

# Multi-Output Prediction Model

Bayesian additive regression tree (BART)

- Univariate, nonparametric prediction tool
- Sum of T 'weak learner' trees
- Dynamically learns important predictors via a sparsity inducing prior

Mathematically, a BART model looks like:

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \sum_{t=1}^T g(x_i; \tau_t, M_t)$$

# Multi-Output Prediction Model

## Shared Forest

- An extension of BART for multivariate responses
- Correlation between responses is utilized via shared tree structures
- Variables that predict  $Y_1$  are likely the same variables that predict  $Y_2$  (though the nature of the relationship may be different!)
- Information sharing  $\Rightarrow$  better predictions for  $Y_1$  and  $Y_2$
- Can model binary and/or continuous outcomes

# Multi-Output Prediction Model

1 binary, 1 continuous

$$Y_i \sim N(\mu_i, \sigma_i^2) \quad (1)$$

$$\delta_i \sim \text{Bernoulli}(\pi_i) \quad (2)$$

- $\begin{pmatrix} \mu_i \\ \pi_i \end{pmatrix}$  modeled jointly using a shared forest model

$$\begin{pmatrix} \mu_i \\ \Phi^{-1}(\pi_i) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^T g(x_i; \tau_t, M_t^\mu) \\ \sum_{t=1}^T g(x_i; \tau_t, M_t^\theta) \end{pmatrix}$$

- Likelihood involves two dimensional response:  $\{Y_i, \delta_i\}_{i=1}^N$

# Multi-Output Prediction Model

2 binary responses

The model structure is the same as described before; however, the likelihood reflects that:

$$\delta_{1i} \sim \text{Bernoulli}(\pi_{1i}) \quad (3)$$

$$\delta_{2i} \sim \text{Bernoulli}(\pi_{2i}) \quad (4)$$

As before, the tree structures will be built using information shared between  $\delta_1$  and  $\delta_2$ , while the location parameters are estimated separately.

# Simulation Study

1 binary, 1 continuous

We run 100 simulations. In each:

- $n_{train} = 500$ ,  $n_{test} = 500$
- We set the number of covariates to  $P = 150$ .
- $x_1, \dots, x_{150} \sim Unif(0, 1)$
- True underlying means based on a modification of the “Friedman function”

$$y_i = 10 \sin(\pi x_{1i} x_{2i}) + 20(x_{3i} - 0.5)^2 + 10x_{4i} + 5x_{5i} + \epsilon_i^Y$$

$$\delta_i = \begin{cases} 1 & \text{if } 5 \sin(\pi x_{1i} x_{2i}) + 25(x_{3i} - 0.5)^2 + 5x_{4i} + 10x_{5i} + \epsilon_i^\delta > 0 \\ 0 & \text{otherwise} \end{cases}$$

where  $(\epsilon_i^Y, \epsilon_i^\delta) \stackrel{iid}{\sim} N(0, 1)$

# Simulation Study

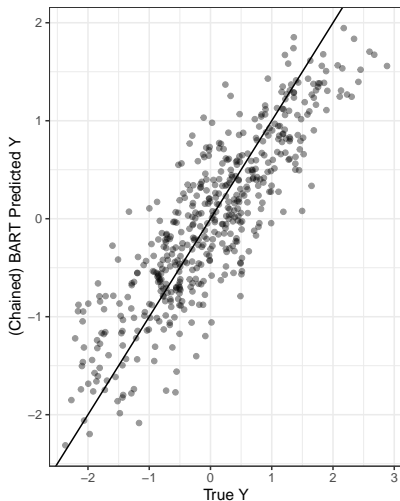
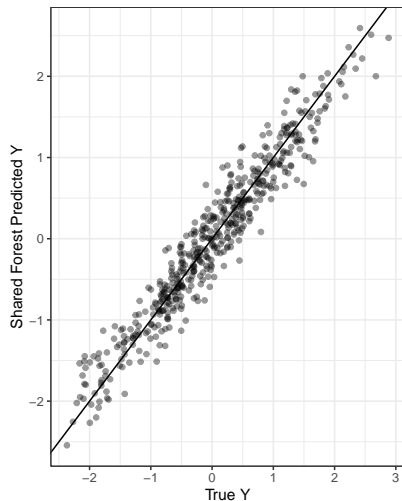
1 binary, 1 continuous

Two quantities may be of interest for prediction

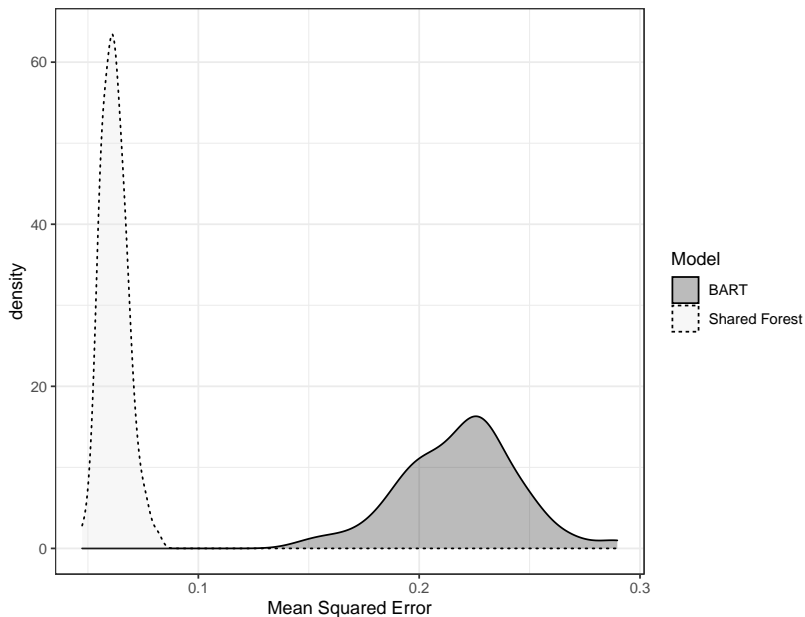
- ①  $E(y^* | x), E(\delta^* | x)$ , where  $*$  indicates those observations come from a test / hold-out sample
  - ex) Predict hours worked for individual, conditional on cellular traits  $x_i$
  - ex) Predict gender of individual, conditional on cellular traits  $x_i$
- ②  $E(y^* | \delta = 1) = \frac{1}{\sum I(\delta_i=1)} \sum_{i:\delta_i=1} E(y_i^* | x_i)$ 
  - ex) Predict mean hours worked for females (avg. across traits observed for females, i.e.,  $\delta = 1$ )
  - ex) Predict mean hours worked for males (avg. across traits observed for males, i.e.,  $\delta = 0$ )

