The document details out the process of analyzing data to assess and compare two potential markets of self-driving ridesharing industry in San Francisco Bay Area

Self-driving ridesharing

(SF Bay Area)

Gaurav Dembla

Oct 21, 2020

Table of Contents

Introduction

Context

Perhaps the most exciting new frontier in driving technology is that someday cars may not require drivers at all. Self-driving cars, once considered a science-fiction daydream, are now taking the streets of San Francisco (SF), New York, Seattle and beyond. Built with the most powerful 3D imaging and AI out there, driverless vehicle developers aim to ensure that riders enjoy a hands-free experience without compromising on personal safety.

Massive Dynamic (MD), a hypothetical start-up in San Francisco, has taken up the daunting task of manufacturing world's most advanced all-electric self-driving vehicles that are not only safe for passengers but also friendly for the environment.

Goal

MD has already started building its fleet of robocars and is planning to deploy it in ridesharing service business in San Francisco Bay Area (San Francisco and surrounding areas). With the help of publicly available Travel Decision survey data from residents of Bay Area, it needs data-driven analysis to carry out following strategic decisions.

1) **Market Selection**: Which of the following markets should MD launch ridesharing in?
    a. Residents of SF, operating within the boundaries of the city, or
    b. Residents of other adjoining and nearby counties to commute into and out of SF.
2) **Marketing Segmentation**: Which specific groups of customers (age, gender, race, or any combination thereof) should MD target as its first set of customers and why?

Data Source

Two sets of data have been sourced to solve the problem at hand.

1) Travel decision survey 2017 data from SFMTA (San Francisco Municipal Transportation Agency). The data contains results of survey conducted in year 2017 upon roughly 800 people commuting in SF Bay Area, comprising of residents from both San Francisco and other nearby counties.
2) Distance of other counties to SF county from Google maps. It will be useful later to figure out potential value of each market in question (more about this later).

Methodology

Market Selection

In order to answer the first question and decide whether ridesharing service should target SF residents commuting within SF boundaries or other Bay Area residents (non-SF) commuting to and from SF, we need to calculate *potential value* of each market based on behavior of (~800) people for a month as captured in the survey.

The survey data has a perfect balance of SF and non-SF residents (401:403) that would ease the comparison and evaluation of the two given markets (refer to assumption *i*).

Potential value of a market is the sum of potential values of all commuters within the market.

$$Market\ Value\ = \sum Potential\ Value_{commuter}$$

$$Potential\ Value_{commuter}\ = \#\ Targetable\ Trips\ \times\ Average\ Distance\ of\ trips$$

Targetable trips for a commuter are the ones that have the potential of being converted to MD ridesharing service. We derive targetable trips for a commuter based on number of trips over the last month and a factor *TTF* (Targetable Trips Factor) that is based upon mode of transportation, residence county and reason of driving own car.

$$\#\ Targetable\ Trips = f(\#\ Trips, TTF)$$

$$TTF\ =\ f(transportation\ mode, residence\ county, driving\ reason)$$

While business rules for initialization of TTF values and underlying rationale and assumptions have been explained in Appendix A, the detailed calculation of targetable trips and potential value for few sample records has been put down in Appendix B.

Average distance of trips is sourced from Google maps. For non-SF residents, it is solely dependent on county of origin. For SF residents, average commute distance within SF has been set as 5 miles (refer to assumption *v*).

County	Avg Distance
Alameda	55
Contra Costa	40
Marin	40
Napa	65
San Mateo	30
Santa Clara	50
Solano	60
Sonoma	80
San Francisco	5

Number of trips of a commuter for each category of transport shall be calculated as follows.

a) Within-SF: Number of trips made by SF residents within city boundaries in last 2 days, multiplied by 11 to get an estimate for last month (refer to assumptions *vi* and *x*).

b) To-and-from SF: Number of trips by non-SF residents to SF and back to home county, excluding within-SF trips made by them (refer to assumption *vii*).

Marketing Segmentation

In order to answer the second question and identify specific groups of commuters that MD should target as its first set of customers, we deep dive into demographics of non-SF residents with regards to potential market value.

We compare potential values of different groups/segments under each demographic dimension – age, gender, income, and race, to figure out *high-contrast dimensions* (which show up some segments much more valuable than others) that could be good candidates for marketing segmentation and targeting.

