Machine Learning Project

HOC(House of Cardinality)

Presentation Outline:

- Exploratory Data Analysis
  - Raw Data Cleaning
  - Visualization
- Model Building
  - Model Selection
    - Assumption Based
  - Feature Engineering
  - Model Implementation
    - Models used
    - Cross Validation
  - Model Validation
- Applications
  - Homeowners’ perspective
  - Kaggle Competition

EDA:

Exploratory Data Analysis:

Raw Data Cleaning:

- Handling Missing Data (Nulls)
  - Continuous impute
- Training and Test Data Synch - categorical variables

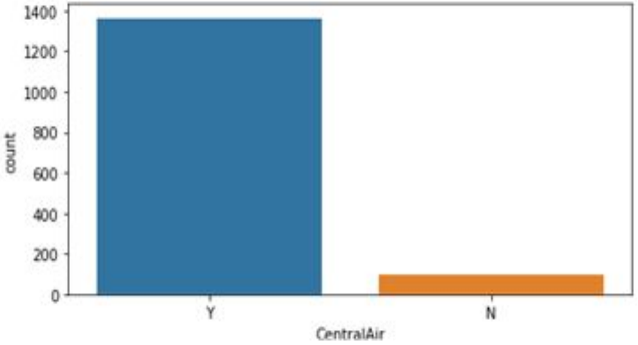
Visualization: Observe Univariate and Bivariate Relationships

- Categorical Nominal
  - Count plot
  - Barplot
  - Boxplot
- Numerical Continuous
  - Univariate Distribution- Outliers
  - Bivariate - Correlation matrix
- Categorical Numerical
  - Mixture of Categorical Nominal and Numerical Continuous

Handling Missing Data:Continuous Variables

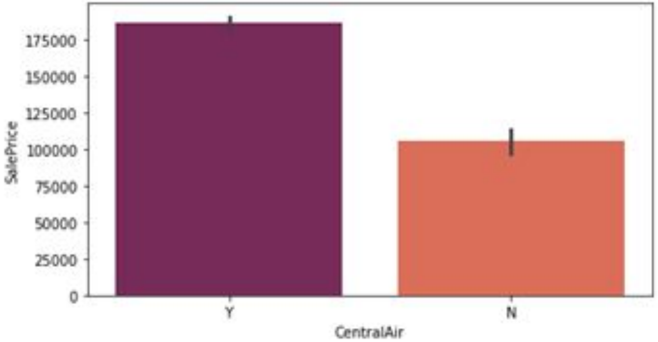
Handling Missing Data:Categorical

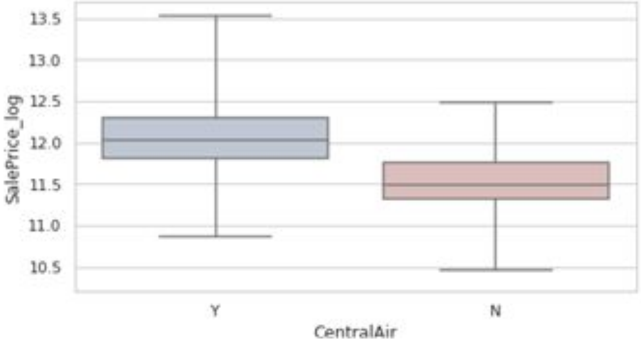
Example Problems:	Solutions:
Variables with 95%+ missing →	Drop the entire column
Basement variables with Null →	Impute with “None” given by data description
Basement related variables incorrect “None” →	Make sure each row is consistent, for example if only one basement variable is “None”, needs imputation
Garage related variables incorrect “None” →	



Countplot: Category vs Count of instances within each subcategory

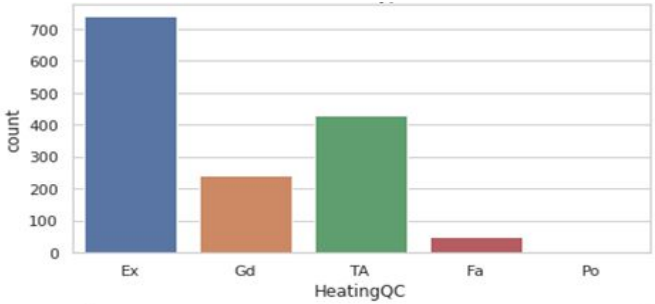
Barplot: Category vs Mean SalePrice within each subcategory



