IBM – Coursera

Data Science Specialization

Capstone project - Final report

Predicting House Price Based on School and Venues Vicinity

Divya Mahesh

Table of content:

I. Problem Statement: 2
II. Data description: 3
III. Methodology: 5
1. First insight using visualization: 5
2. Linear Regression: 6
3. Gradient Boosting Regression (GBR): 6
IV. Results: 7
VI. Conclusion: 10
References: 11
Table of Figures: 12

I. PROBLEM STATEMENT:

The vicinity of schools in a neighborhood is widely believed to be a key determinant of housing prices. However, the strength of the consensus is puzzling, given the formidable empirical challenges facing any homeowner or empirical researcher seeking to answer the question carefully.

Good schools usually come bundled with other neighborhood qualities—such as proximity to employment, shopping and recreational conveniences, and neighborhood peers. Because the home buyers who enjoy (and can afford) such amenities tend to congregate together, it is difficult to isolate the effect of schools from the effect of these other traits that accompany good schools.

The main goal will be exploring the neighborhoods of Greater Toronto Area in order to find out if the average house price of any given neighborhood depends on the number of schools and venues present in the vicinity.

The complexity of this problem lies in cleaning and aggregating diverse data and integrating it.

II. Data description:

GTA neighborhoods were chosen as the observation target due to the following reasons:

- The availability of real estate prices.
- The diversity of prices between neighborhoods.
- The availability of school data.

The type of real estate to be considered is detached homes.

DATA SETS

- Zolo.ca – Toronto Real Estate Website
- TDSB – Toronto District School Board
- FourSquare Data
- Wikipedia

The process of collecting and clean data:

- Scrap the zolo.ca webpage for a list of GTA neighborhoods and their corresponding detached home average price.
- Find the geographic data of the neighborhoods.
- Find the venues present for each neighborhood using Foursquare data
- For each neighborhood, find the total number of elementary, intermediate and secondary schools from TDSB website
- Standardize the average price by removing the mean and scaling to unit variance.

The result dataset is a 2 dimensions data frame:

- Each row represents a neighborhood.
- The last column will be the standardized average price.

	Neighborhood	Average Selling Price	Postalcode	Borough	Latitude	Longitude	Number of Schools	ATM	Accessories Store	Afghan Restaurant	...	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo
0	Danforth	1300000	M4C	East York	43.695344	-79.318389	14	0	0	0	...	1	0	0	0	0	0	0	0	0	0
1	Crescent Town	674000	M4C	East York	43.695344	-79.318389	14	0	0	0	...	1	0	0	0	0	0	0	0	0	0
2	Danforth Village-East York	1100000	M4C	East York	43.695344	-79.318389	14	0	0	0	...	1	0	0	0	0	0	0	0	0	0
3	Beechborough-Greenbrook	826000	M6M	York	43.691116	-79.476013	12	0	0	0	...	1	0	0	0	0	1	0	0	0	0
4	Brookhaven-Amesbury	780000	M6M	York	43.691116	-79.476013	12	0	0	0	...	1	0	0	0	0	1	0	0	0	0

Figure 1 - Final dataset

The dataset has 143 records.

III. Methodology:

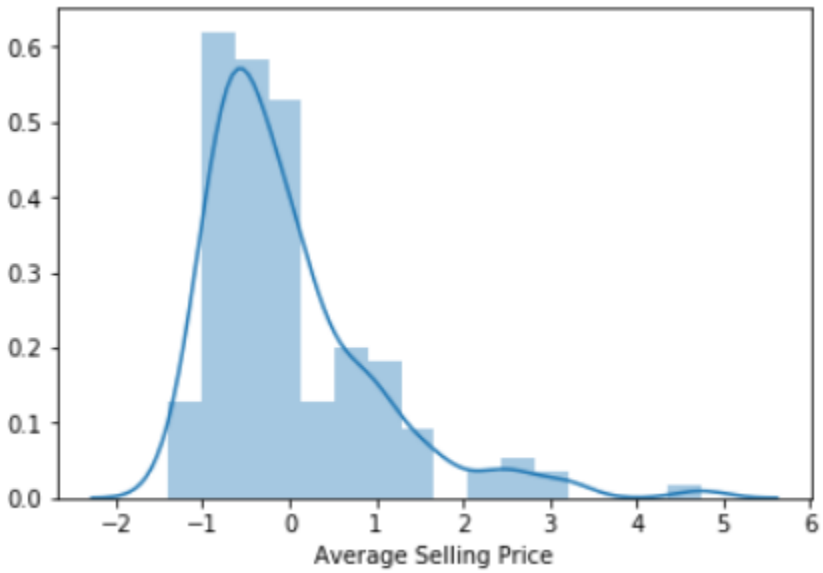
The assumption is that real estate price is dependent on the number of schools and venues available in the neighborhood. Thus, regression techniques will be used to analyze the dataset. The regressors will be the number of schools and venues in the neighborhood. And the dependent variable will be standardized average selling prices.

