An Approach to Accurately Predicting House Prices

Bryan Meriel  
*Humber College*  
*Group 2   
Toronto, Ontario*  
N01496545

Mariana Betancourt  
*Humber College*  
*Group 2  
Toronto, Ontario*  
N01498484

Sandra Maldonado  
*Humber College*  
*Group 2  
Toronto, Ontario*  
N01468706Steven Kinnunen  
*Humber College*  
*Group 2  
Toronto, Ontario*  
N01501178

*Abstract*— *The real estate business has several variables to analyse. One of the most relevant variables and one of the most desirable to predict is the price. Predicting housing prices could be a complex problem due to the number of related variables that can affect it. However, by using models such as K-NN and linear regression, we can find the best price prediction. Having an accurate model to predict housing prices could help real estate agents, constructors, sellers, and buyers to make better decisions. The results suggest that K-NN with the correct value of K will outperform multiple linear regression and regression trees when predicting housing prices.*

Keywords—K-NN, Housing price, Multiple Linear Regression, Regression Trees, MAPE

# Introduction

Home sales in Canada have hit record breaking prices over the past few years. When the COVID-19 pandemic hit, this caused an absolute shutdown across the country. Many businesses were forced to close while other businesses switched to remote working, driving many individuals to purchase new homes [7]. A person working from home is likely to invest in a home which has ample space for an office, or if they intend to run a business from their home, maybe a couple of floors. This new demand for homes quickly increased real estate pricing, but what we wish to understand is, which variables are the most important in predicting the price of a house [7]. Following the discovery of these variables, our goal is to predict a house's price.

Using the housing dataset provided, we will be analysing over 21,000 records and running various algorithms that will help us to ultimately predict the price of a house [5]. It is of the utmost importance for us to ensure our pricing predictions are as accurate as possible, given the fact that we have 21 variables to work with. Therefore, we will be using the Multiple Linear Regression Model, the K-NN Classification Model, and the Decision Tree Regression Model in order to achieve our goal. A later comparison will help determine which algorithm predicts a house's price best, based on specifications. Among these specifications are the number of bedrooms, bathrooms, floors, square footage, whether it has a waterfront or a general view, and so on.

# Literature Review

Qingqi Zhang explores the benefits and drawbacks of using multiple linear regression to predict Boston housing prices [15]. Ordinarily, licensed real estate professionals estimate price based on the specific location, the surrounding area as well as the kind of facilities that are included with the property. However, Zhang notes that personal bias may impact this estimation process and consequently calls for a more impartial method of price estimation. He notes that historically multiple linear regression has performed better at statistical inference rather than prediction. Zhang adopts the Spearman correlation coefficient since it does not rely on the assumption of linearity that Pearson Correlation does. Ultimately, his investigation resulted in a model of limited accuracy and called for investigation into other models.

Montero et al use Madrid housing data to predict housing prices [8]. They explain that previous analyses have used linear methods but that housing prices have many factors that do not conform to a linear model. They use spatial models such as PSSD, GAM and SDM to estimate housing prices. They conclude that accounting for spatial factors related to housing prices results in better predictive accuracy.

Park and Bae analysed Virginia housing data to predict housing prices [9]. They also note traditional “hedonic” regression models (models that understand price as a series of characteristics that cannot be separated from the good itself) tend to fail to capture the non-linear elements present in the data [4]. Whereas most analyses appear to focus on the specific price, they try to classify whether the final price is less than or greater than the list price [8]. To that end, they use Naive Bayes, RIPPER and AdaBoost algorithms. The RIPPER algorithm achieved the lowest overall error rate.

# algorithms

To facilitate our discovery process, we will be using 3 different supervised machine learning algorithms for the purpose of classifying our data and accurately predicting housing prices. As a result of this, we will be able to compare the outcomes from all three algorithms and choose the one which has the highest degree of accuracy.

Considering that we have several variables in our dataset, the Multilinear Regression algorithm is the most appropriate algorithm to begin with. The concept of multiple linear regression involves the estimation of the relationship between multiple independent variables and one dependent variable [1]. For the purpose of this example, the dependent variable will be price and the independent variables will include things like bedrooms, bathrooms, floors, square footage, view, waterfront, and many others. With the help of this algorithm, we will be able to determine how strong the relationship is between these independent variables and the price of a home. We will also be able to determine the price of a home based on the value of these independent variables.

