CLASSIFICATION AND REGRESSION TREE MODELS

Roumen Vesselinov, Ph.D.[1]
Queens College, City University of
New York, USA

ACKNOWLEDGMENTS

The author would like to thank the analysis division at Policy Research Associates, Inc in Delmar, New York; E. Frank, J. Fridriksson, and D. Deliyski from University of South Carolina (USC), Department of Communication Sciences and Disorders (COMD), and J. Mann at the USC School of Medicine and Arnold School of Public Health.

Special thanks to Dr. Valentin Goev from University of National and World Economy (UNWE) for his idea to write this paper.

MOTIVATION

The author has been working on Regression and Classification Tree (CART) models for a long time (see Banks et al, Fridriksson et al, Deliyski et al). This paper tries to answer two new research questions. First, are the CART models comparable to more traditional and standard statistical models like logistic regression, and second, are there any advantages to using the CART models.

METHODS

The logistic regression models are standard statistical tools and they can be found in all graduate texts on statistical methods (e.g. Agresti 2002).

Recursive partitioning methods and CART in particular (Steinberg and Colla, 1995, 1997, Breiman et al., 1984) are part of the data mining family of instruments. Data mining is usually defined as process of identifying valid and understandable, previously unknown, and complex patterns in data. CART is a nonparametric binary recursive partitioning method, which is a very powerful discovery tool for data with complex structure. Data mining and CART in particular are very effective with large datasets with hundreds of thousands of cases. A decision tree (classification or regression) is structured as a sequence of simple questions, and the answers to these questions trace a path down the tree. The end node reached by each case determines the classification or prediction made by the model.

The target variable for a classification tree is categorical (usually with small number of categories), and for a regression tree it is continuous (or ordinal with large number of categories). The goal is to partition the data into relatively homogeneous terminal nodes. The building of the tree structure starts with the first binary split on the most important variable. Then, this "parent" node is split again, etc. The tree with all possible nodes is then pruned in order to get the optimal tree. The optimality is determined by minimizing classification error while producing a parsimonious model. The splitting criterion is Gini impurity coefficient (or similar) and the optimality is based on the predictive accuracy and the penalty for larger trees. On each node the variable that gives the best split (smallest error) is included in the tree. At each node CART produces a list of "competitors" and "surrogates". The former are the predictors which have slightly higher error rate than the chosen variable and the later are the predictors that split the node in a similar way regarding the target variable.

Since some trees may be very sensitive and not robust, an internal validation process is employed in building the optimal tree. Usually a 10-fold cross-validation is used to validate the tree. This procedure

[1] Visiting Assistant Professor, Economics Department, Queens College, City University of New York. For contact, roumen.vesselinov@qc.cuny.edu, or stat@vesselinov.com

ensures independent predictive accuracy for the optimal tree, and the confidence that the resulting tree can be reasonably generalized and used for a completely different set of data. CART is very robust to outliers and other distribution problems and very effective for assessing the reliability of new data predictions. CART is very effective in finding context dependence and high order interaction effects. One variable can appear many times in a tree in different contexts. CART works very well with data in which the baserate of the target variable is very low.

Once the tree is built and internally validated, the importance of each variable can be ascertained. At each splitting point, both the most influential variable and its surrogates are determined by CART. With standard statistical models, the importance of one variable is often "masked" by another variable. For example, in a model using a stepwise procedure, the surrogates would be dropped out of the equation and their actual importance would be obscured. CART solves this "masking" problem by taking into account the improvement measure not only for the primary splitter, but also for the surrogates. The variable importance score is calculated by "looking at the improvement measure attributable to each variable in its role as a surrogate to the primary split. The values of these improvements are summed over each node, totaled for the tree, and scaled relative to the best performing variable" [Steinberg/Colla 1997] forming the variable importance measure. The variable with the highest score is set to 100% and the remaining variables are scaled relative to this variable.

THE STUDY

The original findings were presented in Fridriksson, Frank and Vesselinov (2005). In the current paper we are using the results from the 2005 paper to compare them to the results from a new logistic regression model based on the same variables.

