Property speculation: A potential way to capitalise on the increasing demand for housing in Toronto

Jonathan Looman
February 11, 2020

1 Introduction

1.1 Background

Toronto has seen a tremendous increase in property prices between 2014 - 2019, with some neighbourhoods increasing in value by as much as 142% on average. With immigration only set to increase in the future it can be expected that there will be sustained demand for properties in cities like Toronto and other neighbourhoods will see increases in their property prices when th will lead to increased further increases .A novel way to determine the demand of an area is to find what type of neighbourhoods are the most sort after and finding similar neighbourhoods based on amenities.

1.2 Problem

The problem lies in finding the best place to invest. This project looks to find out whether similarity of neighbourhoods can be used to predict demand for neighbourhoods and thus its increase in property prices

1.3 Interest

This information would be beneficial to those who are:

- Estate agents and prospective property buyers could easily find properties similar to ones that they like if they are unable due to lake of supply or expense.
- Those looking to invest in Toronto's booming property market and are looking for additional factors to decide which properties are the best investment.

2 Data acquisition and cleaning

2.1 Data source

The two components of the project are: the neighbourhood amenity data and the community growth data. The first part, the neighbourhood data, was accessed from Wikipedia to get a list of the neighbourhood names and a csv file of geo locations provided by coursera was used to then find each neighbourhoods latitude and longitude. Finally foursquare was used to get the amenities present in 500m of the neighbourhood. The second part was to initially gain the growth rates of each neighbourhood. The comminity price increases, both in canadian dollar terms and percentage increases were scraped from an estate agencies website. Additionally a list of high growth neighbourhoods were scraped from another estate agency website. Data from both websites originated from the Toronto Estate board.

2.2 Data Cleaning

The postcode was the primary means of referencing neighbourhoods. Thus neighbourhoods which had the same postcode were joined together. Postcodes which did not have neighbourhoods associated with them were dropped. The Geo location CSV file was then merged with the neighbourhoods data. Finally, using one-hot encoding, new columns for each venue category were created and were between zero and one to represent the relative amount of similar venues in the neighbourhood compared to the rest of Toronto. The community growth data was initially scrapped from separate web pages into lists and were joined into one data frame. The growth both in dollar terms as well as in percent terms was then join to the neighbourhood amenity data frame. To account for spelling and errors in either of the data frames fuzzy logic matching was used.

2.3 Feature Selection

Venue categories taken from foursquare are combined and the most common values are listed in columns ranging from 1st most common to 10th most common. This will be used to group the data into similar neighbourhoods as a potential prediction method.

To differentiate neighbourhoods further, the average property price per neighbourhood was calculated from the increase values.

From 10 years annualized growth A popular way in which stock prices are analysed is with momentum. This momentum value is calculated by combining prices from the stocks past but giving more weight to values which are closer to the present date. Momentum is a good indicator of growth as its takes the price increase over a long period into account but gives the most recent data the highest impact in the score.

3 Exploratory Data Analysis

3.1 calculating target variable

Multiple target variables will be analysed. We will try to show through correlation, spacial and property amenity analysis which property price ranges, geographical locations and amenity are in the highest demand.

3.2 Relationship between property price and property growth

Figure 1 plots the Price of property against the percentage growth the property has had over the last 10 years. As is illustrated by the red and yellow circles, the highest property value growth lay in the middle of the property price range, showing that there is a non-linear relationship between a properties price and its growth. This plot also shows that the largest demand is for mid-priced properties.

