PLANNING CARMENUP® COMMERCIAL EXPANSION IN CALIFORNIA

Applied Data Science Capstone course

Final Report

Michael Touati

April 2019

Contents

1. Introduction

1.1. Background

CarmenUp is a large company providing senior and elder care products for home health care equipment and assistive devices for aging parents and caregivers. The company is a full-service home medical equipment firm that uses tailored products and services to empower patients to live their fullest lives. The company is well established in Nevada and Arizona states and now has identify a high potential of growth in the neighboring state of California.
Given that any commercial roll-out in a new state inevitably bring high costs to the company, CarmenUp must plan carefully its deployment and select only the best candidates among the potential locations.

Therefore the senior management of the company hires our services to plan the commercial expansion toward California. Our main goal is to identify best potential location to set up commercial booths in Colorado with highest profiles to lead to increase in sales.

1.2. Business Problem

In this scenario, it is urgent to adopt machine-learning tools in order to assist stakeholders to make wise and effective decisions. As a result, the business problem we are currently posing is how could we provide support to the executive team to choose the best mall locations to deploy CarmenUp commercial booths under viability and best valuation conditions?

To solve this business problem, we are going to isolate California's cities and neighborhoods with wealthiest population and high potential of interest for CarmenUp products and services. Then, we will define a list of the closest public shopping malls around these best potential areas.

Our final output will be a synthesis report with a recommendation to implement booths in the 20 highest-potential malls in California.

2. Data

2.1. Dataset on highest income areas in California

Data on households' income in USA was found on Golden Oak Research Group "U.S. Income Database Kaggle" website (https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations/version/8).
This dataset, originally developed for real estate and business investment research, includes income data of any given geographic location on a neighborhood scale in the US. The dataset contains large amount of data on income and geographic information.

The data includes state name, city, neighborhood id, geographic coordinates, and values for average households' income in 2017. Specifically we will extract data for California State.

2.2. Dataset on mall locations

To explore and target recommended public chopping malls to implement new booths, we will access data through FourSquare API interface and convert them to a dedicated dataset for visualization. By cross-referencing data on California neighborhoods from Golden Oak Research Group dataset and data on nearest shopping malls from FourSquare API interface, we will be able to recommend the most attractive mall locations in term of business to implement new CarmenUp's commercial booths.

3. Methodology

3.1. Main packages and toolkits

To perform this report, we used several Python packages and dependencies as following:

- Pandas - Library for Data Analysis
- NumPy - Library to handle data in a vectorized manner
- JSON - Library to handle JSON files
- Geopy - To retrieve Location Data
- Requests - Library to handle http requests
- Matplotlib - Python Plotting Module
- Sklearn - Python machine learning Library
- Folium - Map rendering Library

3.2. Main dataset collection

First, we read the dataset that we collected from the Kaggle website into a pandas' data frame. The full data represents more than 32000 neighborhoods in USA with geolocation and income data.

	id	State_Code	State_Name	State_ab	County	City	Place	Type	Primary	Zip_Code	Area_Code	ALand	AWater	
0	1011000	1	Alabama	AL	Mobile County	Chickasaw	Chickasaw city	City	place	36611	251	10894952	909156	30.
1	1011010	1	Alabama	AL	Barbour County	Louisville	Clio city	City	place	36048	334	26070325	23254	31.
2	1011020	1	Alabama	AL	Shelby County	Columbiana	Columbiana city	City	place	35051	205	44835274	261034	33.
3	1011030	1	Alabama	AL	Mobile County	Satsuma	Creola city	City	place	36572	251	36878729	2374530	30.
4	1011040	1	Alabama	AL	Mobile County	Dauphin Island	Dauphin Island	Town	place	36528	251	16204185	413605152	30.

Our dataset consists of state name, city, neighborhood id, geographic coordinates, values for average households' income in 2017 and many other information.

The next step is to process the data and extract all the relevant information for our analysis.

3.3. Data preparation and preprocessing

At this stage, we prepare our dataset for the modeling process, opting for the most suitable machine-learning algorithm for our scope. Accordingly, we perform the following steps:

a) Rename the column names.
b) Select data only for the specific state we investigate - California.
c) Clean the irrelevant data for average income (NA or 0 values).
d) Prepare a list of neighborhoods sorted by average income.

After all the steps for preparation and preprocessing, our dataset is clean and ready for further investigation.

