Bangalore Catchment Area Analysis

IBM Applied Analytics Course

Ajit Kumar Jain

14 June 2019

Preface

This project and introduction to Foursquare API gives me this tremendous opportunity to explore foremost and fundamentally important analysis done by retail stores that is catchment area analysis. This analysis is basis for deciding where to open the retail store. Classically and mostly even till now this is done as subjective analysis based on advice, intuition and experience but if this initial decision goes wrong then substantial amount of investment as well as future of store will be stake. I think this is a ideal case for machine learning and I take this opportunity to explore opportunity of Machine Learning in Catchment Area Analysis.

Audience and Stakeholders:
The obvious audience and stakeholders of this effort are senior level executives responsible for growth, diversification and expansion plans. This analysis will also be a good reference for small and medium size retailers. This analysis can also support for promotion planning to think about outlets where natural footfall is not much and additional promotion is required to bump up footfall. This analysis will sufficiently serve the purpose for opening a Grocery store. With the rise of E-Commerce retailers need to make choice of location in much more informed manner and for the convenience of customer as well. For a given city this analysis will clearly separate areas where high footfall is expected vs areas saturated with sufficient number of outlets. So in a saturated as well as low footfall areas high amount of promotion will be required whereas areas with high footfall and low saturation there is opportunity to make more margin and even open a new store.

Project Coverage
City: Bangalore
Scope: Project will not be covering customer demographics or any competitors but will cover analysing various physical infrastructure of an area in sufficient detail. Because target store is a Grocery Store will be looking at PTA (Primary Trade Area) of 500 Meters.
Primary Variables: Pin code Wise Count of Grocery Outlets Stores to be categorized as stand alone, mall, mega store, retail chain.
Secondary Variables: Pin code wise count of Fashion Stores, Electronic Store ,Apartment's, Villas and Temple's, Complementary stores like, coffee shops, restaurants, Lounges, Banquets, Pubs, Cinema etc. Count of office buildings, IT parks, Banks and ATM's, Hospitals, Nursery, School and College, Railway/Metro Station's/bus top, sports facilities and gyms, Furniture store, showrooms, garage etc...
Will try to get as much data possible from Foursquare and upon getting latitude and longitude data of stores data will be converted to a geo tagged data frame and density evaluation will be the main criteria.

Table of Contents

Introduction

This project and introduction to Foursquare API gives me this tremendous opportunity to explore application of Machine Learning in fundamentally important analysis done by retail stores that is catchment area analysis. This analysis can be basis for deciding where to open the retail store. Classically and mostly even till now catchment area analysis is done as subjective analysis based on advice, intuition and experience, but if this initial decision goes wrong then substantial amount of investment as well as future of store will be at stake.

The obvious audience and stakeholders of this effort are senior level executives responsible for growth, diversification and expansion plans. This analysis will also be a good reference for small and medium size retailers. This analysis can also support for promotion planning to think about outlets where natural footfall is not much/competition is fierce and additional promotion is required to bump up footfall. This analysis will sufficiently serve the purpose for opening a medium size super market. With the rise of E-Commerce retailers need to make choice of location in much more informed manner and for the convenience of customer as well. For a given city this analysis will clearly separate areas where high footfall is expected vs areas saturated with sufficient number of outlets vs areas with high competition. So in a saturated as well as high competition areas high amount of promotion will be required whereas areas with high footfall and low saturation there is opportunity to make more margin and even open a new store.

Catchment area analysis not only includes surveying physical establishments but also commercial rates, connectivity, upcoming growth in the locality and also target customer base survey to decide on merchandise mix. Due to resource constrain this project is limited to understanding suitability of locality for opening a new supermarket. This will be done by gathering basic physical establishments data for whole of Bangalore from FourSquare API.

