New Bubble Tea Shop Location Prediction in Santa Clara County

Zijian Zhu

1 Introduction

1.1 Background

Bubble Tea is one the most famous and well known Taiwanese drinks in Asia and even the worldwide. Along with the economic booming for the past 10 years, San Francisco Bay Area, as the central of the Technology and Innovation of the world, attracted millions of immigrants moved in and became one of the highest density areas with Asian population in the US. Hundreds of Bubble Tea Shops opened in recent decade around the Bay Area. Even though the margin profit of a Bubble Tea shop is shrinking due to the raising competition, there are still lots of people trying to get into the business because of the low cost and threshold with relative high return features in opening a Bubble Tea shop. Like coffee shops and restaurants, location is one of the key factors for a Bubble Tea shop to being success. Therefore, taking an initial advantage with an ideal location to start a Bubble Tea shop seems very curial for the stakeholders, especially in current over crowded Bay Area.

1.2 Problem

This project targets to look for some ideal locations to open Bubble Tea shop in Santa Clara County (the heart of Silicon Valley) which might potentially have a higher than average profit and less competition.

1.3 Interest

The stakeholders who are looking to start Bubble Tea business or franchise expansion would be interested in searching for such kind of candidate locations.

2 Data

Based on definition of our problem, the following data will be used are:

First, the dataset contains geographical data of Santa Clara County (the County) such as geographical shape and location data, distribution of population densities. This dataset can be

found and downloaded online which the County divided by census blocks defined in 2010 Census and the dataset includes land size, population, coordinates, and geographical shape. (Figure 1.)

Second, the dataset should include detail information of all Bubble Tea shops in the County and venue attributes surround each Bubble Tea shop. The full list of Bubble Tea shops in the County can be collected by exploring each census blocks using Foursquare API. Detail information of Bubble Tea shops such as rating and likes (we assuming profitability is tightly correlated with popularity and rating of Bubble Tea shops) can be detail searched from Foursquare API. Exploring Bubble Tea shops to gather venue attributes within certain radius limit according to Foursquare API.

