Analysis of Zomato Food Chain

Adhwaith Menon
5C , PES1201800207
PES University , Bengaluru ,
adhwaith2017.menon@gmail.com

Vikshith Shetty
5C , PES1201801555
PES University , Bengaluru ,
vikshith61@gmail.com

Rajdeep Sengupta
5C , PES1201800144
PES University , Bengaluru ,
senguptarajdeep21@gmail.com

I. Introduction and background

The topic chosen for analysis by us is 'Zomato Food Chain Analysis'. In this given report we shall illustrate the various types of analyses we have performed on our dataset. The latter was downloaded from a website known for datasets called 'kaggle.com' (the link of which shall be provided in the reference section). The chosen dataset contains data for various restaurants across the city of Bengaluru (Karnataka, India) stored in the Zomato database. In simpler terms, these are the restaurants from which Zomato organises deliveries to various users.

There are various types of restaurants, all ranging from 'Indian' to 'Chinese' cuisine. Notably, the dataset also features other important attributes such as the average cost of a meal in a given restaurant. Moreover, additional attributes such as location and 'favourite food' are also featured in the dataset.

The problem we wish to solve is twofold - Our analysis wishes to answer the type of questions a entrepreneur would ask - as to whether the setting up of a restaurant is profitable in a certain area or not. Recently due to the Covid-19 Pandemic, it has been noticed that orders via Online Food-Chain companies such as Zomato, Swiggy etc. have only increased. Thus judging the profitability of a venture in this field by the success of the same on Zomato is akin to judging the success of said venture in the real world. Furthermore, this analysis can also serve as an indicator for the average price per meal a customer will have to pay given his/her intentions to eat outside. The models we intend to create shall perform predictions on the dataset and enable all users to view the cost structure of either opening a venture in a said area or wishing to eat in a restaurant of a particular cuisine in said area.

Additionally we shall also back up our claims and model predictions with a plethora of strong visual evidence and thus shall engage in a lot of visualization with various tools.

By solving such issues, perhaps it will encourage entrepreneurs to open up more ventures in the city , and if these are profitable it will certainly be beneficial for both of them as well as the various dependents who rely on such ventures for their livelihood.

Our goal is to perform all the above analyses with the Python language and display it in a format comfortable for the user to read and understand.

II. Previous work

A. Recent Works

We have recently done data analyses on a variety of datasets , all of which have been downloaded from 'kaggle'. A plethora of techniques have been applied on them , ranging from Regression to Decision Trees and to Barplots. We wish to ensure that this current analyses will utilise a variety of analytical techniques that will help in deriving meaning from said data in a smooth format.

B. Assumptions about dataset

A major assumption about the dataset that we have made is that it is official data from Zomato. The dataset was downloaded from the profile of a 'kaggle' user , who was verified by 'kaggle' , yet we assume it has come from the direct source and hence all the data in it is factually correct.

Another important assumption we have made is that the primary source of profit of a said restaurant is via its' Zomato score and revenue generated solely by Zomato.

The above assumption can be supported by the fact that we are right now in the midst of a pandemic and hence most in-person restaurants will not attract a favourable number of customers. Instead the profitability of a restaurant will lie on how well it caters to the large pool of Zomato users as recent studies have shown that the number of Zomato users have increased dramatically over the previous few months.

The second assumption we have made is that the price for foods should not have changed by a significant amount from the year 2019 , the year in which our dataset was collected and uploaded online, to the current year which is 2020.

The above assumption can be supported by the fact that most restaurants generally run the same menu for long periods of time and that due to the current pandemic , have possibly had no time to change the same. Furthermore we make one more assumption that these restaurants are indeed still active and are running on Zomato.

Again this assumption can be supported by the fact that most of these restaurants still exist on the Zomato application and hence are functioning and running. We have also stated that our analysis is only on restaurants that exist on the Zomato database , and not any restaurant in general. This assumption might lead to a few less accurate predictions that our models might make. Nevertheless it will mostly be accurate.

III. PROPOSED SOLUTIONS

The solutions that we have decided to work on are as follows.

A. Preprocessing of data

The first step in any data analysis is the processing of the data we have on hand. Our dataset was fairly large, around 586MB in size.

