Coursera Capstone project:

Suggesting locations to set up a new medical laboratory in Calgary using geolocation data and support vector machine models

Sergio Navarro

2021-09-04

Introduction and business problem

Dynacare Medical Laboratories is a Canadian medical laboratory services company. As part of its expansion projects, the company would like to set up a laboratory in Calgary, Alberta, where it does not provide any services yet. They would like to use data science techniques to determine a suitable location within Calgary for their new laboratories. I will analyze geolocation data with support vector machine models to suggest locations in Calgary.

Data

In this project, there were three main data sets:

1. The target data set: consisting of the information about the neighbourhoods in Calgary (example in table 1)
2. A positive data set: the information of neighbourhoods in major cities across Canada that have a Dynacare laboratory (example in table 2)
3. A positive-unlabelled data set: the information of all the neighbourhoods in major cities across Canada that have a Dynacare laboratory (example in table 3)

To prepare the data sets, I extracted the postal codes of the neighbourhoods of interest from Wikipedia or the Dynacare laboratories (Annex 1). I used the Beautiful soup library to extract the postal codes from the Wikipedia page. Next, I determined the latitudes and longitudes of each postal code using a web based latitude and longitude finder (https://www.latlong.net/). Once I had put together all this data, I obtained the venues for each neighbourhood using the Foursquare API, formatted them as one-hot encoding, and grouped them by their frequencies in the total venues that appear in each neighbourhood. After this last step, they were ready to be processed and analyzed by support vector machine (example in table 4).

Table 1. Calgary's neighbourhoods' informations

PostalCode	Neighborhood	Latitude	Longitude
T2A	Penbrooke Meadows, Marlborough	51.04968	-113.96432
T3A	Dalhousie, Edgemont, Hamptons, Hidden Valley	51.12454	-114.14289
T2B	Forest Lawn, Dover, Erin Woods	51.02533	-113.97890
T3B	Montgomery, Bowness, Silver Springs, Greenwood	51.08963	-114.19751
T2C	Lynnwood Ridge, Ogden, Foothills Industrial, Great Plains	50.98122	-113.99786

Table 2. Neighbourhoods with Dynacare laboratories

Postal code	Neighborhood	Latitude	Longitude
G1C	example lab 1	46.881771	-71.189369
G1E	example lab 2	46.860130	-71.194054
G1M	example lab 3	46.817230	-71.269836
G6W	example lab 4	46.757560	-71.225570
H1K	example lab 5	45.608180	-73.544520

Table 3. Neighourhoods from cities with Dynacare laboratories

PostalCode	Neighborhood	Latitude	Longitude
K2A	Highland Park, McKellar Park,Westboro,Glabar Park,Carlingwood	45.38025	-75.76138
K4A	Fallingbrook	45.46734	-75.47799
K1B	Blackburn Hamlet, Pine View, Sheffield Glen	45.42042	-75.59603
K2B	Britannia,Whitehaven, Bayshore, Pinecrest	45.36172	-75.78945
K4B	Navan	45.41413	-75.40364

Table 4. Venues in Calgary

Neighborhood name	Bakery	Bank	Bar	Breakfast Spot	Construction & Landscaping
Braeside, Cedarbrae, Woodbine	0.000000	0.000000	0.000000	0.000000	0.5
Brentwood, Collingwood, Nose Hill	0.000000	0.000000	0.000000	0.333333	0.000000
Bridgeland, Greenview, Zoo, YYC	0.000000	0.000000	0.000000	0.000000	0.000000
City Centre, Calgary Tower	0.022222	0.022222	0.066667	0.000000	0.000000

Methodology

A data set that containing the neighbourhoods of places where Dynacare has set up laboratories represents a data set with only positively-labeled data. Thus, a one-class support vector machine from the sci-learn library was used to process this data. For the positive-unlabelled data set, a support vector machine from the pulearn library that is based on Elkan-Noto technique was used (1). F1 and Jaccard scores where used to evaluate the models before applying them to the target data. The resulting positive locations were plotted in a map of Calgary for visualization using the folium library.

Results

Two support vector machine models were trained using positive-only and positive-unlabelled (PU) geolocation data sets to determine a suitable location to set up a medical laboratory for DynaCare Medical Laboratories. The Jaccard indexes and F1-scores of the models when tested with different sizes of test data are summarized in tables 5 and 6. The positive-only SVM was trained with a test size of 0.2, while the PU SVM used a test size of 0.25.

On table 7, the results of the suggested locations from each SVM model are summarized. The model trained with only positive data suggested the neighbourhoods corresponding to the following eight postal codes: T2C, T3C, T2E, T2G, T3G, T2L, T3P, and T2Y. The model trained with positive-unlabelled data suggested the neighbourhoods corresponding to the following twenty-one postal codes: T3A, T3B, T3E, T3G, T2H, T3H, T2J, T3J, T2K, T3K, T2L, T3L, T3M, T2N, T3P, T2S, T2V, T2W, T1Y, T2Y, and T2Z. The map in Figure 1 shows the locations suggested by each of the models as well as the locations suggested by both models.

