Prediction of Housing Prices in King County, Washington

Ronny Malkowski1, Nicholas Nikolov1, Alexandria A. Pokorny1

1 Department of Mathematics, Illinois State University, Normal, IL 61761

rmalkow@ilstu.edu, ndnikol@ilstu.edu, aapokor@ilstu.edu

**Abstract**

Determining the price of a home was once an abstract process based on intuition and rough comparison to neighboring homes. However, in the modern day it would not seem to be sufficient to determine profits on mere intuition. Proper housing price modeling within relatively fixed time is affected by various house parameters such as square footage, number of bedrooms, number of bathrooms, etc. This project utilizes several selection and assessment techniques to develop the most superior model for predicting the expected price of a home. Data was collected as housing sales from May 2014 - May 2015 in King County, Washington. Evaluation

**Introduction**

The creation of a regression model for determining housing prices is of high importance for both consumers and appraisers. Consumers benefit from the ability to predict an expected price based on the model and compare the result to the asking price of a home on the market. This gives consumers increased power in the price they offer for a home. Alternatively, having a comprehensive model for appraisers provides a consistent method of pricing to assist in the selling of a home.

Some of the parameters that go into determining the price of a house are square footage, number of rooms, number of bathrooms, number of floors. Other parameters include was it renovated or not, grade from appraiser, condition and more. We have found a data set that includes all of these parameters and more.

A previous model utilized Artificial Neural Networks to predict housing value using square footage, bedrooms, bathrooms, and age. This study compared traditional evaluation techniques with modern ones. It split the data into 18 subsets of growing size and with each subset a subset of the others. In this way, subset T1 being of size 306 was a subset of T2 which was of size 506, etc. The conclusion drawn was that Neural Networks were not as effective at lower sample sizes but proved highly effective at the higher sample sizes (note the study had a dataset of 3,906 housing sales). This provides clear incentive to develop a good neural network to select the ideal model. (Morano, Tajani, & Torre, n.d.)

Another paper titled “Quantifying the Value of a View in a Single-Family Housing Market” took a look at the housing market and found that a good view in a home can effect the homes total value. The authors of the paper performed a multiple regression analysis on homes in Fairfax County, Virginia. Their model looked at nine different parameters. Some of the parameters they looked at are number of bedrooms, number of bathrooms, square footage of the lot the home is on, age of the house, and does the home have a good view or not. The data on good view was either one of two numbers; a 1 if the house has a good view or a 0 otherwise. The conclusion of the paper was that a good view adds an 8% value to the home. This is important to this research because one of the questions to be addressed is does having a basement affect the model, which is a similar to the question to how does a good view affect the price of a home. (Nguyen, & Cripps, 2001)

**Results**

Our goal was to predict housing prices accurately with the highest attainable accuracy which we defined to be the model with the lowest root mean squared error (RMSE) and average deviation. Average deviation was observed as the mean and median deviation (predicted - actual) of predictions made using models trained on a training set against the actual housing prices in the test set.

The data required initial preprocessing before a feasible model-building process could be conducted. During the feature-selection stage we began by omitting variables that would clearly demonstrate no predictive power such as ID. Additionally we omitted date, latitude, and longitude. Date was omitted because the form in which it was written caused read-in problems. We anticipate that our model should not be severely hindered by this because all sales were recorded in the span of one year. Latitude and longitude were omitted because every observation was unique. The only way to properly utilize these variables would be to “round” each one so it could be used as a grid (e.g. all latitudes of the form 37.56\*\*\*\*, where \* represents an integer from 0 to 9). This would be one method of specifying housing location but we made the decision to utilize another provided variable; zip code.

Before we could utilize zip code, it had to be factorized and a set of dummy variables was created. This is clear because zip code is a classifier of house location and regressing a numerical value for zip code onto housing price would be worthless.

Additionally, an initial training and test set was created for evaluation of model performance. The split designated roughly 60% of data into the training set and 40% into the test set.

**Ordinary Least Squares Process**

Before proceeding with development of a proper model, various adequacy measures had to be checked. Initially we observed residuals behavior through normality plot and the residual plot and (see Figure 1 and Figure 2 in Appendix 1). These plots are the result of regressing all variables (after omissions mentioned previously) onto the response, house price. Immediately we can detect the existence of strong outliers in the data. We confirmed this prevalence by finding that the maximum price in our dataset was almost $8 million. This poses a problem for regression analysis as our least squares estimators are being influenced by these large observations. The normality plot confirms this, demonstrating a positive skew of the data.

We then observed multicollinearity effects in the data. The variance inflation factors suggested that there generally was not an immense impact on variance with a maximum VIF being observed with maximum square footage above the basement (i.e., the first, second, third, etc. floor). This predictor appeared to inflate our variance by a factor of 5.22. This inflation could warrant attention but it’s also expected that this variable would be collinear with basement lot square footage and basement square footage. It may be unexpected that a massive house would have a tiny basement. We determined that these variables are important enough to leave in the model. Nor is a vif of 5.22 typically high enough to warrant eliminating the variable.

