**Using Nearby Businesses to Predict Toronto Apartment Rental Prices**

**by Bailey Duncan**

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/monopoly.jpg)

**Table of Contents**

1. [Introduction](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#1---introduction)
2. [Data Description](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#2---data-description)
3. [Methodology](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#3---methodology)
4. [Results](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#4---results)
5. [Discussion](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#5---discussion)
6. [Conclusion](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#6---conclusion)
7. [References](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#7---references)

**1 - Introduction**

Can we predict a Toronto apartment's rental price by knowing the businesses around it? This capstone is an exploration to see what effect businesses have on nearby apartment prices, and to try and determine the trends that lead to more expensive or desirable living space. If we know that certain businesses impact the price of an apartment when they open nearby, we can locate businesses that are more valuable to people. It can help land developers and business owners better forecast how new businesses can impact apartment rental prices in neighborhoods.

[To Top](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#table-of-contents)

**2 - Data Description**

The data used will be a combination of datasets, and data from API calls. The primary dataset is compromised of Toronto apartment rental prices for 2018. The dataset contains the price, address, latitude, longitude, and rooms of 1124 apartments for rent in the Toronto area. This dataset is publicly available on Kaggle[1](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#7---references). The Latitude, Longitude values in the dataset will be used with the Foursquare API to find the businesses nearby each apartment.

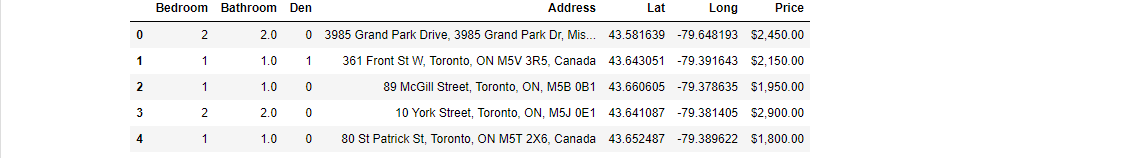
[To Top](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#table-of-contents)

**3 - Methodology**

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

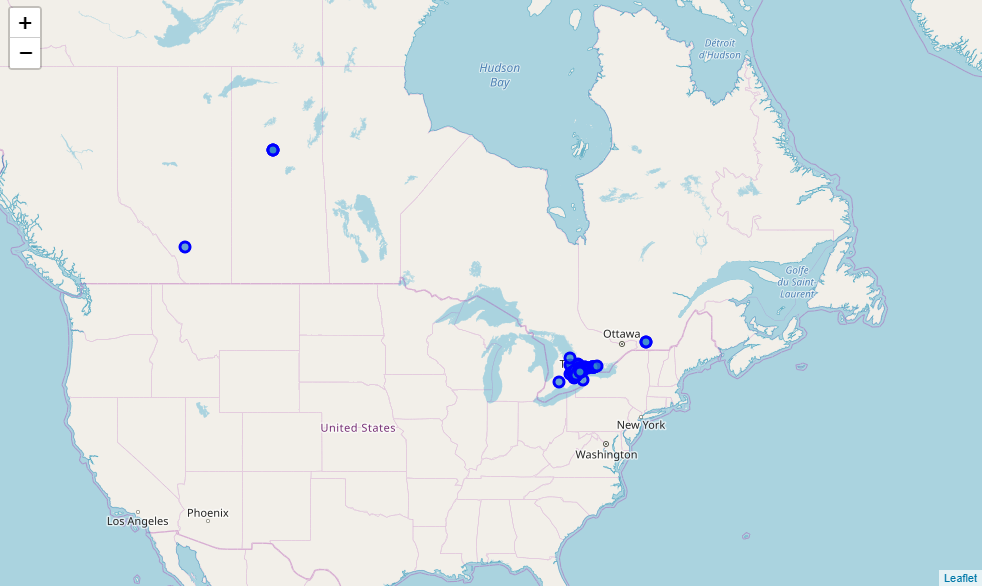
**3.1 - Data Cleansing**

Several steps were taken to prepare the kaggle dataset for the project. The apartment prices were initially string values prefixed by a dollar sign, so they were converted to float values. The address feature was not interesting so I decided to extract the first three letters of the apartment postal code using a regex, and drop the rest of the address information.