On further inspection we found out that there were two columns , notably - 'Favourite Dish' had more than 60% of it's rows as empty. Thus due to the given circumstances , we removed the entire column. Some other columns , we noticed , such as 'Reviews List' were not useful for our analysis and hence we decided to drop this column as well.

Additionally , two other columns - 'online order' and 'book table' had values which were either mismatched or empty. The former two columns had over 65% missing data and thus could by no means be filled. Furthermore these two columns would not be significant to any of our data so far and hence we decided to drop them as well.

Now we were left with a few additional columns , around 14 of them. Amongst these , we noticed that the 'Ratings' columns had around 20% missing data. In our dataset , the values in this column take values between the range (0 - 5). Thus on further analysis , we realised that the missing values can be replaced by two options. The first would be to replace the empty values with the mean of the entire column. The second option would be to replace the values which have- 'NEW' within by a rating of around 5.

Thus in this way , we managed to fill the ratings columns and also added a small tweak - instead of taking just the average we added a very small error factor (around 0.06) which was either subtracted or added to the mean depending on some random value. The above was done so as to ensure that some randomness is present in the dataset.

Now comes the most important columns of all , 'Cost' for each meal. This column had around 35% missing values and being the most important one , could not be arbitrarily removed. Thus to fill in these missing values , we made use of a **Linear Regression** model. The given model was created with packages and takes in certain inputs -

The inputs are as follows - Ratings of said restaurant and The type of restaurant (such as Fine Dining etc.). This is expected to output the approximate cost of the meal in the given restaurant. The following data was then fitted with our model and was then trained and tested. On testing , we realised that the accuracy was 87% which is a fairly decent number. Using this same model , we then fed the values of those attributes which had missing ratings and obtained the required outputs. These were then added to the dataset and it was finally obtained clean.

Lastly , we went through the various links of websites in the required column and removed those links that were inactive or sending an error.

B. Models

- The first model that was prepared was a **Linear Regression** model which was built using scikit learn library and was used to find the values of cost per meal in each of the respective restaurants , with an input of Ratings and Type of restaurant.
- The second and primary model which was constructed was a **Decision Tree**. This model was used to predict the price of a meal in a restaurant , given the - 'Location' , 'Cuisine' , 'Restaurant Type' , such as Fine Dining or Quick Bites. Using these values , we used scikit learn to train these values and fitted it into the model. This given model works two folds , in that it can be used by a budding entrepreneur as well as a regular customer who wishes to have a meal. Thus on testing the data it predicted a very decent accuracy of around 84%. This was then optimised by pruning certain branches of the Tree and reducing the height. This then resulted in an accuracy of around 86.3% , which was satisfactory for us.

C. Evaluations

As explained in the previous paragraph , we trained the model (Decision Tree) to predict the cost of a meal given certain features.

We observed the following while evaluating both models -.

- In the regression model , the cost predicted depended on the ratings and thus was giving a fairly accurate range of around 300-2000.
- The primary reason we chose to use a Decision Tree was that it is very good at predicting categorical data. Other models such as ANN or even regression models (even logistic) are far better for processing numerical sets of data.
- In both our models , we had to make all the categorical data into their respective One Hot Encodings.
- The above was done via the use of scikit learn , which provided a function called 'dummy-vars' that helped us to create the required One Hot Encoding that was needed
- Post this , we then fed the data into the required models and obtained the output.
- As mentioned earlier , to prevent the decision tree from overfitting , we used some pruning and removed a few of the branches , thus resulting in a shorter but thicker decision tree that was able to predict rather accurately.
- Another feature we tried out was to also involve the cuisine type into the input of the decision tree. However , we found that this was not very accurate and often resulted in faulty values being predicted or even arbitrarily large values.
- An argument in support of not performing the above could be that cuisine does not necessarily result in a higher cost or alternatively a higher profit , instead the quality of the restaurant along with location determines the above.
- This feature is very useful for budding entrepreneurs as well , as they can get an idea on where exactly

they can establish a restaurant so as to make a tidy profit.

