Bayesian Zero-Inflated Decision Trees

Lucas Roberts

Senior Data Scientist

The Hartford

rlucas7@vt.edu

May 28, 2016

Abstract

Wait until the end, when I think the draft is done to complete this section

1   Introduction

This chapter provides an overview of the decision tree models that are commonly used in theoretical and applied settings.

1 ¶on applications of count data with decision trees (motivation)

1 ¶on greedy induction We begin by discussing the earliest methods of induction: greedy algorithms.

1 ¶on Bayesian trees Bayesian approaches to building decision trees were not possible until advances in computing power and advances in sampling based approaches became available [8]. While the work of Breiman et al [1] always contained a Bayesian flavor, no probability measure over trees was ever proposed. That is until the groundbreaking works of Denison, Mallick, and Smith [6] and Chipman, George, and Mcculloch [4].

This manuscript proceeds as follows:

Section 2 contains the derivations for the marginal likelihoods of the zero-inflated decision trees. Section 3 displays the results from simulate data. Section 4 presents two case studies, one using the solder data from **? find the ref** [] and the second uses data collected from a study of nematodes in a vineyard in Dobson county North Carolina. Section 5 puts the findings in context with other research, summarizes the results, and points towards future research.

2   The Model Details

2.1   Bayesian Approaches

This section describes the Bayesian approach to decision trees. The methods described in the previous chapter provided an algorithm to fit a decision tree using a greedy algorithm. Besides the observed error, there is nothing to describe the fit of the model, or to provide a measure over decision trees. This section

provides both of these quantities. We begin by defining the model and calculating necessary quantities for the algorithm. Furthermore, there is no explicit model selection, which will be the main contribution of this thesis.

2.1.1 The CGM approach

We begin by defining notation and measures on each quantity of the tree. We assume the tree topology and split rules are conditionally independent. Based on fundamentals of probability we have the following relations:

$$\Pr(\mathcal{T}_i | \underline{y}, X) \propto \Pr(\mathcal{T}_i) \Pr(\underline{y} | \mathcal{T}_i, X) \tag{1}$$

$$\propto \Pr(\mathcal{T}_i) \int_\Theta \Pr(\underline{y} | \mathcal{T}_i, X, \theta) \pi(\theta) d\theta, \tag{2}$$

where $\Pr(\mathcal{T}_i)$, denotes the prior measure on trees and $\Pr(\underline{y} | \mathcal{T}_i, X)$ denotes the integrated likelihood of the tree. Finally, $\Pr(\underline{y} | \mathcal{T}_i, X, \theta)$, and $\pi(\theta)$, denote the tree likelihood and prior measure on node parameters respectively. This is the conditional decomposition defined by CGM [4]. We now proceed to define the aspects of the model described in CGM's paper [4]. The model has two main components, the tree $\mathcal{T}$ with b terminal nodes, and the parameters in each terminal node $(\theta_1, \ldots, \theta_b)$. The two main likelihoods in each terminal node are the normal and the multinomial, for continuous and categorical responses, respectively. Denote the responses in each terminal node as the vector of vectors $Y \equiv (Y_1, \ldots, Y_b)$. Then $Y_i = (y_{i1}, \ldots y_{in_i})$, and the main relation is the independence breakdown

$$f(Y | \mathcal{T}, X, \theta) = \prod_{i=1}^b f(Y_i | \mathcal{T}, X, \theta_i) = \prod_{i=1}^b \prod_{j=1}^{n_i} f(y_{ij} | \mathcal{T}, X, \theta_i). \tag{3}$$

The two likelihoods are now given by

$$f(y_{ij} | \mathcal{T}, X, \theta_i) = N(\mu_i, \sigma_i), \tag{4}$$

and the multinomial likelihood is

$$f(y_{i1}, \ldots, y_{in_i} | \mathcal{T}, X, \theta_i) = \prod_{j=1}^{n_i} \prod_{k=1}^K p_{ik}^{\mathbb{1}(y_{ij}=k)}. \tag{5}$$

In Equation 5, p_{ik} denotes the probability of being in category k in terminal node i and $\mathbb{1}(A)$ denotes the indicator function for the set A.