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/initial_dataset.PNG)

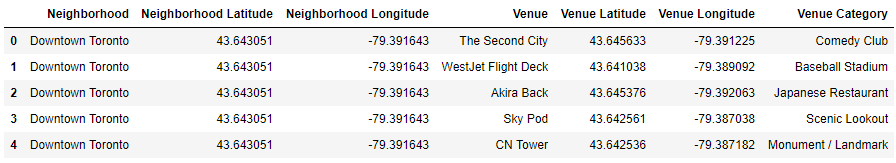
*Figure 1. Image of the uncleaned Kaggle dataset in a pandas dataframe*

That way I could easily group apartments by the postal code area they reside in (it could also be done with the latitude/longitude, but not without a lot of effort setting up neighborhood boundaries). Next the apartments were mapped using their latitude/longitude so I could identify outliers. I found and removed several stray apartments in Montreal, Calgary, Winnipeg and other Ontario cities that should not have been in the Toronto dataset. I had a dataset listing Toronto neighborhoods for each postal code from a previous project[2](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#7---references), so I used that to label the neighborhood and borough of each apartment in the dataset.

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/toronto_apt_capstone_zoom4.PNG)

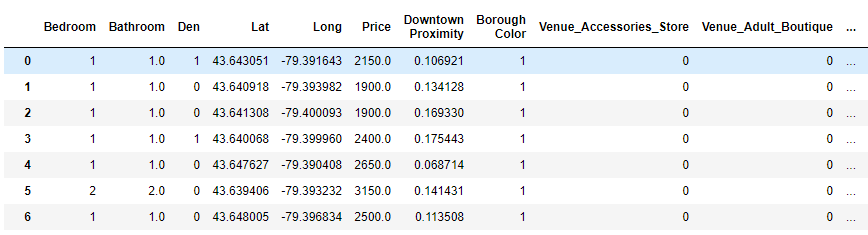
*Figure 2. Stray Apartments in the Dataset*

From this point I looked at the latitude/longitude columns and thought it looked like a less useful feature because of its scale, difficulty to work with, and lack of meaning to us without a map alongside it. I decided to derive a new feature from them called "Distance to Downtown Toronto" by calculating the euclidean distance from City Hall. Then I normalized it with a Min/Max Scaler to shift the values to all be between zero and one. That way all of our features are positive, and most of our features are in the zero-to-one range for any models that are sensitive to scaling.

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/venue_information.PNG)

*Figure 3. Toronto Venue Data*

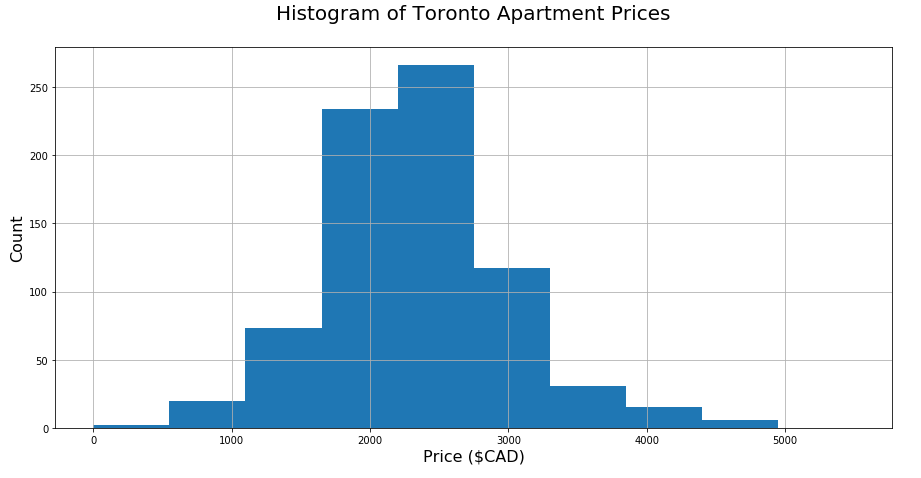
Next I used FourSquare to get the **venues** (FourSquare's term for businesses, amenities, parks, etc) within one kilometer of each apartment. I converted these venues into a one-hot encoded matrix, then aggregated all the rows related to one apartment, and merged it with the prepared kaggle dataframe.