The original study was based on stroke patient data provided by the South Carolina (SC) Office of Research and Statistics. The database comprises of the whole patient population in SC for the 1996-2000 period. All patients who were admitted in hospitals in SC with a diagnosis of stroke were included. The study responds to the acute need to identify and explain the utilization of rehabilitation services (Horn et al., 2000, Rosenfeld, 2002) and who are the people receiving these services. This is extremely important because the primary goal of individuals, who experience disabling stroke, is recovery of the basic skills necessary for effective executions of independent activities of daily living. The purpose of the original study was to investigate the factors associated with utilization of Speech and Audiology Services (SpAS) provided to stroke patients.

Hospital records are very rich source of information but they are not adequately used for research in part because hospital records present certain statistical problems. First, the missing data can impede the estimation of standard statistical models. Second, the nature of the data implies interactions of very high order. In the presence of hundreds of variables there is no satisfactory solution for this problem through standard statistical models. Third, some predictors might be relevant to particular subset of patients but not to all of them, a problem known as "context dependence" which is very hard to accommodate with the standard statistical models. Fourth, outliers and other distribution anomalies cause some problems to the parametric statistical methods, while the CART method deals with them without a glitch and without removing or rescaling them. Fifth, due to the extremely large number of observations the standard statistical models usually end up with most predictor variables being statistically significant and included in the final model even after a stepwise selection which makes the interpretation of the results very cumbersome.

RESULTS

The full CART model of SpAS is presented in the Appendix. As we can see there are only 6 levels of the model. This means that we can classify a patient based on 1 to 6 questions or variables. These questions (except the first one: emergency, non-emergency) may be different for the different paths. Below is presented the first level of the model.

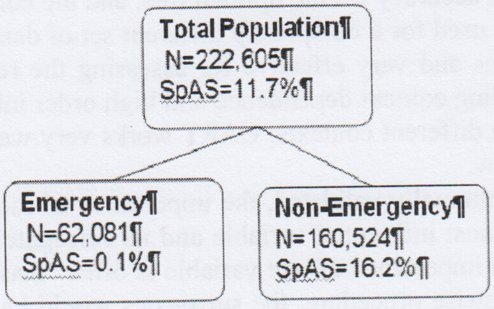


Figure 1. CART Level 1

It shows that overall 11.7% of all patients have received SpAS. Of them, only 0.1% of the patients in the emergency rooms have received these services. The nonemergency patients are enjoying much better services with 16.2% of them receiving SpAS. The second level of the model is presented on Figure 2.

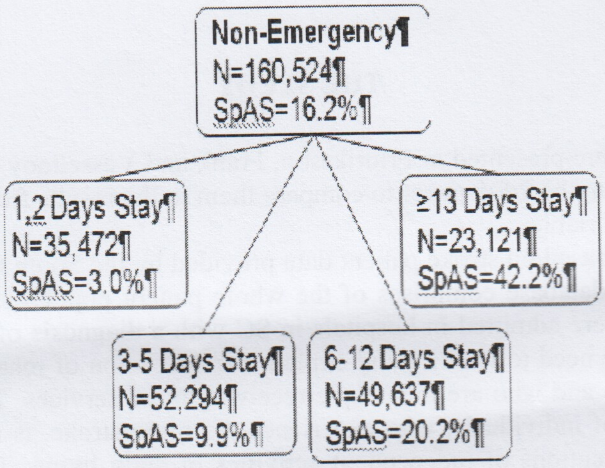


Figure 2. CART Level 2

This means that of all non-emergency patients the largest amount of services receive the ones that stayed 13 or more days in the hospital (42.2%) and the least the patients that stayed for only 1 or 2 days in the hospital (3.0%).

The next 4 levels are interpreted similarly.

There are 20 end nodes altogether. Some of them exhibit high level of service utilization and some exhibit low level of SpAS.

