Northeastern University

Khoury College of Computer Science

CS 5002 Discrete Structures

Instructor: Richard Hoshino

Prediction of Vancouver housing price based on Decision Tree Regression

Group 2

Zhou Chen

Xie Xueyi

Lyu Bochen

Due Date: December 14, 2022

Submitted: December 14, 2022

1． Introduction

Project Context:

Overview

From CS5002, we learned that Decision Tree is a very useful tool when it comes to making major decisions. Since we are all interested in the Vancouver real estate market and will probably purchase a home in the future, exploring Vancouver real estate market with decision tree perfectly aligns with our interests. Therefore, we have decided to make a decision tree model to predict the home prices in Vancouver.

Decision tree is a powerful tool that enables us to do classification and regression analysis. It is a commonly used method for supervised learning in machine learning. There are two common types of decision tree models: decision tree classifier and decision tree regressor. The decision tree regression model is often used in predicting variables with continuous values (such as prices), and the classification model is rather used to predict variables with categorical, non-continuous values (such as "yes/no" in disease detection).

Bochen:

Housing is an essential part of our daily life; it keeps us safe and warm at night. However, not everyone can afford one nowadays. Housing affordability has become a worldwide issue. Sadly, few governments are willing to shoulder their responsibility and do something about the increasing unaffordability of houses due to the intertwined relationship between the housing market and economic growth. More and more people cannot afford to buy a property, and for those who can, most of whom will end up having a huge chunk of debt and have no choice but to lower their living standards. So, it’s becoming a bigger and bigger decision for people to buy a house.

Vancouver, being one of the major cities in Canada, has the worst real estate markets. Not only did the price skyrocket, but it also fluctuates so badly that you would lose tens of thousands of dollars overnight. Due to the fluctuating nature of the housing market, predicting the true value of the house is essential.

And as a programmer myself, I aim to provide a way for the buyer to buy the house wisely by estimating the true value of the house. It can not only help me to decide when to buy a house but also help hundreds of buyers to make the decision either.

For now, our project will focus on the Vancouver West area, due to the limitations of our time and our ability in programming, but hopefully, after finishing my study in Northeastern, I can develop a whole algorithm to not only help the buyers to decide when to buy a house but also provide the government with a model for them to tailor their policy towards the housing market, so that everyone eventually can have a roof over their heads without spending every penny of their life savings.

Chen:

Acquiring knowledge to make the world a better place has been my life’s pursuit. As an investigative reporter, I emigrated to Canada with the hope of enjoying the freedom of the press and a safer society. I shortly found the bitter reality that my experience and achievements were not sufficient and were not recognized by the major news media in Canada. When I finally got the opportunity to continue writing in-depth reports, I was astonished to see that the Canadian government never has met any climate goals and that the journalists reporting the protests of the indigenous communities were being arrested by the police.

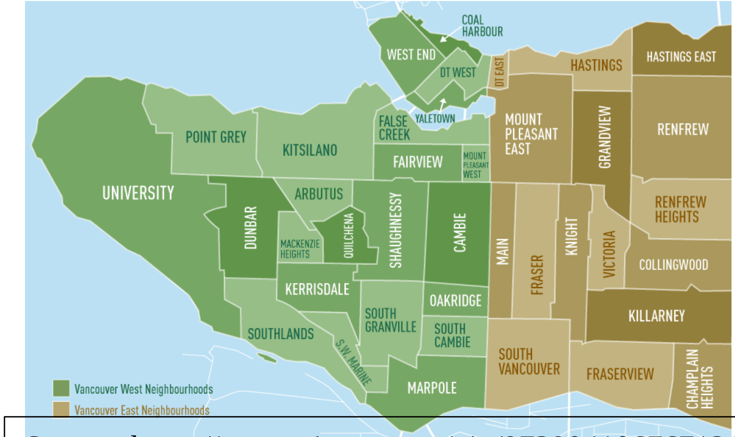
Fortunately, I have determined which kind of person I would love to become in the future. A personal inspiration of mine is programmer Rohana Rezel. While some programmers who I am acquainted with were discussing how to develop a program that would collect information on housing prices to help investors buy homes with the most appreciation potential, on other hand, Rezel, a senior software architect, uses his data analysis skills to find hidden information about house flippers and money launderers, which helps to hamper their activity and aims to potentially solve the city's housing affordability crisis. Rezel's experience has profoundly inspired and bolstered my determination to further learn about computer science to become an expert like him.

