Bangalore Catchment Area Analysis

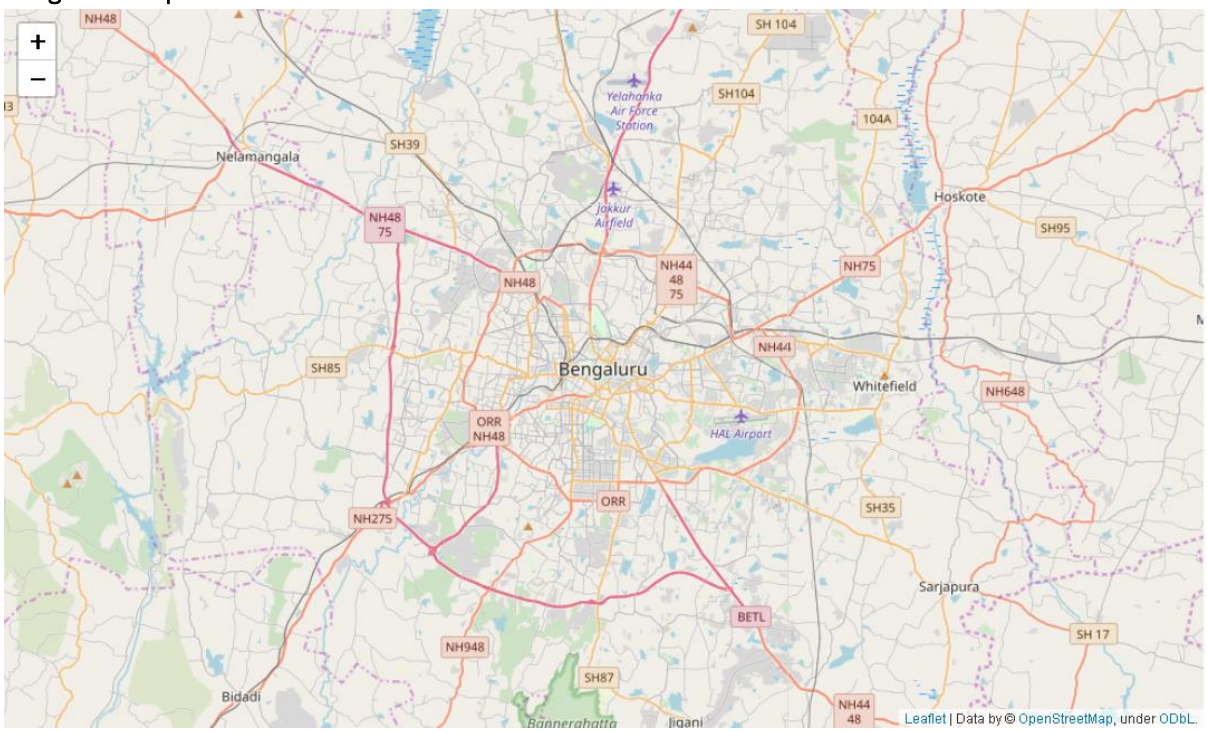
I take this opportunity to explore application of Machine Learning on fundamentally important analysis done by retailers, that is catchment area analysis. This analysis is basis for deciding where to open the retail store and what shall be the merchandise mix. Classically and mostly even till now this is done as subjective analysis based on survey, 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. I think this is an ideal case for machine learning.

The obvious audience 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 as well. Machine Learning in catchment area analysis can also support for promotion planning and to think about outlets where natural footfall is not much and additional promotion is required to bump up footfall. Such analysis will sufficiently serve the purpose for opening a medium size supermarket.

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 we can 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 an opportunity to make more margin and even open a new store.

This analysis is done for Bangalore city and is limited to analysing types and count of physical establishments. While catchment area analysis covers survey of customers, competition, real estate cost, connectivity etc. In that sense looking at only physical venues or infra is small part but aim is only to discover how analytics can be used for catchment area analysis.

