

Boosted Trees

Luke Kim, Wyatt Lansford, Mitchell Neat, Alex Wood

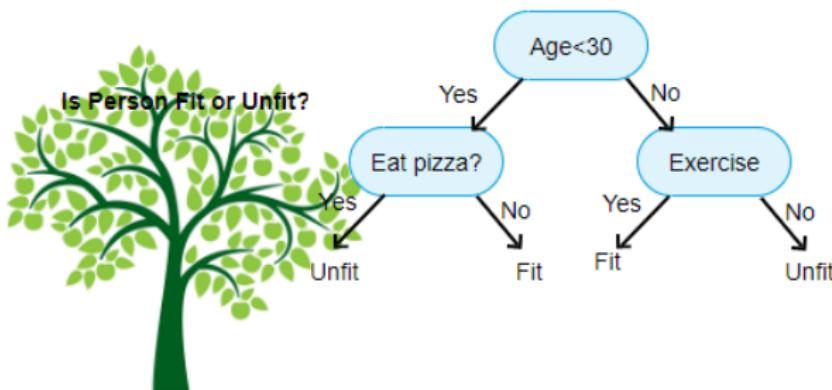
12/3/2019

Outline

- ▶ Boosted Trees
- ▶ Gradient Boosted Trees
 - ▶ Parameters
 - ▶ Math/Algorithm Details
 - ▶ Visual Examples
- ▶ Boston Crime Data
 - ▶ Regression
 - ▶ Trees
 - ▶ Gradient Boosted Trees
- ▶ Summary

What are Decision Trees?

- ▶ Take set of features and split data recursively based on features.
- ▶ Splits are chosen to minimize entropy.



What is Boosting?

- ▶ Combining weak learners (trees) into strong classifiers.
- ▶ Key Aspects:
 - ▶ weighting data
 - ▶ misclassification
 - ▶ regularization

Combined: Boosted Decision Trees

- ▶ Iteratively adding learners (trees) to minimize this objective function of the decision tree.

$$O(x) = \sum_i L(y_i, \hat{y}_i) + \sum_t \Omega(f_t)$$

- ▶ $L(y_i, \hat{y}_i)$ is the loss function
- ▶ $\Omega(f_t)$ is the regularization function

Types of Boosted Trees

- ▶ Ada Boost
 - ▶ One of the original boosted tree methods
 - ▶ Adapts a set of user-specified weak learners to create a strong learner
- ▶ Gradient Boosting
 - ▶ Uses gradient descent to optimize decision tree
 - ▶ Builds weak learners to predict loss of the previous step
- ▶ XGBoost
 - ▶ Specific subtype of gradient boosted trees
 - ▶ “eXtreme Gradient Boosting”

