# Predicting Canadian Housing Prices Using Machine Learning

Oreoluwa Collins
*Faculty of Engineering*
*University of Western Ontario*
London, Ontario
ocollin4@uwo.ca

Olivia Howard
*Faculty of Engineering*
*University of Western Ontario*
London, Ontario
ohoward2@uwo.ca

Tara Laird
*Faculty of Engineering*
*University of Western Ontario*
London, Ontario
tlaird4@uwo.ca

Parsa Moeini
*Faculty of Engineering*
*University of Western Ontario*
London, Ontario
pmoeini@uwo.ca

Laura Tidy
*Faculty of Engineering*
*University of Western Ontario*
London, Ontario
ltidy@uwo.ca

***Abstract* - The Canadian housing market is a major issue that is looming over the heads of young people today. With the rapid increase of housing prices, many young people in Canada are wondering if they will be able to afford a house one day. Using data about the Canadian housing market from the year 2021, this project creates regression models and a neural network to predict housing prices. To find the best fit for the data both linear and polynomial regression models are used. The models are given parameters such as income, city, desired number of bedrooms, and desired number of bathrooms. The models were created to help young adults who are looking to purchase homes in Canada. However, due to the chaotic nature of the Canadian housing market, no exact correlation could be found between the prices of homes and their attributes.**

***Keywords* - Housing, Prices, Prediction, Linear Regression, Polynomial Regression, Neural Network**

## I. INTRODUCTION

In the last few years, Canadians have seen a surge in housing prices. In April of 2023, the Canadian Nation Housing Council found that the amount of affordable housing was decreasing and the country desperately needed more. [1] The need for more affordable housing is echoed when observing the Greater Toronto Area specifically. The GTA is going through an increase of immigrants and a growing population leading to an increase in housing needs [2]. Because of this supply and demand situation, Canadians have seen housing prices climbing.

Ontario's government has heard these demands and have announced new action to build more homes. In February 2024, "the government announced a $123 million investment through the Affordable Housing Innovation Fund, which will build more than 5,000 affordable homes" [3]. Considering the time it will take to build these homes it is interesting to see how much these prices will actually cost. Overall, the federal government has announced they plan to spend billions to get more affordable housing to the country. Some government officials have spoken out against this plan as the type of housing proposed is not possible for certain municipalities [4].

Whether or not the government follows through on this plan, new homeowners are interested in the possible prices of these homes. Hopefully, prices will go down but based on the trends of the last few years it's unlikely. Regression models and neural networks can analyze these trends to learn more about housing prices in Canada. This project aims to specifically use these trends to give some insight on housing prices to new Canadian homeowners who are currently stressed with growing prices.

## II. RELATED WORK

The housing market and housing prices have been the topic of a number of researches across the world in an attempt to explain and predict which factors are most significant in influencing housing prices, including internal direct factors and external indirect factors [5].

Related researches have focused on housing prices in a range of different locations and scales such as, on a large scale studying housing prices across 31 provinces and cities in China [5], or on a more specific scale such as housing prices in Beijing [6]. The focus of our project, however, is specifically the Canadian housing market on a large scale, examining housing prices across 45 of Canada's most populous cities.

The House Price Index (HPI), a weighted repeated sales index, has been commonly used across multiple countries' housing markets to predict future changes in housing prices, however, this index does not take into consideration other factors that influence housing prices such as location, population, and area [6]. With recent advancements in machine learning technology, new techniques are being used to predict future housing prices that include the influence of other factors not considered by the HPI. These new techniques include models and tools such as multivariate linear regression models [5], and the Random Forest ensemble models algorithm [6] . Our project aims to predict housing prices using a machine learning regression model and a neural network, while taking into consideration a variety of factors that influence housing prices including population and median family income.

## III. DATA

The dataset used in this project is taken from Kaggle [7]. Kaggle is an online platform commonly used by data and machine learning scientists where they can collaborate with others and find and publish datasets. The dataset lists Canadian house prices and information from the top 45 populated cities in Canada. The information includes listings with addresses including city and province, number of bedrooms and bathrooms, the city's population, longitude and latitude, and the median household income for that city based on the 2021 Canadian census.