We will then use K-Nearest Neighbors as our second algorithm. Based on proximity, the K-nearest neighbours’ algorithm, or K-NN, assigns a classification or makes a prediction about a data point based on its similarity to the closest neighbour. Even though it can be applied for both regression and classification problems, it is more commonly used for classification problems. Hence, it is based on the assumption that similar points are nearby one another [14]. Using the K-NN algorithm, the probability of the test data belonging to the classes of the ‘K’ training data is calculated, and the class with the highest probability is selected out of the set [2]. It is pertinent to note that the 'K' values for our example will be K=5 and K=10. This K-NN model gives us the opportunity to determine which of the two 'K' values gives the highest prediction accuracy. As soon as this 'K' is determined to be the most accurate predictor, we will use it to further develop our house pricing prediction model.

As our final algorithm, we chose to use Decision Tree Regression. Both classification and regression tasks can be performed using decision trees. The structure is hierarchical starting with the root node, outgoing branches feed into internal nodes, also known as decision nodes. By evaluating the available features, both node types form homogeneous subsets, which are referred to as leaf nodes [13]. All possible outcomes in the dataset are represented by these leaf nodes. In this method, the optimal split points within a tree are identified by conducting a search that employs a divide and conquer strategy. A top-down, recursive process of splitting records is repeated until every record has been classified under a specific class label [13].

# methods

After analysing and cleaning the dataset, the most relevant variables from the dataset to predict the house price were chosen to use in the linear regression, K-NN regression and decision tree regression models. The most important variables were chosen by calculating feature importance, in which the top 6 variables [10] were:

* SQFT Living – total area of living in square feet
* Grade - grade given by the qualifications house system
* Waterfront - a binary variable for whether the house is on the waterfront or not
* Bedrooms - the number of bedrooms in the house
* Bathrooms - the number of bathrooms in the house
* SQFT Lot - total area of the property in square feet.

Using the statistical technique of multiple linear regression - the featured variables above were used to estimate its relationship with the dependent variable, price. Given the results of the model, the measurements of error were assessed and provided a baseline used to compare the performance of other predictive models. For this project, the MAPE of the multiple linear regression model was primarily used to compare the results.

Working with the chosen variables the K-NN regression model was applied by using K=5 and K=10. After comparing the model with each k value and measuring the error, it is possible to assess the MAPE of each K-NN model and compare it to multiple linear regression and the regression tree.

Optimal minimum samples split, and minimum purity decrease optimal values were found by testing several values for each in two steps. After these two iterations, it was found that a tree with a depth of 7, minimum samples split of 14 and minimum impurity decrease of 0 was ideal. The MAPE of the resulting tree was calculated and compared to both multiple linear regression and K-NN.

# results

Chart, histogram

Description automatically generated

**FIGURE 1: RESIDUALS OF MULTIPLE REGRESSION**

The multiple linear regression model uses standardized values of price. Figure 1 shows the distribution of the residuals around 0, with values distributed from around -0.15 to 0.2. As such, the most meaningful way to assess the model is the mean absolute percentage error (MAPE), which takes the absolute value of the percent error between the predicted and actual values. The linear regression model produces the second highest MAPE at 43.22%.

Chart, histogram

Description automatically generated

**FIGURE 2: RESIDUALLS OF K-NN WITH K = 10**

Two K-NN models were used with a k of 5 and a k of 10. Figure 2 shows the residuals of the K-NN algorithm with a K of 10. The values are concentrated around 0, with values from -0.15 to 0.2. The former model produced a value of 43.28%, which was the highest MAPE of all models. However, the latter model produced the lowest MAPE at 42.53%.

Chart, histogram

Description automatically generated

**FIGURE 3: RESIDUALS OF REGRESSION TREE**

Two K-NN models were used with a k of 5 and a k of 10. Figure 2 shows the residuals of the K-NN algorithm with a K of 10. The values are concentrated around 0, with values from -0.15 to 0.2.The former model produced a value of 43.28%, which was the highest MAPE of all models. However, the latter model produced the lowest MAPE at 42.53%.

# discussion

Overall, the K-NN model with a k of 10 produced the most accurate model. This performance may be explained by how K-NN makes no assumptions about the distribution of the data [11]. Conversely, multiple linear regression assumes linearity and independence of observations [11]. As Montero et al explained, linear models do not conform well to the spatial dimensions of housing price predictions. This may account for the model’s performance relative to K-NN. Conversely, while decision tree regression is quite robust and does not rely on assumptions about the data, it tends to overfit to the training data [3]. This may account for why that model was outperformed by the K-NN model with a K of 10.

