Poster – SESUG 2002

Using Decision Trees to Identify Medicare Part B Providers for Audit

Noel McKetty, First Coast Service Options, Jacksonville, FL
Donna Mohr, University of North Florida, Jacksonville, FL

ABSTRACT

Medicare Providers are selected for audit based on a wide variety of tools, including comparisons to their peers in terms of utilization for various procedures. We use decision trees to examine the efficacy of these peer-comparison statistics in predicting which Providers will be shown to have high overpayments – that is, to owe money back to the Trust Fund. We show that a new screening variable, based simply on dollars paid per patient, is a valuable additional tool for identifying providers for audit.

INTRODUCTION

The Centers for Medicare and Medicaid Services (CMS) charges the Program Integrity contractors with monitoring the validity of payments made to providers (physicians, labs, clinics, hospitals, etc.). As one part of their activities, contractors conduct audits of randomly selected claims from 'suspicious' providers. Given limited budgets, and given the expense of an audit both for the provider and the contractor, contractors need to be efficient in selecting providers for audit. Among the routine selection tools are a variety of reports, including comparisons of providers to peers within the same specialty with respect to frequency with which various procedures are billed. Do these measures do a good job of selecting providers for audit?

To answer this, we examined historical archives of results from medical reviews that took place 1998 through 2000. This provided us with an opportunity to use data mining techniques to explore the predictive power of these peer comparisons. Our work focuses on Part B providers, primarily physicians and labs.

BACKGROUND ON THE DATA

This study depends on two sources of information. The primary file is the record of 381 Part B reviews completed between 1998 and 2000. Our interest is in the proportion of paid amounts estimated to be overpayments. This variable – RATIO_PD – is the variable we want to predict. In this archive, the values of RATIO_PD ranged from near 0% to 100%, with a typical value near 100%. Also of interest is the Specialty of the performing provider (PERF_SPEC). In this archive, there were 32 different specialties represented, with an emphasis on 01-General Practice (27%), followed by 11 - Internal Medicine (8.4%), 69 – Clinical Laboratory (6%), and 08 – Family Practice (5.5%).

This archive does not represent a random sample of providers. It represents providers selected for audit for a variety of reasons: tips from employees, government directives, spike billing reports, patient complaints, and aberrant values on peer comparisons. This explains why the overpayment ratios are so high. This observation provides a kind of cautionary note about how well any results can be extended to the general population.

For each provider in the archive, we calculated several values meant to compare their usage to their peers. Peer data was based on the second source of information: all claims filed for Medicare Part B in Florida. Ratio I and II are comparison statistics dictated by the CMS, the government agency which administers Medicare.

Ratio I is a peer comparison statistic computed for each procedure code (X-ray, office visit, injection, etc.). It is the number of services of this code allowed for the provider divided by the total number of beneficiaries seen. To compare a provider's value to that for his peers, a z-score is computed via:

$$z_{ij} = z \quad for \quad provider \quad i, \quad procedure \quad j$$

$$= \frac{provider's \quad ratio \quad I - mean \quad for \quad peers}{standard \quad dev \quad for \quad peers}$$

This value is on a procedure-by-procedure code basis. To have a value that is easier to work with, we summarize these values as

$$TOTWGT1 = \sum_j all \quad z_{ij} \quad that \quad are \quad positive$$

And $MAXWGT1 = \max imum \quad of \quad z_{ij}$

MAXWGT1 is designed to catch a provider who has even one procedure code that is seriously aberrant. TOTWGT1 is designed to catch a provider with several procedure codes that are somewhat high.

Ratio II is the number of services in a particular procedure code billed by the provider, divided by the number of unique beneficiaries receiving this particular service. Whereas Ratio I is geared towards reporting a high overall usage rate among patients as a whole, Ratio II focuses on detecting cases where those patients who get the service do so multiple times. To compare a provider to his peers, z-scores are computed just as for Ratio I. Since we also compute Ratio II on a procedure-by-procedure basis, summary statistics MAXWGT2 and TOTWGT2 are computed just as they are for Ratio I.

Dollars per patient is simply total dollars allowed by Medicare divided by number of unique beneficiaries seen in the time period. Overall dollars per patient within specialties were available, but not dollars per patient for each provider in the peer group. Hence, we could not calculate standard deviations within the peer group. To stabilize variances for the provider to peer comparison, we assume standard deviations are proportional to the square root of mean dollars per patient. An approximate z-score would then be:

$$\text{ZDOL} = \frac{provider \quad \$ \quad per \quad patient - peer \quad \$ \quad per \quad patient}{\sqrt{peer \quad \$ \quad per \quad patient}}$$

STATISTICAL QUESTION –

Given input variables ZDOL, TOTWGT1, MAXWGT1, TOTWGT2, MAXWGT2, can we identify providers who are likely to have high RATIO_PD?