Bangalore Map:



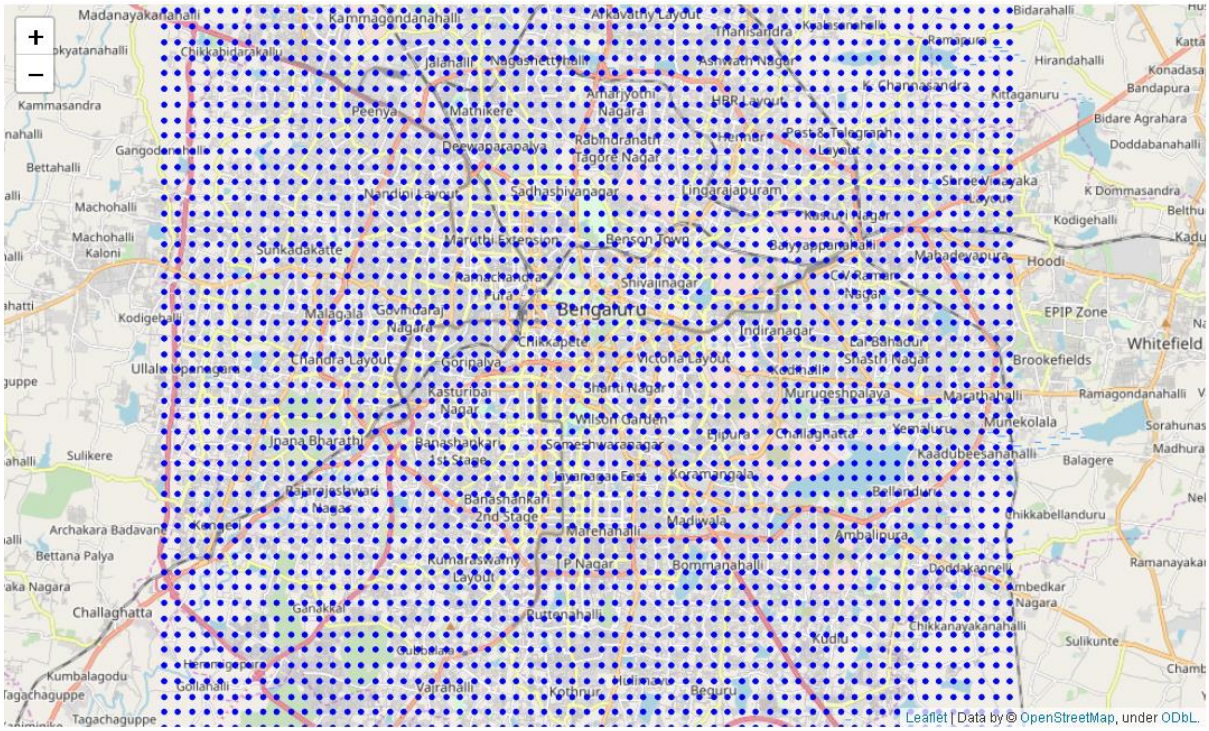
Methodology:

Map of Bangalore was divided into grid of 10000 points (100X100 Latitude X Longitude). Each diagonally adjacent points are 638 meters away so to be on safer side 750 meters was taken as radius. Within radius data of all venues for each point on grid was arranged through FourSquare API.

Sample data of Bangalore grid view:

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	

Bangalore Grid View:

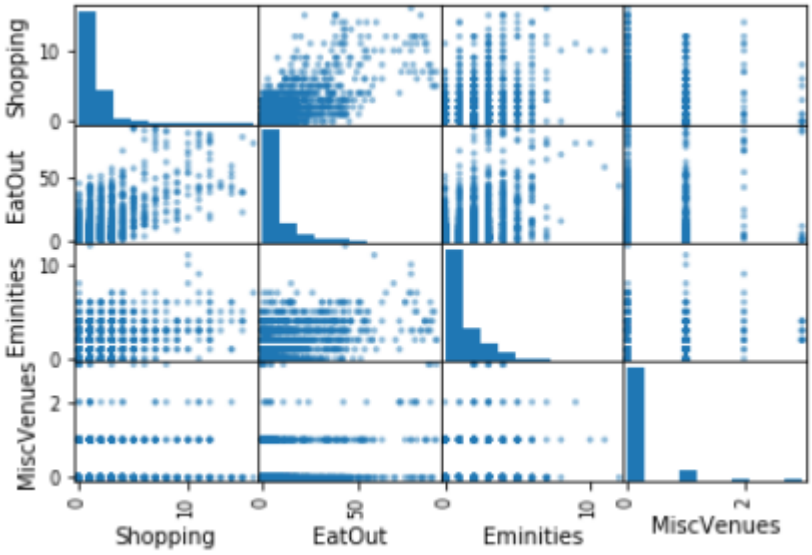


Each point on the grid is an observation and count of venues within a radius of 750 meters is our data. Count of shopping venues is dependant variable while count of any other type of venue is independent variable. Different type of venues are bucketed into four major categories. First is dependant variable which is shopping where all types of big or small shopping venues are bucketed together and count is taken as our data. Next is food venues again big or small as well as bar or pubs are counted in. Then comes the amenities delivered by the area, like connectivity, sports or gym facilities, entertainment etc. are bucketed. Rest all venue types are bucketed as miscellaneous. While we are trying to find out opportunity of coming up with new supermarket in a certain location, grouping all shopping outlets could make sense because typically a supermarket offers a wide range of merchandise ranging from grocery, FMCG, packaged goods, stationary and departmental store as well .Assumption is with a good locality if number of food outlets, amenities etc increase then that location is definitely attractive and number of shopping outlets also shall follow.

