Predicting the Existance of an

Indian Restaurant

in The Neighborhoods of Toronto

Lavanya sai kumar Kantubhukta

June 11 ,2020

1. Introduction

1.1 : **Background** :

Everything that exists in this universe affects every other thing . If one can understand the relation between them one gets the ability to predict the future . The strength of relations vary among different things. The report is about learning the correlation among venues present in an area / Neighborhood. We can say that a restaurant near a shopping mall might have more customers in a day than a restaurant which is located near some Government Building . The presence of different venues in a place can affect each other . In this report we go through the process of finding the correlation of certain venues in an area with a Indian restaurant.

1.2 : **Problem**:

A restaurant near an isolate location draws less customers . where a restaurant near a shopping mall or a movie theatre draws more customers.Our goal of this report is to examine the present data on the list of neighborhoods and venues in the city of Toronto. We are concerned with the presence of an Indian restaurant in a neighborhood . We have to find correlation of the indian restaurant with the other venues . We should train a model that predicts the possibility and presence of an Indian restaurant in that neighborhood.

1.3 : Key idea / Hypothesis : If the model predicts a existance of an Indian Restaurant in a neighborhood . It defines 2 points :

1.The place is suitable for an Indian Restaurant

2.If the place has no Indian restaurant in the area and models predicts one. We can say that the place is suitable of starting a new restaurant.

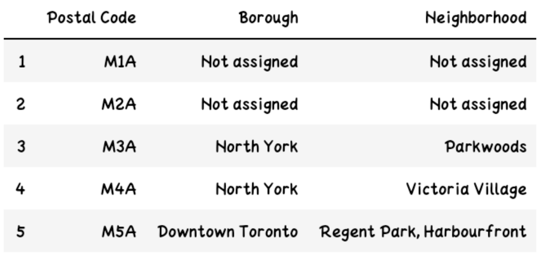
Note : The report focus on the suitability of a place for a new restaurant . It wont go in detail with respective to the financial aspects of starting a restaurant.

In this Report we are going to analyse the venues presence in the neighborhoods of Toronto city and the relation between the venues.

2. Data Acquisation and processing :

2.1: Data sources :

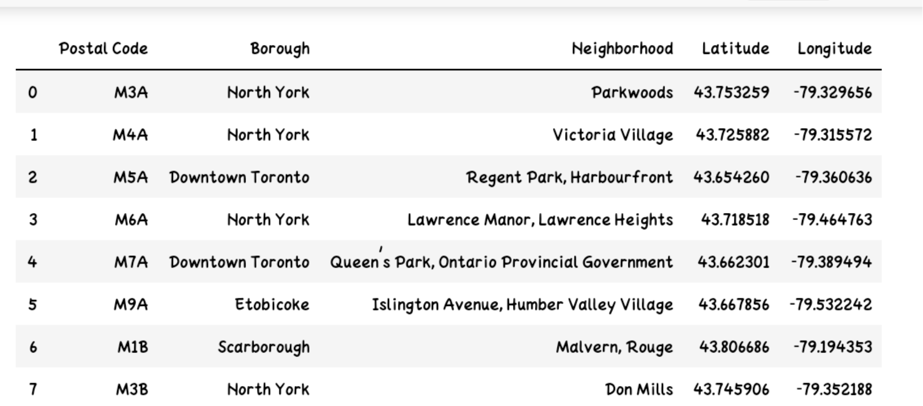
The Borough and neighbourhood data of Toronto city is obtained from wikipedia url : “https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M” .
The data is proccessed from a table containing borough and neighbourhood details .



	Postal Code	Borough	Neighborhood
1	M1A	Not assigned	Not assigned
2	M2A	Not assigned	Not assigned
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Regent Park, Harbourfront

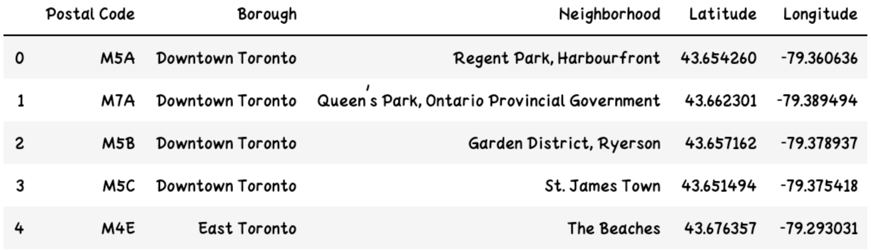
2.2 : Data preprocessing

we than process the data by removing non assigned values. We than download the postal code coordinates from “https://cocl.us/Geospatial_data” . we then assign the coordinates to the neighborhoods corresponding to the respective postal code it belongs.



