Gradient Boosting Machines: A Case Study Predicting Residential Sale Prices in Washington, D.C.

Danielle Totten
University of Colorado – Denver, MATH 6388, Fall 2018

INTRODUCTION

Decision trees are a popular modeling technique

+ Can be used for both regression and classification problems
+ Structure is intuitive and results are easily interpretable
+ Non-parametric, can be used when data doesn't meet assumptions
- Low bias, high variance
- Prone to overfitting, poor predictions on test set

Gradient Boosted Machines

- Ensemble model with decision tree as base
- Combines many "weak" learners to create a "strong" learner
- Grows new trees based on the residuals of previous tree
- Key parameter: Learning rate λ
  - Maximum of 1, "faster" learner
  - Minimum approaches 0, "slower" learner
- Objective: Compare prediction error of case study test set using Gradient Boosted Machines at learning rates of $\lambda = 1, 0.1, 0.01$

CASE STUDY

Data from Kaggle [1]

- Residential housing price data from January 1st, 2014 – July 12th, 2018
- 19333 observations: 75% train, 25% test
- Outcome: Price of Residences
- 14 predictors

Figure 1: Distribution of outcome variable. Residential housing prices have a strong right skew.

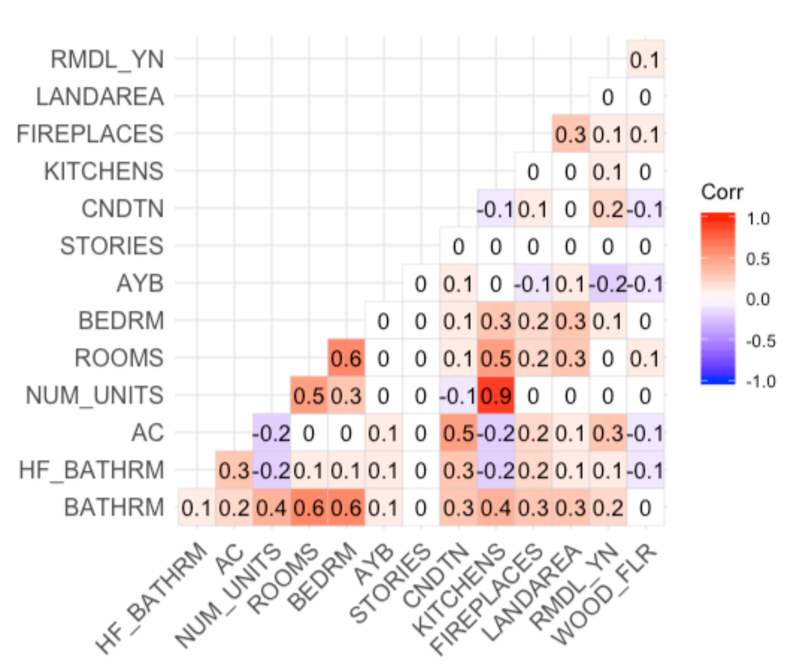

Figure 2: Correlation between predictors. Residential housing prices have a strong right skew.

METHODS

Decision Trees use recursive, binary splitting to divide observations into prediction regions

- All observations begin in a single region, called the root
- Observations are split into two new regions based on predictor X_i and cut point s
- New create largest possible reduction in RSS
- New regions are "internal nodes"

$$R_1(j,s) = \{X|X_j < s\} \text{ and } R_2(j,s) = \{X|X_j < s\}$$

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \widehat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \widehat{y}_{R_2})^2$$

- Repeated until stopping criteria met
- Final regions are terminal nodes or leaf
- All observations in leaf have same predicted outcome