For this project, not all information was valuable and was not included. Data that is not in numerical format is not appropriate for the models used in this project therefore City, Address, and Province was taken out of the dataset. The Median Family Income and Population can be used to determine the location of the listing as these differ from each city. Originally the project hoped to use the Longitude and Latitude information but overall it caused noisy results and therefore was excluded from the dataset. This was a good choice as a comment from the Kaggle dataset page found that the longitude values for Nova Scotia were incorrect.

The original dataset included 35, 768 unique rows. After preprocessing, discussed below, the project uses 33, 298 unique rows. The project shuffles the dataset using the train_test_split function from SKLearn.model_selection to create an 80% training set and 20% testing set.

## IV. DATA ANALYSIS

Before preprocessing the data, it was best to understand how the variables correlated to the house's price. The final dataset included five attributes; Price, Number_Beds, Number_Baths, Median_Family_Income, and Population. A variable dictionary can be found below in Table 1. This was also helpful in visualizing outliers in the data that would need to be removed later.

The last four attributes were looked at separately. Each graph plotted the attribute versus Price. The graph also plotted the coefficient of determination ($R^2$) to understand the correlation between the attribute and Price.

$$R^2 = 1 - RSS/TSS \quad (1)$$

The following four figures are the graphs as discussed. It was found that there was no strong correlation between the price of a house and any of the four input variables. This is likely due to the range of housing available in Canada.

TABLE 1
Data Dictionary

| **Attribute Name** | **Data Type** | **Description** |
|---|---|---|
| Price | float | Cost of house |
| Number_Beds | float | Number of bedrooms in house |
| Number_Baths | float | Number of bathrooms in house |
| Median_Family_Income | float | Median family income of city house located in |
| Population | float | Population of city house located in |

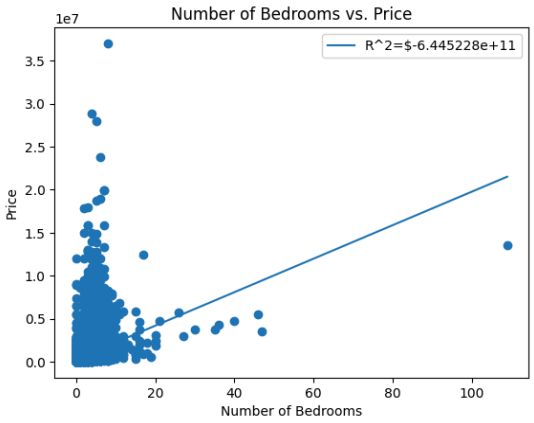


Fig 1. Number of Bedrooms vs. Price

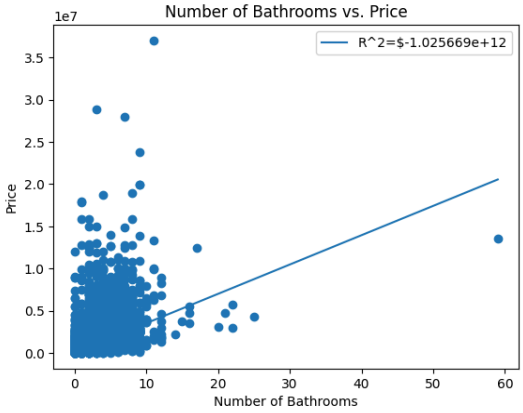


Fig 2. Number of Bathrooms vs. Price

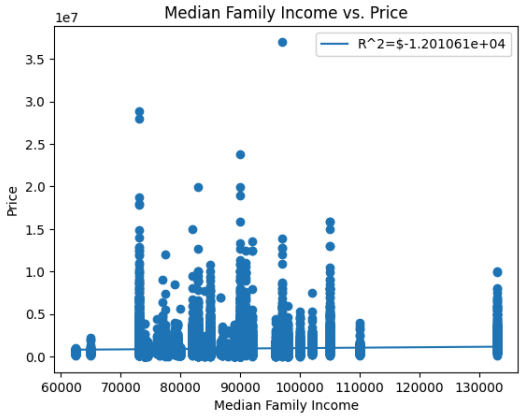


Fig 3. Median Family Income vs. Price

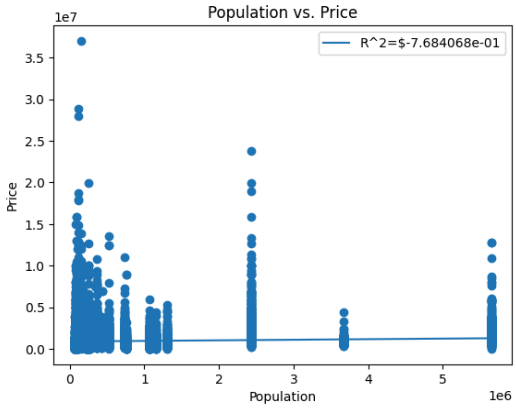


Fig 4. Population vs. Price

## V. DATA PREPROCESSING

### *A. Missing Values*

The dataset was iterated through and any rows missing values for price, number of beds, number of bathrooms, population, or median family income were set for removal. No rows met this criteria so no rows were removed in this step.