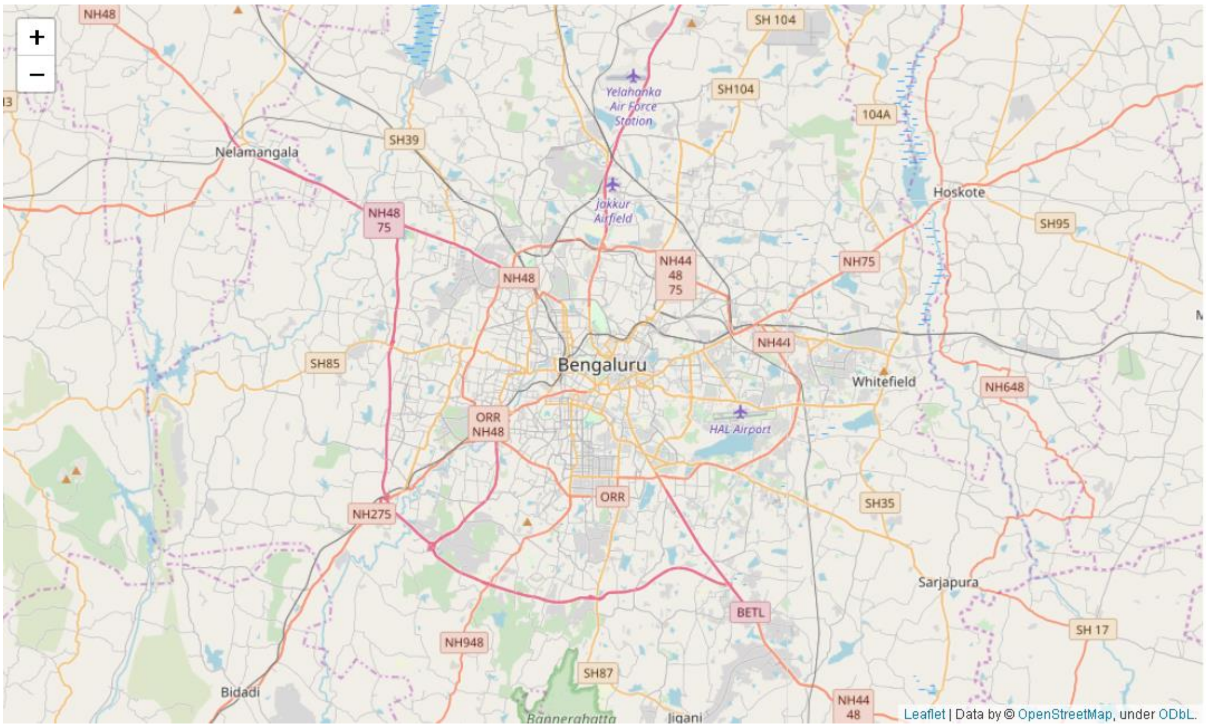
Will try to get as much data possible from Foursquare and upon getting latitude and longitude data of stores data will be converted to a geo tagged data frame and venue category wise density evaluation will be the main criteria to understand a locality.

Changing project scope:
From initial analysis, variable grocery was not found to be strong enough to do analysis upon as number of data points were very less. Further grouping venue buckets into super buckets namely venue for food, shopping, amenities of area and miscellaneous places.

Changing objective statement: To evaluate a locality in terms of shopping on basis number of food venues, amenities and other venues in the area.

Methodology

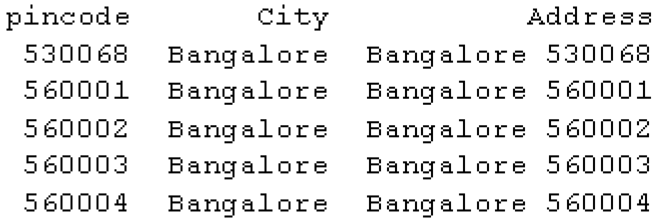
Map of Bangalore:



Steps:

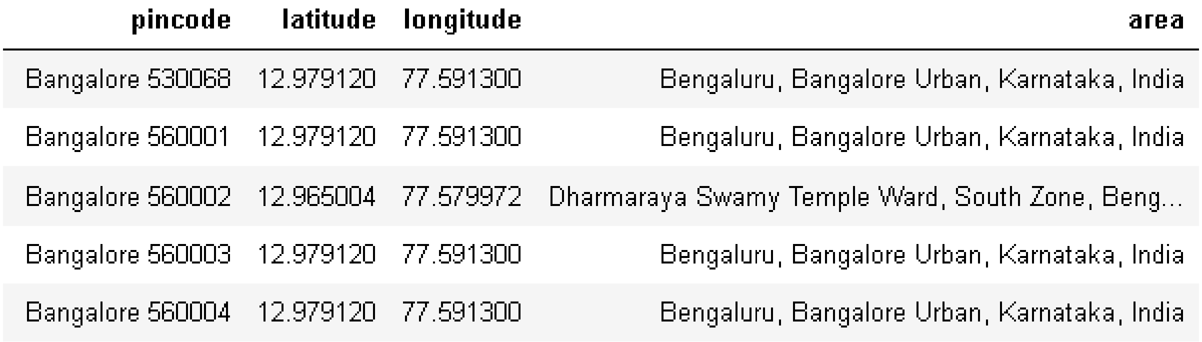
1. Gather latitude and longitude data and make grid for Bangalore city.

Sample Pin-code List: Total 131 pincodes.



```
pincode        City           Address
 530068   Bangalore  Bangalore 530068
 560001   Bangalore  Bangalore 560001
 560002   Bangalore  Bangalore 560002
 560003   Bangalore  Bangalore 560003
 560004   Bangalore  Bangalore 560004
```

Sample: Latitude/longitude data: Complete grid of 10000 points.

