Objective

Develop a machine-learning algorithm that uses mobile phone data to estimate gendered patterns and gaps in key labor market indicators in Ghana.

Labor Market Indicators

Domain	Example Outcome	Outcome Type
Work	Hours worked in last 7 days	Continuous
Labor under-utilization	Under-employed in last 7 days	Binary
Informal Work	Worked informally in last 7 days	Binary
Commuting	Minutes spent during typical morning commute	Continuous

Joint Prediction

- For each indicator, we want to estimate the expected value of the indicator for men and women, or $E(y \mid \delta = 0)$ and $E(y \mid \delta = 1)$.
- Gender is unobserved, so we must jointly predict y_i and δ_i for each person i.
- The previously-used procedure is as follows:
  1. With the training data, estimate a model that predicts gender.
  2. With the training data, estimate two models that predict the indicator of interest, one model for each gender.
  3. Apply the model from (1) to the test data and apply the models in (2) depending on the predicted value from (1).
  4. Calculate gender-disaggregated labor market indicators.

Limitations

1. Theoretical justification
2. Sample splitting
3. Error propagation
4. Correlated outcomes
5. Uncertainty intervals

Single-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:

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

- $g()$ tree function which inputs x and outputs prediction
- τ_t: the t^{th} tree structure in the forest
- M_t: node parameters for t^{th} tree

Multi-Output Prediction Model

Shared Forest

- An extension of BART for multivariate responses
- Correlation between responses is utilized via shared tree structures
- Assumption: variables that predict Y_1 are 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

The shared forest package accomodates heterogeneous outcomes.

ex) Y = hours worked, δ = gender

$$Y_i \sim N(\mu_i, \sigma_i^2)$$
$$\delta_i \sim Bernoulli(\pi_i)$$

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

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

where Φ^{-1} is the probit link function.

Multi-Output Prediction Model

1 binary, 1 continuous

- "Information sharing": likelihood for τ_t involves two dimensional response: $\{Y_i, \delta_i\}_{i=1}^N$
- Algorithm's decisions about tree structure (τ_t) considers what is beneficial to both Y and δ
- Predictions of δ are improved by utilizing information provided by Y, vice versa

Multi-Output Prediction Model

2 binary responses

In our application, we may have two binary responses

ex) δ_1 = gender, δ_2 = employed

$$\delta_{1i} \sim Bernoulli(\pi_{1i}) \tag{1}$$

$$\delta_{2i} \sim Bernoulli(\pi_{2i}) \tag{2}$$

As before, the tree structures will be built using information shared between δ_1 and δ_2, while the location parameters are estimated separately.

$$\begin{pmatrix} \Phi^{-1}(\pi_{1i}) \\ \Phi^{-1}(\pi_{2i}) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{T} g(x_i; \tau_t, M_t^1) \\ \sum_{t=1}^{T} g(x_i; \tau_t, M_t^2) \end{pmatrix}$$

Simulation Study

1 binary, 1 continuous

We run 100 simulations. In each:

- $n_{train} = 500$, $n_{test} = 500$
- Predictors: $x_1, \ldots, x_{150} \sim Unif(0, 1)$
- 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 $(\epsilon_i^Y, \epsilon_i^\delta) \overset{iid}{\sim} N(0, 1)$

Simulation Study

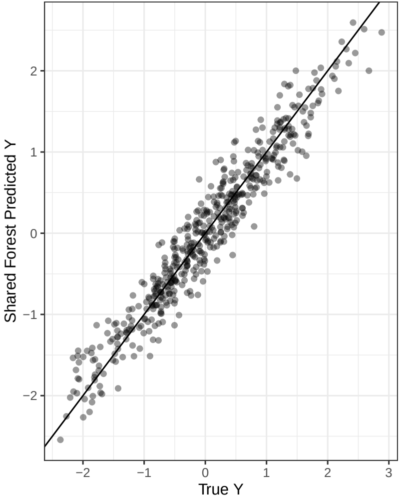
1 binary, 1 continuous

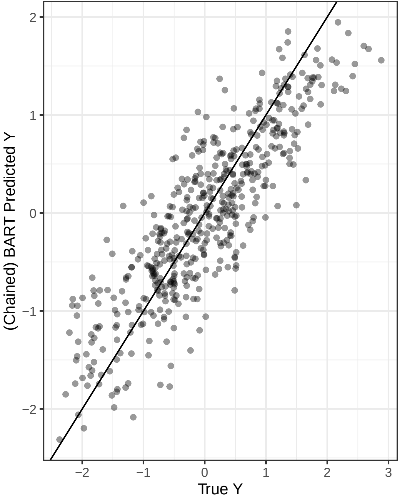
Two quantities may be of interest for prediction

