LIKE PREDICTION & CLUSTERING OF DINING SITES IN INDIAN METRO CITIES

NAME: SRICHANDAN DASH

Date: **16/05/2021**

Email: **srichand@iitg.ac.in**

INTRODUCTION / BUSINESS PROBLEM

India boasts an incredibly diverse collection of restaurants catering to different palettes and appetites. A large part of marketing for a modern restaurant (or any company) is social media, where the number of "likes" that the company can receive will dictate its brand and image to the public. For a new food business owner (or existing company) to open a new restaurant in India, knowing ahead of time the potential social media image they can have would provide an excellent solution to the ever-present business problem of uncertainty. In this case the uncertainty is regarding performance of social media presence.

We can mitigate this uncertainty through leveraging data gathered from FourSquare's API, specifically, we are able to scrape "likes" data of different restaurants directly from the API as well as their location and category of cuisine. The question we will try to address is, how accurately can we predict the amount of "likes" a new restaurant opening in this region can expect to have based on the type of cuisine it will serve and which city in India it will open in. (For the purposes of this analysis, we will contain the geographical scope of analysis to thirteen heavily populated metro cities in India, namely Delhi, Mumbai, Kolkata, Chennai, Bangalore, Hyderabad, Pune, Visakhapatnam, Kanpur, Surat, Patna, Jaipur, Nagpur).

Leveraging this data will solve the problem as it allows the new business owner (or existing company) to make pre-emptive business decisions regarding opening the restaurant in terms of whether it is feasible to open one in this region and expect good social media presence, what type of cuisine and which city of three would be the best. This project will analyse and model the data via machine learning through comparing regression, logistic regression and Support Vector Machines to see which method will yield better predictive capabilities after training and testing. An additional aspect of the project also includes CLUSTERING of existing restaurants in these cities which can help an existing owner decide which city is most like his present city and where can he get similar public response if he decides to build a new restaurant there.

DATA – SCRAPPING & CLEANING

In this section we will first retrieve the geographical coordinates of the thirteen metro cities. Then, we will leverage the FourSquare API to obtain URLs that lead to the raw data in JSON form. We will speerately scrape the raw data in these URLs in order to retrieve the following columns: "Venue_Name", "Category", "Latitude", "Longitude". and "id" for each city. We can also provide another column ("Metro_City_Name") to indicate which city the restaurants are from.

It is important to note that the extracts are not of every restaurant in those cities but rather all of the restaurants within a 10000KM range of the geographical coordinates that geolocator was able to provide. However, the extraction from the FourSquare API actually obtains venue data so it will include venues other than restaurants such as concert halls, stores, libraries etc. As such, this means that the data will need to be further cleaned somewhat manually by removing all of the non-restaurant rows. Once this is complete, we have a shortened by cleaned list to pull "likes" data. The reason the cleaning takes precedence is mainly that pulling the "likes" data is the computing process which takes the longest time in this project so we want to make sure we are not pulling information that will end up being dropped anyways.

The "id" is an important column as it will allow us to further pull the "likes" from the API. We can retreive the "likes" based on the restaurant "id" and then append it to the data frame. Once this is complete, we finally name the dataframe 'raw_dataset' as it is the most complete compiled form before needing any processing for analysis via machine learning.

DATA – FURTHER PREPARATION

The data still needs some more processing before it is suitable for model training and testing. Mainly, the "categories" column contains too many different types of cuisines to allow a model to yield any meaningful results. However, the different types of natural cuisines have natural groupings based on conventionally accepted cultural groupings of cuisine. Broadly speaking, all the different types of cuisine could be reclassified as European, Latin American, Asian, Regional, North American, or casual establishments such as coffee shops or ice cream parlours. We can implement manual classification as there really are not that many different types of cuisines.

As this project will compare linear regression, logistic regression and SVM classifier, it makes sense to have "likes" as both a continuous and categorical (but ordinal) variable. In the case of turning into a categorical variable, we can bin the data based on percentiles and classify them into these ordinal percentile categories. I tried different ways of binning but in the end, splitting the sample into three different bins proved to yield the best classification results from a prediction standpoint. As the last stage of data preparation, it is important to note that the regressors are categorical variables (13 different cities and 6 different categories of cuisines). Hence, they require dummy variable encoding for meaningful analysis. We can accomplish this via one-hot encoding.