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/apt_and_venue_df.PNG)

*Figure 4. Apartments data merged with the Venue Matrix*

**3.2 - Exploratory Analysis**

The bulk of the exploratory analysis was focused on the kaggle dataset.

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/apartment%20price%20histogram.PNG)

*Figure 5. Histogram of Toronto Apartment Prices*

Looking at the apartment prices distribution, we see it forms a nice bell curve centered around a $2500 rental price, ranging from around $500 to $5000. It is skewed right, which makes sense as the apartment prices have a hard limit before $0, but can be any value on the right side of the curve (ie. they extend from 0 to **∞**). This leads to a long trailing tail on the right side of our distribution and a short left side of the distribution.

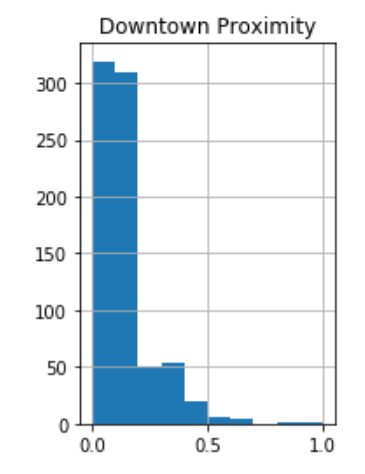
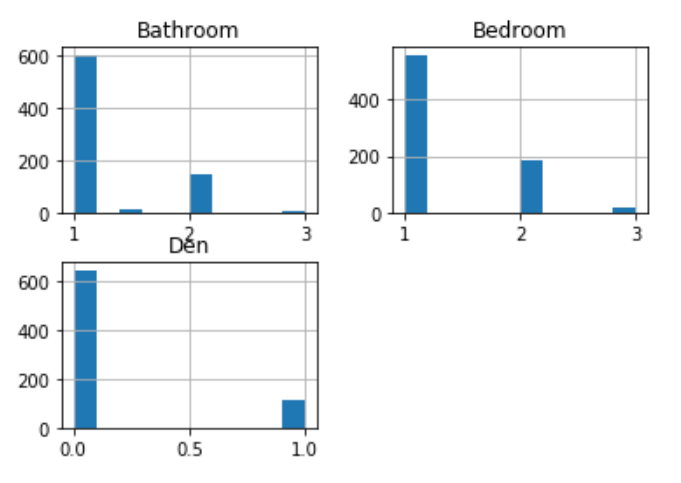
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/downtown_proximity.PNG)[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/bed_bath_den.PNG)

Figure 6.1 Histogram of Toronto Apartment Proximity to Downtown. Figure 6.2 Histogram of Toronto Apartment Rooms

We can see from figure 6.1 that the majority of our apartments are in downtown, and are very close to city hall. From Figure 6.2 we see that most of our apartment are single bedroom, single bathroom, and without a den. That makes sense for most downtown apartments. Lets plot the two against each other and see what kind of trend we get.

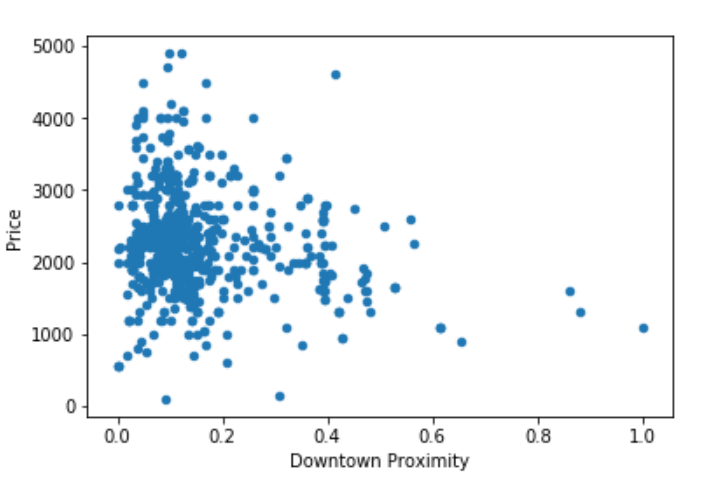
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/downtown_prox_vs_price.PNG)

Figure 7 Toronto Apartment Price vs. Downtown Proximity

Looking at the plot, its very clear that any model we create is going to have trouble identifying prices when a house is close to downtown. Lets filter this plot by our room types and hopefully it will become much clearer.