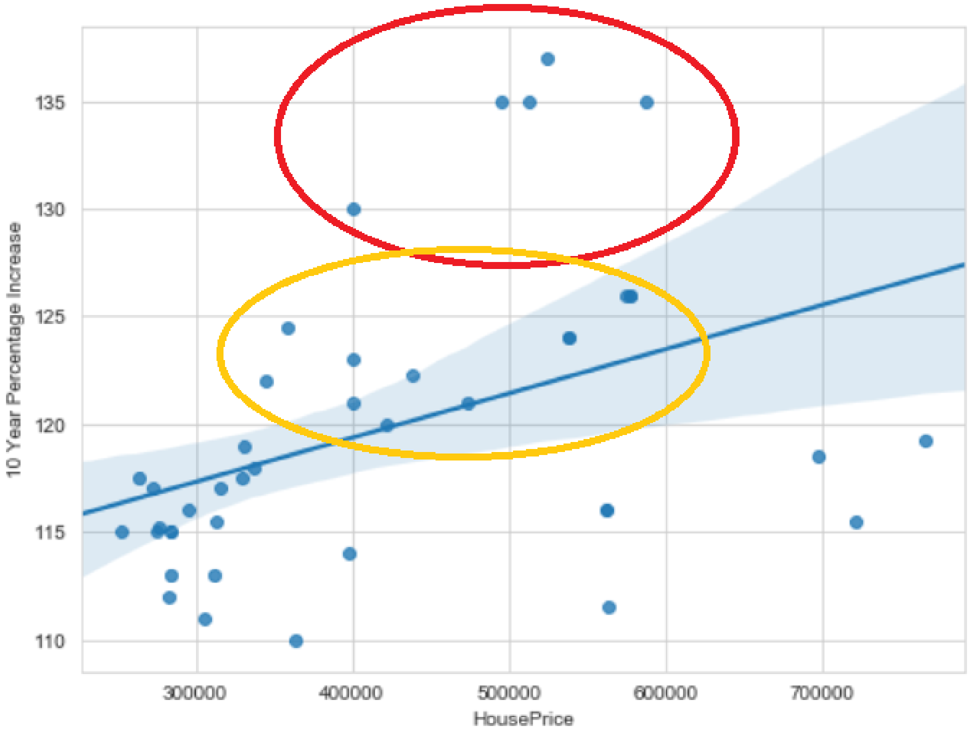


Figure 1: Property prices vs 10 year percentage growth

Figure 2 confirms that majority of mid range property prices had large increases over the past 10 years as a large majority of the mid valued properties fell into the high growth category. Spatially it appears areas in proximity to the coastline had the greatest increase in value. Neighbourhoods in Old Toronto of lower value saw the lowest percentage growth over the past 10 years.

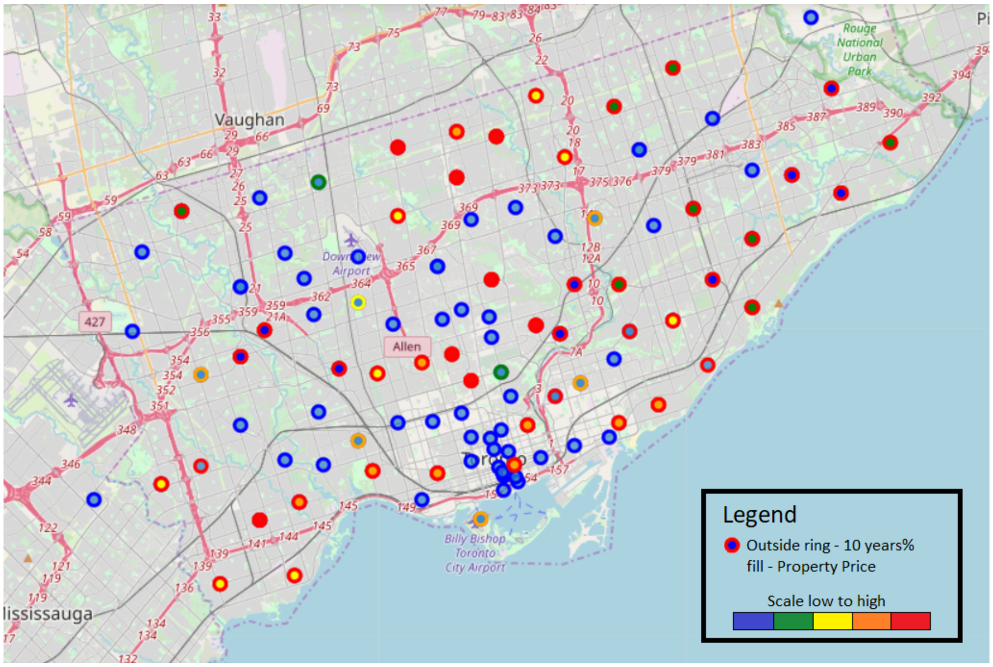


Figure 2: Map of property prices and 10 year percentage growth

3.3 Relationship between property price and momentum

Momentum which weighs recent growth more heavily than growth further in the past shows a very similar relationship to historical growth as depicted in Figure 1. This confirms that middle priced properties tend to have the greatest demand in Toronto.

