Housing Price Analysis

Team 2: Mckinlytic
Karl Hu & Ellen Wang

Table of Contents

Background

- Volume of Real Estate Market
- Preliminary Screening
- Identify Key Taste Preference
- Time saving and efficiency

Data Preparation

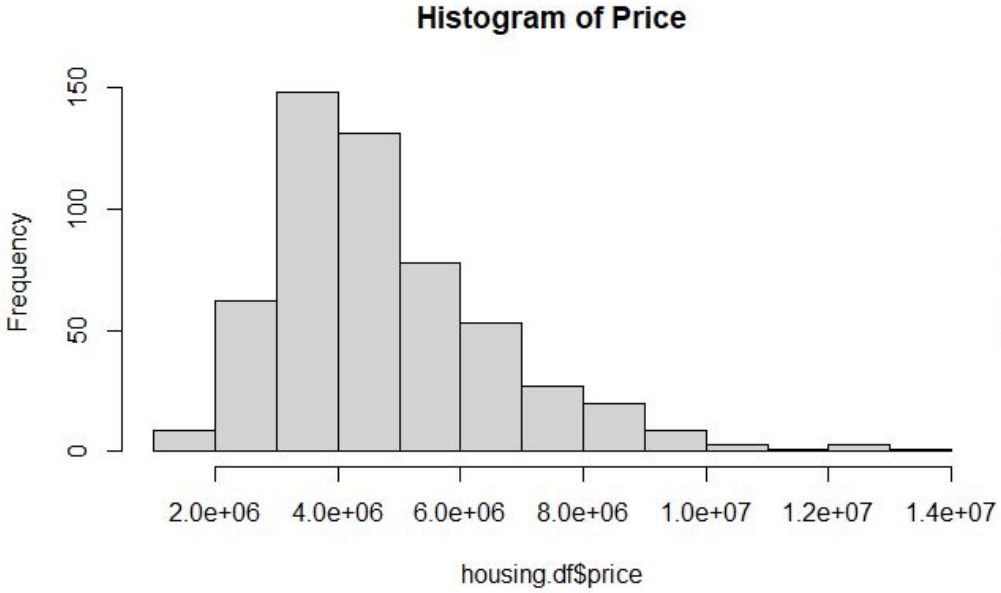
- Quality
- Transformation
- Normalization
- Partition/Validation

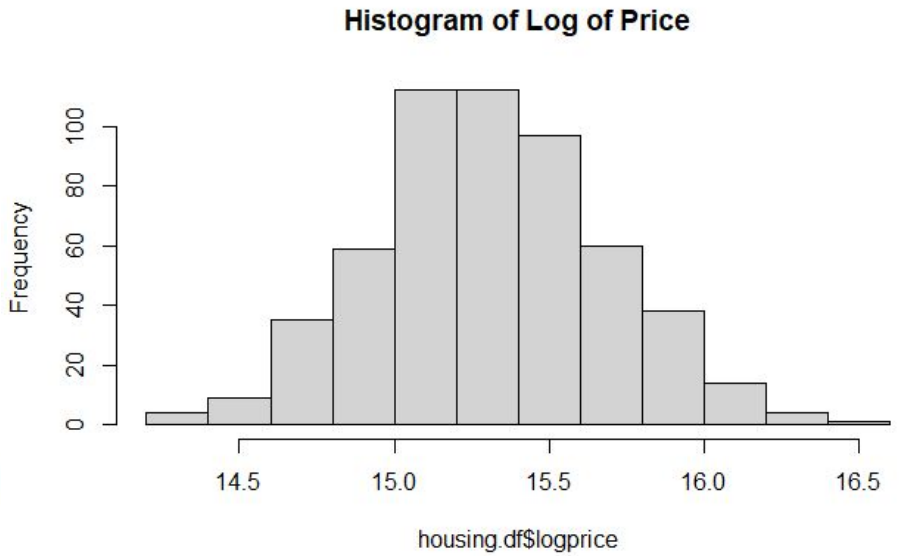
Independent Variables

1. Total area of the property
2. Number of bedrooms
3. Number of bathrooms
4. Number of stories/floors
5. Number of parking spaces
6. Access to the main road
7. Availability of a guest room
8. Availability of a basement
9. Availability of water heaters
10. Availability of air conditioning
11. Postal preferred area status
12. Furnishing status