We now proceed to define the tree prior. We start with a tree consisting of a single node, the root node. We then imagine the tree growing by randomly choosing terminal nodes to split on. To grow a tree we must specify two functions, the growing function and the splitting function. The splitting function is denoted as $p_{\text{split}}(\eta, \mathcal{T})$ and the rule function is defined as $p_{\text{rule}}(\rho | \eta, \mathcal{T})$. The rule function provides a criteria to determine which of the two children nodes observed data go into. If the observed covariate value is less than the rule value, then observations go into the left child node. Similarly, if the observed covariate value is greater than the rule value, then observations go into the right child node. Growing the tree consists of creating two new children from a terminal node and assigning a rule to the terminal node (now a parent of two terminal nodes).

The probability measure on the potential splits of the tree is defined as

$$p_{\text{split}}(\eta, \mathcal{T}) = \alpha(1 + d_\eta)^{-\beta}, \quad \alpha > 0, \beta \geq 0, \tag{6}$$

where d_η denotes the depth of the node η and α, and β are scalars.

$$p_{\text{rule}}(\rho|\eta, \mathcal{T}) \propto \underbrace{\Pr(\text{split on covariate})}_{=(1)} \underbrace{\Pr(\text{split on a value given a covariate})}_{=(2)}. \tag{7}$$

Here CGM recommends using a uniform prior on (1), and splitting uniformly amongst the splitting values (in (2)) that do not result in an empty terminal node. While we choose the same proposal mechanism for quantity (2), the main point of this thesis is to examine and propose alternate specifications for (1). The data sets modeled in CGM [4] and DMS [6], and other modifications in the literature, deal with data with a small number of predictors. In this manuscript we are concerned with a response variable which is a count, which will ultimately explain how we specify quantity (1) in Equation 7.

2.1.2 Integrated Likelihood

We will now focus on the integrated likelihood, which is the quantity

$$\Pr(Y_i|\mathcal{T}, X) = \int_\Theta \Pr(Y_i|\mathcal{T}_i, X, \theta)\pi(\theta)d\theta. \tag{8}$$

To evaluate the integral in Equation 8 we must first define a prior, denoted $\pi(\theta)$, for the parameters in each terminal node. There are two possible priors for the case of the normal likelihood that will result in a conjugate prior/posterior. These are normals and normals and gammas, or normals and inverse gammas depending upon the given parametrization.

2.1.3 The Process Prior

Assuming we have a closed form solution for the integral in Equation 8, we can use Bayes' rule to determine

$$\Pr(\mathcal{T}|Y, X) \propto \Pr(Y|X, \mathcal{T}) \Pr(\mathcal{T}). \tag{9}$$

We now have an effective means of searching the posterior space over trees to determine the high posterior trees. We can do so by using the Metropolis-Hastings rule

$$\mathcal{T}^{i+1} = \begin{cases} \mathcal{T}^*, & \text{with probability } \alpha(\mathcal{T}^*, \mathcal{T}^i) = \min\left(\frac{q(\mathcal{T}^*, \mathcal{T}^i)}{q(\mathcal{T}^i, \mathcal{T}^*)} \frac{\Pr(Y|X, \mathcal{T}^*)}{\Pr(Y|X, \mathcal{T}^i)} \frac{\Pr(\mathcal{T}^*)}{\Pr(\mathcal{T}^i)}, 1\right) \\ \mathcal{T}^i, & \text{with probability } 1 - \alpha(\mathcal{T}^*, \mathcal{T}^i). \end{cases} \tag{10}$$

To evaluate the normalization constant would require summing Equation 9 across all possible trees. This is a sum with $\mathcal{O}(nd\frac{4^h}{h^{3/2}})$ terms, with h denoting the maximum height of the trees, n denoting the number of observations, and d denoting the number of covariates. This is an infeasible sum for most data sets, and for all data sets examined in this thesis. For the function $q(-|-)$, which is called the proposal function, we use q to propose a new tree $\mathcal{T}^*$. In Equation 10, $q(\mathcal{T}|\mathcal{T}^*)$ denotes proposing a new tree $\mathcal{T}^*$, starting from the current tree $\mathcal{T}$. Our proposal mechanism is as follows:

- The grow step chooses at random one of the terminal nodes and proposes to append two new child nodes with a certain probability that could depend on the tree depth, splitting on a chosen covariate.

- The prune step works in reverse of the grow. A terminal node is selected at random and that node and the node's sibling are pruned to the immediate parent of the two child nodes.

- The change step randomly picks an internal node and attempts to change the split rule at the node with that of another observation, possibly on a different covariate.