IV. Experimental results

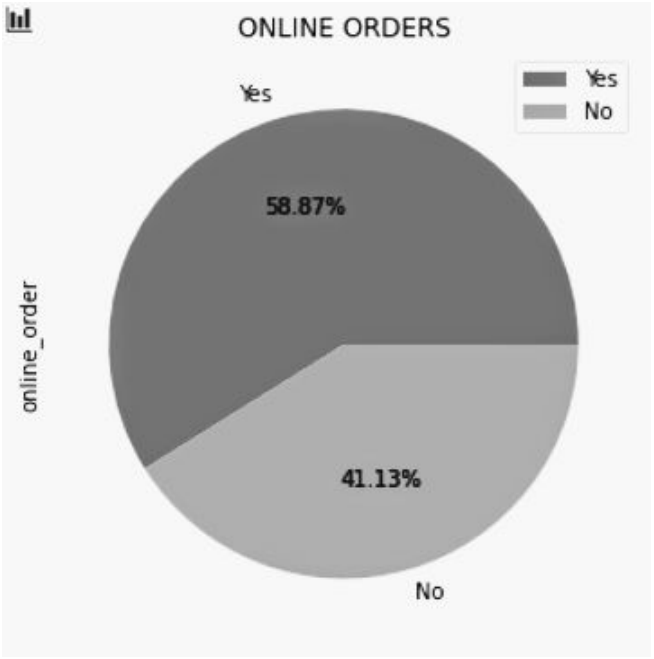
The following are our experimental results along with the various visualization techniques we tried to implement -

A. visualization

A variety of visualization techniques were tried out on our dataset some of which we shall list out here -

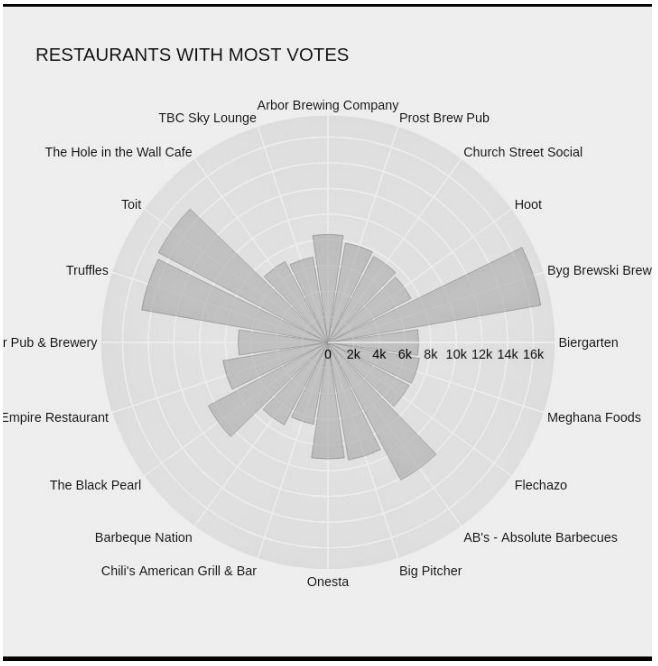
- The first visualization we performed was the creation of a Pie chart that showed the spread of online and offline orders to each restaurant.
- The second one we performed was that of a bar chart which showed the number of different types of restaurants , ranging from Casual to Fine Dining.
- Furthermore this was also made into a Boxcomb Plot which helped in identifying the differences more pronounced.
- Additionally another bar chart representing the ratings of various different types of restaurants grouped into 1-5 was displayed in a basic graph. This helped us understand the overall spread of ratings of the restaurants.
- Another Coxcomb plot of the most voted restaurants was also taken - this provided us with a very generic picture on the quality of restaurants.
- In our plot , we noticed that the 'Black Pearl Restaurant' was given the most number of votes.

The following are some illustrations of our graphs as provided -



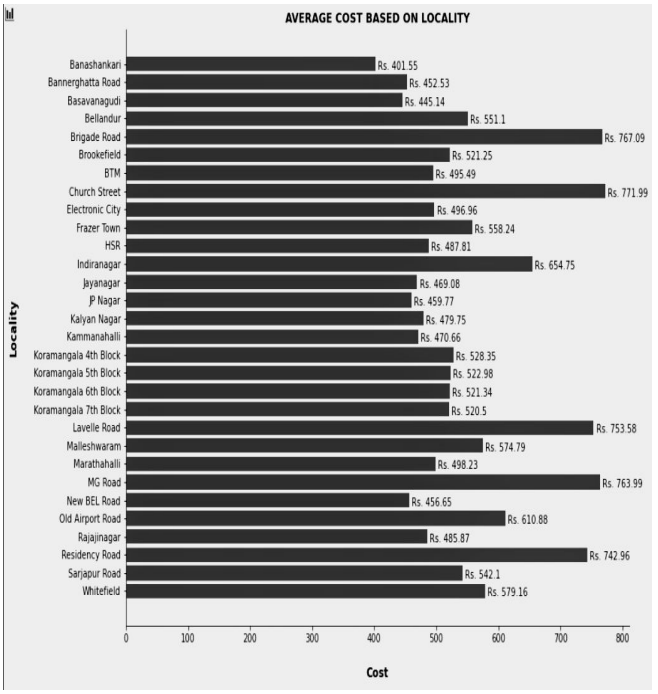
Notice the number of online orders versus the number of offline orders.