This sounds like a regression problem, but the relationship between these input variables and the target variable, RATIO_PD, is likely to be nonlinear and to contain a variety of interactions which one cannot specify in advance. Moreover, even though we adjusted the variables for specialty, it is possible that the nature of their relationship to RATIO_PD might vary by specialty. Hence, we decided to use REGRESSION TREES!

STATISTICAL RESULTS

We used **SAS Enterprise Miner™** to fit a variety of regression tree models. Because the data set is relatively small, by data mining standards, we partitioned the data set to allow 60% of the observations for training and 40% for validation. We skipped the test step. RATIO_PD was, of course, the target variable, and the input variables were SPECIALTY and the peer comparison statistics discussed above.

In Model 1, the training data was selected as a simple random sample using the default random number seed of ‘12345’. The default settings were used in the decision tree control settings. The assessment of the training and validation data sets, using average square error as the assessment measure, is shown in Figure 1. It would indicate that 5 leaves would be a reasonable complexity for this tree. The best decision tree with 5 leaves is shown in Figure 2.

Figure 1. Assessment Plot for Model 1, showing 5 leaves as optimal.

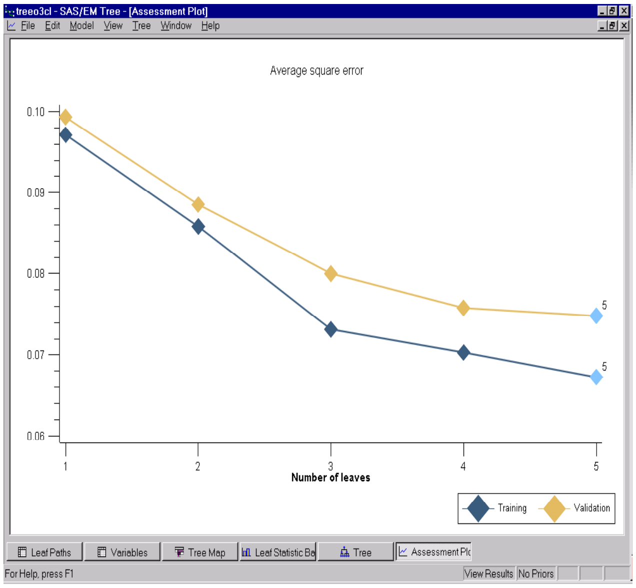
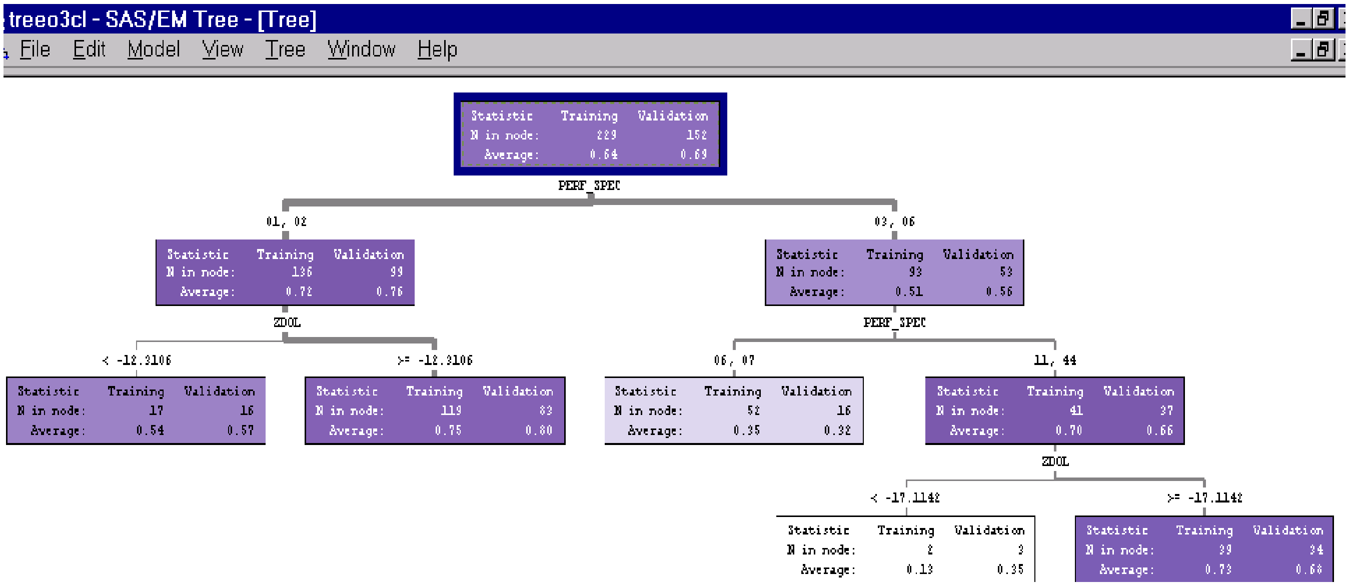


