Paper 089-2013

Using Classification and Regression Trees (CART) in SAS® Enterprise Miner^{TM} For Applications in Public Health.

Leonard Gordon, University of Kentucky, Lexington, KY

ABSTRACT

Classification and regression trees (CART) - a non-parametric methodology- were first introduced by Breiman and colleagues in 1984. In this paper they are employed using SAS® Enterprise Miner™ and several examples are given to demonstrate their use. CARTs are underused (especially in public health) and they have the ability to divide populations into meaningful subgroups which will allow the identification of groups of interest and enhance the provision of products and services accordingly. They provide a simple yet powerful analysis. It is hoped that their value is demonstrated and this will enhance their increased use in data analysis.

Keywords: classification and regression trees, CART

INTRODUCTION

Classification and Regression trees (CART) were introduced by Breiman et al in 1984. The main idea behind tree methods is to recursively partition the data into smaller and smaller strata in order to improve the fit as best as possible. They partition the sample space into a set of rectangles and fit a model in each one. The sample space is originally split into two regions. The optimal split is found over all variables at all possible split points. For each of the two regions created this process is repeated again. Hence some researchers have termed the method recursive partitioning. The major components of the CART methodology are the selection and stopping rules. The selection rule determines which stratification to perform at every stage and the stopping rule determines the final strata that are formed. Once the strata have been created the impurity of each stratum is measured. The heterogeneity of the outcome categories within a stratum is referred to as “node impurity”.
Classification trees are employed when the outcome is categorical and regression trees are employed when the outcome is continuous. Classification trees can take most forms of categorical variables including indicator, ordinal and non-ordinal variables and are not limited to the analysis of categorical outcomes with two categories.

There are three commonly used measures for node impurity in classification trees. They are misclassification error, Gini index and cross-entropy or deviance. While the three of them are similar, the latter two are differentiable and easier to optimize numerically. Additionally, the Gini index and cross-entropy are more sensitive to changes in the node probabilities. As a result, they are favorites in computer algorithms designed for classification trees and used more often. The measure of node impurity in regression trees is least squares. For the reader interested in the theoretical properties and how they are calculated, the book by Hastie et al is a good resource.

CARTs are not as popular compared to traditional statistical methods because of the lack of tests to evaluate the goodness of fit of the tree produced and the relatively short span that they have been around. They are typically model free in their implementation. Howbeit, a model based statistic is sometimes used for a splitting criterion. The main idea of a classification tree is a statistician’s version of the popular twenty questions game. Several questions are asked with the aim of answering a particular research question at hand. However, they are advantageous because of their non-parametric and non-linear nature. They do not make any distribution assumptions and treat the data generation process as unknown and do not require a functional form for the predictors. They also do not assume additivity of the predictors which allows them to identify complex interactions. Tree methods are probably one of the most easily interpreted statistical techniques. They can be followed with little or no understanding of Statistics and to a certain extent follow the decision process that humans use to make decisions. In this regard, they are conceptually simple yet present a powerful analysis (Hastie et al 2009).

