MCW Biostatistics Technical Report 71: Novel pediatric height outlier detection methodology for electronic health records via machine learning with monotonic Bayesian additive regression trees

RA Sparapani

October 13, 2021

1 Introduction

Our novel outlier detection methodology relies mainly on nonparametric machine learning via Bayesian Additive Regression Trees (BART) Chipman et al. (2010); Sparapani et al. (2021); specifically, an extension known as Monotonic BART or MBART Chipman et al. (2016). BART/MBART is an ensemble model of binary regression trees. Ensembles are the best-known predictive models in out-of-sample performance as assessed by an independent validation data set Baldi and Brunak (2001); Kuhn and Johnson (2013), i.e., ensembles will not over-fit to the training data at the expense of predictive performance on the unseen validation data (thus providing robustness to outliers present in the training data). The monotonic constraint of MBART requires that the unknown function to be learned from the data, $f(.)$, whose value is the estimated height, be monotonically increasing with the continuous covariates; in our study, those are age and weight (gender and race are categorical so monotonicity does not apply per se). We chose to model height monotonically by age for obvious reasons. However, fluctuating weights could be a real phenomenon that may require a treatment plan. The reason that we chose to model height as monotonically increasing by weight is to detect subtle outliers in weight, i.e., a height outlier that is flagged by the model may be correctly recorded, but the corresponding weight is an outlier. Therefore, a manual review of height outliers with a time series of height is likely to detect some of the rarer aberrant weights as well. For automated detection of height outliers, we rely on MBART. While others have relied on classical methods, we purposely chose to use machine learning with MBART since it is a natural fit for the problem at hand focused on height values alone without regard to weight.

2 Methods

2.1 Monotonic BART and Outliers

We have *population* predictions of the form $\hat{y}_{ij} = \mathrm{E}[y_{ij}] = \mu + \hat{f}(x_{ij})$ where $i = 1, \ldots, N$ indexes children and $j = 1, \ldots, n_i$ indexes the longitudinal measurements of child i (μ is a constant roughly centering the population: in this case $\mu = \overline{y}$). We need to adjust these predictions up or down for a given subject. Note that the values of y_{ij} are observed; after all, these are the values to be tested for outliers. So let $m_i = \mathrm{median}_j(y_{ij} - \hat{y}_{ij})$ (median rather than mean to be robust to outliers). Now, we make *personalized* predictions like so: $\tilde{y}_{ij} = m_i + \hat{y}_{ij}$. We define the *relative error* of these as $d_{ij} = (y_{ij} - \tilde{y}_{ij})/\tilde{y}_{ij}$. Height outliers are defined as $|d_{ij}| > \delta$ where δ can be determined from the Receiver Operating Characteristic (ROC) curve. Furthermore, the discriminating performance of the method is assessed by the area under the ROC curve.

This methodology has a practical advantage since it is a form of unsupervised learning, i.e.,

whether a particular patient has an outlier is NOT needed to be known for the training cohort (provided that height outliers are relatively rare as they are here) allowing us to amass a much larger data set for training than what would otherwise be possible. Therefore, hundreds, or even thousands, of patients can be employed for training that do NOT need to be manually reviewed to determine outliers. The ground truth knowledge of height outliers is only required for the validation cohort, i.e., we are training $f(.)$ on the relationship between height and the covariates in a robust to outliers ensemble predictive model (the strength of which can be assessed by metrics such as R^2 for both internal and external validity). The ground truth from the validation cohort will be used to assess the discriminatory performance via ROC, and the area under the ROC curve (often called the c statistic). This algorithmic proposal is based on a well-known result of predictive modeling: a continuous outcome, when applicable, will result in better predictive performance even in cases where the ultimate goal is a dichotomous decision like outlier vs. no outlier which can be arrived at via a properly chosen cutoff.

2.2 Marginal effects: Friedman's partial dependence function

BART/MBART does not directly provide a summary of the effect of a single covariate, or a subset of covariates, on the outcome. This is also the case for black-box, or nonparametric regression, models in general that need to deal with this same issue. A very useful development for such complex models, Friedman's partial dependence function Friedman (2001) can be employed here to summarize the marginal effect due to a subset of the covariates. Friedman's partial dependence function is a concept that is very flexible.