	Id_neighborhood	State_Name	City	Latitude	Longitude	Avg_income
3341	60224419	California	San Diego	32.737719	-117.197744	242857
2910	60220539	California	Huntington Beach	33.679397	-118.020316	203910
2157	60213769	California	Los Angeles	34.043325	-118.404305	201716
3677	60227449	California	Menlo Park	37.436442	-122.192328	192856
3814	60228689	California	Palo Alto	37.451186	-122.147024	190876
2147	60213679	California	Los Angeles	34.091809	-118.508627	189728
2547	60217279	California	Manhattan Beach	33.880731	-118.410369	187441
3451	60225409	California	Solana Beach	32.989037	-117.240270	186307
3811	60228659	California	Los Altos	37.391307	-122.107655	184961
2919	60220619	California	Huntington Beach	33.722749	-118.061046	182565
4842	6029789	California	Orinda	37.901562	-122.182610	182528
2591	60217679	California	Agoura Hills	34.160627	-118.773024	182475
2773	60219299	California	Rancho Santa Margarita	33.639447	-117.629962	181956
3745	60228059	California	San Jose	37.309034	-121.739794	181057
4511	6026809	California	Atherton	37.446300	-122.214767	180590
2004	60212389	California	Encino	34.136576	-118.509450	179773
2325	60215279	California	San Marino	34.120246	-118.122234	178029
3678	60227459	California	Portola Valley	37.369709	-122.215844	177665
4838	6029749	California	Lafayette	37.932961	-122.120982	176992
3787	60228439	California	Saratoga	37.258968	-122.007893	176740
2810	60219639	California	Tustin	33.716623	-117.778283	176723
4840	6029769	California	Lafayette	37.868223	-122.104820	176685
2008	60212419	California	Sherman Oaks	34.137986	-118.437649	176545
2310	60215139	California	La Canada	34.206241	-118.195010	176514

3.4. Narrowing research on best candidates

After focusing on California State and cleaning irrelevant data, we remain with a massive quantity of observation.

```
df_data.shape
```

```
(3258, 6)
```

As we can see, 3258 neighborhoods is not yet a relevant scale for considering commercial deployment.

Moreover, the average income in the all dataset is not optimal - ~79 thousand USD:

```
df_data['Avg_income'].describe()
```

```
5]: count      3258.000000
    mean      78654.297115
    std       32967.022720
    min       15740.000000
    25%       53554.250000
    50%       72636.000000
    75%       98369.500000
    max      242857.000000
    Name: Avg_income, dtype: float64
```

For further investigation, we must narrow our research on the best neighborhood candidates based on average income parameters. The internal logic based on the company's own analysis is that more wealthy are the closest inhabitants to the shopping mall, more profitable is the commercial booth on site.

Therefore, we choose to reduce the research on the most profitable neighborhoods only based on their average income data.

We will focus on the top 50 neighborhoods in California.

Obviously, after keeping the best candidates, the average income in the dataset is much higher and in adequacy with the company's target.