Another visualization which we have obtained is as follows-



Once again , notice the most voted restaurants using the Cox-Comb chart. It is evident that Truffles, Toit and Byg Brewski Brew are the ones with the most votes.

Another graph is as follows:



The above graph gives the average cost of dining per locality. Hence, it's crystal clear from such a useful graph that the localities like Brigade Road, BTM, MG Road are more expensive than other areas.

B. *Further Model Analysis*

We analysed both our models separately and obtained a list of results.

In our **Regression** model , the data we obtained was used to fill missing values in the cost table. As mentioned before the train accuracy was 93% and the test data had a 85% accuracy.

More importantly it always predicted values between the range of (200 - 1200). This was consistent with all of our observations.

In most cases the above model works wail and we have not found any cases where it fails as of now.

In our **Decision Tree** model , we noticed that the model fails for a few cases.

In some of the cases, we notice when we search for those types of restaurants that have a type that is relatively unique.

For example , we have one type as - 'Deserts' which is a type for only around 5-6 rows. For such types , no matter the location , the price always comes around the same - 450. This is due to the fact that our data is limited and hence the decision tree is unable to learn.

The other area where our model fails is for predicting those places which do not have many restaurants. For those locations it often gives rather large values which do not match very well with other data.

For example , Hesereghatta is an area with just one restaurant which has a cost of around 350. On putting that location as an input along with any random type ,we obtain a value of around 750 which is significantly higher than expected.

Nevertheless the model works very well for other types of data having a test accuracy of around 86.2% and train accuracy of around 90%.

C. *Tables*.

Sl. No	Values		
	Location	***Type***	***Cost***
1	Church Street	Fine Dining	2600
2	Banashankari	Quick Bites	450
3	Whitefield	Casual Dining	560

The following table shows our predictions for some values in the **Decision Tree** -

Notice the vast difference in the prices. In a real world scenario , this predicted data is actually correct (if one were to consider Bengaluru restaurant prices by observance).

CONTRIBUTIONS

The contribution of each member are given as follows -

- **Adhwaith Menon** - Pre Processing Data and Preparation of the report.
- **Vikshith Shetty** - Creating models and Video Preparation.
- **Rajdeep Sengupta** - visualization and Exploratory Analysis along with Report Preparation.

All team members worked on their respective parts for the given project.

CONCLUSIONS

The strong contribution by each member of the team helped us achieve very useful insights of the data using various visualizations and machine learning models.

The decision tree model designed in this project produces **81.85% accuracy** in predicting the cost of dining in a restaurant given the independent variables 'location' and 'cuisine'.

GITHUB PROJECT LINK

https://github.com/Rajdeep2121/Data-Analytics-Project

REFERENCES

The following are the references we have used for our project -

[1] https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants. (**Dataset link***)*

[2] Tanizaki, Takashi & Shimmura, Takeshi & Takenaka, Takeshi. (2018). Demand forecasting in restaurants using machine learning and statistical analysis.

[3] Customer satisfaction in the restaurant industry; Examining the Model.Perspective-http://www.aessweb.com/pdf-files/2-104-4(1)2014-JABS-1831.pdf.

[4] Nataasha Raul., Yash Shah., Mehul Devganiya.:Restaurant-Revenue Prediction using Machine Learning. In: International Journal of Engineering and Science 6(4), 2016, pp. 91-94.

[5] Dr.Anitha.C, Bidisha Das Bakshi, Varsha Rao: A Survey on Local Market Analysis for Successful Restaurant Yield