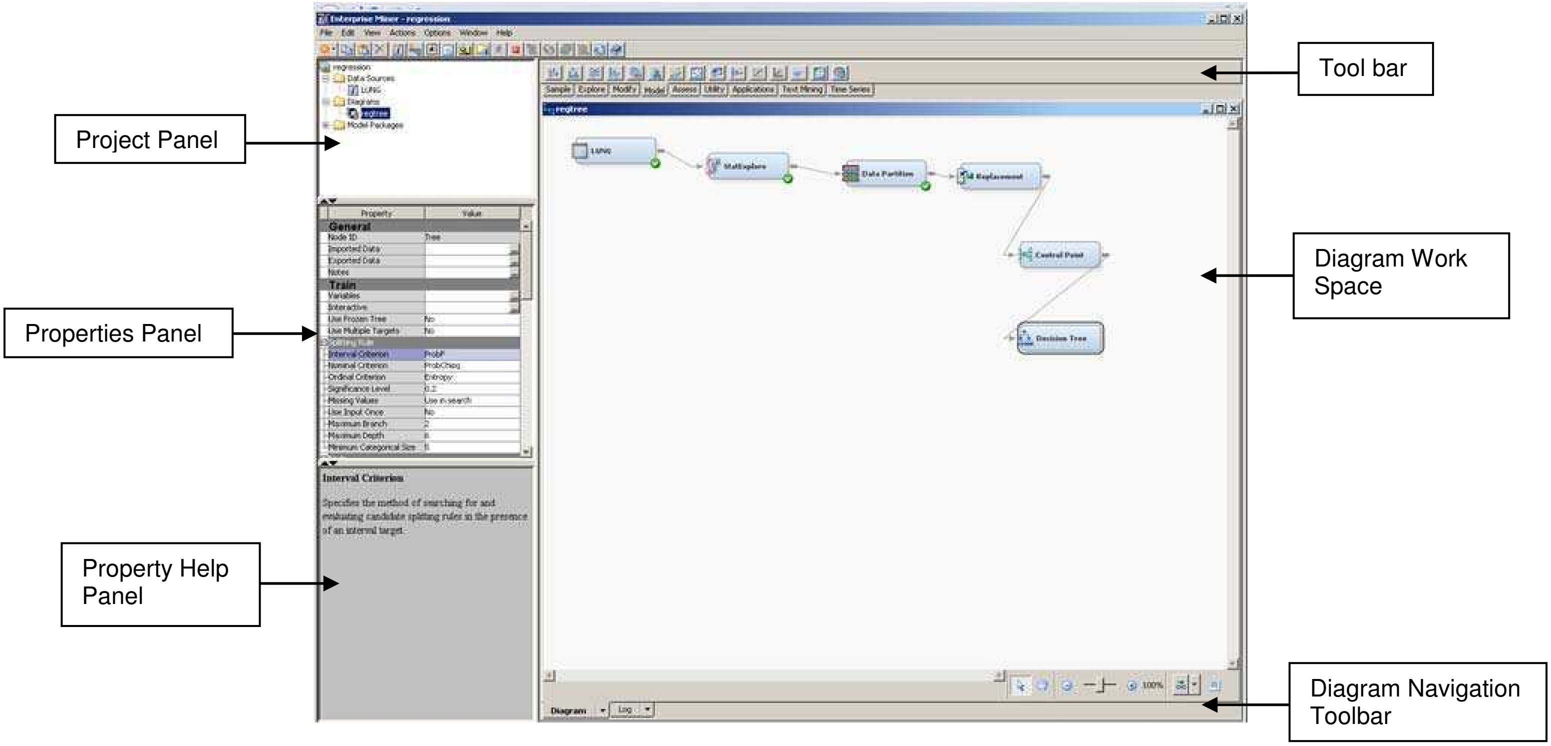
Two illustrations are presented to demonstrate the use of CART. In the first scenario the outcome is categorical and a classification tree will be used. In the second scenario the outcome is continuous and a regression tree will be used. The first data set for this illustration is a publicly available data set on the survival of lung cancer of 228 patients at the Mayo Clinic. The outcome variable was weight loss for the purpose of this analysis. The variables used and the dataset have been defined elsewhere (Gordon 2010).
The second dataset was obtained from the Keokuk County Rural Health Study (KCRHS), a population-based prospective cohort study that began in 1994 to primarily assess the prevalence of injury and respiratory diseases in an agricultural population. The dataset and variables has been described elsewhere (Merchant et al 2005. Merchant et al 2002).The study sample comprises of 565 children who range in age from birth up to 18 years. Children were

selected by stratified random sampling with an oversampling of farms and rural non-farming households. The outcome variable was doctor diagnosed asthma.