Xueyi:

We all live in Vancouver, and it is truly an amazing city. One thing about Vancouver though, is that its housing price is exceptionally high compared to most other Canadian cities. Housing affordability has long been a problem for the coastal city of Vancouver and its neighboring towns and cities. As early as 2015, I remember listening to the radio on my way to a grocery store with some friends, and the interview on the radio was about the housing affordability crisis in Vancouver. Since then, the problem has only deteriorated, despite the implementation of policies such as charging empty home tax on vacant homes or putting restrictions on purchases made by foreign home buyers. With the impact of the COVID-19 pandemic and the upcoming recession, buying a home is getting harder and harder for most people who reside in Vancouver. In our project, we are interested in the house prices in Vancouver, and we decided to use the decision tree model to predict home prices in Vancouver, or more specifically, Vancouver West because our data is for homes in Vancouver West. Policy-wise, there is nothing we could do to intervene with the housing price. But by predicting home prices in Vancouver, we hope to keep ourselves (or anyone who sees and uses our model) aware of the current real estate market and stay informed as much as possible.

Question:

We acquired a dataset that contains information on homes sold in Vancouver. We decided to narrow down our range to Vancouver West, East Vancouver, and Downtown rather than the city of Vancouver or the Metro Vancouver area. So, to conclude our final project with one question, how can we predict the home price of a given property in selected areas in Vancouver based on the distinctive features that a house possesses. To further elaborate on our question, we would like to investigate how we could use decision tree to divide different features of the houses to help us predict the value of the homes in Vancouver's real estate market.



Source: https://www.pinterest.ca/pin/973904106767121357/

2. Analysis

Overall, we decided to obtain the data, clean the data, and load the data on Python for further analysis. To analyze the data, we decided to split our data into training data sets and testing data sets. We made a decision tree with our training data set to predict home prices in Vancouver on our testing data set. Lastly, we validate our model for accuracy by calculating the mean percentage error.

2.1 Preparation

2.2.1 Datasets

We have obtained the latest public data on Vancouver's Westside resale homes from the non-profit organization Open House (https://openhousing.ca/about/).

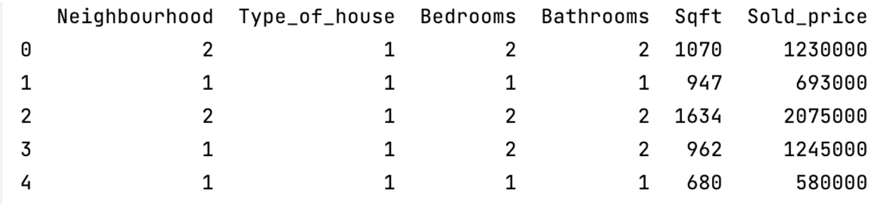
The data was initially parsed off the web with 1000 sets of data stored in a Json file, containing all the information for all types of properties. Due to the nature of our projects, we picked four variables that are most related to the price of a property and cleaned them with libraries provided by python. After the initial data cleaning to remove the redundant data and acquire the data of the variables needed, we cleaned the data a second time. This time, it deleted all the invalid data based on invalid number of bedrooms and bathrooms as well as invalid neighbourhoods. After several cleanups, we converted categorical variables like neighbourhood and type of house to numerical variables as shown in the table below, e.g., Vancouver West Area as 1 and Downtown Area as 2. Finally, we stored the cleaned data in a .csv file, which contains the columns "Neighbourhood, Type of house, Bedrooms, Bathrooms, Sqft, Sold price", with 470 lines of data.

Neighbourhood	Vancouver West (1) or Downtown Area (2)
Type of house	Apartment (1), House (2) or Townhouse (3)
Bedrooms	Number of Bedrooms
Bathrooms	Number of bathrooms
Size	Size of the house in sqft
Sold price	Price as the prediction target in CAD

2.2.2 Data examination

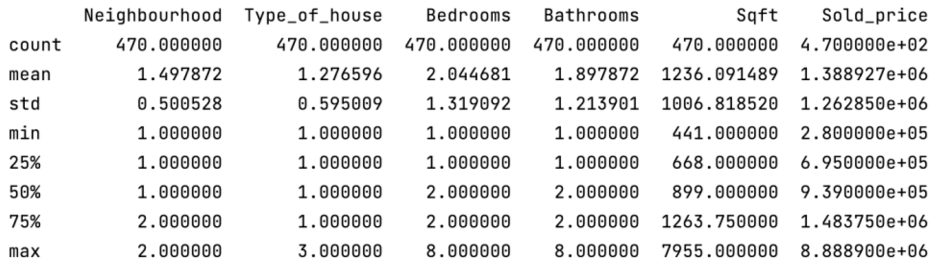
After acquiring our raw data from known real estate websites and cleaning them with regex and pandas library using Python, we analyzed our data first to find out the correction between different variables to provide support to our model.

