



Coursera/IBM

Capstone Project

By Andrew Dahlstrom

Modeling Local Demographics based on Local Venues in Toronto, Canada

October 05, 2019


1. Introduction

This Notebook is my submission for the final project in the IBM Applied Data Science Capstone Course. This assignment requires that I design an imagined scenario which will allow me to demonstrate the techniques and methodologies I have learned in the IBM Applied Data Science Specialization program as well as utilize the FourSquare API and the scikit-learn library for learning algorithms in Python.

The scenario I devised for this project is that I have been contracted by International Health Organization to conduct an exploratory case study with the goal to improve methodologies for building demographic maps. To accomplish this goal, I have chosen to experiment with modeling and predicting local demographics based on local venue data for neighborhoods in the city of Toronto, Canada.

2. Background

Population demographic maps are useful tools for many humanitarian efforts including to manage disease outbreaks, water scarcity, disaster relief efforts, electrical grid expansion, expansion of health or education services etc. Demographic data is not always readily available in some areas so any contribution to methodology that can improve the accuracy of population maps could be useful to humanitarian efforts. The goal of this project is to explore how accurately local venue data can



predict the demographic data for a neighborhood (categorized by postal code) by using relevant machine learning techniques to build a prediction model.

3. Data

3a. Sources

This city of Toronto was selected because of the detailed and recent demographic data available publicly and the large collection of venue data available for each neighborhood. The data for this project has been collected from the following sources:

1. **Toronto postal code data can be found on [Wikipedia]:**
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. **The demographic data for Toronto has been collected from the [2016 Census]:**
<https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&S=22&O=A>
3. **The local venue data will be retrieved from the [FourSquare API]:**
<https://developer.foursquare.com/>
4. **Latitude and longitude geospatial data for postal codes provided by IBM course website.**

3b. Notes

1. The demographic data comes from a 2016 census but the venue data is current meaning that the demographic data is lagging behind the venue data by a couple of years so this will affect the accuracy of the model.
2. For financial reasons I must limit the number of venues I am able to collect data on for each neighborhood. I will therefore find the distribution of the venues for each category within each neighborhood in order to create a venue profile for each neighborhood.
3. I will be exploring how well venue data predicts the age distribution of the population in each neighborhood separated into 5 year increments.

4. Methodology

The first step was to normalize the age data by finding the mean for each age category and then subtracting that from the mean of each postal code. This made it easier to notice differences between the age demographics neighborhoods rather than comparing population sizes between neighborhoods. I also used the sklearn built in preprocessing to normalize the features and labels with the StandardScaler.

Next I experimented with machine learning models from the sci-kit learn library in order to find the best model to predict age demographics based on the venue data. Since there are 21 age categories to predict with continuous output the problem is a multioutput regression problem which narrows down the number of possible potential models available to experiment with. I initially used linear regression models first then k-nearest neighbors and lastly the decision trees models. The two primary metrics I utilized to score and compare the performance of the regression model estimators were explained variance and r squared score (coefficient of determination) because these metrics work well on regression problems and the `r2_score` and `explained_variance_score` provide an objective measure of overall correlation. It is also important to note that I chose the "uniform_average" parameter for both metrics because the venue data which I was able to obtain only included the 100 highest ranked venues for each neighborhood so with this lack of data, weighting venue categories differently would not be appropriate.

Lastly, I used a different approach to explore the relationship between venue categories and age demographics by using clustering algorithms. These machine learning algorithms from the sci-kit learn library clustered neighborhoods together based on their age demographic data then compared the venue categories in each cluster based on their frequency and novelty. I initially leveraged the Affinity Propagation clustering algorithm to predict the appropriate number of clusters because this is a convenient feature of this particular clustering algorithm and since the dataset is small. I then used the standard KMeans clustering to form the clusters and the following metrics to score the model (max score = 1). Homogeneity describes how well clusters contain only members of a single label. Completeness measures how well all members of a given label are assigned to a particular cluster. V-measure represents the harmonic mean between homogeneity and completeness.

5. Results

After finding wide variations in the results of the same model re-training on the data multiple times, I investigated the venue data collection and found that by adding and setting the parameters of time/day to any in the search method then I could achieve a more consistent venue profile for each neighborhood. Despite the deficits in the supply of data, several of the models performed surprisingly well.

