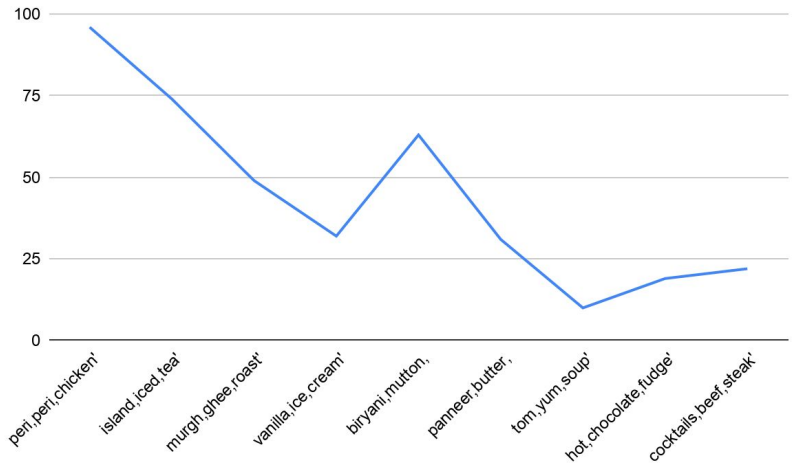
To understand why the restaurants situated in the above three localities fare really well, we have calculated the ratio of ordering online to dine in for each location. A typical value for most locations across bangalore falls in the range of 1.8 to 2.8. Yet the mentioned localities have a ratio of only 1.09,1.08 and 1.07 respectively. Upon analysing the demographics, we have found that these places are known as Bangalore's hotspots for boosting revenue. With posh boutiques and hip eateries strewn all over the place, these areas are also iconic spots for celebrating auspicious occasions like New year. The target audience here is celebrities, critics and other social media influencers who keep restaurateurs on the tip of their toes.

Church Street Social Interiors

The above set of pictures were taken from the same restaurant cum bar named Church Street Social. The stark contrast between both rooms in the same restaurant is what makes it stand out. A pub is never usually an ideal place for anyone seeking peace and quiet. Yet, they have separate workspaces, presumably for freelancers.

Figure 10 : Top trigrams mined from food reviews for CS,MGR,RR





From the above graph we can see that from the trigrams ‘place,hangout,friends’ and ‘one,best,places’ to ‘food,really,good’, there is a sharp decline. This tells us that customers look for a unique restaurant theme and are not really picky when it comes to food.

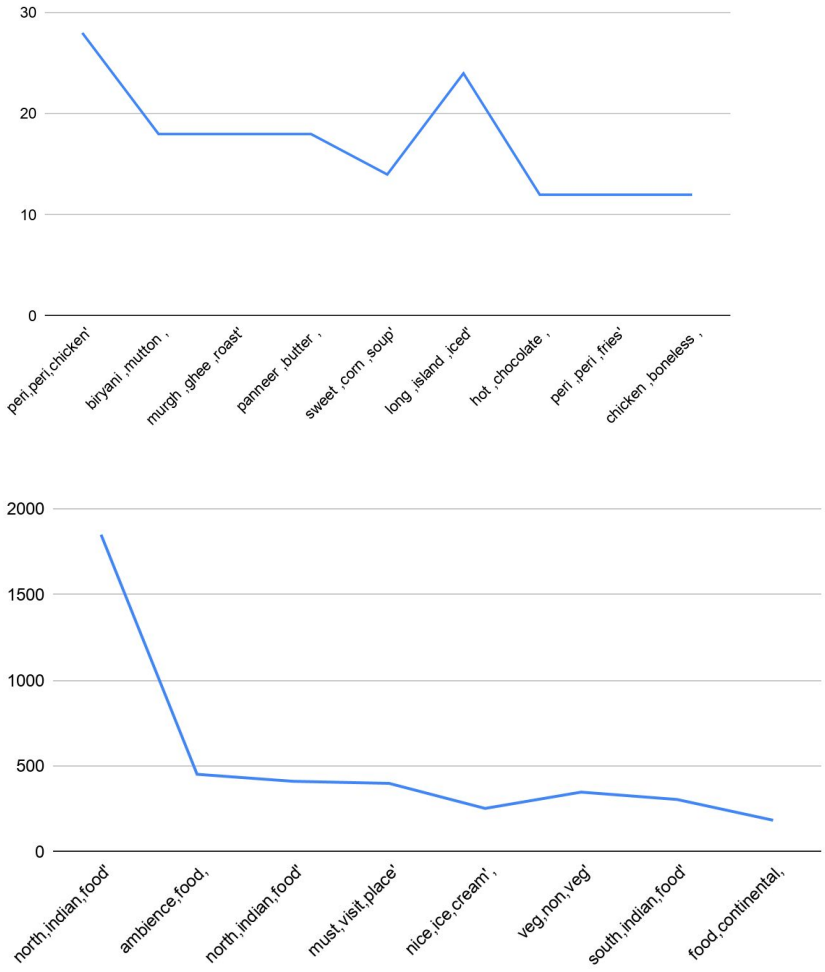
Old Airport Road :