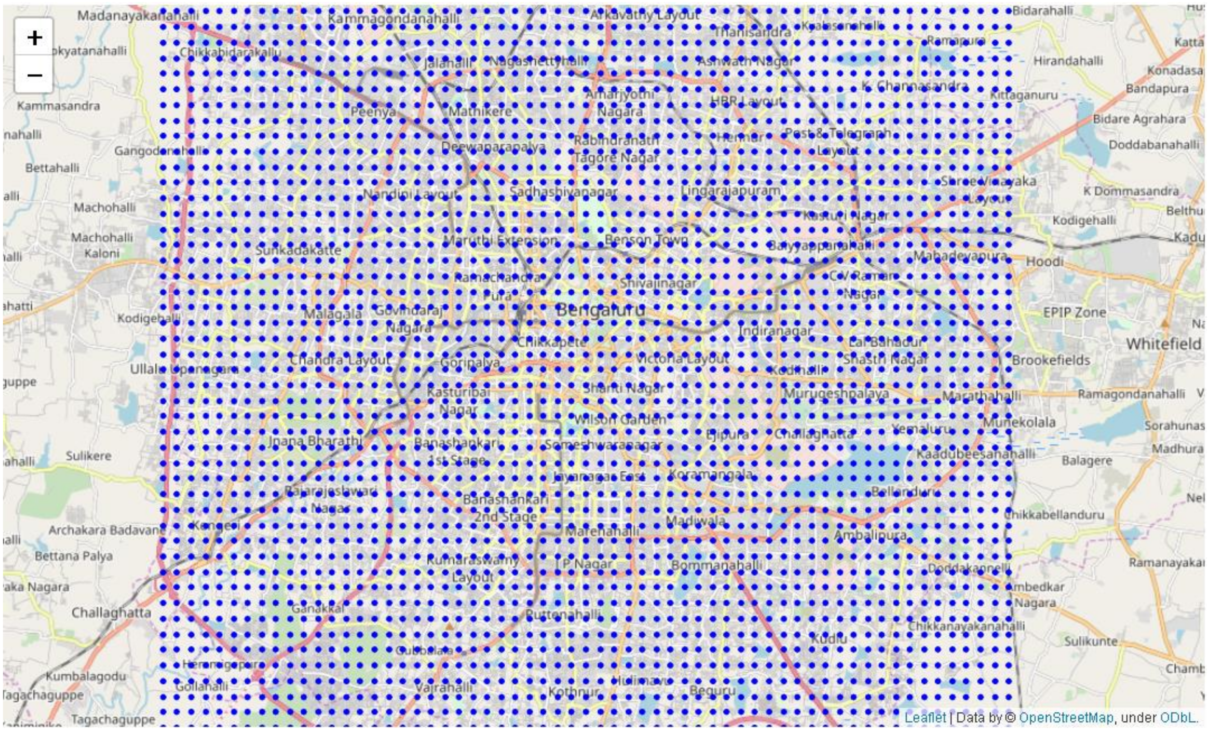


pincode	latitude	longitude	area
Bangalore 530068	12.979120	77.591300	Bengaluru, Bangalore Urban, Karnataka, India
Bangalore 560001	12.979120	77.591300	Bengaluru, Bangalore Urban, Karnataka, India
Bangalore 560002	12.965004	77.579972	Dharmaraya Swamy Temple Ward, South Zone, Beng...
Bangalore 560003	12.979120	77.591300	Bengaluru, Bangalore Urban, Karnataka, India
Bangalore 560004	12.979120	77.591300	Bengaluru, Bangalore Urban, Karnataka, India

Grid view of Bangalore:

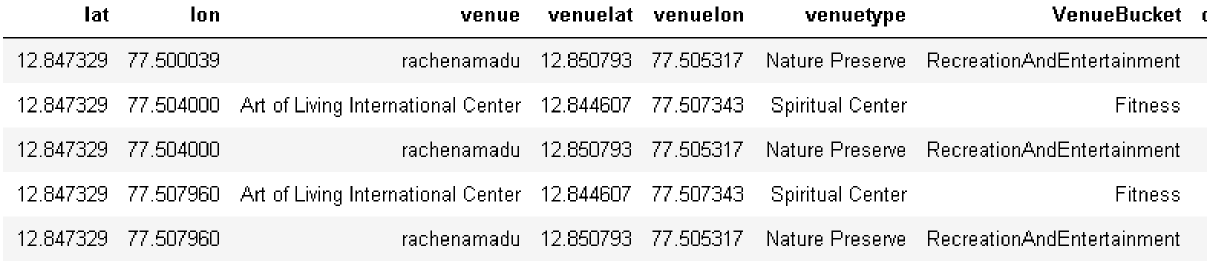
From the Latitude and Longitude information gathered, the range of Latitude and Longitude is divided into 100 parts to make grid of Latitude and Longitude consisting of 10000 points. This grid is trimmed to boundaries of outer ring road as most of the development and business is concentrated within this boundary. Distance between each diagonally opposite point is 638 meters. To be on safer side will be gathering information of venues within a radius of 750 meters around each point on grid.

