Imputation of Missing Data with Bayesian Additive Regression Trees

Todd Burrows

Supervised by Emmanuel Ogundimu,
Department of Mathematics, Durham University

11th February 2026

Contents

Introduction to the Problem of Missing Data

- Historically, **missingness** assumed accidental and thus the process that caused it was ignored.[1]
- **Donald Rubin** formalised the analysis of missing data allowing considerations for **imputing** the data.
- **Multiple Imputation** proposed to capture the uncertainty caused by the missing data.[2]

[1] Rubin, 1976
[2] Rubin, 1978

Tree Based Regression

Classification and Regression Trees [3]

- Statistical **classification** or **regression** which partitions the predictor space into subgroups.
- Once built, **predicts** a response given a set of covariates.
- **Non-parametric** and can model high level **interactions** and **non-linearities**.

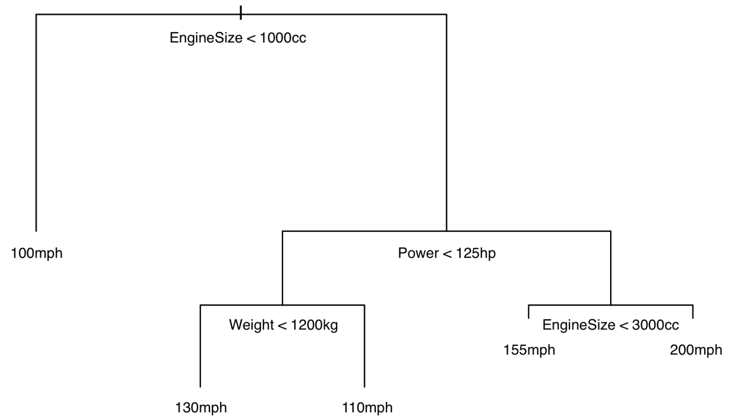

Figure. A simple regression tree to predict the top speed of a vehicle, given the engine size, power and weight.

[3] Formalised by Breiman et al., 1984

Bayesian Additive Regression Trees

An Introduction to BART [4]

- BART utilises a **sum-of-trees** structure, $Y = \sum_{j=1}^{m} g(\boldsymbol{X}; T_j, M_j) + \epsilon$,

  where for tree j
  - T_j is the **tree structure**.
  - $M_j = \{\mu_{1j}, \mu_{2j}, ..., \mu_{b_j j}\}$ is the set of b_j **terminal parameters**.
- Trees built via a **Bayesian Backfitting** algorithm.
- Can form the basis of an **imputation** model.

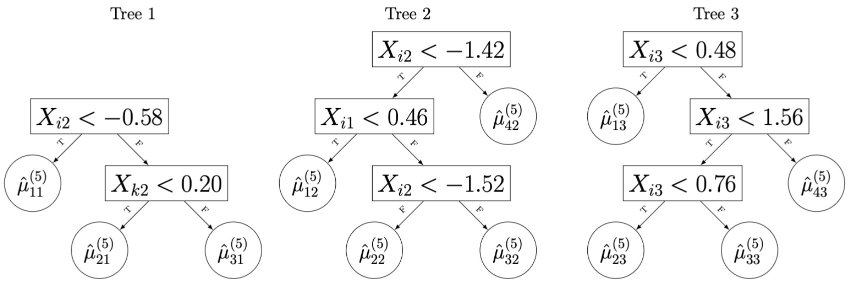

Figure. A sum-of-trees structure. Extracted from Tan and Roy, 2019

[4] Chipman et al., 2010

- Strong influential **regularising** priors guide the tree building process, tailoring the fit.

- **The Tree Prior** – $p(T_j)$.
- **The Terminal Parameter Prior** – $p(\mu_{ij} \mid T_j)$.
- **The Error Variance Prior** – $p(\sigma)$.