	id	Venue_Name	Category	Latitude	Longitude	Metro_City_Name	Likes	categories_classified	ranking
0	54783eab498e910f8bd1781d	Naturals Ice Cream	Ice Cream Shop	28.634455	77.222139	Delhi, India	48	casual	2
1	5662936e498e19a9801a663f	Amritsari Lassi Wala	Snack Place	28.657325	77.224138	Delhi, India	7	casual	1
2	54dc85c7498ef8f9ab9b3c08	Tamra	Restaurant	28.620543	77.218174	Delhi, India	37	american	2
3	519ba450498eb0c559152d94	HOTEL SARAVANA BHAVAN	South Indian Restaurant	28.632319	77.216445	Delhi, India	98	regional	2
4	4cb876d7f50e224bd2d6e6fb	Sagar Ratna	Indian Restaurant	28.635487	77.220650	Delhi, India	38	asian	2

METHODOLOGY

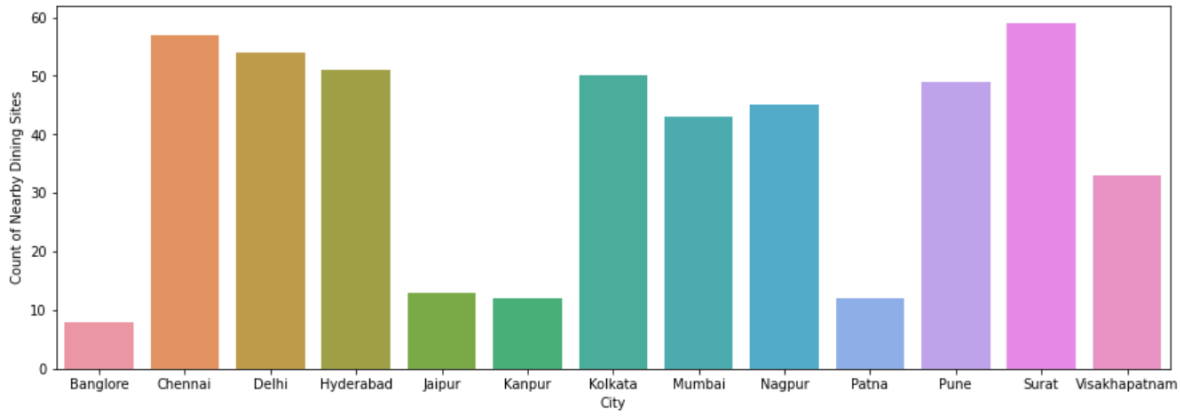
For analysis of the data, we will use Seaborn and Matplotlib. We will be using Folium for plotting maps. Exploratory analysis done involves comparison between average likes and category, count of dining sites and city, average likes and city. Inferences drawn from these are explained in further sections. After this regression and clustering are implemented.

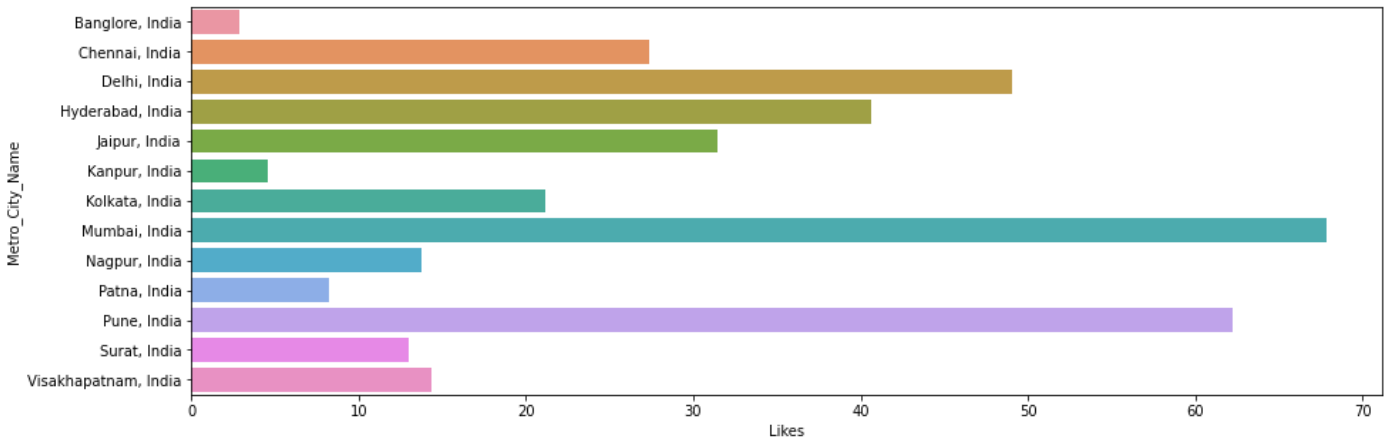
The clustering of the dining venues will be done via K Means Clustering. We will be using the Sci-Kit Learn Package for the same. In the Clustering section we will use the categories and likes of each venue to assign them a cluster. Thus, the ONE HOT ENCODED venue category will be utilized for K means clustering.