Grid view of Bangalore:



2. At each point of grid gather location venue data.
3. Bucket venues on similar traits.

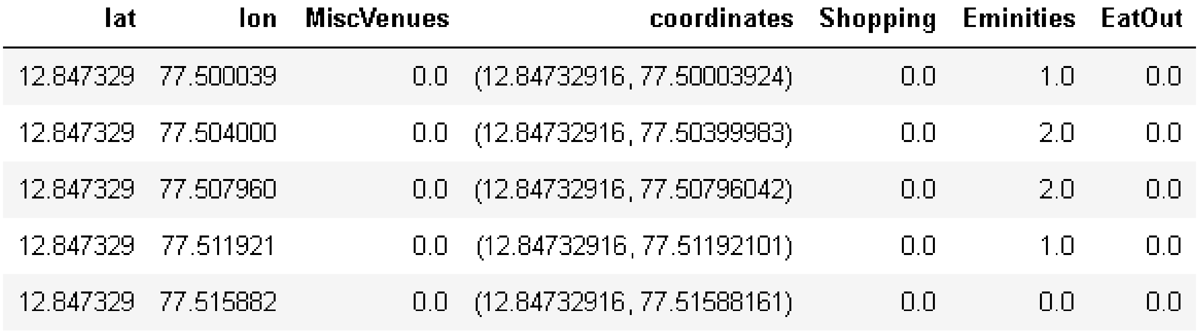
Sample data post step two and three:



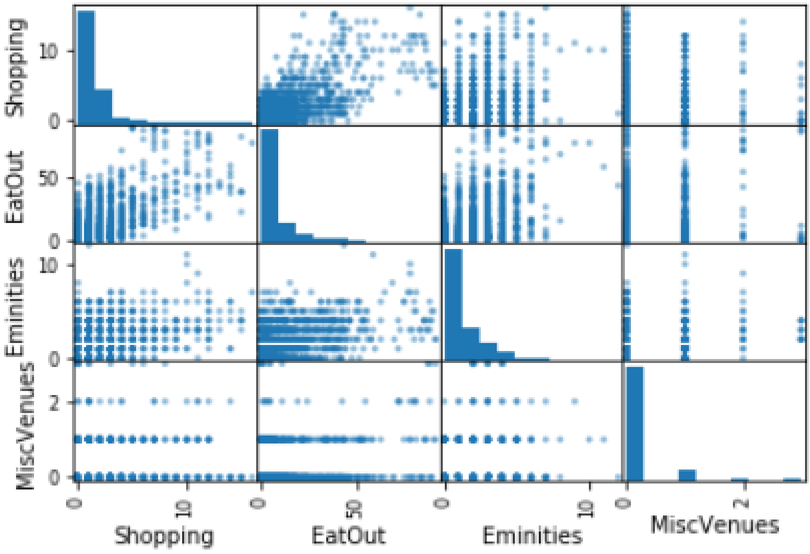
lat	lon	venue	venuelat	venuelon	venuetype	VenueBucket	c
12.847329	77.500039	rachenamadu	12.850793	77.505317	Nature Preserve	RecreationAndEntertainment	
12.847329	77.504000	Art of Living International Center	12.844607	77.507343	Spiritual Center	Fitness	
12.847329	77.504000	rachenamadu	12.850793	77.505317	Nature Preserve	RecreationAndEntertainment	
12.847329	77.507960	Art of Living International Center	12.844607	77.507343	Spiritual Center	Fitness	
12.847329	77.507960	rachenamadu	12.850793	77.505317	Nature Preserve	RecreationAndEntertainment	

4. Compare relation of Shopping variable to other variables like food, amenities etc.

Sample of Bucketed and Transformed Data:



	lat	lon	MiscVenues	coordinates	Shopping	Eminities	EatOut
	12.847329	77.500039	0.0	(12.84732916, 77.50003924)	0.0	1.0	0.0
	12.847329	77.504000	0.0	(12.84732916, 77.50399983)	0.0	2.0	0.0
	12.847329	77.507960	0.0	(12.84732916, 77.50796042)	0.0	2.0	0.0
	12.847329	77.511921	0.0	(12.84732916, 77.51192101)	0.0	1.0	0.0
	12.847329	77.515882	0.0	(12.84732916, 77.51588161)	0.0	0.0	0.0



From the below scatter plot it can be seen that variable shopping can related to EatOut(Food Venues) variable and to a little extent to Amenities present in the area. Taking MiscVenues(Miscellaneous Venues) from the analysis.