Table 2 : Top 3 restaurants of Old Airport Road

Restaurant Name	Location	Rating	Votes	Approx.cost for two
The Pizza Bakery	Old Airport Road	4.8	1,781	₹1,200
Big Pitcher	Old Airport Road	4.6	9,300	₹1,800
Smoke House Deli	Old Airport Road	4.6	5,446	₹1,600

Old Airport Road is neither a residential nor a commercial spot. Being an area focused on real estate, it attracts wealthy businessmen and their families. These people look for a luxurious place to stay along with an in-built restaurant that serves high quality food. They do not mind paying an exorbitant amount of money as evident from the table below, depicting the top three hotels with high ratings, high votes and extremely high costs. Restaurateurs seeking to set shop here must have enough capital investment to start a luxury hotel as the target audience here do not just want a place to eat , but a cozy place to spend the night as well.

Figure 11 : Top trigrams mined from food reviews for Old Airport Road



Base Model

The base model chosen was the Decision Tree.

Classification Report for Decision Tree :

	precision	recall	f1-score	support
0	0.85	0.82	0.84	1244
1	0.96	0.97	0.97	9594
2	0.96	0.94	0.95	3256
accuracy			0.95	14094
macro avg	0.93	0.91	0.92	14094
weighted avg	0.95	0.95	0.95	14094

Since this is an imbalanced dataset for classification, we look at f1-score(micro avg which is same as accuracy in this case). In the base model, text based reviews were ignored and only 3 bins were used.

Model performance measures used for evaluating models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is an imbalanced dataset.. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

Accuracy :

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.
Considering the response rate (conversion rate) of our dataset which is ~16%, accuracy is not a valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 84%. Hence other model performance measures need to be evaluated.

Sensitivity or recall :

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

Specificity :

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

Precision :

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions. If precision is low, it implies that the model has a lot of false positives and vice-versa.

F1-Score :

F1 is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

Model Tuning and Finalization

We perform sentiment analysis on the reviews in order to ascertain the general sentiment of the reviewer and feed the input to the model. This is done in order to increase the robustness of the model and thereby generalize it. The process is as follows :

1) **Tokenization** is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

2) **Lemmatization** is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

3) Removal of **Stopwords**, Words with one letter, empty tokens and punctuation.

We then use the **VADER**(Valence Aware Dictionary and Sentiment Reasoner) Sentiment Intensity Analyzer for sentiment analysis. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

It provides a Positive Score, Negative Score, Neutral Score and a Compound score. The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1(most extreme positive).

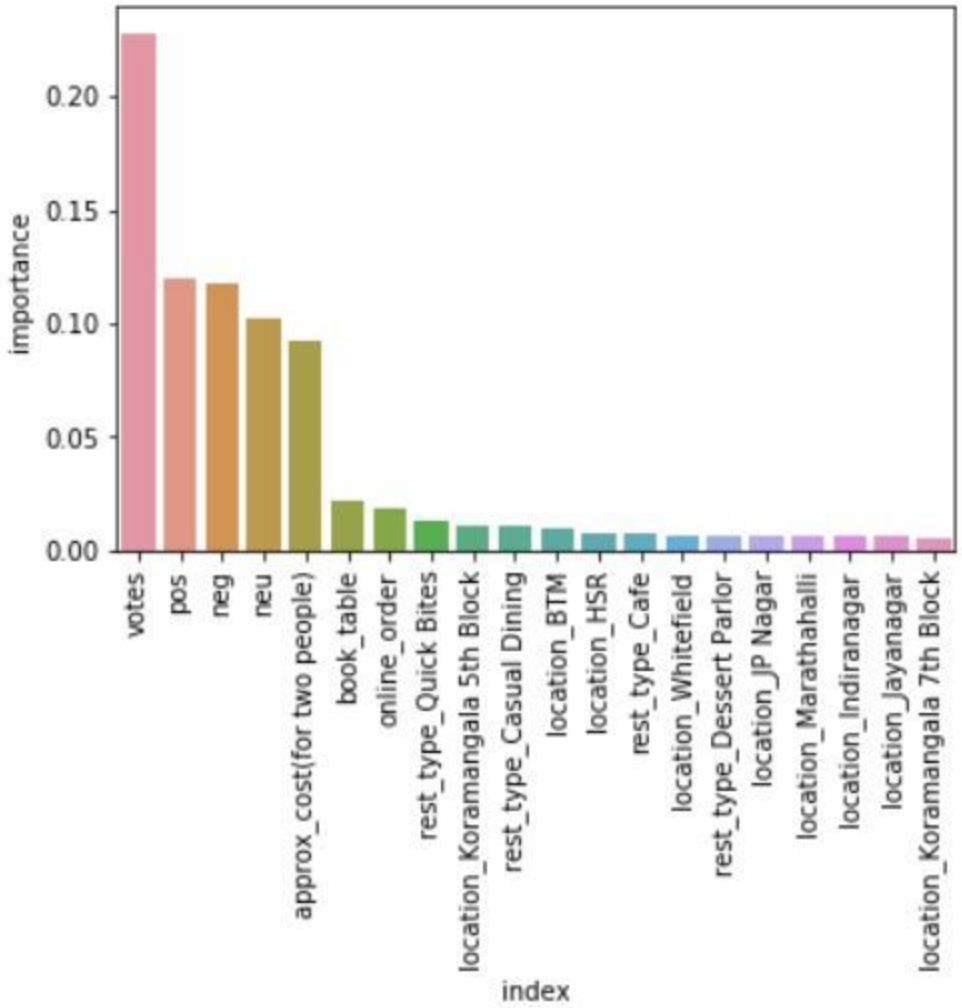
Positive Sentiment - Compound Score > = 0.05
Neutral Sentiment - Compound Score > = -0.05 and < 0.05
Negative Sentiment - Compound Score < = -0.05