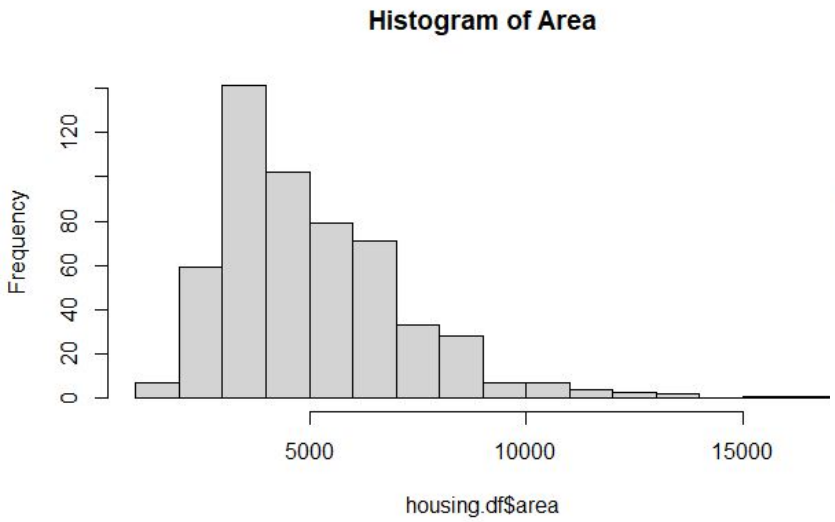
Dependent Variable: Log_Price (Regression), Price Level (Classification)

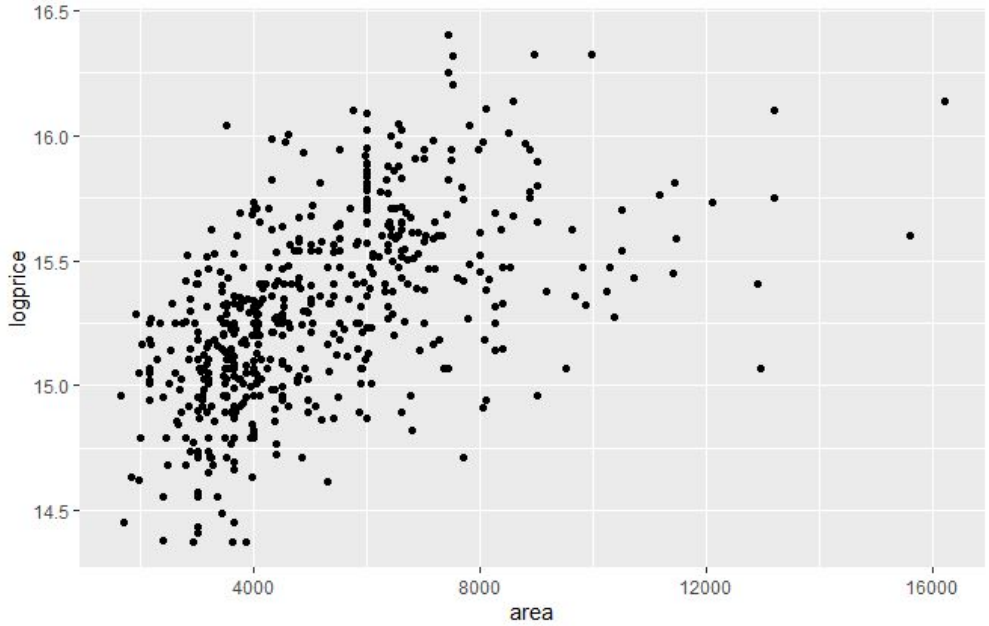
Descriptive Statistics





Descriptive Statistics - cont.





