OBA 455 Final Project

Housing Price Prediction

A house with a large body of water

Description automatically generated with low confidence

March 14, 2022

Daniel Warner

**Introduction**

The real estate market is always changing due to many different economical variables, interest rates, and technological advancements. Thus it is often difficult to accurately realize the value of a home. Inaccurate price listing could lead to you or your client vastly undervaluing the home or listing it at an unreasonable price leading to it sitting on the market for too long. However, through predictive modelI are able to take information from previous listings to predict housing prices using different variables specific to the house we are concerned about.

**Problem Description**

Currently real estate agencies determine the value of a home by doing a competitive market analysis also known as CMA. This factors in property location, condition of the property, and the condition of the housing market in that area. Some common factors that real estate agencies use are:

* Age of the home
* Square footage
* Number of bedrooms and bathrooms
* Whether any recent updates have been made

Although these are the standard metrics to decide the value of a home, it can be difficult for real estate agencies to figure the individual driving factors of a home. For example if home was to gain a bathroom how much would that increase it’s value? What is the price of a home that has a waterfront opposed to one that doesn’t? What do cities that are expensive to live in have that cheaper cities don’t? This knowledge could help consumers and real estate agencies figure out how to increase the value of a property and what changes should be made to do so. In essence, I believe that most people would assume making an addition to a home would increase it’s value, but we’d like to create a model that allows them to see just how much of an effect these different aspects have on the price of a home, and which ones are the most influential.

**Dataset**

*Exploration*

The dataset that I chose was found on Kaggle and it contains 4,600 rows and 18 variables. All of the data comes from the state of Washington. Below is a description of the variables as well as their ranges in our dataset.

Date: The day that the house was listed (May 2014 - July 2014)

Price - The price of the home ($0 - $26,590,000).

Bedrooms - Number of bedrooms (0 - 9).

Bathrooms - Number of bathrooms (0 - 9).

Sqft\_living - Total square footage (370ft - 13,500ft).

Sqft\_lot - Square footage of the lot the house is built on (638ft - 1,070,000ft).

Floors - Number of floors (1 - 3.5).

Waterfront - (0 indicates property had waterfront, 1 indicated it did).

View - 1-4 rating describing the view from the property (1 indicated poor view, 4 indicated great view).

Condition - 1-5 rating describing the condition the home was in (1 indicated poor condition, 5 indicates great condition).

Sqft\_above - Square footage of the house above ground (370ft - 9410ft).

Sqft\_basement - Square footage of the basement. (0ft -4820ft)

Yr\_built - Year that the house was made (1900 - 2014)

Yr\_renovated - Year that the house was renovated ( 0 if not renovated, year renovated if renovated)

Street - Street address of home.

Country - All were from the USA, specifically in Washington state.

*Cleaning/Manipulation*

Given these variables there were some changes I made to ensure their usefulness in our models. First of all I changed the Yr\_renovated variable to a binary variable where 1 represents a home that has been renovated and 0 represents a home that has not been renovated. I did this to ensure homes that had not been renovated didn’t skew the effect of the variable due to their value being zero. Next I utilized the date variable to extract the year and subtracted the yr\_built to create the variable Age\_when\_listed. Along with these two variables I had one major categorical variable and that was City. For this I had to factor the variable for its use in the modeling process.

In addition to these variable changes I made the decision to cut down our dataset based on the price of homes. Initially I removed just our major outliers, those under $100,000, and those over $5,000,000. This helped in removing any 0 values and outliers on the low end, and also high end outliers like the one home in our dataset with a price of over $26 million. After further testing with our models I found some success in bringing down error measures when I limited our data to homes priced between $100,000 and $1,000,000.

**Models**

*K-NN Regression*

Along with the data cleaning mentioned previously, in the K-NN model I then had to make each city output a binary variable, then normalize each one of these. After normalizing all significant data and using price as the output, I cross validated the data using the caret program. Using K-fold with a K of 10 I cross validated the data which gave us the following measures.

K-NN Error Measures

A picture containing text, font, document, black and white

Description automatically generated

The K-value with the smallest RSME occurred at seven giving us the optimal K. However this still gave us the highest RSME and MAE of the three models.