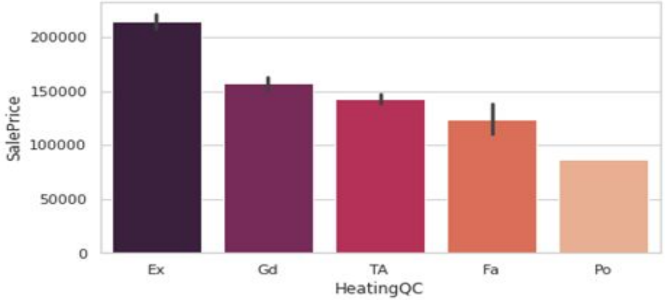


Boxplot: Category vs Distribution of Log_Price within each subcategory

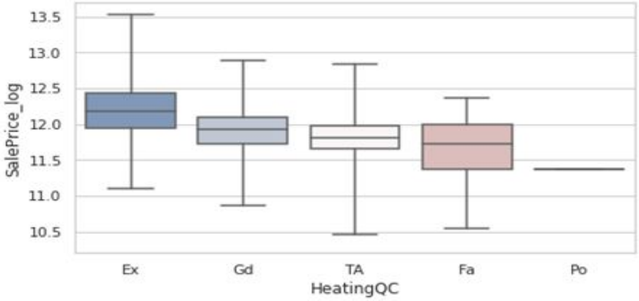
Ideal Categorical Variable: HeatingQC



- Dominant no more than 50% of data
- Subcategories are represented
- Need to examine the proportions in test set



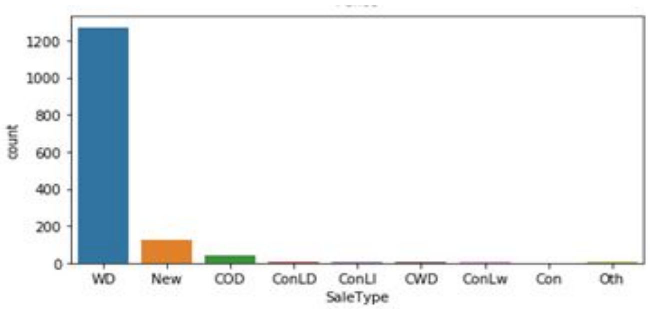
- Ordinal categorical variable subcategories exhibit sequential behavior w.r.t. Mean of SalePrice



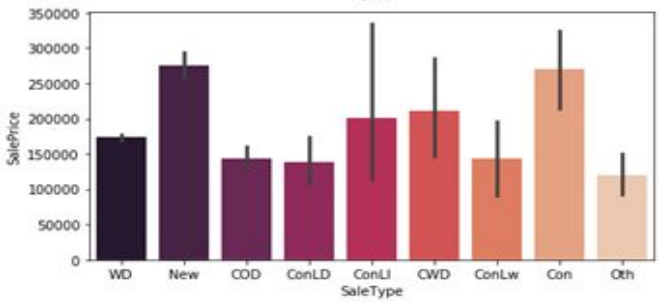
- Distributions of subcategories seem normal
- Medians of subcategories descend sequentially

Categorical Variable that needs to be Feature Engineered: SaleType

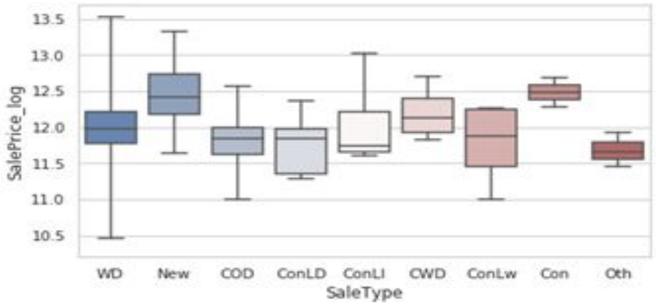
- Dominant group cover well over 50% of date
- Collapse subcategories that are underrepresented
- Need to examine the proportions in test set