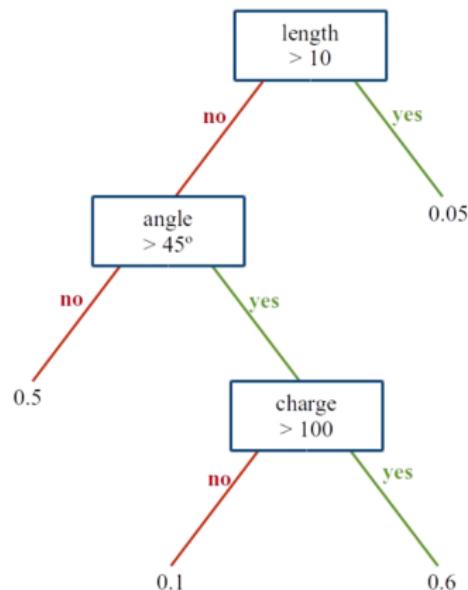
Generic Tree Parameters

- ▶ Maximum depth

- ▶ Maximum features

- ▶ Number of Trees

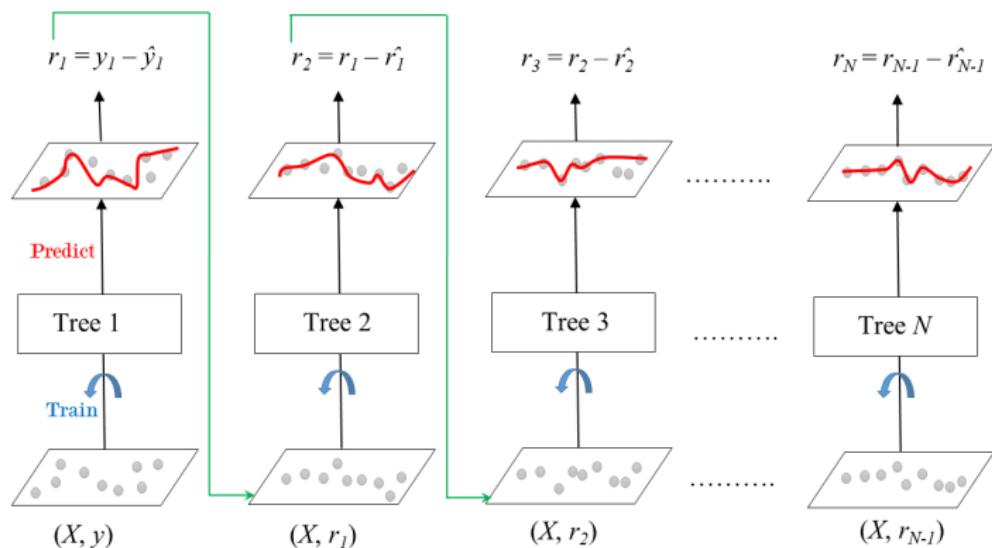
- ▶ Minimum samples per leaf



GBT Parameters

- ▶ Loss Function
- ▶ Learning Rate
- ▶ Subsample Size
- ▶ Number of Trees
 - ▶ Generic tree parameters apply here

Gradient Boosted Tree



GBT Algorithm

- ▶ Step 1: Initialize Model Values
 - ▶ Values should be constants

$$F_0(x) = \operatorname{argmin} \sum_{i=1}^n L(y_i, \gamma)$$

- ▶ Step 2: Fit Weak Learners
 - ▶ (more details on next slide)
- ▶ Step 3: Output $F_M(x)$
- ▶ Step 4: Prune (optional)

Step 2 in Detail

1. Compute pseudo-residuals

$$r_{im} = -\left[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}\right]_{F(x)=F_{m-1}(x)} \text{ for } i = 1, \dots, n.$$

2. Fit base learner (tree) to psuedo-residuals

Train the tree using training set (x_i, r_{im}) for $i = 1, \dots, n$

3. Compute multiplier γ_m by solving 1D optimization problem

$$\gamma_m = \operatorname{argmin} \sum_{i=1}^n L(y_i, F_{m-1} + \gamma h_m(x_i))$$

4. Update model

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Detailed Algorithm

Algorithm 4: Gradient tree boosting

Input : Data set \mathcal{D} .

A loss function L .

The number of iterations M .

The learning rate η .

The number of terminal nodes T

1 Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg \min_{\theta} \sum_{i=1}^n L(y_i, \theta)$;

2 **for** $m = 1, 2, \dots, M$ **do**

3 $\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}^{(m-1)}(x)}$;

4 Determine the structure $\{\hat{R}_{jm}\}_{j=1}^T$ by selecting splits which maximize
 $Gain = \frac{1}{2} \left[\frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} - \frac{G_{jm}^2}{n_{jm}} \right]$;

5 Determine the leaf weights $\{\hat{w}_{jm}\}_{j=1}^T$ for the learnt structure by
 $\hat{w}_{jm} = \arg \min_{w_j} \sum_{i \in \hat{I}_{jm}} L(y_i, \hat{f}^{(m-1)}(x_i) + w_j)$;