Below is a table generated by head() function within NDFrame. It gives us an overview of our datasets and how we use Dummy Variables to convert categorical variables to numerical variables to fit our machine-learning model.



```
   Neighbourhood  Type_of_house  Bedrooms  Bathrooms  Sqft  Sold_price
0              2              1         2          2  1070     1230000
1              1              1         1          1   947      693000
2              2              1         2          2  1634     2075000
3              1              1         2          2   962     1245000
4              1              1         1          1   680      580000
```

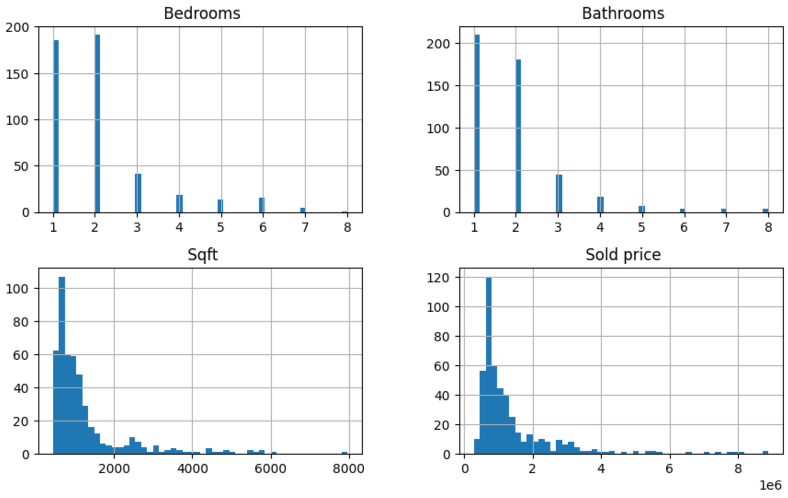
We also called the built-in function describe() to give us a descriptive statistic about our datasets. It summarizes our central tendency, dispersion and shape of our dataset's distribution. We can again verify that our data cleaning is done appropriately, as the mean, minimum and maximum value is within the proper range.



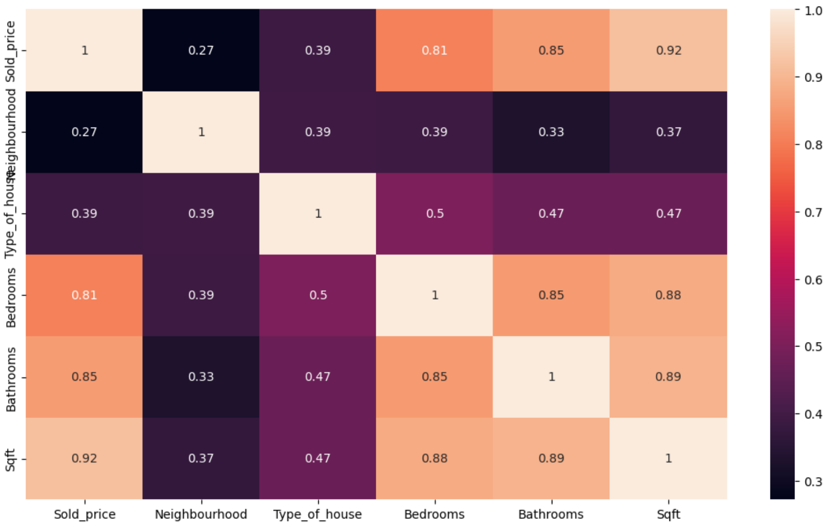
```
       Neighbourhood  Type_of_house    Bedrooms   Bathrooms         Sqft    Sold_price
count     470.000000     470.000000  470.000000  470.000000   470.000000  4.700000e+02
mean        1.497872       1.276596    2.044681    1.897872  1236.091489  1.388927e+06
std         0.500528       0.595009    1.319092    1.213901  1006.818520  1.262850e+06
min         1.000000       1.000000    1.000000    1.000000   441.000000  2.800000e+05
25%         1.000000       1.000000    1.000000    1.000000   668.000000  6.950000e+05
50%         1.000000       1.000000    2.000000    2.000000   899.000000  9.390000e+05
75%         2.000000       1.000000    2.000000    2.000000  1263.750000  1.483750e+06
max         2.000000       3.000000    8.000000    8.000000  7955.000000  8.888900e+06
```

.