I am going to look at the bedrooms, bathrooms and dens to identify correlations and trends without introducing our additional venue data. Lets use the same plot as above to see if those characteristics can help us identify different price brackets close to downtown.

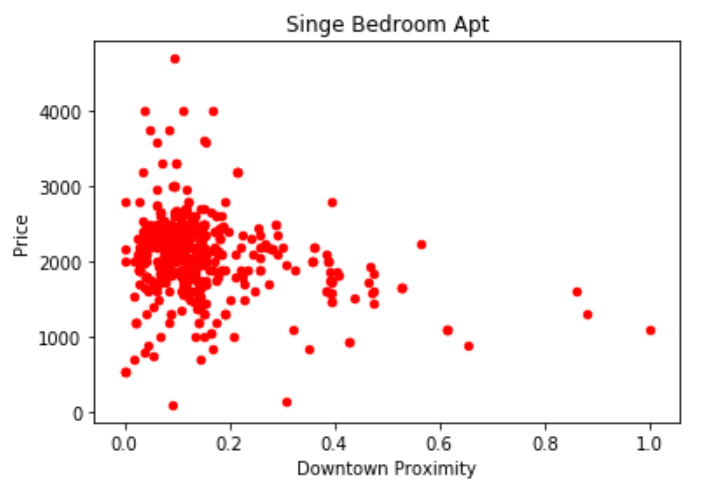
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/single_bedroom_price.PNG)

Figure 8. Single Bedroom Apartments: Price vs Proximity to Downtown

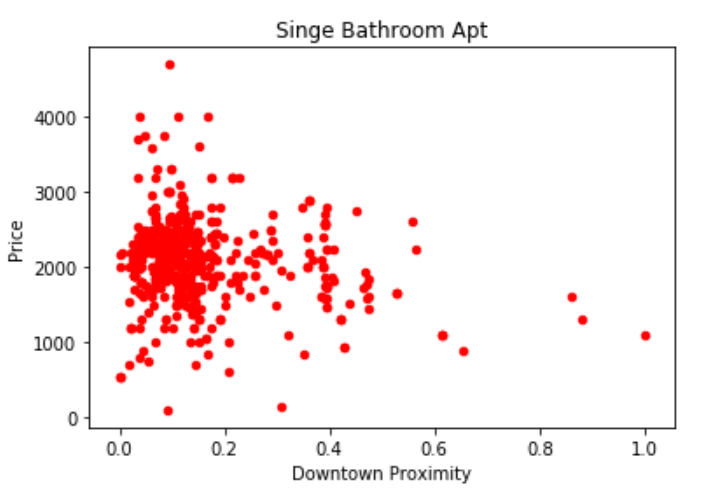
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/single_bathroom_price.PNG)

Figure 9. Single Bathroom Apartments: Price vs Proximity to Downtown

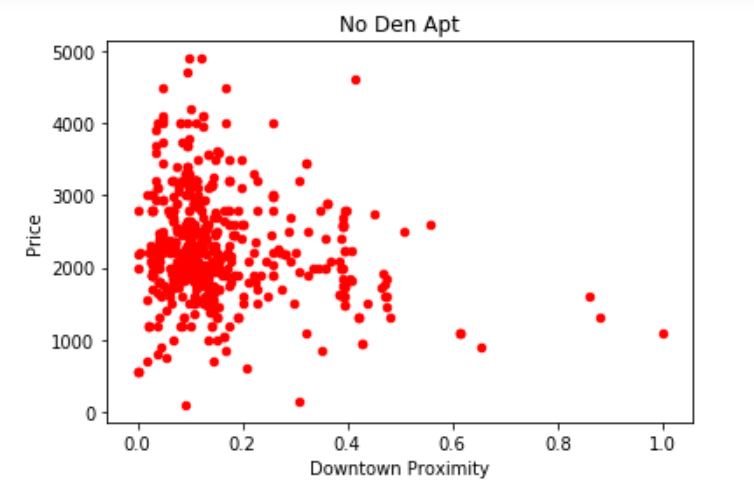
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/no_den_price.PNG)