This paper aims to demonstrate the value of CART especially in fields like public health where segmentation is important for the identification of at risk groups in order for timely intervention. Additionally, it is hoped that this demonstration will enhance their increased use.

METHODS

It is assumed that the user is familiar with the SAS® Enterprise Miner™ graphical user interface (GUI). However, some explanations are provided for those unfamiliar with the tool. Figure 1 highlights some of the important features of the GUI. More explanations are provided in the Getting Started with SAS® Enterprise Miner™ 7.1 manual.



Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Criterion	ProbF
Nominal Criterion	ProbChisq
Ordinal Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5

Interval Criterion

Specifies the method of searching for and evaluating candidate splitting rules in the presence of an interval target.

Fig.1- The SAS® Enterprise Miner™ GUI

With an available analysis data set, the user creates a new project, a library and a data source using the SAS® Enterprise Miner™ workstation. The project contains the process flow diagrams and information that pertains to them. The library is where the data set is stored and this is a storage location in your operating environment. This can be a company server, the computer or some other external storage device. The data source stores the metadata for the input data set. The procedures for creating these are provided in the Enterprise Miner manual. For the purpose of this illustration the projects are called classification and regression for the classification and regression tree respectively. The library is called SGF2013. Note that the library name does not allow spaces.
The creation of the data source is important as this is where the variable roles are defined. Depending on the nature of the target variables (categorical versus continuous) a classification or regression tree respectively will be produced. The analyst also gets to decide whether they want to use the whole data set or a sample of the data.

Once the project, library and data source have been created the data is ready for analysis. Prior to building the tree exploratory analysis can be done in Enterprise Miner. The analyst also has an opportunity to partition the data set, into a training, validation and test subsamples. The training subset is used to fit various competing models, the validation is used to judge between and choose a final model and the test subset is used to evaluate the chosen model. For this illustration the classification data set was not divided but the regression data set was. The analyst also has the opportunity to deal with missing values by either replacing them or defining how they should be used. Once all the exploration has been done the tree is ready to be built.
The tree that is built can be trained and pruned automatically or interactively. For the purpose of this illustration the trees will be trained and pruned automatically. Once the decision tree node has been selected the rules can be

defined. The analyst can choose the splitting and stopping rules, the maximum number of branches from a node, the maximum depth, the minimum strata size, number of surrogate rules and several other rules that are allowed.

In SAS® Enterprise Miner^{TM}, the whole CART process is driven by a process flow diagram that is created by dragging nodes from the toolbar into the diagram workspace. Once created the process flow diagram can be easily updated or modified. The first node of the process flow is the input data node. This is usually the data source that was created along with the project and library. The StatExplore node follows the data input node. This node allows for exploratory analysis of the data. . The next node is the data partition node. This allows the analyst to divide the dataset into training, validation and test sub datasets. The three datasets and how they are used have been described above. Then follows the replacement node which allows the analyst to replace or recode certain values in the datasets. The next node is the control point node which is used to simplify the process flow diagram by reducing the number of connections between multiple interconnected nodes. This node is followed by the decision tree node. In this node the splitting criterion can be set so that Enterprise Miner is aware that we want to construct a CART diagram which is one of several decision tree methods available in the software. For categorical variables the splitting criterion is set to Gini reduction and for interval or continuous variables the splitting criterion is set to variance reduction. Nodes are connected by an arrow which is created by dragging a pencil from the right edge of the preceding node to the left edge of the succeeding node. For this process flow diagram the input data node is connected to the StatExplore node which is connected the data partition node etc. When a node is run all the other nodes preceding it in the process flow will run in the order that they were presented. See the results section below for a figure of the process flow diagram that was created for one of these projects.

The CART algorithm fits the data by constructing a full grown tree. From the original tree a sequence of subtrees are found. The final subtree is chosen based on a complexity tuning parameter with values greater than or equal to zero. When the parameter is zero the fully grown tree is used. For small values of this parameter large trees will be produced and vice versa. The best tree is chosen by selecting the optimal number of leaves for minimizing the impurity function as chosen by the splitting criterion.