Table 1

Patterns for SpAS for Stroke Patients

Node (#)	**SpAS (%)**	**Patients who received a *low* amount of SpAS …**
1	2.8	Stayed one or two days in hospital in inpatient setting.
3	3.3	Stayed 3-5 days in hospital on their first visit without major diagnosis.
5	6.3	Stayed 3-5 days in hospital as inpatient, on their second or later visit.
7	8.2	Stayed 6-12 days in hosp, on 1st or 2nd visit as inpatient, no major diagnosis.
		Patients who received a *high* amount of SpAS …
2	39.6	Stayed 1-2 days in hospital in rehabilitation.
13	47.8	Stayed 6-12 days in hospital on their third or later visit in rehabilitation.
6	56.4	Stayed 3-5 days in hospital in rehabilitation on their 2nd or later visit.
9	65.4	Stayed 6-12 days in hospital, on their first or second visit in rehabilitation.
20	78.7	Stayed 13 or more days in rehabilitation.

COMPARISON[1]

The quality of classification models like CART and logistic regression models is traditionally assessed by the following criteria presented in Table 2.

Table 2

Criteria for Evaluation

	Predicted value	
True (Actual) Value	0 (No SpAS)	1 (Yes, SpAS)
0 (No SpAS)	A	C
1 (Yes, SpAS)	B	D

Sensitivity = D/(C+D) or Sensitivity=True Positive/All Positive;
Specificity=A/(A+C) or Specificity=True Negative/All Negative;
False Negative = B/(A+B); False Positive = C/(A+C)
Total Correct = (A+D)/(A+B+C+D)

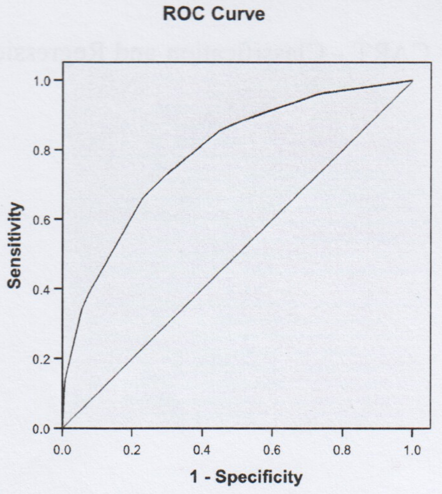
The results of the comparison between the CART model and the logistic regression model for SpAS are presented in Table 3.

Table 3

Comparison Results

Criteria (%)	**CART Model**	**Logistic Regression**
Sensitivity	72.9	77.5
Specificity	70.2	75.1
Total Correct	70.6	75.3
False Negative	4.4	3.8
False Positive	25.0	24.9
ROC Area Under the Curve (AUC)	79.0	84.0

AUC is the overall criterion that is usually used for comparison and any classification model with AUC about 70% or more is considered adequate. The ROC for the CART model is presented in Figure 3.



Diagonal segments are produced by ties.

Figure 3. ROC for the CART Model

[1] The full logistic regression model includes 15 stepwise selected variables. It is not presented here because of the page limitations but it is available from the author upon request.

CONCLUSION

The results of the comparison of the CART model and the logistic regression model show that they are comparable and both have good statistical qualities. The only problem with both models is that the false negatives are about 25 % but this is due to the fact that the baserate is relatively low, 11.7 %.

The advantages of the CART models are that in addition to the predicted values we can trace a path of development to each particular end node. For example, it was expected that the longer the patients stay in the hospital the bigger the probability of receiving SpAS (e.g. Node 20). But there are some unexpected and revealing results as well. For example there is a group of people (Node 2) who spent only one or two days in hospital but they received considerable amount of SpAS.

In conclusion, CART models should be used in similar circumstanced in addition to or instead of the standard statistical models like the logistic regression models because they perform on average at least as well as the standard models and in addition they give the researchers a chance to discover the different paths that patients go to receive the different prediction scores.

References

Agresti, A. (2002): Categorical Data Analysis, 2nd ed., Wiley-Interscience.

