2nd National Level Students' Research Conference on
"Innovative Ideas and Invention with Sustainability in Computer Science and IT-2021"
In association with International Journal of Scientific Research in Computer Science, Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Identification of Proper Restaurant Place : By Extracting Customer Reviews

Vikas Shukla[1], Soumya Bajpai[1], Sachin Bhoite[2]
[1]School of Computer Science, Faculty of Science, MIT-WPU, Pune, Maharashtra, India
[2]Assistant Professor School of Computer Science, Faculty of Science, MIT-WPU, Pune, Maharashtra, India

ABSTRACT

Now a days ordering food online by using different application is normal for us . Some of the popular names of the apps are as ZOMATO, SWIGGY , UBER EATS etc. Many customers' visits a restaurant based on reviews given by the customer or user in different apps. The growth of restaurants and hotel is mainly based on the reviews provided by the customers. India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotels. In this paper we will , collect data from a online application and check in which area peoples are more attached to a particular type of cuisine. So, this paper focuses on analysing the Zomato restaurant data for a city in India to predict the place for new restaurant by using various supervised Machine Learning algorithms based on existing data. And build predictive model with best accuracy.

Keywords: Zomato, Science and Technology, Machine learning algorithms, accuracy.

I. INTRODUCTION

Now a days everyone has a unique style and taste of eating some peoples like South Indian food some likes Punjabi Food and some likes Marwari food .In growing country like INDIA restaurants , resort and hotel play a big role and there business depends upon the reviews given by the customer .

So basically the main objective of this project is to find in a city like BANGALORE , PUNE , DELHI etc. In particular area which cuisine is more famous and what people eat more. Higher ratings lead to higher profit margins. Notations of the ratings usually are stars or numbers scaling between 1 and 5.

Different machine learning algorithms like SVM, Linear regression, Decision Tree, Random Forest can be used to predict the most rated food in a particular area.

II. RELATED WORK

Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing, Feature.

James Huang, Stephanie Rogers, and Eunkwang Joo have used LDA & Python scripts to improve restaurants ratings they said" Based off of the Online LDA algorithm, we have been able to show what users care about most in their reviews of restaurants, and have been able to pinpoint the areas of interest for specific restaurants. Overall, it turned out that

users care most about service, and subsequently value, take out, and decor" [1]

Gayathri.T1 used J48 , Multilayer Perceptron and Naïve Bayes to improve and review the rating of hotels and obtained highest 84.42% accuracy in Naïve Bayes. She says "To find the trigger to improve ratings, a sample record of poor rating is taken and modified to reduce the price range to one. This sample record is tested on J48 Zomato model. It was found that there was no change in rating. Whereas when the country code was changed there was change in rating " [2]

V. B. Raut and D. D. Londhe, "Opinion Mining and Summarization of Hotel Reviews," The reviews are scraped from ww.tripadvisor.com using web crawler. The reviews are distinguished into positive and negative polarity using sentiwordnet and various machine learning algorithm are used to check their accuracy. [3]

"Random Forest Model for Predicting Grayscale Digits on Images" Stanley Ziweritin1 , Uchenna Chikwendu Anyimukwu Ugboaja & Chidiebere Moses Osu3.They uses Random forest technique and says "The experimental results proved to be highly accurate in prediction and errors converges as required when more trees are grown and added to the forest space. We therefore; evidently conclude that the RF classification tree model performed better as required than the random regression model". [4]

Chandresh make his project and concluded this "From this data processing, we can get this following conclusion: There are 105624 restaurants registered on Zomato Apps based in India. Almost 18% of registered restaurants in India are located at New Delhi (19611 restaurants), 13% in Mumbai (14026 restaurants), 7% in Bengaluru (8507 restaurants). 1080 Restaurants in India having Excellent rating type. 11695 Restaurants having Very Good as Rating Type and 336 with Poor Rating. 104 Restaurants having rating above 4.8 and 84 above 4.5" [5]

Coronary Illness Prediction and Analysis of Various Machine Learning Techniques Banumathi P. And , Miraclin Joyce Pamila J.C. discuss about Various Techniques in their conclusion they said "The proposal uses the Random, IJSRCSE All Rights Reserved 33 forest and neural network for anticipating coronary illness and it compared against different ML procedures. Experimental results on foreseeing coronary illness show that the proposed random forest gives 90.16% accuracy which is better compared to artificial neural network and also high f measure esteem than ML procedures this because it constructs a decision tree for every 13 attributes. [6]

III. DATA SET DESCRIPTION

For this analysis, we will be using Zomato Bangalore Restaurants dataset present on kaggle. The dataset contains all the details of the restaurants listed on Zomato website as of 15th March 2019.

IV. RESEARCH METHODS

We used observation, exploratory analysis and proposed experiment method for this work using Machine Learning Linear regression algorithm.

4.1 About supervised Machine learning classification algorithms.

1. LOGISTIC REGRESSION

Logistic Regression is perfectly match with the placement prediction objective of the research. In case of placement prediction the problem is simple binary problem. The sigmoid function will give result 1 or 0 for placement prediction. It maps predictions to the probabilities.