Regression Analysis

Multiple Regression

Best Model

Regression Tree

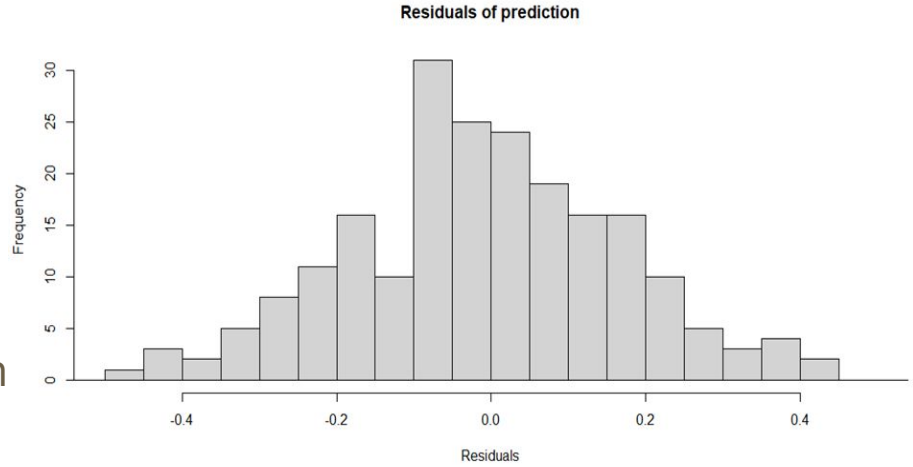
Penalty: 0
Min-Leaf: 21

Random Forest in R

Memory Shortage

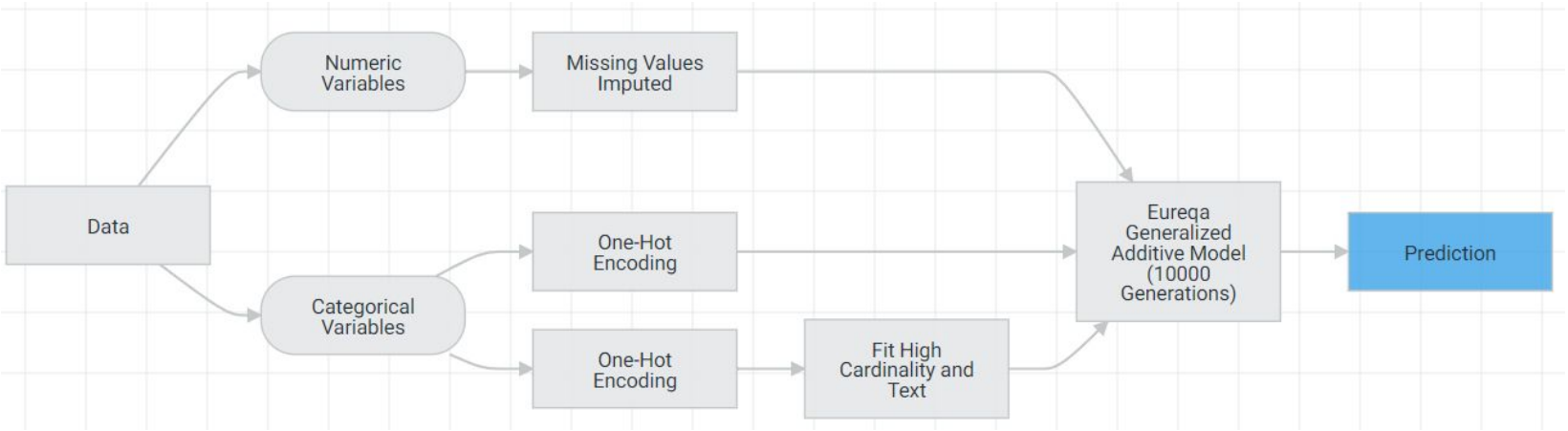
Multiple Regression - cont.

- Started with simple multilinear regression
- Removed insignificant variables
- Removed 7 outliers
- Feature Engineering and Interaction Terms
- Cluster Variables
- RMSE 0.2069
- R-Squared 0.6907