- Mean of subcategories vary with no distinct grouping
- May need domain knowledge to supplement regrouping



- Distributions of subcategories are not normal
- Median Log of Sale Price of subcategories vary with no distinct grouping



Model Building

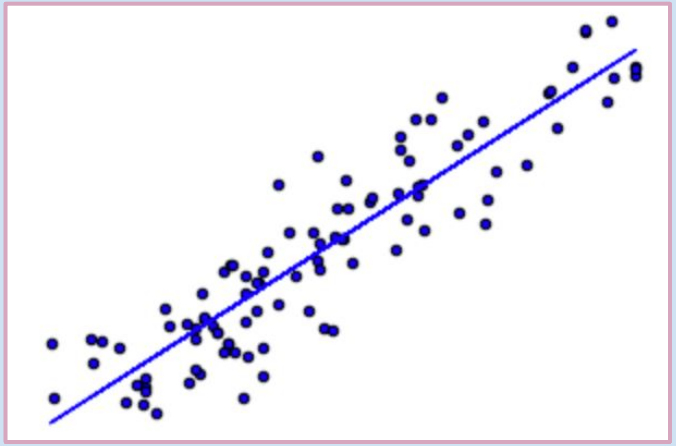
Model Building: Model Selection

Assumption 1:

- Linear “data generator”: the relationship between the independent variables and the the mean of house price is linear

Feature Engineer objectives:

- Transform Y to reinforce the normality assumption
- Transform X to avoid Multicollinearity to reinforce the independence assumption
- Trial and Error transform to improve accuracy: prepare two datasets



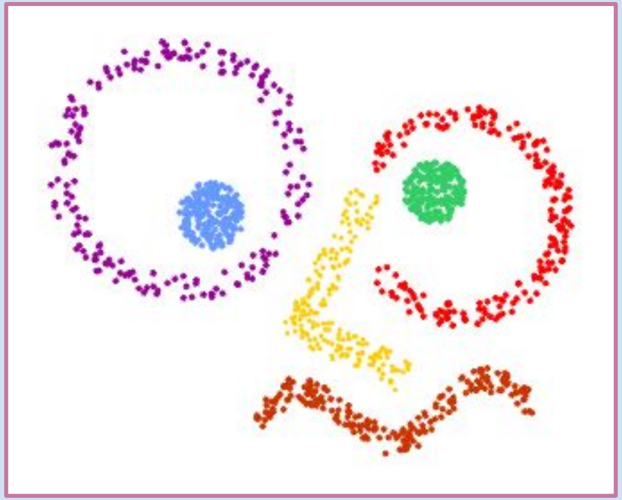
Model Building: Model Selection

Assumption 2:

- No assumption on the relationship between house price and independent variables
- Emphasis on modelling for sake of accuracy of prediction

Feature Engineer objectives:

- Prepare dataset to feed into tree based nonlinear model



Model Building: Model Implementation

- Split training set with labeled Y into test and train 80-20
- Model Building Tools:
  - R Lm library Multiple linear regression
  - R glmnet library Ridge and Lasso
  - R Xgboost
- Perform Cross Validation for Ridge and Lasso and Xgboost
  - Hyperparameter tuning

Model Building: Model Validation

Is Multiple Linear Regression sufficient to model the relationship between X and Y?