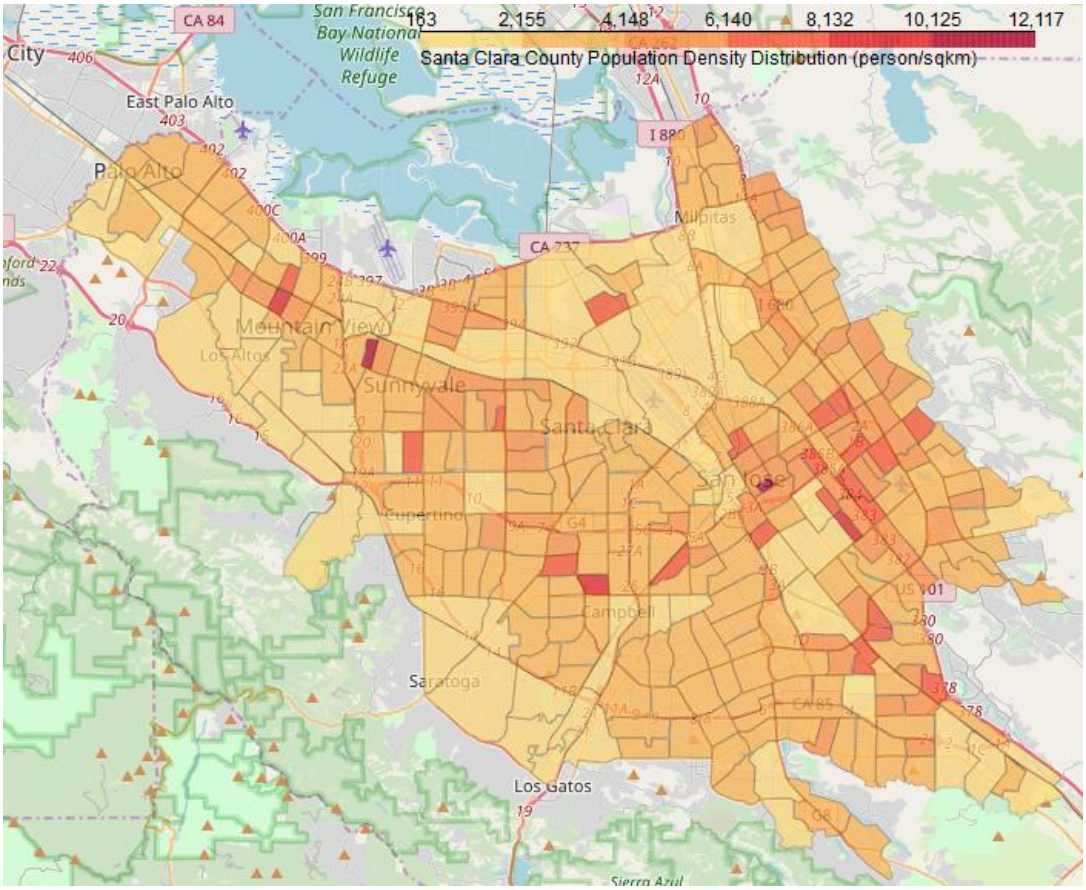

Figure 1. Santa Clara County Population Density Distribution

Third, the dataset also needs demographic data by postal codes in the County. The data can be found and downloaded from Santa Clara County Public Health Department website

Fourth, we define the County consisted by neighborhoods which are 1,000-meter radius circles. All the addresses of the neighborhoods' centers can be located by using geocoders. The venue attributes in neighborhoods will be explored by using Foursquare API. (Figure 2.)

With the data above, we would be able to discover the relation between Bubble Tea shop's rating and venue attributes surround the Bubble Tea shop in order to predict a rating of Bubble Tea shop in a certain location.

