Paper 110-27

Tree-Based Models: Identification of Influential factors under Condition of Instability

Pavel Brusilovskiy and Yilian Yuan
IMS HEALTH, Plymouth Meeting, PA

Abstract

The objective of the paper is to determine an overall importance of inputs in a series of runs of SAS® Enterprise Miner Tree node under condition of instability of tree-based models. We discuss simple and practical four step iterative approach of dealing with instability: derive a Pareto- optimal set of trees; test hypothesis for concordance of the selected inputs that corresponds to tree-based models in the Pareto set; if the hypothesis of concordance is accepted, form overall importance of the union of selected inputs, otherwise try to narrow Pareto set, or add new runs, and go to step one. The importance could be measured on nominal, ordinal or interval scales. Testing for concordance is performed by SAS macros MAGREE and INTRACC. The application of this approach is illustrated by pharmaceutical market research problem of the identification of influential factors for prescribing Viagra (308 inputs, 2000 observations). Skill level of Intended audience is IV.

Introduction

In pharmaceutical marketing research, we often want to identify the influential factors of prescribing certain prescription products, so that the marketing efforts can be targeted to those doctors who have potential to be high volume prescribers. Many factors have influences on doctors prescribing behavior: patient population, doctors characteristics such as specialty, years in practice, practice setting, geographic location, managed care influence, and inclination to prescribe certain similar drugs, etc. Therefore, there are a lot of factors to be considered in this type of analysis.

When the number of input variables is large enough, it is common that the tree-based models are unstable, due to the nature of tree-based algorithms, high redundancy of the input space, complexity of the relationship among target and inputs, noise in the data, etc. This makes it very difficult to identify the influential factors of doctors prescribing behavior and rank /rate them. Since a tree-based model is unstable, each run will produce a different list of influential factors with different ranking/rating. The tree instabilit y is not an issue in the predictive modeling when the objective is to predict who is a potential high volume prescriber. On the contrary, it refines the model – one can use different boosting/arcing/bagging algorithms to improve accuracy of the prediction (see, for example, Bauer and Kohavi, 1999). However, these algorithms do not help if the goal of the decision tree analysis is to identify drivers of prescribing behavior. In this paper, we suggest an approach to aggregate the results from multiple runs when a tree-based model is unstable.

In running each decision tree, we assume that the usage of a new random seed (in the random sampling procedure to form training and validation data sets) results to a new tree with a high probability. We suggest a simple, yet practical, approach of dealing with instability of the decision tree analysis that allows us to identify rank and/or rate of influential factors according to their overall importance in a series of runs of corresponding software, for example, Tree node of SAS Enterprise Miner. We use Viagra prescribing behavior study as an example to illustrate our approach. Viagra data includes 308 inputs and 2000 observations.

Methodology

The objective of this approach is to determine an aggregated ranking and/or rating of inputs from multiple runs of tree-based models. Here, the term 'tree-based model' refers to any model that can be developed within Tree node of SAS Enterprise Miner.

The approach deals with three different types of data, including the original data set and three auxiliary data sets. The auxiliary data sets for a regression tree are formed from the information that is: (i) displayed in Summary tab of Tree node (criteria data set), (ii) displayed in Score tab – Variable Selection button (inputs importance data set), (iii) generated by subject matter expert(s) in order to reflect different values of different tree-based models for a decision maker (judgments data set). Different tree-based models may have different value for a decision maker. One may compare tree based models in terms of their explanatory capability, complexity, reasonableness, etc. The approach could be treated as a three- step process in the case of equally weighted tree-based models, or as a four-step process, if you are planning to use expert judgment to determine weights of the tree-based models in the driver identification procedure:

-derive a Pareto-optimal set of tree-based models, based on the criteria data set
-test the hypothesis for concordance of the inputs, corresponding to the tree-based models in the Pareto set, using the inputs importance data. If the hypothesis is accepted, then go to the next step. Otherwise try to narrow a Pareto-optimal set, or generate additional trees, and go to the step one again.
-compute weights for each tree-based model from the Pareto set, utilizing the judgments data set (optional step). Otherwise all tree-based models are treated equally important.
-form final aggregated ranking/rating of inputs employing inputs importance data set and tree-based model weights. Inputs with positive aggregated importance (rating) and/or aggregated ranks are regarded as identified influential factors (drivers) of a phenomenon under consideration.