With this relation in mind lets gather our data and first visualize what we are looking at makes sense or not. From the below scatter plot it can be clearly seen that our independent variable (shopping) strongly relates to number of food outlets (EatOut) and weakly with number of amenities. There seems to be no relation with Miscellaneous Venues.

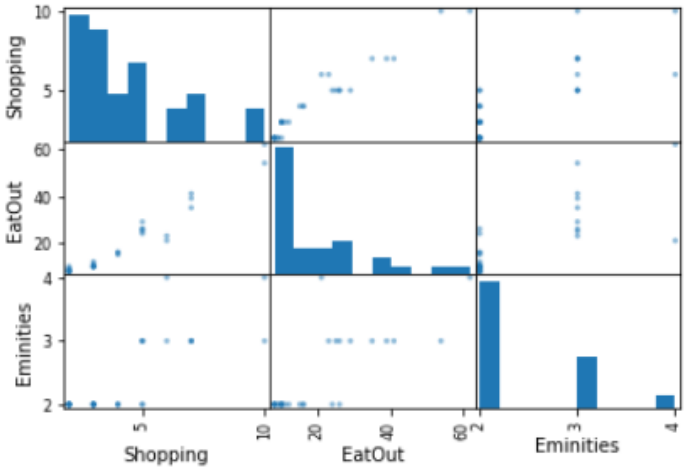
Visualization and sample data of variables;

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



Lets clear out the clutter by taking median (with 10% tolerance) value of EatOut and Amenities variable at each level of Shopping variable we get below picture with clear and strong relation among variables.

Visualize filtered data for relation among selected variables:



This relation is used to predict (RegShops) ideal number of shopping outlets for a given locality. Thus we can calculate additional shopping outlets that can be opened (ShopDiff).

Sample output of regression:

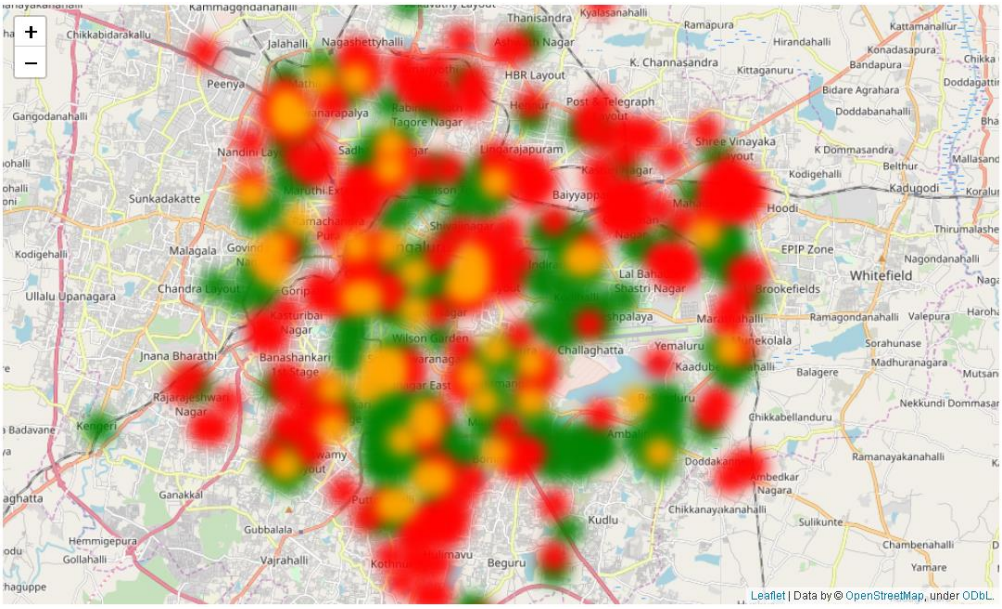
lat	lon	MiscVenues	coordinates	Shopping	Eminities	EatOut	RegShops	ShopDiff
12.847329	77.500039	0.0	(12.84732916, 77.50003924)	0.0	1.0	0.0	0	0.0
12.847329	77.504000	0.0	(12.84732916, 77.50399983)	0.0	2.0	0.0	1	1.0
12.847329	77.507960	0.0	(12.84732916, 77.50796042)	0.0	2.0	0.0	1	1.0
12.847329	77.511921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

Variable ShopDiff is used to tag a locality as High-Competition, Saturated or a Opportunity location. This information is further transformed to be used for density based clustering basis latitude/longitude and on strict parameters.