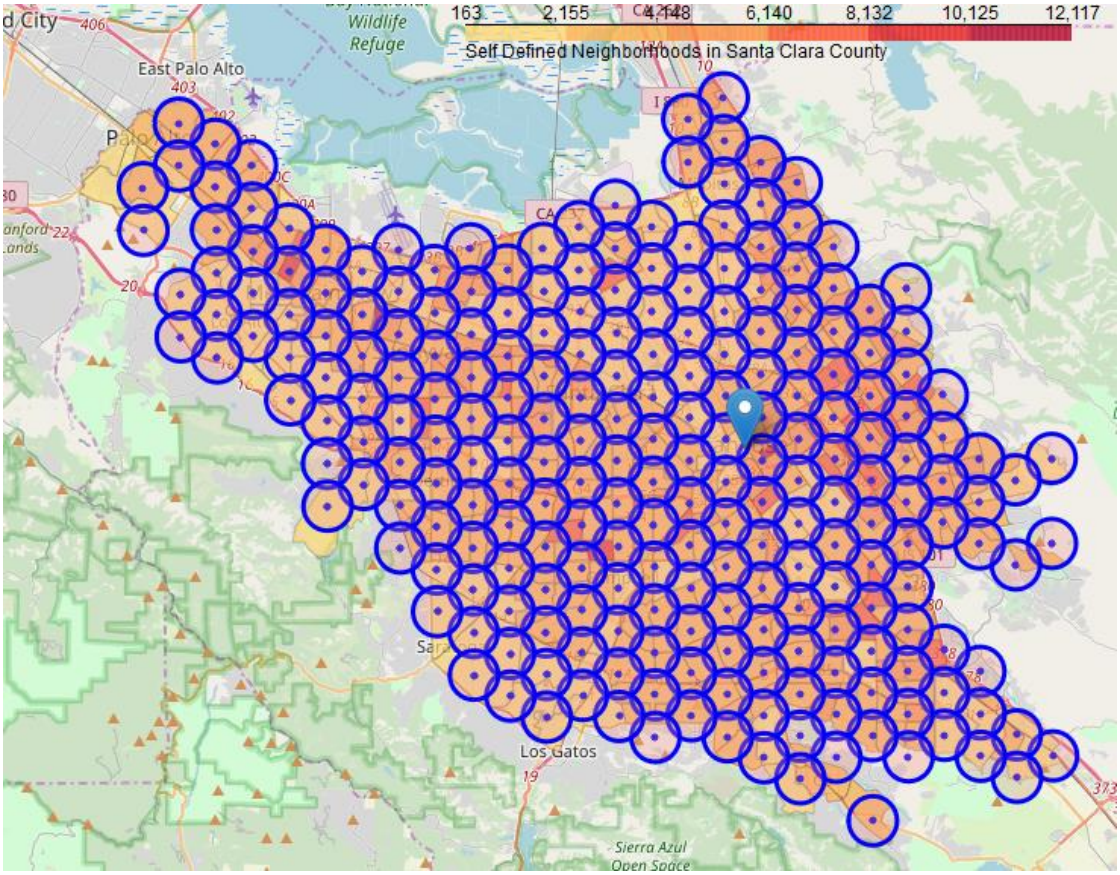

Figure 2. Neighborhoods defined in Santa Clara County for Study

3 Methodology

3.1 Exploratory Data Analysis

Before building up prediction model, we made an assumption as mentioned in introduction section that the profitability of a Bubble Tea Shop is highly correlation with its popularity. Therefore, we define popularity score, which consist with 25% weighted of Tip Counts, 25% weighted of Like Counts and 50% weighted of Rating, as a new dependent variable and normalize the variable with Minimax method.

3.1.1 Relationship between Bubble Tea Shop's score and Population density

We are curious if there is a continuous relationship between Bubble Tea Shop's popularity score and population or the density that the shop location at. Thus, we apply single linear regression with popularity score depends on population. The F-test results conclude that there is no significant statistical correlation neither between blocks densities and Bubble Tea Shop scores nor between population around Bubble Tea Shop and its score. (Figure 3.)

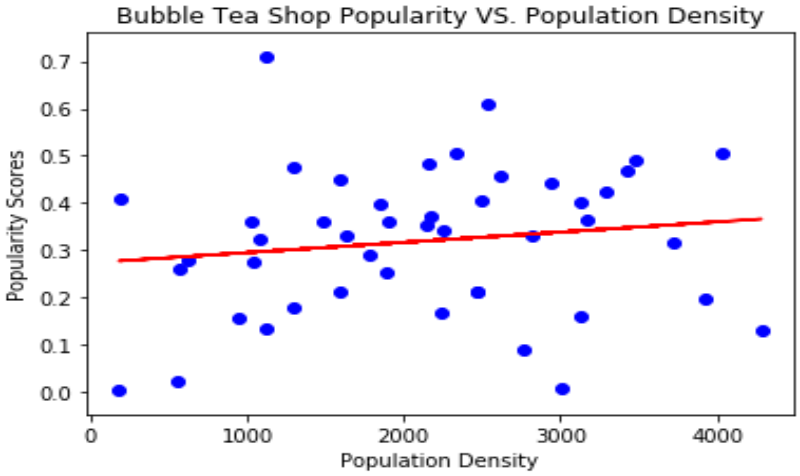

Figure 3. Bubble Tea Shop Popularity vs. Population Density

F-test results conclude that there is no significant statistical correlation neither between blocks densities and Bubble Tea Shop scores nor between population around Bubble Tea Shop and its score. The results of linear regression indicate that the model is not fitting the data well. (F-score: 0.2476, P-value: 0.6212)

According to the choropleth above, however, owners prefer to open Bubble Tea Shops close to or locate at blocks with relatively higher density. This will be a factor to be considered as ideal candidates. (Figure 4.)

