**Costco Stores Growth Potential Prediction**

Rong Zheng

Mar 11, 2019

**1.Introduction**

**1.1 Background**

Costco is rapidly growing U.S. Warehouse chain store that have been getting more and more popularity as well as business success. What makes Costco so successful? More members? unique lower price products or strong supply chain providers? This is wide topic for data science analysis, related to products analysis, customers analysis and supply chain analysis and so on. In this project, I focused on Costco new stores location analysis. For the next three year, there would be average 20 stores opening in new locations. As any potential risks associated with building new stores, the more stores open, the higher the risk that company will build a store in an undesirable location. So, picking a store location is very important with the goal of driving more traffic and boosting sales.

**1.2 Problem**

This project is to predict Costco stores growth potentials by analyzing the existing Costco stores to generate similar future store locations within any given U.S. State so that Costco can grow there to reach its potentials.

In the interests of this project, machine learning using in place to locate current Costco warehouse store city information to find similar city within a given state. The city information includes population, family size, median household income and other factors. K-Nearest Neighbors (KNN) algorithm is used to make recommendations of the future store locations for a given State.

**1.3 Interest**

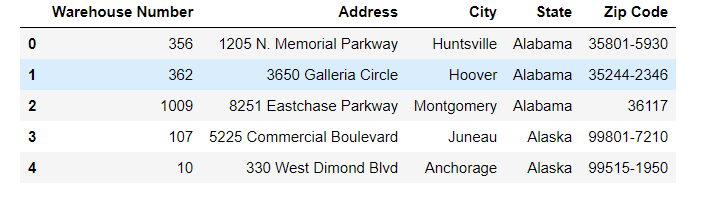
Actually, the executive team of Costco and some commercial investors would be very interested in accurate prediction of new store location, giving more confidence in their decisions. For competitive advantage and business value, the executive team need to know where their potential customers are. For others who are interested in investing company such as commercial investors or small business owners may also be interested it, they need to know where they will be over the 10-to-25 years lifetime of the investment they make in physical space. Definitely, no matter which location is selected, it must drive the development of the local economy.

**2. Data acquisition and cleaning**

**2.1 Data sources**

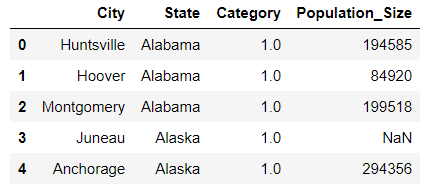
Originally, the datasets of this project came from four areas. One is “current existing Costco warehouse location”, named Costco-locations.csv（<https://github.com/swinton/Visualize-This/blob/master/ch08/geocode/costcos-limited.csv>）， that is including addresses of each existing location, zip code, state, city, warehouse number(Table 1.Costco\_location)

Table 1. Costco\_location



The second data source I searched “Top 1000 Nationwide\_2018 Population Data” (<https://www.biggestuscities.com/>). As the location of fast-moving consumer goods stores, it must be in a city with a large population and a good economic growth, this is also a reason that I wanted to use this data source as one of dataset for this project. So, I used BeautifulSoup, grabbed data from website and joined this table with Table 1. to look for which existing locations have been in this list, I found that from the 456 cities that Costco has store, 183 of them does not have population data in this dataset! The value of Population\_size column is null (Table2.Biggest Cities). This is interesting finding, based on my analysis solution, I decided not using the dataset.

Table 2. Biggest Cities

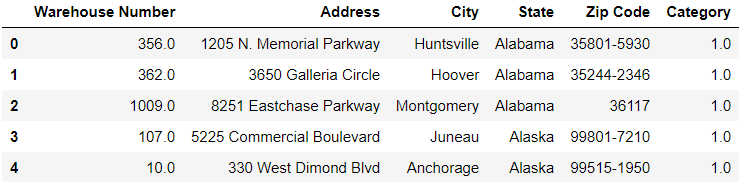


The third dataset is from Census Data API with income and household info ( <https://www.nhgis.org/>), that is including median family income, number of vehicles, number of family members. The forth data source I used is geocode to map locations. Additional, I used this source for labeling. (from <https://www.winsightgrocerybusiness.com/retailers/heres-where-costco-opening-2019> and <https://www.savingadvice.com/articles/2019/01/08/1052240_new-costco-locations-2018.html>)