K value vs RSME Graph

A picture containing diagram, line, text, plot

Description automatically generated

*Linear Regression*

For this model I utilized backward elimination to figure out which variables were the most significant in relation to price. Out of the 13 variables 11 of them were significant when it came to the overall price of a house. The sqft\_above and sqft\_basement were the summation of the sqft\_living variable so I decided it was best to remove the sqft\_living variable from the regression model to ensure the model reflected the effect of both above ground spaces and basement spaces. It was also necessary to split this model into 70% training and 30% validation and set the seed to 30. The model generated an adjusted R squared of 0.6959. Although the R-squared value does not reflect a perfect model, it provided us with a number of significant variables to draw conclusions on.

Summary of Linear Regression

A picture containing text, font, screenshot, white

Description automatically generated

In an attempt to better our accuracy measures for the linear regression model, I used K-Fold cross validation and set the K equal to 10.

Linear Regression Summary of City Variable

A screenshot of a computer

Description automatically generated with medium confidence​​

Mean/SD RMSE and MAPE for K-Fold Cross Validation of Linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
| Mean RMSE | SD RMSE | Mean MAPE | SD MAPE |
| 110677 | 6257 | 19.7 | 1.04 |

*Regression Tree*

For this model there were a few things that had to be done before running it. To begin, while running the model in early stages of our work I had issues with some of the cities in the dataset. Both Medina and Yarrow Point only had one row of data present in the dataset so they had to be excluded from the model. For this model the data was split into 70% training, and 30% validation in line with both of our other models. I began by running a simple regression tree and plotting the CP values to find what our optimal amount of splits would be. For this tree I decided that an nsplit of 27 would be ideal because the error begins to stabilize around this point.

Plot of CP Table for Regression Tree

A graph of size of tree

Description automatically generated with low confidence

From here I moved on to actually getting a visual on the tree itself. Here I can visualize the significance of many variables and attempt to draw conclusions that will help us determine answers to our overarching question. In this model the most significant variables were City, Sqft\_Above, Sqft\_Basement, and View. Along with finding out the most significant variables for this model I also collected error measures before moving on to the cross validation.

Regression Tree output (Shortened to 7 splits for visual clarity)

A picture containing text, diagram

Description automatically generated

Accuracy Measures for simple Regression Tree model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ME | RMSE | MAE | MPE | MAPE |
| 4093.218 | 126977.6 | 93925.04 | -6.81149 | 22.54763 |

Finally for this model I performed a k-fold cross validation to try and reduce the error and improve the ability of the model to predict more accurately the price of homes. The cross validation was run with K equal to 10 and I recorded the mean and standard deviation of the RMSE and MAPE values for the 10 iterations of our tree. I found mild success bringing down the RMSE and MAPE values slightly leaving us with the following error measures.

Mean/SD RMSE and MAPE for K-Fold Cross Validation of Regression Tree

|  |  |  |  |
| --- | --- | --- | --- |
| Mean RMSE | SD RMSE | Mean MAPE | SD MAPE |
| 123480 | 5876 | 21.4 | 0.689 |

**Analysis Results**

Mean RMSE and MAPE values for cross validation of each model

|  |  |  |
| --- | --- | --- |
| Model | RMSE | MAPE |
| K-NN Regression | 128333.4 | - |
| Linear Regression | 110677 | 19.7 |
| Regression Tree | 123480 | 21.4 |

*Model Analysis*

After finding error measures and performing k-fold cross validations on all three of our models I found the best model for our specific problem is the Linear Regression model. I found a MAPE of 19.7, an RMSE of 110677, and an R Squared value of 0.6959. This was our model with the lowest error measures and the R Squared signified that our model and it’s variables accounted for around 70% of the variation seen in price for this data. While this isn’t a perfect result I did find that many of our variables were significant in predicting the price of homes giving us a lot of options in recommending which factors should be focused on when evaluating what factors are most important.

While most of our variables showed a high level of significance there were some that stood out with very high coefficients, the first of which was waterfront. The waterfront variable was a binary variable (1 or 0) and determined that if a house had a waterfront it would see a $129,600 increase in price. Along with this I found bathrooms to be an extremely strong predictor in price showing an increase in bathroom providing a $37,100 increase in price. City was our factored variable and had a variety of cities that when compared to Auburn had a notable effect on price as well.