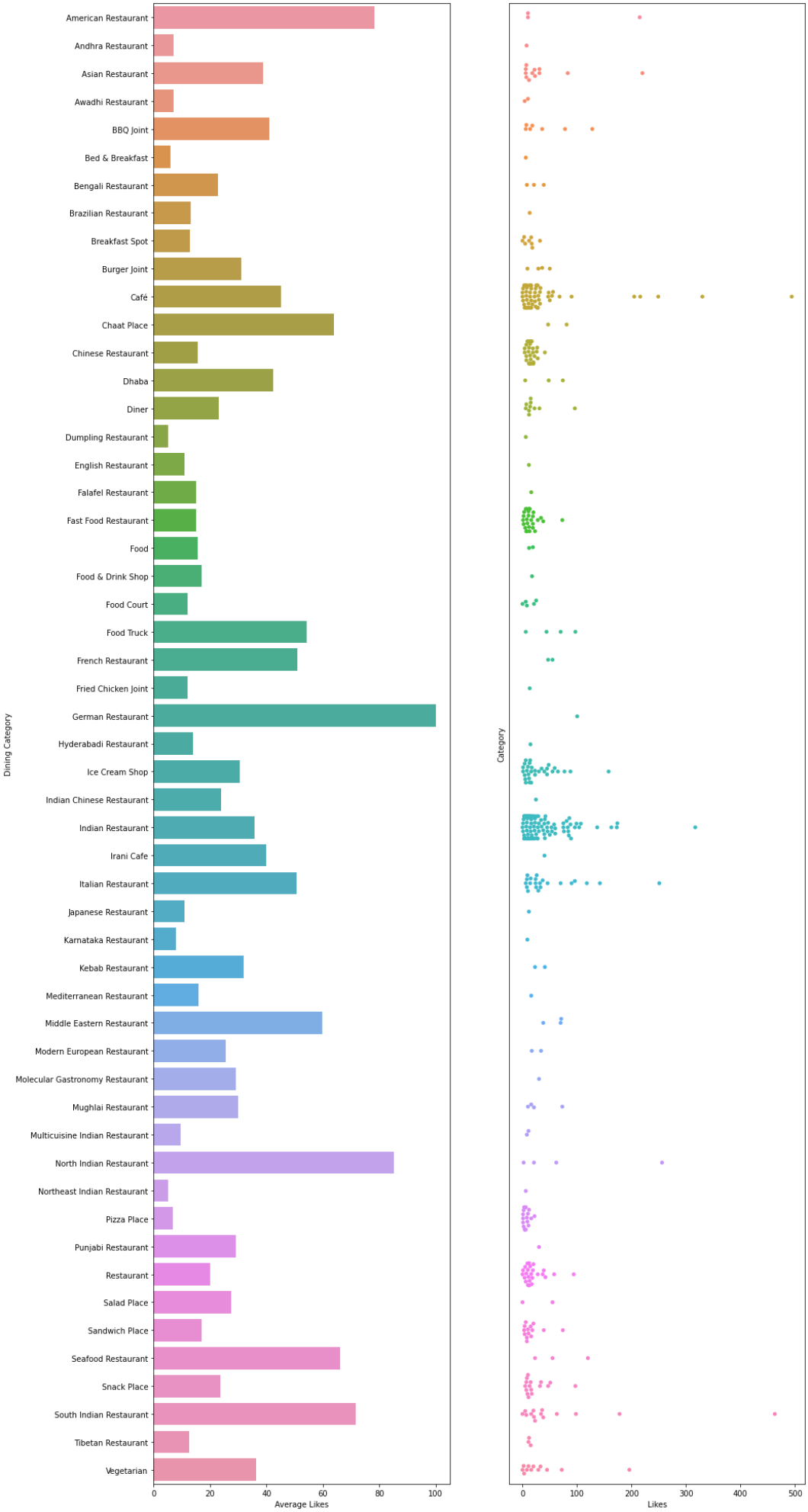
This project will utilize both linear and logistic regression machine learning methods to train and test the data. Namely, linear regression will be used to predict the number of "likes" a new restaurant in this region will have. We will utilize the Sci-Kit Learn Package to run the model.

We can also utilize logistic regression as a classification method rather than direct prediction of the number of likes. Since the number of "likes" can be binned into different categories based on different percentile bins, it is also potentially possible to see which range of "likes" a new restaurant in this region will have.

