**Opening a Gym in America’s 4th Fittest City**

Alex Dowd

9/5/2020

1. **Introduction**
   1. **Background**

Washington, DC is the fourth fittest city in America according to USNews, the suburb of Arlington, VA was ranked #1 by Runner’s World in 2020 and the DC Policy Center estimates that approximately 80% of adults in the city have a gym membership. This makes the area hyper competitive for opening a successful fitness franchise. I was approached by John Smith, CEO of GloboGym Inc, to help advise where in the city he should open his next franchise.

* 1. **Problem**

Washington, DC and its surrounding suburbs in Maryland and Virginia are very healthy, as measured by gym, health club and running club memberships. It is a highly competitive environment with very high rent prices (6th in the US according to WTOP.com). This project aims to reveal underdeveloped by growing sections of the area (DC, Maryland and Virginia - DMV) which are good candidates for a new fitness center. John Smith’s requirements are to identify an area that is 300 meters from the next closest gym and 100 meters or less from at least 2 dining options.

* 1. **Interest**

This data would be very relevant to any fitness individual or company who is looking to open a new business in the Washington, DC area.

1. **Data Acquisition and Cleansing**
   1. **Data Sources**

This analysis will incorporate data from multiple public data sources. Geographic information use to tell where gyms currently are will come from the Foursquare API. Demographic data will come from Open Data DC and the DC Ward dataset. Additional demographic and individual health data will come from the 500 Cities Project. Additional contextual data will come from various sources and will be cited accordingly.

* 1. **Data Cleansing**

I prepared the Jupyter notebook entirely within IBM Watson studio. The data files were read from Watson as well. After renaming some of the columns to make them shorter, the first step was to sort the data by total population to see which Ward contained the most people overall.

A screenshot of a cell phone

Description automatically generated

After this step, I sorted the data by population over 65 in Ascending order and found that Ward 1 had the lowest population (7.5%) over 65. This is significant because older people are less likely to purchase gym memberships than younger people.

A screenshot of a cell phone

Description automatically generated

Following this step, I performed an income analysis. DC is an expensive city and gym memberships regularly approach $100/month. All of this data came from the DC government’s website with income and ages broken down by Ward. The income analysis found that 21% of Ward 1 earns more than $100K/year and more than 1/3 earn more than $75K/year.

A picture containing drawing

Description automatically generated

Following the income analysis, I plotted where current gyms were in Ward 1. I called on the Foursquare venues API and based the query on a 1,000-meter radius around the Target located on 14th Street NW. Target is close to centrally located in the Ward, which is located in the central part of the city. The results returned 30 venues categorized as gyms by Foursquare within 1,000 meters of the Target.

A close up of a map

Description automatically generated

The Foursquare map is below

A close up of a map

Description automatically generated

After plotting where gyms currently are, since the Globo Gym cannot be within 500 meters of another gym, I ran a K-Nearest-Neighbors algorithm to evaluate and later predict categories of gyms. This is significant because commercial gyms are distinctly different from apartment gyms and hotel gyms in that the latter don’t have membership options and do not compete for business with the former. The KNN analysis required considerable cleansing.

Other than the latitude and longitude columns, all other columns had to be converted to floats/integers in order to use in SKLearn. The names all contained spaces, so I could not use the LabelEncoder() function native to SKLearn. Instead, I had to run 2 cleansing operations. First, I had to replace all spaces with “\_” using the **str.replace()** function. Next, because you cannot cast strings to floats, I had to convert the object data types to ‘category’ using the **astype()** function. Finally, I assigned the ‘category’ columns to a variable and ran **cat.codes** using **.apply(lambda x: x.cat.codes).**

1. **Modeling**

With all features assigned a numeric value, I ran the KNN machine-learning algorithm using the train\_test\_split function from sklearn.preprocessing. After assigning the training and testing values, I used KNeighborsClassifer from sklearn.neighbors to build the model. The KNN model showed that the optimum K is 3, with a test accuracy of 83%. A limitation of the model here is that there just aren’t that many gyms. Even expanding the radius to 5,000 meters, which covers most of the city, my results only pulled 50. I ran the model with the population in Ward 1, which was 30 gyms. To help compensate for the small sample size, I included Name, Category, Cross Street, Address, City, Postal Code, Longitude and Latitude in my feature set. This gives the model more data to work with, since categories of gyms (college gyms, for example) are impacted by location. See below for a visualization of the K-predictor function.

**KNN Model**

A screenshot of a social media post

Description automatically generated

1. **Conclusion**

I plotted the locations of gyms in Ward 1 and have recommended that John Smith opens his Globo Gym franchise in the vacinity of Lamont St and Mt. Pleasant St NW. This will satisfy the requirements of being at least 300 meters from another gym and 100 meters from 2 restaurants or eating places. The KNN model will categorize current and future gyms to help John Smith decide whether another gym is actually competition to him or not. Due to the limited number of gym categories and the ease of plotting their locations, the model should be very accurate at determining these values. A map of the reccommended area is below.

Next page

**Wide View Map**

The purple circle is the build area, the red circle is the 300-meter zone around that, and the blue dots are existing gyms.

A close up of a map

Description automatically generated

Next page

**Close View Map**

In this view, you can identify 3 business labeled as restaurants with the fork and knife icon, as well as at least 1 grocery store denoted by the basket icon.

A close up of a map

Description automatically generated