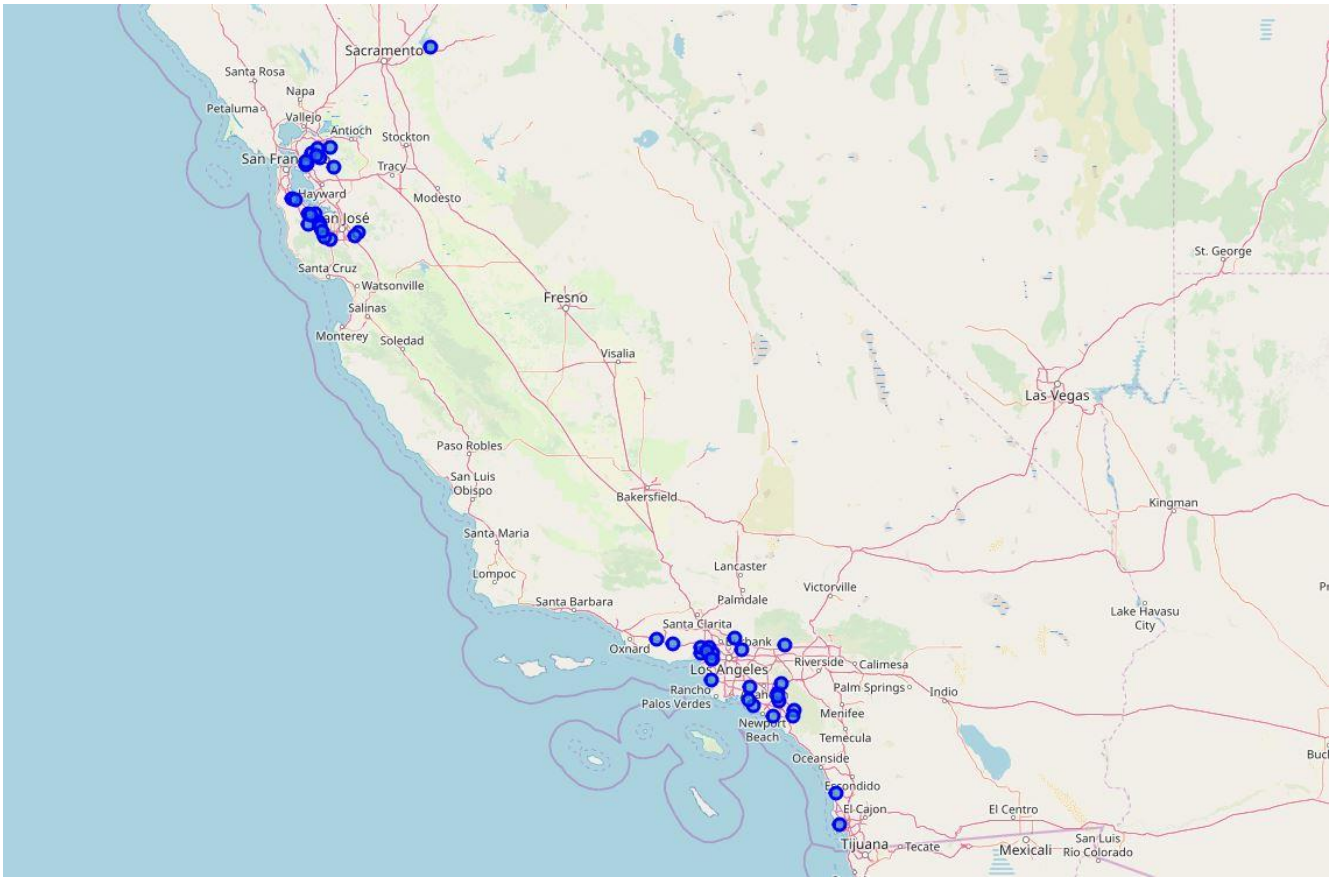
```
df_data['Avg_income'].describe()
```

```
4]: count        50.000000
    mean     177814.940000
    std       12961.177678
    min      165140.000000
    25%      169092.250000
    50%      175919.000000
    75%      181731.250000
    max      242857.000000
    Name: Avg_income, dtype: float64
```

The average income increased from 79 to 178 thousand USD.

3.5. Clustering the best locations

The resulting map of our dataset plotted on State of California map is as following:



We can easily observe that many of the 50 neighborhoods are in a very close area to other points. A further analysis based on that dataset would probably lead to identify same mall locations several times and as a result, many duplicate candidates. Thus, we decided to create geographic clusters for close neighborhoods.

At this point, we need to choose the best clustering algorithm.

The k-means algorithm is likely the most common clustering algorithm. It groups N observations (i.e., rows in an array of coordinates) into k clusters. However, k-means is not an ideal algorithm for latitude-longitude spatial data because it minimizes variance, not geodetic distance. There is substantial distortion at latitudes far from the equator, like those of this data set. The algorithm would still “work” but its results are poor and there is not much that can be done to improve them.
Moreover, more rows near a given location in the data set means a higher probability of having more rows selected randomly for that location.

Even worse, due to the random seed, many locations would be missing from any clusters, and increasing the number of clusters would still leave patchy gaps throughout the reduced data set.

