Applied Data Science

Capstone project

Objective

This exercise/project is being done to fulfill the requirement of applied data science capstone project on coursera. This capstone project is part of IBM applied data science specialization. In this view the key objective of the project is to apply concepts and tools learnt during this specialization. Consequently, I will be making (and calling out explicitly) suitable assumptions in other parts of the project so as to focus on the application of data science tools and not distracted by other complexities.

Introduction

Background

As on demand services are becoming popular and have been accepted as a new way of life by millenials, competition is growing fierce to occupy larger share of pie.

Intelligent and sustainable supply strategies become founding stone for these business - especially in the scenarios where item being delivered are generic in nature and can be delivered from multiple points without significant difference in quality.

Some of the businesses that fall in this category are - grocery, medicine, liquor etc. These products are manufacturing brand specific and typically will have have wider offline presence. Other factors like shelf life, storage capability etc make these supply chains more open to “re-design” and dynamic.

In this project, I will be working to suggest a supply strategy for an imaginary on-demand grocery delivery company. This strategy tries to find balance between service quality and cost of service for delivering generic items in an on-demand fashion.

Interest

Having worked in hyperlocal logistics systems in the past, I have worked on some of these systems in the past. Given hyperlocal on-demand systems are still in quite early stage, there is significant scope for improvement/maturity in these services. Plus, there is little doubt that these on-demand economy, with dynamically deployable business strategies, are systems for future.

Problem Statement

For this project I will be working with the following problem statement:

“What supply strategy should an on-demand grocery delivery adopt in order to optimize between customer experience and cost of fulfilling orders.”

Scope & Assumptions

Even before we start to get into details of such open ended problems, it is important to define the scope of the investment to ensure depth in selected scope. For this projects:

- I will be limiting the scope of this problem to “Groceries” only. Each category is additive and can be obtained by repeating similar steps.
- Venues and places data available on foursquare. It is quite possible that foursquare data is not complete for certain areas/regions.
- I will be limiting our analysis to 2 geographies - Bangalore and San francisco. Needless to say, each geography will have its own nuances and need due tuning.
- Uniform distribution of demand in a city - This is done to contain the scope of project. If we have business data of prior demand distribution, it can be easily superimposed on the given methodology.
- I will be limiting business model to include only fixed cost, distance cost and cost of business loss. Further, even for these costs, I will be using very simplistic assumptions and models. This is done to ensure focus on learning objective of this exercise and not to get loss in business complexities.
- Cross grid effect on selection of stores has been ignored to reduce complexity. In the real world, having some number of stores in a grid will affect selection in surrounding grids as well.

In the spirit of the above scope, i have made following implicit assumptions as well

- Demand distribution is a proportional function of residential society complexes.
    - Weights are given for different types of residences to budget for number of people living there.
- Other business parameters (Availability, Fill rates, commissions etc.) is assumed to be constant throughout stores.
- Each store has infinite capacity to fulfill orders, so as order volume increases
- Business loss curve is assumed to be simplistic, in the real world, we would need to do a lot of A/B experiment and bayes probabilistic models (at the least) to get more refined assumption for this cost.
- Each store can service anywhere in the city.

- Business hours are assumed to be standard for all stores and customers (8 AM to 8 PM). I will be ignoring use cases around early morning or late night deliveries etc.

Problem Formulation

I will be Breaking down city areas into 2 digit geo hashes. Each grid will be roughly a square of 1 km*1 km. Then I will be using foursquare APIs to get venue locations and other details. I will be using the popularity indicators to proxy for the customer experience wrt service quality. Next, per order cost is calculated by choosing 'n' best possible stores in each grid and daily order volume at different level. I have formulated "depth of options" (another aspect of customer experience) in the objective cost function by putting a penalty for business loss due to less number of store options available within a certain radius.

Once this basic analysis is done, i will be Clustering (using k means method) these grids into 3 buckets: Low, medium and high demand density areas using data about different types of venues in each grid. Thereafter I will be finding optimal strategy for each cluster type.

