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)

Creating the most superior model based on the data will occur in several steps which will be elaborated upon following their presentation. First Data will be imported and preprocessed, a full regression model will be fit, and initial issues will be addressed. Next, we will seek and possibly remedy influential outliers. The model will then enter the subset selection process. During the subset selection model evaluation will take place. The superior model will be selected and presented as final.

The preprocessing step will begin by splitting the observations into a training set and test set. Most likely the split will be 70% Training Data will offer a large dataset to train the data but still provide a sufficiently large test set of 6,483 observations. A final important step in the preprocessing stage will be feature scaling in which we normalize data. Although not applicable to simpler subset selection methods like Best Subset, Forward Subset, and Backward Subset Selection, it will be necessary if employing more advanced techniques like Elastic Net, Lasso, or Ridge Regression. This is due to the constraining nature of the methods on the regressors being affected by scale.

Following the data preprocessing step a full regression model will be fit with all 19 regressors. This step has already been conducted. At present, all regressors appear significant although a singularity issue appeared due to perfect collinearity in the square footage variables. This is apparent because total square footage is linearly related to basement square footage and living square footage. R automatically drops basement square footage but additional analysis will be conducted to alleviate this issue and possibly utilize this variable. Other multicollinearity issues must be explored. High Variance Inflation will negatively impact our model and thus must be addressed before the model enters proper subset selection. Otherwise, the initial model appears quite strong with an .

Outliers and high-leverage points may not be influential if only a few such points exist. But if a large subset of the dataset features outliers or high-leverage points it might be worth analyzing how influential the effect is. It is hoped, however, that outliers will not cause a substantial impact on the model with such a large subset.

After outliers have been detected and decisions made on how to address any issues they may pose, the model will enter subset selection. Various methods exist for subset selection. This is the step expected to take the most time. We will attempt to create the most superior model possible using classical methods such as Forward, Backward, and Best Subset Selection as well as more modern techniques like Lasso, Ridge Regression, or a combination of the two through Elastic Net. We will also seek to utilize artificial neural networks (ANN) to select a superior model. The ANN will not be used in addition to traditional stepwise subset selection. The ultimate goal is to compare the best model from several different methods of model selection. More specifically, we will fit three models using Lasso, Ridge Regression, and Elastic Net. We will then fit a model using stepwise selection and a model using best subset selection. Lastly we will establish an ANN which will begin with the full initial model and select the most superior model. The models will be evaluated against the test data on the basis of which model has the lowest squared error. Lastly, as a clarifying note, elastic net will be iteratively evaluated at different values of alpha with the hope that a superior model can be found somewhere between pure Lasso and pure Ridge Regression. We anticipate this method being the number one contender for ANN.

The goal is to create a regression model with a high adjusted coefficient of determination and low mean squared error. The superior model and any additional findings will be presented. Findings may include initial model issues, parameters of interest, outlier influence, dropped parameters and subsequent discussion.

**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>