Classification and Regression Tree Analysis

Jake Morgan

`jakem@bu.edu`

Boston University School of Public Health
Department of Health Policy & Management

Contents

Executive summary

Classification and Regression Tree Analysis, CART, is a simple yet powerful analytic tool that helps determine the most "important" (based on explanatory power) variables in a particular dataset, and can help researchers craft a potent explanatory model. This technical report will outline a brief background of CART followed by practical applications and suggested implementation strategies in R. A short example will demonstrate the potential usefulness of CART in a research setting.

1 Background and Application

As computing power and statistical insight has grown, increasingly complex and detailed regression techniques have emerged to analyze data. While this expanding set of techniques has proved beneficial in properly modeling certain data, it has also increased the burden on statistical practitioners in choosing appropriate techniques. Arguably an even heavier burden has been placed on non-statistician health practitioners – in university, government, and private sectors – where statistical software allows for immediate implementation of complex regression techniques without interpretation or guidance. In response to this growing complexity, a simple tree system, Classification and Regression Tree (CART) analysis, has become increasingly popular, and is particularly valuable in multidisciplinary fields.

1.1 CART in Public Health

CART can statistically demonstrate which factors are particularly important in a model or relationship in terms of explanatory power and variance. This process is mathematically identical to certain familiar regression techniques, but presents the data in a way that is easily interpreted by those not well versed in statistical analysis. In this way, CART presents a sophisticated snapshot of the relationship of variables in the data and can be used as a first step in constructing an informative model or a final visualization of important associations. In a large public health project, statisticians can use CART to present preliminary data to clinicians or other project stakeholders who can comment on the statistical results with practice knowledge and intuition. This process of reconciling the clinical and statistical relevance of variables in the data ultimately yields a more well informed and statistically informative model than either a singularly clinical or statistical approach. For example, complex regression models are routinely presented in economics literature with little introduction or explanation as the audience is familiar with these techniques, and are more interested in the specific application of a particular technique or unexpected result. In public health, however, this method of presentation is not motivating for practitioners without statistical expertise and who need to know the mechanism of the health effect to determine clinical relevance or craft an effective intervention. On the other hand, if the data were explained purely by narrative description and anecdote – or excluding variables with statistically significant explanatory power without reason – it would be interpreted as lacking scientific rigor. The benefit of CART is to visually bridge interpretation and statistical rigor and facilitate relevant and valid model design.

While CART may be unfamiliar in some areas of public health, the concept is firmly rooted in health practice, particularly in epidemiological and clinical settings. The seminal work on classification and regression trees was published in book form by Breiman, Friedman, Olshen, and Stone in 1984 under the informative title *Classification and Regression Trees*, and they open the text with an example from clinical practice seeking to identify high risk patients within 24 hours of hospital admission for a myocardial infarction [10]. This example proved to be particularly relevant – hundreds of studies have since emerged using CART analysis in clinical settings investigating myocardial infarctions.[1] Often, these studies have many

---

[1]Over 300 results in PubMed, over 700 in Google Scholar.

confounding variables which independently may not strongly predict a given outcome, such as heart attack, but together are important. CART analysis can guide medical researchers to isolate which of these variables is most important as a potential site of intervention. <mark>Public health research, especially concerning behavioral factors, often have some intuition regarding the most important predictors, which may explain why the method has been often absent from public health literature.</mark> One review of CART in public health has a much more pessimistic rational for the lack of use: "a general lack of awareness of the utility of CART procedures and, among those who are aware of these procedures, an uncertainty concerning their statistical properties." [6] This report will highlight some of the possible CART procedures and present their statistical properties in a way which is helpful to public health researchers.