Detailed methodology

Step 1: I will start with defining boundaries for a city.
Step 2: Break down city areas into 2 digit geo hashes.
    - Each grid will be roughly a square of 1 km*1 km.
    - For SF I will be having about 108 grids and for Bangalore I will have 560 grids.
Step 3: Getting available grocery stores in each of these grids.
Step 4: For each grid I will pick **'n'** (=(1,2,3,4)) most popular grocery stores.
Step 5: Calculate estimated value of store to customer distance (d^) of delivery:
    - Assuming demand is equally distributed in grid.
    - Further divide a grid into 10*10 sub-grids (3 significant digits)
    - **d** = sum (distance metrics (node of sub-grid, nearest store)
    - distance metrics = 1.4* aerial distance
        - Other possible distance metrics can be time based, network based etc. see more below: https://www.esri.com/videos/watch?videoid=oJDEKd84iPo&title=using-territory-design-distance-parameters
    - Using ride hailing companies rate card in these cities to approximate for the cost of per km travelled for an order. (lets say rate per km be **r**)
Step 6: Other costs considered in model:
    - Costs of having a partner store:

- Monthly Fixed cost (**F**) it includes sign up, pics, inventory mapping, technical integrations etc)
    - Assumed to be $10000 in SF and INR 20000 in Bangalore
- Daily Fixed cost (**V**) per store per day it includes daily wages and other daily recurring cost.
    - Assumed as hourly min wage * 2 people * x (assuming wages will be about 50% of daily variable cost)
- Cost of business loss per order (**BL**): If we have higher density on stores on out platform then it provides more options to customers to choose, consequently will result in higher conversion.
    - Of Course it would be of diminishing incrementality in nature.

Step 7: I will be working with 3 scenarios (daily order (**DO**) demand of 10k, 50k, 100k, 500k in each city) for each value of 'n' and try to work out the cost per order.

- Per delivery cost (PDC) is defined as:

PDC = ((F/30+V)/DO) + BL + d*r

I will be filling out following table that will help us suggest a specific strategy for each city at a given volume.

Scenario	n	PDC in SF (USD)	PDC in BLR (INR)
DO = 10,000	1		
	2		
	3		
	4		
DO = 50,000	1		
	2		
	3		
	4		
DO = 100,000	1		
	2		
	3		
	4		

DO = 500,000	1		
	2		
	3		
	4		

Step 8: Next I will analyze the contribution of distance cost or Business loss cost increasing with volume (intuitively it should as fixed cost would be discounted by higher order volume), then I will be double clicking on “High” demand clusters.

Step 9: Clustering (using k means method) these geohashes in 3 buckets: Low, medium and high demand density areas. I will be proximating of population living in each grid by number of houses and other utilities in those areas.

Step 10: Repeat steps 2 to steps 8 with one additional significant digit.

Step 11: Final conclusion and suggestion those can be feed into business and product strategy.

