Zomato Data Analysis and Restaurant Recommendation

Rahul D. Makhija
Computer Science & Engg.
PES University
Bengaluru , India
rmrahulmakhija74@gmail.com

Rishab K.S.
Computer Science & Engg.
PES University
Bengaluru , India
rishab12360@gmail.com

Rohan Mallesh
Computer Science & Engg.
PES University
Bengaluru , India
rohanmrb@gmail.com

***Abstract*—The main aim of the this project is to develop a model to recommend restaurants to users by analysing the impact of various factors that are obtained from the Bangalore Zomato dataset have on restaurant decision of a customer. The various implementation techniques which are previously published and the unique attributes in the dataset are also discussed. Several graphs and plots are modeled to obtain key insights and inferences form the data and to understand how the features are related to one another.**
The paper also proposes a different implementation technique to recommend k restaurants based on the users history using cosine similarity function with multiple attributes present in the data to overcome the shortcomings of the various techniques cited in this paper

***Index Terms*—recommendation, data analytics, restaurant**

I. Introduction

At present there is vast flow of information on the internet and it continues to grow exponentially. Despite the benefits of such humongous data, the increasing flood of information on the web creates a need for selecting content according to the end user's preferences. In this case, it is important to alter the information to a limited amount based on the current user/customer preferences in order to assist them in making the correct decision.

In recent times, Recommender Systems (RSs) are used in variety of areas with notable examples taking the form of shopping (Amazon), movies (Netflix), travel (TripAdvisor), restaurant (Yelp), people (Facebook), and articles (TED). Most of the RSs approaches rely on a single-criterion rating (overall rating) as a primary source for the recommendation process. However, the overall rating is not enough to gain high accuracy of recommendations because the overall rating cannot express fine-grained analysis behind the user's behavior. With the tremendous success of large deep learning-based recommender systems, in better capturing user-item interactions, the recommendation quality has been significantly improved.

This project aims to recommend 'K' similar restaurants to a user based on the variety of features that helps us to achieve a deep understanding of the process that makes a user choose certain restaurants over others.

II. Literature Survey

A. Multi-Criteria Review-Based Recommender System

The main aim of [1] was to survey a bunch of existing research papers which explore generic rating based recommender models as well as research papers on how the user-generated reviews can be utilized as an alternative and valuable source in the recommendation process through merging them with Multi-Criteria Recommender System (MCRS) to enhance the accuracy of the Recommender System (RS)'s performance. Generally, the rating function in the MCRS is described as follows:
R: Users x Items = R0 x R1 x .... x Rk ,where R0 is the overall rating and R1, R2,....Rk is the rating values for each singular criterion.
The 2 main types of Multi-Criteria Recommender Systems surveyed are:

1) *MCRS using Explicit User Preferences*:
In this type, the user gives ratings to each of the item's features with or without the rating of the whole item. The user's preferences are known directly from the users' ratings on the items' features.
2) *MCRS using Implicit User Preferences*:
In this type, users provide opinions only on the item's feature that they are interested in through writing reviews that express their feelings or opinions about their experiences with the items. This type of approach is claimed to be more accurate in determining the users' preferences.

User Generated Reviews
There are many fields involved in processing textual reviews and extracting the valuable information from the reviews such as natural language processing, text mining and opinion mining (or sentiment analysis). In this survey, the authors are more interested in the involvement of sentiment analysis with RS because the sentiment analysis field will help in determining the user's preferences by analyzing the user's sentiment behind their reviews.
Overall, [5] helps researchers to gain more understanding about the multi-criteria review based recommender system and encourage them to explore the implicit values of the reviews and utilize them in future studies.

B. A Restaurant Recommendation System by Analyzing Ratings and Aspects in Reviews

The aim of [2] was to design a restaurant recommender system based on a novel model that captures correlations between hidden aspects in reviews and numeric ratings. The authors were motivated by the observation that a user's preference against an item is affected by different aspects discussed in reviews. Their method first explores topic modeling to discover hidden aspects from review text.Later profiles were created for users and restaurants separately based on aspects discovered in their reviews. Finally, they used regression models to detect the user-restaurant relationship.

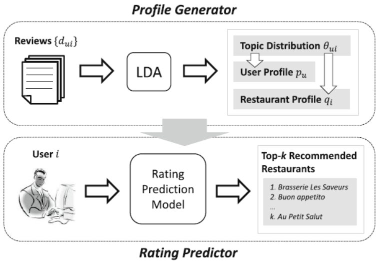

Fig. 1: System Flowchart

Users and restaurants were mapped to a common latent space ***S*** discovered from the review text. The hidden aspects in the review text may well reflect the latent factors that affect the user ratings.To find the hidden aspects and to construct the latent space, *Latent Dirichlet Allocation (LDA)* was applied to the reviews.