As we all know, machine learning feeds on the dataset that we provided and incrementally build a model upon that, so it is not inherently objective. We have to take possible bias into consideration. That’s why we plotted a histogram for our variables. It gives us an intuitive look at the data that we fetched from the internet and how it may shape the model as it is. From the histogram, we can have a vague perspective on what the bias might be and how our model will operate the best when predicting the prices of Vancouver houses. For example, most of our bedrooms’ data are between 1 and 2, it means that when predicting the prices, the model will operate best when the predicted houses have 1 or 2 bedrooms. It might also have a bias towards 1 or 2 bedrooms when predicting the prices.



Finally, we drew a heatmap to visualize the correlation between different variables. We think the most important part is the correlation matrix of predicting factors (e.g., bathrooms, sqft) and sold prices. We can see from the graph that bedrooms, bathrooms and sqft have a strong linear relationship with sold prices, which might explain why sqft are the root nodes when splitting the data within decision tree regression. When we look at the correlation between neighbourhood, type of house and sold price, the correlation coefficient is small. However, that’s counterintuitive, as we know that neighbourhood and type of house plays an important role in property prices as well. This might be caused by machine learning bias as we’ve mentioned earlier, because we do not have enough datasets to eliminate this bias when constructing our model.



2.3 Methodology

First, we used the Python pandas library to load the CSV file content. Specifically, we used the read_csv() function to read our CSV file into data frame. In our search for methods that would generate decision trees, we came across the scikit-learn library, and its DecisionTreeRegressor model meets all our needs for the scope of this project. Therefore, we decided to use the DecisionTreeRegressor from scikit-learn to train our model for home price prediction.

In our decision tree regression model, our prediction target (dependent variable, also called "y" in our code) is the sold price in our CSV file. Independent variables (called "x") used to predict house sold prices include neighborhood, type of house, number of bedrooms, number of bathrooms, and area in square feet (sqft).

Before we trained our model, we decided to split our data into two sets, one for training and one for testing, because data splitting is crucial to avoiding bias. The scikit-learn library has a package called model_selection, and the train_test_split() function allows us to achieve data splitting with ease. We set 20% of all data for testing, and the remaining 80% for training. Both dependent and independent variables were split into training and testing sets. The split is random and is done only once, meaning that we use the same training and testing data sets to train our model each time the program is run.

While building the model, we first define our model, and then use the train set of data to fit in the model. The maximum depth of a tree indicates how deep the tree can be. The deeper the tree is, the more splits it has and the more information it can capture about the data. We set the maximum depth of our tree model to different values from 1 to 100 to test out which maximum depth would give us the best prediction accuracy. We set the parameter random_state to a nonnegative integer to get a consistent random model. We then use the model to predict y (home sold prices) on our testing data set.

To examine the accuracy of our model, we used mean percentage error as an indicator. We took the absolute value of the difference between predicted values and actual values, converted these errors to percentage in relation to actual values, and averaged these percentage values. At last, we visualized our tree by using another function from the scikit-learn library: the plot_tree() function from the tree package.

