King County House Price Prediction Project

Price Prediction Project

Cescily Metzgar, Averia Padgett

Department of Mathematical Sciences

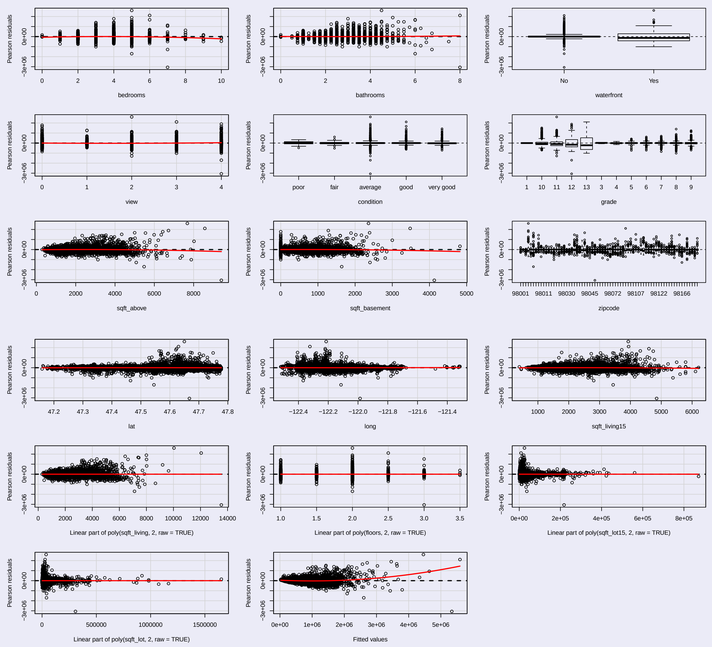
Appalachian State University
Boone, North Carolina

Overview

- This poster examines housing data from King's County and aims to predict prices of houses based on different variables.
- To ensure accurate predictive power for future observations, the data are split into a training set (80%) and a test set (20%).
- Root mean squared error of the test set is used as a measure of model adequacy.
- All computations and graphs are created with the open source software R [1].

Residual Plots

- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.
- Using the residual plots we found variables that need to be made non-linear to create a more randomly distributed residual plot so that the variables can be used in a linear model.



- In the above residual plots the variables sqft_living, floors, sqft_lot15, and sqft_lot were squared which resulted in an almost perfectly distributed residuals for each variable.

Basic Models Used

- Backwards Elimination: The backwards elimination approach takes mod2 as the full model and works backwards to eliminate variables until it produces the best model, resulting in the lowest AIC possible.
- Forward Selection: The forward selection approach takes a null model, a model that is basically empty, and adds variables individually until the best model is produced, resulting in the lowest AIC possible.
- Interaction Model: Examines the interactions between various variables.

King County Real Estate

```
Error in get("f", environment(CoordMap$train)):  object
'f' not found
```

Backwards Elimination

- Below is the model backwards elimination produced. After running a summary of the model it produced an Adjusted R-squared value of 0.8416.

```
Step:  AIC=413806.9
price ~ date + bedrooms + bathrooms + waterfront + view + condition +
    grade + sqft_above + yr_built + yr_renovated + zipcode +
    lat + long + sqft_living15 + poly(sqft_living, 2, raw = TRUE) +
    poly(floors, 2, raw = TRUE) + poly(sqft_lot15, 2, raw = TRUE) +
    poly(sqft_lot, 2, raw = TRUE)
```

Forward Selection

- Below is the model forward selection produced. After running a summary of the model it produced an Adjusted R-squared value of 0.8416, the same Adjusted-R squared backwards elimination produced.

```
Step:  AIC=413806.9
price ~ poly(sqft_living, 2, raw = TRUE) + zipcode + waterfront +
    grade + view + condition + date + sqft_above + poly(floors,
    2, raw = TRUE) + bathrooms + sqft_living15 + yr_built + yr_renovated +
    poly(sqft_lot, 2, raw = TRUE) + long + bedrooms + poly(sqft_lot15,
    2, raw = TRUE) + lat
```

Interaction Model

- Below is the interaction model created to predict house price using various interaction terms. The following interaction terms were included in this model: bedrooms:bathrooms, sqft_living15:sqft_living, condition:grade, yr_built:yr_renovated, and zipcode:long. This model resulted in an Adjusted-R squared value of 0.8606 which is higher than both backwards elimination and forward selection.

```
interaction.mod <- lm(price ~ poly(sqft_living, 2, raw = TRUE) +
zipcode + waterfront + grade + view + condition + date + sqft_above
+ poly(floors, 2, raw = TRUE) + bathrooms + sqft_living15 +
yr_built + yr_renovated + poly(sqft_lot, 2, raw = TRUE) + long +
bedrooms + poly(sqft_lot15, 2, raw = TRUE) + lat + bedrooms
:bathrooms + sqft_living15:sqft_living + condition:grade + yr_built
:yr_renovated + zipcode:long, data = housedata)
```

References

[1] R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2016.