Gender	Value
Male	39,576
Female	29,850
Non-Binary	574

Age	Value
25-34	16,824
35-44	14,960
45-54	14,249
55-64	12,578
65+	5,967
Unknown	3,128
18-24	2,291

Race	Value
White	33,615
Asian	12,426
Hispanic/ Latino	8,870
African American	8,116
Refused	6,058
Mixed (Unspecified)	71
Other	5

Income	Value
100k-200k	19,855
75k-100k	12,723
35k-75k	12,646
More than 200k	10,418
Unknown	9,856
Less than 15k	2,018
15k-25k	1,485
25k-35k	997

While gender and age do not seem to show up high contrast among different groups, race and income do highlight one group significantly more than the others, thereby qualifying for the final set of attributes that could be useful for marketing segmentation.

Conclusion

Market Selection

Upon calculating potential value of each market, we compare the two markets and notice that even though "SF residents" market witnesses roughly 4 times as many trips as "non-SF residents" market, the fact that non-SF commuters have to experience much longer distances than SF residents is simply driving the potential value of the market up.

Market	# Commuters	# Trips	# Targetable Trips	Potential Value
SF residents	401	25,674	6,578	32,890
non-SF residents	403	6,204	1,615	70,000

As a result, it would be wise for MD to **operate ridesharing service for non-SF residents** first, providing long-distance commuters convenience for trips from their residence county to SF and back. Apart from accessing the more lucrative market, MD would also be able to carve out a unique value proposition to make an efficient and formidable entry into already competitive ridesharing industry. After all, it is not hard to comprehend that long-distance commuters have long been devoid of both convenience and economy thus far, and MD could be on its way to completely shake the present state of affairs and disrupt the market.

Marketing Segmentation

Upon creating a heatmap of potential value by race and income of commuters (as shown below), we notice that a particular group stands out among the lot – **white people making more than $100k per year**. I recommend this group to be the first one to be targeted as potential customers of electric ridesharing service.

	Race							
Income	Asian	African American	Hispanic/Lati..	White	Native American	Other	Refused	Mixed (Unspecified)
Less than 15k	916	28	757	318				
15k-25k	966		396	124				
25k-35k	231	138	300	270				60
35k-75k	1,029	2,450	2,390	5,925	838	6		11
75k-100k	2,703	2,664	2,783	4,412			162	
100k-200k	3,682	2,772	1,481	11,548			372	
More than 200k	1,432	61	450	7,277			1,200	
Unknown	1,469	6	314	3,744			4,324	

Alternatively, depending on the availability of marketing resources, we could either go more granular by selecting just *white people earning $100k-200k per year*, or zoom out and select just *white people* in general, discarding any consideration of income whatsoever.

Assumptions

i. Actual number of SF residents moving within SF boundaries is roughly equal to actual number of non-SF residents making trips to and from SF, mimicking the (1:1) ratio of number of SF and number of non-SF residents available in the survey data. In other words, the distribution of people responding to the underlying survey is a fair representation of all commuters in Bay Area, irrespective of gender, mode of transport, county residence, etc. However, we will need to check upon this assumption and most likely relax it when the project makes it way to real-life product decision.

ii. Purpose of a trip (work, school, social, etc.) does not influence selection of MD ridesharing service over an alternate option, and hence, has not been considered for setting up TTF values.

iii. Commuters' alternative mode of transport (Q20) has not been considered in the identification of targetable trip factor, as MD offerings are still not in the market; it is possible that if MD offerings were in the market and were quite compelling, more commuters would lean towards the ridesharing service.

iv. TTF values have been set up based on best guesses and can be appropriately changed based on more facts when they pop up.

v. Average distance of commute within SF has been estimated to be 5 miles, which is just half of the farthest two points in SF county (calculated from Google maps). Though there is a potential to set it to a more appropriate value, it would not change my recommendation unless it was set to at least 10.7, which is more than the longest possible trip within the city boundaries, and hence, not feasible.

vi. We are willfully ignoring within-SF trips made by non-SF residents, since the question statement specifically mentioned to consider residents of SF for this market. However, we can relax this assumption when scope of markets is reconsidered, as the survey data has information about non-SF residents making within-SF trips.

