**Artificial Intelligence Project**

**Greater Toronto Area Housing Price Predictor: Linear-Regression Model**

CP468 - Group #12

Mohamad Mansour 200449150

Haris Ejaz 200713450

August 12, 2023

Table of Contents

# Abstract

Our project aims to predict the housing prices of the Greater Toronto Area (GTA) housing market. This was done through training and machine learning models on the 2022 listings in the GTA, and used to predict prices using different parameters. Data preprocessing was performed on the dataset to make the model more accurate and remove any unnecessary fields. Then upon investigation of the parameters we were able to find a relationship and split the data into training and test sets on a linear regression model. We then evaluated the accuracy of the model by running it on the test dataset and comparing the actual price of the property vs the predicted price. Further, to improve the model we tried to minimize mean absolute error, and the root mean squared value.

By creating an accurate model, realtors, buyers and even sellers can enter metrics of a property to find out a predicted value of how much a property should be valued at considering similar properties in the area. For our purposes the model can make accurate purposes but has room for improvement by introducing new parameters to make improved predictions.

# Introduction

Living in the Greater Toronto Area (GTA) has gotten more and more expensive over the years. It is one of the fastest growing real-estate markets in North America. Toronto being the economic and cultural center of Canada, attracts thousands of people yearly in its ever growing housing market. Over the last few decades we have seen the population of the GTA grow immensely. But with the high property demand and low inventory of affordable homes, we have seen houses get sold for hundreds of thousands over their list prices, and have “bidding wars” making it difficult for newcomers or first time buyers. This is why we found the need to create a machine learning model that will be able to accurately help people in the housing market evaluate the property value, so that they can ensure they are getting a fair value for the property.

# Project Description

## Dataset

The dataset we used was downloaded from Kaggle (Singh,  2022). The data set came with 6 fields Price, Region, Address, Bedrooms, Bathrooms, Pricem (PriceInMillions). These are the fields we had to analyze and train our model to create a predicted valuation of each property. See figure 1 below (*Note 7324 records before data cleaning and preprocessing).

(7324, 6)

| | price | region | address | bedrooms | bathrooms | pricem |
|---|---|---|---|---|---|---|
| **0** | 799000 | Ajax, ON | 2 ROLLO DR, Ajax, Ontario | 3 | 3 | 0.799000 |
| **1** | 989000 | Ajax, ON | 717 OLD HARWOOD AVE, Ajax, Ontario | 2 | 1 | 0.989000 |
| **2** | 999900 | Ajax, ON | 52 ADDLEY CRES, Ajax, Ontario | 3 | 4 | 0.999900 |
| **3** | 799900 | Ajax, ON | 249 MONARCH AVE, Ajax, Ontario | 3 | 3 | 0.799900 |
| **4** | 899999 | Ajax, ON | 18 MONK CRES, Ajax, Ontario | 3 | 3 | 0.899999 |

**Fig. 1** - Initial project dataset read into Python

**Data Consideration**

We limited our dataset so that we could remove any outliers. We decided not to consider any homes besides single detached homes, which is restricting the dataset down to a specific type of homes as apartments, condos and complexes would fluctuate the dataset and make it difficult to model.

**Data Preprocessing**

We initially decided to remove one field being “address”, as this would not be useful information on the house relative pricing in location. Further, knowing that a small portion of the homes in the GTA would be high end mansions, these would play as outliers in the data and could skew the regression value higher. Thus the data set was further restricted to homes with prices higher than 500k and less than 2 Million. Additionally, we had to pre-process the region field of the dataset which had each property “city, ON”, which was not usable for training our linear regression model. Hence, we had to use the Pandas library called “pd.get_dummies”, to split the region into separate columns making the dataset more readable for the computer. Feature scaling was also done to reduce variance of the dataset.

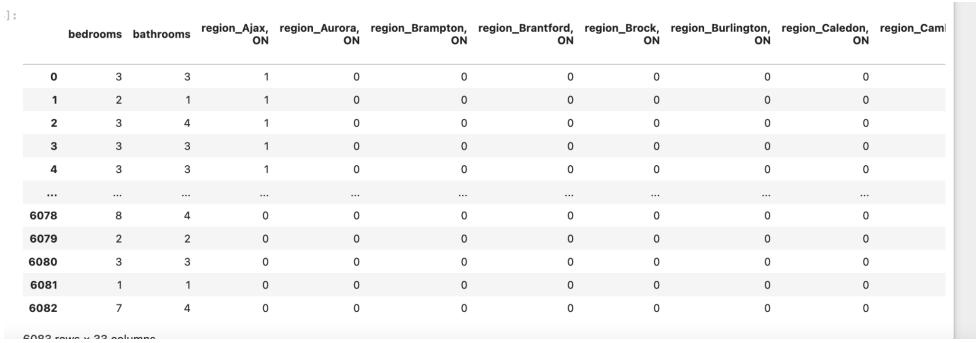
We ensured to handle empty records (if any) by deleting any records with NULL entries from the dataset.  and made sure to reset the index to match the number of records. After preprocessing

the data we had to reset the index to match the number of records. It was important to preprocess data as it helps remove outliers from the dataset and can improve the accuracy of the model.

# Methodology

The machine learning algorithm will be a Linear regression prediction algorithm. Since the result of the predictions are not classifications but could be in the real numbers (Within context range). A regression algorithm will be used to best represent the given data.