- The swap step randomly selects an internal node that is not the root node and proposes to swap the split rules of the parent-child pair. If both child nodes' split on the same covariate, then both children and the parent node's rules are swapped.

- The rotate step randomly chooses a left or right rotation move. Then this step randomly chooses an admissible internal node and rotates.

The rotate operation for binary trees was first introduced in Sleater and Tarjan [12] and was introduced into Bayesian decision trees in GL [9]. A good introduction and several practical uses of the rotate move can be found in Cormen, Lieserson, Rivest and Stein [5]. The proposal of Gramacy and Lee [9] only allows a rotate move for the specific case when a swap move is proposed and the parent child pair both split on the same covariate. We modify this and allow rotate to be a separate operation of the transition kernel and not a special swap move case. The proposal mechanism of CGM uses the grow, prune, change and swap moves only. We also allow swap moves in our proposal. In addition, neither of these papers included weights on each covariate in their examples or model specifications. They sampled each covariate and split value uniformly, at random.

The probability measure on the tree is defined as

$$\Pr(\mathcal{T}) = \prod_{\eta \in \mathcal{T}} p_{\text{rule}}(\rho|\eta, \mathcal{T}) p_{\text{split}}(\eta, \mathcal{T}). \tag{11}$$

The probability measure on each split, here denoted $p_{\text{split}}(\eta, \mathcal{T})$, uses Equation 6. Similarly the measure on each rule, here denoted $p_{\text{rule}}(\rho|\eta, \mathcal{T})$, uses Equation 7. All that is left to specify is the likelihood model in each node and the prior structure for the parameters in each node. This is done in the next subsection.

2.1.4 Node Likelihoods and Priors

CGM discuss three models. Two of the models use Gaussian priors and Gaussian likelihoods and one of the models uses a Dirichlet prior and a multinomial likelihood. The two Gaussian models differ in that one has a single variance and the other has a different variance for each node. As noted by Lee [11], in a greedy optimization context, sometimes the data suggest a different model than either a Gaussian or a multinomial-Dirichlet. If the experiment suggests analyzing data using an alternate model, the Bayesian context easily handles these alterations, once the corresponding likelihood and prior are specified. In the case of Lee [11], a zero inflated poisson (ZIP) model was proposed to analyze the solder data. Our Bayesian model can easily handle extensions such as this and also permits covariate selection, provided the integrated likelihood is available in closed form.

We begin with the Gaussian likelihood and Gaussian prior model. We define the likelihood as

$$N[y_{ij}|\mu_i, \sigma^2]. \tag{12}$$

Also, we define the prior for μ_i as

$$N[\mu_i|\bar{\mu}, \sigma^2]. \tag{13}$$

Furthermore, we define the prior for σ^2 as

$$\text{Inv-Gamma}(\sigma^2|\alpha, \beta). \tag{14}$$

4

All that remains is to evaluate the integral

$$\prod_{i=1}^{b} \int_{0}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} N[y_{ij}|\mu_i,\sigma^2] N[\mu_i|\bar{\mu},\sigma^2] \text{Inv-Gamma}(\sigma^2|\nu/2,\nu\lambda/2) d\mu_i d\sigma^2. \tag{15}$$

For this specific prior and likelihood we can explicitly calculate the marginal likelihood. Being able to marginalize the node parameters explicitly allows us to implement a Metropolis-Hastings algorithm without resorting to complicated, specialized algorithms, or numerical integrations. Straightforward analytic manipulations yield the solution to Equation 15 written here in Equation 16

$$\frac{ca^{b/2}}{\prod_{i=1}^{b}\sqrt{n_i+a}} \times \left(\sum_{i=1}^{b} \left(\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \right) + \frac{(\bar{y}_i-\bar{\mu})^2(n_i a)}{n_i+a} \right)^{-(\nu+n)/2}. \tag{16}$$

Assuming instead that the variances might change from node to node, then the stated model is misspecified. Let us denote the variance in each node as σ_i^2 and keep all other notations from the stated model specification. Then the model is specified using

$$N[y_{ij}|\mu_i,\sigma_i^2]. \tag{17}$$

Also, we define the prior for μ_i as

$$N[\mu_i|\bar{\mu},\sigma_i^2]. \tag{18}$$

Furthermore, we define the prior for the σ_i^2s as

$$\text{Inv-Gamma}(\sigma_i^2|\nu/2,\nu\lambda/2) \tag{19}$$

and now we evaluate the integral equation

$$\prod_{i=1}^{b} \int_{0}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} N[y_{ij}||\mu_i,\sigma_i^2] N[\mu_i|\bar{\mu},\sigma_i^2] \text{Inv-Gamma}(\sigma_i^2|\nu/2,\nu\lambda/2) d\mu_i d\sigma_i^2. \tag{20}$$