vii. We are knowingly ignoring inter-county trips originating from SF by SF residents because the problem statement specifically mentioned to consider non-SF residents for this market. However, even if the scope of markets were reconsidered and expanded, we could not go very far as the survey data does not have any information about SF residents making trips out of SF to other counties. That said, there is a piece of information that could come handy if we were to absolutely consider inter-county trips originating from SF - ratio of number of people exported out of SF (for work) and number of people imported into SF (for work), which is roughly 1:3 and deciphered from 2016 Commute Flows between Bay Area Counties.

viii. Assuming MD launches offerings that are not costlier by a substantial amount than driving own car, we can reasonably persuade drivers to switch to ridesharing service to avoid hassles of driving (especially, during rush hours) and parking own car. The higher cost of ridesharing must be balanced by more convenience associated with the service as perceived by the drivers.

ix. It is safe to assume that MD can make ridesharing more economical than regular taxis and modern commuting options such as Uber/Lyft, easily luring commuters from the incumbent services to MD ridesharing service. However, the competitors in this segment will inevitably respond to MD offerings in some manner or the other to prevent complete loss of their market share. Hence, we assume that MD shall be able to capture only 50% of this segment easily, thereby setting TTF as 0.5.

x. What SF residents experienced in last 2 days is a good reflection of what they must be experiencing over weekdays. As a result, we need to multiply the number of rides in last 2 days with 11 to get an estimate of total trips over 22 weekdays of the last month.

Codebase

The required data wrangling and analysis to unearth insights has been conducted in Jupyter notebook after installing standard Anaconda distribution for Python 3. The entire code has been organized into one python notebook (*self_driving_ride_sharing_SF.ipynb*) and uploaded on Github repository.

Appendix

A – Initialization of Targetable Trips Factor (TTF)

Upon looking deeper into the reasons of commuters for driving their own cars, it led me thinking that not all trips would call for (and hence, be good targets for) ridesharing service. As a result, a new term has been coined to derive a subset of all trips that could potentially be taken up with MD ridesharing – *Targetable Trips Factor* (TTF).

As shown in the chart below, TTF values have been initialized considering following factors (refer to assumptions *ii*, *iii*, and *iv*).

a) Category of transport (customized higher-level hierarchy of mode of transport),
b) Within SF vs. to-and-fro SF trip, and
c) Reason for driving own car

			SF	non-SF	
Category of Transport	**Mode of Transport**	**Reason for "Drive own vehicle"**	**TTF**	**TTF**	**Rationale**
Car	1=Drove my vehicle alone 2=Drove my vehicle with others 3=Drove car share	Soft (defined below)	0.75		Though such commuters did not cite any of the hard reasons for driving own car, it is unreasonable to assume that all their trips would qualify for ridesharing service
		Hard (defined below)	0.25		Though such commuters cited atleast one of the hard reasons for driving by themselves, it is reasonable to assume that some of their trips could still qualify for ridesharing service
Taxi	4=Uber, Lyft, etc. 5=Regular taxi		0.5		Since Uber/ Lyft/ taxis will definitely respond to MD offerings, we assume that MD shall be able to capture only 50% of this segment easily (hence, TTF as 0.5) - refer to assumption *ix* .
Public	6=Public transportation (e.g. BART, VTA, Amtrak)		0.25	0.05	Public transport is usually cheaper and faster than driving own car or taking Uber/ taxi; more so, for longer distances (non-SF residents) than for shorter distances (SF residents). For some non-SF residents who live far from the nearest public station/ stand, it might be an inconvenient and only option to choose from. Since survey data does not indicate reasons for commuters to choose public transport over other options, it is safe to assume that few commuters would still be willing to use ridesharing service; more of SF residents than of non-SF residents.
Mass Private	7=Private bus or van		0		No incentive for commuter to ditch this facility because it is mostly free (provided complimentary by employers) or economical (as part of local incentives, such as SRP initiatives) and convenient for precise pickups/ dropoffs
Legs	8=Bicycle		0		We assume that these options must be getting used for short distances, and hence, commuter would most likely not have a strong motive to switch to ridesharing service
	9=Walk				
Others	10=Scooter/ Motorcycle		0		Ignore these options as there is no tangible information available about them
	11=Other (specify)				
	12=Don't know / Don't remember				

To elaborate further with an example, commuters who drive by themselves or take taxi can be more easily persuaded to switch to more convenient alternative of MD ridesharing service than others who take private bus or use bicycle (refer to assumption *viii*). Even for car drivers, ones that have soft reason for driving are more malleable (and hence, easier to be convinced to switch to MD offerings) than ones that have hard reasons for driving by themselves (and hence, difficult to persuade to use ridesharing, let alone MD offerings).