RESULTS

Figure 2 shows the final tree that was built using the classification tree. It was mentioned that this data was not split into a training validation and test data set. This was done for several reasons. Primarily that data set was not very large. There were only 565 children. Furthermore, childhood asthma was borderline rare for this data set.

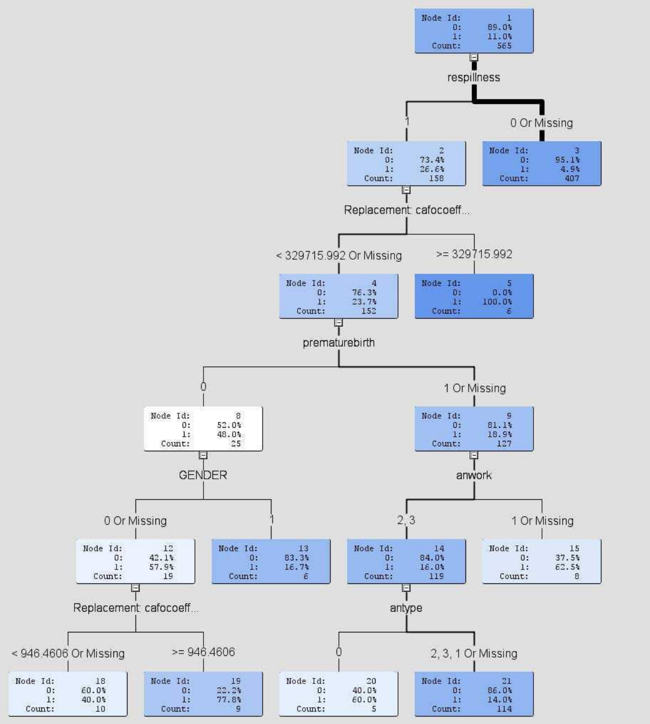

Node Id: 1 0: 89.0% 1: 11.0% Count: 565
respillness
1
0 Or Missing
Node Id: 2 0: 73.4% 1: 26.6% Count: 158
Node Id: 3 0: 95.1% 1: 4.9% Count: 407
Replacement: cafocoeff...
< 329715.992 Or Missing
>= 329715.992
Node Id: 4 0: 76.3% 1: 23.7% Count: 152
Node Id: 5 0: 0.0% 1: 100.0% Count: 6
prematurebirth
0
1 Or Missing
Node Id: 8 0: 52.0% 1: 48.0% Count: 25
Node Id: 9 0: 81.1% 1: 18.9% Count: 127
GENDER
anwork
0 Or Missing
1
2, 3
1 Or Missing
Node Id: 12 0: 42.1% 1: 57.9% Count: 19
Node Id: 13 0: 83.3% 1: 16.7% Count: 6
Node Id: 14 0: 84.0% 1: 16.0% Count: 119
Node Id: 15 0: 37.5% 1: 62.5% Count: 8
Replacement: cafocoeff...
antype
< 946.4606 Or Missing
>= 946.4606
0
2, 3, 1 Or Missing
Node Id: 18 0: 60.0% 1: 40.0% Count: 10
Node Id: 19 0: 22.2% 1: 77.8% Count: 9
Node Id: 20 0: 40.0% 1: 60.0% Count: 5
Node Id: 21 0: 86.0% 1: 14.0% Count: 114

Fig 2. Classification Tree for prediction of Asthma

Below are two different trees that were produced for different proportions when the data was divided into the training, validation and test datasets. This illustrates the important of sample size in decision tree methodology.