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	43.745906	-79.352188

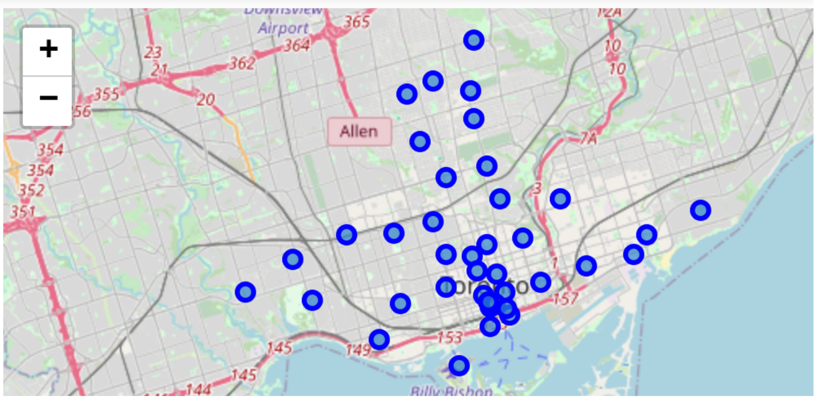
For the analysis we only work on neighbourhoods in toronto city. we cluster the neighbourhoods corresponding to toronto city.



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

2.3 Plotting on map

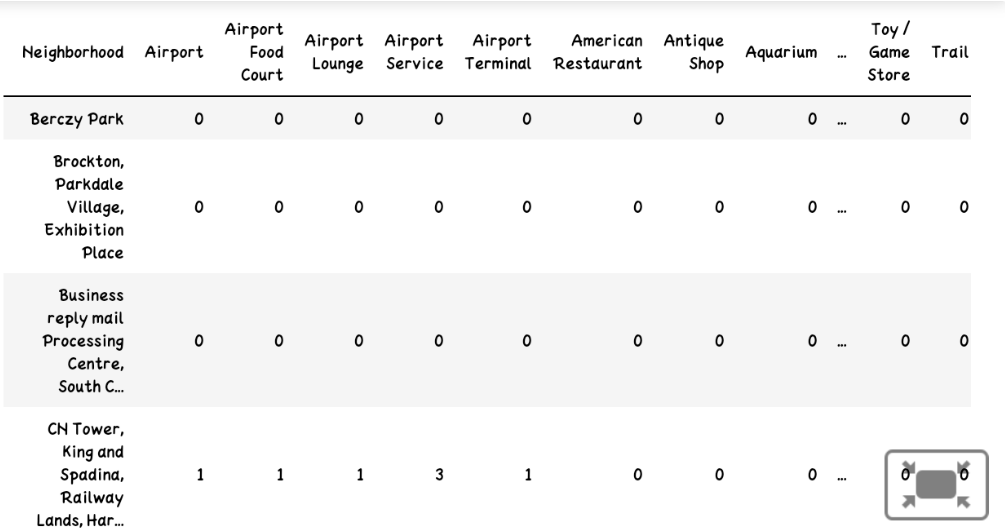
The neighborhoods are superimpose on Toronto city map using folium library.



Map of Toronto with neighborhoods superimposed

2.4 : Retreiving venues locations present in each neighborhood

We retreive the venues locations of every neighborhood using the foursquare api . The data consists of a set of venues presencd in each neighbourhood and are stored in the pandas data frame. The neighborhood name are used and index and the colums are various venues . the values contains number of such venues in a neighborhood.



Neighborhood	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Toy / Game Store	Trail
Berczy Park	0	0	0	0	0	0	0	0	...	0	0
Brockton, Parkdale Village, Exhibition Place	0	0	0	0	0	0	0	0	...	0	0
Business reply mail Processing Centre, South C...	0	0	0	0	0	0	0	0	...	0	0
CN Tower, King and Spadina, Railway Lands, Har...	1	1	1	3	1	0	0	0	...	0	0

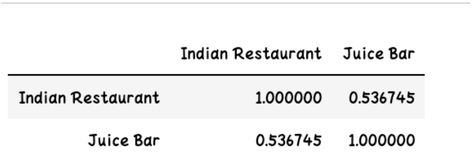
Neighborhoods and venues

3. Exploratory Data Analysis

3.1 : Correlation between venues and Indian restaurant

We now start analysing acquired data, first we find the venues which correlate with the Indian restaurant ina all neighborhood. Here Indian restaurant is dependent variable and remaining venues are

independent variables
we use pandas built in function to find the correlation coefficient of venues with indian restaurant.
Eg : lets consider the the venue **Juice Bar** , let find the pearson correlation of Juice Bar with Indian Restaurant .