**2.2 Data cleaning**

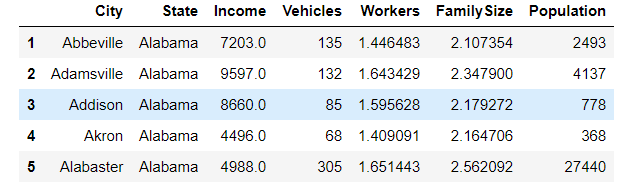
An assumption has been made here to leverage the existing stores as “select”, and other cities in U.S.as “not-selected” for clustering categorization differentiation. Further on, to help break down the number of “ selected” stores into more classes, another assumption is made to use the timing when these stores were opened to be the differences of the classes: 0-city has no Costco stores;1-city that has Costco stores opened 5 years ago:2-city that has Costco stores opened 2 years ago; 3-city that has Costco stores opened less than 2 years. So, I got table location\_labels

Table 3. Location\_labels



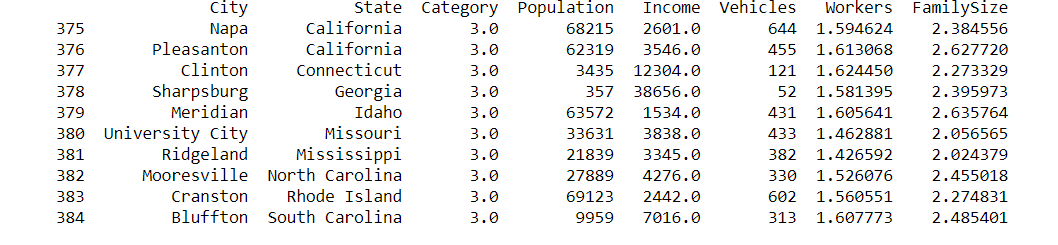
The following steps are to extract data, clean and calculated workers, family size, population based on raw data, and run aggregation to prepare for better format and got this table 4 census\_city, the first five rows here

Table 4. Census\_city



Then, I merged this table with Table 1, categorized all cities with 0,1,2 and 3 for preparing KNN model (Result 1. City\_dataset).

Result 1. City\_dataset



**2.3 Future selection**

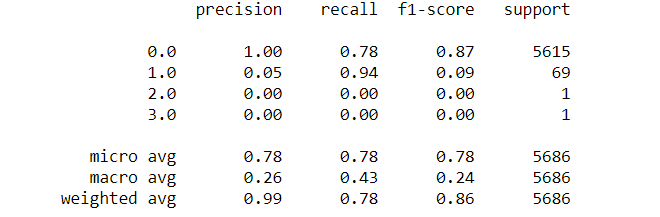
For the limitation of finding more accurate data, the features of this project focused on customers area. There are totally five features using in my project, that is population, median family income, vehicles, workers of each family and family size.

**3. Predictive Modeling**

K-Nearest Neighbor (KNN) Classification was used to generate recommended cities for Costco to open their future stores. For this project, due to the short amount of time, one state (Oregon) is selected to be exclude outside of the training and testing. All stores from other states are used for KNN training and testing to create the classifier. Once the classifier is created, cities from Oregon are used for prediction.

So, first I selected all rows with all features except Oregon to train learning, and I got this result 2. Classification\_report

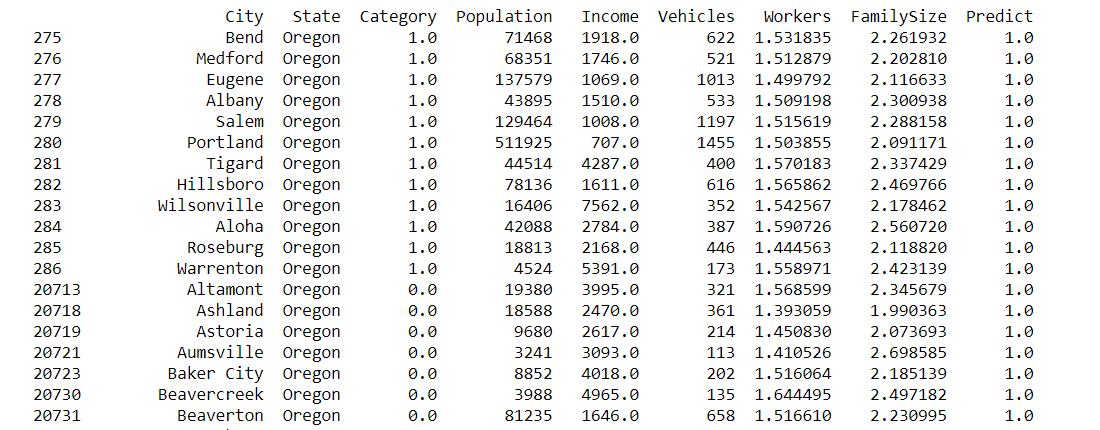
Result 2. Classification\_report



The above result is reasonable on the category 0 but bad for category 1. For 2 and 3, there are not enough data in the training set. For future considerations, the category might need to combine into just 0 and 1 or find more data to user different categorization strategy.