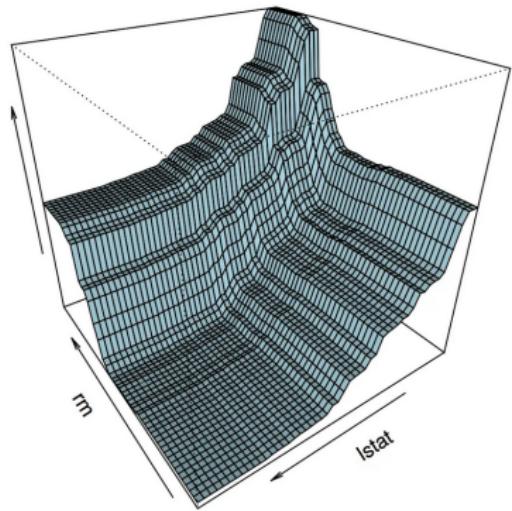
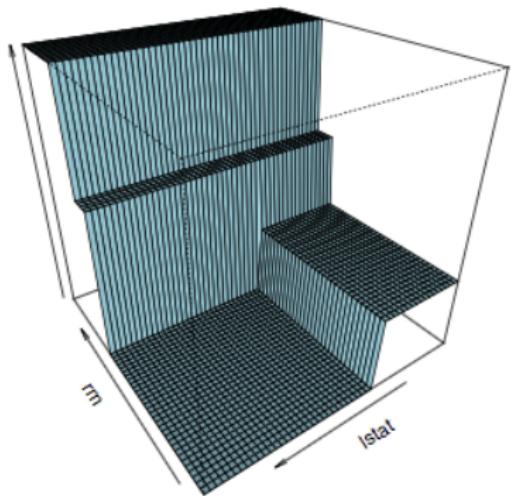
6 $\hat{f}_m(x) = \eta \sum_{j=1}^T \hat{w}_{jm} I(x_i \in \hat{R}_{jm})$;

7 $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$;

8 **end**

Output: $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^M \hat{f}_m(x)$

Visual Example



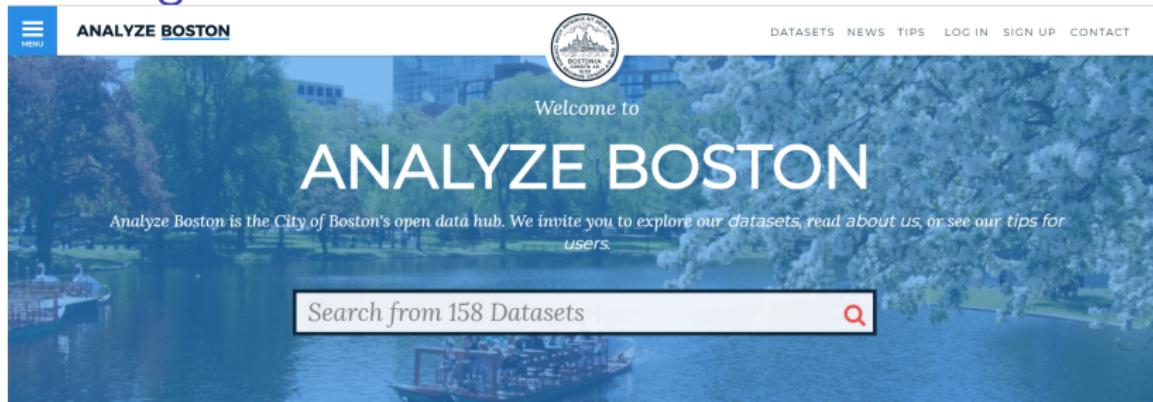
Why?

- ▶ Tree Boosting ‘beats’ the *curse of dimensionality* by not relying on any distance metric, as well as employing a large amount of inherent regularization through boosting parameters, tree parameters, and randomization parameters
 - ▶ Information becomes diluted if there are too many dimensions represented.
- ▶ Deeper trees allow capturing of interactions of features automatically.

Why?

- ▶ Boosting fits the data against the errors of the previous trees, allowing for a self-correcting of incorrect predictions
- ▶ Using optional methods such as gradient descent allows you to minimize much more complicated loss functions
 - ▶ If using a learning rate, this also introduces the concept of shrinkage
 - ▶ Shrinkage brings slower convergence and therefore generally improved performance
- ▶ Boosting trees generally perform well for both classification and regression

Introducing the Dataset



ANALYZE BOSTON

menu

DATASETS NEWS TIPS LOG IN SIGN UP CONTACT

Welcome to

ANALYZE BOSTON

Analyze Boston is the City of Boston's open data hub. We invite you to explore our datasets, read about us, or see our tips for users.