Sample of transformed data for clustering:

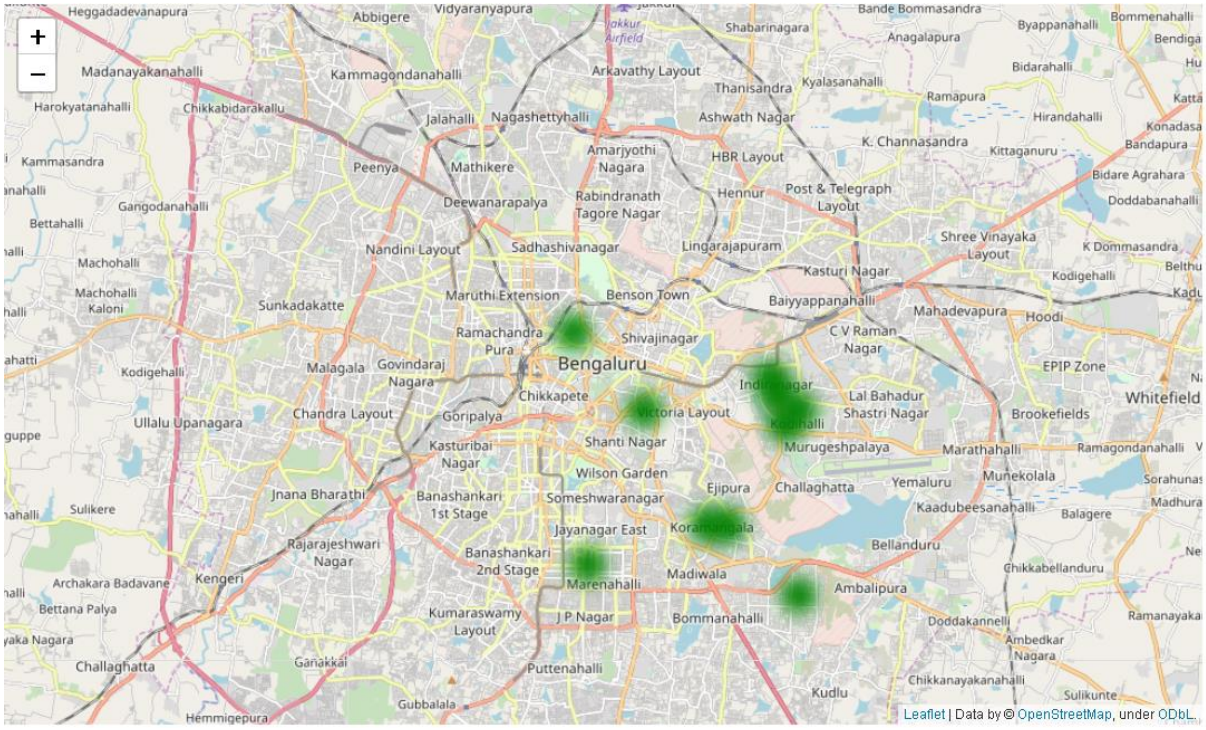
lat	lon	ShopDiff	Remark	newlat	newlon	InDecisive	Opportunity	Saturated	HighCompetition	c
12.847329	77.500039	0.0	InDecisive	0.0	0.031685	1	0	0	0	
12.847329	77.504000	1.0	InDecisive	0.0	0.035645	1	0	0	0	
12.847329	77.507960	1.0	InDecisive	0.0	0.039606	1	0	0	0	
12.847329	77.511921	0.0	InDecisive	0.0	0.043566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

By the time we reach this stage a lot of data cleaning took place which I am not discussing. Being said that as you can see in above table variable 'InDecisive' is the one that we cannot use for clustering as there are a lot of observations that are illogical. Below is the visualization when we plot output of clustering on Bangalore map. Red are the locations with high competition, orange is saturated and green is opportunity area.



Listing down top fifteen localities that are most promising and with a visual on map as well.

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



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