The next step is to predict on a selected state (using “Oregon” in this case), how many Costco store locations would the classifier recommend (Result 3. Result\_data).

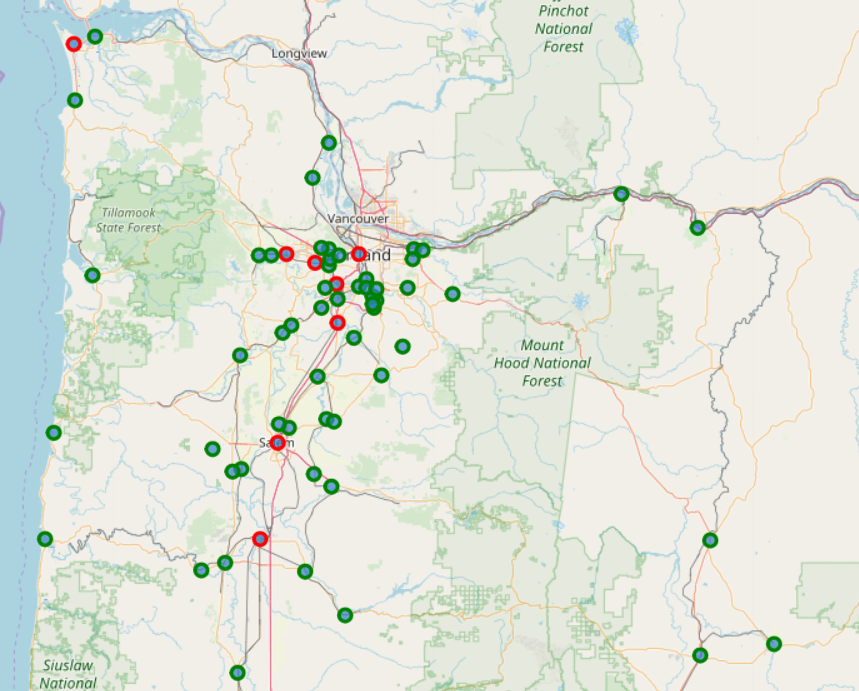
Result 3. Result\_data



The results indicate that there are many cities the classifier recommends to open a Costco store (result value = 1) where many other cities are not recommended (result value = 0). The limitation here seems that there is no result value with 2 or 3. Probably due to the same problem where there are no enough training data for these two labels.

Finally, I used matplotlib and folium mapping the locations (map 1. Map\_result)

Map 1. map\_result



In the above graph, the RED color cycle indicating existing Costco locations while the GREEN color cycle indicates new future locations for Costco Warehouse.

The good news is that all existing Costco warehouse locations in Oregon state are correctly predicted by the classifier. That’s 100% accuracy! It gives confidences that other locations predicted by the classifier have meaningful considerations.

Due to time limits, I have only completed the Prediction for one state (Oregon State). In order to complete the predictions for all states in U.S., repeat the above steps to first exclude that specific state during training, then run the predict function of the classifier to predict that state.

Once all states are predictors, the total cities are the full potentials of what Costco should or could open their warehouse in. That indicates the Costco Stores Growth Potentials.

**4. Conclusions**

In this study, I identified population, growth, family size, workers, vehicles, income, city, state features that affect location choice. I built KNN model, predicted where the next locations would be in Oregon. By testing, the accuracy of this model training is 100% accuracy.

**5. Future Considerations**

Due to the limitation of the project schedule as well as the availability of the data sources, this project only focuses on limited set of features for city predictions. There are multiple areas for future enhancement:

1. Additional City Information

More city information would be very helpful to add as features to assist the classification analysis. For example, the property or housing price in the area, crime rate for safety, number of children on average family, etc.

1. Costco Warehouse Financial Information

When more Costco warehouse financial information are available, it would help to make multiple classification more accurate, for example, the average sales or revenue per location, the type of Costco warehouse location as grocery, gas station, vision or other.

1. Different Categorization Strategy

For Categories ==2 or 3, there are no enough data for the training set. In the future, we need to search for more data. Another categorization strategy could be to leverage the Costco store sales performance.

( I searched data from Foursquare, but I didn’t find any data that is suitable in this project)