A Pareto-optimal set is a set of equally good solutions in which each solution has at least one objective that is better than any other solutions in the set (Steuer, 1986). In other words, a Pareto-optimal set is a set of non-dominated, or non-inferior alternatives. A Pareto-optimal set of tree-based models is created on several criteria calculated on the validation data set. For example, for a regression tree, SAS Enterprise Miner Tree node calculates two criteria: average squared error and R-squared. For classification tree with a binary target, it makes sense to use sensitivity, specificity, accuracy, etc.

The importance measurements of the input variables generated by the tree-based models are stored in a data set with n +1 columns, where n is the number of tree-based models in a
To use original interval-scaled values of the importance, produced by Tree node for each tree-based model, it is necessary to employ SAS Macro INTRACC to calculate intraclass correlation coefficients to test rater agreement (or reliability) on continuous responses. This macro produces six intraclass correlations discussed in Shrout and Fleiss (1979) and in McGraw and Wong (1996). Along with this, it calculates the reliability of the ratings mean, using the Spearmen-Brown prophecy formula to examine the effect obtaining more raters would have on the reliability of the mean. If the hypothesis of concordance is rejected, modify the Pareto set (for example, try to narrow the Pareto set by using expert knowledge or by computing Kemeny median (Kemeny, 1959; Kemeny and Snell, 1972) that is a subset of Pareto set, or try to extend the Pareto set, adding several new trees, and then go to step one. If the modification of the Pareto set did not lead to acceptance of the hypothesis of concordance, then the results of the decision tree analysis are inconsistent, and probably should not be used at all.

Tree-based models in the Pareto set may have different values for a decision maker. The simplest way to take this fact into account is to make use of models' ranking, generated by subject matter expert(s) or decision maker(s). The ranking can be converted to weights according to one of the rules, discussed in Barrette and Baron (1996). In particular, the rank-sum rule is the most popular one (Brusilovskiy and Hernandez, 1997). More sophisticated ideas of incorporating expert judgment is considered by Brusilovskiy and Tilman (1996). Weights are used to calculate the final aggregated importance of inputs and identify influential factors - inputs with positive aggregated importance.

Pareto-optimal set. The first column is the variable names of the inputs from the original data set to be analyzed. Each tree-based model from the Pareto-optimal set is represented in this data set by one column that reflects the importance measurements of the corresponding input variables. The concept of partial lists (Dwork et al., 2000) provide a practical way to handle inputs selected by tree-based algorithm in a series of runs.

The importance can be measured on a nominal scale, for example, using two classes with the labels INPUT or REJECTED, defined by the column 'Role' in Score tab – Variable Selection button of Tree node. The importance can be measured on an interval scale, for example, see the values of the column 'Importance' in Score tab – Variable Selection button of Tree node. Finally, the input importance can be measured on an ordinal scale, for example, by translating the values of the importance, produced by Tree node, into ranks. Examining the concordance of the inputs selected by tree-based models, a special interpretation is useful: each tree-based model from the Pareto set plays the role of a rater, each input (observation in the inputs importance data) plays the role of a subject, and the value that a rater assigns to a subject (column in the inputs importance data) is the input importance, considered in nominal, ordinal or interval scales. If rater responses are treated as nominal or ordinal, then the agreement among multiple raters could be tested by the SAS Macros MAGREE. In particular, SAS Macro calculates the kappa statistic (for measurements on nominal or ordinal scales) and Kendall's coefficient of concordance (for measurements on an ordinal scale).

Other approaches to order candidate models in the Pareto set is discussed by I. Das (1999).

The construction of an optimal aggregation of input variables' importance depends upon several factors, and first of all, upon a measurement scale of input importance, and upon a distance measure between rankings/ratings. A good discussion of new ideas within axiomatic approach could be found in Sadovsky (2001), and within heuristic approach - in Dwork et al. (2000).