Banks, S., Robbins, P., Silver, E., Vesselinov, R., Steadman, H., Monahan, J., Mulvey, E., Appelbaum, P., Grisso, T., Roth, L. (2004): Multiple Models Approach to Violence Risk Assessment Among People with Mental Disorder, *Criminal Justice and Behavior*, Vol. 31, Number 3, June 2004, pp.324-340.

Breiman, L, Friedman, J, Olsen, R, Stone, C. (1984): Classification and Regression Trees. Pacific Grove: Wadsworth.

Deliyski, D., Shaw, H., Evans, M., Vesselinov, R. (2006): Regression Tree Approach to Studying Factors Influencing Acoustic Voice Analysis. *Folia Phoniatrica et Logopaedica*, 58 (2).

Fridriksson, J., Frank, E., Vesselinov, R., (2005): Utilization of Speech-Language Pathology and Audiology Services in Stroke Patients, *Journal of Medical Speech-Language Pathology*, Volume 13, Number 4, pp. 235-243.

Horn, W., Yoels, W., Bartolucci, A. (2000): Factors associated with patients' participation in rehabilitation services: a comparative injury analysis 12 months post-discharge. Disability and Rehabilitation, 22(8), 358-362.

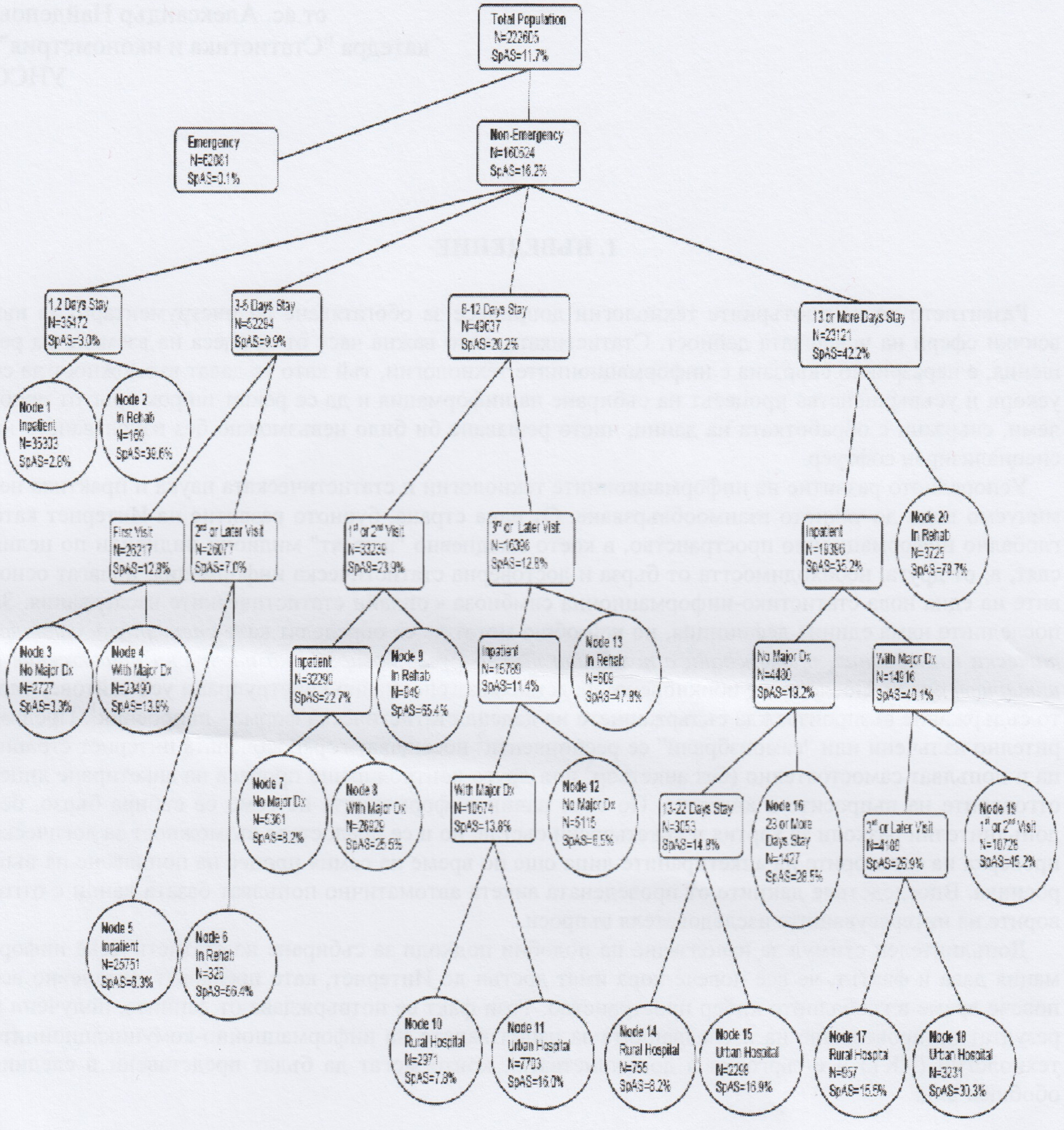
Rosenfeld, M. (2002): Report on the ASHA Speech-Language Pathology Health Care Survey, American Speech-Language-Hearing Association.

Steinberg, D., Colla, P. (1995): CART - Tree-Structured Non-Parametric Data Analysis. San Diego, CA: Salford Systems.

Steinberg, D., Colla, P. (1997): CART - Classification and Regression Trees. San Diego, CA: Salford Systems.

APPENDIX

CART Model of SpAS


Total Population N=222605 SpAS=11.7%
Emergency N=62081 SpAS=0.1%
Non-Emergency N=160524 SpAS=16.2%
1.2 Days Stay N=35472 SpAS=3.0%
3-5 Days Stay N=62294 SpAS=9.9%
6-12 Days Stay N=49637 SpAS=20.2%
13 or More Days Stay N=23121 SpAS=42.2%
Node 1 Inpatient N=35303 SpAS=2.8%
Node 2 In Rehab N=169 SpAS=39.6%
First Visit N=26217 SpAS=12.8%
2nd or Later Visit N=26077 SpAS=7.0%
1st or 2nd Visit N=33239 SpAS=23.9%
3rd or Later Visit N=16396 SpAS=12.8%
Inpatient N=19396 SpAS=35.2%
Node 20 In Rehab N=3725 SpAS=79.7%
Node 3 No Major Dx N=2727 SpAS=3.3%
Node 4 With Major Dx N=23490 SpAS=13.9%
Inpatient N=32290 SpAS=22.7%
Node 9 In Rehab N=949 SpAS=55.4%
Inpatient N=15789 SpAS=11.4%
Node 13 In Rehab N=609 SpAS=47.8%
No Major Dx N=4480 SpAS=13.2%
With Major Dx N=14916 SpAS=40.1%
Node 7 No Major Dx N=5361 SpAS=8.2%
Node 8 With Major Dx N=26929 SpAS=25.5%
With Major Dx N=10674 SpAS=13.8%
Node 12 No Major Dx N=5115 SpAS=6.5%
13-22 Days Stay N=3053 SpAS=14.8%
Node 16 23 or More Days Stay N=1427 SpAS=28.5%
3rd or Later Visit N=4188 SpAS=26.9%
Node 19 1st or 2nd Visit N=10728 SpAS=45.2%
Node 5 Inpatient N=25751 SpAS=6.3%
Node 6 In Rehab N=325 SpAS=56.4%
Node 10 Rural Hospital N=2971 SpAS=7.8%
Node 11 Urban Hospital N=7703 SpAS=16.0%
Node 14 Rural Hospital N=755 SpAS=8.2%
Node 15 Urban Hospital N=2299 SpAS=16.9%
Node 17 Rural Hospital N=957 SpAS=15.5%
Node 18 Urban Hospital N=3231 SpAS=30.3%