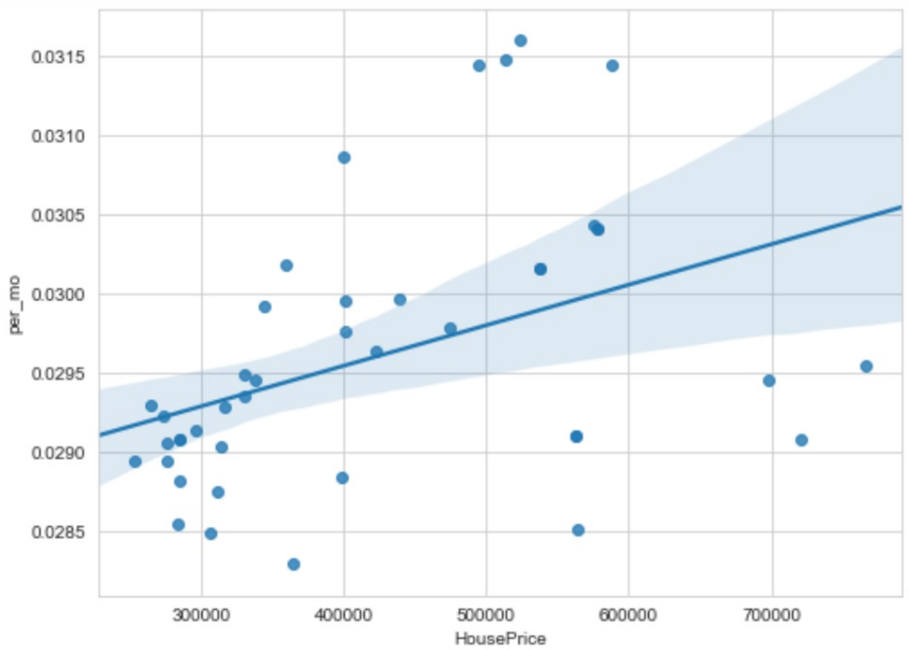
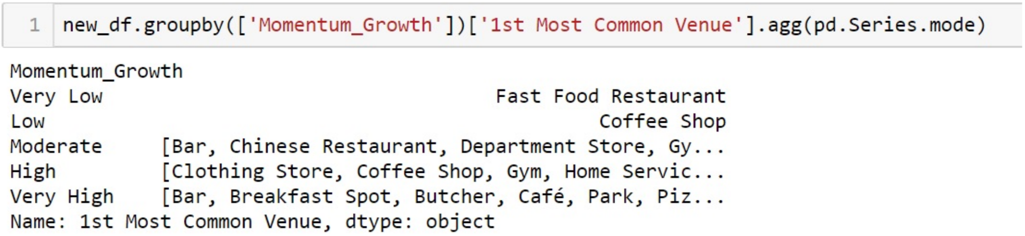


Figure 3: Property prices vs percentage momentum

3.3.1 Relationship between growth and Venue types

For a property to increase in value there must be an increased demand for it. Figure 4 shows that areas which have increased demand do not have one venue type which is the most popular in contrast to lower growth neighbourhoods which do have one type of restaurant as their favourite. The lowest growth neighbourhoods have 'Fast Foods' as their most common venue type which could dictate the taste of people looking to buy properties there.



```
new_df.groupby(['Momentum_Growth'])['1st Most Common Venue'].agg(pd.Series.mode)
```

```
Momentum_Growth
Very Low                                 Fast Food Restaurant
Low                                               Coffee Shop
Moderate     [Bar, Chinese Restaurant, Department Store, Gy...
High         [Clothing Store, Coffee Shop, Gym, Home Servic...
Very High    [Bar, Breakfast Spot, Butcher, Café, Park, Piz...
Name: 1st Most Common Venue, dtype: object
```