The Bayesian Ridge linear regressor in particular was able to explain about 43% of the local demographic data based on the local venue category data. In the sample table examining the model coefficients for the age group 0-4 years, the highest positively weighted venue categories were Latin American Restaurant, Brewery, Egyptian Restaurant and Farm. Some other relationships which occurred in this model were that the gym and athletic venues are more popular among the ages 20-34 and the BBQ Joint ranking #1 between the ages 45-74 and then is gradually replaced by asian and mediterranean restaurants.

The clustering model seemed to favor grouping people who are close in age together rather than more diverse combinations which is what you might expect if certain venue categories' appeal greatest to a particular age group and then gradually appeal less to people older or younger than that group following a distribution. To support the findings for the age group in the linear regressor for group 0-4 years the clustering model also found the Farm and Egyptian Restaurant were associated with the groups 0-4 and 5-9 years. The Latin American Restaurant was associated with the 10-14 and 15-19 years groups and the Brewery was not associated with younger age groups in the cluster model. The clustering model also presented clusters with dominate 10 year range age groups and associated with specific venue categories which overlap those seen in the linear regressor.

6. Discussion

This exploratory study provides lots of space for further discussion and exploration. Due to the small sample size of only 96 neighborhoods and the large number of demographic and venue categories (21 and 226 respectively), it is difficult to train a robust or generalized model on such a small yet diverse dataset. It is also important to note that there are a limited number of regression models available that are compatible with the multioutput extension compared to the models available for

single output regression or classification problems. It would of course be useful to include more cities and larger venue category datasets in a future study but given the results from this small study some useful observations can still be made.

The relationship between the demographic and venue data can be more nuanced. For example, when examining the Bayesian Ridge regression model the estimator coefficients for the age group 0-4 show venues that might not be intuitively associated with young children such as Brewery or the ethnic restaurants. One possible explanation is that the children are not the ones using these venues but perhaps their parents are. This appeared to be true for the ethnic restaurants but in the case of the brewery, it only appeared highly associated with the 0-4 and 5-9 age groups. There were 150 Breweries included in the venue data which means that it was not simply an outlier case.

The clustering model provided a novel perspective of the relationship by associating some new venue categories with different age groups such as Comedy Club associated with people in their 20s, Performing Arts Venue for people in their 30s and Sports Venues for people in both age groups. The Circus is a big attraction for people in their early 40s and various Asian cuisine restaurants are highly associated with people over 60 years old. Some of these findings are also supported in the linear regression model. The Comedy Club was weighted highly for people in their late 20s and the Circus was weighted heavily for people in their early 40s. The linear regressor also supports the trend of Asian cuisine restaurants being popular with older adults.

One distinct type of venue category that is highly associated with many different age groups is restaurants of various cuisines. Given how ethnically diverse the city of Toronto is, it's useful to examine its history of population growth and influxes of immigration from different ethnic populations at different periods in time. How long ago an influx of a particular Ethnic population occurred may explain the popularity of the respective cuisine with the current age of those patrons. Understanding the importance of historic demographic changes for the city of Toronto may be useful for Toronto but may not be consistent across other cities in Canada or even less consistent with other countries. Consequently, even though different types of restaurants are highly associated with different age groups, they may not generalize well and perhaps should be merged into broader categories or simply as restaurants in order to pursue a model which could be applied more generally. This same rationale could be used to combine other venue categories into broader venue categories thus significantly shrinking the total number of feature categories especially if the neighborhood sample size is small.

7. Conclusion

The purpose of this case study was primarily exploratory with the goal to understand novel approaches to improve methodologies for building demographic maps. Several different machine learning techniques were utilized to predict neighborhood age demographic data based on venue category data in the city of Toronto in order to better understand the relationship between the two. For this type of multioutput regression problem it seems that a linear regression model such as Bayesian Ridge may produce estimators with the highest prediction efficacy. The clustering approach reveals novel associations as well as supporting findings from the regression model making the two approaches complementary for this type of problem.

While the regression and clustering approaches can be mutually beneficial for modeling local demographics based on local venues, there are still unique local patterns of venues that can arise in response to historic, ethnic and other factors affecting the local population. In order to achieve a model which can be generalized for predicting age demographics in different cities, forming broader venue categories may be useful to overcome the unique local differences. This study has demonstrated that valuable insight into local age demographics can be gained from local venue data and combined with the techniques described in this study can help improve the overall methodology for constructing demographic maps.