*Variable Analysis*

In our search for the most significant factors in the pricing of a home I also came across some issues with some variables. First of all was the year built variable. Because we are dealing with listings that happen in 2014 I subtracted the year built variable from the date variable to get the age of the homes when they were listed and created the age\_when\_listed variable. I found an issue with this variable because the older a house was at the time of listing actually increased the price. Although the dataset isn’t small, it could be considered small for the wide range of ages of houses in the dataset. As stated before the dataset has houses that were built up to over 100 years before they were listed. If I had more data to work with or a smaller range of ages it would be possible that we could see this coefficient flip and become more consistent with what we would expect out of a variable like this.

Another variable I wanted to note as being out of the ordinary was the bedroom variable. We would expect that the addition of a bedroom would cause a significant increase in price, but our chosen model listed it as a significant variable with a negative coefficient. Again, it looks like we have a really wide range of values in the bedroom variable in our dataset (0 - 9), and after looking into it further I think I have a couple ideas of what could be going wrong here. The first is that there may be some data entry/collection error in this variable. Some rows showed houses with really high square footage but a very low number of bedrooms or smaller houses with many bedrooms. Because the dataset does include real data that can help us locate these specific houses online, I was able to check some of our records and concluded that there were some errors in the data concerning the amount of bedrooms houses had. The second reason is that some of the listings may be empty lots. Some rows featured a high square footage of the lot but no bedrooms at all. This could also be due to data entry/collection errors but may just be empty lots being listed in the same place as houses.

**Recommendations**

After running the three different regression models and assessing our linear regression model to be our most accurate due to its low MAPE level of 19.5. Implementing this predictive model will allow real estate companies to accurately list and relist houses based on their specifications. This also lets real estate agents gauge what houses are worth their time by giving them strong indicators on the house's price, allowing the agency to target more expensive homes. When gauging a house’s price using this model realtors should pay attention to the following variables.

* Waterfront: If clients house is considered a waterfront property price increases by an average of $129,600
* Number of Bathrooms: On average each bathroom increases price by $37,100
* City: Houses in Seattle, Mercer, and Bellevue, increased prices relative to Auburn by on average $239,00, $392,800, and $307,900 respectively.

Also using our model to predict average price listing based on the city we are able to observe average price relative to Auburn. This gives real estate agents insight on what cities contain the highest and lowest price homes compared to those in Auburn, letting them focus their attention on cities with listings that will net them higher profits. We recommend our agents look for listings in Seattle, Mercer, Bellevue as they have higher average listing prices and a significant number of listings within our dataset.

Finally for realtors that have listings in cheaper markets such as Enumclaw and Federal Way they could use this model to find ways to increase the price of their homes. One alternative is adding a bathroom, on average the price rose by $37,100 with an extra bathroom. Another is looking to renovate or fix up the home, for each condition rating (0-5) home prices rose by $24,900. Thus, if these price increases outweigh the costs, it could be beneficial for agents to suggest renovations in order to increase the price of houses they are selling.

All of this considered, we have come up with the following managerial insights for real estate agencies in Washington looking to make additional profits on the sales of homes in this region.

* When compared to houses in Auburn, there is a steep incline in price in Seattle, Bellevue, Mercer. Real estate agencies should focus on these areas if they wish to target specific locations with high end homes.
* The strongest predictors of an increase in price were being on the waterfront, additional bathrooms. Homes with these features are prime for targeting higher prices.
* Renovations to existing properties could be beneficial if the predicted gain outweighs the cost, we suggest improving the condition or making an addition to the home.

**Conclusion**

As discussed earlier this model could be used by a real estate agency to decide what factors are most important when pricing a home, by home owners to determine whether or not it would be worth it to make additions to their home, or by home buyers searching to see what aspects of a home they should pay attention to to ensure they are getting a fair price. For the purpose of answering our specific question however, we wanted to be able to point to a few key factors in house pricing that could help a real estate agency increase overall profits. After exploring three models and performing cross validations on those models, the Linear Regression model was our best fit for this task.