Figure 10. Apartments without a Den: Price vs Proximity to Downtown

We see from these three figures that the prices look somewhat linear right of 0.4 scaled distance units, however everything from 0 - 0.4 units is nowhere near a linear trend. Even combining these attributes, we are going to have a lot of error with a regression model unless we introduce more data. This is where the venue data will help us to get better predictions.

### 3.3 Use of Machine Learning

An array of regression algorithms were used to benchmark the performance on the dataset. I started with simple linear algorithms, and worked my way up to ensemble methods like Random Forest and Gradient Boosted Decision Trees. Each model was evaluated by taking the mean squared error of the model price prediction against the actual rental price.

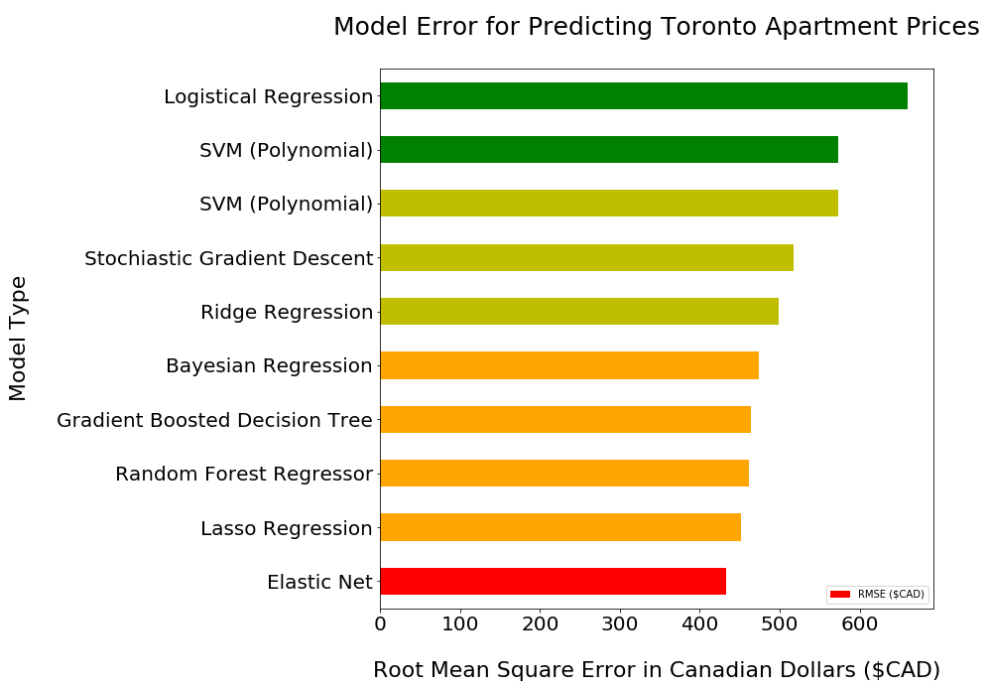
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/model%20evaluation.PNG)

Figure 11. Benchmarking Machine Learning Algorithms

The elastic net performs the best out-of-the-box, so that is the model we will select. It has a root mean square error (RMSE) of $432 and a R-squared correlation of 61% on the test set.

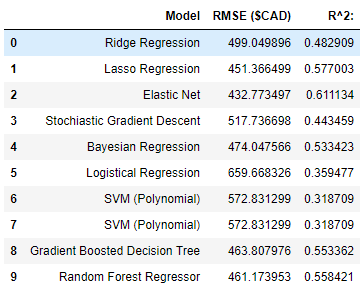
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/model_results.PNG)

Figure 12. Benchmarking Machine Learning Algorithms

Next I tuned the model hyperparameters to make the model the best it can be, evaluating its improvement with the test set. I started by doing a random search of the parameters to find a good guess for each, and then used grid search to find the best parameters starting with those guesses.