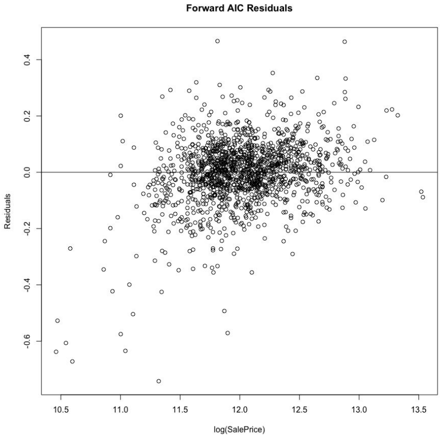
- Perform residual analysis on Multiple Linear Regression
    - Coefficient statistical significance

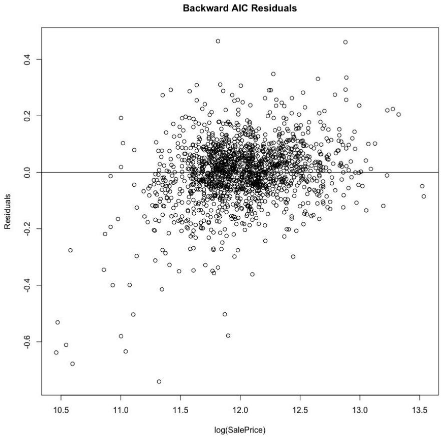
Is Linear Regression with penalized terms sufficient to model the relationship between X and Y?

- Cross Validation Approach
- AIC, BIC Criteria

Multiple Linear Regression Residual Analysis

Forward AIC vs. Backward AIC



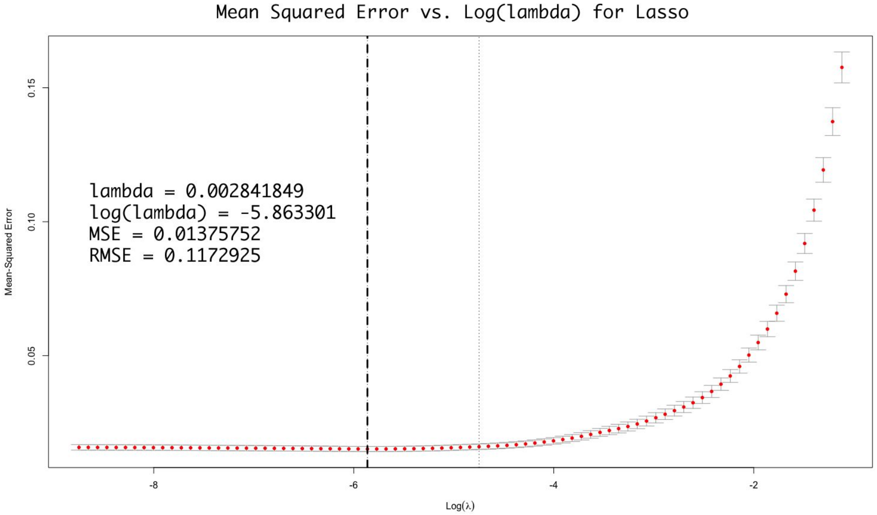


Forward: RMSE: 0.1160, 55 coefficients, R^2: 0.9116

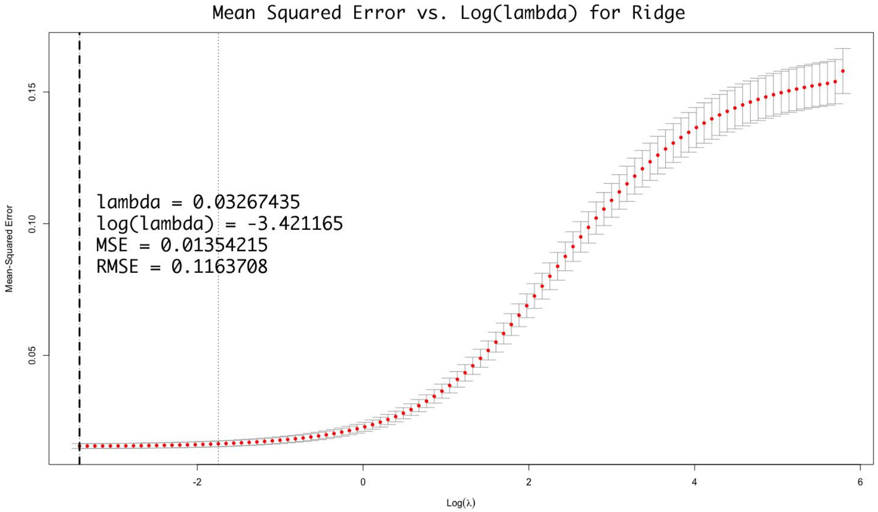
Backward: RMSE: 0.1159, 56 coefficients, R^2: 0.9117