Python data science tools will be used to help analyze the data. Completed code can be found here: https://github.com/divyamahesh94/IBM-Capstone-Final-Project/blob/master/Predicting%20House%20Price%20Based%20On%20School%20Vicinity.ipynb

1. First insight using visualization:

In order to have a first insight of GTA real estate average price between neighborhoods, there is no better way than visualization.

The medium chosen is distribution plot.



2. Linear Regression:

Linear Regression was chosen because it is a simple technique. And by using Sklearn library, implementing the model is quick and easy. Which is perfect to start the analyzing process.

The model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

The result doesn't seem very promising. R2 score is 1, which means the model may be overfitted.

```
R2-score: 0.9880565498302533
Mean Squared Error: 0.010095485970080774


MAE: 0.0372075917102445
MSE: 0.010095485970080774
```

Figure 2 - Linear Regression result

The results seem to suggest that there is an overfitting.

Looking back further to the dataset, its dimensions sizes is not enough with only 144 samples. Logical steps to take are either collecting more samples. But since there are no other public source available, increasing sample size is not possible now.

And that's why Gradient Boosting Regression is chosen to analyze the dataset in the next part.

3. Gradient Boosting Regression (GBR):

"Boosting" in machine learning is a way of combining multiple simple models into a single composite model. This is also why boosting is known as an additive model, since simple models (also known as weak learners) are added one at a time, while keeping existing trees in the model unchanged. As we combine more and more simple models, the complete final model becomes a stronger predictor. The term "gradient"

in "gradient boosting" comes from the fact that the algorithm uses gradient descent to minimize the loss.

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value.

This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual. This residual predicted by a weak model is added to the existing model input and thus this process nudges the model towards the correct target. Repeating this step again and again improves the overall model prediction.

Again, R2 score and MSE are used to see how well the model fit the dataset.

```
R^2 is:
 0.9978553335892455


MAE: 0.011931057266387133
MSE: 0.0018128304093501474
```

Figure 4 - GBR scores

The result is promising as it shows improvement over the simple Linear Regression.

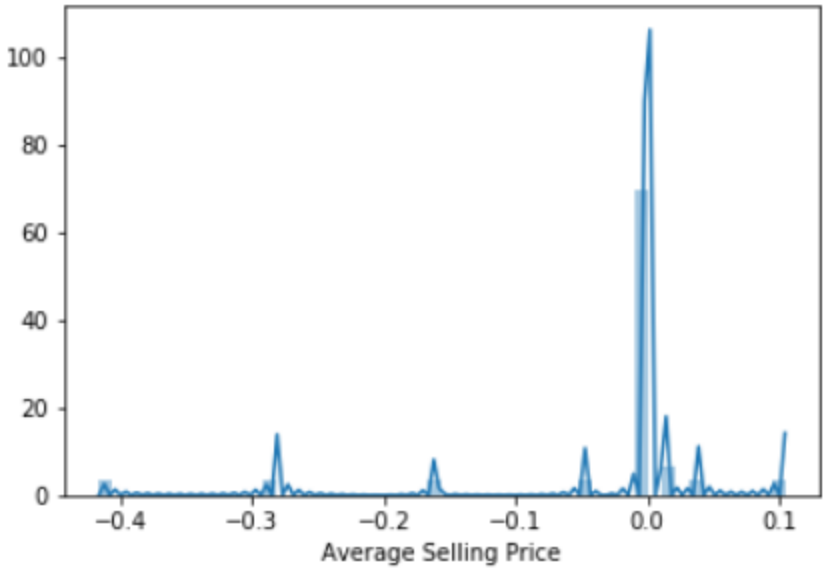
The insight is still consistent compared to the Linear Regression's.