2   Theory and Implementation

The statistical processes behind classification and regression in tree analysis is very similar, as we will see, but it is important to clearly distinguish the two. For a response variable which has *classes*, often 0–1 binary,[2] we want to organize the dataset into groups by the response variable – classification. When our response variable is instead numeric or continuous we wish to use the data to *predict* the outcome, and will use regression trees in this situation. A richer mathematical explanation follows, but, essentially, a classification tree splits the data based on homogeneity, where categorizing based on similar data, filtering out the "noise" makes it more "pure" – hence the concept of a **purity criterion.** In the case where the response variable does not have classes, a regression model is fit to each of the independent variables, isolating these variables as nodes where their inclusion decreases error. A useful visualization of this process, created by Majid (2013), is Figure 1.
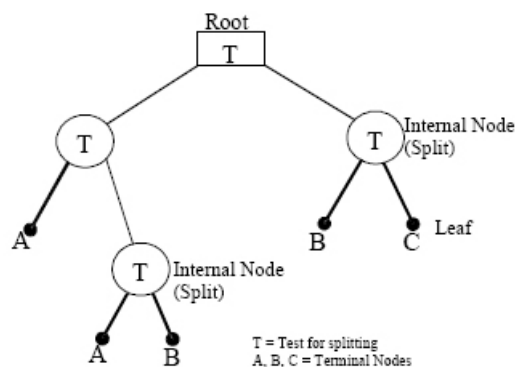

Figure 1: The Structure of CART

The premise of our investigation is fairly simple – given factors $x_1, x_2, x_3, ..., x_n$ in the domain X we want to predict the outcome of interest, Y. In Figure 1, the graphic is the

---

[2]Technically, classification under CART has a binary response variable and a variation on the algorithm called C4.5 is used for multi-category variables. This distinction is not of supreme importance in this discussion, however.

domain of all factors associated with our Y in descending order of importance. In traditional regression models, linear or polynomial, we develop a single equation (or *model*) to represent the entire data set. CART is an alternative approach to this, where the data space is partitioned into smaller sections where variable interactions are more clear. CART analysis uses this **recursive** partitioning to create a tree where each node – T in Figure 1 – represents a cell of the partition. Each cell has attached to it a simplified model which applied to that cell only, and it is useful to draw an analogy here to conditional modeling, where as we move down the **nodes,** or **leaves,** of the tree we are conditioning on a certain variable – perhaps patient age or the presence of a particular comorbidity. The final split or node is sometimes referred to as a leaf. In Figure 1, A, B, and C are each terminal nodes (leaves), implying that after this split, further splitting of the data does not explain enough of the variance to be relevant in describing Y. Using the notation from *Encyclopedia of Statistics in Quality and Reliability*[7], we mathematically represent this process:

Recalling we want to find a function $d(x)$ to map our domain X to our response variable Y we need to assume the existence of a sample of n observations $\mathcal{L} = \{(x_1, y_1), ..., (x_n, y_n)\}$. As in standard in regression equations, our criterion for choosing $d(x)$ will be the mean squared prediction error $E\{d(x) - E(y|x)\}^2$, or expected misclassification cost in the case of the classification tree. For each leaf-node l and c training samples in the regression tree, then, our model is just $\hat{y} = \frac{1}{c} \sum_{c=1}^{c} y_1$, "the sample mean of the response variable in that cell[12]" which creates a piecewise constant model. In the case of the classification tree with leaf node l, training sample c and $p(c|l)$, the probability that an observation l belongs to class c, the **Gini index node impurity criterion**[3] $(1 - \sum_{c=1}^{C} p^2(c|l))$ defines the node splits, where each split maximizes the decrease in impurity. Whether using classification or regression, reducing error – either in classification or prediction – is the principal driving statistical mantra behind CART.

2.1 Implementation in *R*

Implementation of CART is primarily carried out through R statistical software and is fairly easy to execute. The two families of commands used for CART include `tree` and `rpart`. `tree` is the simpler of the commands, and will be the focus of the applied examples in this report due to ease of use. The benefit to `rpart` is that it is user-extensible[4] (a desired commodity in the R community) and can be modified to incorporate surrogate variables more effectively. Additionally, `tree` has a useful option to more easily compare deviance to GLM or GAM models, which is more difficult in the `rpart` formulation. The `tree` command is summarized as:[5]

```
tree(formula, data, weights, subset,
   na.action = na.pass, control = tree.control(nobs, ...),
   method = "recursive.partition",
   split = c("deviance", "gini"),
```

---

[3] There are entropy measures that are functionally equivalent.