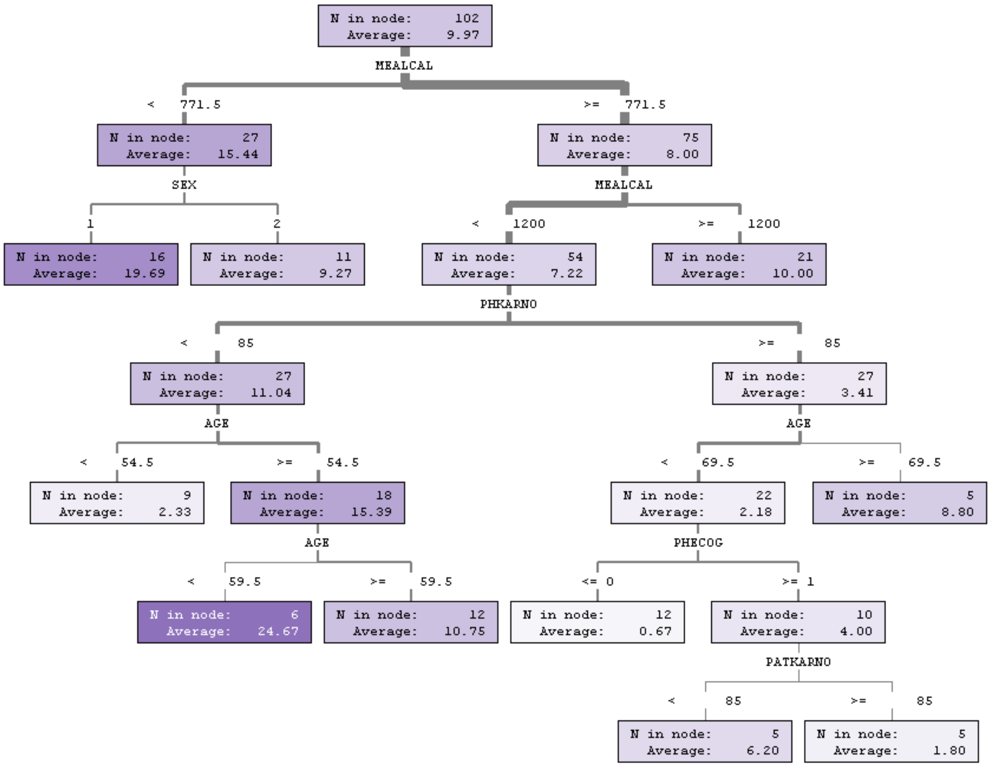


Fig 3. Regression Tree with data divided

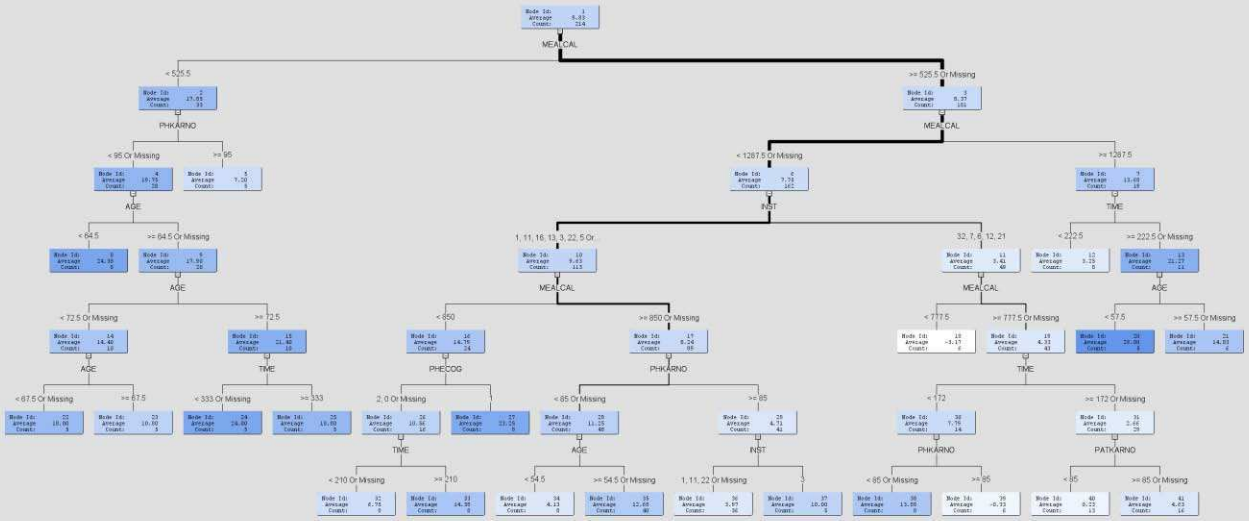
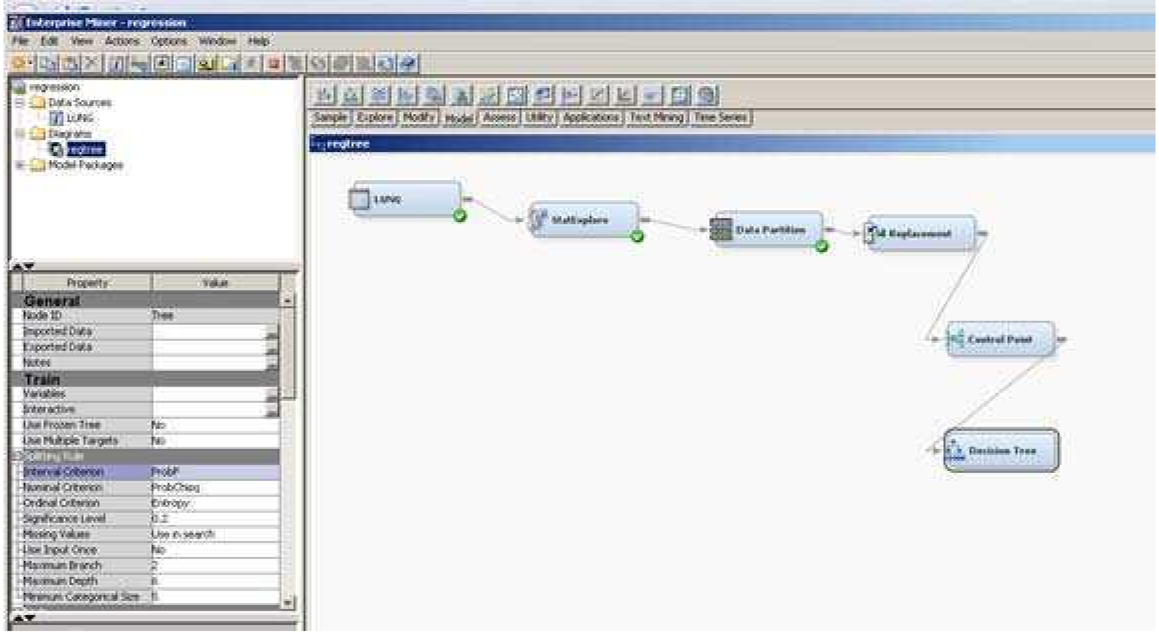


Fig 4. Regression Tree with whole data

The English rules displays the interpretable node definitions for the leaf nodes in a tree model.there are three different nodes in a tree model. The root node, the internal node and the leaf nodes. The root node is the top node of the tree with all the observations. The internal nodes are the non-terminal nodes with the splitting rules. And the leaf nodes are the terminal nodes with final classification for a set of observations. An example of English rules for a tree model is shown in the appendix.

Figure 5 shows the process flow diagram for building one of the trees.



Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Criterion	ProbF
Nominal Criterion	ProbChisq
Ordinal Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5

Fig.5. Process Flow Diagram

DISCUSSION

CARTs are useful tools in decision tree methodology. They allow for the segmentation of a population in meaningful subsets. There are several advantages of CART. One of the advantages is that trees are more efficient at dealing with high dimension data than parametric regression techniques (Westreich 2010). Additionally, they are able to flexibly deal with missing data. There are two ways that they can deal with missing data. The first way is to treat missing observations as a new category. This will allow the difference between missingness and non-missingness of the variables to be seen. The second way is to construct surrogate variables. For a given split, if the original variable is missing, a surrogate variable that mimics the behavior of the original variable will be used for that split. In SAS® Enterprise Miner™ the analysts has the option to define more than one surrogate variable. As such they are not limited and can efficiently manage missing data issues

CARTs are easier to interpret, explain and implement compared to the results obtained from logistic and multiple regression. Finally, they give a pictorial view of the results obtained. They support the age old adage that a picture is worth a thousand words. This helps to involve the general public in the research process and provides them with a clearer understanding of the results of studies. Additionally, they are useful as a time and effort saving technique in terms of making, checking or explaining model assumptions.

One of the strengths of the CART analysis (especially in public health) is that they have the ability to efficiently segment a population into meaningful subsets. This allows researchers to identify sectors of the population that are most likely to be involved with a particular health behavior and adequately target and maximize the distribution of public health resources (Lemon et al 2003). As a result, high risk populations are easily identified and resources allocated accordingly.

Another strength of CART is that they are useful in messy situations and settings in which over fitting is not so problematic. As a result it is advised that they are used in areas where over fitting is not a problem.

There are many physical and social interactions in epidemiologic data. CARTs are capable of indentifying associations, including higher-order interactions that would otherwise go unnoticed. Since the distributions of variables in epidemiologic studies are often not known, CARTs provide a model void of any assumptions about the distribution of the variables, preventing model misspecifications. Additionally, they will help identify some of the many physical and social interactions that are encountered in the field.

In general CART offer a complimentary perspective to traditional regression methods and they are further advantageous in that classification trees can accommodate categorical responses with more than two categories without the complexities offered by other modeling techniques. Finally, CART is non-parametric in that they do no assume a distribution for the data. There are no distribution and parameter requirements.

However, for all their advantages they are not a panacea for issues encountered in data analysis. They are not without disadvantages. One of the limitations of CART analysis is that one is not able to force variables into the model. This is a problem especially in epidemiological studies where we want to control for certain risk factors even when they are not necessarily significant in a model. If the tree process does not deem that risk factor important it might be omitted from the variables in the tree. As a result, they cannot be used for the estimation of average effects in which case the traditional regression methodology will be more appropriate. However, this can be overcome by building the tree interactively instead of automatically.

Because of the ease with which classification trees can be constructed, it is easy to get carried away and perform “data dredging” by just entering all possible independent variables into the analysis (Lemon et al 2003). It is cautioned that classification trees be used with a theoretical consideration of which independent variables to consider.

Another disadvantage of trees is that they can be unstable. Small changes in the data can result in completely different tree. This is because it a particular split changes then all the other splits that are under it change as well.

CONCLUSION

This paper utilizes classification and regression trees (CART) and demonstrates their usefulness for data analysis especially in the field of public health. They have the ability to provide a simple yet powerful analysis of problems in which they are applied.

REFERENCES

Breiman L., Friedman J.H., Olshen R.A. and Stone C. J. (1984). Classification and Regression Trees (2nd Ed.). Pacific Grove, CA; Wadsworth.

Charnigo, Richard.(2009). Data Mining in Public Health. http://www.richardcharnigo.net/CPH636S09/index.html. [Accessed November 2012].

Gordon L. (2010). Using the SAS® System and SAS® Enterprise Miner™ for Data Mining: A study of Cancer Survival at Mayo Clinic. http://support.sas.com/resources/papers/proceedings10/217-2010.pdf . Accessed November 2012.

Hastie, Trevor, Tibshirani, Robert and Friedman Jerome. (2009).The Elements of Statistical Learning. New York, New York; Springer.

Lemon, S C., Roy, Jason and Clark Melissa A. (2003). Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. *Annals of Behavioral Medicine* 26(3):172-81.