The final model also uses 5 bins instead of the 3 in the base model.

We thereby create 4 new columns.

neg	Negative VADER score
neu	Neutral VADER score
pos	Positive VADER score
compound	Compound VADER score

Figure 12 : Variable Importance Plot :



Classification report for the Tuned Decision Tree :

	precision	recall	f1-score	support
0	0.86	0.84	0.85	647
1	0.91	0.90	0.90	2316
2	0.92	0.92	0.92	4552
3	0.92	0.93	0.93	2727
4	0.88	0.85	0.86	433
accuracy			0.91	10675
macro avg	0.90	0.89	0.89	10675
weighted avg	0.91	0.91	0.91	10675

Classification report for Random Forest:

	precision	recall	f1-score	support
0	0.92	0.85	0.89	647
1	0.94	0.92	0.93	2316
2	0.93	0.96	0.95	4552
3	0.96	0.95	0.96	2727
4	0.96	0.88	0.91	433
accuracy			0.94	10675
macro avg	0.94	0.91	0.93	10675
weighted avg	0.94	0.94	0.94	10675

Classification report for XGBoost Algorithm:

	precision	recall	f1-score	support
0	0.92	0.84	0.88	866
1	0.93	0.91	0.92	3016
2	0.93	0.95	0.94	5996
3	0.95	0.95	0.95	3644
4	0.92	0.86	0.89	569
accuracy			0.93	14091
macro avg	0.93	0.90	0.92	14091
weighted avg	0.93	0.93	0.93	14091

As in the base model, since this is an imbalanced dataset for classification, we look at f1-score(micro average which is same as accuracy in this case). We observe that the Random Forest classifier gives the best results followed closely by the XGBoost Classifier.

Recommendations and Actionable Insights

1) The restaurants of Banashankari had scathing reviews in spite of high ratings. The number of votes was low for those restaurants which misled people to believe that the restaurants were good. Hence,we would recommend zomato to primarily sort restaurants by number of votes and then sort them by ratings, thus reducing the probability of the customer being dissatisfied with their experience.

2) To set up shop in residential areas, restaurants are recommended to specialize in a variety of cuisines,especially fast food, to cater to the large number of students studying in the educational institutions erected there. Invest in good infrastructure such that customers get to enjoy their meal in a peaceful environment.

3) Running a restaurant in a commercial area is no walk in the park.The target audience here happen to be celebrities,critics and other prominent people.With such tough audience, restaurants need to step up their game and go beyond quality food and enhance the customer experience with exceptional customer service and unique themes. For instance, Church Street Social has separate workspaces for freelancers to exercise their creativity while still satisfying boisterous party-goers in adjacent party rooms. The versatility of this bar-cum-cafe is what sets it apart from most typical restaurants as it serves customers the whole experience.

4) If restaurateurs have set their eye on areas popular for real estate,it is not advisable to build a restaurant. Businessmen tend to frequent these areas looking for great places to stay. Hence, one must have the capital investment to be able to build a luxury hotel with a five-star restaurant embedded within it while they are free to implement a surge in their prices as customers here do not mind paying more for a luxurious stay.

Conclusion

In this project we attempt to analyze the trends in the operations of Zomato in the city of Bangalore. Through exploratory data analysis, we've discovered a multitude of actionable insights which sheds light on the currently observable trends across demographics and the respective locales. A model has been built in order to predict the ratings of a restaurant based on a multitude of integral factors owing to its popularity amongst the general public. We observe that effective prediction to a good degree can be done through the usage of a few columns alone. Tuning the model to perform at close to its peak capacity does require text analytics in the form of sentiment analysis based on the reviews which was done by the VADER sentiment analyzer. We observe that the Random Forest classifier gives the best results followed closely by the XGBoost Classifier. The trend analysis is extremely useful for zomato in and of itself along with inadvertently helping upcoming restaurants and existing franchises to optimize their operations in the future and fix existing issues to the fullest extent possible.

Future Scope

A more effective modality of text analytics would be to use the word2vec word embedder. In our case this would prove ineffective due to the fact that we have words like 'Biryani' which have a Sanskrit based root.. A solution to this problem would be to create a custom lexicon in order to accomodate Sanskrit root words.