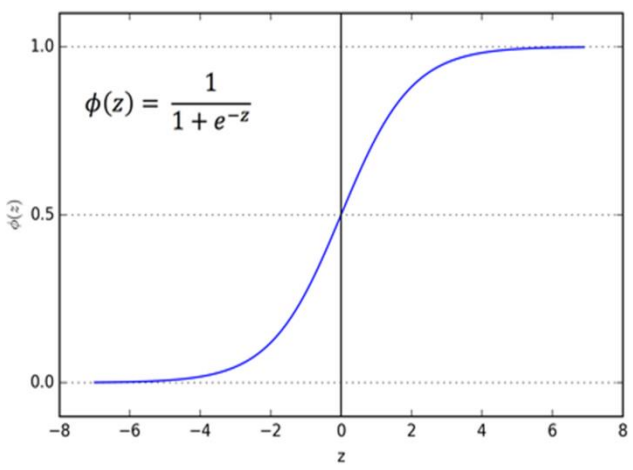


Fig. 1. Logistics Regression Function

The logistics regression equation is

Y = e^(B0 + B1X1 + . . . + BnXn) / 1 + e^(B0 + B1X1 + . . . + BnXn)

Where Y is dependent variable (predicted output), B0 is bias term, X is independent variable, n = number of observations, B1 is coefficient to input value X. Logistic regression is a linear regression but logistic regression uses more complex cost function compare to linear regression.

2. DECISION TREE

It works by creating a tree-like structure by dividing the dataset into several smaller subsets based on various conditions. The components of the decision tree are the decision nodes, leaf nodes and the branches. Nodes having multiple branches are the decision nodes, nodes with zero branches are called the leaf nodes, and the topmost node is called the root node of the tree. The nodes are connected to each other through branches. The root and decision nodes are created by computing the entropy and information gain for the dataset [5].

3. RANDOM FOREST

Random Forests builds many classification trees. While classifying a new vector from an input vector, the input vector kept down each of the trees in the forest. Then each tree gives a classification, i.e. tree "votes" for that class. The forest chooses the classification having the most votes [6]. Random forests are fast, and it do not over fit. Random Forest Classifiers are more accurate and efficient as the computation is done with a number of classifiers and the best among them is chosen by the model.

4. SUPPORT VECTOR MACHINE

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a nonprobabilistic binary linear classifier [8]. In SVM, each data item is plotted as a point in n dimensional space with the value of each feature being the value of a particular coordinate. Then, classification carried out by finding the hyper-plane that differentiates the two classes. When two data classes are linearly separable, many hyper-planes could be drawn to separate the two classes. All the hyper-planes can classify the data into two classes, and the best among all the hyper-planes is selected by the SVM classifier for the prediction model. One reasonable standard for judging the quality of these hyper-planes is via their margin lengths [7].

4.2 Pre-processing

- Cleaned the data which is not for our use.
- Attributes with null values were dropped from ratings columns and were replaced in the other columns with a numerical value.
- Label Encoder – For categorical variables this is perfect method to convert them into numeric values , best used when having multiple categories . We had various categorical values converted them into numeric for further use in algorithms .

4.3 Feature Selection

Let's start by deleting the unnecessary or redundant features. For data analysis, we do not need the

contact details of the restaurant so, deleting the following features :

url, phone, Name , Online_order ,Book_table, Rest_type , Approx. cost of two people, Listed_in(type)Listed_in(city)

Some of these columns may look like they are important but all of the same information could be found in other columns with lesser complexity. The Columns being used are as follows:

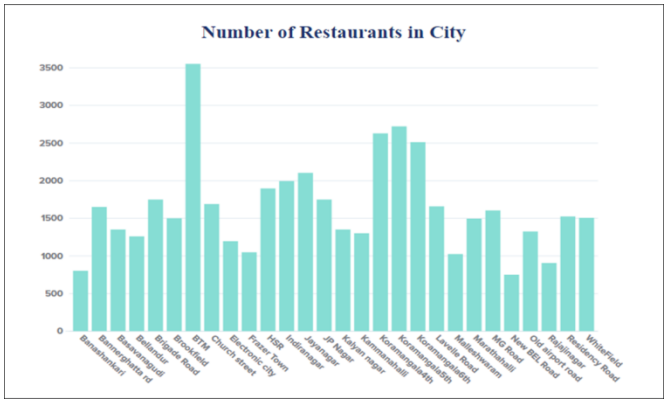
Address, Dish_like, Menu, Review_list Votes Location ,Cuisine

V. EXPLORATORY DATA ANALYSIS

We predominantly used Python scripts. Specifically, we used the Gensim Python Library, which is a topic modeling tool for documents. We used PyGal for data visualization.

In this graph we will see in which particular area more Hotels and Restaurants are there which get highest booking in whole Bangalore.