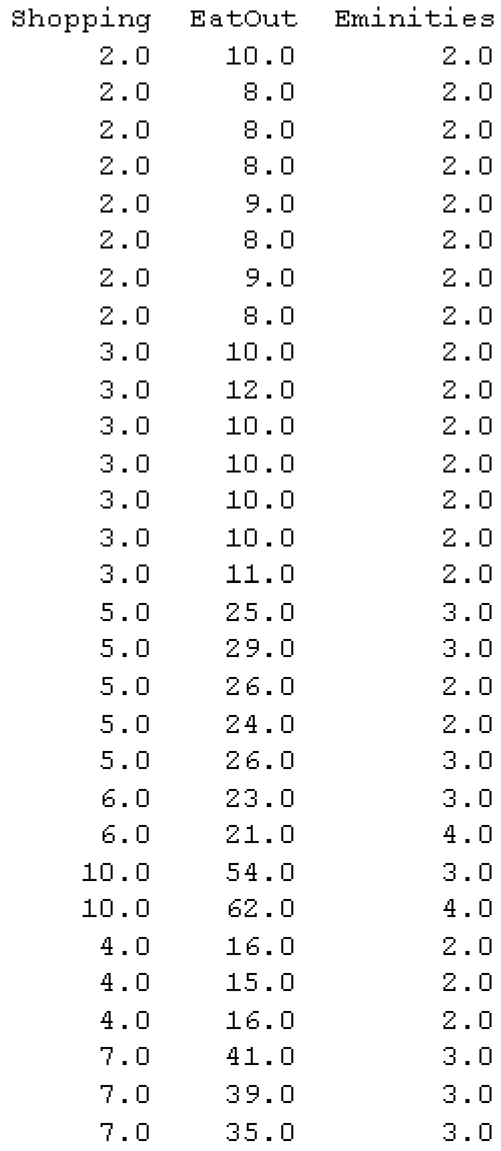
5. EDA: Clean data by taking out noise among variables and doubtful localities.

Foursquare data does not seem good enough as in any locality of 750 meters radius there are always a few amenities, food venues and shopping venues. Thus taking out observations i.e. Latitude Longitude points in grid with less than two shopping venues, less than 3 food venues or less than 1 amenity. Taking three variables i.e. shopping, eatout and amenities in separate dataset to build regression model.

6. EDA: Select variables sufficiently related to shopping variable.

As we saw data is very dispersed and any amount of transformation or feature scaling will not serve the purpose. Taking median (with 10 percent tolerance) of Eatout and Amenities variable at each level of Shopping variable.

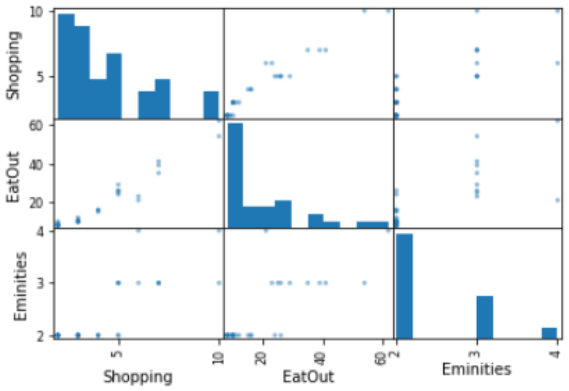
Clean and filtered data:



```
 Shopping  EatOut  Eminities
      2.0    10.0        2.0
      2.0     8.0        2.0
      2.0     8.0        2.0
      2.0     8.0        2.0
      2.0     9.0        2.0
      2.0     8.0        2.0
      2.0     9.0        2.0
      2.0     8.0        2.0
      3.0    10.0        2.0
      3.0    12.0        2.0
      3.0    10.0        2.0
      3.0    10.0        2.0
      3.0    10.0        2.0
      3.0    10.0        2.0
      3.0    11.0        2.0
      5.0    25.0        3.0
      5.0    29.0        3.0
      5.0    26.0        2.0
      5.0    24.0        2.0
      5.0    26.0        3.0
      6.0    23.0        3.0
      6.0    21.0        4.0
     10.0    54.0        3.0
     10.0    62.0        4.0
      4.0    16.0        2.0
      4.0    15.0        2.0
      4.0    16.0        2.0
      7.0    41.0        3.0
      7.0    39.0        3.0
      7.0    35.0        3.0
```

Data has shrunk considerably to 30 observations only but relation between variables is now very clear and linear. Amenities does not show very good relation but it seems there are levels to the variable Amenities thus will be keeping it for regression modelling.

Plotting filtered data:



7. Unsupervised Regression Model: Create model with selected variables and on filtered data.

Coefficients:
Intersection- 0.3007132638252754
EatOut-   0.1367758
Amenities-   0.51209101

Using the model to predict number of shopping venues at each location (point in grid). Predicted values are in variable RegShops.