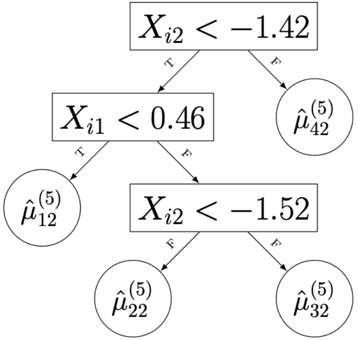

Figure. A single tree. Extracted from Tan and Roy, 2019

The BART Process

Tree Building

- The **backfitting** algorithm draws each tree (T_j, M_j) conditional on $\left[(T_{(j)}, M_{(j)}), \sigma\right]$.
- The conditional distribution of the j^{th} tree depends only on the other trees and the training data via,

$$R_j \equiv y - \sum_{k \neq j} g(\boldsymbol{x}; T_k, M_k).$$

- Tree design proposed iteratively by a **stochastic search** procedure that repeatedly perturbs the trees in one of four ways.

- **Grow**
- **Prune**
- **Change**
- **Swap**

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: **for** each column, c in k **do**
6: Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: **end for**
10: Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	–	–
86.4	3.10	–	4.24	–
–	–	6.98	5.25	74
70.5	4.53	7.15	4.73	85
–	4.40	12.00	3.01	–
65.0	4.92	–	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	–	–
–	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

MISSBART IMPUTATION METHOD

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

```
Initialise the missing values with a simple mean imputation
k ← vector of the indices of the columns in X in order of increasing missingness
while not γ do
    X_old^imp ← current imputed matrix
    for each column, c in k do
        Fit a BART model: y_obs^(c) ~ x_obs^(c)
        Predict y_mis^(c) using x_mis^(c)
        X_new^imp ← new imputed matrix, with updated y_mis^(c)
    end for
    Update stopping criterion γ
end while
return the imputed matrix X^imp
```

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	**4.57**	**77.7**
86.4	3.10	**5.91**	4.24	**77.7**
72.4	**4.38**	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	**77.7**
65.0	4.92	**5.91**	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	**4.57**	**77.7**
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: $\quad\quad$ Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\quad\quad$ $\boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	**4.57**	**77.7**
86.4	3.10	**5.91**	4.24	**77.7**
72.4	**4.38**	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	**77.7**
65.0	4.92	**5.91**	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	**4.57**	**77.7**
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: $\quad\quad$ Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\quad\quad$ $\boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	**4.53**	7.07	**4.57**	**77.7**
86.4	**3.10**	**5.91**	4.24	**77.7**
72.4	**4.38**	6.98	5.25	74
70.5	**4.53**	7.15	4.73	85
72.4	**4.40**	12.00	3.01	**77.7**
65.0	**4.92**	**5.91**	5.87	116
70.5	**4.53**	0.98	7.31	81
79.6	**4.02**	3.82	**4.57**	**77.7**
72.4	**4.40**	9.02	5.33	42
64.6	**4.95**	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: $\quad\quad$ Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\quad\quad$ $\boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	**4.57**	**77.7**
86.4	3.10	**5.91**	4.24	**77.7**
72.4	**4.92**	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	**77.7**
65.0	4.92	**5.91**	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	**4.57**	**77.7**
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: $\quad\quad$ Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\quad\quad \boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	**4.57**	**77.7**
86.4	3.10	**5.91**	4.24	**77.7**
72.4	**4.92**	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	**77.7**
65.0	4.92	**5.91**	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	**4.57**	**77.7**
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}_{\text{old}}^{\text{imp}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}_{\text{obs}}^{(c)} \sim \boldsymbol{x}_{\text{obs}}^{(c)}$
7: $\quad\quad$ Predict $\boldsymbol{y}_{\text{mis}}^{(c)}$ using $\boldsymbol{x}_{\text{mis}}^{(c)}$
8: $\quad\quad \boldsymbol{X}_{\text{new}}^{\text{imp}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}_{\text{mis}}^{(c)}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	4.57	77.7
86.4	3.10	5.91	4.24	77.7
72.4	4.92	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	77.7
65.0	4.92	5.91	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	4.57	77.7
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: $\quad\quad$ Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\quad\quad$ $\boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	**4.57**	**77.7**
86.4	3.10	**6.98**	4.24	**77.7**
72.4	**4.92**	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.4	4.40	12.00	3.01	**77.7**
65.0	4.92	**4.10**	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	**4.57**	**77.7**
72.4	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

missBART Imputation Method

- Based on the popular **missForest** method of **Stekhoven and Bühlmann, 2012**.