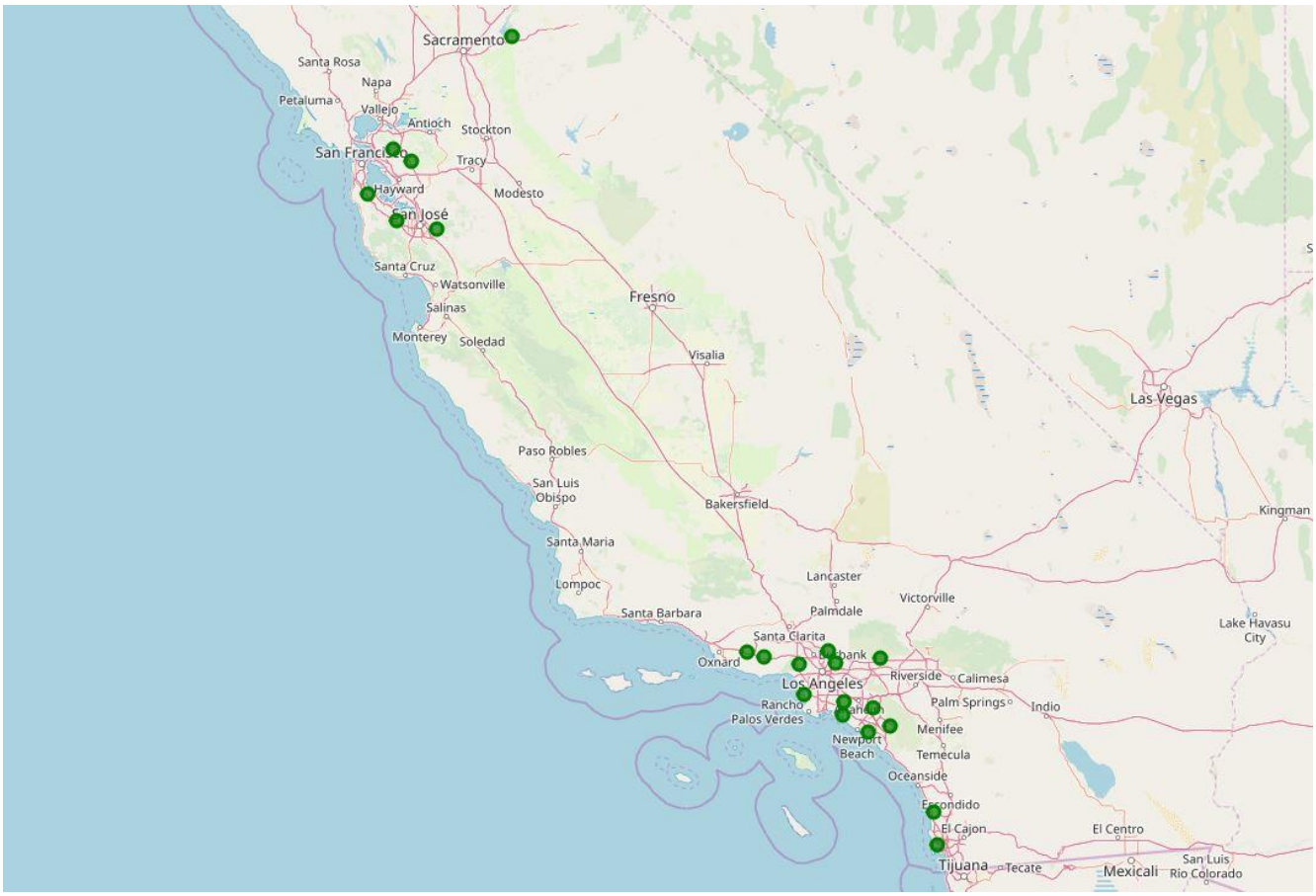
Instead, we decided to use an algorithm that works better with arbitrary distances: scikit-learn's implementation of the DBSCAN algorithm. DBSCAN clusters a spatial data set based on two parameters: a physical distance from each point, and a minimum cluster size. This method works much better for spatial latitude-longitude data.

The critical parameter to cluster geographic points with DBSCAN is the epsilon parameter. It defines the maximum distance that points can be from each other to be considered a cluster.

In order to limit our analysis to 20 location candidates, we choose to set the epsilon value to 10 kilometers. This value is also a reasonable distance for traveling to nearest shopping mall.
Thereby, we reduce the number of locations from 50 neighborhoods to 20 clusters centroids in California.

centroid_id	lat	lon
0	32.737719	-117.197744
1	33.722749	-118.061046
2	34.108023	-118.456964
3	37.371934	-122.101819
4	33.880731	-118.410369
5	32.989037	-117.240270
6	37.879801	-122.136483
7	34.160627	-118.773024
8	33.639447	-117.629962
9	37.309034	-121.739794
10	34.120246	-118.122234
11	33.778983	-117.785431
12	34.206241	-118.195010
13	38.682534	-121.061749
14	37.564961	-122.369169
15	33.595088	-117.829038
16	34.198459	-118.929552
17	34.152156	-117.717679
18	33.820512	-118.050022
19	37.799282	-121.965094

The resulting points are stored in a data frame and can be plotted on the map.



