MATH 218 Final Project

Predicting Housing Prices in Ames, Iowa

Team Haydoja

Data Description

- Ames Housing Data - publicly available on Kaggle
  - https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data
- 1460 observations, 79 variables
  - Combination of continuous, categorical, and discrete features
- Example features:
  - Neighborhood: Physical locations within Ames city limits
  - SalePrice: Price of house
  - LotArea: Lot size in square feet
  - YearBuilt: Original construction date
  - FullBath: Number of full bathrooms above ground
  - OverallCond: Rates the overall condition of the house

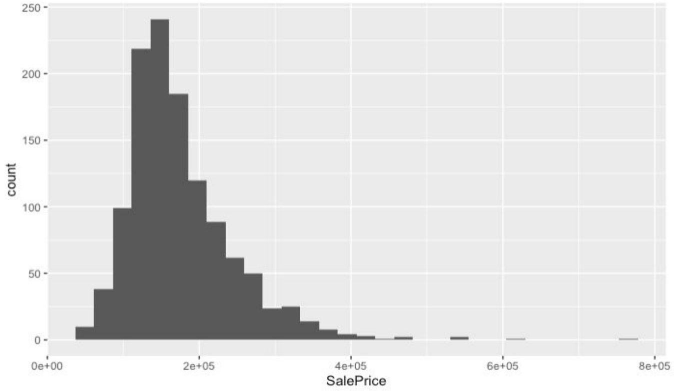
Feature Derivation / Data Cleaning

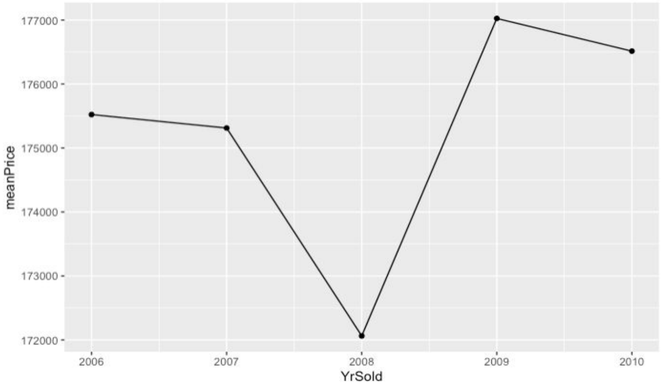
- Picked 23 variables from the 79 to focus on for the sake of simplicity

- Added log transformation to ‘SalePrice’ to account for right-skew

- Scaled numeric features using scale()

- Added binary feature ‘Sold_08’ to account for houses sold during housing market crash





Research Question 1: Which statistical learning method most accurately predicts housing prices?

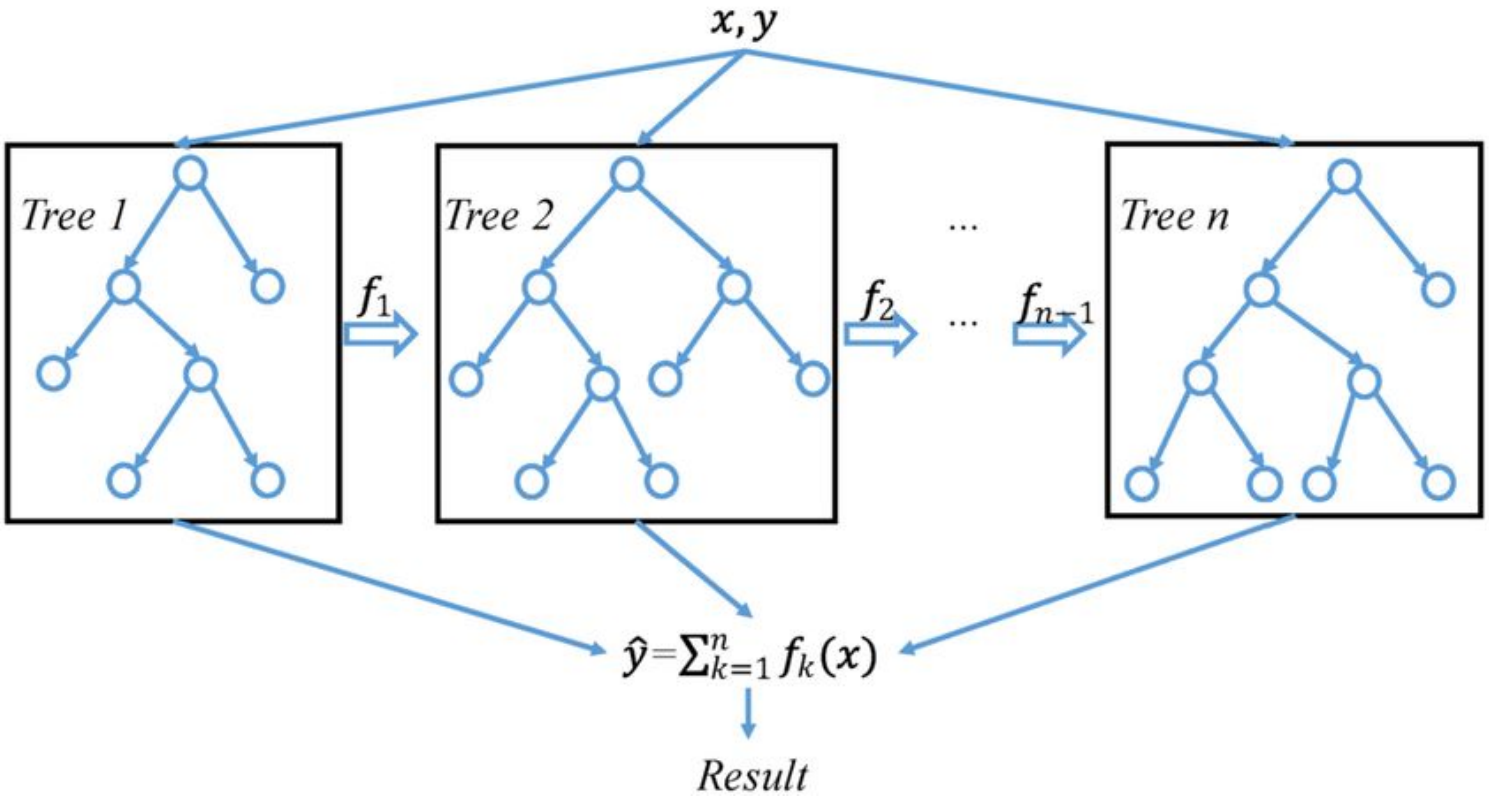
Methodology

- Models:
  - Simple Linear Regression
  - K-fold CV Linear Regression
  - Ridge Regression
  - Lasso Regression
  - XGBoost

- 80:20 train/test split
- Assessed model by measuring the RMSE on the test set

	Statistical Method	RMSE Error
	Simple Linear Regression	$17,498.74
	K-fold CV	$17,950.85
★	Ridge Regression	$16,196.87
	Lasso Regression	$17,882.54
	XGBoost	$18,810.44

XGBoost: General Architecture



Research Question 2: Which statistical learning method most accurately classifies houses by neighborhood?

Methodology

- Models:
    - Naive Bayes
    - Decision Tree
- 80:20 train/test split
- Assessed model by measuring test set misclassification rate

Statistical Method	Misclass. Error
Naive Bayes	~ 0.56
★ Decision Tree	~ 0.45

Limitations

- Limited number of observations
- Challenging to select which of the 79 available features were the most important
- RMSE changes significantly based on how we split the data (exception: k-fold CV)
- Interpreting coefficients with categorical data