The result of computing the integrals in Equation 20 is

$$\prod_{i=1}^{b} \pi^{n_i/2}(\lambda\nu)^{\nu/2}\sqrt{\frac{a}{n_i+a}} \frac{\Gamma((n_i+\nu)/2)}{\Gamma(\nu/2)} \times \left(\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \frac{(\bar{y}_i-\bar{\mu})^2(n_i a)}{n_i+a} + \nu\lambda \right)^{(n_i+\nu)/2}. \tag{21}$$

These are the two "regression" models for Bayesian decision trees given in CGM [4].

The classification model discussed in CGM [4] defines the likelihood, prior, and integrated likelihood as

$$y_{i1},\ldots,y_{in_i}|\mathcal{T} \sim \text{Multinomial}(Y_i|\underline{n},\underline{p}), \tag{22}$$

$$\underline{p}|\mathcal{T} \sim \text{Dirichlet}(\underline{p}|\underline{\alpha}), \tag{23}$$

and

$$\Pr(Y|\mathcal{T},X) = \left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \right)^b \prod_{i=1}^{b} \left(\frac{\prod_{k=1}^{K}\Gamma(n_{ik}+\alpha_k)}{\Gamma(n_i+\sum_{k=1}^{K}\alpha_k)} \right), \tag{24}$$

respectively.

If we wanted to model the data using a different data generating process, for example a ZIP. We could do so by specifying a different likelihood, prior, and computing the integrated likelihood. For a ZIP model this is possible using gamma priors for the rate (λ) and beta priors for zero inflation components (ϕ).

2.1.5 A Bayesian Zero Inflated Poisson Model

Lee and Jin [11] reconsidered impurity functions in light of the connection to likelihood functions. Lee and Jin [11] proposed to use likelihood functions instead of impurity functions that model the data generating process. Towards this end they considered the soldering data from Chambers and Hastie [2]. The response of interest in this case is a collection of counts on manufactured circuit boards. This response has many zero's and Lee and Jin [11] propose using a zero inflated (ZIP) poisson likelihood to model the measured counts. Lee and Jin [11] optimized using a greedy algorithm and they found the fit and holdout prediction to be better using the ZIP model in each terminal node. If we are to use a Bayesian approach to this problem we need to define the likelihood, the prior, and the integrated likelihood. We now define these three quantities.

The likelihood for a single observation is

$$f(y|\lambda, \phi) \propto \mathbb{1}(y = 0)(\phi + (1 - \phi)\exp(-\lambda)) + \mathbb{1}(y > 0)\left(\exp(-\lambda)\frac{\lambda^y}{y!}\right). \tag{25}$$

The priors for λ and ϕ are

$$\pi(\phi, \lambda) \propto \underbrace{\frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{=\text{A beta prior}} \times \underbrace{\frac{\lambda^{\alpha_\lambda-1}\exp(-\lambda/\beta_\lambda)}{\Gamma(\alpha_\lambda)\beta_\lambda^{\alpha_\lambda}}}_{=\text{A gamma prior}}. \tag{26}$$

Now we need to calculate the integrated likelihood, which means we must evaluate

$$\int_0^1 \int_0^\infty \left(\mathbb{1}(y = 0)(\phi + (1 - \phi)\exp(-\lambda)) + \mathbb{1}(y > 0)\left(\exp(-\lambda)\frac{\lambda^y}{y!}\right)\right) \tag{27}$$

$$\frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\lambda^{\alpha_\lambda-1}\exp(-\lambda/\beta_\lambda)}{\Gamma(\alpha_\lambda)\beta_\lambda^{\alpha_\lambda}}d\lambda d\phi.$$