Data acquisition and description

Data sources

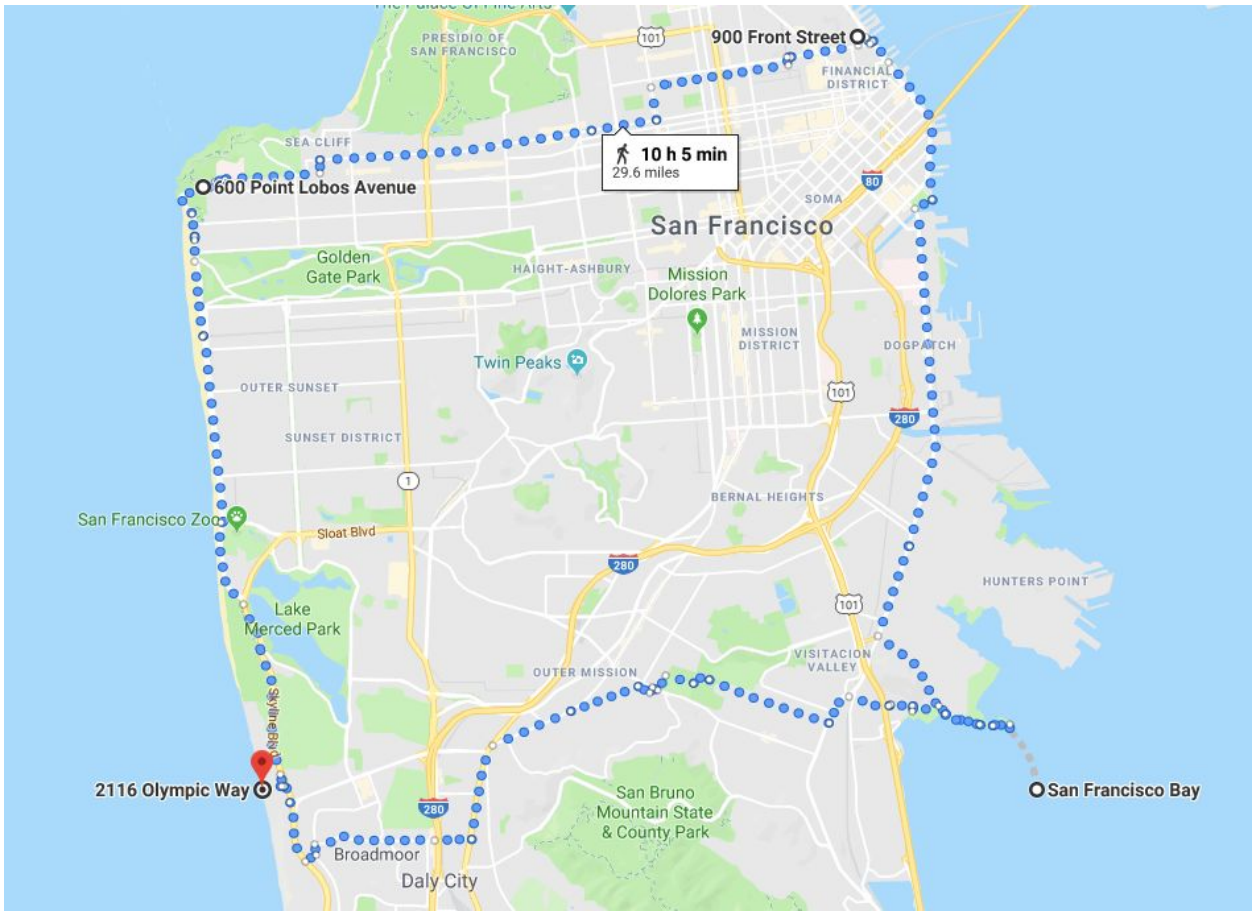
No.	Name	Objective	Source
1	City boundary data	Defining geographical boundary for scope	Approximation using Google maps.
2	Divide the city into grids	I will be selecting stores within each grid (as explained in the methodology)	Plotting using python library.
3	Store location	Getting available grocery stores in each of these grids.	Foursquare venue API to get location of grocery stores.
4	Demand density	Divide grids into Low, medium and high demand density areas.	Foursquare venue API to get locations of houses, all kinds of stores etc.
5	Popular stores	Determining popular stores in each grid.	Foursquare venue detail API.

6	Per km cost of travel	To build into business cost	Uber rate card for SF, Rapido rate card for Blr
7	Cost of business loss	To build into business cost	Industry reports