Application

The application of this approach is illustrated by the analysis to identify influential factors for prescribing Viagra, using doctor's demographic and prescription data, such as physician age, specialty, years in practice, practice setting, geographic location, and inclination to prescribe certain other drugs.
The objective was to identify the influential factors that are associated with higher level of Viagra prescribing behavior. The input data set has 308 input variables, and among them there are 10 variables that reflected the doctor's demographics (7 of them were categorical variables), and 298 interval scaled variables that described doctor's prescription activity. The target variable was number of Viagra prescriptions. In the example we used 10 different random seeds for simple random sampling of the data with fixed percentages: 65% for train, and 35% for validation, and F-test as a splitting criterion, which generated 10 different regression trees.

Table 1. Viagra Prescription Study: Criteria Data and Pareto Set Determination

Tree ID	**Random Seed**	**Average Squared Error**	**R-squared**	**Dominated Trees**	**Element of Pareto Set**
1	**12345**	1013.6993	0.6332	3	x
2	**7760**	952.9744	0.4354		
3	**9556**	1030.2332	0.5817		
4	**3922**	852.0042	0.4986	2, 5	
5	**5677**	906.4417	0.4887	2, 7	
6	**5297**	825.2153	0.5190	2	
7	**1161**	1054.6481	0.4308		
8	**1496**	823.3705	0.5547	2, 4, 6	x
9	**558**	850.5508	0.5567	5, 7	x
10	**7230**	786.0792	0.4977	2, 5, 7	x

Table 2. Viagra Prescription Study: Input Importance Data Set

		Random Seed							
		12345		**1496**		**558**		**7230**	
	Input	**Inputs' Importance**	**Rank**	**Inputs' Importance**	**Rank**	**Inputs' Importance**	**Rank**	**Inputs' Importance**	**Rank**
1	_31440	1.000	1	1.000	1	1.000	1	1.000	1
2	_31410	.6167	2	.0000	8	.0000	7	.0000	14
3	_99002	.4394	3	.0000	8	.0000	7	.0910	11
4	_39100	.3906	4	.0000	8	.0000	7	.0000	14
5	_31141	.2750	5	.0000	8	.3123	4	.2655	5
6	YRSEXP	.2669	6	.0000	8	.0000	7	.0000	14
7	_09110	.1881	7	.0000	8	.2078	5	.0000	14
8	_61660	.1691	8	.0000	8	.0000	7	.0000	14
9	_64610	.1599	9	.0000	8	.0000	7	.0000	14
10	_99018	.1521	10	.0000	8	.0000	7	.0000	14
11	_15180	.1336	11	0.3721	3	.3321	3	.3837	4
12	_52250	.0740	12	.0000	8	.0000	7	.0000	14
13	_39210	.0000	13	.0986	6	.0000	7	.0576	13
14	_61690	.0000	13	.0000	8	.0000	7	.0000	14
15	_85000	.0000	13	.7876	2	0.1504	6	.1128	10
16	_99025	.0000	13	.0000	8	.0000	7	.7609	2
17	_31110	.0000	13	.3035	4	.0000	7	.0000	14
18	CEN_MSA	.0000	13	.1980	5	.0000	7	.1250	9
19	_76140	.0000	13	.0583	7	.0000	7	.0000	14
20	_24330	.0000	13	.0000	8	.7174	2	.0000	14
21	_01200	.0000	13	.0000	8	.0000	7	.3930	3
22	_02120	.0000	13	.0000	8	.0000	7	.2419	6
23	_02110	.0000	13	.0000	8	.0000	7	.1562	7
24	_99015	.0000	13	.0000	8	.0000	7	.1477	8
25	_99032	.0000	13	.0000	8	.0000	7	.0733	12

The criteria data set is described in Table 1. We can see that the tree with ID=10 dominates trees with ID number two, five, and seven, etc. Only four trees, marked by 'x', are included in the Pareto set.

The two simplest trees from the Pareto set are represented in Graph1 and Graph2. A reader can view the structure of two others in Graph 3. We can declare that trees in the Pareto set are the most dissimilar and simultaneously possess the best quality in terms of two criteria at hand.

Table 3. Viagra Prescription Study: Test for Concordance of Inputs, Selected by Trees from the Pareto Set, Using Ordinal Scale for Importance

MAGREE macro
Kendall's Coefficient of Concordance for ordinal response

Coeff of Concordance	F	Num DF	Denom DF	Prob>F
0.45151	1.64639	23.3333	46.6667	0.0732

The importance of inputs, produced by trees from the Pareto set, is displayed in Table 2. The input YRSEXP is the number of years since obtaining MD, and the input CEN_MSA is population size of the MSA where a doctor practices. Other inputs starting with "_" are the prescription volume of other drug classes.