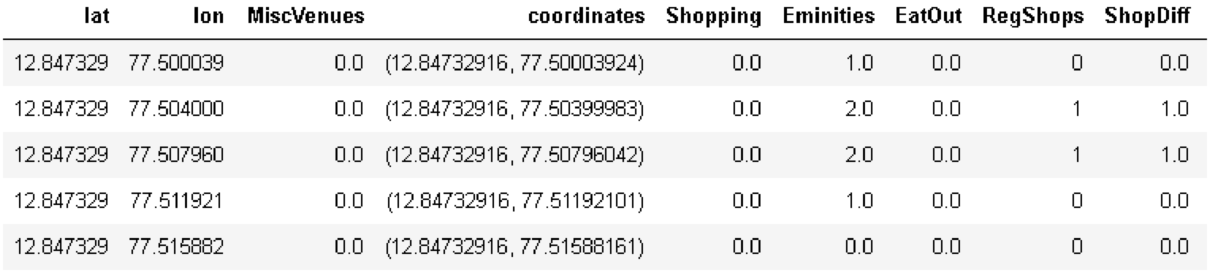
8. Apply model to complete data categorize locations as Opportunity, High Competition etc.

Variable ShopDiff is created which is difference between predicted shopping venues (RegShops) and actual shopping venues.

Tagging locations as, if number of actual shopping venues is within twenty percent tolerance of predicted venues count then that location is saturated, if greater than twenty percent then competition is high, if actuals is les then twenty percent of predicted then there is opportunity.

If in any observation if any two variables among Shopping, EatOut and Amenities is zero then location is marked indecisive. Also tagging indecisive as per earlier rule that at any location there should be at least two shopping, three eatout and one amenity venue.

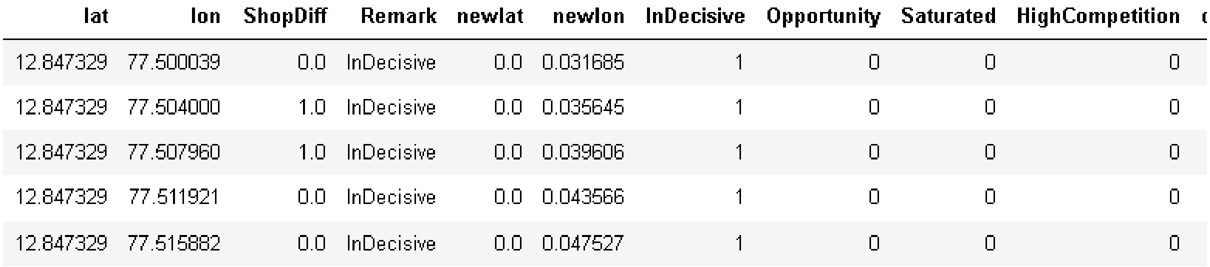
Sample data post prediction:



lat	lon	MiscVenues	coordinates	Shopping	Eminities	EatOut	RegShops	ShopDiff
12.847329	77.500039	0.0	(12.84732916, 77.50003924)	0.0	1.0	0.0	0	0.0
12.847329	77.504000	0.0	(12.84732916, 77.50399983)	0.0	2.0	0.0	1	1.0
12.847329	77.507960	0.0	(12.84732916, 77.50796042)	0.0	2.0	0.0	1	1.0
12.847329	77.511921	0.0	(12.84732916, 77.51192101)	0.0	1.0	0.0	0	0.0
12.847329	77.515882	0.0	(12.84732916, 77.51588161)	0.0	0.0	0.0	0	0.0

Next transforming variable similar to one hot encoding to prepare for DBSCAN clustering.

Sample of transformed data ready for clustering:

