Ruslan Ibragimov
Data Science Applied Capstone - Coursera
IBM Professional Certificate

Research on effects of different types on venues on the price of property in the neighborhoods of Toronto

12.09.2019

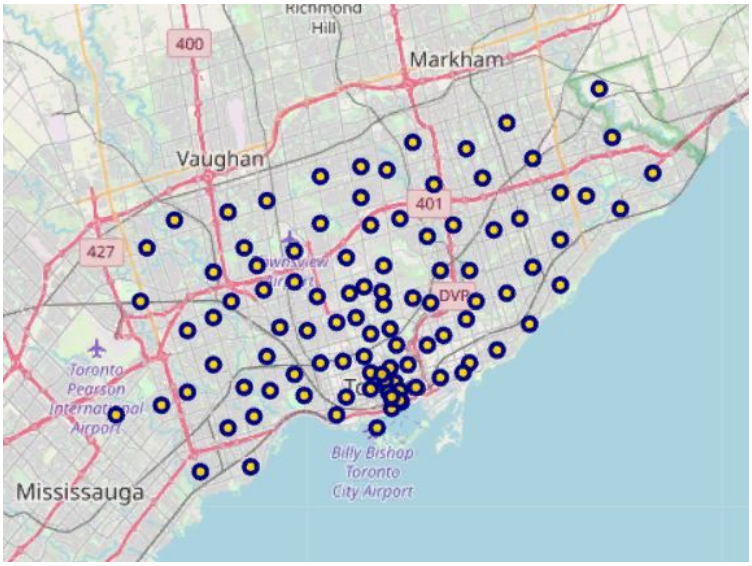


fig-1. Neighborhoods of Toronto.

I. Introduction and Business Understanding

1. Introduction

In this paper we will cover the analysis of impact of different venues in area on the price of private property in different neighborhoods of Toronto (Canada). Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. City is full of different venues, that affect the price of housing one way or another.

2. Problem

As it is commonly known, the price of housing in specific area is strongly dependent on the number of different venues, namely shops, restaurants and et criteria. The variation and the accessibility of them can be one of the first things to consider for customer during real estate purchase.

3. Interest

For better management and market research, it would be useful to know what kinds of venues influence the price more, than the others, therefore it explains the importance of this research. To do it we will use open datasets (**Data**) and different methods (**Methodology**). The main target audience would be city service workers, businessmen and people interested in buying or sell their property.

II. Data

1. Data Sources

We are going to need the next two datasets:

1. Neighbourhoods longitudes and latitudes from Toronto administration open portal (https://open.toronto.ca/dataset/neighbourhoods)
2. Dataset provided by city of Toronto administration on house prices in neighborhoods. (https://open.toronto.ca/dataset/wellbeing-toronto-economics)

2. Data Understanding

At the end we need to get the dataset that has the next format:

Name of neighborhood	Venues encoded by one hot encoding	Price

Features for evaluation will be venues encoded by one hot encoding, while labels will be present as price.

Features will give us understanding about what kind of venues is present in area of neighborhood, so we can connect it with the price.

The dataframe above is going to be created using above mentioned two sets:

1. We will use **Foursquare API** in order to get information of venues in neighborhoods, using longitude and latitude from the table.
2. We will export housing price from the second data set and merge it with the frame from the first step.
3. Then we will drop unused columns (‘Borough’ or ‘Post code’) and finish wrangling of dataset.

The complete dataset will have a look like this:

	Neighborhood	Price	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Warehouse Store	Wine Bar
1	Agincourt South-Malvern West	317508	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
2	Alderwood	251119	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
3	Annex	414216	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
4	Banbury-Don Mills	392271	0.066667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
5	Bathurst Manor	233832	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

3. Data Cleaning

We are going to need to clean the data, due to the large presence of missing values not for all neighborhoods we have got information about longitude and latitude. We also have a lot of unrequired data in both tables, so we need to get rid of this as well. In principle, data cleaning is not going to be very complex for this task, as datasets provided are good enough, But on different stages of model development we will need to perform dealing with missing values anyway and we make an assumption, that deleted data will not be crucial for this research.

III. Analysis of features and features selection

1. Analysis of features

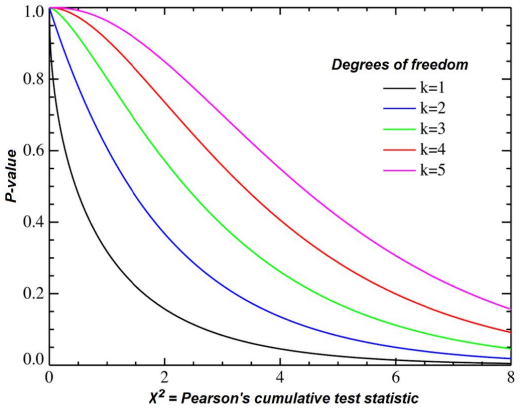
We have the wide range of features with different presence on the streets of Toronto.