missBART Algorithm

Require: $\boldsymbol{X}$ an $n \times p$ matrix, stopping criterion γ

1: Initialise the missing values with a simple mean imputation
2: $\boldsymbol{k} \leftarrow$ vector of the indices of the columns in $\boldsymbol{X}$ in order of increasing missingness
3: **while** not γ **do**
4: $\quad \boldsymbol{X}^{\text{imp}}_{\text{old}} \leftarrow$ current imputed matrix
5: $\quad$ **for** each column, c in k **do**
6: $\quad\quad$ Fit a **BART** model: $\boldsymbol{y}^{(c)}_{\text{obs}} \sim \boldsymbol{x}^{(c)}_{\text{obs}}$
7: $\quad\quad$ Predict $\boldsymbol{y}^{(c)}_{\text{mis}}$ using $\boldsymbol{x}^{(c)}_{\text{mis}}$
8: $\quad\quad$ $\boldsymbol{X}^{\text{imp}}_{\text{new}} \leftarrow$ new imputed matrix, with updated $\boldsymbol{y}^{(c)}_{\text{mis}}$
9: $\quad$ **end for**
10: $\quad$ Update stopping criterion γ
11: **end while**
12: **return** the imputed matrix $\boldsymbol{X}^{\text{imp}}$

Table. Data based on the Theoph dataset [R Core Team, 2024].

Y	X_1	X_2	X_3	X_4
70.5	4.53	7.07	**5.16**	**40.2**
86.4	3.10	**6.98**	4.24	**88.4**
65.2	**4.92**	6.98	5.25	74
70.5	4.53	7.15	4.73	85
72.7	4.40	12.00	3.01	**11.7**
65.0	4.92	**4.10**	5.87	116
70.5	4.53	0.98	7.31	81
79.6	4.02	3.82	**8.33**	**12.2**
71.1	4.40	9.02	5.33	42
64.6	4.95	0.25	0.85	68

- **missForest/BART** produces a single complete dataset.
- Is this a disadvantage?

Discussion and Research

Imputation Models

- **missForest/BART** produces a single complete dataset.
- Is this a disadvantage?
- Yes – not multiple imputation! [5]

[5] **van Buuren, 2018**

Discussion and Research

Imputation Models

- **missForest/BART** produces a single complete dataset.
- Is this a disadvantage?
- Yes – not multiple imputation! [6]

- Bootstrap the data and produce multiple complete datasets.
- Compare missForest vs missBART vs both bootstrapped.
- Utilise BART's intrinsic **Bayesian probability model** to generate multiple imputations from **posterior predictive** draws and combine via **Rubin's rules**.

[6] **van Buuren, 2018**

- Evaluating accuracy alone **fails** to detect **variability** issues.
- Record the **bias**.
- Calculate the **coverage**.

- **Imputation is not prediction**. [7]

[7] van Buuren, 2018

References and Questions

Thank you for listening, any questions? [8]

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). ***Classification and regression trees* (1st ed.).** Chapman; Hall/CRC. https://doi.org/10.1201/9781315139470

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010).**Bart: Bayesian additive regression trees.** *The Annals of Applied Statistics*, *4*(1). https://doi.org/10.1214/09-aoas285

R Core Team. (2024). ***R: A language and environment for statistical computing*.** R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rubin, D. B. (1976).**Inference and missing data.** *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1978).**Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse.** *Proceedings of the Survey Research Methods Section of the American Statistical Association*. https://api.semanticscholar.org/CorpusID:197861764

Stekhoven, D. J., & Bühlmann, P. (2012).**Missforest—non-parametric missing value imputation for mixed-type data.** *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tan, Y. V., & Roy, J. (2019).**Bayesian additive regression trees and the general bart model.** *Statistics in Medicine*, *38*(25), 5048–5069. https://doi.org/https://doi.org/10.1002/sim.8347

van Buuren, S. (2018). ***Flexible imputation of missing data, second edition* (2nd ed.).** Chapman; Hall/CRC. https://doi.org/10.1201/9780429492259