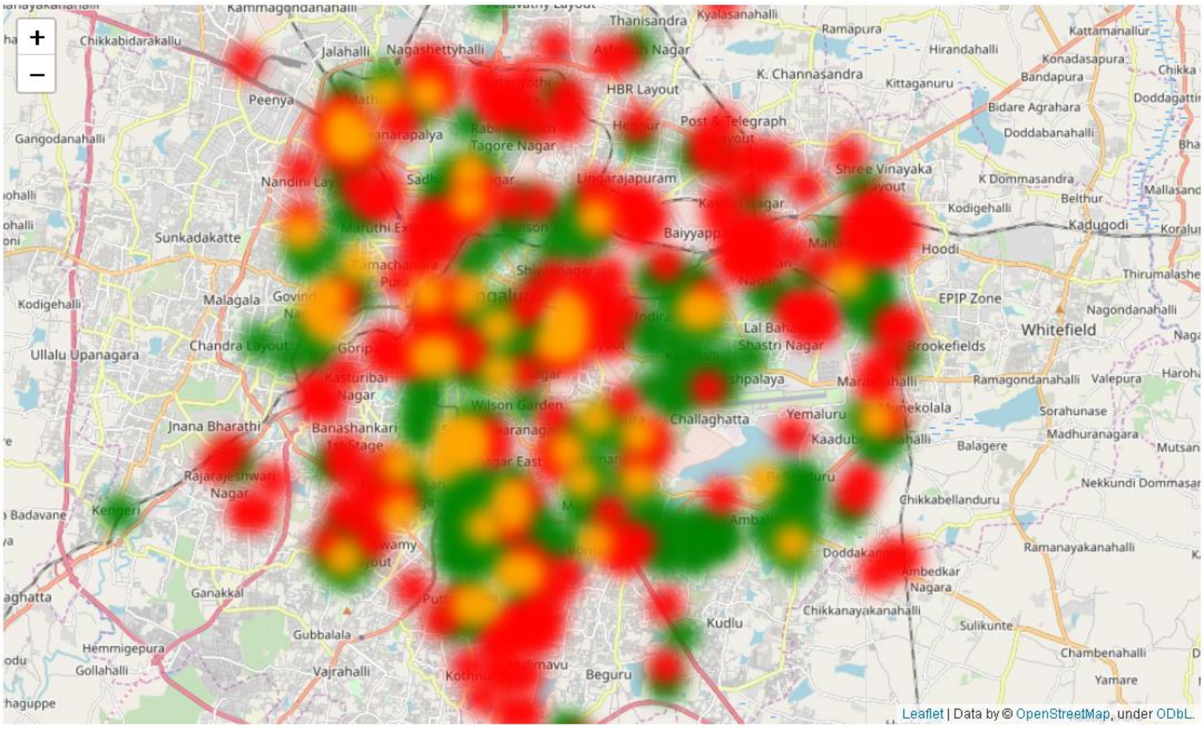


lat	lon	ShopDiff	Remark	newlat	newlon	InDecisive	Opportunity	Saturated	HighCompetition
12.847329	77.500039	0.0	InDecisive	0.0	0.031685	1	0	0	0
12.847329	77.504000	1.0	InDecisive	0.0	0.035645	1	0	0	0
12.847329	77.507960	1.0	InDecisive	0.0	0.039606	1	0	0	0
12.847329	77.511921	0.0	InDecisive	0.0	0.043566	1	0	0	0
12.847329	77.515882	0.0	InDecisive	0.0	0.047527	1	0	0	0

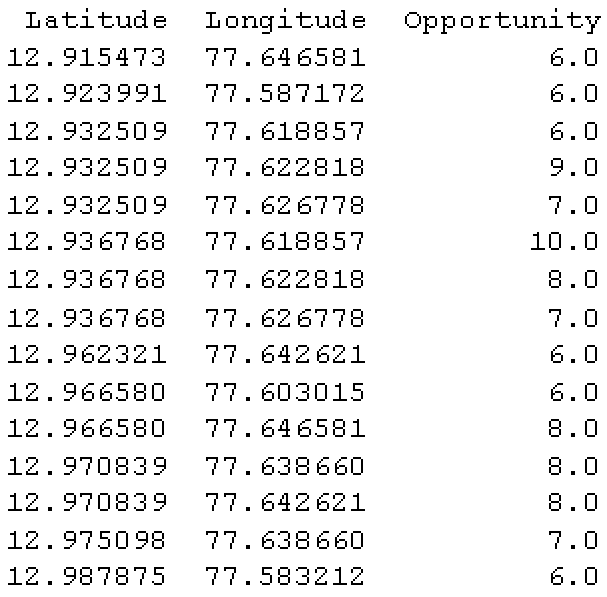
9. Unsupervised Clustering: Apply DBSCAN to output of Regression tagging.

DBSCAN is performed to takeout outliers and be sure of the solution by setting min sample parameters to 5. Which means at least five points in any cluster should show up to be marked as Opportunity, High competition etc.

Output of DBSCAN plotted on Bangalore Map.



10. Recommend locations with good opportunity of opening supermarket.



```
 Latitude  Longitude  Opportunity
12.915473  77.646581          6.0
12.923991  77.587172          6.0
12.932509  77.618857          6.0
12.932509  77.622818          9.0
12.932509  77.626778          7.0
12.936768  77.618857         10.0
12.936768  77.622818          8.0
12.936768  77.626778          7.0
12.962321  77.642621          6.0
12.966580  77.603015          6.0
12.966580  77.646581          8.0
12.970839  77.638660          8.0
12.970839  77.642621          8.0
12.975098  77.638660          7.0
12.987875  77.583212          6.0
```