Penalization vs Error

Lasso

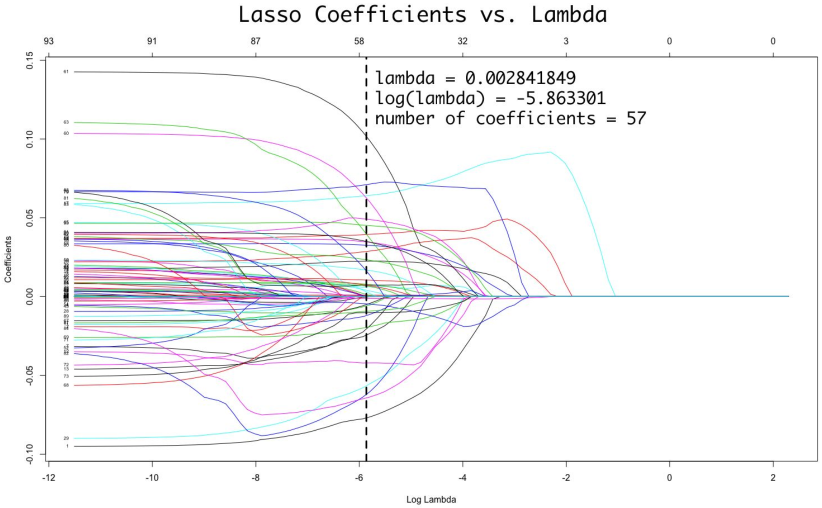


Ridge



Variables Survived

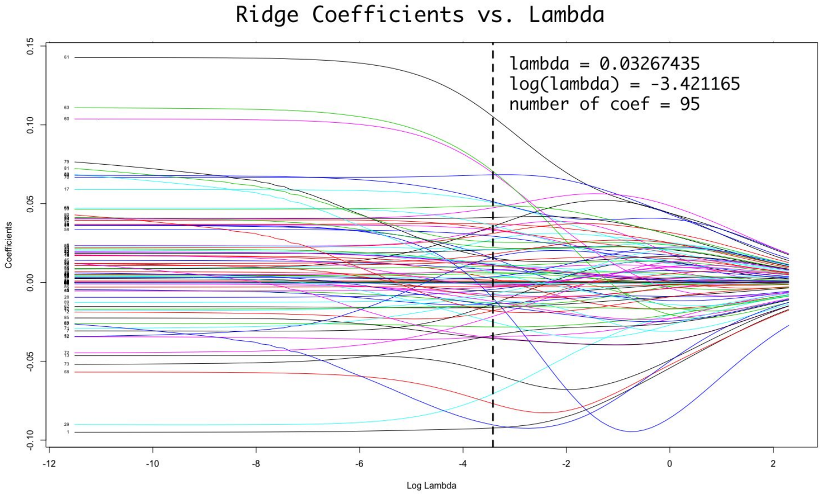
Lasso



Lasso Coefficients vs. Lambda

lambda = 0.002841849
log(lambda) = -5.863301
number of coefficients = 57

Ridge



Ridge Coefficients vs. Lambda

lambda = 0.03267435
log(lambda) = -3.421165
number of coef = 95

Model Building: Model Validation

Any violation of the multiple linear regression assumptions?

For example homoscedasticity, then we need to conduct non-parametric analysis instead of parametric modelling

XGBOOST??