3.6. Explore data on shopping malls close to the centroids

Now we will create a request to identify the closest shopping malls to our most valuable locations centroids. We will use Foursquare API to get our data. Foursquare is used for getting the venues available in a particular neighborhood and their venue details.

We use Foursquare API request to extract all the shopping mall locations in the immediate vicinity of our 20 location candidates. According to various researches on the established methodology on vicinity and consumption's

habits,[1] we have chosen to determine a radius of 5 miles around each clustered location - 8 kilometers.

The final output of the 20 API requests is a dataset of shopping mall and their geographic coordinates.

3.7. Present final list of best shopping malls for commercial deployment

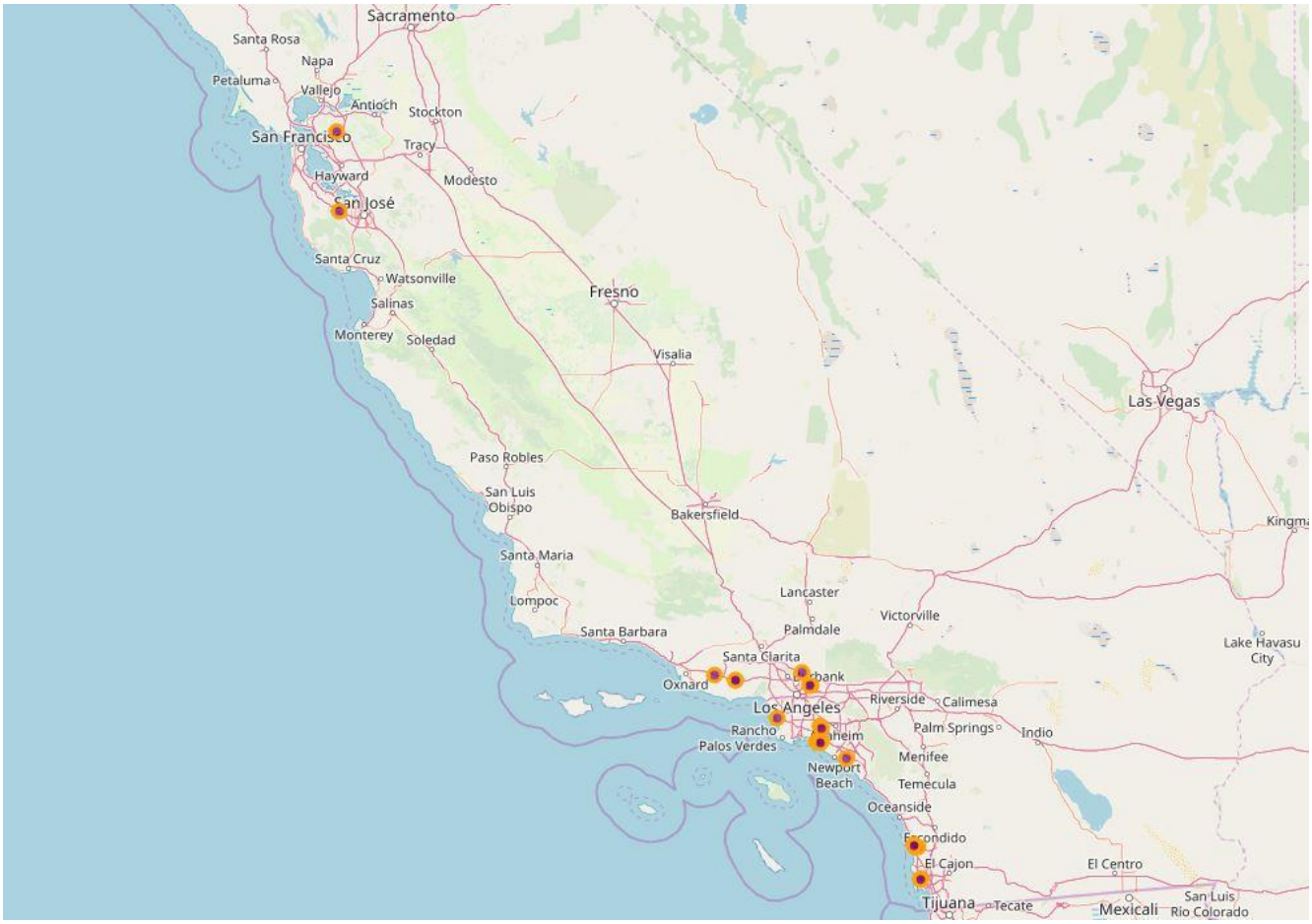
Our final list consist of 22 most valuable shopping malls in California.