[8] https://github.com/toddburrows

Appendix

Appendix A: The BART MCMC Cycle

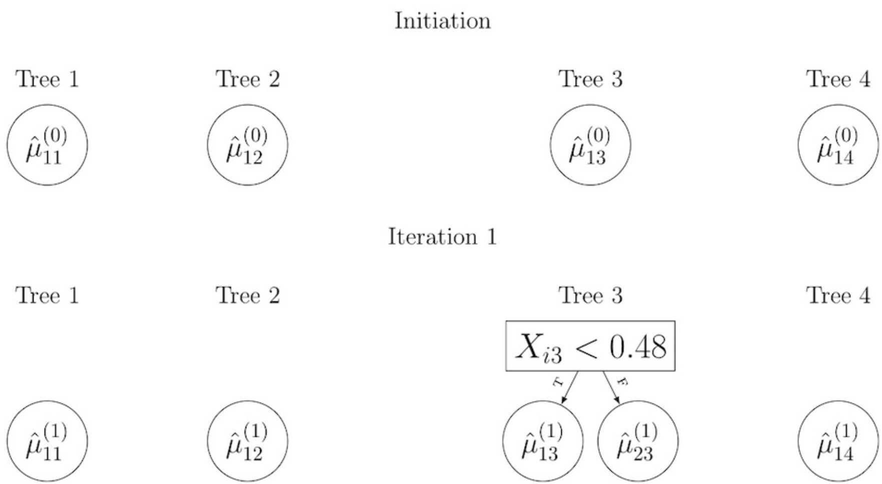

Figure. Initialisation and Iteration 1 of a BART model. Extracted from **Tan and Roy, 2019**

Appendix

Appendix A: The BART MCMC Cycle

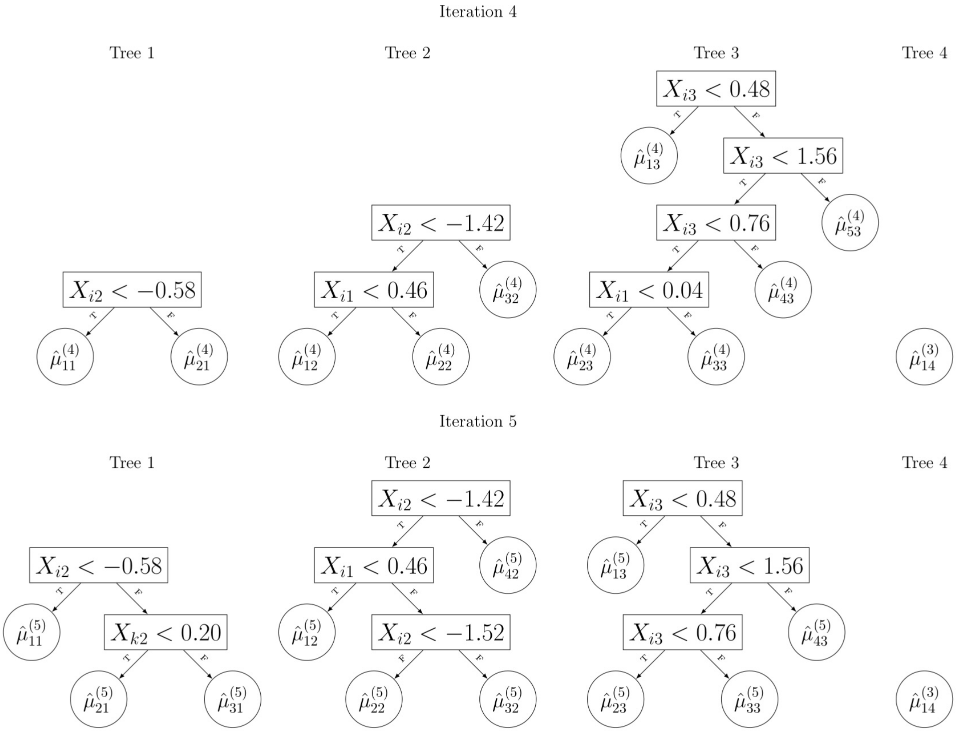

Figure. Iteration 4 and 5 of a BART model. Extracted from Tan and Roy, 2019