[4] That is, *R* is designed so that any user can expand or add to its capabilities with the proper programming knowledge.

[5] Adapted from the R help files.

```
      model = FALSE, x = FALSE, y = TRUE, wts = TRUE, ...)
```

where `formula` defines where the classification or regression model should be entered in the form $Y \sim x_1 + ... + x_n$ and `data` refers to the dataset to use. This level of specification is adequate for many simple tree analyses, but it is clear that the other options allow `tree` to be a powerful tool. `Weights` allow a user to assign a vector of non-negative observational weights, `subset` allows a subset of data to be used, `na.action` is the missing data handler (the default is to skip missing values), `control` allows users to manually determine certain aspects of fit, `method` allows the users to assign other CART methods where the default is the previously discussed "recursive.partition," `split` allows users to define alternative node splitting criterion, and `model` can be used to define numerical nodule endpoint, for example labeling a binary response of "y" as "True."
`Rpart` is similar, with the general command structure:

```
   rpart(formula, data, weights, subset,
      na.action = na.rpart, method,
      model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)
```

where the definitions are the same as above. It should be obvious that the implementation is almost identical, except for subtle shifts in coding. For example, in `tree` one specifies classification analysis in the formula line by specifying the dependent variable as a factor, i.e.:

$$\texttt{tree}(factor(Y) \sim \sum(X), \texttt{ data=}mydata)$$

while in `rpart` classification is specified in the method option:

$$\texttt{rpart}((Y) \sim \sum(X), \texttt{ method=``class'', data=}mydata)$$

Similarly, the only difference between the regression implementation in `tree`:

$$\texttt{tree}((Y) \sim \sum(X), \texttt{ data=}mydata)$$

and `rpart`:

$$\texttt{rpart}((Y) \sim \sum(X), \texttt{ method=``anova'', data=}mydata)$$

is removing the "factor" specification from `tree` and changing the `model` statement in `rpart` to "anova." These methods are easily applied to a sample data set.

2.2   An Applied Example

Using a dataset of patient visits for upper respiratory infection at Boston Medical Center, we can demonstrate the potential benefits of using CART analysis. Consider the issue of overprescribing of antibiotics for viral respiratory infections. A potential question might be the factors associated with a physician's rate of prescribing – if we were to construct an intervention against high prescribing, where might be a good location? To answer this question, the dataset provides each physician's rate prescribing – a continuous variable – so we will implement the `tree` function in R:

```
Regression Tree:    mytree=tree(prescribing_rate~+physician_specialty+
       patient_ethnicity + patient_insurance +...+ comorbidites,
       data = mydata)
```
The command summary(*mytree*) gives us the relevant statistical results:
```
   mytree=tree(prescribing_rate~patient_gender+physician_specialty+
      patient_ethnicity + patient_insurance +...+ comorbidites,
      data = mydata) Variables actually used in tree construction:
   [1] "private_insurance" "black_race"
   Number of terminal nodes:  3
   Residual mean deviance:  0.02441 = 167.5 / 6864
   Distribution of residuals:
     Min.      1 Q     Median    Mean    3 Q      Max
   -0.546   -0.1155   -0.00597    0    0.1328   0.5245
```

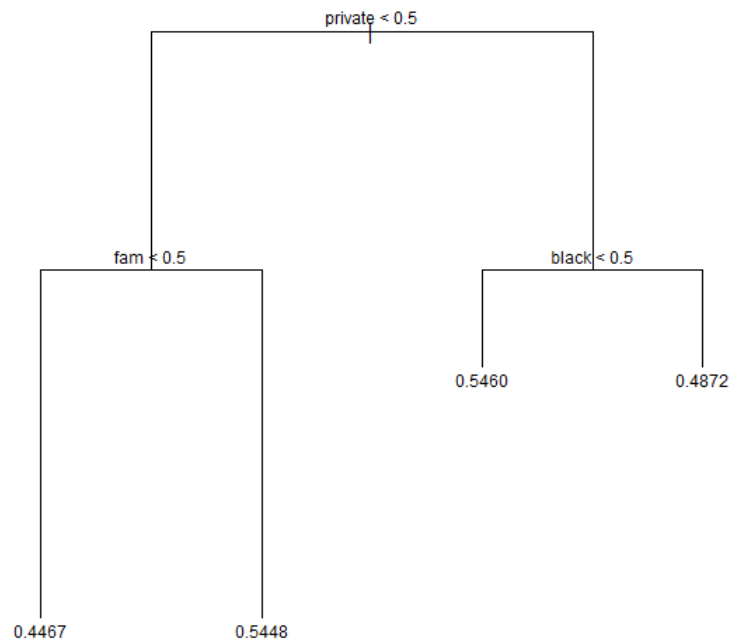And the associated visualization in Figure 2:

Figure 2: Regression Tree Analysis

Here, CART reveals that the most important factors associated with higher prescribing rates is the percent of privately insured patients a physician has, the percent of black patients a physician has, and whether or not the physician specializes in family medicine. For each node, the right branch of the node is conditional on the node being true, and the left is

conditional on the node being false. For example, the first node is the proportion of the physician's patients which are privately insured, so the right branch is only those data with privately insured patients, and the left includes patients without private insurance. Conditional on private insurance proportion, physician specialty in family medicine or the race mix of patients becomes important. The numbers at the bottom of the terminal branches indicate the mean of the prescribing rate in each data subset. For example, privately insured and black patient populations had a 48% average physician rate of prescribing, while visits without private insurance and with a family medicine doctor were associated with a 54% average physician prescribing rate. Both terminal branches conditional on the presence of private insurance are higher than those associated with other types of (often public) insurance, which could indicate prescriptions are easier to cover under a private plan, or might indicate different rates of demand among those who hold private insurance. Among those without private insurance, the presence of a family medicine doctor raises the average provider prescribing rate by approximately 10%, indicating that family medicine doctors systematically prescribe most antibiotics than non-family medicine doctors. The summary of the model shows a deviance of 0.0244, or a root mean square of 0.16 (16%), which indicates that our model is actually pretty good given its simplicity, suggesting that the nodes identified explain a good deal of the variation in prescribing rates. This also helps to explain while our tree only has two levels: most of the variation has been described by this relationship.

Our regression tree analysis suggested that family medicine specialists may prescribe systematically differently than other specialties. We may want to know what type of patients choose to see a family medicine doctor to see to what degree patient characteristics may be driving the difference among family medicine doctors. For this question, we can use a classification tree based on the binary response of whether or not a physician is a family medicine practitioner. Figure 3 depicts this relationship.

```
myclass=tree(family_med ~ patient_gender + patient_age +
      patient_ethnicity + patient_insurance +...+ comorbidites,
      data = mydata)
```

The command summary(*myclass*) gives us the relevant statistical results:

```
Classification Tree:
  myclass=tree(family_med ~ patient_gender + patient_age +
  patient_ethnicity + patient_insurance +...+ comorbidites,
  data = mydata)
  Variables actually used in tree construction:
  [1] "black" "private_insurance" "patient_over_45"
  Number of terminal nodes:  4
  Residual mean deviance:  1.054 = 7234 / 6863
  Misclassification error rate:  0.2403 = 1650 / 6867
```

