House Prices Segmentation

Mettilda Kaimathuruth
ABSTRACT

- A 10,000 sq. ft. house in San Francisco, CA vs. a similar house in Stillwater, Oklahoma, would show a stark difference in the real estate price of the house.
- Even in a single city, the cost of two 10,000 sq. ft. houses would differ based on different factors.
- There are a lot of factors that go into the final sale price of the house, such as the condition of the house, proximity to schools and parks, proximity to public transport, and so on.
- Understanding the underlying factors that go into creating the price of each house will help marketers price these houses most effectively.
- The goal is to build a segmentation model to identify differentiating factors for houses which are deterministic in the final house price.

METHODS

- Data Preparation - Handle Outliers, Handle Skewness, Missing Values
- Variable Selection – Decision tree, Stepwise Regression were used to understand variable importance with respect to the target variable
- Different Statistical Models - Five models were used – Decision Tree, Multiple Linear Regression, LASSO Regression, Gradient Boosting and Ensemble model
- Model comparison module was used to identify the best model based on least Average squared error
- Final set of predictor variables were used as base variables to cluster the houses into different segments to understand different profiles for the houses that were sold and their differentiating factors

RESULTS

- Model comparison shows that LASSO regression has the least average squared error among all models and hence is chosen as the best model.
- Average sale price for a house in the dataset is 180,412 and Root Avg. Squared error for LASSO model is 22,720.
- LASSO regression also gave an adjusted R-square of 93%, which is an indicator of the variance explained by model.
- Important factors determined by the model were:
  - Neighborhood
  - MS Subclass (1-Story, 2-Story, Duplex)
  - Lot size (Square Feet)
  - Number of Bedrooms Above Ground
  - Basement Exposure (walkout or garden level walls)
  - Garage Capacity (in terms of number of cars)
  - Exterior covering on house
Using these predictors as bases and hierarchical cluster analysis using Ward’s method, 3 unique clusters were obtained with distinct profiles.

The 3 different clusters have unique differentiating factors which are highlighted below:

- **RESULTS CONTINUED (CLICK TO EDIT)**

- **REFERENCES**

3. Bourassa, Steven C; Cantori, Eva; Hoelsi, Martin, The Journal of Real Estate Research; Apr-Jun 2010; 32, 2; ProQuest
4. Vasilios Piskandararas, Rangan Gupta, Periklis Gogasa, Theophilos Papadimitrou - Forecasting the U.S. real house price index - Department of Economics, Democritus University of Thrace, Greece. Department of Economics, Pretoria University, South Africa

Special thanks to Dr. Miriam McLaugh for her continued guidance and support.
SAS® GLOBAL FORUM 2018

April 8 - 11 | Denver, CO
Colorado Convention Center

#SASGF