Profile Generator:

Let, d_{ui} = denote the review of restaurant *i* by user *u*. LDA is applied on review corpus d_{ui} to discover ***K*** topics.Let ϑ_{ui} denote the topic distribution of d_{ui} generated by LDA. D_{u} is defined as the set of reviews written by user *u*, and D_{i} as the set of reviews written for restaurant *i*. Each user *u* is associated with a profile p_u, which is a vector from *S*.In system *S*=$[0,1]^{\text{K}}$, for a given user their profile is defined as

$$p^{'}_{uj} = \frac{\Sigma_{\text{i}}\theta_{\text{uij}}}{|D_{\text{u}}|} \qquad p_{uj} = \frac{p^{'}_{\text{uj}}}{\Sigma_{\text{j}}p^{'}_{\text{uj}}}, j \in [1, K]$$

p_{uj} is defined as the distribution on the *j*th topic for user *u*, and ϑ_{uij} is defined as the distribution on the *j*th topic for review d_{ui}. Similarly profile q_{ui} for restaurant *i* is defined as,

$$q^{'}_{ij} = \frac{\Sigma_{\text{u}}\theta_{\text{uij}}}{|D_{\text{i}}|} \qquad q_{ij} = \frac{q^{'}_{\text{ij}}}{\Sigma_{\text{j}}q^{'}_{\text{ij}}}, j \in [1, K]$$

In summary, profile p_u / q_i is the normalized average topic distribution over all the reviews of a given user *u*.

Rating Predictor:

For a given user *u*, rating $\hat{r}_{\text{ui}}$ for restaurant *i* is predicted. Recommendations are then based on $\hat{r}_{\text{ui}}$ of restaurants that *u* has not visited. The rating prediction model is build on linear/logistic regression to model the relationship between ratings $\hat{r}_{\text{ui}}$ and topic distributions ϑ_{ui} of d_{ui}.

Linear Regression : $\hat{r}_{\text{ui}} = W^{\text{t}}\ \vartheta_{\text{ui}} + \varepsilon_{\text{ui}}$, $W = (W_1, W_2..., W_k)$ where W_{j} is the weight of *j*th topic and ε_{ui} is the error variable.

Logistic Regression : A multinormal logistic regression model is built as :

$$Pr(\hat{r}_{\text{ui}} = n) = \frac{e^{\beta^{\text{T}}{}_{\text{n}}\theta_{\text{ui}}}}{1 + \sum_{n^{'}=1}^{N-1} e^{\beta^{\text{T}}{}_{\text{n}^{'}}\theta_{\text{ui}}}}$$

where n = 1,2...N-1 and $\beta_{\text{n}} = (\beta_{\text{n1}}, \beta_{\text{n2}}...\beta_{\text{nk}})$ are the weights.

Given a user *u* and restaurant *i* that *u* has not rated, topic distribution $\hat{\theta}_{\text{ui}}$ is estimated based upon p_u and q_i as :

$$\hat{\theta}_{\text{uij}} = p_{uj}q_{ij} \qquad \theta_{\text{uij}} = \frac{\theta^{'}_{\text{uij}}}{\Sigma_{\text{j}}\theta^{'}_{\text{uij}}}, j \in [1, K]$$

ˆϑ_{ui} is then fed into one of the learned regression model to predict $\hat{r}_{\text{ui}}$.

The model was tested on reviews crawled from *Dianping*.It consisted of 1,168,420 reviews written by 316,702 users for 42,274 restaurants in Shanghai. Both linear and logistic regression were used, which were denoted as R-Linear and R-Logistic respectively. Evaluation metrics used were MSE (Mean Squared Error) and ACC (Accuracy). For evaluation purpose, the dataset was split into two subsets for training and test purpose respectively with the ratio 9 to 1 and test data selected randomly.

TABLE I: MSE and Accuracy of Dianping Data

Models	MSE	Accuracy
R-Linear	0.69	52.3%
R-Logistic	0.79	52.0%

C. Restaurant Rating Prediction using Regression Models

In [3], multiple independent input attributes from the Zomato Restaurant dataset are used to predict the rating of a restaurant. It is also a comparative study between the performance of various regression models to arrive at a conclusion as to which one is best for this dataset.

The metrics used to measure each of the regression models are regression score, absolute error, mean squared error and root mean squared error.