1. $E(y_i^* \mid x_i)$,$E(\delta_i^* \mid x_i)$, 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. $E(y^* \mid \delta = 1) = \frac{1}{\sum I(\delta_i = 1)} \sum_{i:\delta_i=1} E(y_i^* \mid x_i)$
   - 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., $\delta = 0$)

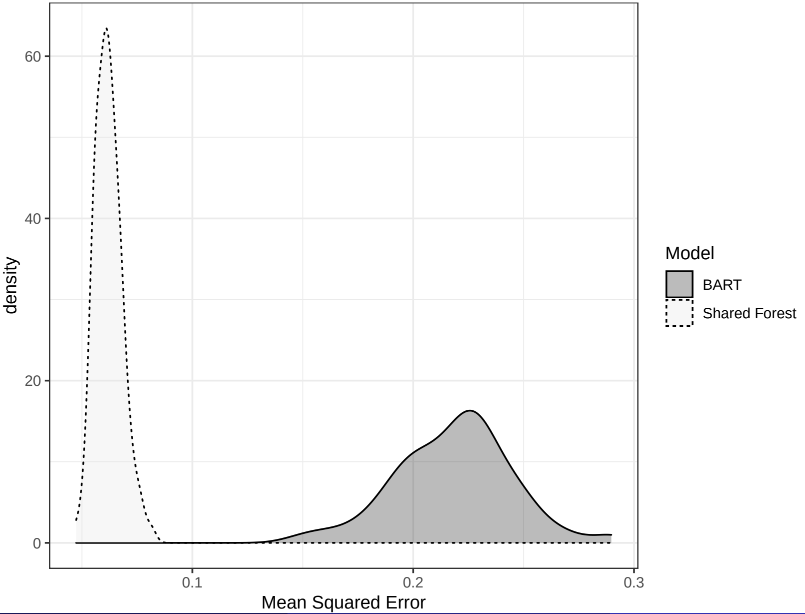
Predicting Individual Y^*

One typical simulation

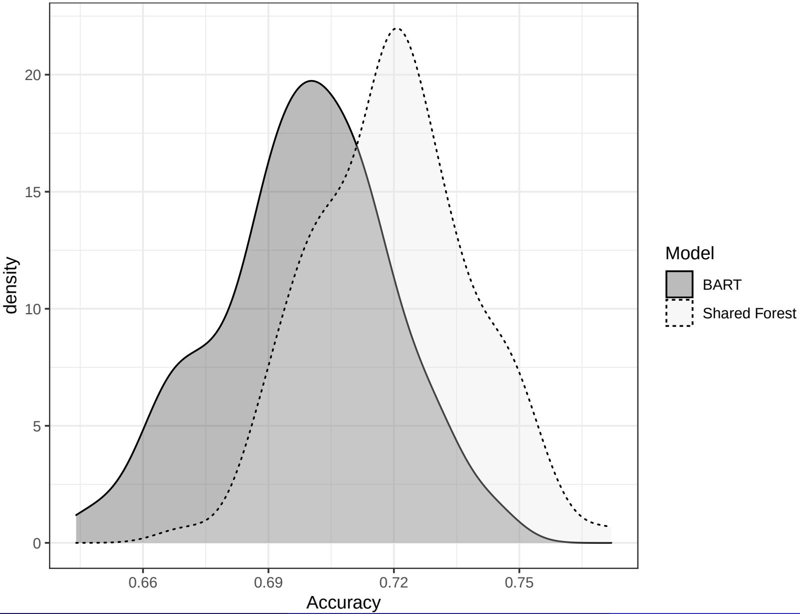




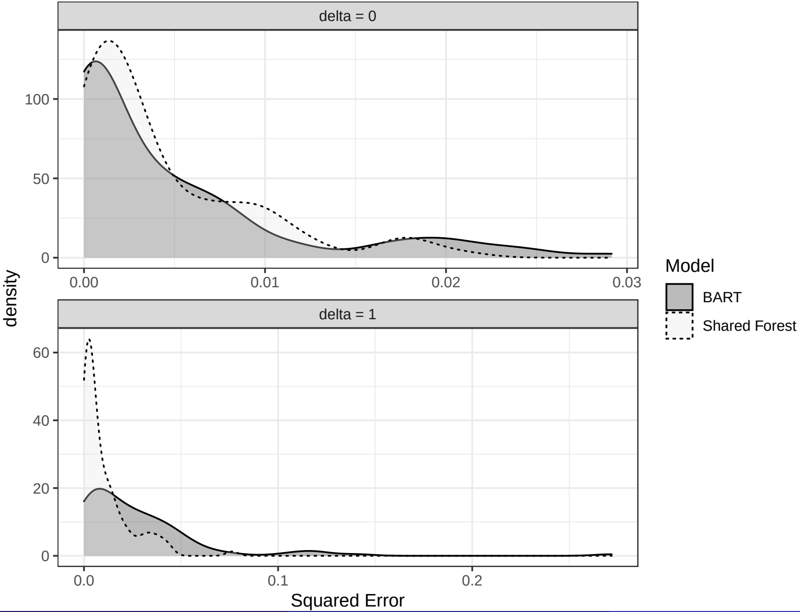
Predicting Individual Y^*



Predicting Individual $\delta*$



Predicting $E(Y^* \mid \delta^*)$



Predicting $E(Y^* \mid \delta^*)$

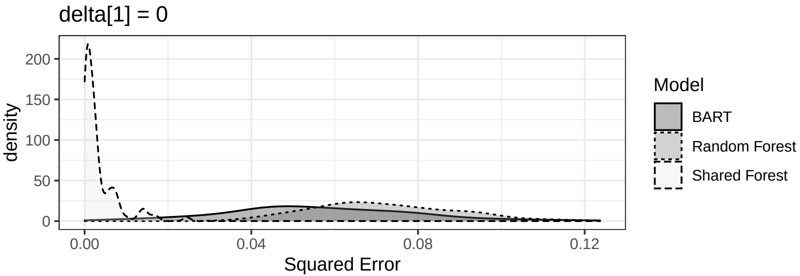
δ	Model	MSE
0	BART	0.004875
0	Shared Forest	0.004330
1	BART	0.028464
1	Shared Forest	0.010399

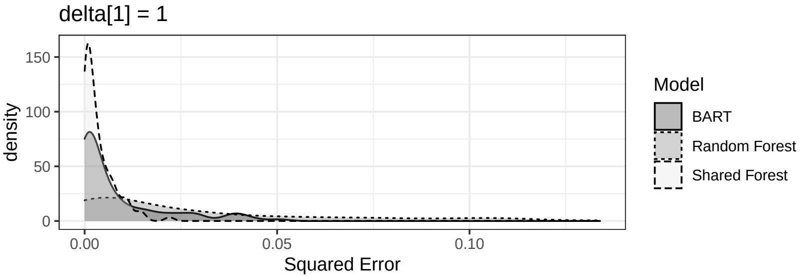
Table: Mean of the squared errors (across simulations) comparing the true $E(Y^* \mid \delta^*)$ to the estimated $\hat{E}(Y^* \mid \hat{\delta})$.

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.
    (δ_1 = gender, δ_2 = 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 $E(\delta_2^* \mid \delta_1^*)$





Predicting $E(\delta_2^* \mid \delta_1^*)$

Model	MSE	δ_1
BART	0.056580	0
Random Forest	0.071952	0
Shared Forest	0.003252	0
BART	0.008327	1
Random Forest	0.027744	1
Shared Forest	0.003514	1