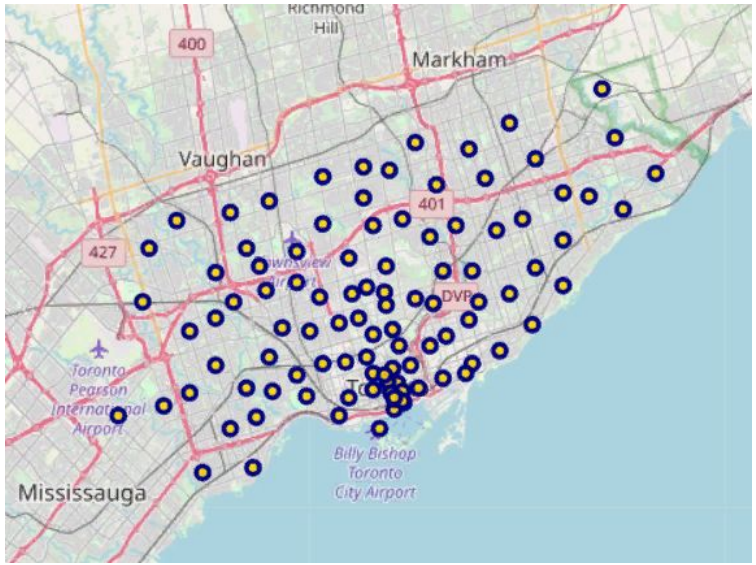
Research on effects of different types on venues on the price of property in the neighborhoods of Toronto

IBM Professional Certificate - Capstone Project

Introduction

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.

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.



Data Acquisition and Cleaning

Data Sources

We are going to need the next two datasets:

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

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.

Analysis of features and features selection

Chi-Squared test as the feature selector

I have been using Chi2 test in order to determine the influence of the feature on the target variable - the price.

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

Classification and test of features

Conformation

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’.

Results of classification

Out of three classifiers (kNN, SVC and D3) the best performing model was 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.

The accuracy of kNN classifier: 55%

The accuracy of SVC: 55%

The accuracy of D3 classifier: 65%

Conclusions and future development

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.

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.