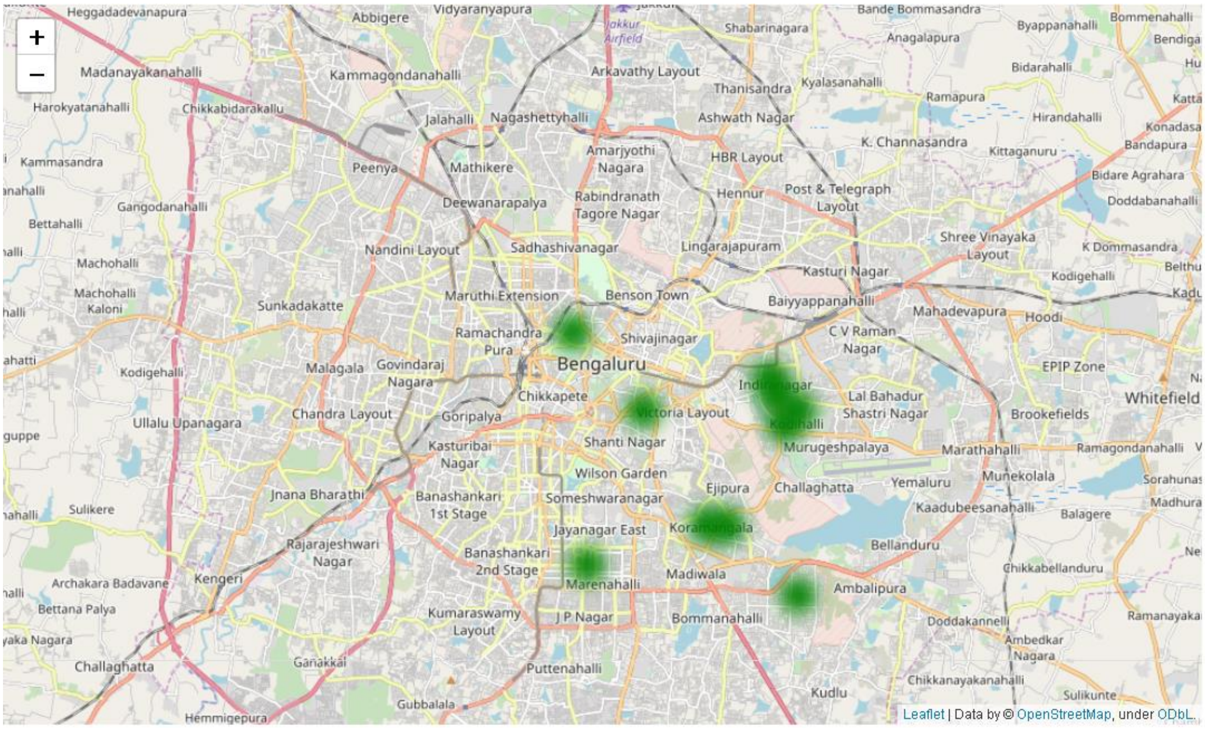
Data from Foursquare API is not sufficient as tagging might be good but I seriously feel number of venues from API output are not correct. Google API could have served better along with footfall data but google tagging might not be that great. For not have to do with Foursquare API and results are indicative and not deterministic.

Note: As we are interested in finding opportunity to open medium size supermarket and supermarket serves merchandise ranging from grocery to non-food items of departmental stores thus all kind of shops were bucketed together in one variable called shopping.

Results

Output from DBSCAN is filtered for top 15 localities which are best for opening a new supermarket. Latitude, longitude along with number of supermarkets that can be opened are indicated below.

Map view of top 15 locations coordinates listed for in methodology section.



Discussion

With this basic analysis we are able to say what localities are good option for new retail store but this analysis is not extensive but only indicative. We are able to indicate geocode with radius yet how can we make this practice more sharp indicating which road and nearest landmark. How can current data issues be resolved and new data sources can be exploited. Right now only two data sources are used, one is for pin codes and another for venue list for given geocode and radius. Definitely external data sources for customer demographics, foot fall for each venue, real estate prices with potential upcoming localities are good data sources to be integrated along with Machine Learning.

Conclusion

With all resource constrains and data issues, Machine Learning model was able to come up with list of localities (geocode + radius) which can be potential destination to open your next supermarket. Further survey will be required on top five localities indicated in results section to be sure of connectivity, future developments, restate cost etc..

With better and more diverse data sets Machine Learning can deliver enormous value to catchment area analysis.

Reference

Data Sources:

1. https://foursquare.com/
2. https://www.mapsofindia.com/pincode/india/karnataka/bangalore/
3. http://www.mybengaluru.com/resources/2066-Pincode-Area-Wise.aspx

Machine Learning:

1. https://scikit-learn.org/stable/
2. https://python-visualization.github.io/folium/

Acknowledgement

With great pleasure I can say that I am very happy to be part of this course. Learning was great in terms of theory, programming and flatform knowledge as well. I also appreciate IBM Watson for extending the resources and make this learning a success.

Thanks So Much,
Ajit Jain.