Table 4. Viagra Prescription Study: Influential Factors and their Aggregate Importance on Ordinal Scale

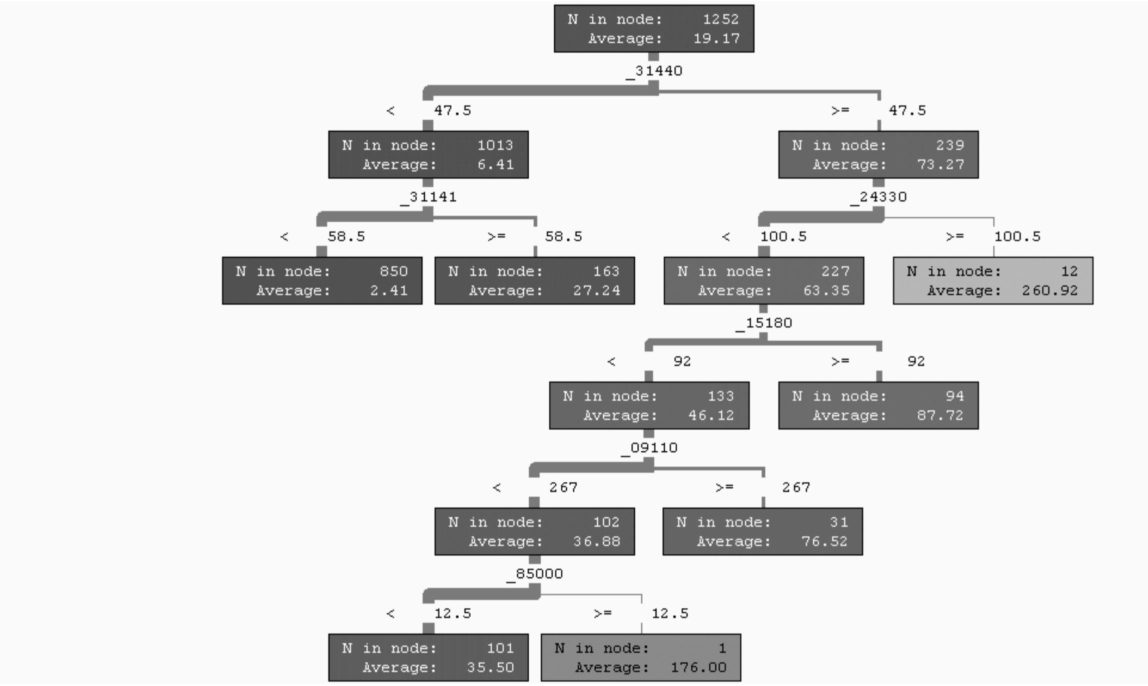
Obs	Influential factor	Aggregated importance
1	_31440	1
2	_15180	2
3	_31141	3
4	_99002	4
5	_99025	5
6	_31410	6
7	_85000	6
8	_01200	6
9	_39100	9
10	_09110	10
11	CEN_MSA	10
12	_02120	10
13	YRSEXP	13
14	_02110	13
15	_99015	15
16	_61660	16
17	_24330	16
18	_64610	18
19	_31110	18
20	_99018	20
21	_39210	20
22	_99032	22
23	_52250	23
24	_76140	23
25	_61690	25

The hypothesis of the concordance of inputs, selected by different trees in terms of the importance, was accepted. In other words, we can say that there were concordance relationships among different tree-based models from the Pareto set (see Table 3). The simplest way to construct aggregated importance for the inputs is to obtain the average rank over the ranks resulted from tree-based models for each input. This approach is used to generate results in Table 4.

