STT 3851: Statistical Data Analysis II

Predicting Housing Prices in King County, WA

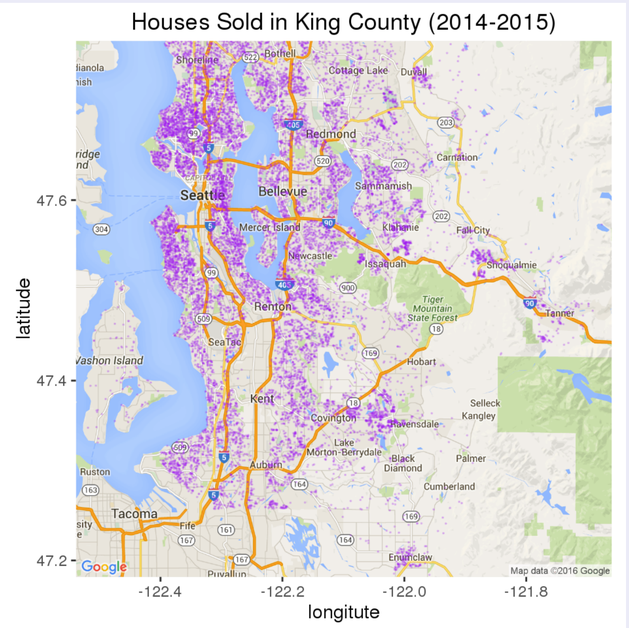
Andrew Sullivan, Tucker Southern, Hannah Laws

Department of Mathematical Sciences

Appalachian State University
Boone, North Carolina

Overview

- The goal of this project was to predict prices for houses in King County, Washington.
- Data was examined from 17384 houses sold in the county between 2014 and 2015 in order to construct a pricing model.
- Models were constructed through exploratory analysis and forward AIC selection, and then tested using k-fold cross-validation.
- All computations and graphs are created with the open source software R [5].

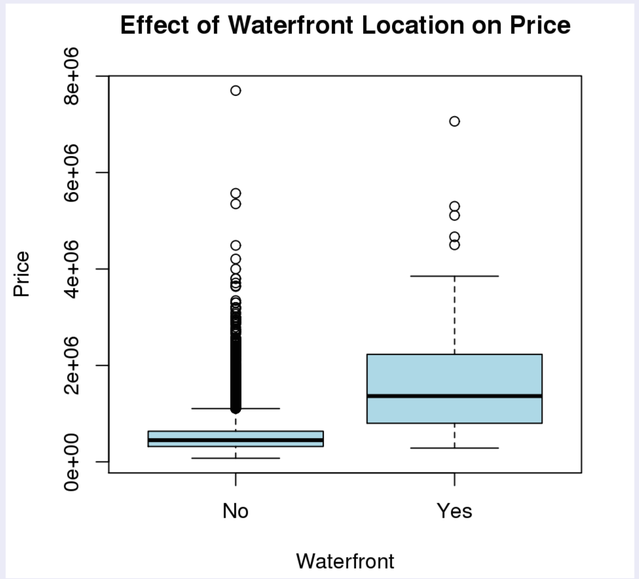


Data Formatting

- The variables *Waterfront*, *Condition*, *Grade*, and *Zipcode* were converted from numeric values to factors.
- The variable *YearRenovated* was set to the corresponding *YearBuilt* for any houses that were missing *YearRenovated* values - that is, any houses that had not been renovated had their renovation dates re-set to the dates they were originally built.
- The variables *Grade* and *Condition* were collapsed to account for limited observations and limited distinct effect in their lower categories.
- Finally, a new variable, *LotSize*, was introduced based on established realtor lot categories [4].

Exploratory Analysis

- Exploratory analysis was performed by examining plots and single-variable regressions for various variables on to *Price*, as well as variable interactions which were assumed to be significant (such as the interaction between *Bedrooms* and *Bathrooms* on to *Price*) [3].
- Variables and interactions which looked to have strong correlation were later added to the model and tested for significance.
- The boxplot below shows the effect of waterfront location on house price:



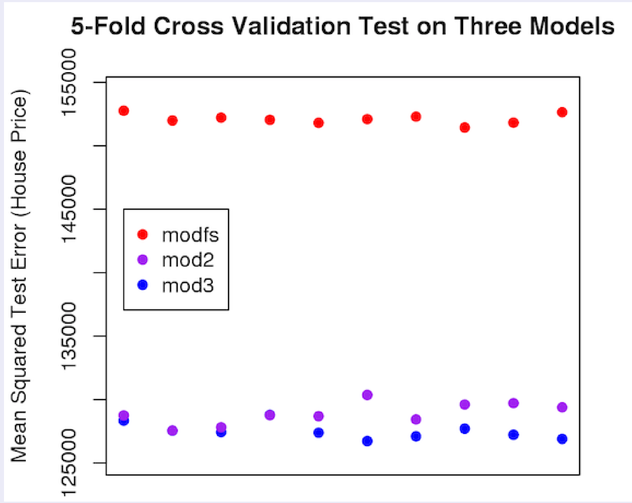
Model Creation

- The first model was created by comparing AIC in a forward stepwise algorithm [6].
- The second model included polynomials based on residual plots of the forward-selected model [2], and interaction terms based on exploratory analysis.
- The final model was a simplified version of the second, dropping features with low significance [3]. It included these features and interactions:

$$E(price) = b_0 + b_1 Grade + b_2 Zipcode + b_3 SqftLiving^2 + b_4 Waterfront + b_5 View + b_6 LotSize + b_7 Condition + b_7 SqftAbove^2 + b_8 YearBuilt + b_9 YearRenovated + b_{10} Floors + b_{11} SqftLiving15 + b_{12} SqftLot15 + b_{13}(SqftLiving : SqftLot) + b_{15}(Bedrooms : Bathrooms) + b_{16}(Waterfront : SqftLiving) + b_{17}(Waterfront : SqftLot) + b_{18}(Lat : Long) + b_{19}(Zipcode : SqftLiving)$$

Model Selection

- For k-fold cross-validation, the original data is divided into k equal sized subsamples. One of these subsamples is set aside as the validation data sample for testing the model created from using the training data. The training data is comprised of the rest of the subsamples, which is used to create the model that will be tested with the validation data subsample. This is repeated k times so that each subsample is used as the validation set once and the errors are then combined to produce a single estimation.
- We used 5-fold cross-validation to select the best model [1]. The simplified final model consistently performed the best, having the lowest mean squared test error.
- Results of ten repetitions of cross-validation on all three models are plotted below:



References

[1] Angelo Canty and Brian Ripley.
boot: Bootstrap Functions (Originally by Angelo Canty for S), 2016.
R package version 1.3-18.

[2] John Fox and Sanford Weisberg.
car: Companion to Applied Regression, 2015.
R package version 2.1-0.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors.
An introduction to statistical learning: with applications in R.
Number 103 in Springer texts in statistics. Springer, New York, 2013.

[4] Iain Pardoe.
Modeling Home Prices Using Realtor Data.
Journal of Statistics Education, 16(2), 2008.

[5] R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2015.

[6] Brian Ripley.
MASS: Support Functions and Datasets for Venables and Ripley's MASS, 2015.