Figure 3 indicates that the most important predictors of choosing a family medicine doctor are patient's race, age, and insurance status. The largest patient share of family medicine doctors is young, black patients, among whom 42% have a family medicine physician. Among black patients *over* 45 that number is almost half, 24%. Among non-black patients, those without private health insurance are more likely to choose a family medicine practitioner than those with private insurance. These results are somewhat more difficult to interpret because of the nature of the family medicine doctor. For example, we might expect more
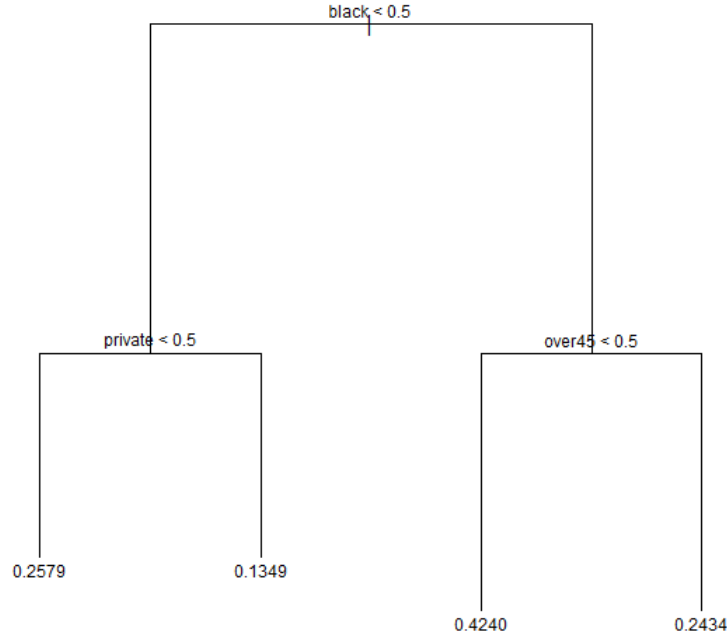
Figure 3: Classification Tree Analysis

young people to see family medicine doctors – family medicine treats children – if they have not been motivated to seek a different PCP. High prescribing among family medicine doctors is consistent with the public literature, which suggests that the broader training associated with family medicine (compared to a specialized internal medicine practitioners) may lead to more liberal prescribing habits.

The interventions suggested in our classification diagram may include more rigorous training programs during family medicine residency, or some sore of continuing education program. It might also be important to make sure patients are connected with an appropriate PCP and develop closer connections to care. While a classification tree analysis alone may not be convincing enough to implement certain interventions, this exercise is hypotheses generating, suggesting further areas of empirical research. The misclassification error rate here is 24%, which while slightly high is not uncommon in published studies using classification trees.

2.3 Stata and SAS in Brief

Review of the literature reveals that R is by far the most widely used statistical implementation of CART. The reasons for this are varied: R is freely available and user-modifiable, and for this reason the constantly changing landscape of CART is quickly rolled out as a

program extension; R was the first widely available program to handle CART, so has the most robust collection of trouble-shooting guides and technical support[6]; The implementation is simple – three lines of code will return a robust CART analysis. Increasingly, however, implementations for Stata and SAS are emerging, although these programs are often long and cumbersome. For SAS, implementation must occur in a separate module SAS Enterprise Miner (EM)[17]. EM is driven by a click-and-point interface and a digital workspace where users drag nodes from a toolbar to specify slitting criteria, partion rules, and other program rules. The program then implements CART to statistically grow a tree. For some, the simple interface and colorful flowchart of decision making rules may make the programming of CART more intuitive and clear. EM, however, is not included with the base version of SAS – the desktop module costs around \$20,000 for a single seat license[18]. Stata has the capability to implement a form of CART for failure data [19], but does not have the ability to run CART on cross sectional data as is widely popular. Nick Cox, lecturer at Durham University and one of the most prolific and well known Stata programmers noted in 2010 that "in general, the scope for this kind [CART] of work in Stata falls much below that in other software such as R" [20] and little has changed since then. The reference section provide the relevant literature for interested readers.

3   Extensions – Bagging and Boosting

There are many slight variations in CART analysis based on almost any data structure permutation we could imagine. If you decide to apply CART in your own work, Google and R web forums will prove to be very helpful in determining which slight adjustments need to be made now that you are familiar with the basic structure and terminology associated with the procedure. It is appropriate to discuss two commonly used variations: bagging and boosting.