Let j index the observed zero counts. Furthermore, let $\bar{y}_+$ denote the average of the non-zero counts and n_0 and n_+ denote the number of zeros and non-zeros in the data respectively. Now we assume that the observations are $i.i.d.$ and simple calculations lead to the conclusion that

$$\Pr(Y|X, \mathcal{T}) = \left[\sum_{j=0}^{n_0} \binom{n_0}{j}\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+j)\Gamma(n_0+\beta-j)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_0)}\left(\frac{n_0-j+\beta_\lambda^{-1}}{\beta_\lambda}\right)^{\alpha_\lambda}\right] \tag{28}$$

$$+ \frac{\Gamma(\alpha_\lambda+n_+\bar{y}_+)}{\Gamma(\alpha_\lambda)\beta_\lambda^{\alpha_\lambda}}(n_+ + 1/\beta_\lambda)^{\alpha_\lambda+n_+\bar{y}_+}.$$

2.2 ZIP Derivations

In this section we provide the derivation for the integrated likelihood for the Bayes ZIP (zero inflated poisson) tree model. Now let us define some notation: j will index either all observations or only the observed zero count observations if the upper limit is n_0 then j will index observed zero counts only, if the upper index limit is n then all observations are indexed. Also j' will index the non-zero observations. The total number of non-zero observations is denoted n_+, so that $n_+ + n_0 = n$. Finally let $\bar{y}_{i+}$ denote the sample mean of the non-zero count observations in terminal node i.

$$\Pr(Y|X,\mathcal{T}) = \prod_{i=1}^{b} \int_0^1 \int_0^\infty \prod_{j=1}^{n_i} \left[\mathbb{1}[y_{ij}=0](\phi+(1-\phi)\exp{(-\lambda)}) + \mathbb{1}[y_{ij}>0]\frac{\exp{(-\lambda)}\lambda^{y_{ij}}}{y_{ij}!} \right] \pi(\phi_i,\lambda_i)d\lambda_i d\phi_i$$

$$= \prod_{i=1}^{b} \int_0^1 \int_0^\infty \underbrace{\prod_{j=1}^{n_0}(\phi+(1-\phi)\exp{(-\lambda)})\pi(\phi_i,\lambda_i)d\lambda_i d\phi_i}_{=(1)}$$

$$+ \prod_{i=1}^{b} \int_0^1 \int_0^\infty \underbrace{\prod_{j'=1}^{n_+}\frac{\exp{(-\lambda)}\lambda^{y_{ij'}}}{y_{ij'}!}\pi(\phi_i,\lambda_i)d\lambda_i d\phi_i}_{=(2)}\,.$$

We will first tackle (1), then tackle (2).

$$(1) = \int_0^1 \int_0^\infty \prod_{j=1}^{n_0}(\phi+(1-\phi)\exp{(-\lambda)})\pi(\phi_i,\lambda_i)d\lambda_i d\phi_i$$

$$= \int_0^1 \int_0^\infty (\phi+(1-\phi)\exp{(-\lambda)})^{n_0}\pi(\phi_i,\lambda_i)d\lambda_i d\phi_i$$

$$= \int_0^1 \int_0^\infty \sum_{j=1}^{n_0}\binom{n_0}{j}\phi^j(1-\phi)^{n_0-j}\exp{(-(n_0-j)\lambda)}\pi(\phi_i)\pi(\lambda_i)d\lambda_i d\phi_i.$$

Now we take $\pi(\phi_i)$ to be a beta(α,β) prior and $\pi(\lambda_i)$ to be a gamma$(\alpha_\lambda,\beta_\lambda)$ prior. This simplifies matters greatly.

$$\int_0^1 \int_0^\infty \sum_{j=1}^{n_0}\binom{n_0}{j}\phi^j(1-\phi)^{n_0-j}\exp{(-(n_0-j)\lambda)}\frac{\Gamma(\alpha+\beta)\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}\frac{\lambda^{\alpha_\lambda-1}\exp{(-\lambda/\beta_\lambda)}}{\Gamma(\alpha_\lambda)\beta_\lambda^{\alpha_\lambda}}d\lambda_i d\phi_i$$

$$= \sum_{j=1}^{n_0}\binom{n_0}{j}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha_\lambda)\beta_\lambda^{\alpha_\lambda}}\underbrace{\int_0^1 \phi^{j+\alpha-1}(1-\phi)^{\beta+n_0-j-1}d\phi_i}_{\text{a beta kernel}}\underbrace{\int_0^\infty \lambda^{\alpha_\lambda-1}\exp{(-(n_0-j+\beta_\lambda^{-1})\lambda)}d\lambda_i}_{\text{a gamma kernel}}$$

$$= \underbrace{\sum_{j=1}^{n_0}\binom{n_0}{j}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha_\lambda)\beta_\lambda^{\alpha_\lambda}}\frac{\Gamma(\alpha+j)\Gamma(\beta+n_0-j)\Gamma(\alpha_\lambda)}{\Gamma(\alpha+\beta+n_0)}(n_0-j+\beta_\lambda^{-1})^{\alpha_\lambda}}_{=(1)}\,.$$