	Mall name	Mall Latitude	Mall Longitude	Category
0	Fiesta de Reyes	32.755402	-117.197340	Shopping Mall
1	Loma Square	32.747835	-117.205684	Shopping Mall
2	Sunset Beach Farmer's Market	33.723658	-118.076168	Shopping Mall
3	Harbour Pointe	33.722870	-118.041826	Shopping Mall
4	Huntington Harbour Mall	33.717400	-118.051013	Shopping Mall
5	Rancho Shopping Center	37.360838	-122.097417	Shopping Mall
6	Metlox - Manhattan Beach	33.885976	-118.408416	Shopping Mall
7	Flower Hill Promenade	32.981108	-117.251224	Shopping Mall
8	Polo Plaza	32.983147	-117.229405	Shopping Mall
9	Lomas Santa Fe Plaza	32.995811	-117.257901	Shopping Mall
10	Solana Beach Town Centre	32.994596	-117.259020	Shopping Mall
11	La Fiesta Square	37.891135	-122.119931	Shopping Mall
12	Twin Oaks Shopping Center	34.155830	-118.756868	Shopping Mall
13	Agoura Meadows Shopping Center	34.153153	-118.758430	Shopping Mall
14	San Marino Mailbox	34.111989	-118.127592	Shopping Mall
15	Johnny Was	34.116496	-118.138741	Shopping Mall
16	La Canada Town Center	34.204855	-118.199547	Shopping Mall
17	Newport Coast Shopping Center	33.608041	-117.826874	Shopping Mall
19	The Village at Newbury Park	34.191336	-118.935064	Shopping Mall
20	Cypress Square	33.831866	-118.046176	Shopping Mall
21	Cypress Plaza	33.815810	-118.035155	Shopping Mall

[1] Miles, S., & Paddison, R. (1998). Urban consumption: an historiographical note. *Urban studies*, *35*(5-6), 815-823.

4. Results

We managed to identify our best potential locations for commercial deployment. In order to improve the final output of the analysis we choose to create a new map with the final locations of the malls with name, latitude and longitude of each mall.



We clearly identified 22 location malls. In a general perspective, one can observe that the company can define three main areas for commercial expansion in California:

A. San Francisco Bay in the North (2 shopping malls identified);
B. The coastal strip from Thousand Oaks to Newport Beach in the Los Angeles region (14 shopping malls identified);
C. Southern coastal strip from Solana Beach to San Diego (6 shopping malls identified).

5. Discussion

In this section, we will present final observations on our results and then present leads for further research.

5.1. Final observations

Based on the results, we recommend concentrating the commercial expansion effort on the Los Angeles area that includes 64% of the identified best locations (14 into 22 malls).

Such a recommendation should ensure economies of scale in planning and effective commercial costs, as well as optimal resource saving.

5.2. Further researches

Given more time and resources, one would like to explore other ways to create Folium maps, including different manners to map data, various labels, and incorporating information such as distance or shopping mall rating.

Furthermore, an interesting approach for additional and complementary reports could be to explore each mall in depth and access Foursquare reviews, photos, and tips to identify the best booth location inside the shopping mall. This capstone project did not include an in-depth machine learning section – perhaps there could be used for algorithms to predict crowd size, best times to operate the commercial CarmenUp booth, and other problems that can be solved using predictive models.

Using client-friendly and intuitive tools and open access data on the Internet we successes to offer a solution for channeling efforts and investments, minimize entry costs and maximize potential profit.

6. Conclusion

Starting from an extremely wide area of 423 970 km² (the all state of California), our machine-learning approach allows us to create a complete ready-to-use commercial deployment plan under uncertainty conditions.
The main task was to minimize the risks of business failure under an established plan of targeting new clients in a whole state of the USA.

It is important to emphasize that the present model is built in an open-source and upscaling approach such that it is completely open to offer solution for rollout in any state in the United States and even can be used in a much closer resolution. Moreover, this machine-learning guidebook is adaptable to other sources of data, future updates and changes in configuration.