Figure 4: Most Common venue types by growth bracket

4 Model Development ad Results

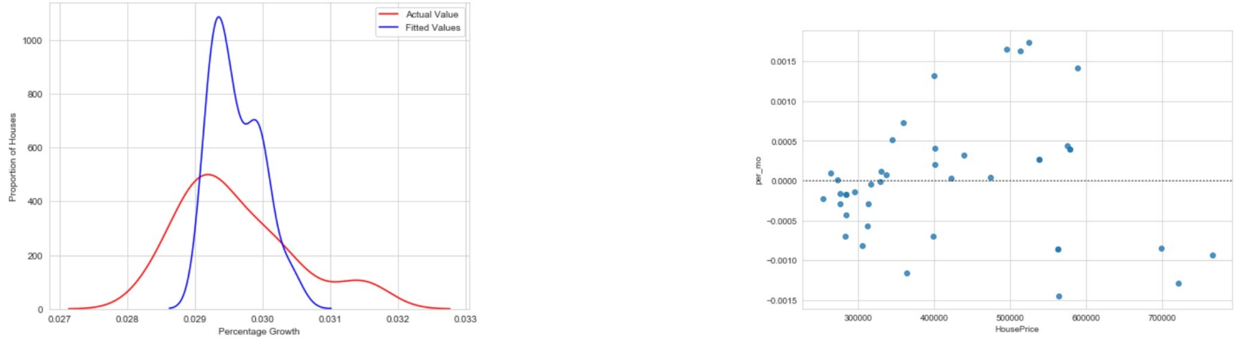
Multiple models will look to predict growth of an area based on its price and on the venues present in the area.

4.1 Linear Model

A good starting point from any model development is to try the simplest way first as it has limited costs and gives a good baseline to work from. The linear model used to price of a house to predict its growth. As can be seen in Figure 5a the linear model is slightly biased and is not percise. Additionally it has a second small local peak far outside the true range. The residual plot shown in Figure 5b shows a potential non linear relationship but the evidence is weak.

4.2 None Linear Model

As discussed in section 3, a nonlinear relationship exists between property price and growth. Additionally in section 3.3 it is shown that there is a relationship between growth and a neighbourhoods venue types. These two relationships will be used to predict property growth in a multiple regression.



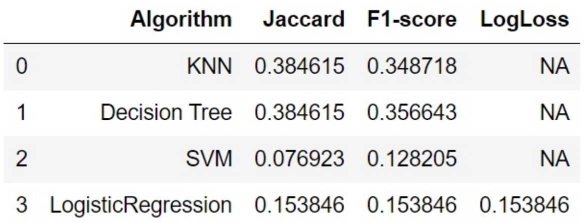
(a) Fitted vs actual values of Growth per House.

(b) Residual Plot of Growth.

Figure 5: Linear Model Performance.

4.3 Decision Tree, Support vector Machine and K Nearest Neighbours

Average neighbourhoods property growth rates were binned into categories ranging from very low to very high. Three classification machine learning approaches were then taken to see if using property prices and the relative frequency of venue categories could be used to classify a neighbourhoods growth. As can be seen in Figure 6, all three methods were unable to accurately predict growth from the independent variables.



	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.384615	0.348718	NA
1	Decision Tree	0.384615	0.356643	NA
2	SVM	0.076923	0.128205	NA
3	LogisticRegression	0.153846	0.153846	0.153846

Figure 6: Prediction Scores

5 Discussion

5.1 Usable information from research

From the exploration section of this project, there are numerous insights which could be used by prospective homeowners to help make informed decisions about which properties can be expected to have the best growth. Neighbourhoods whose average property price lay in the mid-range can be expected to have the best growth. Additionally, neighbourhoods closer to the coast saw larger growth to similarly priced neighbourhoods further from the coast or in the old town, as shown in Figure 2
Unfortunately, using venue types is not a very accurate way to predict property growth. There is weak evidence to suggest that there is a lower demand for neighbourhoods which have large numbers of fast food resultants as shown in Figure 4. Potentially a neural network or a random Forrest approach would be better able to combine the large amount of sparse data present in the relative venue category columns.

6 Conclusion

The results show that it is difficult to predict growth from venue types using machine learning approaches of a decision tree, SVM, KNN, linear and non-linear regression. There has been some insight gained into the type of properties that are in demand in Toronto such as the desired price range and geographical location. Using a Neural Network approach may help better find trends in the venue category data but it appears that more data about other attributes of the neighbourhoods are required to make accurate assessments of the growth.