Table 5. Jaccard and F1 scores for the positive only SVM model

Test size	Jaccard score	F1 Score
0.05	0.400000	0.571429
0.10	0.400000	0.571429
0.15	0.533333	0.695652
0.20	0.578947	0.733333
0.25	0.541667	0.702703
0.30	0.517241	0.681818
0.35	0.558824	0.716981
0.40	0.552632	0.711864
0.45	0.465116	0.634921
0.50	0.520833	0.684932

Table 6. Jaccard and F1 scores for the PU SVM model

Test size	Jaccard score	F1 Score
0.05	0.500000	0.666667
0.10	0.733333	0.846154
0.15	0.916667	0.956522
0.20	0.714286	0.833333
0.25	0.965517	0.982456
0.30	0.900000	0.947368
0.35	1.000000	1.000000
0.40	0.973684	0.986667
0.45	0.560976	0.718750
0.50	1.000000	1.000000

Table 7. Final results for the suggested neighbourhoods in Calgary. A “1” indicates it is a suggested location. A “-1” indicates it is not a suggested location.

Postal Code	Neighborhood	Latitude	Longitude	Positive -only Result	PU Result
T3A	Dalhousie, Edgemont, Hamptons, Hidden Valley	51.12454	-114.14289	-1	1.0
T2B	Forest Lawn, Dover, Erin Woods	51.02533	-113.97890	-1	-1.0
T3B	Montgomery, Bowness, Silver Springs, Greenwood	51.08963	-114.19751	-1	1.0
T2C	Lynnwood Ridge, Ogden, Foothills Industrial, Great Plains	50.98122	-113.99786	1	-1.0
T3C	Rosscarrock, Westgate, Wildwood, Shaganappi, Sunalta	51.04492	-114.13070	1	-1.0
T2E	Bridgeland, Greenview, Zoo, YYC	51.07029	-114.04284	1	-1.0
T3E	Lakeview, Glendale, Killarney, Glamorgan	51.02038	-114.13822	-1	1.0
T2G	Inglewood, Burnsland, Chinatown, East Victoria Park	51.03712	-114.04642	1	-1.0
T3G	Hawkwood, Arbour Lake, Citadel, Ranchlands, Royal Oak	51.13818	-114.20157	1	1.0
T2H	Highfield, Burns Industrial	50.99195	-114.06099	-1	1.0
T3H	Discovery Ridge, Signal Hill, West Springs,Christie Park,	51.04314	-114.19478	-1	1.0
T2J	Queensland, Lake Bonavista, Willow Park, Acadia	50.94264	-114.03868	-1	1.0
T3J	Martindale, Taradale, Falconridge, Saddle Ridge	51.11655	-113.94819	-1	1.0
T2K	Thorncliffe, Tuxedo Park	51.10199	-114.07128	-1	1.0
T3K	Sandstone, MacEwan Glen, Beddington, Harvest Hills	51.15312	-114.07394	-1	1.0
T2L	Brentwood, Collingwood, Nose Hill	51.09035	-114.12176	1	1.0
T3L	Tuscany, Scenic Acres	51.12323	-114.24007	-1	1.0
T2M	Mount Pleasant, Capitol Hill, Banff Trail	51.07077	-114.08753	-1	-1.0
T3M	Cranston, Auburn Bay, Mahogany	50.88795	-113.95621	-1	1.0
T2N	Kensington, Westmont, Parkdale, University	51.05859	-114.10576	-1	1.0
T2P	City Centre, Calgary Tower	51.04860	-114.07407	-1	-1.0
T3P	Symons Valley	51.17748	-114.10508	1	1.0
T2R	Connaught, West Victoria Park	51.04092	-114.07484	-1	-1.0
T2S	Elbow Park, Britannia, Parkhill, Mission	51.02494	-114.07260	-1	1.0
T2V	Oak Ridge, Haysboro, Kingsland, Kelvin Grove, Windsor Park	50.98051	-114.09409	-1	1.0
T2W	Braeside, Cedarbrae, Woodbine	50.94794	-114.10660	-1	1.0
T2X	Midnapore, Sundance	50.88608	-114.04169	-1	-1.0
T1Y	Rundle, Whitehorn, Monterey Park	51.08058	-113.96087	-1	1.0
T2Y	Millrise, Somerset, Bridlewood, Evergreen	50.91110	-114.09638	1	1.0
T2Z	Douglas Glen, McKenzie Lake, Copperfield, East Shepard	50.92148	-113.96479	-1	1.0