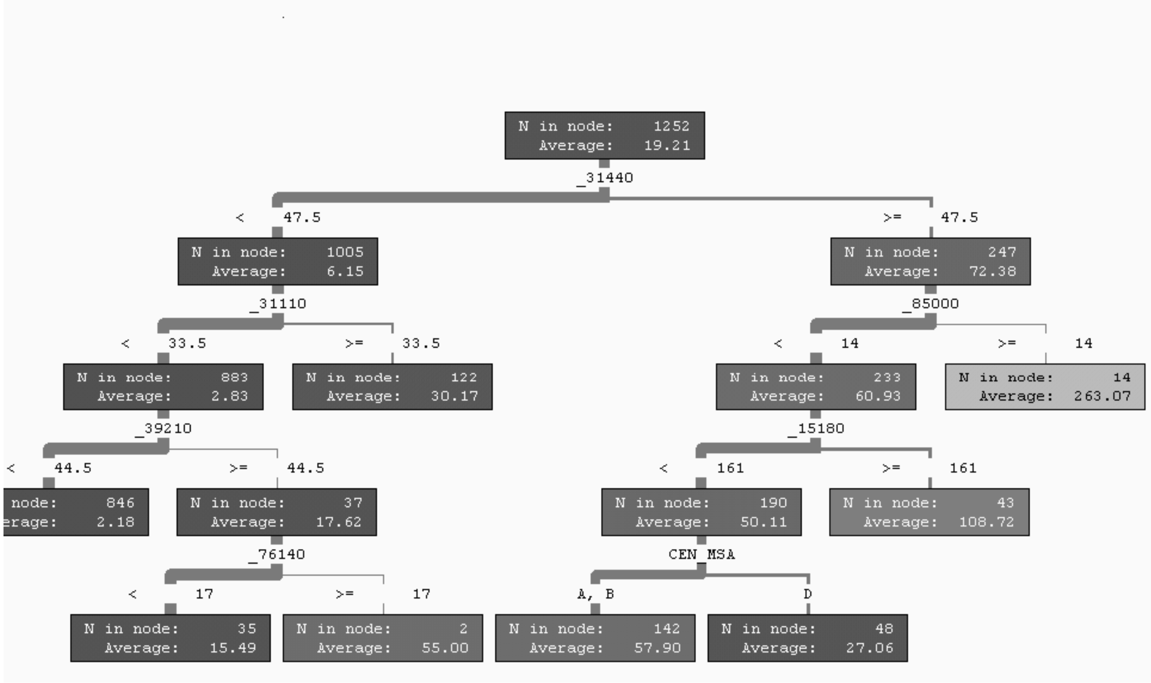
Discussion

In this example, the tree-based models were not stable. Each run resulted in different measurement of the input importance (different list of selected predictors by decision tree algorithm) that makes it difficult to pin point which variables are influential.

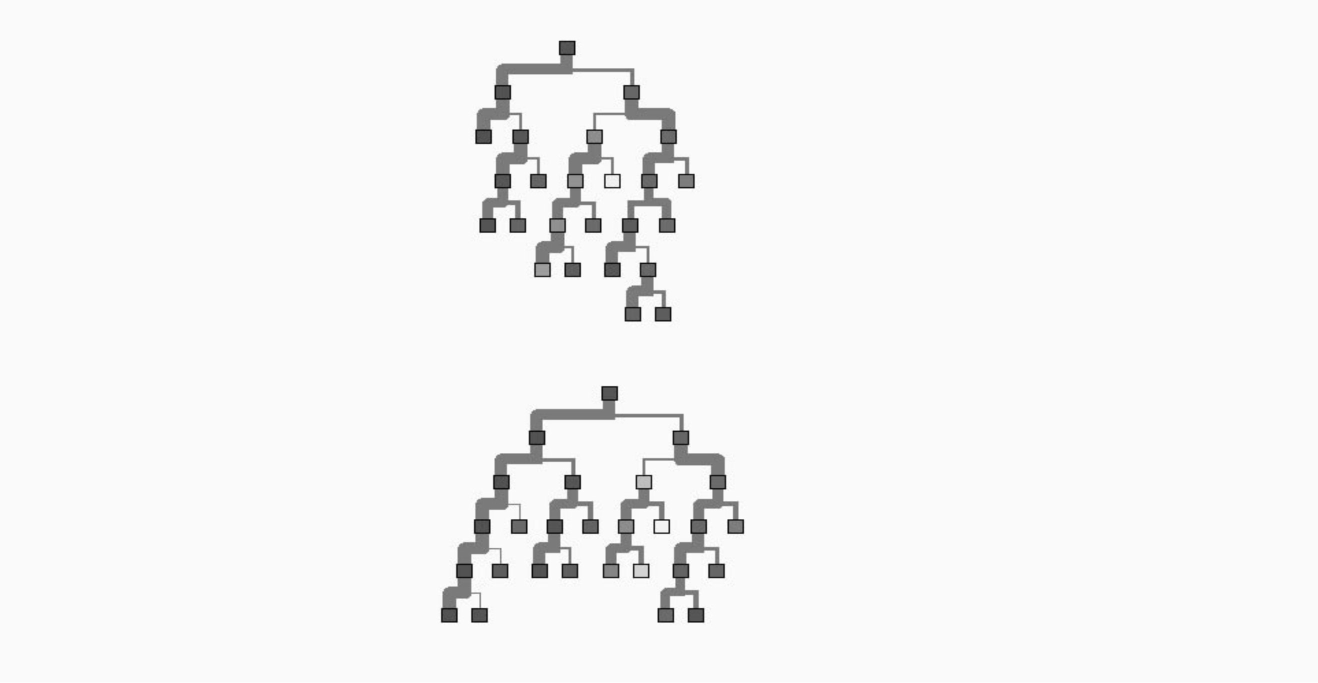
Graph 1. Tree with Random Seed 558 (statistics are displayed only for training data set)