Data description

City boundaries and dividing into grids

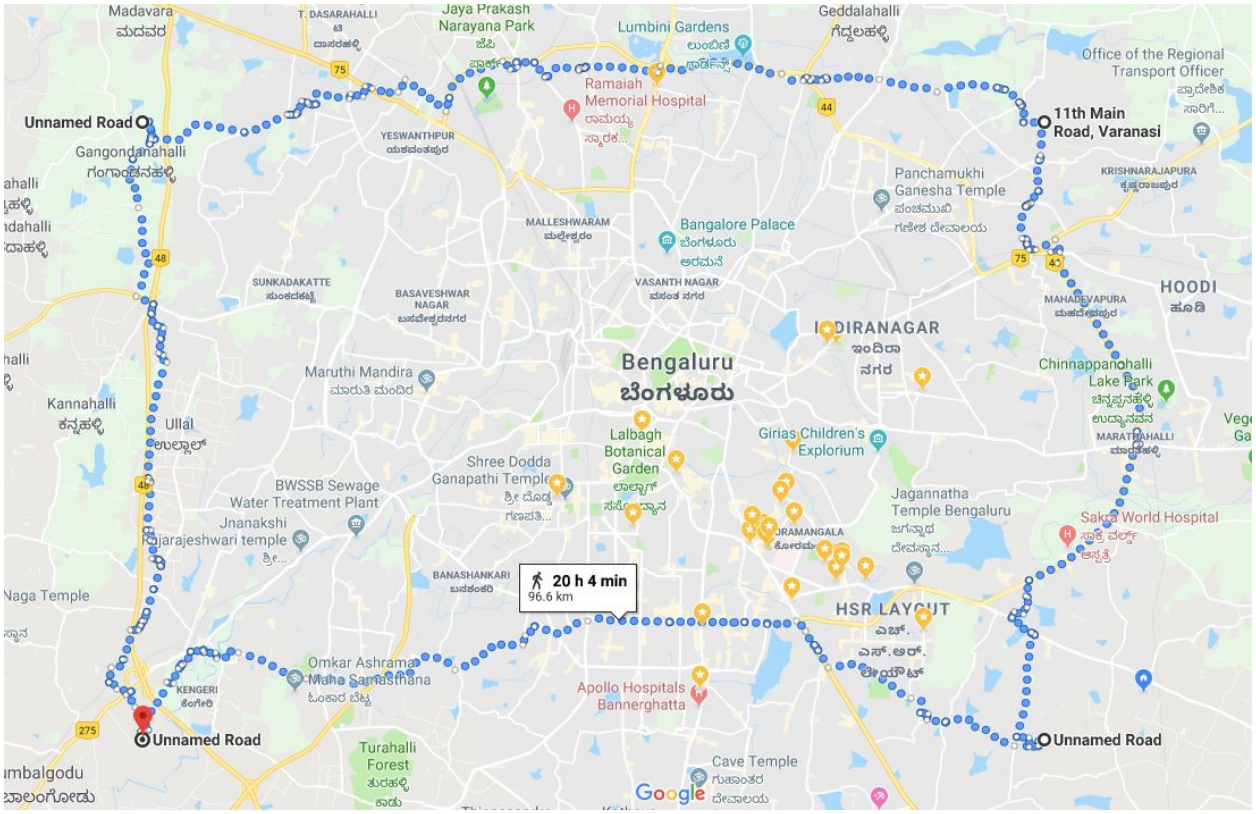
- City 1: San francisco
    - San francisco boundaries points:
      { (37.708277, -122.501808, 37.708746, -122.379993, 37.808400, -122.408343, 37.782380, -122.512448)} (Map link)



- boundary coordinates: (37.70, -122.50), (37.70, -122.37), (37.80, -122.40), (37.78, -122.51)

- Number of grid edges: 13, 10, 11, 8
    - 12*9 = **108 grids**
- Length of an edge: **0.8 to 1 kms**

- City 2: Bangalore
    - Bangalore boundary points: (12.899756, 77.471630), (12.895662, 77.683785), (13.036880, 77.683210), (13.036939, 77.477115) (Map link)



    - Grid coordinates: (12.89, 77.47), (12.89, 77.68), (13.03, 77.68), (13.03, 77.47)
    - Number of grids edges: 21, 14, 21, 14
        - Number of grids: 21*14 = **294 grids**
    - Length of an edge: **1 kms**

Foursquare APIs

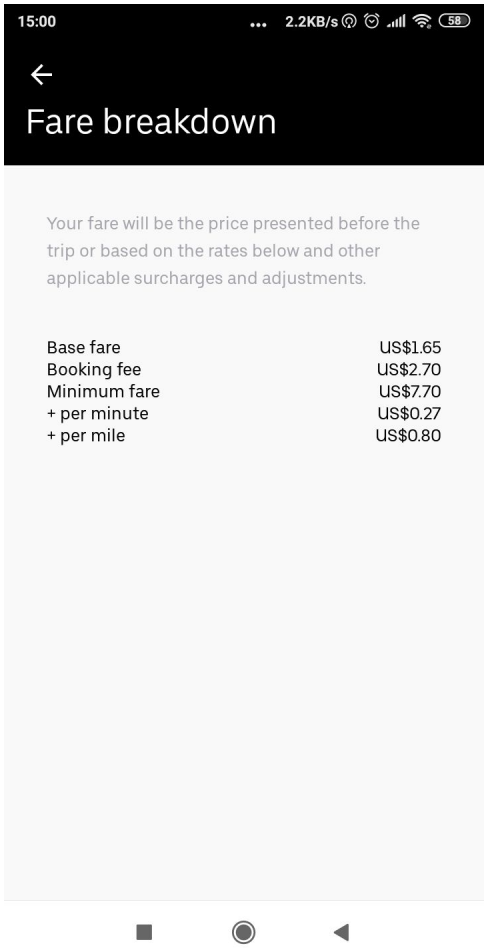
API name	Doc link	Fields of interest	Objective*
Venue details	https://developer.foursquare.com/docs/api/venues/details	Popular, rating, created_At, Like, dislikes.	5
Indicate venue incorrect	https://developer.foursquare.com/docs/api/venues/flag	Venue ID and problem	5