Bagging was developed by Breiman and appeared shortly after his seminal work defining the field, and has gained traction following the increased interest in bootstrapping and similar procedures in statistical analysis, and it is useful to think of it as bootstrapping for tree analysis.[7] The name derives from **b**ootstrap **agg**regat**ing** and involves creating multiple similar datasets, re-running the tree analysis, and then collecting the aggregate of the results and re-calculating the tree and associated statistics based on this aggregate. This technique is often used as cross-validation for larger trees a user wishes to prune and where different versions of the same tree have vastly different rates of misclassification. In general, the procedure will improve the results of a highly unstable tree but may decrease the performance of a stable tree. Using the package `ipred` the procedure is easy to implement, and we will briefly present it here using the data from our applied example:

```
mybag=bagging(family_med~patient_gender+patient_age+
    patient_ethnicity + patient_insurance +...+ comorbidites,
    data = mydata,nbagg=30,coob=T)
```

---

[6]These files can be widely found with simple Google searches. The `rpart` help file is available at <http://cran.r-project.org/web/packages/rpart/rpart.pdf>

[7]The other common bootstrapping procedure for trees is so-called "Random Forest" which is similar and will not be covered in depth here

where `nbagg` specifies that the procedure will create 30 full datasets to aggregate and `coob` specifies the aggregation selection technique, here the averaged model ("out-of-the-bag"). Calling on the command `mybag$err` will return the new misclassification error, in our case 23.9% – an unimpressive 0.01 decrease in misclassification.

Boosting is a technique that seeks to reduce misclassification by a recursive tree model. Here, classifiers are iteratively created by weighted versions of the sample where the weights are adjusted at each iteration based on which cases were misclassified in the previous step – many "mini-trees" which exhibit continuously decreasing misclassification. This technique is often applied to data which has high misclassification because it is largely uninformative, or a "weak learner." In these data sets classification is only slightly better than a random guess (think misclassification only slightly less than 50%[8]), since the data are so loosely (perhaps because of confounders) related [15]. Implementing boosting is either done through a complex series of packages in R or some third-party software specializing in decision tree analysis. Our results indicate that our data are not weak learners, so we will not implement boosting here; in our case, bagging is much more appropriate. Generally, the data structure will indicate whether boosting or bagging is more appropriate. For more information on boosting in R, consult the `adabag` function.

---

[8]Since a random guess will be correct 50% of the time.

4 Required Reading

4.1 Introduction to CART

Simple Overview of CART with Step-by-Step Intuition: Loh, WeiYin. "Classification and regression tree methods." *Encyclopedia of statistics in quality and reliability* (2008).

R Implementation with Additional Tools for Robustness Analysis: Shalizi, Cosma "Classification and Regression Trees" unpublished report on data mining procedures available at http://www.stat.cmu.edu/ cshalizi/350/lectures/22/lecture-22.pdf

A Beginner's Introduction to Bagging and Boosting: Sutton, Clifton D. "Classification and regression trees, bagging, and boosting." Handbook of Statistics 24 (2005).

4.2 Advanced Readings

Application to Pharmaceuticals: Deconinck, E., et al. "Classification of drugs in absorption classes using the classification and regression trees (CART) methodology." *Journal of Pharmaceutical and Biomedical Analysis* 39.1 (2005): 91-103.

A Review of CART in Public Health: Lemon, Stephenie C., Peter Friedmann, and William Radowski "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression." *Annals of Behavioral Medicine* 26.3 (2003): 172-181.

Applications in Epidemiology: Marshall, Roger J. "The use of classification and regression trees in clinical epidemiology." *Journal of Clinical Epidemiology* 54.6 (2001): 603-609.

Applied Public Health example – Smoking: Nishida, Nobuko, et al. "Determination of smoking and obesity as periodontitis risks using the classification and regression tree method." *Journal of Periodontology* 76.6 (2005): 923-928.

The Classic: Breiman JH, L. Olshen, RA Friedman, and Charles J. Stone. "Classification and Regression Trees." Wadsworth International Group (1984).

4.3 Implementation References

SAS: Gordon, Leonard "Using Classification and Regression Trees (CART) in SAS Enterprise Miner^{TM}For Applications in Public Health." *SAS Global Forum 2013*

Stata (failure only, includes code): v an Putten, Wim. "CART: Stata module to perform Classification and Regression Tree analysis." *Statistical Software Components* (2006). Code and associated materials can be found at: http://econpapers.repec.org/software/bocboc and http://fmwww.bc.edu/repec/bocode/c/cart.ado

R: Official help guide for `rpart` available at http://cran.r-project.org/web/packages/rpart/rpart.pdf

References

1. Deconinck, E., et al. "Classification of drugs in absorption classes using the classification and regression trees (CART) methodology." *Journal of Pharmaceutical and Biomedical Analysis* 39.1 (2005): 91-103.