A modified regression model was fit to the data excluding variables with residuals outside of 4 standard deviations of the median. The model was tested against a test sample which included the omitted residuals. The RMSE came out to 143,030.6. On average this predictive model deviated from the actual test set with a mean of $90,954.82 and a median of $64,129.50. This is an improvement on the baseline regression model which was fit with all regressors and no pre-processing. The baseline model demonstrated an RMSE of 151,035.4 with average price deviation with a mean of $97,386.32 and median of $68,650.84. This makes clear that the model trained with outliers was influenced to a significant degree. The noticeable differences between mean and median price deviation shows a non-normal pattern. This can be seen in the data which has very evident outliers (the highest costing home in the dataset sold for almost $8 million).

The next logical step is to perform subset selection. The backward-selection subset reduced the predictive variables from 84 to 68 but had little significant impact on prediction with an RMSE of 143,366.8 and average predicted mean deviation of $91,193.73 and median deviation of $64,106.65. The forward selection model pushed all variables into the model. We felt that traditional subset selection would not provide notable improvement to our model. Surprisingly, the adjusted R2 is almost the same in the summary: 0.8062 for the unreduced model and 0.8057 for the reduced model. This result is surprising since some predictors appeared very insignificant and the adjusted R2 did not punish the model excessively.

Further, concerns over the robustness of this algorithm arose when it was realized that the algorithm was omitting a few of the zip code dummy variables. Although these variables may not have been deemed as important for the data provided, they would make future predictions using omitted zip codes less meaningful. The only other non-significant variable was sqft\_lot15, a description of the square footage of the living room in 2015. Thus we determined a linear fit on all predictors (forcing every available zip code into the model) except for sqft\_lot15 and acquired an adjusted R2 of 0.8061 but a slight improvement to overall accuracy. RMSE was 143,053.7, mean average deviation in prediction was $90,962.02, and median deviation was $64,130.55. It is not likely that much improvement can be derived from this.

Following initial subset selection, investigation was conducted on the form of the data. Of principal interest was the relationship of individual predictors on the price of homes. Naturally this could not be graphically investigated for the zip code variables but we did research the quantitative variables and discovered certain patterns that suggested potential benefit from conducting a data transformation. One of the more telling variables was the plot of grade to price (See Figure 4 Appendix 2) . The data shows a potentially nonlinear pattern with non-constant variance. This model inadequacy did not stop at grade (a subjective determination of home quality). We discovered this pattern even with the number of bedrooms plotted against price (See Figure 3 Appendix 2) . The result found was very surprising. We found that the minimum price of homes in our dataset increased slightly but persistently with additional bedrooms but overall the pattern did not persist and peaked at 8 bedrooms even though there were many homes with more bedrooms. Notice the outlier in the plot and the described pattern as well.

We then decided to investigate the effect of a logarithmic transformation on the price and tested the model’s performance on the test set. This offered substantial improvement over what we had previously. Our adjusted R2 was 0.8774 (notably higher than all previous models). We achieved an RMSE of 122,718, a mean deviation from true price of $71,695.63, and a median deviation of $44,084.43. This median deviation was roughly $20,000 more precise than previous models.

The final model’s normality plot can be observed in Figure 6 Appendix 2. Rather than solely rely on the normal Quantile-Quantile plot we include a histogram of the residuals as well to really clarify the distribution observed (See Figure 5 Appendix 2).

Overall we can see that the skew from the initial model is gone and we are left with some more extreme values than expected but the overall shape still appears relatively acceptable. Ultimately, we feel that further work on this regression model will not lead to substantial benefit.

**Artificial Neural Network Process**

The usage of artificial neural networks has demonstrated substantial strength in its ability to predict in various scenarios. It has been praised for its ability to optimize the accuracy of regression problems as well as its ability to perform multi-factor classification.

In order to create a successful model, it is necessary to utilize a sufficiently developed package. nnet() is insufficient for an actual project due to its limited usage. neuralnet() was attempted but its slow speed prevented its use (fast training is important for cross-validating results). Ultimately we decided to utilize the package h2o which allowed us to make use of all threads in our CPU.

H2o’s deep-learning function uses a feedforward trained with stochastic descent using back-propagation. It allows analysts to adjust a plethora of parameters such as overall architecture (number of layers), activation functions, and loss function.

We began by testing a simple model. Initially we considered utilizing NeuralNet but found the progress quite slow. For this reason we decided to work with H2O which is one of the most utilized packages for Fortune 500 companies. We began by simply one-layer model fits.

