SAS® GLOBAL FORUM 2018

USERS PROGRAM

House Prices Segmentation

Mettilda Kaimathuruth

April 8 – 11 | Denver, CO

#SASGF

House Prices Segmentation

Mettilda Kaimathuruth

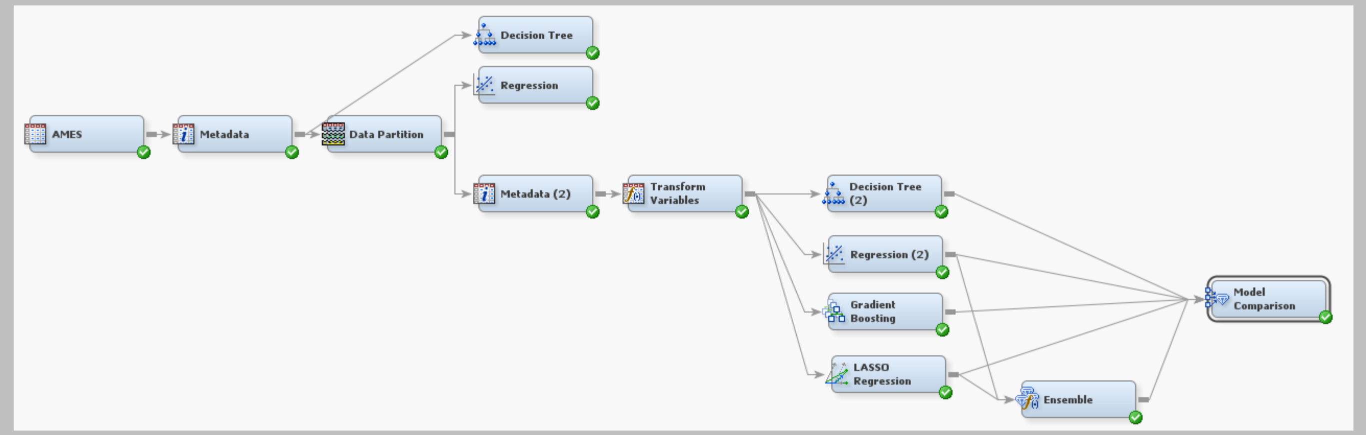
Oklahoma State University

ABSTRACT

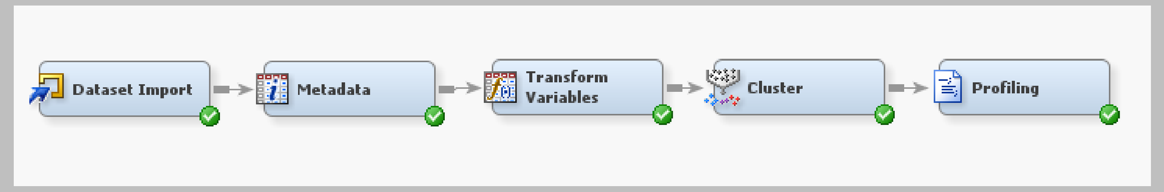
- A 10,000 sq. ft. house in San Francisco, CA vs. a similar house in Stillwater, Oklahoma, would show a stark difference in the real estate price of the house
- Even in a single city, the cost of two 10,000 sq. ft. houses would differ based on different factors
- There are a lot of factors that go into 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

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Valid: Root Average Squared Error	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	LARS	LARS	LASSO ...	SalePrice	SalePrice	5.1622E8	22720.37	150689.8	7.222E11	4.1152E8	20285.89
	Ensmbl	Ensmbl	Ensemble	SalePrice	SalePrice	5.1717E8	22741.27	150224.9	7.196E11	4.1004E8	20249.45
	Reg2	Reg2	Regressi...	SalePrice	SalePrice	5.1907E8	22783.19	149760	7.188E11	4.0955E8	20237.28
	Tree2	Tree2	Decision ...	SalePrice	SalePrice	1.0079E9	31747.17	198983.6	1.303E12	7.4238E8	27246.6
	Boost	Boost	Gradient ...	SalePrice	SalePrice	1.0318E9	32122.42	248360.3	1.596E12	9.0948E8	30157.57

- 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

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	411517219.37	516215283.76
DIV	Divisor for ASE	1755.00	1170.00
MAX	Maximum Absolute Error	150689.78	160905.54
NOBS	Sum of Frequencies	1755.00	1170.00
RASE	Root Average Squared Error	20285.89	22720.37
SSE	Sum of Squared Errors	722212719994.32	603971881996.45

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	90	1.023168E13	1.136853E11	261.93
Error	1664	7.222127E11	434022067	
Corrected Total	1754	1.095389E13		