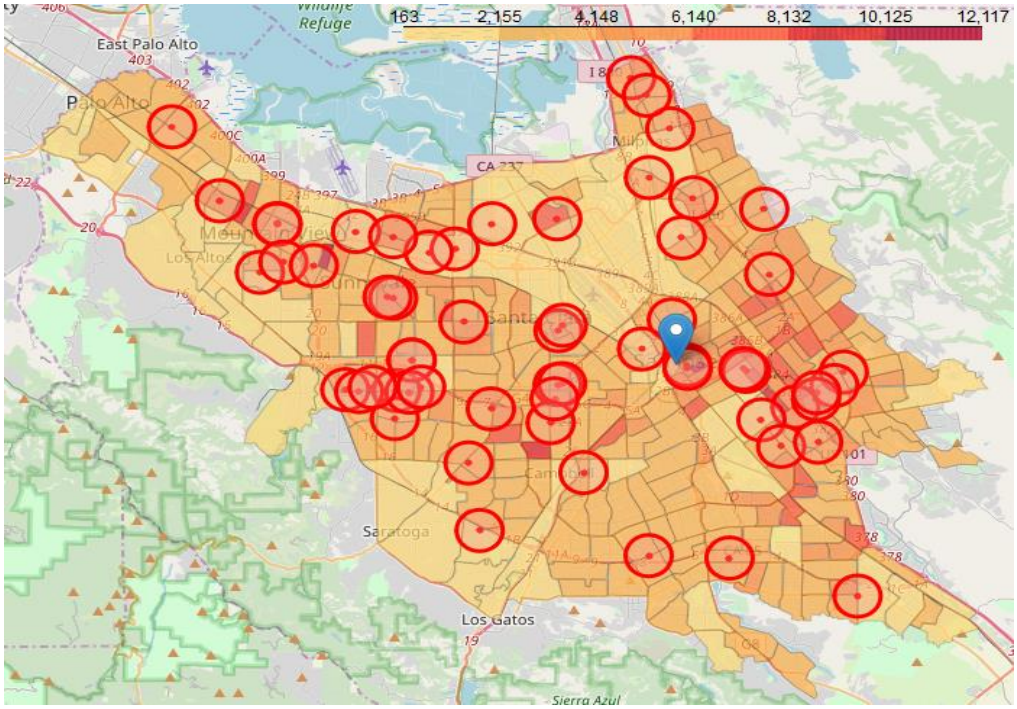

Figure 4. Current Bubble Tea Shops Distribution in Santa Clara County

3.1.2 Relationship Between Asian Percentage and Bubble Tea Shop Score

We hypothesized that the location's Asian demographical percentage might be correlated with the popularity of a bubble tea shop, because the bubble tea is original from Asia. The blocks were categorized into two group by average Asian percentage 42% of all postal code area over the County. The F-test results indicate the average Scores of Bubble Tea Shop are statistical significantly different between the locations Asian percentage above and below the all over average 42%. (F-score: 4.3107, P-value: 0.0437) We also surprisingly find that Scores and Asian percentages have inversed relationship. (Figure 5.)

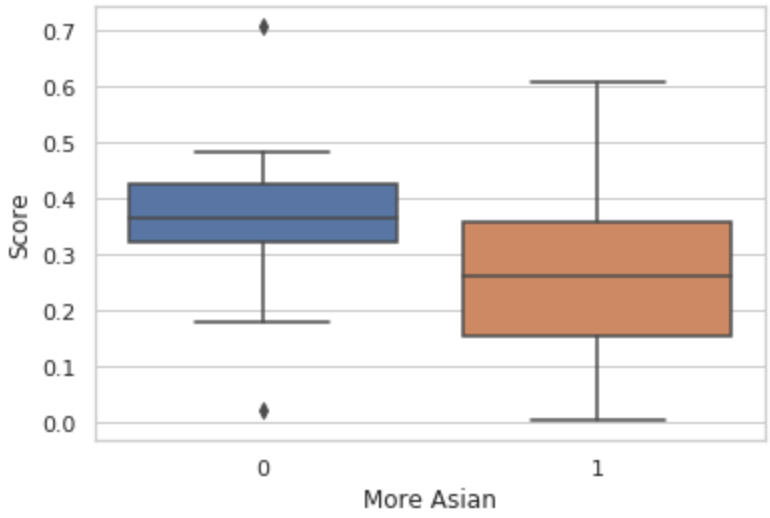

Figure 5. Average Score Comparison between two group of Asian percentages

3.2 Prediction Modeling

In general, there are two types of model in Machine Learning for prediction, Regression and Classification. Based on the problem we defined and the features of dataset on hand, Classification models would be sufficient and straightforward to solve the problem. We summarize the problem as whether a new Bubble Tea Shop would have an outstanding score in a given location and we make the assumption that the score better than 75 percent of existing Bubble Tea Shops' scores as outstanding. Therefore, the target model will be built for address a binary classification prediction.