The first model fit was utilized to perform an evaluation on the process of fitting neural networks to the data. The advantage we have with neural networks is that outliers needn’t be trimmed as they are in regression analysis. In fact, typically results would be worsened by omitting data that could help the neural network learn from the outliers. The models fit were all feedforward neural networks trained with stochastic gradient descent using backpropagation and using an absolute loss function (as opposed to MSE). The first model was trained with a single layer of 100 neurons activated using the rectifier function over 30 epochs and yielded promising results. RMSE was 118,360, mean deviation from price was $69,377.81, and median deviation was $42,760.74. This model was overall superior in its ability to predict housing prices.

Next we studied various neural network architectures in order to procure a superior model. We discovered that a superior model could be developed using a rectifier-activated hidden layer of 25 neurons passing to a rectifier-activated hidden layer with 19 neurons. This model demonstrated further improvement. During the search process it demonstrated an ability to achieve median deviation from true price of $40,641.025 (during the initial search). When run over 300 epochs this model demonstrated an ability to bring median deviation down to $41,345.82. Even mean deviation was improved to $69,588.98 and RMSE came out to 118,745.3.

The results of this show that the artificial neural network has a lot of potential in predicting housing prices. The worst neural network tended to perform as well as the selected linear regression model. Although the difference wasn’t immense, there are further parameters that would be worth considering in order to further reduce error. Nonetheless, for most houses in the dataset, the neural network predicted quite well (often predicting within $1000 of the actual house price).

**Bibliography**

Morano, P., Tajani, F., & Torre, C. M. (n.d.). Artificial intelligence in property valuations An application of artificial neural networks to housing appraisal, 7.

Newsome, B. A., & Zietz, J. (1992). Adjusting Comparable Sales Using Multiple Regression Analysis - The Need for Segmentation. *The Appraisal Journal; Chicago*, *60*(1), 129.

Nguyen, N., & Cripps, A. (2001) Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks, *Journal of Real Estate Research, Vol. 2*

Rodrigez, M. & Simans, C.F. (1994) Quantifying the Value of Single Family Housing Markets, *The Appraisal Journal.*

Sirmans, G. S., MacDonald, L., Macpherson, D. A., & Zietz, E. N. (2006). The Value of Housing Characteristics: A Meta Analysis. *The Journal of Real Estate Finance and Economics*, *33*(3), 215–240. https://doi.org/10.1007/s11146-006-9983-5