2. Friedman, Jerome H., and Jacqueline J. Meulman. "Multiple additive regression trees with application in epidemiology." *Statistics in Medicine* 22.9 (2003): 1365-1381.

3. Hess, Kenneth R., et al. "Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma." *Clinical Cancer Research* 5.11 (1999): 3403-3410.

4. Kramer, Stefan, and Gerhard Widmer. "Inducing classification and regression trees in first order logic." *Relational Data Mining.* Springer-Verlag New York, Inc., 2001.

5. Lawrence, Rick L., and Andrea Wright. "Rule-based classification systems using classification and regression tree (CART) analysis." *Photogrammetric Engineering & Remote Sensing* 67.10 (2001): 1137-1142.

6. Lemon, Stephenie C., Peter Friedmann, and William Radowski "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression." *Annals of Behavioral Medicine* 26.3 (2003): 172-181.

7. Loh, WeiYin. "Classification and regression tree methods." *Encyclopedia of statistics in quality and reliability* (2008).

8. Marshall, Roger J. "The use of classification and regression trees in clinical epidemiology." *Journal of Clinical Epidemiology* 54.6 (2001): 603-609.

9. Nishida, Nobuko, et al. "Determination of smoking and obesity as periodontitis risks using the classification and regression tree method." *Journal of Periodontology* 76.6 (2005): 923-928.

10. Breiman JH, L. Olshen, RA Friedman, and Charles J. Stone. "Classification and Regression Trees." Wadsworth International Group (1984).

11. Pang, Herbert, et al. "Pathway analysis using random forests classification and regression." *Bioinformatics* 22.16 (2006): 2028-2036.

12. Shalizi, Cosma "Classification and Regression Trees" unpublished report on data mining procedures available at http://www.stat.cmu.edu/ cshalizi/350/lectures/22/lecture-22.pdf

13. Stark, Katharina DC, and Dirk U. Pfeiffer. "The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology-an example." *Intelligent Data Analysis* 3.1 (1999): 23-35.

14. Steinberg, Dan, and Phillip Colla. "CART: classification and regression trees." The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series (1997): 179-201.

15. Sutton, Clifton D. "Classification and regression trees, bagging, and boosting." Handbook of Statistics 24 (2005).

16. Temkin, Nancy R., et al. "Classification and regression trees (CART) for prediction of function at 1 year following head trauma." *Journal of neurosurgery* 82.5 (1995): 764-771.

17. Gordon, Leonard "Using Classification and Regression Trees (CART) in SAS Enterprise Miner™For Applications in Public Health." *SAS Global Forum 2013* (2013)

18. Thompson, Wayne on a LinkedIn SAS forum. Accessed 2013, availibe at http://www.linkedin.com/gro is-cost-SAS-Enterprise-35222.S.57868066

19. van Putten, Wim "CART: Stata module to perform Classification And Regression Tree analysis" *Statistical Software Components* (2006)

20. Cox, Nicholas J. in response to a Stata listserv question available at http://www.stata.com/statalist/a 02/msg00986.html

Glossary

Gini Impurity Criterion: The Gini in this context is a measure of misclassification, calculating how often some random data element from the data set would be incorrectly specified if it were randomly specified – so how much better is the classification from random assignment.

Leaf: Lower most node on the tree, also called a terminal node.

Node: A variable that significantly contributes to model explanatory power (based on the splitting criterion).

Recursive Partitioning: The method by which CART operates – subsequently dividing the data into smaller portions to isolate most important variables.

Root: Top most node on the tree – the most important explanatory variable.

Splitting (Purity) Criterion: In CART some criteria needs to be used to determine the "best" primary way to split the data. While there are different criterion available, the Gini (im)purity criterion is most widely used. In CART, this process of splitting makes makes the tree more "pure."