Four classification models, as Logistic Regression, K-Nearest Neighbor, Support Vector Machines (SVM), and Decision Tree, has been built and tested. Among the all algorithms we have performed, Logistic Regression model turn out the best test results (both accuracy and F1 score over 80%). What's more, the logistic regression model would provide probability information for every target location. (Table 1.）

Test Score	Logistic Regression	K-Nearest Neighbor	SVM	Decision Trees
Jaccard Score	0.8125	0.6875	0.7500	0.6250
F1 Score	0.8062	0.5602	0.6859	0.5938

Table 1. Classification Model Test Scores Comparison

4 Result and Discussion

By preforming the prediction with our trained logistic regression model, there are 38 qualified candidates found out. Moreover, we put additional constrains such the location without any currently active bubble tea shop and Asian percentage lower than average (42%). The final 17 potential neighborhoods were list below. (Table 2.)

	Neighborhood ID	Address	Probilities
0	303	3354, Gawain Drive, Alum Rock, San Jose, Santa...	0.510759
1	628	Saint Joseph Avenue, Creston, Los Altos, Santa...	0.509505
2	613	620, Cree Drive, Almaden Valley, San Jose, San...	0.506897
3	612	Curie Court, Almaden Valley, San Jose, Santa C...	0.505697
4	632	Junipero Serra Freeway, Los Altos Hills, Santa...	0.503966
5	528	606, Los Olivos Drive, Santa Clara, Santa Clar...	0.503062
6	365	2487, Bambi Lane, Alum Rock, San Jose, Santa C...	0.502733
7	503	146, Kittoe Drive, Whisman Station, Mountain V...	0.502403
8	540	Schwab Residential Center, 680, Serra Street, ...	0.502084
9	553	Communications Hill, Willow Glen, San Jose, Sa...	0.502000
10	586	2600, Booksin Avenue, Dry Creek, San Jose, San...	0.501976
11	675	6046, Paso Los Cerritos, Almaden Meadows, Alma...	0.501779
12	333	National Hispanic University, 14271, Story Roa...	0.501037
13	555	Willow Glen High School, Dry Creek Road, Dry C...	0.500983
14	677	Noddin Elementary School, Gilda Way, San Jose,...	0.500915
15	538	4064, Solana Drive, Barron Park, Palo Alto, Sa...	0.500700
16	469	645, Bernal Avenue, Sunnyvale, Santa Clara Cou...	0.500285

Table 2. Final List of Predicted Locations

What's more, according to the distribution map (Red: existing bubble tea shop, Blue: qualified location, Purple: potential location, Green: final list), If we take avoiding to overlap the 1 kilometer range with current existing Bubble Tea Shops and as closer to a high density area as possible into considerations, there are Seven better potential candidates list as following addresses: (Figure 6.)

1. 606, Los Olivos Drive, Santa Clara, Santa Clara County, California, 95050, USA
2. Willow Glen High School, Dry Creek Road, Dry Creek, San Jose, Santa Clara County, California, 95125, USA

3. 2600, Booksin Avenue, Dry Creek, San Jose, Santa Clara County, California, 95125, USA
4. Communications Hill, Willow Glen, San Jose, Santa Clara County, California, USA
5. 3354, Gawain Drive, Alum Rock, San Jose, Santa Clara County, California, 95127, USA
6. National Hispanic University, 14271, Story Road, Alum Rock, San Jose, Santa Clara County, California, 95127, USA
7. Saint Joseph Avenue, Creston, Los Altos, Santa Clara County, California, 94024-6833, USA

