Capstone Project - Battle of Neighbourhoods (Week 5)

1. Introduction

The desire to progress professionally and always go further is something innate in the human being. However, this progression is not without sacrifices. Many times, in order to progress, you have to get out of your comfort zone, reaching the extreme of having to move to another city or to another country.

The fact of moving to another city has a less impact than if we do it internationally, since as a general rule, by staying in our own culture, our basic needs are usually covered wherever we go. When the change is international, the cultural change is extremely greater, and at this time, choosing the area where you live will make us adapt better or but or that we manage to live at least more or less the same as when we were in our own country

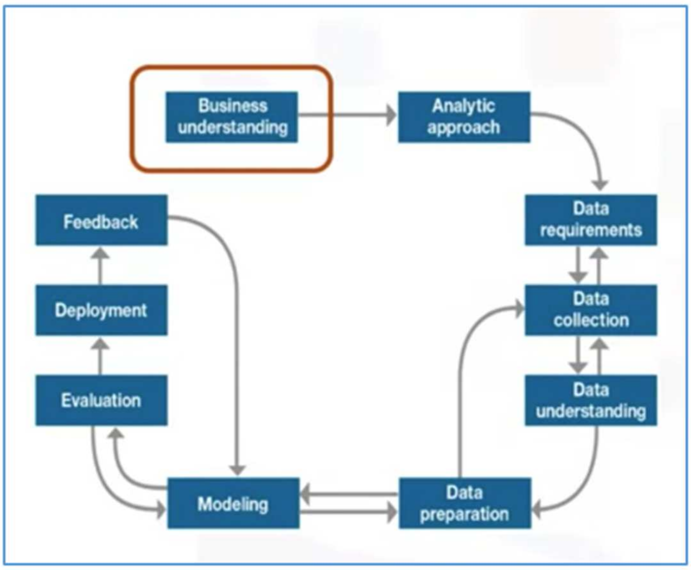
The Smith family (SF from now on) for work reasons had to move to Toronto. Due to the rapid incorporation to the new job, they did not have time to evaluate all the alternatives offered by their new house, which caused multiple changes of house later until they found the right one.

Due to their bad experience, Tom and Allison (who work for 2 very important multinationals) thought it would be very interesting to create an algorithm that allows their future colleagues (expatriates or not) or even other people, to choose quickly and accurately the best area in which to live based on their preferences.

If everything goes as expected, they will patent and commercialize this solution in order to monetize their model.

2. Methodology

In order to raise this challenge, SF relies on the *foundational data science methodology*. The methodology is highly iterative and never ends.



2.1 Business understanding

Very important stage where the goal is defined asking in the most appropriate way the principal questions.

The main questions to answer were:

- What are your basic needs?
- What strengths and opportunities should the neighborhood have?
- What can discard a zone do?

2.2 Analytic approach

Helps identify what type of patterns will be needed to address the question most effectively. The chosen analytic approach determines the data requirements. Specifically, the analytic methods to be used require certain data content, formats and representations, guided by domain knowledge.

In order to analyze the preferences and show the appropriate solutions, SF relied on the machine learning so-called *recommender systems*.

In an initial stage, the algorithm was personalized, so they used a *Content-based recommender system* algorith, but with the passage of time and when more information is available, they will be able to use a hybrid system based on *Collaborative filtering*.

2.3 Data requirements

Prior to undertaking the data collection and data preparation stages we define the data requirements for this reccommeder systems. This includes identifying the necessary data content, formats and sources for initial data collection.

In SF case, the main sources that they used were the following:

- Borough/PostalCode data Wikipedia→ https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Geospatial data for Toronto → http://cocl.us/Geospatial_data)
- Foursquare API to obtain information about venues and facilities→ https://foursquare.com/

- Random user data to train the model and show the final results.

2.4 Data Collection

Determine whether or not the sources selected have what we need. We evalute if we have to defer the use of some kind of data that it is not available at the moment.

After do it, SF can conclude that we have all the information needed to go ahead with the project.

Wikipedia

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with
suggesting that Canada Post may have reserved the M0 FSA for high volume add

Postal Code	Borough	
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Governm
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens

Geospatial

Postal Code,Latitude,Longitud
M1B,43.8066863,-79.1943534
M1C,43.7845351,-79.1604971
M1E,43.7635726,-79.1887115
M1G,43.7709921,-79.2169174
M1H,43.773136,-79.2394761
M1J,43.7447342,-79.2394761
M1K,43.7279292,-79.2620294
M1L,43.7111117,-79.2845772
M1M,43.716316,-79.2394761
M1N,43.692657,-79.2648481
M1P,43.7574096,-79.273304
M1R,43.7500715,-79.2958491

Foursquares

[7]:

	Neighbourhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	American Restaurant	Antique Shop	...	Vegetarian / Vegan Restaurant	Video Game Store
0	Malvern, Rouge	0	0	0	0	0	0	0	0	0	...	0	0
1	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	...	0	0
2	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	...	0	0
3	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	...	0	0
4	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	...	0	0

5 rows × 271 columns

2.5 Data Understanding and Preparation

SF used scrapping techniques to extract the dataframe with the postal codes, borough and neighborhood of the Wikipedia address indicated above. They took advantage the information extracting in order to clean up all fields with #N/A data o “Not assigned information”

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
11	M3B	North York	Don Mills

SF then downloaded the Toronto geographic information using Pandas in order to assign each neighborhood a latitude and longitude. Later, the family decided to merge these two tables into 1 where all the content was collected.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848

Next step was to connect with their Foursquare developer profile to extract all the information about venues and other facilities in the area. SF decided to put a limit of 100 stores per neighborhood and the range from the center of the neighborhood at 500 m.

```
                                         neighborhood_longitude))
LIMIT = 100
radius = 500
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
```

Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	5	5	5	5	5	5
Alderwood, Long Branch	6	6	6	6	6	6
Bathurst Manor, Wilson Heights, Downsview North	22	22	22	22	22	22
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	25	25	25	25	25	25
...	...	...	...	...	...	...
Willowdale, Willowdale East	33	33	33	33	33	33
Willowdale, Willowdale West	5	5	5	5	5	5
Woburn	4	4	4	4	4	4
Woodbine Heights	8	8	8	8	8	8
York Mills West	3	3	3	3	3	3

With the information collected, SF made a new table where all the categories of the neighborhood and the density of services in the area appeared as a ranking. They could build a clasification of each neighbourhood in which they could know which venue was more common.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	C
0	Agincourt	Lounge	Breakfast Spot	Clothing Store	Latin American Restaurant	Skating Rink	Yoga Studio	Doner Restaurant	Diner	Discount Store	Dist
1	Alderwood, Long Branch	Pizza Place	Coffee Shop	Athletics & Sports	Pub	Gym	Eastern European Restaurant	Electronics Store	Dumpling Restaurant	Drugstore	
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Pet Store	Frozen Yogurt Shop	Bridal Shop	Shopping Mall	Diner	Sandwich Place	Deli / Bodega	
3	Bayview Village	Café	Japanese Restaurant	Bank	Chinese Restaurant	Department Store	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Sandwich Place	Coffee Shop	Restaurant	Sushi Restaurant	Café	Indian Restaurant	Japanese Restaurant	Pub	

2.6 Modelling

Most important part after developing models is the calibration stage where we will see if our model answer in appropriate way the question. Training set (data in which the outcome is known) are used in predictive models in order evaluate the model.

With the information extracted in the previous table, a score was assigned to each neighborhood in each of the venues, and based on those scores a percentage that will give the final grade for the neighborhood depending on the user's preferences.

	Neighborhood	Salon / Barbershop	Sandwich Place	S L
0	Adelaide, King, Richmond	0.01	0.000000	0.
1	Agincourt	0.00	0.250000	0.
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.00	0.000000	0.
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.00	0.083333	0.
4	Alderwood, Long Branch	0.00	0.100000	0.

For training reasons, SF generated a random user and assigned him a list of 10 categories available in the city that answered the questions shown in 2.1. SF created a table with the categories as the columns and 1 row, where the values are 1 if the user has the category and 0 in contrary case. This will result in a user profile that will be used in the recommendation system.

Bus Station	Bus Stop	Business Service	Butcher	Cafeteria	Café	Cajun / Creole Restaurant	Ca Sto
1	0	0	0	0	0	0	0

Then was the time to evaluate the user's preference data with the recommendation matrix created. SF compared the user random profile to the table with the neighbourhoods and the mean of value for the amount of venues of each category in it.

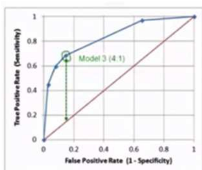
They multiply both matrix and apply a sum for each row. As the result they get a new matrix with the neighbourhoods and the score for each one of them. The higher the score the better the neighbourhood matches.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Score
0	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923	0.142857
1	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029	0.142857
2	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577	0.111111
3	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter	43.781638	-79.304302	0.100000
4	M4P	Central Toronto	Davisville North	43.712751	-79.390197	0.090909

2.7 Evaluation

Model evaluation is performed during model development and before the model is deployed. Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.

From this result, SF could see that the 2 best neighbourhoods for our first user were “North York Flemingdon Park, Don Mills South” and “East Birchmount Park, Ionview, Kennedy Park”. 2 areas had the same score, but the difference amount the 5 neighbourhoods is not big. A probable reason is that categories, which our user chose was more or less common, they don’t include anything extraordinary as “Airport Food Court”.



2.8 FeedBack

The model need to be improved, trainning with more users and looking for a hybrid model in a nearly future.