Get venue recommendation	https://developer.foursquare.com/docs/api/venues/explore	Venue ID, location, name and categories	3,4

*objectives mentioned in “Data source” table above.

Cost per km

Uber SF rate card: USD 0.84 per km (0.8 + 0.27 USD per mile)



Bangalore bike taxi rate card: INR 5 per km.

Min Wages

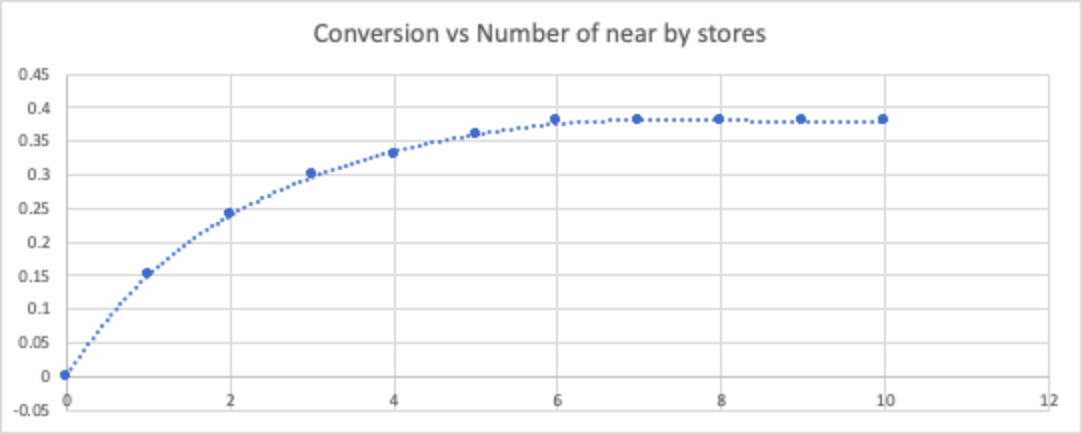
California: https://www.dir.ca.gov/dlse/faq_minimumwage.htm

Bangalore min wages:
https://www.livemint.com/news/india/rs-375-minimum-wage-plan-junked-as-govt-opts-for-rs-2-hike-1563035733771.html

Cost of lost business

Let x be (number of stores nearby): x is defined as the number of stores within distance metric (defined above) of 4 kms.

Conversions (defined loosely as users placing order viz a viz users opening app for the purpose of this project) will improve as x increases initially. Of Course it would be of diminishing incrementality nature.



X	Conversion	Profit per order in SF (in USD)	Profit per order in BLR (INR)	business loss SF (USD)	business loss BLR (INR)
0	0	0.5	10	0.19	3.80
1	15%	0.5	10	0.12	2.30
2	24%	0.5	10	0.07	1.40
3	30%	0.5	10	0.04	0.80
4	33%	0.5	10	0.03	0.50
5	36%	0.5	10	0.01	0.20
6	38%	0.5	10	0.00	0.00

7	38%	0.5	10	0.00	0.00
8	38%	0.5	10	0.00	0.00
9	38%	0.5	10	0.00	0.00
10	38%	0.5	10	0.00	0.00

Data cleaning

<TBD>

Exploratory Data analysis

<TBD>
<Relationships>

Unsupervised learning

<TBD>

Conclusions

<TBD>

Future Directions