Search from 158 Datasets

A search bar with a magnifying glass icon.

	OFFENSE_DESCRIPTION	OCCURRED_ON_DATE	STREET	Lat	Long
6	AUTO THEFT - LEASED/RENTED VEHICLE	2019-10-12 15:12:43	ALLSTON ST	42.35237	-71.13510
7	AUTO THEFT	2019-10-12 04:41:52	SAINT JAMES AVE	42.34948	-71.07640
8	AUTO THEFT - LEASED/RENTED VEHICLE	2019-10-02 08:08:49	N MEAD ST	42.38185	-71.06655
9	INVESTIGATE PROPERTY	2019-09-29 06:39:00	WILMORE ST	42.27796	-71.09246
10	INVESTIGATE PERSON	2019-09-29 03:45:00	NASHUA ST	42.36769	-71.06586

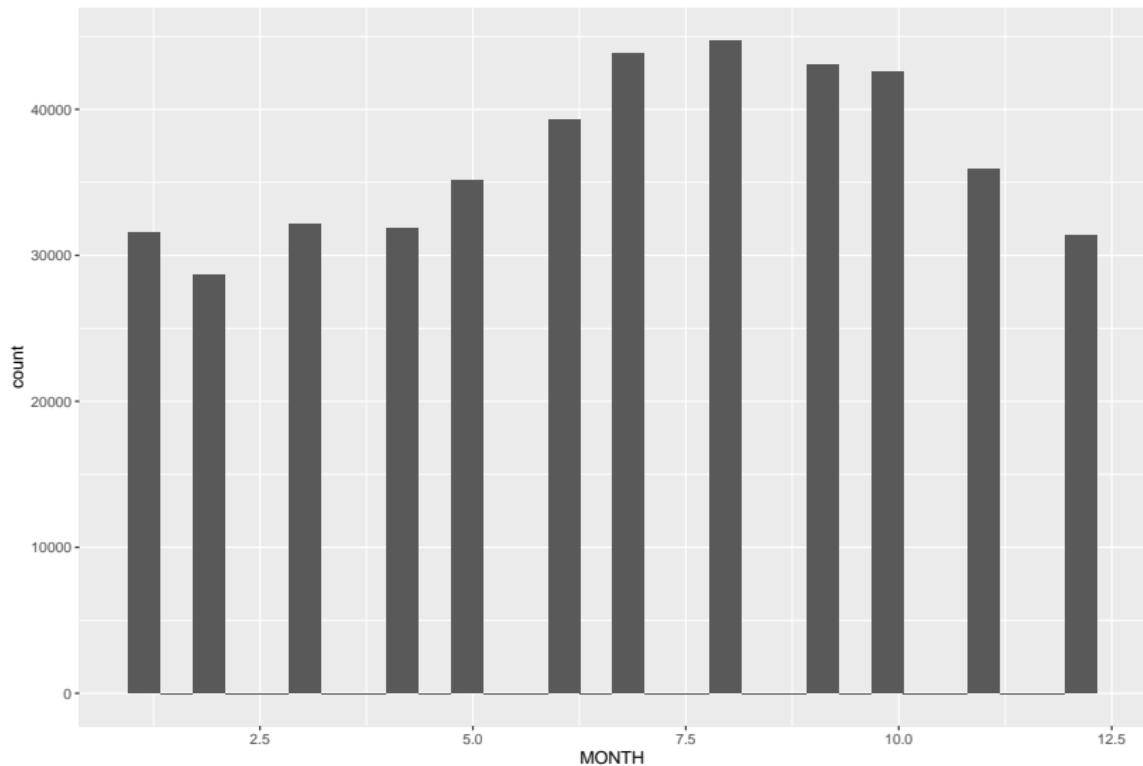
Crime incident reports provided by Boston Police Department (BPD) Incident reported from August 2015 to current are reported.

This dataset contains over 440,000 incidents. For our purpose, we will be observing reports in 2018 (98888 incidents).