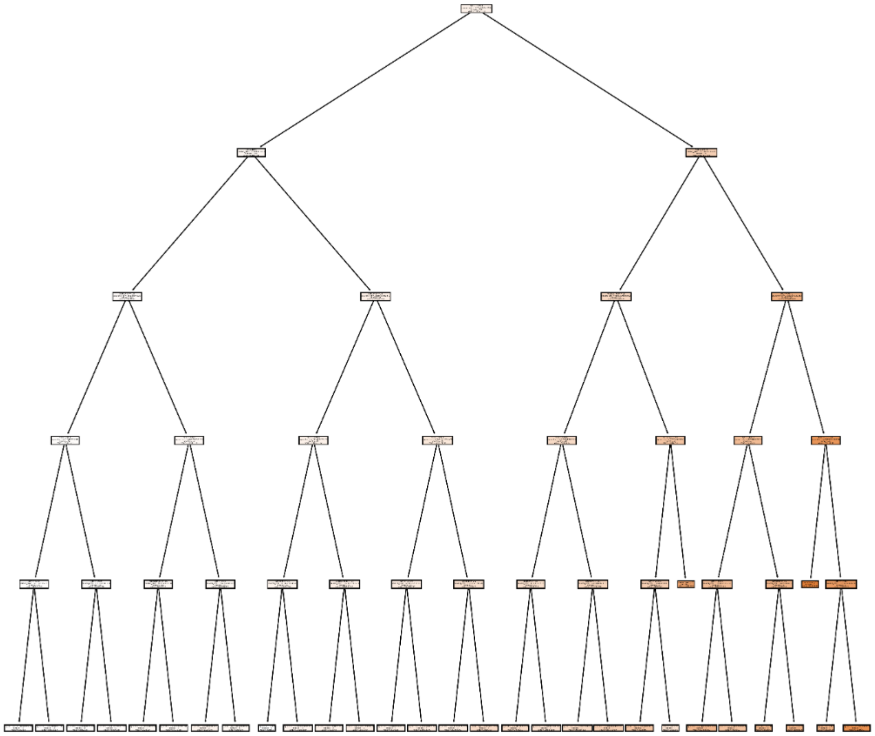
2.4 Results and Discussion

We successfully created smaller subsets by splitting the original datasets in the meantime incrementally developing a decision tree regression model with sklearn – a machine learning library built for python, then predicting the housing price in the Vancouver area with the remaining testing dataset.

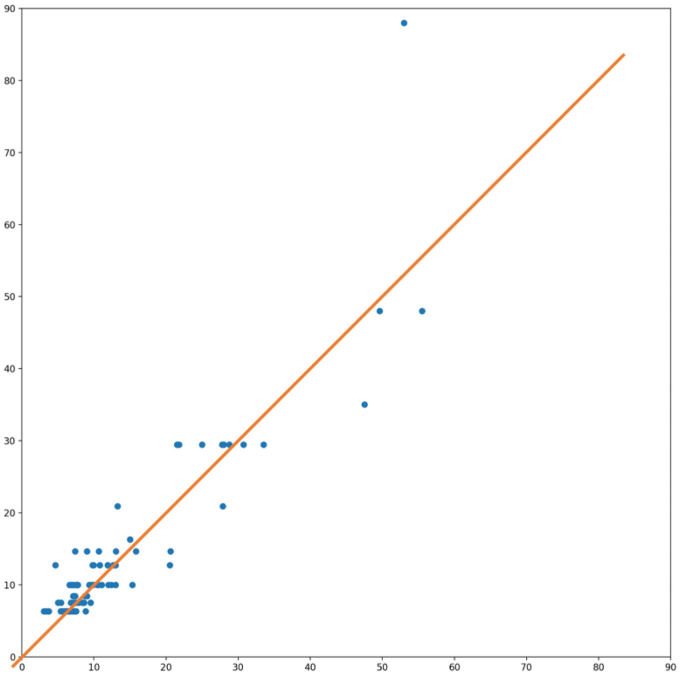
By testing the maximum depth of our tree model to different values from 1 to 100, we find out that when the maximum depth equals 5, we can get a minimum mean percentage error of 15.5551%.

max_depth	2	3	4	5	6	7
Accuracy	77.2357%	81.6043%	82.6461%	84.4449%	83.9333%	83.3470%
max_depth	8	9	10	20	50	100
Accuracy	81.1205%	80.1597%	79.2664%	79.2146%	79.2146%	79.2146%