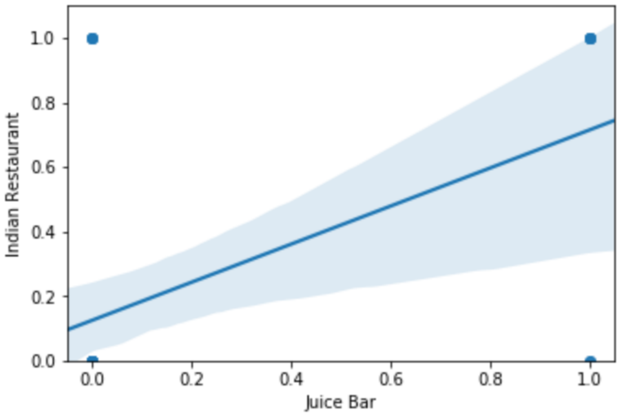


	Indian Restaurant	Juice Bar
Indian Restaurant	1.000000	0.536745
Juice Bar	0.536745	1.000000

correlation coeffcient bw indian restaurant and juice bar

we can find the value is nearly **0.537** . There is a partial correlation between the two variables and we can use this variable for our further analysis.

Lets plot a scatter plot between them :

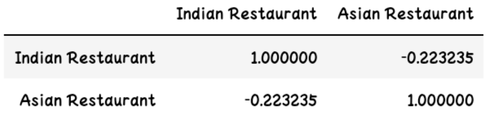


Scatter plot

we can say that there is a positive relation between the two variables.

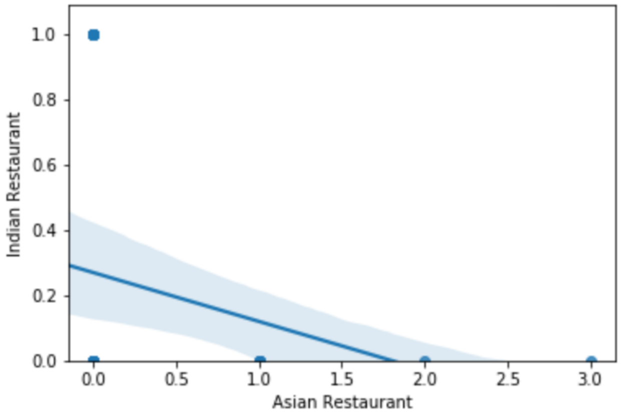
Eg2:

lets take another variable i.e, Asian Restaurant and compare the correlation coeffcient with Indian Restaurant.



	Indian Restaurant	Asian Restaurant
Indian Restaurant	1.000000	-0.223235
Asian Restaurant	-0.223235	1.000000

The correlation coeffcient value is **-0.223** . and the scatter plot between variables is shown below.



Scatter plot

From the plot we can conclude the there is negative relationship between variables . So we can’t use this variable for further analysis.

We then find the top 25 venues with are in positive correlation with the Indian Restaurant. The top 4 venues obtained are shown in table 1.0

Venue	**Correlation Coeff**
Juice Bar	0.5367450401216934
Liquor Store	0.5367450401216934
Sandwich Place	0.47718429649788857
Pharmacy	0.45484950135257857

table 1.0

3.2 **Model development :**

In this section we use logistic regression and linear regression to develop a model .And find out which model is best suited for our analysis. The available data is splitted into training and testing set.

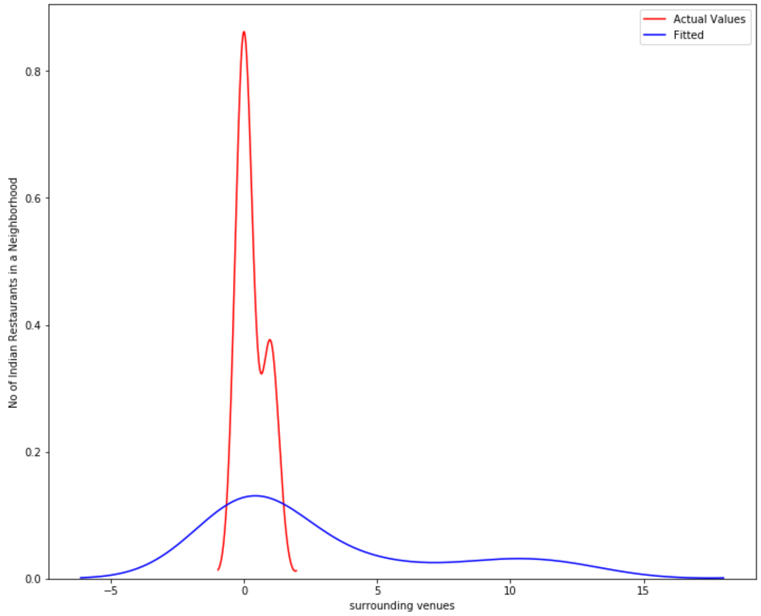
The y variable for the model development is taken from the values of no of Indian restaurant in a neighborhood, and the values for x variables are taken from the series of 25 venues which showed possitive correlation with y variable. The data is consists of 39 rows each rows represents a neighborhood and each Column represent the top 25 venues