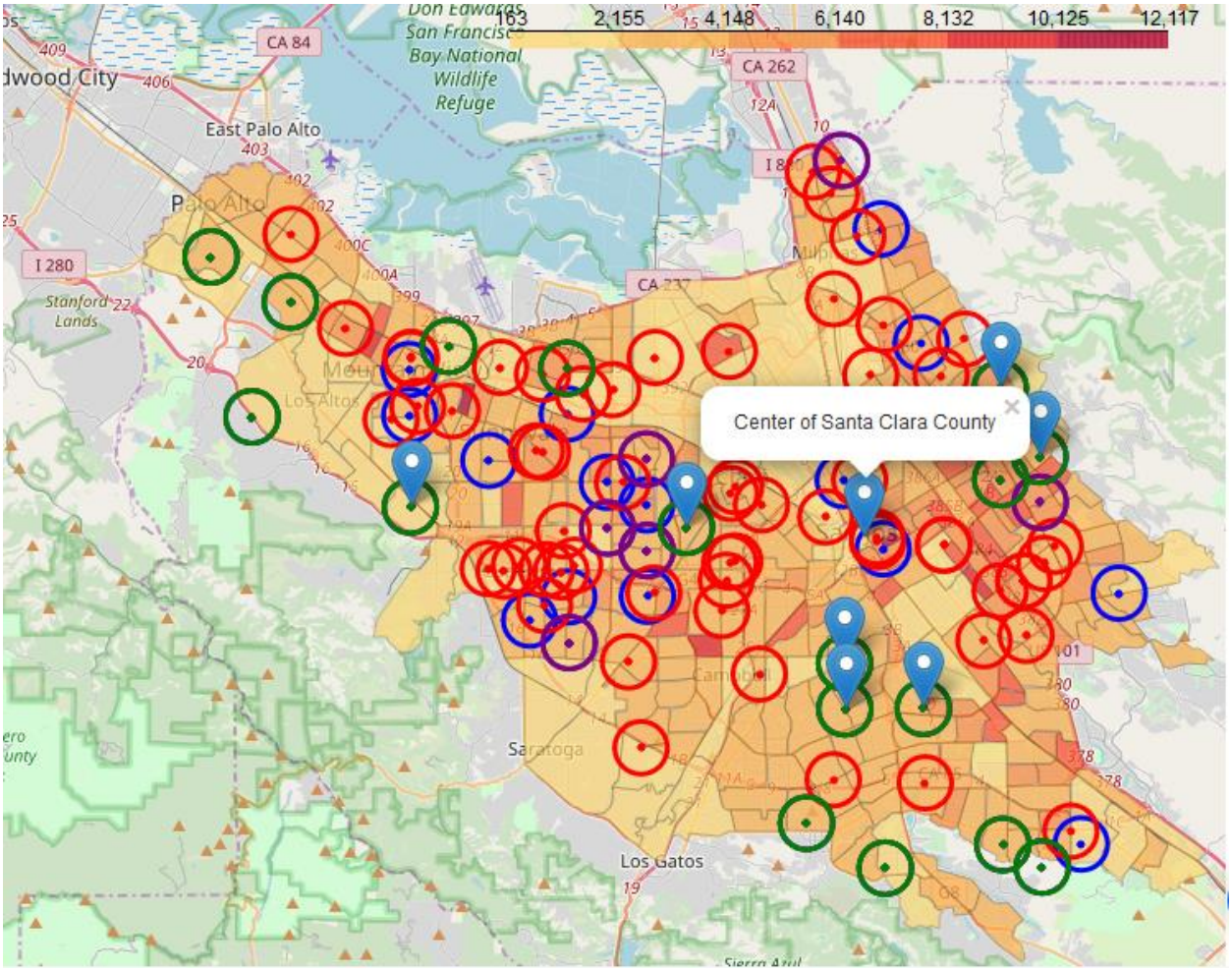

Figure 6. Predicted Location Results Distribution

5 Conclusion

Purpose of this project was to identify existing Bubble Tea Shops in Santa Clara County in order to aid stakeholders in narrowing down the search for optimal location for a new Bubble Tea Shop. We gathered and cleaned 2010 National Censes Data to identify density distribution of the area,

Foursquare API to explore all currently active Bubble Tea Shops as well as to identify Venue Attributes of both Bubble Tea Shops and defined neighborhoods in 1 km radius circle, and demographical date from the County Heathy Department. We also performed and compared four different Classification Machines Learning algorithms and decided to use Logistic Regression as the prediction model of this project. We selected the top 100 most frequent venues around Bubble Tea Shops as target features. The final prediction was preformed to come up 37 qualified candidate locations. The additional requirements, as Bubble Tea Shop currently existing or not and lower Asian population or not, were taken into account to narrow down the candidate list to 16 locations and the addresses were listed. Five out of sixteen locations were suggested due to further conditions considered such as avoiding to overlap the 1 kilometer range with current existing Bubble Tea Shops and as closer to a high density area as possible.

Final decision on optimal Bubble Tea Shop location will be made by stakeholders based on specific characteristics of neighborhoods and locations, preference of stakeholders, taking into consideration additional factors like amiabilities of plazas or shopping malls, accessibilities (major roads or freeways), rental prices, social and economic dynamics of the neighborhood etc.