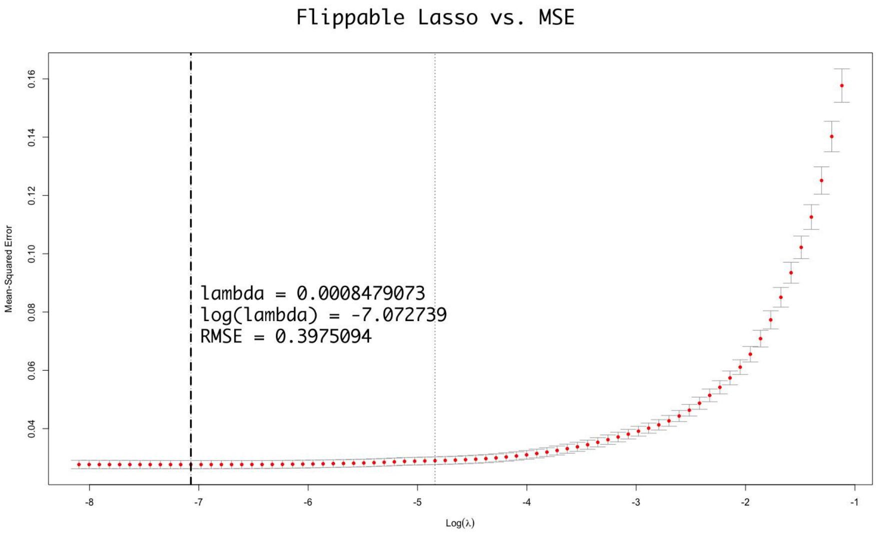
Application

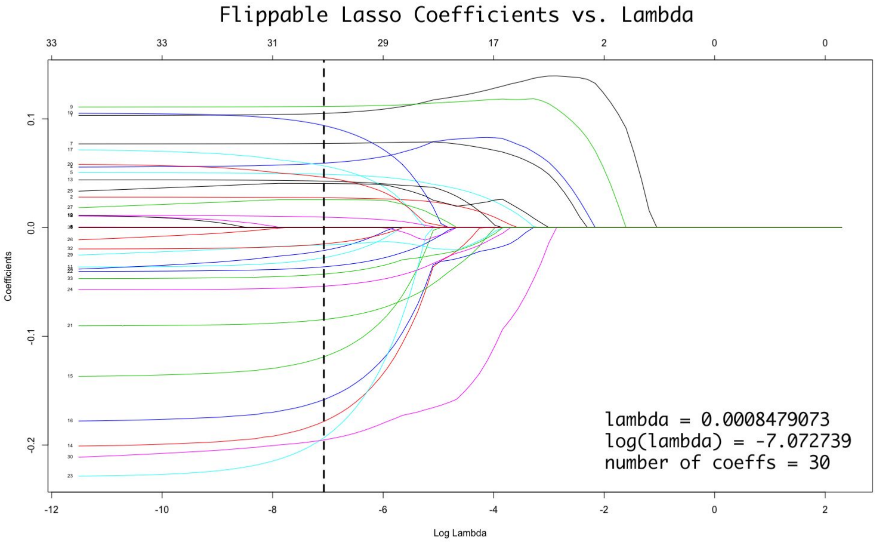
Application: Home Improvement

- What are the Features that homeowners can control to make their homes more valuable?
- Key features
  - Unfinished basement
  - Overall quality
  - Central Air
  - Gas heating
  - Number of rooms
- Base price dependent
  - Neighborhood
  - Age
  - Zoning

Lasso CV and Penalization plots: Controllable Variables Only

Of the 17 variables started with, only one (# of kitchens) was dropped: most of the variables that homeowners can improve are proven to be significant in improving sale price.





EXTRA SLIDES

Avoid Multicollinearity to reinforce the independence assumption

Example Problems:	Solutions:
Dummy variables	Remove first column after examining the means
One column is a linear combination of another column	Combine columns and remove one
When to create interaction terms??	
PLOTS???	

Able to justify each step: assumption linear relationship:

Baseline model MLR- goodness of fit and residuals (random spread)

do we have good evidence to reject the null hypothesis?

Have a model without low P-values variables?

Automatic

Forward and backward selection based on AIC BIC

Lasso penalized- decreased the error, look at bias vs variance

Normalize??

Split train test for training dataset- cross validation

- Tune the hyperparameter

Ridge coefficients will never go to 0.

Assumption nonlinear relationship:

XGBoost (tree model)-

don’t dummify anything (labels don’t matter)

Chisq test for subcateg