We used python as the language due to its scalability and machine learning libraries/software packages. For instance, one of the Python packages we used in data preprocessing, as mentioned above, is called Pandas.



| | bedrooms | bathrooms | region_Ajax, ON | region_Aurora, ON | region_Brampton, ON | region_Brantford, ON | region_Brock, ON | region_Burlington, ON | region_Caledon, ON | region_Cam |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 1 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 2 | 3 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 4 | 3 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 6078 | 8 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 6079 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 6080 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 6081 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 6082 | 7 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

**Fig. 2** - Splitting of regions

Moreover, we used another Python library known as Scikit-learn (Sklearn), which is a machine learning python software package, which was used to import the LinearRegression model and train the dataset. We knew we needed to be able to make a prediction based on some parameters and this is where we came across the linear regression function from Scikit-learn. According to the documentation, “LinearRegression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation” (Sklearn Documentation, 2023). This is exactly the model we needed to train our dataset accurately and make the prediction. Also, we imported another Scikit-learn library known as StandardScaler. This was used in preprocessing of the data to reduce the unit of variance of our model.

```
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
x_scaled = scaler.fit_transform(x_train)
reg = LinearRegression()
reg.fit(x_scaled, y_train)
```

**Fig. 3** - Code snippet calling linear regression model on training data

In training our model we had to make a split for the training and testing. For our model we were able to split the data with 70% of the data for training the linear regression model and the rest 30% for testing and validation. This means the large majority of data would be used to train the model and we ran the test on a smaller portion of new unseen data for the model. For our project we were able to use the

# Results

After running the training sets, we produced the following graph of comparing the actual price of the houses compared to the predicted prices from the model (Refer to figure 4 below) . The data is set in millions, we placed a line of best fit of our data (line in red) and the Green line is the ideal between actual and predicted for reference.

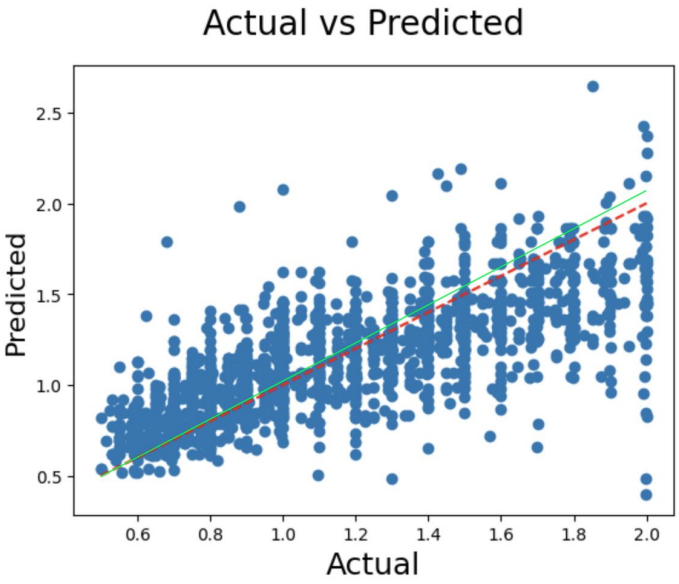



**Fig. 3** - Graph predicted price vs actual (price in millions)

The R- squared score (See figure 4 below) was approx 0.556. Meaning the model has captured about half the variability of the dependent variables around the regression line. Since the regression line is straight. Overfitting the  training data when it varies is not possible, So with a R score of half the model is deemed satisfactory.

```
R-squared score (Linear Regression): 0.5559247908683462
```

**Fig. 4** - R score value from linear regression model

Since we used a linear regression model, we can measure the performance and accuracy of the model using the Mean Absolute Error. We imported the mean_absolute_error function from the scikit-learn library. For our model the result shows the absolute average difference between the actual cost of the home vs predicted. A lower score indicates an accurate model with predictive power similar to ours. Thus, we can say the model is working for our purposes and it isn't overfitting or underfitting the data.

```
Mean absolute error for trainning: 0.19696181296778756
Mean absolute error for test: 0.20278668858507143
```

**Fig. 5** - Mean squared value from linear regression model

| | Actual Price | Predicted Price | Difference |
|---|---|---|---|
| 5591 | 0.929000 | 1.030319 | -0.101319 |
| 64 | 0.799998 | 1.016159 | -0.216161 |
| 87 | 1.199999 | 1.016159 | 0.183840 |
| 1024 | 1.295990 | 0.851120 | 0.444870 |
| 1970 | 0.999900 | 1.110885 | -0.110985 |
| ... | ... | ... | ... |
| 5994 | 0.599000 | 1.141403 | -0.542403 |
| 3833 | 1.179000 | 1.317428 | -0.138428 |
| 4916 | 0.849900 | 0.857956 | -0.008056 |
| 4131 | 0.549800 | 0.726608 | -0.176808 |
| 4307 | 0.699900 | 0.726608 | -0.026708 |

[1825 rows x 3 columns]

**Fig. 6** - Actual price vs predicted price tabled form

# Conclusion

Knowing the future home prices can help greatly with planning the Canadian housing industry. Since there are many more factors that will affect housing prices that were not taken into account, this model is not to be used for commercial purposes. However, it serves as a good example how a simple implementation of machine learning with sound data preprocessing can make accurate predictions on Canadian future homes.

For further improvements we could try to make our model more accurate and include a bigger set of data. For example, we limited the dataset to houses between 500k and 2 million due to the limited parameters of the dataset not being able to accurately predict above 2 million dollar properties. Hence, to combat this we could include more parameters in our dataset such as square footage, price-per square foot, nearby amenities, lake access etc. These parameters would be able to more accurately predict the price of more expensive homes and reduce the variance of the model.

# References

Singh, M. (2022, May 13). *Greater Toronto Area Property Prices*. Kaggle. https://www.kaggle.com/datasets/mangaljitsingh/torontoproperties?resource=download

*Sklearn.linear_model.linearregression*. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

*Sklearn.preprocessing.StandardScaler*. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html