Since the "likes" are binned into multiple (more than 2) categories, the type of logistic regression will be multinomial. Additionally, although the ranges are indeed discrete categories, they are also ordinal in nature. Therefore, the logistic regression will need to be specified as being both multinomial and ordinal. This can be done through the Sci-Kit Learn Package as well. The SVM will also be implemented via Sci-Kit Learn Package and a *sigmoid* kernel is used.






American Restaurant
Andhra Restaurant
Asian Restaurant
Awadhi Restaurant
BBQ Joint
Bed & Breakfast
Bengali Restaurant
Brazilian Restaurant
Breakfast Spot
Burger Joint
Café
Chaat Place
Chinese Restaurant
Dhaba
Diner
Dumpling Restaurant
English Restaurant
Falafel Restaurant
Fast Food Restaurant
Food
Food & Drink Shop
Food Court
Food Truck
French Restaurant
Fried Chicken Joint
German Restaurant
Hyderabadi Restaurant
Ice Cream Shop
Indian Chinese Restaurant
Indian Restaurant
Irani Cafe
Italian Restaurant
Japanese Restaurant
Karnataka Restaurant
Kebab Restaurant
Mediterranean Restaurant
Middle Eastern Restaurant
Modern European Restaurant
Molecular Gastronomy Restaurant
Mughlai Restaurant
Multicuisine Indian Restaurant
North Indian Restaurant
Northeast Indian Restaurant
Pizza Place
Punjabi Restaurant
Restaurant
Salad Place
Sandwich Place
Seafood Restaurant
Snack Place
South Indian Restaurant
Tibetan Restaurant
Vegetarian
Dining Category
0
20
40
60
80
100
Average Likes
Category
0
100
200
300
400
500
Likes

RESULTS

K MEANS CLUSTERING RESULTS

K Means Clustering is implemented here with 5 clusters. The data used comprises ONE-HOT-ENCODED data of different dining venue category and city. The resulting clusters are plotted on a graph using Folium and are further analysed. With this a business owner, looking to open a new food centre in a different city can know which location to choose from and which location and category matches perfectly to his present location and category. Then he can choose the best location based on likes or demographics.

LINEAR REGRESSION RESULTS

A linear regression model was trained on a random subsample of 70% of the sample and then tested on the other 30%. To see if this is a reasonable model. the residual sum of squares score, and variance score were both calculated. Given the low variance score, this is probably not a valid/good way of modelling the data. Therefore, we move on to logistic regression.

LOGISTIC REGRESSION RESULTS

A multinomial ordinal logisitc regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%. To see if this is a reasonable model, its accuracy score and log-loss were calculated (77% and 0.55 respectively). Although this is not a perfect prediction, an accuracy of 77% between the training set and test set is a reasonable result. The classification report is also printed later on below. Given the modestly accurate ability of this model, we can also run the model on the full dataset.

SUPPORT VECTOR MACHINE

The SVM is implemented with a sigmoid kernel and default value of gamma parameter. The model is trained with 80 percent of training data and offers an accuracy score of 73% and ROC-AUC score of 0.71. Thus, we can now predict the rank of a dining location and hence have an idea of the range in which the total likes for that location would lie. Although this is not a perfect prediction, an accuracy of 73% between the training set and test set is a reasonable result.

DISCUSSION

The clustering gives insights on which category of restaurant or which city to choose depending on the present city and category of the business owner. The first thing to note is that given the data, logistic regression presents a better fit for the data over linear regression. Using logistic regression we were able to obtain a Accuracy Score of 77%, which although not perfect, is more reasonable than the low variance score obtained from the linear regression. As stated before, please note that for the purposes of this project, we are assuming that likes are a good proxy for how well a new restaurant will do in terms of brand, image and by extension how well the restaurant will perform business-wise. Whether or not these assumptions hold up in a real-life scenario is up for discussion, but this project does contain limitations in scope due to the amount of data that can be fetched from the FourSquare API.

As such, to obtain insights into this data, we can proceed with breaking down the results of the logistic regression model. The results showed that the model is better at predicting if a restaurant will fall into the best or worst percentile of likes. This allows us to roughly predict the potential performance of the business opportunity. Different binning methods for the classes were attempted, but the use of 3 bins yielded the best Accuracy Score.

CONCLUSION

In conclusion, after analyzing restaurant "likes" in metro cities of INDIA from 486 restaurants, we have developed a general classification model for which "ranking" of likes a new restaurant will potentially fall into based on its characteristics.