### *B. Negative Values*

The dataset was iterated through and any rows with negative values for price, number of beds, number of bathrooms, population, or median family income were set for removal. No rows met this criteria so no rows were removed in this step.

### *C. Outliers*

Upon observation of the data, severe outliers in price (either less than Q1 - 1.5*IQR or greater than

Q3 + 1.5*IQR) were found, so those rows were deleted (2490 rows total).

## VI. METHODS

### A. Research Objective

The research objective was to create a model that can predict housing prices reliably. Our standard for reliable prediction was an $R^2$ value greater than 0.3 between the test set and model predictions.

### B. Linear Regression

The first model made was a linear regression model with no regularization parameter, with weights calculated using the following formula:

$$w = (((X^T)*X)^{-1})*(X^T)*y. \quad (2)$$

The final linear equation was the following:

$$Y = 49326.18917802287 + 2.0953919556080556 * \text{Median\_Family\_Income} + 0.052226317962013644 * \text{Population} + 153942.54571027576 * \text{Number\_Baths} + 35851.970205297774 * \text{Number\_Beds} \quad (3)$$

### C. Polynomial Regression

The second model was very similar to the first, with the addition of two new variables: Number_Beds^2 and Number_Baths^2. This model was also made with no regularization parameters, with weights calculated using the same formula as used in the previous model:

$$w = (((X^T)*X)^{-1})*(X^T)*y. \quad (4)$$

The final linear equation was the following:

$$Y = 61893.85877840128 + 2.11606624852552 * \text{Median\_Family\_Income} + 0.052158285961444784 * \text{Population} + 3194.2256040243665 * \text{Number\_Baths}^2 + 135452.6277665263 * \text{Number\_Baths} + -454.15092427510535 * \text{Number\_Beds}^2 + 39954.68215223448 * \text{Number\_Beds} \quad (5)$$

### D. Neural Network

The third and final model used was a sequential neural network. The first layer was a four-node input layer for the four input variables (Number_Beds, Number_Baths, Median_Family_Income, Population). The next layer was eight Dense nodes with ReLU activation, followed by four Dense nodes with ReLU activation, and finally with one Dense node with no activation function predicting the price.

## VII. EXPERIMENTAL RESULTS

None of the three models met the previously mentioned criteria of having an $R^2$ value greater than 0.3 between the test set and model predictions. All three had a nearly identical testing root mean square error. Both regression models had a nearly identical training root mean square error. The neural network had the lowest $R^2$ value and highest training root mean square error, implying it is the least accurate of the three models. The polynomial regression model had the best $R^2$ value by a small margin. None of the three models were a good fit for the data.