A. Linear Regression model:

Linear regression fits the data to a straight line for single independent variable models and an hyper plane for multiple independent variables. It is used for continuous values of data.

The dataset is divide into 75% (29 rows)for training and (10 rows) 25% for testing. The model is trained with train data nd the tested with the test data .

Then the y values of predicted test data and original values are compared. This can be visulized using distribution plot.



Distribution plot

In the above plot the red curve indicates actual values of test data and y value indicates predicted values of test data. The plot proves that the model is not a right fit for our analysis. And the R square value is found to be -43.5 . Which is out of range for a R square value.

This explains that linear regression is not a good model for out data.

Note : Range of R square is [0,1] and the value 1 represents best fit.

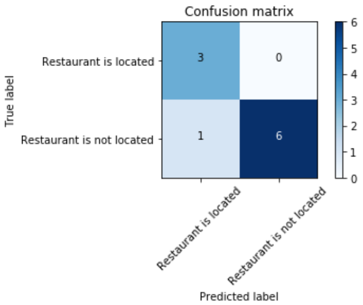
B. Logistic Regression

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

Logistic regression fits a special s-shaped curve by taking the linear regression and transforming the numeric estimate into a probability with the following function, which is called sigmoid function

The train data is used for fitting the model and tested by the test data. The yhat of test data are predicted using the model . We then evaluate the accuracy of the model by calculating jaccard index.

In our present model the **jaccard index** value is found to be **0.9** nearly **90** accuracy . Which depicts the model is a good fit for the data. We can analyse the predicted and real values of dependent variable using confusion matrix given below.



Confusion matrix

In the above matrix the values represent one neighborhood each from test data. The total neighborhood are 10 for test data. Let's analyse the matrix.

Look at first row. The first row represents the no of neighborhoods where an Indian restaurant is located .

Out of 10 test sample neighborhoods 3 neighborhoods have an Indian restaurant located . Out of those 3 the classifier(model) predicted all of them .

Let's move to the second row which depicts no of neighborhoods which doesn't have a Indian restaurant.

There are 7 neighborhoods out of 10 test samples . Out of these 7 the model predicted 6 of them true and 1 false . We consider this as error of model for second row.

Out of 10 test samples model predicted 9 samples as correct . We can say the model is a good fit for our data .

4. **Conclusion**

Let's recall of hypothesis and our goal of this analysis. Our hypothesis is that the location and the existance of a certain venues can affect the boon for starting new venues. This trend can be analysed from the existing venues data set . In our analysis we focussed on the existance of Indian restaurant in various neighborhoods . We believed that the venues located around the restaurant has some affect on the survival of restaurant.

We trained a logistic Regression model which provides insight on the probability of existance of an Indian restaurant in a neighborhood. The model successfully predicted . Hence we can say certain venues co exists and benefits each other mutually. This can be accepted when we considered an example A shopping mall next to a restaurant can affect the customers visiting the restaurant . The shopping mall customers might visit restaurant. Vice versa the customers of a shopping mall can visit the shopping mall after the food.So venues collectively attract customers.

Our model predicted the existence of an Indian restaurant based on the existance of othervenues in a neighborhood. If one wants to start an Indian restaurant in a neighborhood and the model predicts existance of an Indian restaurant in the neighborhood . There is chance that other venues help in attracting customers towards the neighborhood.
If there is no restaurant in that neighborhood than selecting that neighborhood for new startup of a restaurant best choice. Since the model is predicting one there.

5. Future directions

In addition to the existance of certain venues we can also further push our analysis toward the most visited venues in neighborhood. And adding the rent data of the shops in the neighborhood can help to choose best neighborhood for specific startups.

References :

1. The data processing and visualization are done with the help of course material provided by coursera on data science.