In the future, it may be worthwhile to implement multiple linear regression using the Spearman correlation coefficient as [15] did to produce a more accurate model. Moreover, focusing on algorithms such as RIPPER or AdaBoost and approach price predictions as a classification problem (such as different price ranges). Finally, [9] approached price predictions by forgoing linear models altogether and using models that are specifically designed to deal with non-linear data such as PSSD, GAM and SDM. Any of the aforementioned approaches may merit consideration in future analyses of price prediction.

# conclusion

Ultimately, it appears that amongst multiple linear regression, K-NN and regression trees that K-NN with the correct K value will produce the best results for predicting house price. Linear regression’s assumption of linearity is likely violated as the literature notes that housing prices are not the result of linear relationships. Moreover, while regression trees do not suffer the same limitations as linear regression, they may suffer in performance due to overfitting the training data. K-NN also does not make assumptions about the data and in this case had the best performance.

Despite the superior performance of K-NN relative to linear regression and regression trees, existing literature suggests there are superior methods to predict price. When price is considered as categorical, Naive Bayes, AdaBoost and RIPPER may offer viable alternatives that account for the limitations of more traditional methods. Conversely, PSSD, GAM and SDM could allow price prediction to remain continuous and offer a means to account for the non-linear relationships present in the housing data.

# references

1. Bevans, R. (2022, November 15). Multiple linear regression: A quick guide (examples). Scribbr. Retrieved November 27, 2022, from https://www.scribbr.com/statistics/multiple-linear-regression/
2. Christopher, A. (2021, February 3). K-Nearest Neighbor. Medium. Retrieved November 27, 2022, from https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4
3. Decision tree advantages and disadvantages: Decision Tree Regressor. EDUCBA. (2021, September 23). Retrieved November 25, 2022, from https://www.educba.com/decision-tree-advantages-and-disadvantages/
4. Hedonic regression method. Corporate Finance Institute. (2022, November 4). Retrieved November 25, 2022, from <https://corporatefinanceinstitute.com/resources/data-science/hedonic-regression-method/>
5. House sales in King County, USA. Kaggle. (2016, August 25). Retrieved November 24, 2022, from https://www.kaggle.com/datasets/harlfoxem/housesalesprediction
6. Kohli, S., Godwin, G.T., Urolagin, S. (2021). Sales Prediction Using Linear and K-NN Regression. In: Patnaik, S., Yang, XS., Sethi, I. (eds) Advances in Machine Learning and Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-5243-4\_29
7. Lord, R. (2020, December 7). 'we haven't seen anything like this before': Pandemic boosts canadian real estate. Global News. Retrieved November 27,2022, from https://globalnews.ca/news/7502577/coronavirus-pandemic-canadian-real-estate/
8. Montero, J.-M., Mínguez, R., & Fernández-Avilés, G. (2017). Housing price prediction: Parametric versus semi-parametric spatial hedonic models. Journal of Geographical Systems, 20(1), 27–55. https://doi.org/10.1007/s10109-017-0257-y
9. Park, B., & Bae, J. K. (2015). (working paper). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, (Vol. 42, Ser. ISSN 0957-4174, pp. 2928–2934).
10. Spagnola, J. (2020, December 15). Determining significant features in a house sale. Medium. Retrieved November 26, 2022, from https://medium.com/swlh/determining-significant-features-in-a-house-sale-b501cfd01077
11. Vadan, F. (2020, April 23). Problems with multiple linear regression, in R. Medium. Retrieved November 27, 2022, from https://towardsdatascience.com/problems-with-multiple-linear-regression-in-r-bef5940518b
12. Vatsal. (2022, May 20). K nearest neighbours explained. Medium. Retrieved November 22, 2022, from https://towardsdatascience.com/k-nearest-neighbours-explained-7c49853633b6
13. What is a decision tree? IBM. (n.d.). Retrieved November 27, 2022, from https://www.ibm.com/topics/decision-trees
14. What is the K-nearest neighbors algorithm? IBM. (n.d.). Retrieved November 27, 2022, from https://www.ibm.com/topics/K-NN
15. Zhang, Q. (2021). Housing price prediction based on multiple linear regression. Scientific Programming, 2021, 1–9. https://doi.org/10.1155/2021/7678931