Reason for "Driving car"	Type
Driving and parking is faster than other modes of travel (transit, biking, and walking)	Soft
Parking was available close to my destination	Soft
I needed to carry something	Hard
Parking at my destination was free	Soft
I needed to make multiple stops before returning home	Hard
Driving and parking is safer than other modes of travel (transit, biking, and walking)	Soft
I was traveling with children	Hard
Parking at my destination was cheap	Soft

B – Calculation of Potential Value

Let's wrangle the raw survey data in a way to get to a more efficient format as shown below.

respnum	county	sf_resident	car_reason_hard	car_trips	taxi_trips	public_trips
2584	San Mateo	0	0	0	8	4
167	Contra Costa	0	0	4	0	36
6218	Marin	0	0	20	0	30
3509	San Francisco	1	0	0	22	22
1677	Solano	0	1	30	0	0
1143	San Francisco	1	1	55	0	11

Appendix A provides us with information to set TTF values for each commuter and each category of transport (car, taxi, public, private, legs and others) based on residence county (SF vs. non-SF) and car driving reason (hard vs. soft).

We input the TTF values under green attributes below – one for each category of transport. We ignore private, legs and others as their TTFs are 0.

Car TTF is dependent on commuter's driving reason

Taxi TTF is same for all

Public TTF is different for SF vs. non-SF residence county

respnum	county	car_reason_hard	car_trips	car_ttf	taxi_trips	taxi_ttf	sf_resident	public_trips	public_ttf
2584	San Mateo	0	0	0.75	8	0.5	0	4	0.05
167	Contra Costa	0	4	0.75	0	0.5	0	36	0.05
6218	Marin	0	20	0.75	0	0.5	0	30	0.05
3509	San Francisco	0	0	0.75	22	0.5	1	22	0.25
1677	Solano	1	30	0.25	0	0.5	0	0	0.05
1143	San Francisco	1	55	0.25	0	0.5	1	11	0.25

We then multiply number of trips under the given category of transport with corresponding TTF value to obtain targetable number of trips for the given category.

taxi_trips	taxi_ttf	targ_taxi_trips
8	0.5	4
0	0.5	0
0	0.5	0
22	0.5	11
0	0.5	0
0	0.5	0

Once we have number of targetable trips for all the three relevant categories of transport, we sum them up to obtain total targetable trips for each commuter.

respnum	county	car_trips	taxi_trips	public_trips	targ_car_trips	targ_taxi_trips	targ_public_trips	targ_trips
2584	San Mateo	0	8	4	0	4	0.2	4.2
167	Contra Costa	4	0	36	3	0	1.8	4.8
6218	Marin	20	0	30	15	0	1.5	16.5
3509	San Francisco	0	22	22	0	11	5.5	16.5
1677	Solano	30	0	0	7.5	0	0	7.5
1143	San Francisco	55	0	11	13.75	0	2.75	16.5

Now, we get average road distance between each county to SF and average commute distance within SF. Notice that the inter-county distance is as high as 60 miles for Solano to SF. Upon multiplying number of targetable trips with average distance, we obtain potential value for each commuter.

respnum	county	targ_trips	avg_travel_dist	value
2584	San Mateo	4.2	30	126
167	Contra Costa	4.8	40	192
6218	Marin	16.5	40	660
3509	San Francisco	16.5	5	82.5
1677	Solano	7.5	60	450
1143	San Francisco	16.5	5	82.5

Summing up potential values for all the commuters in a market yields market's potential value. For the given example below, market value of "SF residents" is 165 and that of "non-SF residents" is 1428.

respnum	sf_resident	targ_trips	avg_travel_dist	value
2584	0	4.2	30	126
167	0	4.8	40	192
6218	0	16.5	40	660
3509	1	16.5	5	82.5
1677	0	7.5	60	450
1143	1	16.5	5	82.5

Potential Value of "non-SF residents" market = **1428** (=126+192+660+450)

Potential Value of "SF residents" market = **165** (=82.5+82.5)