Exploratory Analysis

```
## `stat_bin()` using `bins = 30`. Pick better value with
```

Histogram of Crimes Reported by Month



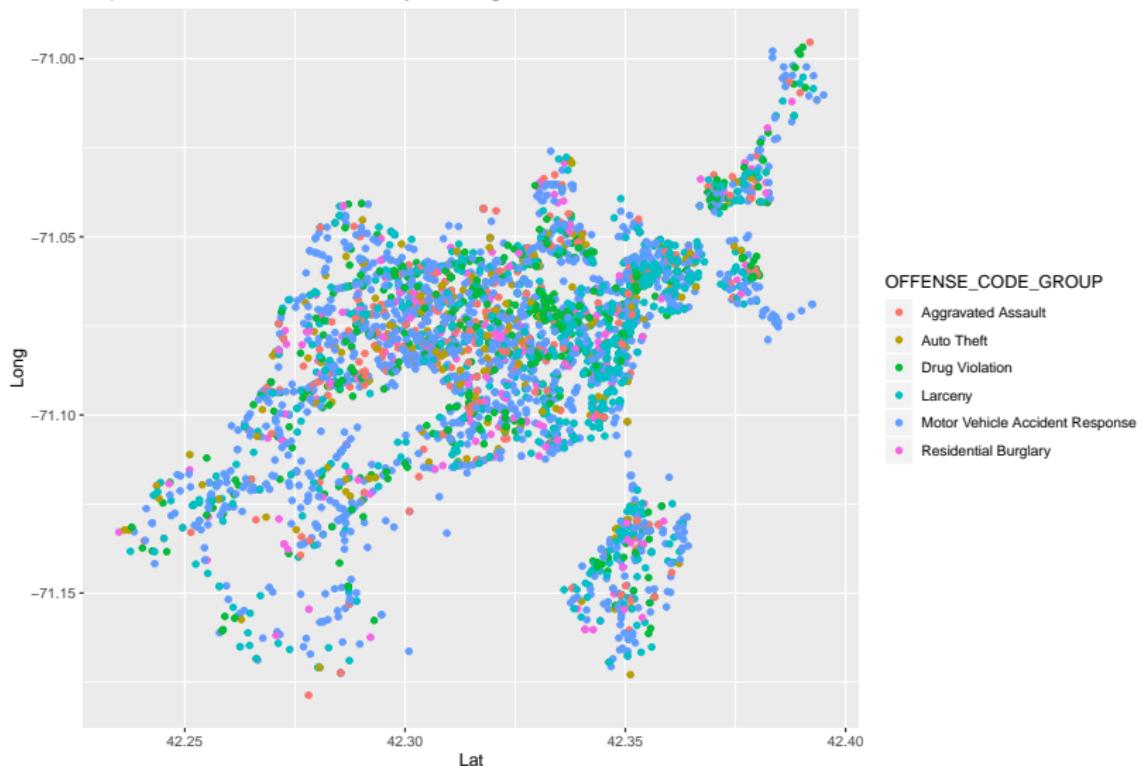
Exploratory Analysis (cont.)

There are 61 different types of offense code group. We have picked 6 interesting crimes reported in our data.

Var1	Freq
Motor Vehicle Accident Response	1905
Larceny	1376
Drug Violation	824
Aggravated Assault	396
Auto Theft	239
Residential Burglary	234

Exploratory Analysis (cont...)

Reported Crime Incident location by Lat/Long



Question we want to answer...

Could we predict what kind of offense(OFFENSE CODE GROUP)?

Regression Attempts

Trying to fit regression models to predict which crimes are being committed did not have great results. When trying to predict based on location we have an r squared of

```
## [1] 4.95167e-06
```