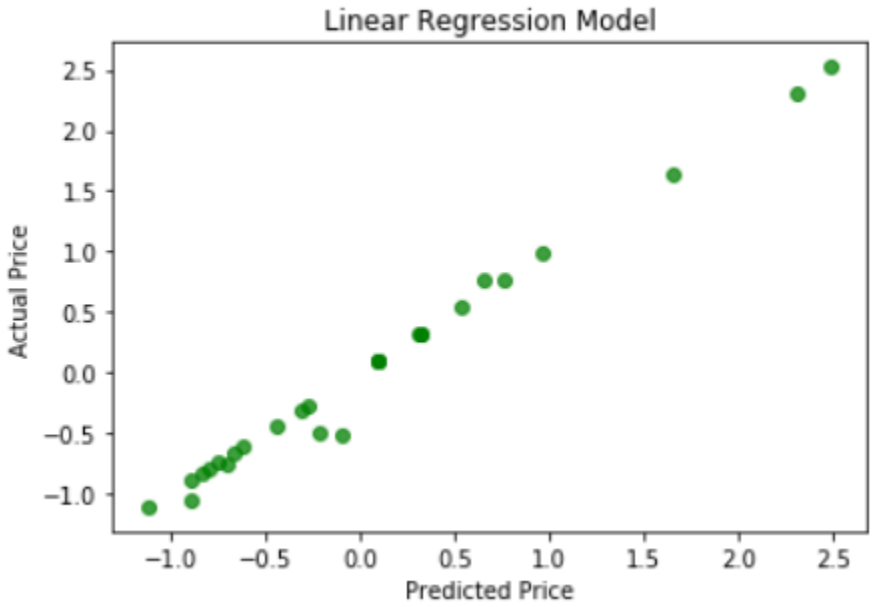
IV. Results:

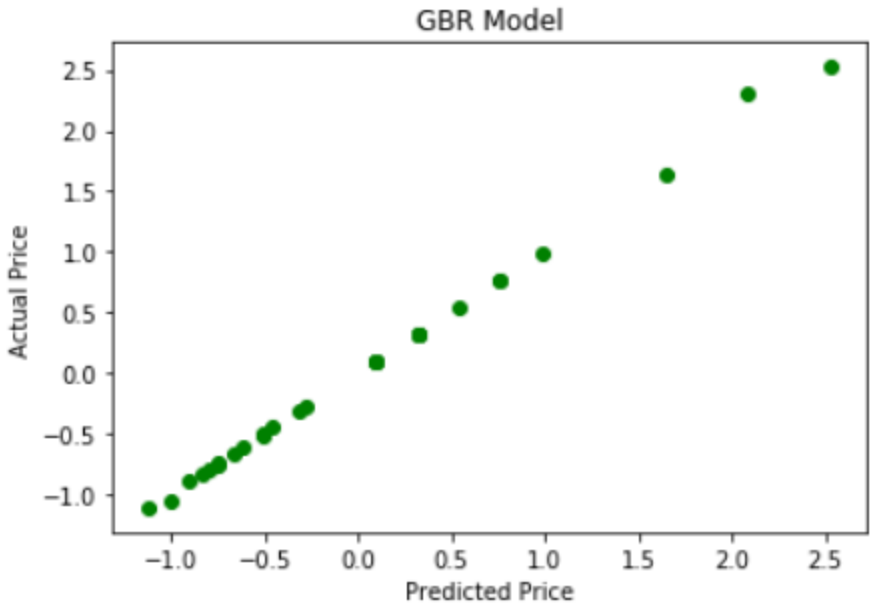
Even though the scores seem to be improved after applying a more sophisticate method, the model is still not suitable for the dataset. Thus, it can't be used to precisely predict a neighborhood average price.

Explanations for the poor model can be:

- The real estate price depends on number of factors.
- The data is incomplete (small sample size, missing deciding factors).
- The machine learning techniques are chosen poorly.


100
80
60
40
20
0
−0.4
−0.3
−0.2
−0.1
0.0
0.1
Average Selling Price


Linear Regression Model
Actual Price
Predicted Price
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


GBR Model
Actual Price
Predicted Price
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

V. Conclusion:

We focused our efforts on finding out if the number of schools and venues present in the vicinity affects the price of a property. As natural next step, we would like to gather more data on the quality and rankings of these schools and built a model to predict prices.

References:

https://blog.paperspace.com/implementing-gradient-boosting-regression-python/

Table of Figures:

Figure 1 - Final dataset 4
Figure 2 - New York city real estate price spread between neighborhoods ..... **Error! Bookmark not defined.**
Figure 3 - Linear Regression result 6
Figure 4 - PCR scores 7
Figure 5 - Coefficient list in original size ..... **Error! Bookmark not defined.**