1. Number of restaurants in whole City

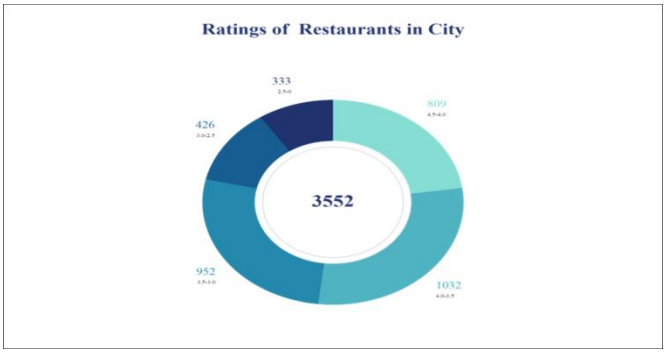


BTM has the highest number of restaurants, followed by Koramangala 7th Block. New BEL Road has the least number of restaurants, followed by Banashankari.

It seems that the main foodies live in BTM and Koramangala.

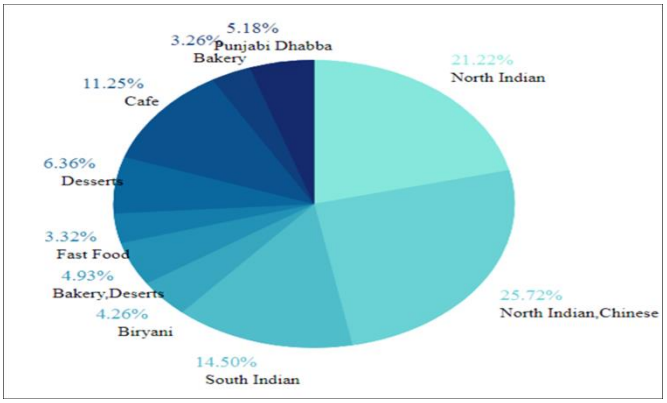
Now we will look for Restaurants RATINGS in BTM

2. Rating wise Restaurants



We can see that the number of restaurants with the rating between 3.5 and 4 are the highest. We will look into its dependencies further.

3. Ten Favourite cuisines of Peoples in BTM



As we can see in the above chart in BTM area Locals are fan of North Indian food & Chinese food ,also there is a good percentage of peoples attracting towards South Indian food.

4. Approx percentage of population in BTM ordering Online food

By the above graph we can conclude that there is good population there who are ordering food online using different applications.

5.1 Key Findings

ONLINE ORDERS	Percentage
Yes	46.6%(approx)
No	53.4%(approx)
Cuisine	Percentage (Approx)
North Indian+Chinese	25.72%
North Indian	21.22%
South India	14.50%

VI. RESULT AND DISCUSSION

In this research we have considered a particular area and by reviews of customer conclude that what is the favourite cuisine of people living there. This will help business personals to open a new restaurants there . Such analysis is essential part of planning before establishing a venture like that of a restaurant. Having an prior idea about the area and public choice of the area helps to set up a business as per the immediate choices of the people and hence profiting the business faster. We have proposed the use of Supervised Machine Learning Algorithms to predict the proper place for new restaurant.

VII. CONCLUSION

On the basis of past customer reviews of restaurants and food .We concluded that in a particular area which in a particular area which type food type is more liked and hence, which menu can be more profitable for a restaurant .This will help out a business owner to easily establish a good profit-earning restaurant.

main motive of a business is to attract people of their zone as per their interests. Zomato dataset helped us to create classification model for restaurant rating. It was found that Multilayer perception work well with this dataset.

In the above paper in BTM area a person can open a new restaurant of North India, Chinese & South Indian food.

VIII. REFERENCES

[1]. “Improving Restaurants by Extracting Subtopics from Yelp Reviews” James Huang, Stephanie Rogers, Eunkwang Joo, Social media expo 2014.

[2]. “Data mining of restaurant review using WEKA “Gayathri.T1International Journal of Engineering and Techniques - Volume 4, Issue 3, May - June 2018, ISSN: 2395-1303

[3]. V. B. Raut and D. D. Londhe, "Opinion Mining and Summarization of Hotel Reviews," 2014 International Conference on Computational Intelligence and Communication Networks, Bhopal, 2014, pp. 556-559. doi: 10.1109/CICN.2014.126

[4]. Random Forest Model for Predicting Grayscale Digits on Images Stanley Ziweritin , Uchenna Chikwendu Anyimukwu Ugboaja , Chidiebere Moses Osu Vol.8, Issue.6, pp.01-07, December (2020) E-ISSN: 2320-7639

[5]. Zomato Data Analysis (chandresh189.github.io)

[6]. Coronary Illness Prediction and Analysis of Various Machine Learning Techniques Banumathi P. , Miraclin Joyce Pamila J.C. Vol.8, Issue.3, pp.26-33, June (2020) E-ISSN: 2320-7639

[7]. Himanshu Sonawane (2019) , ‘Student Admission Predictor’, MSc Research Project Data Analytics, National College of Ireland Project Submission Sheet – 2017/2018 School of Computing

[8]. Leo Breiman and Adele Cutler, Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/c c_home.htm

[9]. Syed Ahmed1, Aditya Zade2, Shubham Gore3, Prashant Gaikwad4, Mangesh Kolhal5 (2017) 'Smart system for placement prediction using data mining', International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; Volume 5 Issue www.ijraset.com

[10]. Support vector machine, from Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Support_vector_machine