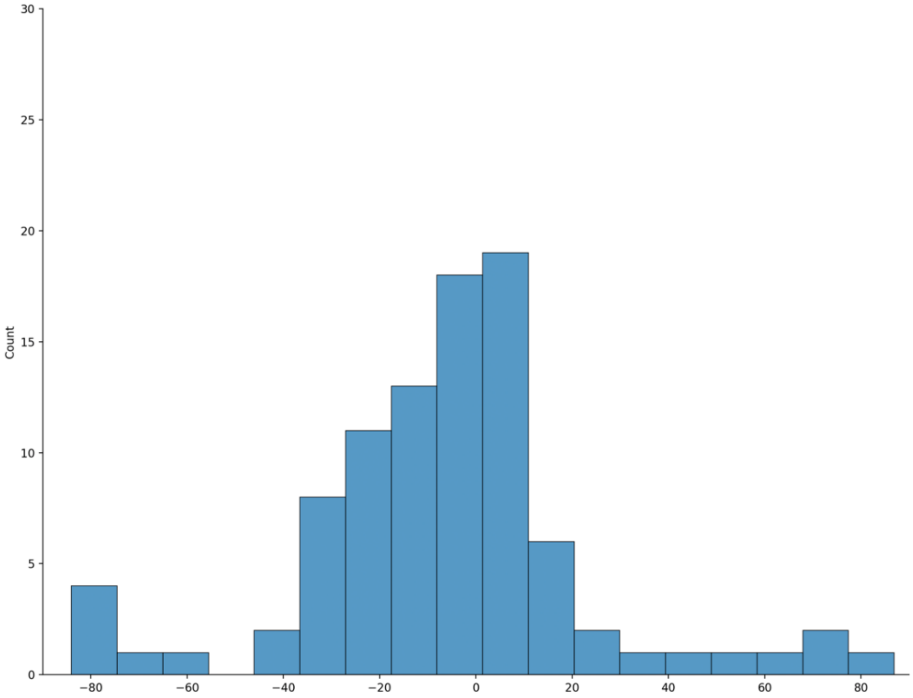
By combining the above conditions, we obtain the optimal decision tree state with a max depth of 5 and minimum mean percentage error of 15.5551%. Finally, we drew the decision tree regression diagram with python to visualize our project.



The following scatter graph shows a relationship between the real price and the predicted price, we can see that most of the dots are around y=x, which means the predicted price is not far from the actual price.



Below is the histogram of the difference between predicted price and the actual price. We can tell that most data reside around 0, is symmetric about the original point. It means that a great majority of predicting prices are accurate with a few exceptions.



3．Conclusion

3.1 Individual Learning Experience

Xueyi: We learned that even though the topic of machine learning sounds intimidating, there are libraries and packages readily available for us to use, and their applications are fairly straightforward. In addition, we taught ourselves how to use the DecisionTreeRegressor model from the scikit-learn library by carefully examining online resources available to us. Identifying reliable resources and learning something new are essential transferable skills in the field of computer science, and they will be carried on to the rest of our degree and career.

Chen: House prices are a very complex issue with many factors that can cause it to change, and it is very exciting that we can find patterns and obtain promising results with a simple model and limited data. We found that in the field of machine learning, Decision Tree is only one of the tools used to analyze the prediction, and there are many other very useful tools that can be used together to bring more complex and accurate results, which are worth exploring further.

Bochen: Though being a small project to utilize everything we've learnt from discrete structures, it really is a learning experience for me. I learnt what decision tree regression is and what different attributes it has from decision tree classification, and how to utilize both models in python. Moreover, we build a prediction model based on machine learning from scratch and it cannot be done without searching for literature and materials on the internet. It taught me that computer science is not about what we've learnt from textbooks, but what we're going to learn by ourselves. This experience greatly improved our ability to search for useful information on the internet and utilize it in the process. I think it will be extremely beneficial to us to learn how to retrieve information and distinguish what we need from what we're provided.

3.2 Conclusion, Limitations, and Future Direction

Overall, we decided to obtain the data, clean the data, and load the data on Python for further analysis. To analyze the data, we decided to split our data into training data sets and testing data sets. We made a decision tree with our training data set to predict home prices in Vancouver on our testing data set. Lastly, we validated our model for accuracy by calculating mean percentage error. To address our question, we successfully developed a decision tree regression model, trained it with 80% of our data, and measured its performance with the remaining 20% of our data. The accuracy of our model is 84.44%, and this was achieved by calculating the mean percentage error of our prediction model first.

In conclusion, we did a decent job considering it is our first time doing a project of such sort involving machine learning techniques. Our biggest restriction was the time frame of the project; it limited our work, and we believe that we could do a more refined job had we had more time. It was also limited by the amount of data we had, we did not have a large sample size and having more data for future modeling would help increase our model accuracy. Our project is an introductory beginner-level project; we produced decent results, but we did overlook some important aspects of the decision tree regression model, such as the algorithm and the code behind it. Future work can involve the use of more independent variables, training with more data points to increase our prediction accuracy and exploring some other models and possibly compare which model can predict home prices in the most accurate manner by validating and evaluating them horizontally.