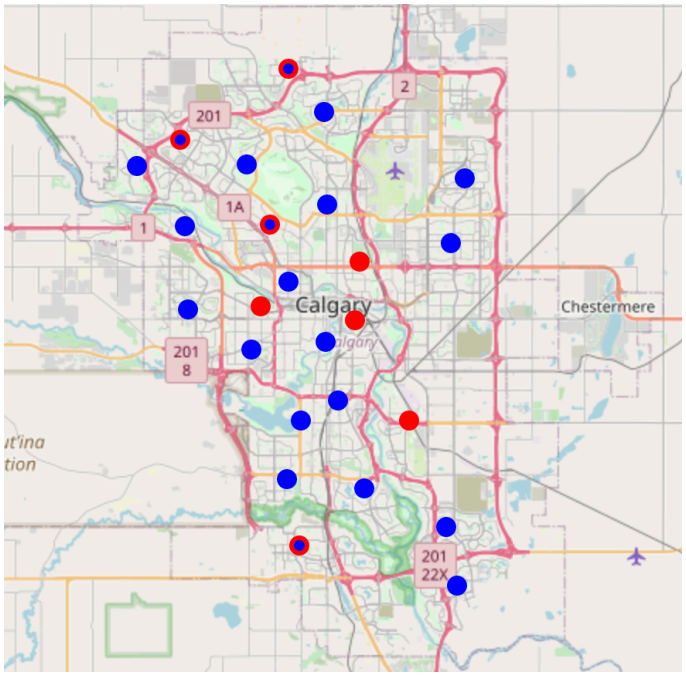

Figure 1. Labelled map with the neighbourhoods suggested by the two SVM models. Locations suggested only by the positive-only model are coloured in red. Locations suggested only by the pu model are coloured in blue. Locations suggested by both are coloured by blue and have a red contour.

Discussion

Analyzing geolocation data with machine learning can have many uses. In this project, geolocation data was analyzed with vector support machine models to determine potential locations to set up a laboratory in Calgary for Dynacare Medical Laboratories. The models were trained with positive-only and positive-unlabelled data and the results were plotted on the map of Calgary. Both models found different groups of potential locations that overlapped in some neighbourhoods. The locations suggested by the positive-only model show a vertical pattern along the middle of the city, but most of the suggestions are in the center. The suggestions of the positive-unlabelled model are more numerous and expanded across the western side of the city. It's possible that when then SVM model was trained with PU data, the thresholds to consider a location as positive was expanded.

Most of the suggestions of the models are found on the left side of the city. Perhaps the western side is more developed than the eastern side. Nevertheless, both models have the majority of their common suggestions in the northwestern part of Calgary. There is one common at the south of the city.

It is often the case that in data science project the data scientist has to work without knowing certain business information. Sometimes, to work around this, they are forced to make certain assumptions. I made several assumptions about the data used in this project. Firstly, I assumed that a certain neighbourhood is a good spot for a laboratory simply because Dynacare has set up one in there. To truly discern this, the geolocation data would have to be coupled with performance data of that laboratory and the relationship between these two data types would have to be evaluated. Also, I assumed that neighbourhoods in major cities without a Dynacare laboratory as considered to have an "unknown" status regarding their potential as good spots for a new laboratory. I would require more information from the business to know if there was a particular reason to not set up a laboratory there in relationship to its geolocation characteristics.

Precisely because of this lack of information, I had to work with positive-only and positive-unlabelled data sets to train the algorithms. Indeed, this represents a real-life situation where PU learning is applicable. While some PU learning-based SVM models already exist, such as the Elkanato-based used in the project, there are still some hurdles to be surmounted when

applying these models. One of the main current hurdles of PU learning models is the need for efficient evaluation models. As there are no negatives to test models in a PU data set, some typical evaluation methods such as the confusion matrix cannot be used to evaluate a PU model.

In this project, Jaccard indexes and F1-scores for tests sizes from 0.05 to 0.50 were tested in both SVM models. A test size of 0.20 was used for the positively-only SVM because it scored the highest in both evaluation methods. For the PU model, a test size of 0.25 was chosen because it was one of the test sizes that was in the middle and scored the highest in both evaluation methods. For some test sizes, the scores were 1.0. I stayed away from believing the accuracy of these evaluations as they may be the result of overfitting. Staying below 1.0 simply meant that some real-positives will be labelled as negatives. Since this project aims to reduce the number of possible locations for a new laboratory, it was ok to lose some potential places as an experimental error.

Some recommendations to expand this project and deal with some limitations are to use other classification models, implement more evaluation methods, and explore the qualities of the suggested locations.

Conclusion

Support vector machine models can be used to analyze geolocation data and find suitable places to set up new laboratories for Dynacare Medical Laboratories. These models can be implemented with data sets containing only positive-only data or positive and unlabelled data. This project represents a situation in real life where PU learning is applicable. Thus, it is important to develop methods to analyze positive-unlabelled data.

References

1. Elkan, Charles & Noto, Keith. (2008). *Learning classifiers from only positive and unlabeled data.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 213-220. 10.1145/1401890.1401920.
2. Mordelet, Fantine & Vert, Jean-Philippe. (2010). *A bagging SVM to learn from positive and unlabeled examples*. Pattern Recog. Lett. 37. 10.1016/j.patrec.2013.06.010.

Annex 1: websites used to extract postal codes, and latitudes and longitudes

Dynacare Medical laboratories:
https://www.dynacare.ca/find-a-location.aspx

Ottawa postal codes:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_K

Toronto postal codes:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Montréal and Laval postal codes:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H

Winnipeg postal codes:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_R

Calgary postal codes:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T

For latitudes and longitudes:
https://www.latlong.net/