Merchant, J.A., Naleway A.L., Svendsen E.R. et al. (2005). Asthma and Farm Exposures in a Cohort of Rural Iowa Children. Environmental Health Perspectives 113(3):350-56.

Merchant J.A., Stromquist A.M., Kelly K.M., Zwerling C, Reynolds S.J., Burmeister L.F. (2002). Chronic disease and injury in an agricultural county: the Keokuk County Rural Health Cohort Study. Journal of Rural Health 18(4):521–35.

Migongo, Alice W., Charnigo, Richard, Love, Margaret M. et al. (2012). Factors Relating to Patient Visit Time With a Physician. *Medical Decision Making* 32: 93-104.

SAS Institute Inc 2011. Getting Started with SAS® Enterprise Miner™ 7.1. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Leonard Gordon
Enterprise: University of Kentucky
Address: 725 Rose Street
City, State ZIP: Lexington, KY 40536
Work Phone: 859-218-2097
Fax: 859-257-6430
E-mail: leonard.gordon@uky.edu

APPENDIX

These are the English rules for the classification tree

```
*------------------------------------------------------------*
 Node = 3
*------------------------------------------------------------*
if respillness IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 3
 Number of Observations = 407
 Predicted: asthma=0  = 0.95
 Predicted: asthma=1  = 0.05

*------------------------------------------------------------*
 Node = 5
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND Replacement: cafocoeffecient >= 329716
then
 Tree Node Identifier   = 5
 Number of Observations = 6
 Predicted: asthma=0  = 0.00
 Predicted: asthma=1  = 1.00

*------------------------------------------------------------*
 Node = 13
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND prematurebirth IS ONE OF: 0
AND Replacement: cafocoeffecient < 329716 or MISSING
AND GENDER IS ONE OF: 1
then
 Tree Node Identifier   = 13
 Number of Observations = 6
 Predicted: asthma=0  = 0.83
 Predicted: asthma=1  = 0.17

*------------------------------------------------------------*
 Node = 15
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND prematurebirth IS ONE OF: 1 or MISSING
AND anwork IS ONE OF: 1 or MISSING
AND Replacement: cafocoeffecient < 329716 or MISSING
then
 Tree Node Identifier   = 15
 Number of Observations = 8
 Predicted: asthma=0  = 0.38
 Predicted: asthma=1  = 0.63

*------------------------------------------------------------*
 Node = 18
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND prematurebirth IS ONE OF: 0
AND Replacement: cafocoeffecient < 946.461 or MISSING
AND GENDER IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 18
 Number of Observations = 10
 Predicted: asthma=0  = 0.60
 Predicted: asthma=1  = 0.40

*------------------------------------------------------------*
 Node = 19
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND prematurebirth IS ONE OF: 0
AND Replacement: cafocoeffecient < 329716 AND Replacement: cafocoeffecient >= 946.461
AND GENDER IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 19
 Number of Observations = 9
 Predicted: asthma=0  = 0.22
 Predicted: asthma=1  = 0.78

*------------------------------------------------------------*
 Node = 20
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND prematurebirth IS ONE OF: 1 or MISSING
AND anwork IS ONE OF: 2, 3
AND antype IS ONE OF: 0
AND Replacement: cafocoeffecient < 329716 or MISSING
then
 Tree Node Identifier   = 20
 Number of Observations = 5
 Predicted: asthma=0  = 0.40
 Predicted: asthma=1  = 0.60

*------------------------------------------------------------*
 Node = 21
*------------------------------------------------------------*
if respillness IS ONE OF: 1
AND prematurebirth IS ONE OF: 1 or MISSING
AND anwork IS ONE OF: 2, 3
AND antype IS ONE OF: 2, 3, 1 or MISSING
AND Replacement: cafocoeffecient < 329716 or MISSING
then
 Tree Node Identifier   = 21
 Number of Observations = 114
 Predicted: asthma=0  = 0.86
 Predicted: asthma=1  = 0.14
```