Suppose that we have a continuous outcome, y, along with a corresponding vector of covariates, $\boldsymbol{x}$. We use S to denote the indices of the covariates in the subset and the collection itself, i.e., define the row vector for test setting h as $\boldsymbol{x}_{hS} = [x_{hj}]$ where $j \in S$. Similarly, we denote the complement of the subset as C with $S \cup C$ spanning all covariates. The complement row vector for training observation i is $\boldsymbol{x}_{iC} = [x_{ij}]$ where $j \in C$. Therefore, the marginal expectation is defined as $\mathrm{E}\left[y|f, do(\boldsymbol{X}_{hS} = \boldsymbol{x}_{hS})\right] = \mu + f_S(\boldsymbol{x}_{hS})$. An analytic expression for f_S is not evident so we resort to Friedman's partial dependence function (FPDF). Of course, you could simply fit a BART function depending only on the variables in S; however, that would require many fits for all possible subsets, S, which we would often choose to avoid.

The nonparametric marginal dependence function is defined by fixing the subset at a test setting while aggregating over the training observations of the complement covariates: $f_S(\boldsymbol{x}_{hS}) = N^{-1} \sum_{i=1}^{N} f(\boldsymbol{x}_{hS}, \boldsymbol{x}_{iC})$. Other marginal functions can be obtained in a similar fashion.

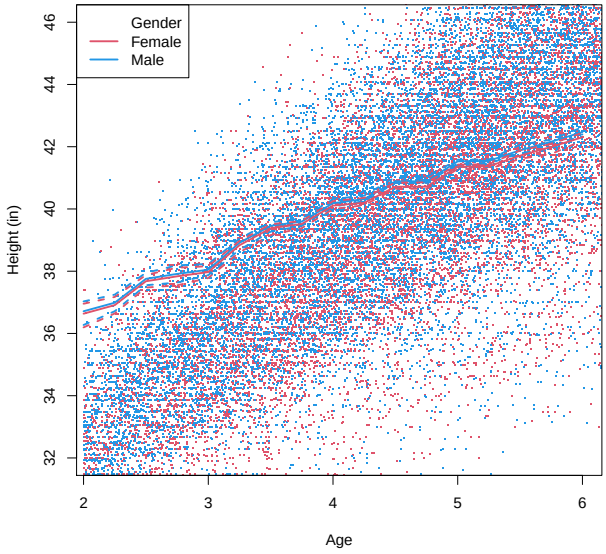

Figure 1: Friedman's partial dependence function when the strength of the relationship between age and weight is mistakenly ignored. In this figure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal effects.

2.3 Marginal effects and dependent variables

Friedman's partial dependence function works well when there are only weak relationships between the covariates. However, when there are strong relationships, such as between age and weight here; then, we need to extend this approach which we illustrate via our example.

We adopt the following notation for our variables: a for age, g for gender, r for race, w for weight and y for height. If we are interested in the marginal effects due to age and gender, then we could consider the FPDF $f_y(a, g) = \mathrm{E}\left[y|do(a, g)\right]$, i.e., the expected height on a grid of specified values for age and gender; however, this will yield unsatisfactory results: see Figure 1.

To do this the right way, first consider the likely monotonic relationship between age, gender and weight denoted as follows: $\tilde{f}_w(a, g) = \mathrm{E}\left[w|do(a, g)\right]$. So, in order to generate a corresponding grid of weight values to our age and gender settings, we need to fit an intermediate MBART model: $w_{ij} = \tilde{f}_w(a_{ij}, g_{ij}) + \tilde{\epsilon}_{ij}$ where $\tilde{f}_w \overset{\text{prior}}{\sim}$ MBART. Now, we arrive at a conditional appropriate for variables with a strong dependency: $f_y(a, g) = \mathrm{E}\left[y|do(a, g, w = \tilde{f}_w(a, g))\right]$; see Figure 2.

Since the age and weight relationship is strong, we arrive at the same issue if we want the marginal effect of weight and gender. And the corresponding remedy is employed here as well; see Figure 3 and Figure 4.