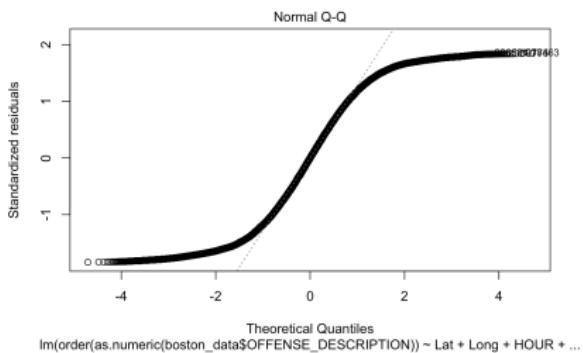
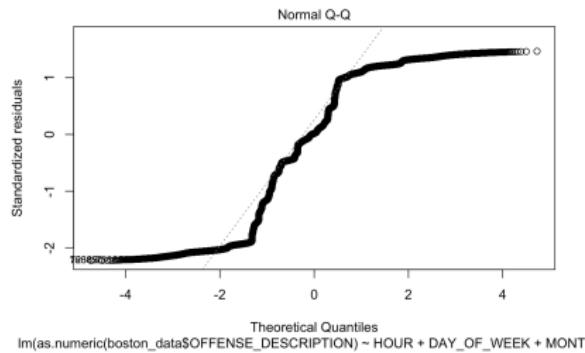
We then thought time may be a significant factor so we tried to predict type of crime based on the month, day of week, and time of day. This again was unsuccessful and resulted in a r squared value of

```
## [1] 0.0004018476
```

Finally we chose to make a model with any many factors from the data that we could. This is the best regression we could come up with and it still performed very poorly with an r squared value of

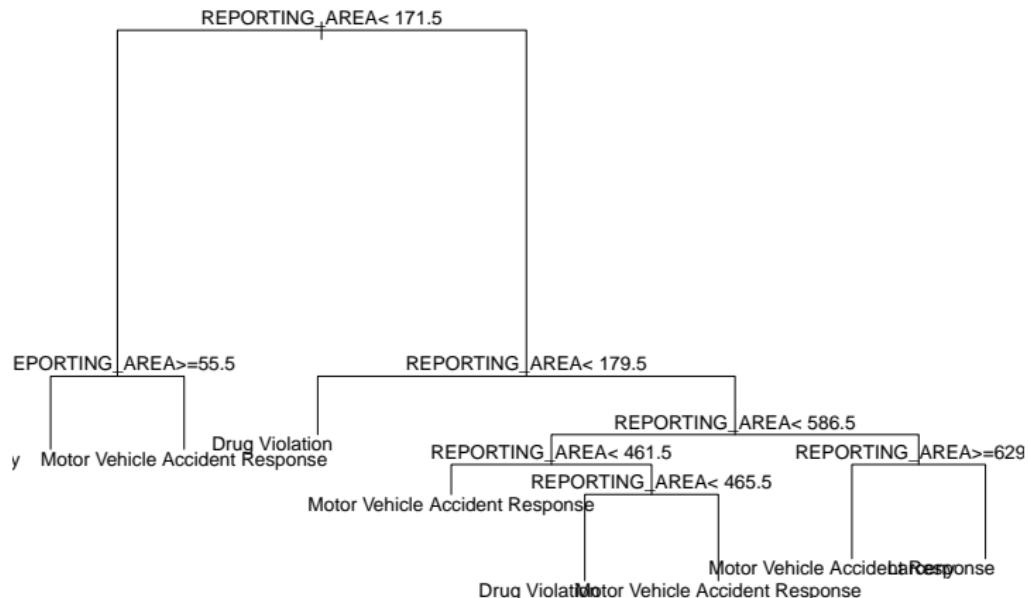
```
## [1] 0.003054367
```

Regression Attempts (cont. . .)



Regression is not the best choice for classification.