[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/test_results.PNG)

Figure 13. Test Set Prediction Results on Tuned Model

After hyperparameter tuning we get the test set RMSE down to $299, with an accuracy of 86%! Finally we test our tuned model on the validation set and got 74% accuracy with a MSE of $341.

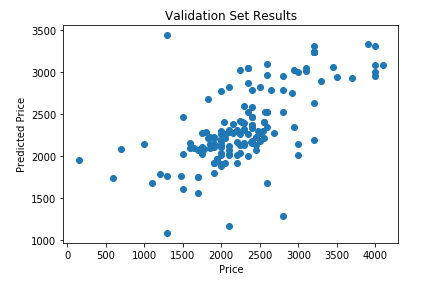
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/validation_results.PNG)

Figure 14. Validation Set Prediction Results on Tuned Model

[To Top](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#table-of-contents)

## 4 - Results

So far we have shown that we can use machine learning models to predict Toronto apartment prices using the nearby venues and apartment characteristics. From our machine learning model we can look at the most important features to understand the key factors that drive our model. The highest weighted features (**key contributors**) and the lowest weighted features (**key detractors**) will be extracted from our model.

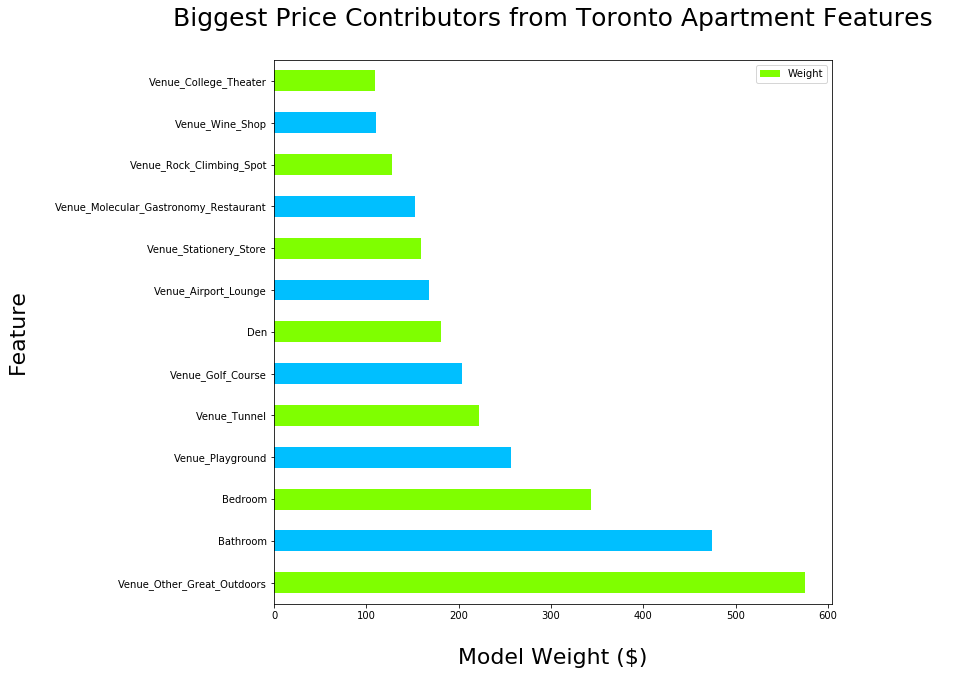
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/contributors.PNG)

Figure 15. Key Contributors to Toronto Apartment Prices

After seeing our key contributors, we see that apartments with more Dens, Bathrooms, and Bedrooms have higher prices. That is no surprise. However we also see that most of the contributors are related to outdoors activities, like parks, golf courses, playgrounds, outdoor scenery, and rock climbing. Other notable businesses are wine stores and molecular gastronomy restuarants.