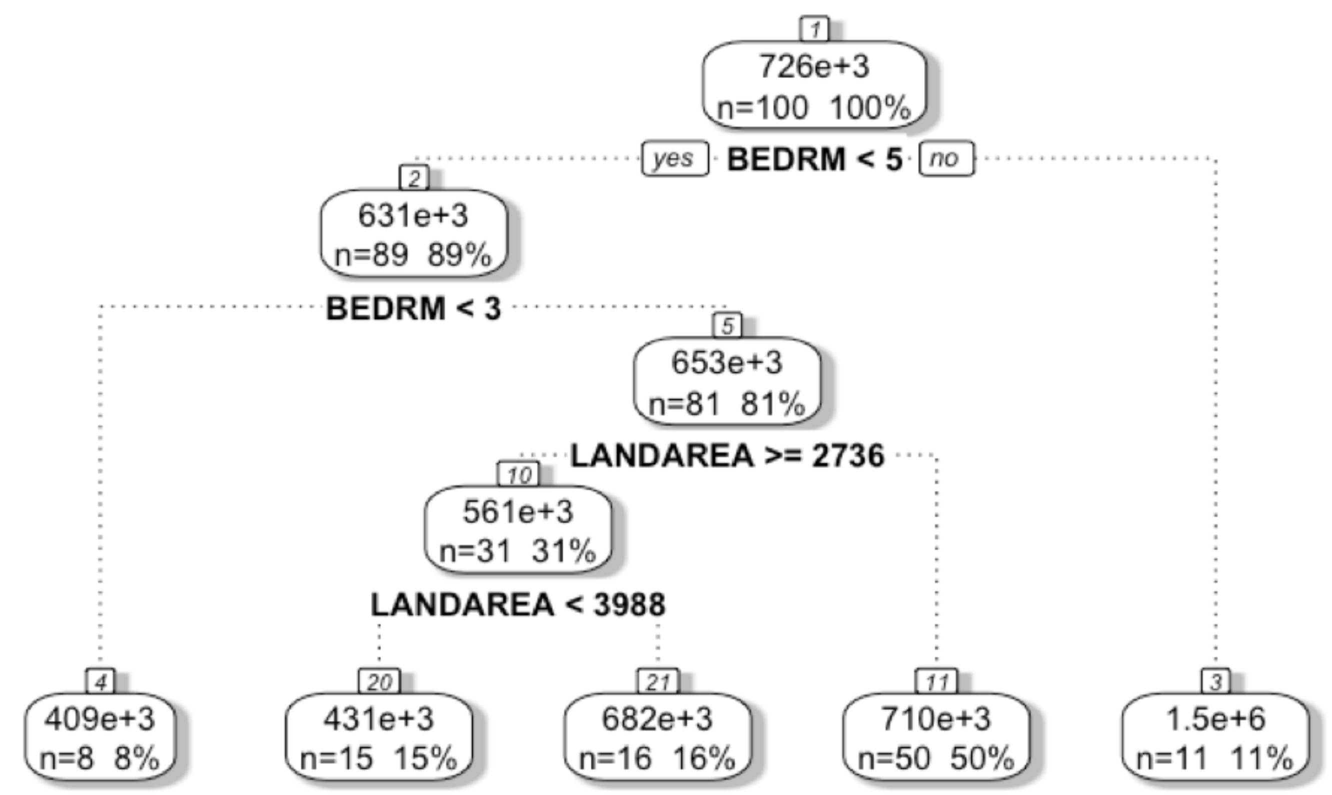

Figure 3: Example of simple decision tree. n=100 observations, 2 predictors: land area and number of bedrooms, stopping criteria: split with less than 20 observations.

Gradient Boosting improves performance over a single decision tree

- Grows m trees sequentially based on residual prediction error in terminal nodes of previous tree

$$F_M(x) = F_{M-1}(x) + \lambda\rho_M \sum_{j=1}^{J} b_{jm} 1(x \in R_{jm})$$

Application: Using Gradient Boosting Machines to predict Residence Prices in Washington, D.C.

- Grew trees based on 3 learning rates: $\lambda = 1, 0.1, 0.01$
- 5000 trees grown at each learning rate
- Trees grown with a depth of 1, no interactions between variables
- Error calculated at each iteration with RSS with negative gradient

$$-g_{im} = y_i - f_{m-1}(x_i)$$

RESULTS

Figure 4: Error reduction by number of trees at 3 learning rates.. Error is reduced more quickly for faster learners but becomes unstable as the number of trees increase.

Gradient Boosted Machines improved MSE against base model

- Highest learning rate $\lambda = 1$
  - MSE reached minimum after growing a fewer number of trees (<200)
  - Evidence of overfitting: training set MSE began increasing after minimum, MSE unsteady at higher number of trees
- Middle learning rate $\lambda = 0.1$
  - MSE decreased quickly
  - MSE remained fairly steady
- Lowest learning rate $\lambda = 0.01$
  - MSE decreased slowly
  - MSE had not reached minimum at 5000 trees

CONCLUSIONS

- GBMs reduce error compared to decision trees
- Slower learners required more trees, but MSE reductions were more smooth
- Minimum MSE of lowest learning rate was not reached within 5000 trees

REFERENCES

Hastie et al. *The Elements of Statistical Learning*. Springer. 2009.

James et al. *An Introduction to Statistical Learning.* Springer. 2013.

Mostafa et al. *Learning From Data.* 2012.

Friedman, Jerome. *Greedy Function Approximation: A Gradient Boosting Machine.* Annals of Statistics. **2009**. *29,* 1189.