Housing Market Analysis

Seth Chart

Data

Our data was collected from the King County house sales between May 2, 2014 and May 27, 2015.

- https://www.kingcounty.gov/

Goal

Produce a model to predict the price of a house from publicly available information.

We will:

- Identify appropriate predictors for our model.
- Interpret our model.
- Provide actionable guidance.

Predictors: Zip Code

We categorized King County zip codes as high priced, medium priced, or low priced based on the median price of houses that they contained. This provided a rough geographical predictor of price.

Predictors: Size of Neighboring Houses

Our data contained the average square footage of the 15 nearest neighboring houses for each house. This provides a measure of the typical size of a house in the neighborhood.

Predictors: Bathroom to Bedroom Ratio

We computed the number of bathrooms per bedroom for each house represented in our data. This provides a measure of the configuration of a house that does not depend heavily on the size of the house.

The model

Our model for the price P of a house takes the following form:

$$P = 10^{-0.34 - 0.29x_1 - 0.11x_2 + 1.74x_3 + 0.09x_4}$$

Where

- x_1 and x_2 indicate the house is in a low and medium priced zip code respectively.
- x_3 is the transformed average size of neighboring houses.
- x_4 is the transformed number of bathrooms per bedroom.

The model

We obtained our model by running a least squares linear regression on the base 10 logarithm of price, which we call y.

y is the power of 10 that is equal to the price of a house.

- If y = 5, then the price is 10^5 = 100,000
- If y=6, then the price is 10^6 = 1,000,000.

Because house prices range from a few hundred thousand to several million dollars, y is a better for linear regression.

The model

- Our model describes 64% of the variation in the log price y.
- The average error for our model is $266,000.
- Our model generalizes well from training data to testing data.
- The parameters in our model are significantly non-zero.
- Model residuals are uniform over the range of predictions, but are not normally distributed.

Conclusions

- The size of neighboring houses is the single strongest predictor of price.
- The prices of houses in a medium priced zip codes are on average 50% of the price in high priced zip codes.
- The prices of houses in a medium priced zip codes are on average 78% of the price in high priced zip codes.

Guidance

- This model will not be sufficiently precise to evaluate pricing of individual houses.
- It does provide a viable benchmarking tool for aggregate taxation.
- We recommend that this model be used to determine if high priced zipcodes are bearing their share of property tax.

Future Work

- It would be interesting to see how a nearest neighbors model performs in comparison to this linear model.
- A more refined predictor based on location would probably improve the predictive power of this model.

Points of interest

Object Oriented Programming: In python, creating and evaluating a linear regression model requires accessing methods from several libraries with differing approaches to managing the data and outputs. To facilitate a more streamlined workflow, we wrote a class that encapsulates the necessary linear regression functionality

Points of interest

Test Driven Design: In this project I tried to implement test driven design, creating unit tests for each function. This practice helps to ensure that code is doing what we intend and makes refactoring code safer, since we can easily verify that refactored code is still working.

Thank You

https://github.com/sethchart/Housing_Analysis