# Predicting Individual $Y^*$

One typical simulation

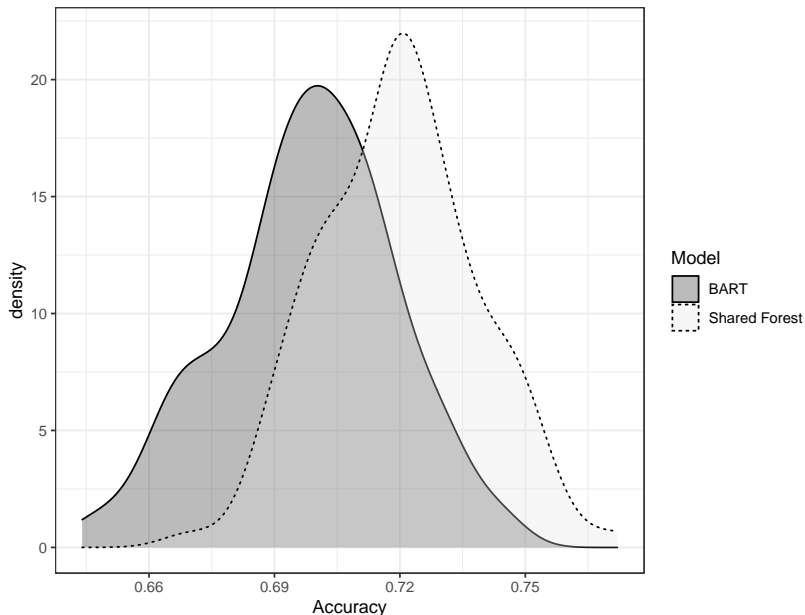


# Predicting Individual $Y^*$

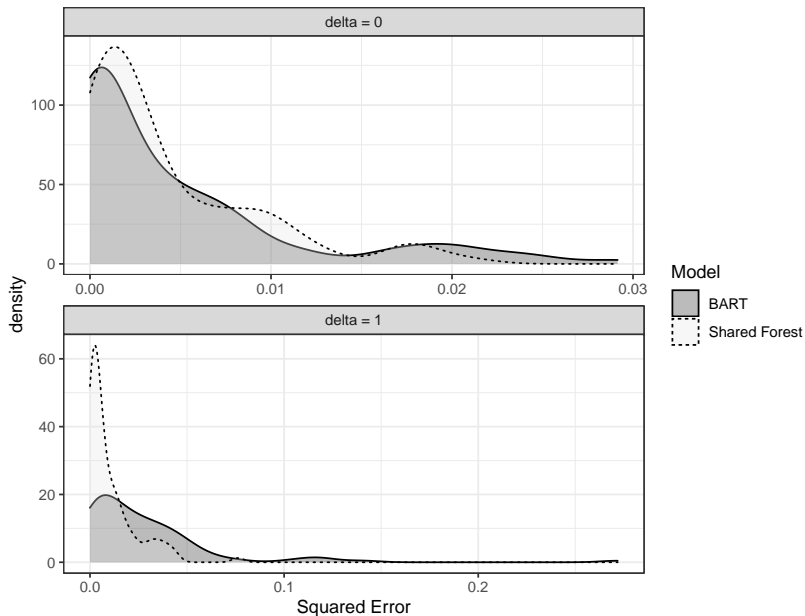




# Predicting Individual $\delta^*$



# Predicting $E(Y^* | \delta^*)$



# Predicting $E(Y^* | \delta^*)$

$\delta$	Model	MSE	L_bound	U_bound
0	BART	0.004875	0.000004	0.023207
0	Shared Forest	0.004330	0.000007	0.017697
1	BART	0.028464	0.000042	0.119264
1	Shared Forest	0.010399	0.000011	0.041591

**Table:** Mean of the squared errors (across simulations) comparing the true  $E(Y^* | \delta^*)$  to the estimated  $\hat{E}(Y^* | \hat{\delta})$ .

## 2 binary responses

- In this context, we are interested in predicting  $P(\delta_2^* = 1 \mid \delta_1^*)$ .  
ex) Probability a randomly selected female is employed.  
( $\delta_1 = \text{gender}$ ,  $\delta_2 = \text{employment}$ )
- We measure model performance by looking at the squared errors comparing the true population (conditional) means to the estimated.
- The simulation study is identical, but with two binary responses. Importantly, these responses again have differing mean functions.

# Predicting $E(\delta_2^* | \delta_1^*)$