**Appendix 1**

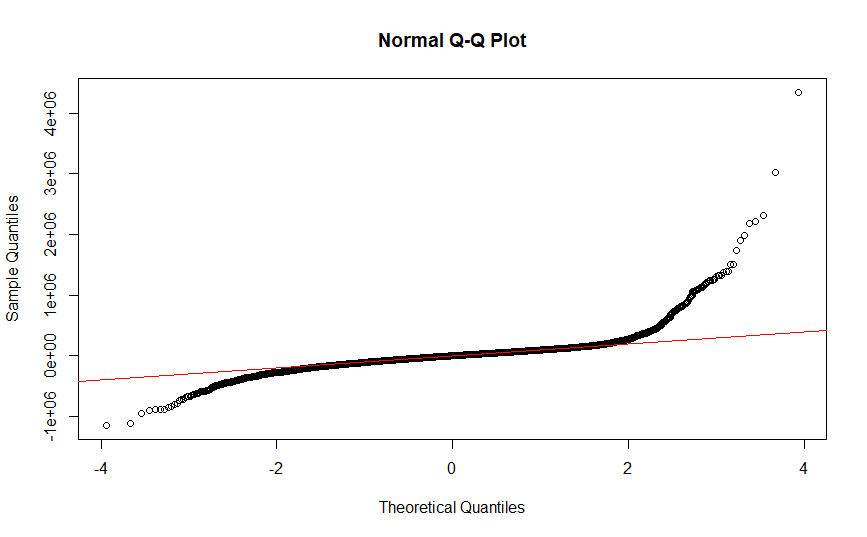
Figure 1

Figure 2

**Appendix 2**

Figure 3

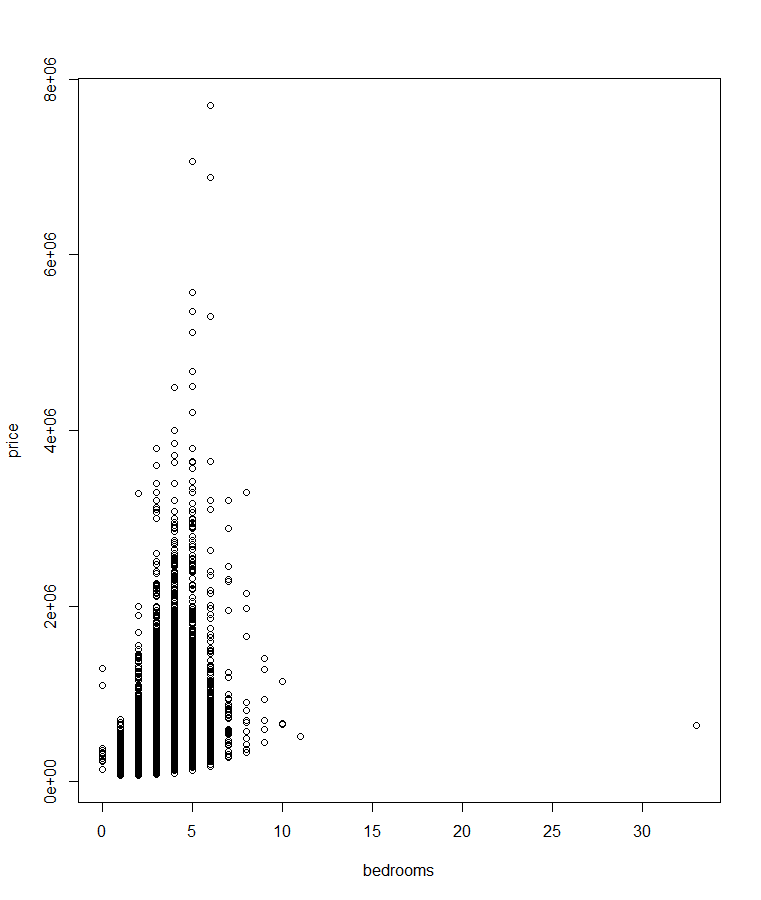


Figure 4

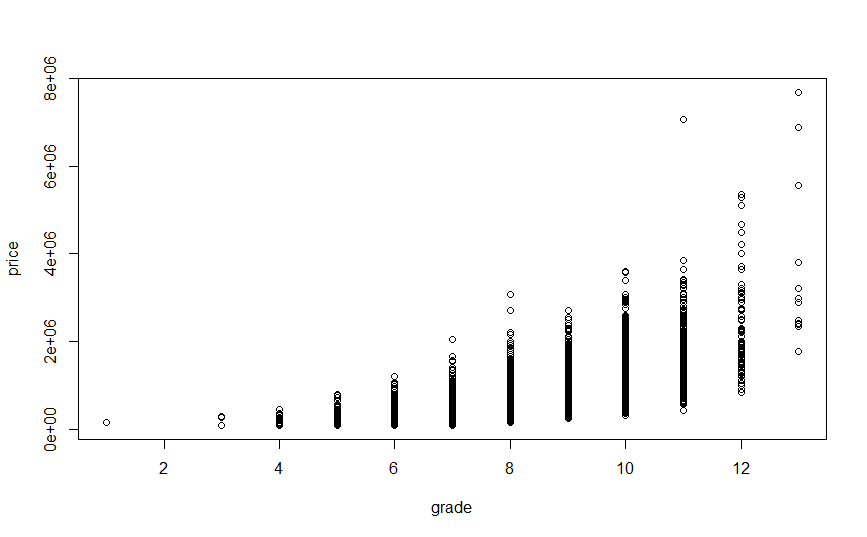


Figure 5

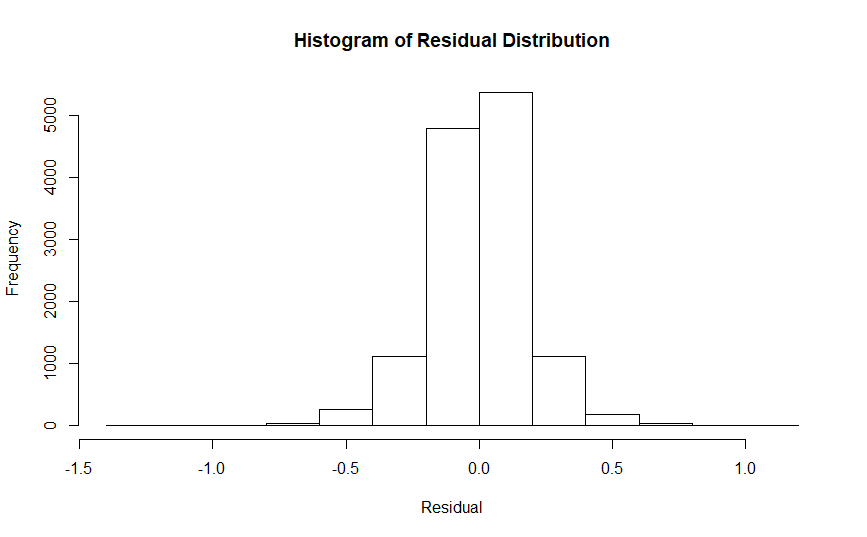


Figure 6