D. Hybrid Restaurant Recommender

In [4] , the model tries to combine the advantages of both, collaborative filtering and knowledge based filtering. Collaborative Filtering method of mining data involves filtering of information or patterns done using techniques involving collaboration among users' viewpoints, data sources, user ratings, etc.

Knowledge based Mining builds its recommendations based on the knowledge of the items explicitly. Unlike Collaborative filtering methods that use user's past history, this technique focuses on the knowledge stored at the back end about users and products to actively reason out what products meet the user's requirements.

A hybrid model combining the advantages of both these

TABLE II: Evaluation Metrics for Regression Models

Type of regression	Regression Score	Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	0.72	0.36	0.27	0.58
Random Forest Regression	0.92	0.127	0.083	0.288
Ridge Regression	0.73	0.345	0.267	0.517
Lasso Regression	0.73	0.352	0.268	0.5185
KNN Regression	0.80	0.2401	0.2004	0.4476
SVM Regression	0.73	0.353	0.287	0.536
Elastic Net Regression	0.75	0.3451	0.268	0.518
Bayesian Regressor	0.72	0.347	0.268	0.518

methods are implemented. The model will choose either of the techniques based on the situation. The collaborative approach is used if the system knows the user well and knowledge based approach is used if the user is new to the system.

Although reported to be highly accurate,the drawback with the model is the computation cost and the cost to search in the user database.

Methods such as data partitioning and multi threading are discussed to improve on the above mentioned drawbacks.

E. Analysis of Zomato Services using Recommender System Models

In [5] ,the authors have proposed three different approaches which include TF-IDF based approach which is a content based collaborative filtering approach where they feed the cuisine type and cost as inputs to the model and cosine similarity is used as a metric to determine and return recommendations based on the input restaurant by the user, the second approach is based on Apriori algorithm where they recommend food items based on the frequency of the occurrences of the dish in the data set, they calculate a support vector and association rules are formed. Minimum confidence is used as a metric in this model to recommend the food items in a particular restaurant. The third approach is deep neural approach where they use several of the Natural Language Processing skills like Word2Vec and FastText and pass in a custom corpus of data containing the food items and and the menu to analyze the various restaurants and use cosine similarity as a metric to recommend similar restaurants.

The shortcomings of all the three approaches are that all the approaches are very data-heavy and require a lot of data to be know about the restaurant to recommend a particular restaurant to a user. Additionally too less features are considered while making a recommendation to a user in every model, thus the recommendations may not be accurate to meet the user's requirements such as location of the restaurant. All the models fail on a new restaurant because there are very few reviews, ratings and information about the new restaurant which makes it harder for the model to derive any knowledge about that restaurant and recommend it to new user

III. Problem Statement

On initial analysis we found the need to find a relation between the various attributes affecting the rating and the popularity of a restaurant in Bangalore. The problem statement involves using various attributes present in the data set to find 'K' similar restaurants to the restaurants liked by the user, given as an input and to recommend the same to the user.

The dependency of the rating of the restaurant with it's cuisine can be very useful to determine the type of customers that are likely to visit the restaurant. We can also obtain useful insights by deriving a correlation between the locality in which the restaurant is located and the average cost for a meal in that particular restaurant to determine type of people living the neighbourhood and to recommend relevant restaurants based on their specific food taste.

Unlike the conventional approaches of recommending a restaurant based on rating and cost alone, we will be aiming to develop a relation between the various attributes provided in the 'Zomato Bangalore Restaurants' data set and to consider the attributes which would affect the ratings of the restaurant consequentially the user's choice of the restaurant.

IV. Approach

The recommendation will done by subjecting the various attributes in the data set to a cosine similarity function where we construct a vector in 'n' dimension space with respect to the attributes considered and then obtain the 'K' nearest vectors by means of calculating the angle between the current vector representing the input restaurant's features and other restaurant's feature vector in the 'n' dimensional space, we will be subjecting the 'K' nearest vectors to ratings and cost filtering before returning the final output list containing the recommended restaurants.

The parameters we will be focusing on majorly include the location of the restaurant, primary cuisine of the restaurant, best dish that is served in that restaurant, rating of the restaurant, average cost for a meal, confidence in rating of the restaurant which will be derived from the number of votes submitted and by analysing reviews submitted to the restaurant. The model will be designed to accept a list of restaurants from the user as input and use this list to recommend further restaurants to the user by constructing a vector for each restaurant in the list and comparing it with the other restaurants in the data set to return the 'K' similar restaurants.