Now with the first piece simplified we move on to piece (2).

$$(2) = \int_0^1 \int_0^\infty \prod_{j'=1}^{n_+} \frac{\exp(-\lambda)\lambda^{y_{ij'}}}{y_{ij'}!} \pi(\phi_i, \lambda_i) d\lambda_i d\phi_i$$

$$= \int_0^\infty \prod_{j'=1}^{n_+} \frac{\exp(-\lambda)\lambda^{y_{ij'}}}{y_{ij'}!} \pi(\lambda_i) d\lambda_i$$

$$= \int_0^\infty \frac{\exp(-n_+\lambda)\lambda^{n_+\bar{y}_{i+}}}{\prod_{j'=1}^{n_+} y_{ij'}!} \pi(\lambda_i) d\lambda_i$$

$$= \int_0^\infty \frac{\exp(-n_+\lambda)\lambda^{n_+\bar{y}_{i+}}}{\prod_{j'=1}^{n_+} y_{ij'}!} \frac{\lambda^{\alpha_\lambda - 1}\exp(-\lambda/\beta_\lambda)}{\Gamma(\alpha_\lambda)} d\lambda_i$$

$$= \frac{\int_0^\infty \exp(-(n_+ + \beta_\lambda^{-1})\lambda)\lambda^{n_+\bar{y}_{i+} + \alpha_\lambda - 1} d\lambda_i}{\Gamma(\alpha_\lambda)\prod_{j'=1}^{n_+} y_{ij'}!}$$

$$= \underbrace{\frac{\Gamma(n_+\bar{y}_{i+} + \alpha_\lambda)(n_+ + \beta_\lambda^{-1})^{n_+\bar{y}_{i+} + \alpha_\lambda}}{\Gamma(\alpha_\lambda)\prod_{j'=1}^{n_+} y_{ij'}!}}_{=(2)} .$$

And the result is shown.

3 Simulations

4 Two Case Studies

4.1 The Solder Data

In this section we analyze the data from the soldering experiment from Chambers and Hastie [3]. The data contains 900 rows and 6 columns. The data was collected from an industrial experiment of manufactured circuit boards and the number of soldering defects present on the circuit boards. These data were analyzed by Lambert and were the motivating application behind the development of the zero-inflated Poisson regression model [10]. The data are available in the R package *faraway* or you may use a balanced subset of 720 observations available in the R package *rpart* [13]. We use the full data from the *faraway* package [7]. The columns in the data are: Opening, Solder, Mask, PadType, Panel, and skips. The skips are the response variable of interest which represent the number of skips in the soldering of the circuit board. The Panel variable is a numeric variable taking on the values 1, 2, and 3. The remaining variables are all categorical with the number of levels for each factor ranging from 2 levels to 10 levels.

5 Discussion

References

[1] L. Breiman, J.H. Friedman, C.J. Stone, and R.A. Olshen. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[2] John M Chambers, Trevor Hastie, et al. *Statistical models in S*. Chapman & Hall London, 1992.

[3] John M Chambers and Trevor J Hastie. *Statistical models in S*. CRC Press, Inc., 1991.

[4] H.A. Chipman, E.I. George, and R.E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, pages 935–948, 1998.

[5] T.H. Cormen, C.E. Lieserson, R.L. Rivest, and Stein C. *Introduction to algorithms*. The MIT press, 2001.

[6] D.G.T. Denison, B.K. Mallick, and A.F.M. Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

[7] Julian Faraway. *faraway: Functions and Datasets for Books by Julian Faraway*, 2016. R package version 1.0.7.

[8] Alan E Gelfand and Adrian F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[9] R.B. Gramacy and H.K.H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[10] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[11] S.K. Lee and S. Jin. Decision tree approaches for zero-inflated count data. *Journal of applied statistics*, 33(8):853–865, 2006.

[12] D.D. Sleator and R.E. Tarjan. Self-adjusting binary search trees. *Journal of the ACM (JACM)*, 32(3):652–686, 1985.

[13] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2015. R package version 4.1-10.