Regular Tree Fit



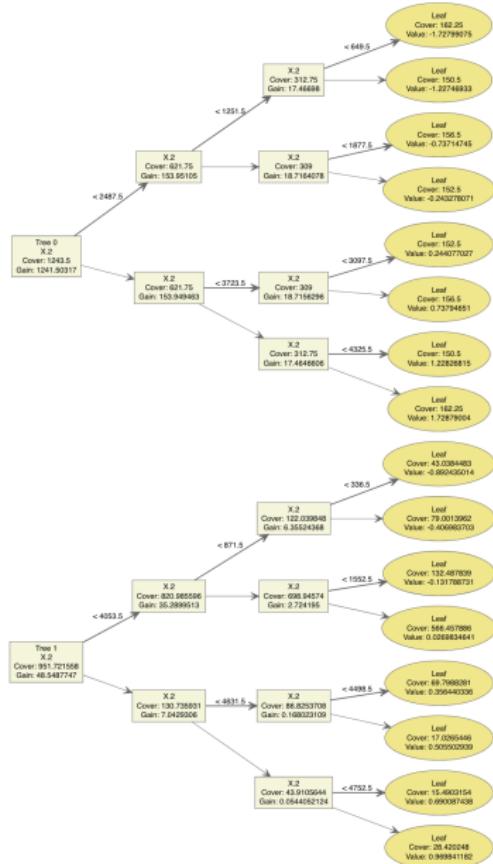
GBT Model

```
matdat <- data.matrix(test)

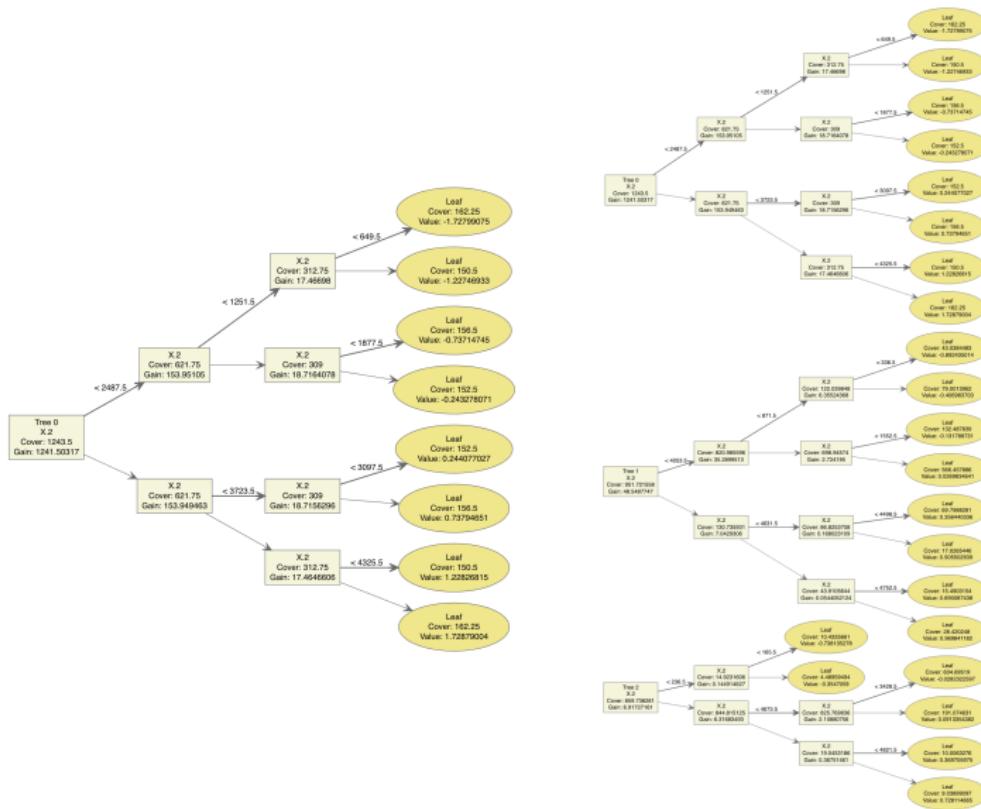
bst <- xgboost(data = matdat, label =
(test$X)/length(test$X), nrounds = 2, objective =
"binary:logistic")
```

- ▶ data: training dataset
- ▶ label: vector of response values ([0:1])
- ▶ nrounds: max number of boosting iterations
- ▶ objective: returns gradients with given prediction and dtrain

GBT Plot



Changing Depth



Differences

Expected Loss of Basic Tree: 0.6170084

Expected Loss of Gradient Boosted Tree: 0.2500000

Summary

- ▶ The Boosted Tree method is a powerful machine learning algorithm used to improve prediction results from a decision tree.
- ▶ The general idea is to “compute a sequence of decision trees in which each successive tree is built for the prediction results of the preceding tree.”
- ▶ Boosted Trees are far better for classification problems than logistic regression.