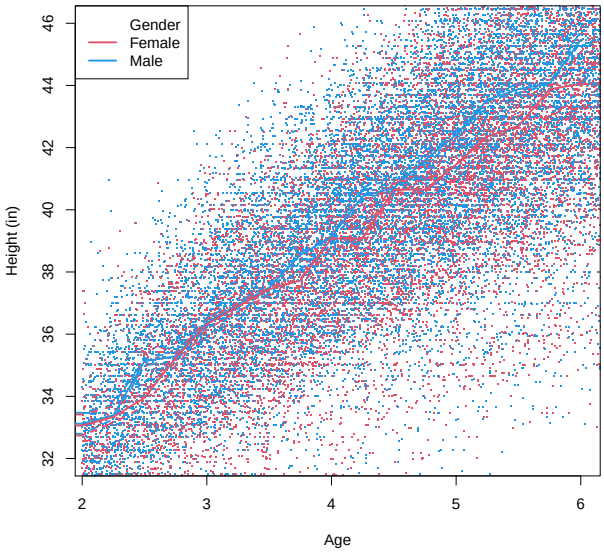

Figure 2: Friedman's partial dependence function when the strength of the relationship between age and weight is properly accounted for. In this figure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal effects.

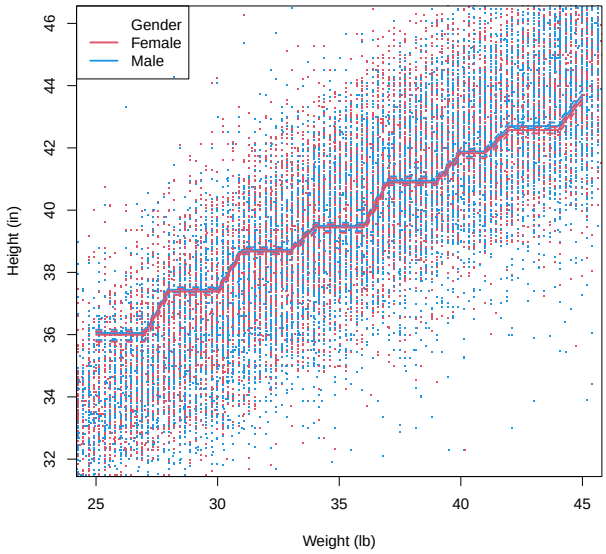

Figure 3: Friedman's partial dependence function when the strength of the relationship between age and weight is mistakenly ignored. In this figure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal effects.

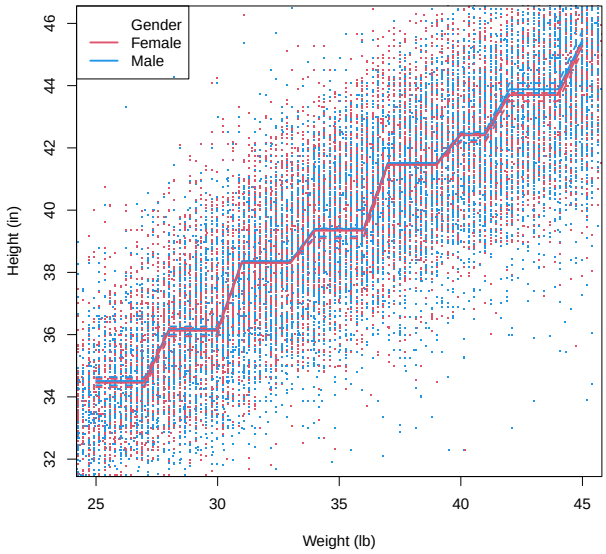

Figure 4: Friedman's partial dependence function when the strength of the relationship between age and weight is properly accounted for. In this figure, males are blue dots/lines and females are red dots/lines with 95% credible intervals around the marginal effects.

References

P Baldi and S Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 2nd edition, 2001.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010. doi: 10.1214/09-aoas285.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. High-dimensional nonparametric monotone function estimation using BART. *arXiv preprint arXiv:1612.01619*, 2016.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. URL `http://www.jstor.org/stable/2699986`.

M Kuhn and K Johnson. *Applied Predictive Modeling*. Springer-Verlag, New York, NY, 2013. doi: 10.1007/978-1-4614-6849-3.

Hang TT Phan, Florina Borca, David Cable, James Batchelor, Justin H Davies, and Sarah Ennis. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. *Scientific reports*, 10(1):1–9, 2020.

R Sparapani, C Spanbauer, and R McCulloch. Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: the BART R Package. *Journal of Statistical Software*, 97(1):1–66, 2021. doi: 10.18637/jss.v097.i01.