Optimal Data Analysis
Vol. 8 (April 14, 2019), 84-93

2155-0182/10/$3.00

Optimizing Suboptimal Classification Trees: Matlab® CART Model Predicting Probability of Lower Limb Prosthesis User's Functional Potential

Paul R. Yarnold, Ph.D. and Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC Linden Consulting Group, LLC

After *any* algorithm which controls the growth of a classification tree model has completed, the resulting model must be pruned in order to explicitly maximize predictive accuracy normed against chance. This article illustrates manually-conducted maximum-accuracy pruning of a classification and regression tree (CART) model that was developed to predict the functional capacity of lower limb prosthesis users.

Recent research[1] used a CART model (Figure 1) to "…assist with the rehabilitation teams' care planning, providing probabilities of functional potential for the lower limb prosthesis user" (p. 2). The nonrandomized study compared samples of *limited* (class L, N=123) *vs*. *unlimited* mobility (class U, N=431) ambulatory patients.[2] The CART model[3] used eight attributes to define nine predicted patient strata: classification accuracy ranged from 53.8% to 96.7%.

Table 1 summarizes the classification accuracy obtained using CART to classify the total sample of 554 observations in training analysis. *Sensitivity* results indicate the model correctly classified 77.24% of 123 limited ambulatory people—this percent of predictive accuracy compares well *vs*. chance for which, if defined as a uniform random number, 50% accuracy is expected.[4] The model also correctly classified 90.26% of 431 unlimited ambulatory people—comparing *very* well *vs*. chance.

The *e*ffect *s*trength for *s*ensitivity (ESS) index (a function of the mean sensitivity across classes) is used to summarize model overall classification accuracy after adjusting for the performance expected chance: ESS=0 is the accuracy expected by chance; ESS=100 is perfect accuracy; and ESS<0 is accuracy worse than expected by chance.[4]

The rule-of-thumb used to qualitatively summarize effect strength after adjusting for chance is: ESS<25 is a relatively weak effect; ESS<50 a moderate effect; ESS<75 a relatively strong effect; and ESS$\geq$75 indicates increasingly strong levels of effect size.[5-8]

For the model in Figure 1, training ESS= 67.5—a *relatively strong* effect.

2155-0182/10/$3.00

Figure 1: Fully-Grown Matlab® CART Classification Tree Model[1]

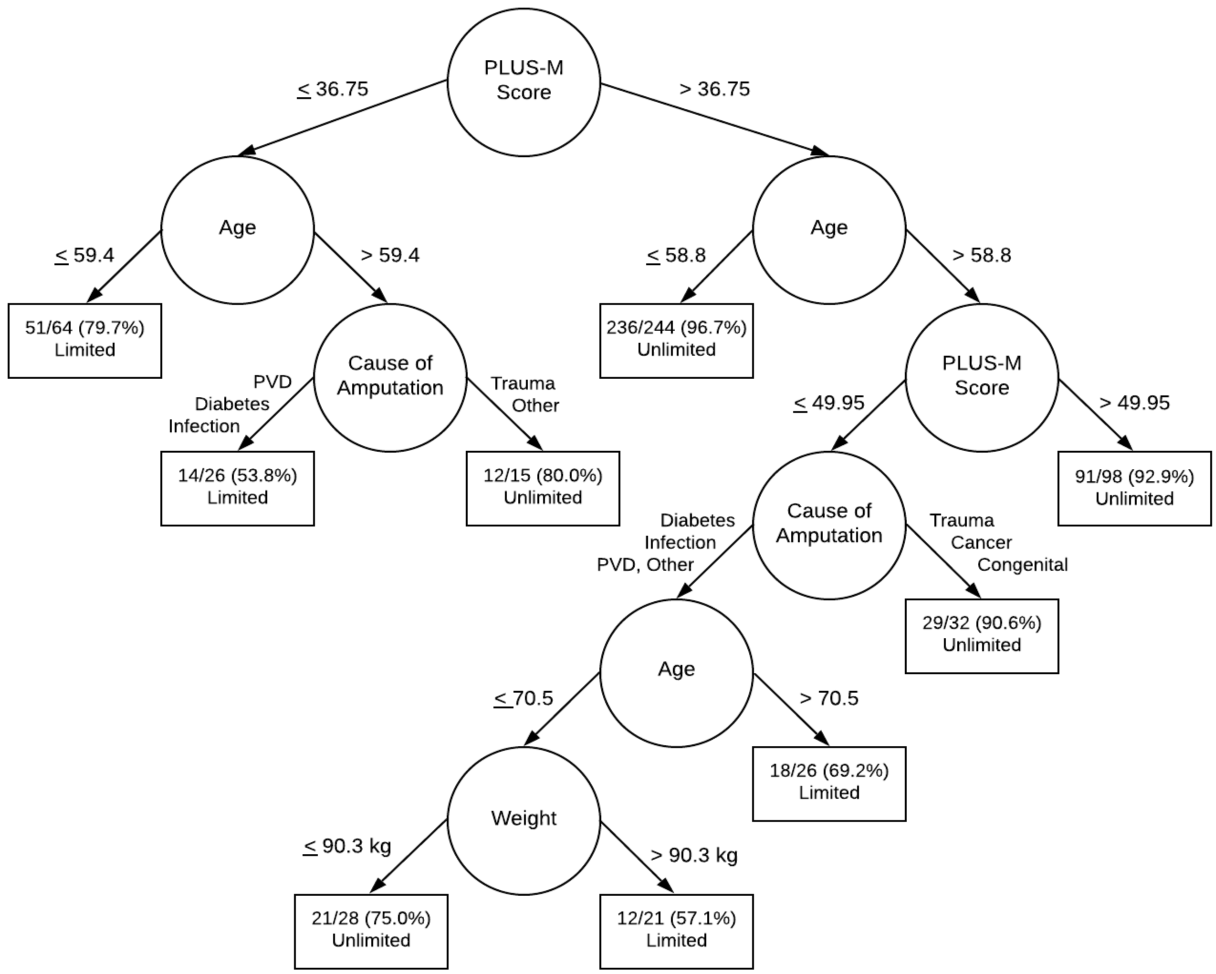


Table 1: Confusion Table for Fully Grown Matlab CART Model in Figure 1 (*Sens*=Sensitivity; *PV*=Predictive Value)

	Predicted		
Actual	Limited	Unlimited	*Sens*
Limited	95	28	77.24
Unlimited	42	389	90.26
PV	69.34	93.29	

Predictive value findings in training analysis indicate the model is correct 69.34% of the time that it makes a point prediction that an observation is from class L (137 of such point predictions were made), and 93.29% of the time when predicting an observation is from class U (417 such predictions). The *e*ffect *s*trength for *p*redictive value (ESP) index, a function of mean PV over class categories, summarizes the model omnibus chance-adjusted PV *for the application*: unlike sensitivity, model PV varies over base rate and thus is estimated for different base rates.[9] Qualitative strength is determined as for ESS: here ESP=62.63—a relatively strong effect. In novometric theory, 95% exact discrete confidence intervals are obtained for the model and for chance (for all performance measures): overlap of CIs indicates the absence (lack) of statistical significance.[10]

2155-0182/10/$3.00

ESS and ESP indices assess translational chance-adjusted accuracy obtained by the model when used in application with the entire sample or with individual subjects, respectively. When considered from an applied perspective, a fully-loaded model which explicitly maximizes the empirically-achievable ESS (or ESP) offers the most information available regarding alternative pathways toward and away from the outcome.

However, when evaluated from a theoretical perspective such models are considered over-fit: the sought-after model reflects both explanatory *power* (strongest possible ESS) and *parsimony* (fewest possible outcome strata). Theoretical quality of an empirical model is defined in terms of the discrepancy (distance) between achieved *vs*. the corresponding perfect model.[11] This is quantified by the D (distance) statistic that norms ESS for parsimony: smaller D values indicate better combinations of accuracy and parsimony, and D=0 indicates a perfect model (number of strata is a function of measure granularity and attribute distributions).[12]

For the fully-grown S-PLUS tree model, ESS normed for parsimony is D_{ESS}=4.33, so 4.33 additional strata having equivalent mean ESS are needed to obtain a "perfect" model (D_{PV}=5.37). All fully-grown tree models require optimal pruning to explicitly maximize ESS.[13]

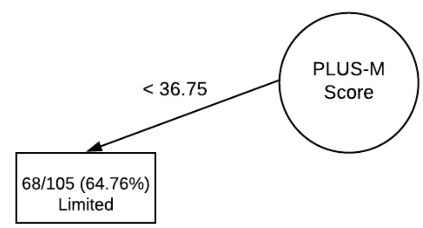
Optimal Pruning to Maximize ESS

The optimal pruning algorithm consists of two simple steps which are straightforward to apply manually. Optimal pruning is the *only* way to *ensure* that a classification tree model explicitly maximizes ESS for the sample—no matter what algorithm was used in its development. Optimal pruning was previously demonstrated for CART and CTA tree models.[13-16]

The first step of optimal pruning requires identifying all sub-branches of every emanating branch. Imagine a left-hand branch having three nodes: A (root), B (middle attribute), and C (end of branch). There are two nested sub-branches: one involving only nodes A and B (C collapsed into B), the other involving only node A (C and B collapsed into A). Here the *left* branch having *three* nodes (A, B, C) is called L3; the trimmed *left* branch having *two* nodes (A, C collapsed into B) is L2; and the trimmed *left* branch with only *one* node (C and B collapsed into A) is L1. Imagine also the hypothetical tree model has a right-hand branch with two nodes: A (the sides share the root attribute) and D (end of branch). The *right* branch having *two* attributes (A, D) is called R2, and the trimmed *right* branch with *one* attribute (D collapsed into A) is called R1.

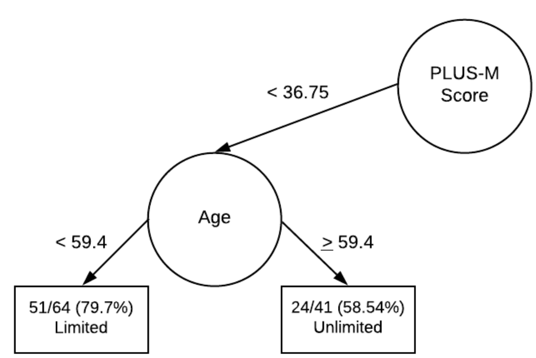
The first step in optimal pruning in this application is seen in Figures 2A through 3F.

Figure 2A: L1 Sub-Branch and Confusion Table



Actual	*Predicted* Class L	Class U
Class L	68	---
Class U	37	---

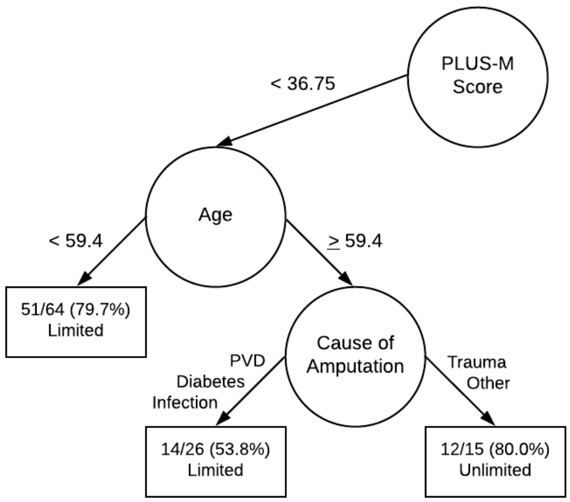
Figure 2B: L2 Sub-Branch and Confusion Table



Actual	*Predicted* Class L	Class U
Class L	51	17
Class U	13	24

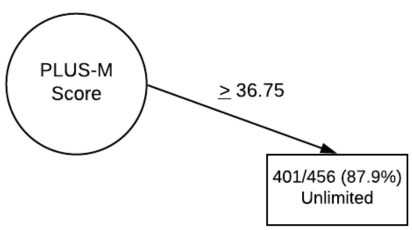
2155-0182/10/$3.00

Figure 2C: L3 Sub-Branch and Confusion Table



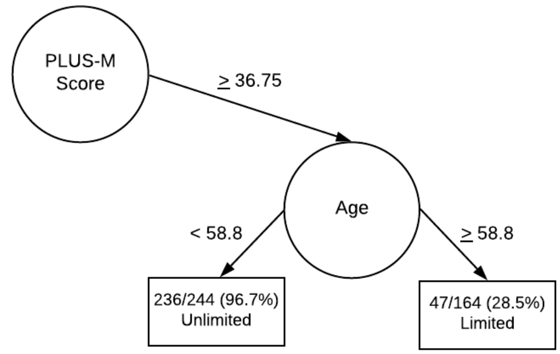
	Predicted	
Actual	Class L	Class U
Class L	65	3
Class U	25	12

Figure 3A: R1 Sub-Branch and Confusion Table



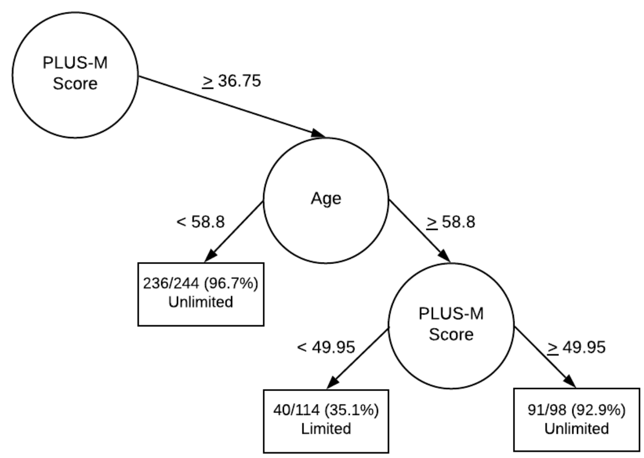
	Predicted	
Actual	Class L	Class U
Class L	---	55
Class U	---	401

Figure 3B: R2 Sub-Branch and Confusion Table



	Predicted	
Actual	Class L	Class U
Class L	47	8
Class U	117	236

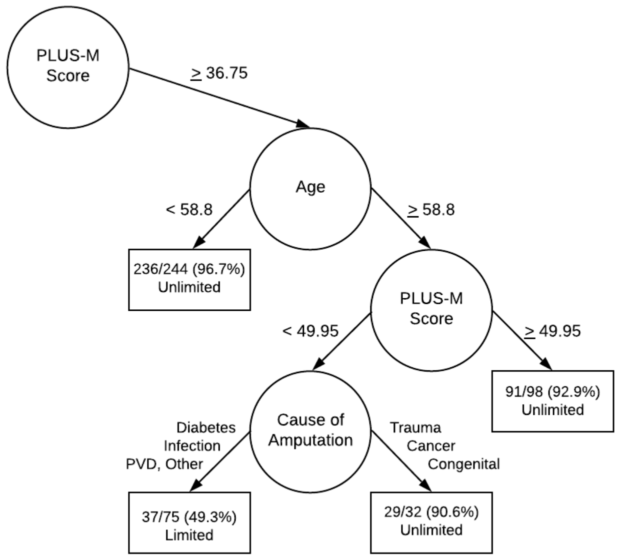
Figure 3C: R3 Sub-Branch and Confusion Table



	Predicted	
Actual	Class L	Class U
Class L	40	15
Class U	74	327

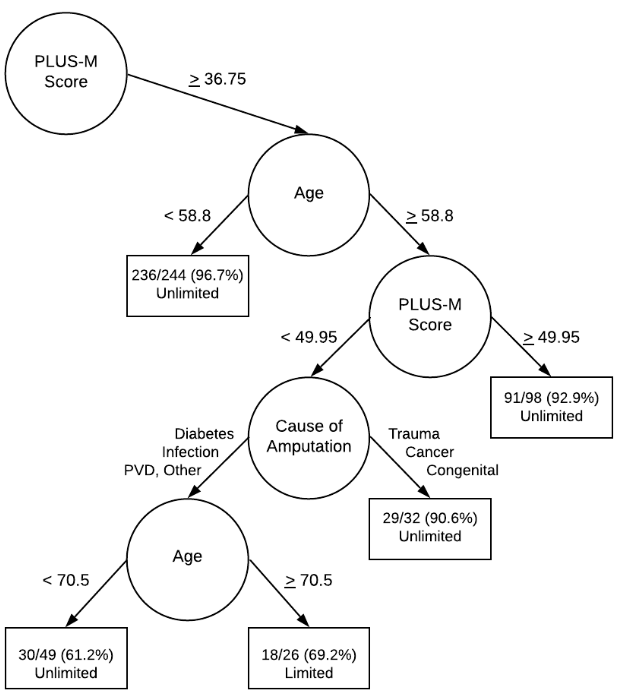
2155-0182/10/$3.00

Figure 3D: R4 Sub-Branch and Confusion Table



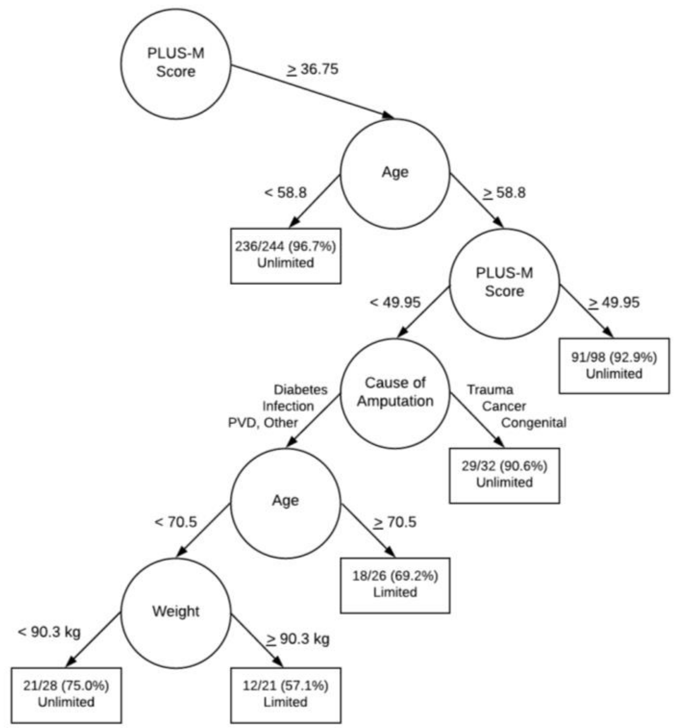
	Predicted	
Actual	Class L	Class U
Class L	37	18
Class U	38	356

Figure 3E: R5 Sub-Branch and Confusion Table



	Predicted	
Actual	Class L	Class U
Class L	18	37
Class U	8	386

Figure 3F: R6 Sub-Branch and Confusion Table



	Predicted	
Actual	Class L	Class U
Class L	30	25
Class U	17	377

Step Two of optimal pruning requires creating a confusion table (rows indicate actual class category, columns indicate class category predicted by the model) for all 18 combinations of left and right sub-branch: {L1-R1, L1-R2 … L1-R6}, {L2-R1, L2-R2 … L2-R6}, {L3-R1, L3-R2 … L3-R6}.

2155-0182/10/$3.00

Table 2 gives integrated confusion tables and associated ESS and D statistics for every unique combination of left and right branches: **red** font highlights the specific combinations yielding strongest ESS and D statistics.

The model (combination) with greatest ESS has greatest *translational* significance—providing the most accurate classification possible given present knowledge. The model having lowest associated D has greatest theoretical significance—the closest approximation to a perfect model given present knowledge.

Table 2: Classification Results for Every Combination of Left (L1-L3) and Right (R1-R6) Sub-Branch

Model	Actual	Predicted Class L	Predicted Class U	Statistics
L1-R1	Class L	68	55	ESS=46.8, **D=2.27**
	Class U	37	401	
L1-R2	Class L	115	8	ESS=54.0, D=2.55
	Class U	154	236	
L1-R3	Class L	108	15	ESS=62.5, D=2.40
	Class U	111	327	
L1-R4	Class L	105	18	ESS=68.0, D=2.36
	Class U	75	356	
L1-R5	Class L	86	37	ESS=59.5, D=4.09
	Class U	45	386	
L1-R6	Class L	98	25	ESS=67.1, D=3.43
	Class U	54	377	
L2-R1	Class L	51	72	ESS=55.6, D=2.40
	Class U	13	425	
L2-R2	Class L	98	25	ESS=46.3, D=4.63
	Class U	130	260	
L2-R3	Class L	91	32	ESS=54.1, D=4.24
	Class U	87	351	
L2-R4	Class L	88	35	ESS=59.7, D=4.05
	Class U	51	380	

2155-0182/10/$3.00

Model	Confusion Table	
L2-R5	Predicted	
Actual	Class L	Class U
Class L	69	54
Class U	21	410
	ESS=51.2, D=6.66	

Model	Confusion Table	
L2-R6	Predicted	
Actual	Class L	Class U
Class L	81	42
Class U	30	401
	ESS=58.9, D=5.58	

Model	Confusion Table	
L3-R1	Predicted	
Actual	Class L	Class U
Class L	65	58
Class U	25	413
	ESS=47.1, D=4.49	

Model	Confusion Table	
L3-R2	Predicted	
Actual	Class L	Class U
Class L	112	11
Class U	142	248
	ESS=54.6, D=4.15	

Model	Confusion Table	
L3-R3	Predicted	
Actual	Class L	Class U
Class L	105	18
Class U	99	339
	ESS=62.8, D=3.56	

Model	Confusion Table	
L3-R4	Predicted	
Actual	Class L	Class U
Class L	102	21
Class U	63	368
	ESS=68.3, D=3.25	

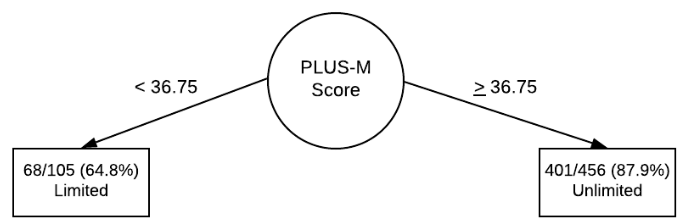
Model	Confusion Table	
L3-R5	Predicted	
Actual	Class L	Class U
Class L	83	40
Class U	33	398
	ESS=59.8, D=5.37	

Model	Confusion Table	
L3-R6	Predicted	
Actual	Class L	Class U
Class L	95	28
Class U	42	389
	ESS=67.5, D=4.33	

Explicitly optimized models are the combination(s) of left and right sub-branches with associated confusion table yielding the maximum ESS, and yielding minimum D.

As seen in Table 2, the L1-R1 combination illustrated in Figure 4 has lowest D=2.27, indicating that 2.27 additional effects with mean ESS=68.3 are needed to obtain a theoretically perfect model.[12]

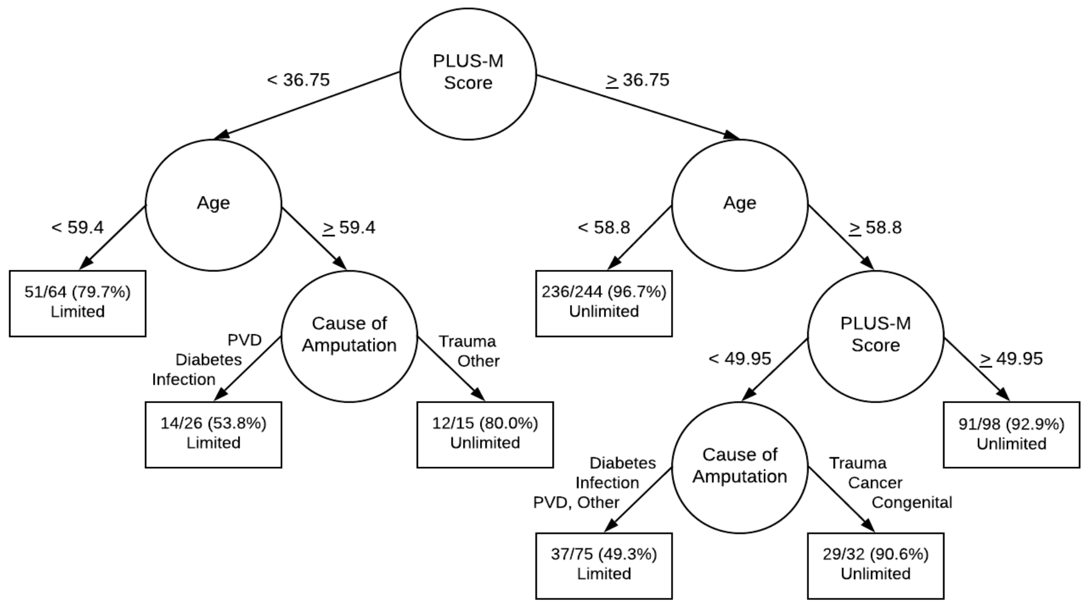
Figure 4: Minimum D "L1-R1" Model



Viewed theoretically, moderate accuracy obtained using Prosthetic Limb Users Survey of Mobility T-Score (PLUS-M™) score to discriminate L *vs*. U samples stands alone among the variables studied: any other variable produces a model with greater D. Mediocre accuracy hints at scoring imprecision attributable to use of a single rather than multifactorial index, and of need to identify factors which affect ambulatory status that are not assessed on the PLUS-M.

2155-0182/10/$3.00

Figure 5: Maximum ESS “L3-R4” Model Discriminating Ambulatory Status



Viewed translationally, Table 2 shows the L3-R4 combination (Figure 5) has the greatest ESS (68.3), a relatively strong effect. Data which are available for this study[1] (Figure 1) can’t address stability of training *vs*. validity accuracy of models discussed herein. However, age cutpoints are defined to the first decimal and the PLUS-M cutpoints to the second: observations having values bordering these cutpoints may be misclassified in validity analysis.

Left- and right-hand sides of the L3-R4 model are highly parallel. Age enters both sides of the root node, and the cutpoint values vary by 1.02%. This configuration theoretically could be modeled linearly if it was known that a cutpoint of 36.75 points on PLUS-M score mediated the effect, and the cutpoint for age lies somewhere between 58.8 and 59.4 years.

Unique to the right-hand-side of the model, PLUS-M score next enters the model for older observations (at this point a linear model becomes unusable). The cutpoint here is notably higher (no normative data) than occurred for the root variable: 92.9% of patients with a score of at least 49.95 points were summarily classified as members of class U.

Both sides of the model end (the left on the right, the right on the left) using cause of amputation as the final attribute: points of agreement between left- and right-hand sides of the model included patients with diabetes and infection were classified as class L, and those with Trauma were classified as class U. The use of a scale category “other” (as used here) is criticized elsewhere[17] as being an imprecise measure which fosters paradoxical confounding.

It is noteworthy that both the Limited (K1, K2) and Unlimited (K3, K4) class categories are agglomerations of an ordered K-index.[1] Agglomerating categories can induce Simpson’s paradox whereby analysis findings are obscured, exaggerated, or reversed.[18-20] *Au contraire*, the ODA paradigm neither requires or recommends agglomerating class variables or attributes: the use of ordered and multicategorical class variables and attributes is straightforward.[21-23]

Finally, although it is an inherent and immitigable problem for paradigms involving linear models, it is nevertheless also noteworthy that the model in Figure 1 was developed for a sample that necessarily excluded 71.7% of cases identified in the initial data extraction from the

2155-0182/10/$3.00

analysis because they were missing data on *any* of the predictor variables used in the study. In contrast, in the ODA paradigm a case is only excluded from an analysis if the case is missing data on a variable being used in an analysis.

References

[1]Wurdeman SR, Stevens PM, Campbell JH (2019). Mobility analysis of AmpuTees (MAAT 4): Classification tree analysis for probability of lower limb prosthesis user functional potential. *Disability and Rehabilitation: Assistive Technology*, 11 Feb 2019. https://doi.org/10.1080/17483107.2018.1555190

[2]The original study employed 20% of the total sample as a "training sample" to construct the CART model, and used the remaining 80% as a "hold-out sample" to validate the model.[1] The present article demonstrates how to optimize the predictive accuracy of the model derived in training analysis—which requires an illustration of the model. Validating the model requires the hold-out sample, which is not provided in the original article.

[3]https://www.mathworks.com

[4]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[4]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC, APA Books.

[5]Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo Estimation of two-category UniODA two-tailed *p*. *Optimal Data Analysis*, *1*, 43-45.

[6]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[7]Bryant FB (2010). The Loyola experience (1993-2009): Optimal Data Analysis in the Department of Psychology. *Optimal Data Analysis*, *1*, 4-9.

[8]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

[9]Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*, *66*, 660-672.

[10]Yarnold PR (2018). Comparing exact discrete 95% CIs for model *vs*. chance ESS to evaluate statistical significance. *Optimal Data Analysis*, *7*, 82-84.

[11]Yarnold PR (2019). The structure of *perfect* optimal models with a two-category class variable and four or fewer endpoints. *Optimal Data Analysis*, *8*, 21-25.

[12]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *5*, 171-174.

[13]Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis*, *1*, 23-29.

[14]Yarnold PR (2016). Pruning CTA models to maximize PAC. *Optimal Data Analysis*, *5*, 58-61.

[15]Yarnold PR (2019). Maximizing classification accuracy of CART® recursive partitioning tree models using optimal pruning. *Optimal Data Analysis*, *8*, 26-29.

[16]Yarnold PR (2019). Maximizing the accuracy of a CART tree model predicting missing data. *Optimal Data Analysis*, *8*, 33-37.

2155-0182/10/$3.00

[17]Yarnold PR (2018). Minimize usage of binary measurement scales in rigorous classical research. *Optimal Data Analysis*, *7*, 3-9

[18]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, *56*, 430-442.

[19]Bryant FB, Siegel EKB (2010). Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of *Melendez v. Illinois Bell*. *Optimal Data Analysis*, *1*, 176-198.

[20]Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, *5*, 41-52.

[21]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, *2*, 177-190.

[22]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5*, 65-73.

[23]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, *3*, 19.

Author Notes

No conflict of interest was reported.