Graph 2. Tree with Random Seed 1456 (statistics are displayed only for training data set)



Graph 3. Structure of Trees with Random Seeds 12345 and 7230



Using the approach suggested in this paper, we identified the top 25 influential factors among 308 inputs that associated with prescribing Viagra and ranked them according to their aggregated importance. For example, inputs _99002 and _99025 representing the prescription volumes of Allegra and Prilosec, were identified as good predictors of Viagra. These two products were highly promoted to consumers through Direct-to-Consumer (DTC) advertisements. This result may indicate that the doctors write highly DTC supported products are more responsive to patient's request and hence are likely to write more prescriptions for Viagra

References

Cynthia Dwork, Ravi Kumar, Moni Naor and D. Sivakumar (2000), Rank Aggregation Methods for the Web , http://www.wisdom.weizmann.ac.il/~naor/PAPERS/rank_www10.html

Kemeny, J. (1959), Mathematics without Numbers, *Daedalus* 88, 571-591

Kemeny, J., and Snell, J. (1972), Mathematical Models in Social Science, The MIT Press

SAS MAGREE Macro, http://ftp.sas.com/techsup/download/stat/magree.html

SAS INTRACC Macro, http://ftp.sas.com/techsup/download/stat/sas

Eric Bauer and Ron Kohavi (1999), An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants, *Machine Learning*, Vol.36, Nos. ½, July/August 1999, 105-139

Shrout, P.E. and Fleiss, J.L. (1979), Intraclass Correlations: Uses in Assessing Rater Reliability, *Psychological Bulletin*, Vol. 86, 2, 420-428

Pavel Brusilovskiy and Leo Tilman (1996), Incorporating Expert Judgment into Multivariate Polynomial Modeling, *Decision Support Systems*, Vol. 18, 199-214

Alexey L. Sadovsky (2001), Multi-Objective Optimization and Decisions Based on Ratings Methods of Preference Ranking, http://www.sci.tamucc.edu/~sadovsky/rankfinal.htm

Steuer, R.L. (1986), Multiple Criteria Optimization: Theory, Computation and Application, Wiley, New York

Indraneel Das (1999), A Preference Ordering Among Various Pareto Optimal Alternatives, *Structural Optimization*, vol. 18, no.1, pp 30-35

Pavel Brusilovskiy and Robert Hernandez (1997), Multi-Criteria Evaluation of Alternatives: Usage of Objective Data, Subjective Measurements and Expert Judgment, Proceedings of NESUG'97, Baltimore, 590-596

Barrette Bruce E. and F.H. Barron (1996), Decision Quality Using Ranked Attribute Weights, *Management Science*, November 1996, 1515-1523

McGraw, K.O. and Wong, S.P. (1996), forming Inferences about some Intraclass Correlation Coefficients, *Psychological Methods*,1 (1), 30 – 46

Contact Information

Pavel Brusilovskiy, (610) 834-4533,
pbrusilovskiy@us.imshealth.com
Yilian Yuan, (610) 834-5177, yyuan@us.imshealth.com
Fax: 610-834-5690
IMS Health, 660 West Germantown Pike, Plymouth Meeting, PA 19462