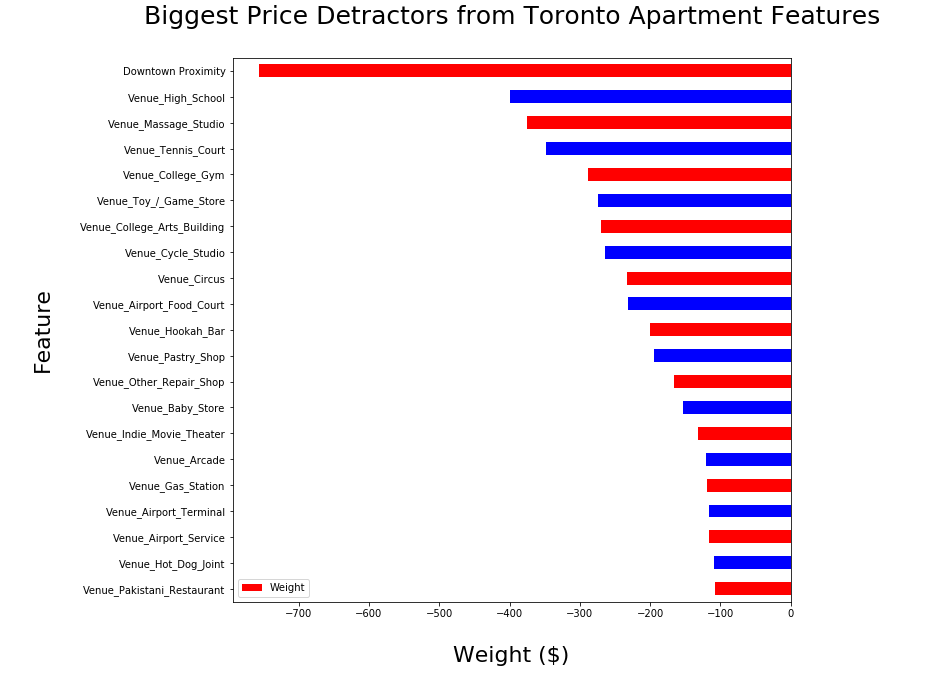
[](https://github.com/BKDuncan/Coursera-Capstone/blob/master/images/detractors.PNG)

Figure 16. Key Detractors to Toronto Apartment Prices

Based on the biggest detractors, we see that apartments close to the airport, far from downtown, and near high schools have the lowest prices. Some of the businesses that detract from the price are car mechanics, hot dog stands, massage studios, toy & game stores, arcades, Pakistani restuarants, and hookah bars.

[To Top](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#table-of-contents)

## 5 - Discussion

Based on the results of our model and its contributors/detractors there are several recommendations we can make.

For apartment building developers, I would recommend developing apartment buildings in neighborhoods with lots of parks, nature, outdoor activities, and as close to downtown as possible, but not near the airport. The units will have the highest price when these conditions are met and they are not surrounded by car mechanics, hot dog stands, high schools, etc.

As an owner of an apartment complex, you can drive prices higher, and make higher demand by investing in more green space, parks, playgrounds, and outdoor activities in the neighborhood.

As a tenant looking for an apartment, the price will be lowest if you live further from downtown or closer to the airport. By looking for apartments close to high schools, or without green space nearby, you can find apartments with a similar number of rooms for a lower rental price.

In terms of expanding the model and generalizing it for other cities, we would need significantly more apartment data. Because the data we had was mostly centered around downtown Toronto, it might not be a great indicator for the surrounding cities. As well, the Toronto apartment prices will differ significantly from the apartment prices of other cities due to demand, population, and other factors.

[To Top](https://github.com/BKDuncan/Coursera-Capstone/blob/master/Capstone%20Report.md#table-of-contents)

## 6 - Conclusion

In this capstone project, we used Toronto apartment location, rooms, and nearby businesses to train an elastic net regression model to predict the monthly rental price of an apartment. The model can predict the price on average within $341 CAD on our validation set, with 74% accuracy.

By extracting our most important features, we found out that apartments far from the airport, close to downtown, near high schools, and with outdoor activities like green space, scenery, and parks are higher value apartments for the same number of rooms. This allowed us to make recomendations to tenants, apartment complex owners, and apartment building developers to evaluate apartment value, and possibly increase it by investing in nearby ammenities, and businesses operating in the neighborhood.