Root MSE	20833
Dependent Mean	179336
R-Square	0.9341
Adj R-Sq	0.9305
AIC	36750
AICC	36760
SBC	35491
ASE (Train)	411517219
ASE (Validate)	516215284

- 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

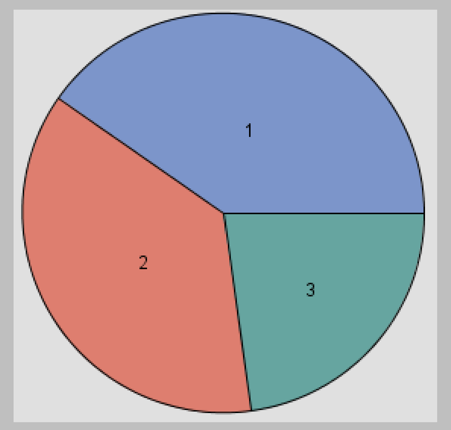
House Prices Segmentation

Mettilda Kaimathuruth

Oklahoma State University

RESULTS CONTINUED (CLICK TO EDIT)

- 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:

SEGMENT	No. of Houses	Lot Size (Sq. Ft.)	Ground Living Area	1st Floor Sq. Ft.	2nd Floor Sq. Ft.	Basement Sq. Ft.	Enclosed Porch Sq. Ft.	Garage (No. of Cars)	Sale Price	Year Built	Year Remodeled	MS Sub Class	Neighborhood
Total		10,104	1,494	1,155	334	438	23	1.76	180,412	1971	1984		
1	1180	9,237	1,659	1,093	564	454	8	2	201,500	1989	1992	More than half the houses 2-STORY 1946 & NEWER	College creek, Gilbert, Sawyer West, Somerset
2	1078	11,604	1,350	1,349	1	566	16	2	187,359	1978	1985	All houses are 1-STORY 1946 & NEWER	North Ames, College Creek, Sawyer
3	667	9,213	1,434	952	464	203	62	1	131,875	1929	1970	Most houses are 1-1/2 STORY FINISHED and 1-STORY 1945 & OLDER	Old Town, Brookside, Iowa DOT and Railroad, Edwards

CONCLUSIONS

- In the final model it is seen that Neighborhood, MSSubClass (type of dwelling such as 1-Story, 2-Story, Duplex), Lot area, No. of Bedrooms, Basement Exposure (which refers to walkout or garden level walls), Garage car capacity, Exterior covering on house, significantly affect the valuation of the home.

Segmentation highlights:

- Houses in segment 1 have less than average lot area, but bigger ground living area, 1st floor area and 2nd floor area, with garage parking capacity for 2 cars and have a higher mean sale price. Most of these houses are 2-story houses that were built post 1946 and in College creek, Gilbert, Sawyer West, Somerset neighborhoods
- Houses in segment 2 have bigger than average lot area and 1st floor area, but no 2nd floor area, with garage parking capacity for 2 cars and have an average sale price as the population. All of these houses are 1-story houses that were built post 1946 and they are located mostly in North Ames, College Creek, Sawyer neighborhoods
- Houses in segment 3 have smaller than average lot area, 1st floor area and basement area, higher than average 2nd floor area and enclosed porch area, with garage parking capacity for 1 car. Their mean sale price is much lower than the rest of the houses. All of these houses are 1.5-story houses that were built prior to 1945 and they are located mostly in Old Town, Brookside, Iowa DOT and Railroad, Edwards neighborhoods

- Houses that were remodeled more recently, with bigger ground living area, basement, garage capacity of at least 2 cars and in College Creek, Gilbert, Sawyer West and Somerset neighborhoods have higher sale price than average.
- Lot area and enclosed porch aren't that important, however houses with 2-story have higher sale price compared to 1-story houses with similar features

REFERENCES

1. The Ames Housing dataset, compiled by Dean De Cock for use in data science education https://ww2.amstat.org/publications/jse/v19n3/decock.pdf
2. Y. Feng and K. Jones, "Comparing multilevel modelling and artificial neural networks in house price prediction," 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, 2015, pp. 108-114. doi: 10.1109/ICSDM.2015.7298035
3. Bourassa, Steven C;Cantoni, Eva;Hoesli, Martin, The Journal of Real Estate Research; Apr-Jun 2010; 32, 2; ProQuest - Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods
4. Vasilios Plakandarasa, Rangan Guptab, Periklis Gogasa, Theophilos Papadimitrioua - Forecasting the U.S. real house price index . Department of Economics, Democritus University of Thrace, Greece. Department of Economics, Pretoria University, South Africa
5. Yusof, Aminah Md and Syuhaida Ismail. "Multiple Regressions in Analysing House Price Variations." (2012).

Special thanks to Dr. Miriam McGaugh for her continued guidance and support

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF