Optimal Data Analysis
Vol. 8 (January 23, 2019), 33-37

2155-0182/10/$3.00

Maximizing the Accuracy of a CART Tree Model Predicting Missing Data

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

A classification tree pruning methodology that maximizes effect strength for sensitivity is demonstrated for a model developed using classification and regression analysis to identify factors which predict missing data.

Imagine a classification tree model initiated by a root variable with two emanating branches that partition the sample into left- and right-hand sides.[1] Identifying the tree model with greatest effect strength for sensitivity[2] (ESS) requires identifying all sub-branches for both emanating branches. Consider a left-hand branch with three nodes: A (root), B (middle attribute), and C (end of branch). This branch has two nested sub-branches: one with nodes A and B (C collapsed into B), and one with only node A (C and B collapsed into A). The left branch with three nodes (A, B, C) is called L3; the trimmed left branch with two nodes (A, C collapsed into B) is L2; and the trimmed left branch with one node (C and B collapsed into A) is L1. Imagine also that the model has a right-hand branch with nodes A (sides share the root attribute) and D (end of branch): the right branch with two attributes (A, D) is R2, and the trimmed right branch with one attribute (D collapsed into A) is R1. Optimal pruning requires obtaining a confusion table, in which rows are actual class category and columns are predicted class category, for each unique combination of the left and right (sub)branches. In this example six unique combinations are L1-R1, L2-R1, L3-R1, L1-R2, L2-R2 and L3-R2. The optimized CART model is the combination of (sub)branches with the corresponding confusion table returning the highest value for ESS.[3,4]

Maximizing ESS of a Classification Tree Model, Predicting Presence *vs.* Absence of Missing Data, Obtained by CART Analysis

A study investigated the use of classification and regression tree (CART) analysis to identify structure underlying missing data in an occupational health data set consisting of questionnaire responses, medical tests, and findings of environmental monitoring.[5] Figure 1 is the resulting ("original") CART model: data types are medical (Type 1), follow-up medical (Type 2), and hygiene/environmental exposure (Types 3-6). Three data collection sites are indicated using dummy codes (1-3). As seen, the root node of the CART model has two emanating branches: the right-hand side has two additional nodes, and the left-hand side has one additional node.

Schematic illustrations of L1 (root node) and L2 (root and second node), as well as their corresponding confusion tables, are respectively presented in Figures 2A and 2B. Similarly, Figures 3A-3C present schematic illustrations of R1-R3 and their corresponding confusion tables.

2155-0182/10/$3.00

Figure 1: Original CART Model

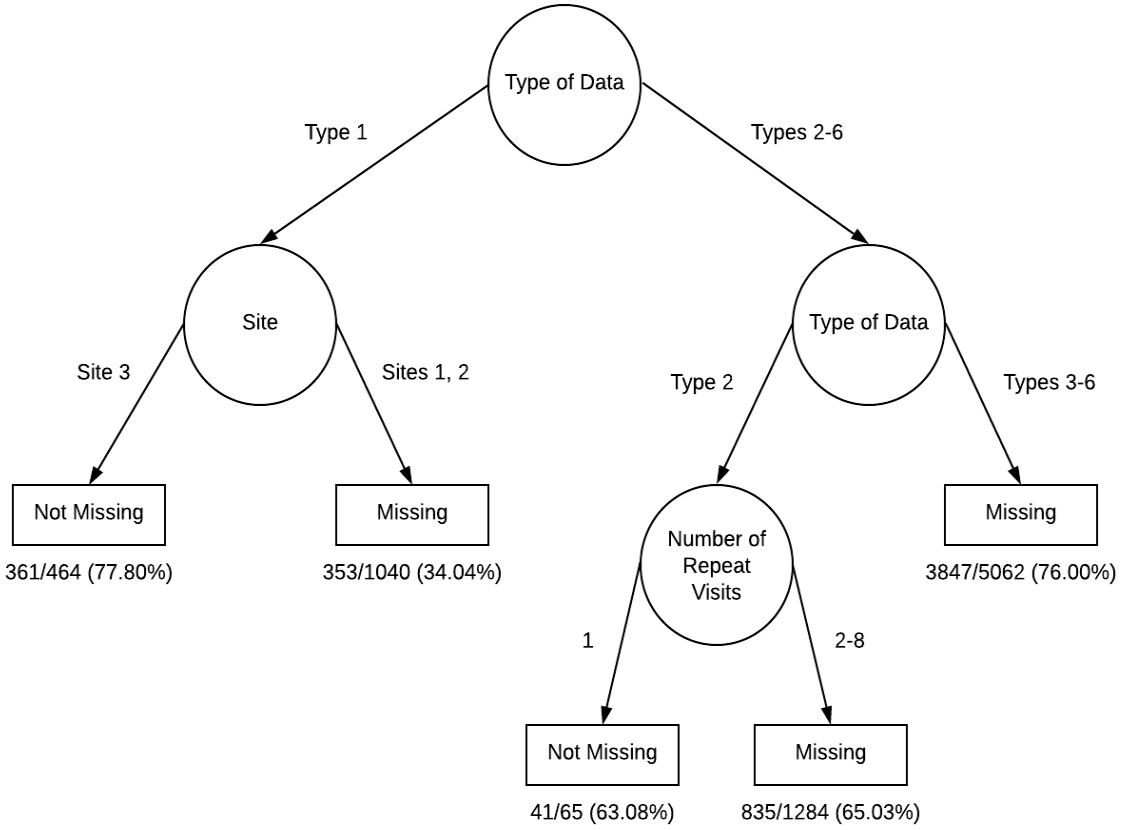
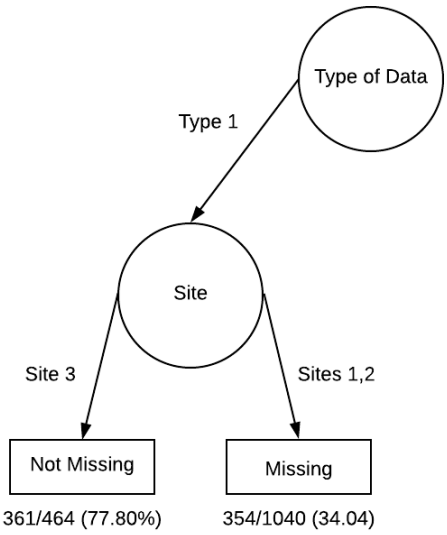


Figure 2A: L1 Sub-Branch and Confusion Table

	Predicted	
Actual	Not Missing	Missing
Not Missing	1047	0
Missing	457	0

Figure 2B: L2 Sub-Branch and Confusion Table



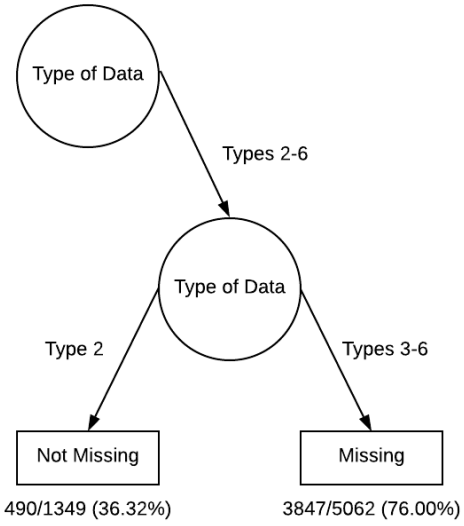
2155-0182/10/$3.00

	Predicted	
Actual	Not Missing	Missing
Not Missing	361	686
Missing	103	354

Figure 3A: R1 Sub-Branch and Confusion Table

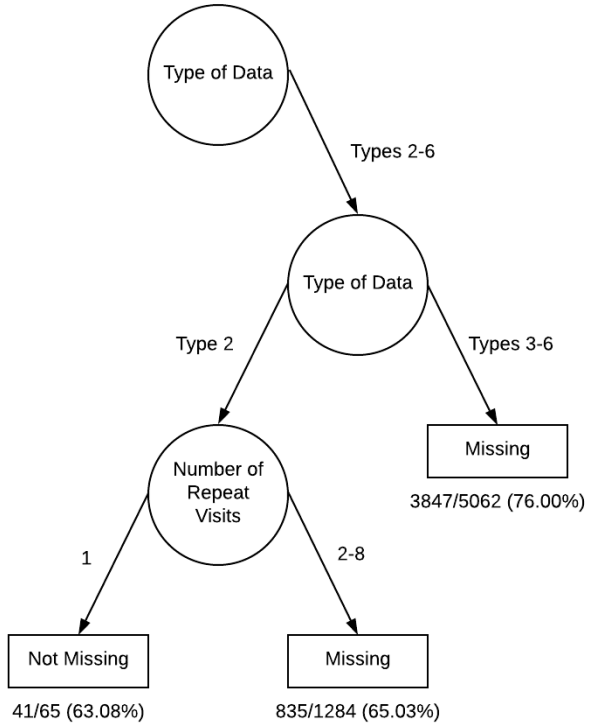
	Predicted	
Actual	Not Missing	Missing
Not Missing	0	1705
Missing	0	4706

Figure 3B: L2 Sub-Branch and Confusion Table



	Predicted	
Actual	Not Missing	Missing
Not Missing	490	859
Missing	1215	3847

Figure 3C: L3 Sub-Branch and Confusion Table



	Predicted	
Actual	Not Missing	Missing
Not Missing	41	24
Missing	1664	4682

Table 1 gives integrated confusion tables for all six possible combinations of left (L1, L2) and right (R1-R3) sub-branches, and their ESS.

Table 1: Classification Results for Every Combination of Left (L1, L2) and Right (R1-R3) Sub-Branch

Model	Confusion Table	
L1-R1	Predicted	
Actual	Not Missing	Missing
Not Missing	1047	1705
Missing	457	4706

ESS=29.2

2155-0182/10/$3.00

L1-R2

Actual	Predicted: Not Missing	Predicted: Missing
Not Missing	1537	859
Missing	1672	3847

ESS=5.6

L1-R3

Actual	Predicted: Not Missing	Predicted: Missing
Not Missing	1088	24
Missing	2121	4682

ESS=66.7

L2-R1

Actual	Predicted: Not Missing	Predicted: Missing
Not Missing	361	2391
Missing	103	5060

ESS=11.1

L2-R2

Actual	Predicted: Not Missing	Predicted: Missing
Not Missing	851	1545
Missing	1318	4201

ESS=11.6

L2-R3

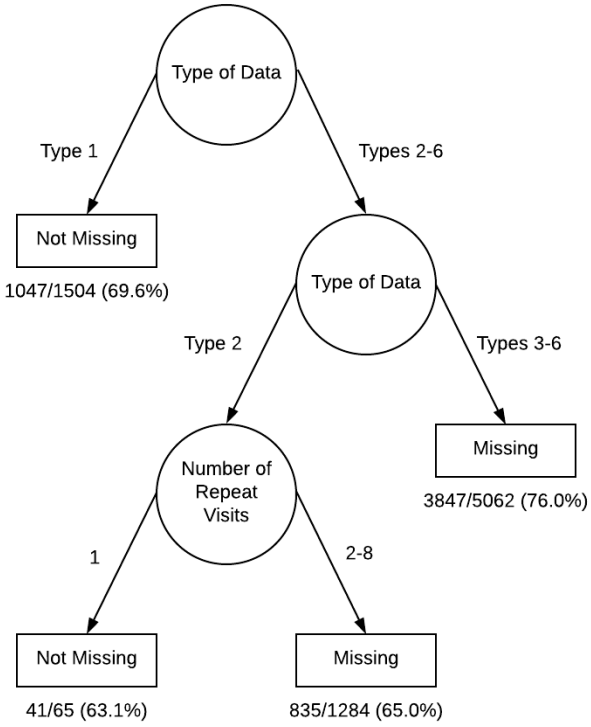
Actual	Predicted: Not Missing	Predicted: Missing
Not Missing	402	710
Missing	1767	5036

ESS=10.2

The combination L1-R3 (Figure 4) has the greatest mean sensitivity (83.33%), yielding optimized ESS=66.7— corresponding to a relatively strong effect.[2]

By definition the model which yields the maximum value of ESS performs best compared against chance and thus is the preferred model for translational application.[2] Highest ESS does not necessarily imply the model is best when it is considered from a theoretical perspective, which includes the criterion of parsimony—this is the purpose of the D (distance) statistic.[6,7]

Figure 4: Optimized CART Model



D norms ESS for number of attributes in the model, indicating the number of additional effects with equivalent mean ESS needed to attain an errorless model.[1,6] Presently, for L1-R1, D=4.85; for L1-R2, D=5.55; for L1-R3, D=2.0; for L2-R1, D=24.0; for L2-R2, D= 30.4; and for L2-R3, D=44.1. The L1-R3 model has *lowest* D among the six CART models therefore L1-R3 is closest to representing a theoretically optimal solution. Contrasted with legacy methods, optimal analyses can't be surpassed with respect to ESS (accuracy normed against chance) or to D (ESS normed for parsimony).[5] It is becoming increasingly clear that empirical findings obtained by legacy methods are rarely optimal.

References

[1]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.

2155-0182/10/$3.00

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

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

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

[5]Tierney NJ, Harden FA, Harden MJ, Mengersen KL (2014). Using decision trees to understand structure in missing data. *BMJ Open*. http://dx.doi.org/10.1136/bmjopen-2014-007450

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

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

Author Notes

No conflict of interest was reported.