American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Volleyball Court	Warehouse Store	Wine Bar	Wine Shop	Wings Joint
0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
0.066667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

fig-2. In total we have 279 types of venues.

However, from looking at this data, we can notice that not all of them are encountered often enough, on the opposite a lot of them are encountered very rarely and not even in all neighborhoods. To work with this data and determine, how does it influence the price, we need to apply some transformations, namely extract data only where it is present and clean it before usage. After this we can proceed to Chi2 calculation.

2. Chi-Squared calculation



"A chi-squared test, also written as χ2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories."

Chi-Squared test - Wikipedia

I have been using Chi2 test in order to determine the influence of the feature on the target variable - the price. To do it we used feature_selection module from scikit library. Sometimes it was not possible to calculate Chi2 score, so we have also got some missing values.

```
1. Feature: Art Museum Chi-Square: [nan] p-value: [nan]
2. Feature: Arts & Crafts Store Chi-Square: [0.01584799] p-value: [0.99996877]
3. Feature: Asian Restaurant Chi-Square: [0.58367541] p-value: [0.99998616]
4. Feature: Athletics & Sports Chi-Square: [0.36947123] p-value: [0.9990846]
```

The example of output is below. Full fule.

After this we can finally determine which features can be considered as "good" and which can be considered as "bad". We create the table, consisting features and corresponding Chi2 score and p-value, then we clean it from nan values, remove definitely "bad" features with value less than 0.7 and after data processing we will have the next table:

	Name of venue	Chi-Square	P-value
202	Park	10.651422	1.000000
209	Pizza Place	4.422612	1.000000
229	Sandwich Place	3.837578	1.000000
210	Playground	3.836209	0.974466
59	Coffee Shop	3.655064	1.000000
45	Café	3.505373	1.000000
223	Restaurant	3.426532	1.000000
96	Fast Food Restaurant	3.142190	1.000000
69	Convenience Store	2.404999	1.000000
148	Indian Restaurant	2.154149	0.999998
109	Furniture / Home Store	1.956862	0.998642
155	Italian Restaurant	1.730417	1.000000
156	Japanese Restaurant	1.713890	1.000000
141	Hotel	1.641055	0.990111

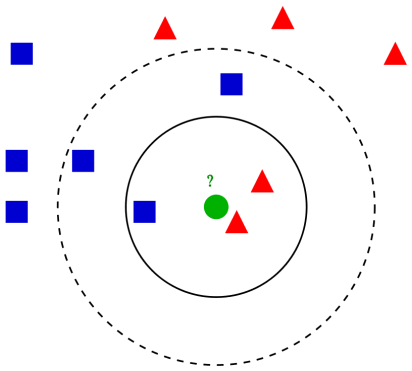
At the end we have selected 38 features that have been deemed to be good. Now we need to test them.

IV. Build of prediction model and results

1. Confirmation

In order to confirm, that we have selected the right features, we can try to build a prediction model. To do it we will use the part of our set as the test data (20 neighborhoods) and use the rest as the training data. We also need to convert continuous numeric values (in column 'Price') to categorical ones, making the assumption that prices below 260000 are considered to be **'Low'**, price above 450000 are considered to be **'High'** and everything in the middle is considered to be **'Medium'**.

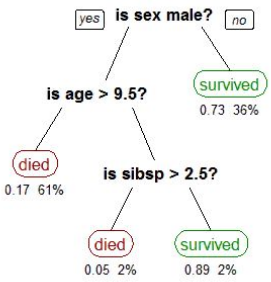
2. Classifier: kNN (k-Nearest Neighbors)



The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning). The distance can, in general, be any metric measure: standard Euclidean distance is the most common choice.

After training of our algorithm, we have tested it and got accuracy equal to **55%**

3. Classifier: D3



The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

Decision Tree on Wikipedia

After training of our algorithm, we have tested it and got accuracy equal to **65%,** which is the best one for this project.

4. Classifier: SVC

“Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.”

Support-vector machine - Wikipedia

The accuracy of this algorithm was equal to 55%.

V. Discussion

Despite the fact, that selected features seemed to be logical (the proximity to parks, fast food and restaurants can indeed influence the price of the property) it is hard to draw conclusions about price only basing on the types and numbers of venues, due to this all of three classification algorithms showed relatively poor results. However, it is still bigger, than accuracy of random selection, so we can confirm our statement, that selected types of venues have impact on the price. The best classifier was a Decision Tree and it can explained by the nature of this classifier: Each number of specific type of venues will bring result closer to one of the price labels.

To sum up, we can say, that our hypothesis was valid and found features indeed have influence on the price, but to more precise determination of the price we need to have some future development.

VI. Future Development

It is possible to increase the the precision of classification by better work with data - first of all we can group venues internally (‘Asian Restaurants and American Restaurants’ into category ‘Restaurants’) as it will help us to improve the results by losing less data. We also can cluster all neighborhoods into different clusters, so we will have less noise data and probably more precision on classification.

In generally, our current results can be used for brief analysing of the price or with future development it can be used as one of the components for economical calculator and calculation of the price of property.