The assumptions made are - all restaurants that are part of the same locality are fairly close located when compared to the restaurants which are listed in a different locality, with respect to the ratings field even though we have the ratings for all the restaurants in the data set we cannot use this directly to recommend restaurants to users based on this single parameter, with respect to the average cost of the restaurant and it's locality we assume it will have an influence on the people living in the neighbourhood thus affecting the category (Quick Bites, Dining, Bar, Pub) to which the restaurant belongs to. Also the menu that is served in that particular restaurant will affect user’s preferences. The final assumption made as a part of this project is that the dish liked by a customer in a restaurant corresponds to the famous or the best dish served in that particular restaurant because we will be using that dish liked, to recommend similar restaurants to user.

All the papers and references reviewed focus on recommending restaurants primarily on two factors the rating of the restaurant and the cost of the restaurant using various techniques to capture the features from user history, recommending based solely on two factors will often not return accurate results as there are a lot of other factors involved while considering to visit a restaurant which include cuisine of the restaurant, the location of the restaurant and many more. We propose a model that will prioritise these attributes over rating and cost to give better recommendations to the user the model will take each attribute individually for example the best dish served in the restaurant and draw a comparison with all the other restaurants in Bangalore which will serve the same food and will output ’K’ similar restaurants which will be filtered based on the ratings of the restaurants to recommend restaurants to the user.

V. Dataset

The dataset used can be found on kaggle under the title Zomato Bangalore Restaurants. The initial, unclean version of the dataset consists of 51,717 rows and 17 columns. The following steps were taken for data cleaning

TABLE III: Dataset Information after cleaning

Column	Count	Null/Not-Null	Dtype
name	51654	non-null	object
online_order	51654	non-null	object
book_table	51654	non-null	object
rate	41627	non-null	float64
votes	51654	non-null	int64
location	51633	non-null	object
rest_type	51427	non-null	object
dish_liked	23627	non-null	object
cuisines	51609	non-null	object
approx_cost	51309	non-null	object
reviews_list	51654	non-null	object
menu_item	51654	non-null	object
type	51654	non-null	object
neighbourhood	51654	non-null	object

A. Dropping columns and rows

The columns in the dataset which are not required for further analysis such as ’url’, ’address’ and ’phone’ were dropped. After dropping these columns, presence of duplicate rows was observed and they were dropped as well.

B. Column Renaming

Few columns were not appropriately named and may mislead the exploratory data analysis. This was fixed by renaming these columns to more suitable names.

C. Removing inconsistencies

The data type of the ’rating’ and ’approx cost’ columns were changed to support numerical analysis. Formatting of these columns was also done.

VI. Initial Insights

In order to better understand the data some of the attributes were independently analyzed and compared to find any co-relation or trends or dependency of any kind by subjecting to different regression and correlation models and modelling various graphs to depict the relationship between the attributes.

Location : The type and the people in the neighborhood can greatly influence the type and number of restaurants in a locality especially in the rapidly developing outskirts of the city in the analysis below in firgure 2 we find that much of the restaurants in Bangalore is concentrated in BTM layout, HSR layout and Koramangala