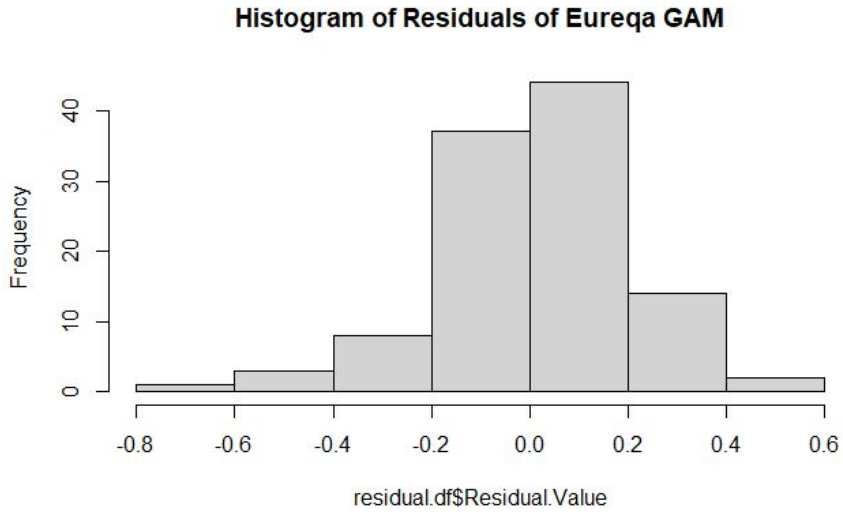
Datarobot - Best Model

Eureqa Generalized Additive Models (Eureqa GAM) - 10000 Generations

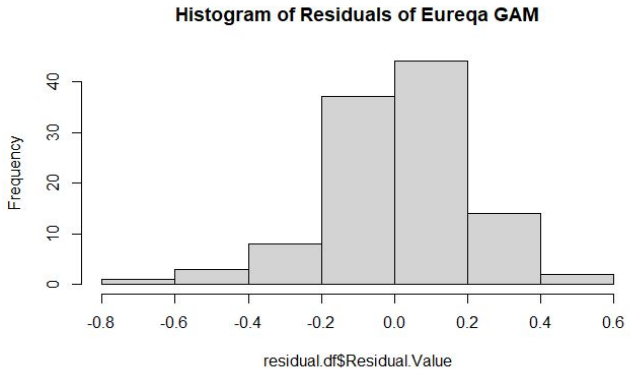
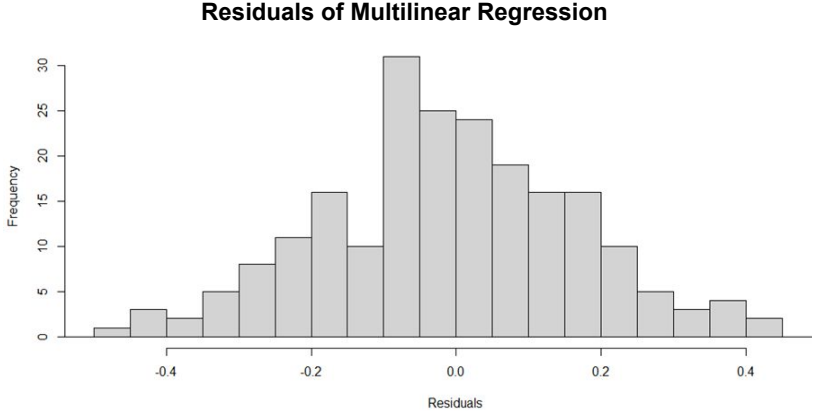


Eureqa GAM - cont

- Unique partition logic
- Complexity: 14
- Error: 0.010(Surrogate MSE)
- RMSE 0.1986
- R-Squared 0.6627



Regression - DataRobot Comparison

DR vs Human	Multilinear Regression	Eureqa GAM
Feature List	Removed 5 variables; Feature Engineering Cluster Variables	All Features
Residuals		
R-Squared	0.6907	0.6627
RMSE	0.2069	0.1986

Classification - Non-Ensemble Models

Logistic Regression*

Target Variable

- Price level

KNN

Grid Search

- KNN: number of k
- Tree: cp, minsplit

Classification Tree

Best Model

- Best non-ensemble: logistic regression

Best Hand-Crafted Model - Logistic Regression

- Started with simple logistic regression on all the variables (excluding price)
  - Removed price to avoid multicollinearity
- Regress on cluster variables
- Split data into training set (60%) and validation set (40%)
- Removed insignificant variables
  - Furnishing status, basement, parking, mainroad, bedrooms
- Removed 19 outliers
  - Greatly improved the accuracy of the model
  - Also suggested by DataRobot
- Checked confusion matrix
  - Overall accuracy = .9221, F1 score = .9161

Classification - Ensemble Models

	Random Forest	Bagged	Boosted
	0.8629	0.8183	0.8653
Hyperparameters	Value Range for Grid Search	Optimal Values	
• Boos • Mfinal • Coeflearn (for Boosted)	• [TRUE, FALSE] • [1,10] • [‘Breiman’, ‘Zhu’, ‘Freund’]	• Boos = FALSE • Mfinal = 8 • Coeflearn = ‘Breiman’	

Classification - DataRobot Comparison

	Best Hand-Crafted Model	Best DataRobot Model
Model of Choice	Logistic Regression*	Light Gradient Boosting Tree
Feature List	Removed 6 variables; Cluster variables	All Features (excluding price)
Top Factors By Importance	1. Air conditioning 2. Hot water heating 3. Guestroom	1. Area 2. Stories 3. Bathrooms
Confusion Matrix	172 14 10 112	47 17 1 44
AUC	0.9831	0.9372

Conclusions & Insights

- Model Building
  - Less is more
    - Insignificant variables
    - Outliers
  - Tradeoffs between interpretability vs accuracy

- Results
  - Which variables are most important in predicting house price:
    - area
    - Bathrooms
    - airconditioning

Thank you!