House Prices

Advanced Regression Techniques

W207 - Applied Machine Learning - Final Project
Yatin Majmudar, Anu Sankar, Eugene Tang
Dec 17, 2018

Introduction

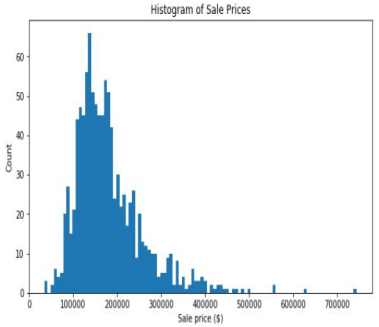
- The goal of this project is to use the features of a residential home to predict the sales price for each house (“SalePrice” variable) in the test dataset scored on the rmse(root mean square error).

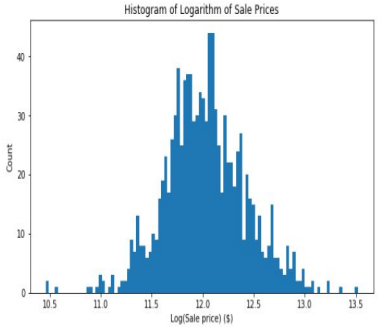
- Kaggle Competition Dataset Iowa House Prices
- The dataset contains home sales made between 2006 and 2010.
- 79 explanatory variables.
- Training dataset - 1460 rows
- Testing dataset - 1459 rows.

- *This data was originally compiled by Dean De Cock as Ames Housing dataset for use in data science education.*

Approach

- Split the given Training dataset into train (70%) and dev (30%)
- Exploratory Data Analysis
  - Missing Data
  - Explore the outcome variable - SalePrice
    - SalePrice is skewed to the right
    - Log(SalePrice) is normally distributed
  - Additional Variables
    - Age of the house (BuiltAge)
    - Years since the last remodeling (RemodelAge)
  - Correlation Matrix
    - Numeric and Ordinal Predictor variables with SalePrice and Log(SalePrice)
  - Univariate and Bivariate Analysis
    - Of numeric and ordinal predictor variables that are highly correlated with SalePrice and Log(SalePrice)
  - Apply Transformation to some predictor variables

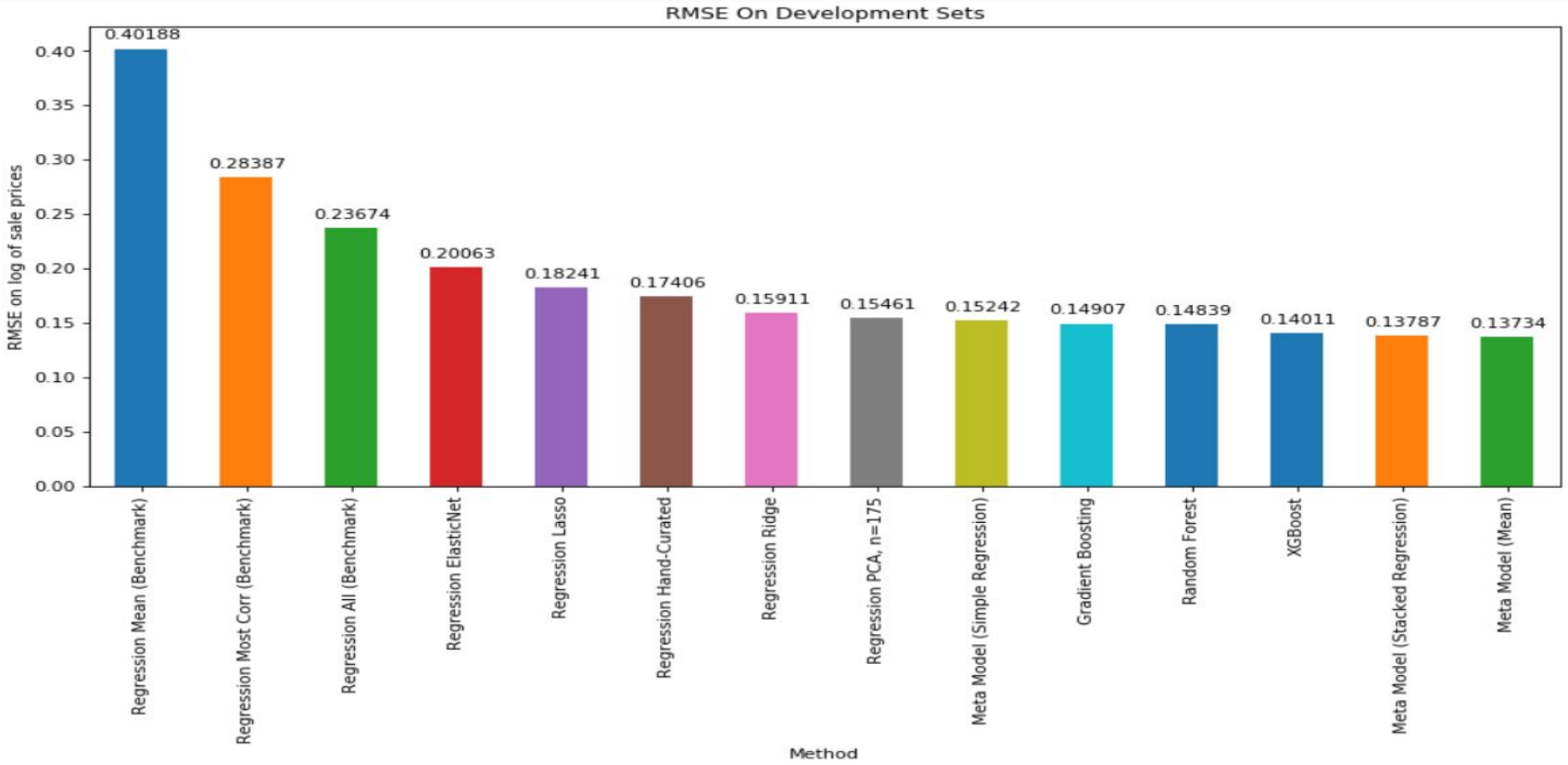




PreProcessing

- New Outcome Variable
  - Log(SalePrice)
- Transformations / Feature Engineering
  - Log(SalePrice)
  - BuiltAge - Age of the house
  - RemodelAge - Years since the last remodeling
  - Log(GrLivArea)
  - Log(LotArea)
  - Log(1stFlrSF)
- Dummy Variables
  - Created dummy variables for categorical and ordinal variable columns

Model Performance



Next Steps

- Reducing Feature Set (292 predictors, including dummies)
  - Remove features that are highly correlated to each other
    - Removing some of these provides a noticeably small improvement, but nothing major
    - LotFrontage, LotShape, GarageYrBlt, GarageCars, PoolArea, PoolQC, TotRmsAbvGrd
  - Removing features are sparsely populated
    - PoolQC, PoolArea, Alley, MiscFeatures
- Feature Engineering
  - Combine total square footage of the house (finished and unfinished)
  - Separate owned vs. rental properties
- Additional Models
  - Different models for different neighborhoods or different years (w/macroeconomic factors)
  - Neural Networks
  - SVM

Backup: After removing strongly correlated predictors