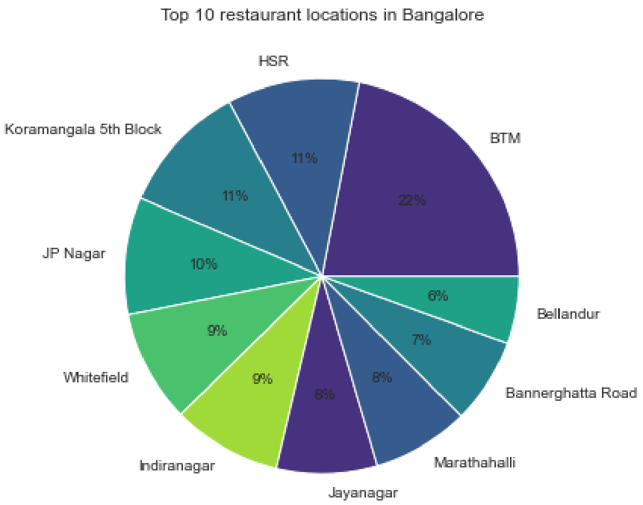

Fig. 2: Top 10 restaurant location in Bangalore

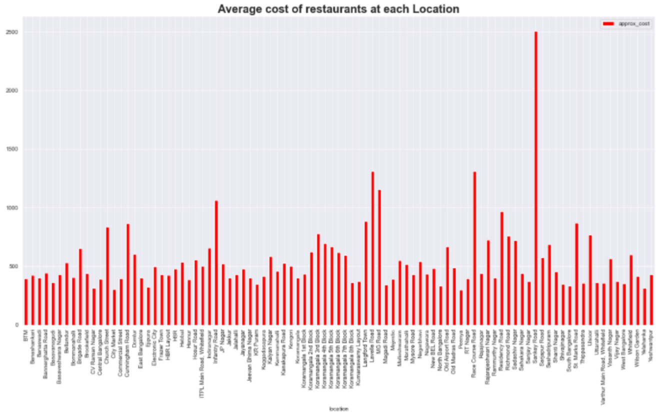

Fig. 3: Average Cost v/s Location of Restaurant

Cost : The cost of a restaurant can depend on various factors including location, cuisine and luxury offered at the restaurant, but the location of a restaurant can greatly influence the cost of the restaurant especially in the older parts of the city where some of the iconic landmarks and restaurants are located which can be clearly observed in the graph above (figure 3) where we see a higher average cost per meal in locations such as Sankey Road, M.G. Road and Race Course Road whereas restaurants located in industrial areas such as Peenya, Magdi Road and City Market in Bangalore have a relatively lower average cost per meal

Cuisine : The primary Cuisine served at a particular restaurant can impact the restaurant in many ways as it can be the primary reason for a person to visit a particular restaurant. In the below diagram (figure 4) we can observe the top 10 cuisine offered by the restaurants in Bangalore where North Indian is the most common cuisine available followed by Chinese and South Indian.

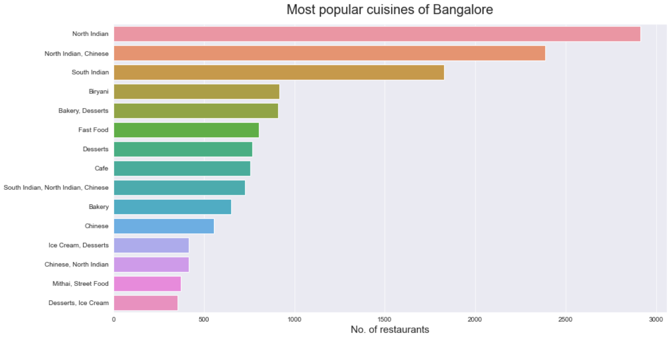

Fig. 4: Top 10 cuisine choices in Bangalore

Upon further analysing the cost and cuisine attributes a very strong co-relation was observed between the two variables as observed in the figure 5 where as the choice of cuisines offered by a restaurant increased the average cost and the rating of the restaurant also increased accordingly indicating the wide variety and luxury being offered by the restaurant

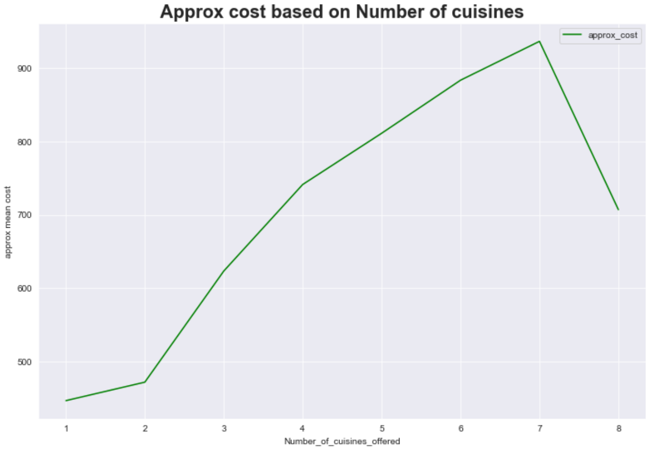

Fig. 5: Average Cost v/s Number of Cuisines

References

[1] Sumaia Mohammed Al-Ghuribi and Sharul Azman Mohd Noah, ” Multi-Criteria Review-Based Recommender System - The State of the Art ”

[2] Gao, Yifan, et al. ”A restaurant recommendation system by analyzing ratings and aspects in reviews.” International Conference on Database Systems for Advanced Applications. Springer, Cham, 2015.

[3] J. Priya, ”Predicting Restaurant Rating using Machine Learning and comparison of Regression Models,” 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.238.

[4] Dwivedi, Prerna, and Nikita Chheda. ”A hybrid restaurant recommender.” International Journal of Computer Applications 55.16 (2012).

[5] A. Sarkar, A. Baksy and V. Kirpalani, ”Analysis of Zomato Services using Recommender System Models,” 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498534.