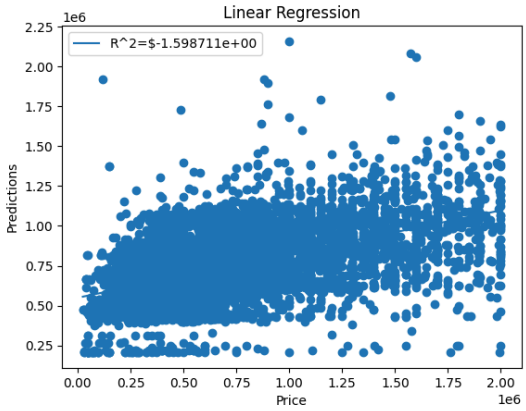


Fig 5. Price vs. Linear Regression Model Predictions

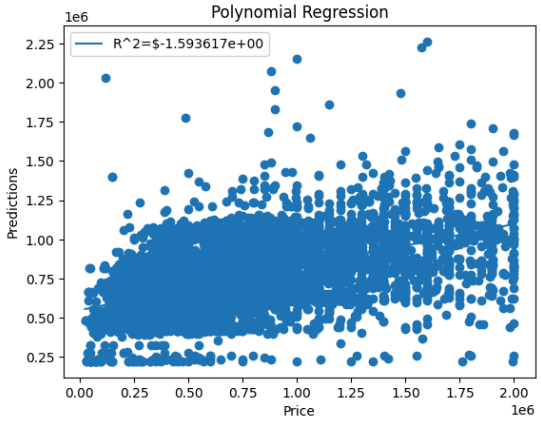

Fig 6. Price vs. Polynomial Regression Model Predictions

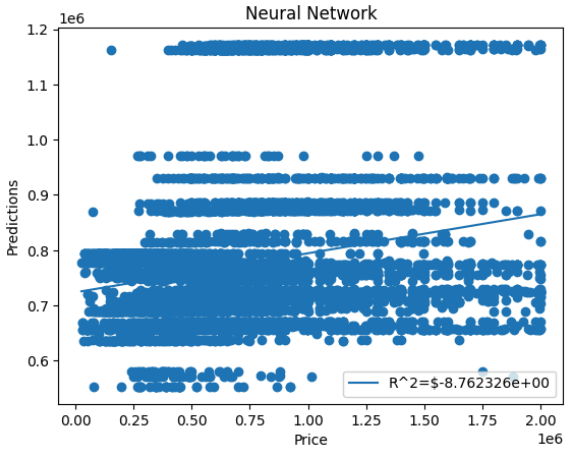


Fig 7. Price vs. Neural Network Predictions

TABLE 2
Model Evaluations

| **Model** | **Training Root Mean Square Error** | **Testing Root Mean Square Error** | **$R^2$ of Test Set and Model Predictions** |
|---|---|---|---|
| Linear Regression | $3.64 \times 10^5$ | 3.58 x $10^5$ | -1.60 |
| Polynomial Regression | $3.64 \times 10^5$ | 3.58 x $10^5$ | -1.59 |
| Neural Network | $4.45 \times 10^5$ | 3.58 x $10^5$ | -8.76 |

## VIII. CONCLUSIONS

In conclusion, all three models produced in this project were a poor fit for the data. We speculate that noise in the data set made it very difficult to identify consistent patterns between the price of a house and the four input variables (number of bedrooms, number of bathrooms, median family income, and population).

Going forward this project could continue to be worked on with new data. It is possible with different or having more input variables a correlation could be found to Canadian housing prices. Having data across multiple years would also be helpful in finding trends in the market. With more time and more data hopefully, a model can be produced that can assist with the prediction of housing prices.

Based on this project and the results it seems the Canadian housing market is extremely inconsistent. Houses across the nation have varying prices with no correlation to size or location. While the government strives for affordable housing [1] this research implies that new buyers would also benefit from consistent pricing.

## REFERENCES

[1] Boutilier, C. (2024, January 12). *Housing policy news: Top stories of 2023*. Canadian Centre for Housing Rights. https://housingrightscanada.com/housing-policy-news-top-stories-of-2023/

[2] Bawuah, Peter. (2024). *CAN ONTARIO'S HOUSING CRISIS BE FIXED? UNDERSTANDING THE ORIGINS AND PROPOSED SOLUTIONS TO AN ONGOING CRISIS IN HOUSING AFFORDABILITY*. Major Papers. https://scholar.uwindsor.ca/major-papers/292

[3] Canada, D. of F. (2024, February 27). *Government announces new action to build more than 5,000 affordable homes and strengthen competition to low...* Canada.ca. https://www.canada.ca/en/department-finance/news/2024/02/government-announces-new-action-to-build-more-than-5000-affordable-homes-and-strengthen-competition-to-lower-prices-for-canadians.html

[4] Le Couteur, M. (2024, April 4). *"a significant overreach": Canada housing plan draws provincial pushback*. CTVNews. https://www.ctvnews.ca/politics/a-significant-overreach-canada-housing-plan-draws-provincial-pushback-1.6833275

[5] Jiang Y. & Qiu L. (2022, February 3). *Empirical study on the influencing factors of housing price ——based on cross-section data of 31 provinces and cities in China*. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050922001922

[6] Truong, Q. et al., (2020, July 27). *Housing price prediction via Improved Machine Learning Techniques*. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050920316318

[7] Larcher, J. (2023, October 29). *Canadian house prices for top cities*. Kaggle. https://www.kaggle.com/datasets/jeremylarcher/canadian-house-prices-for-top-cities