Figure 2. Decision tree for Model 1, using 5 leaves.


treeo3cl - SAS/EM Tree - [Tree]
File Edit Model View Tree Window Help

Statistic	Training	Validation
N in node:	229	152
Average:	0.64	0.69

PERF_SPEC

01, 02

Statistic	Training	Validation
N in node:	136	99
Average:	0.72	0.76

03, 06

Statistic	Training	Validation
N in node:	93	53
Average:	0.51	0.56

ZDOL

< -12.3106

Statistic	Training	Validation
N in node:	17	16
Average:	0.54	0.57

>= -12.3106

Statistic	Training	Validation
N in node:	119	83
Average:	0.75	0.80

PERF_SPEC

06, 07

Statistic	Training	Validation
N in node:	52	16
Average:	0.35	0.32

11, 44

Statistic	Training	Validation
N in node:	41	37
Average:	0.70	0.66

ZDOL

< -17.1142

Statistic	Training	Validation
N in node:	2	3
Average:	0.13	0.35

>= -17.1142

Statistic	Training	Validation
N in node:	39	34
Average:	0.73	0.68

Note that for categorical lists, like Specialty, only a partial list of values is printed in the Decision Tree labeling. That is, the labeling for a split will usually contain just the first one or two values in a branch.

This tree begins by splitting providers according to specialty. The left branch contains specialty ‘01’ General Practitioners and a variety of other providers, while specialty ‘11’ Internal Medicine fell in the right branch having a generally lower value of RATIO_PD. Within the first group, dominated by General Practitioners, the next decision is to split on ZDOL. Those having lower ZDOL tend to have lower RATIO_PD. Within the right group, the next decision is to further subdivide the providers by specialty. Within the sub-branch dominated by Internal Medicine (Perf_Spec=11), a good idea is to split on ZDOL, with those with low ZDOL again tending to have low RATIO_PD.

Hence, this first review of a decision tree shows that within many specialties, information based on dollars per patient is helpful in identifying providers who should be audited. Information based on Ratio I and Ratio II did not show as helpful, unless one explores further into the less stable parts of the tree.

Distressingly, however, these results are not stable. We ran Model 2 with the same parameter settings as Model 1, and on a data partition of similar structure, but generated with a different random seed. An assessment plot for Model 2 (Figure 3), would lead to a clear recommendation of no more than 2 leaves rather than 5. Figure 4 shows the decision tree for Model 2. Note that the single split even differs from Model 1 in the choice of specialties within the two nodes, though at least we find that General Practitioners and Internal Medicine still fall in separate groups.

Figure 3. Assessment Plot for Model 2, showing 2 leaves as recommended.

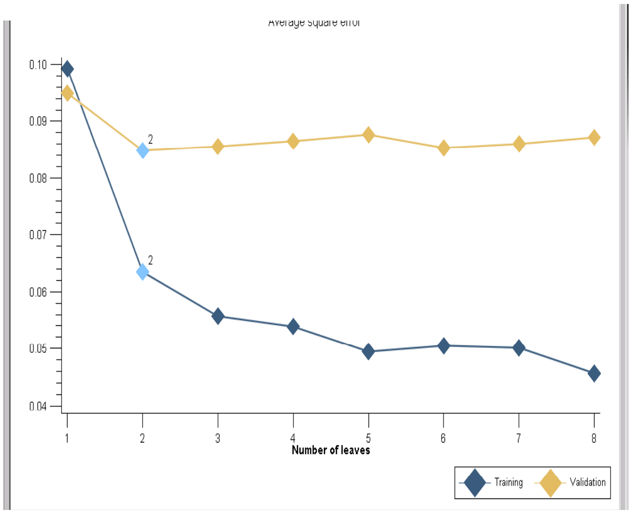
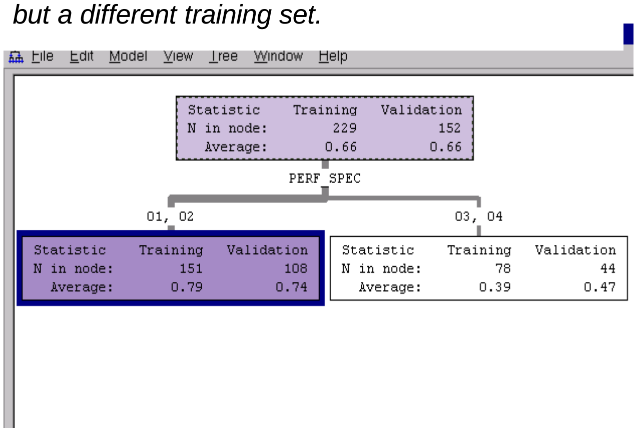


Figure 4. Decision Tree for Model 2, which uses the same model parameter settings as Model 1, but a different training set.



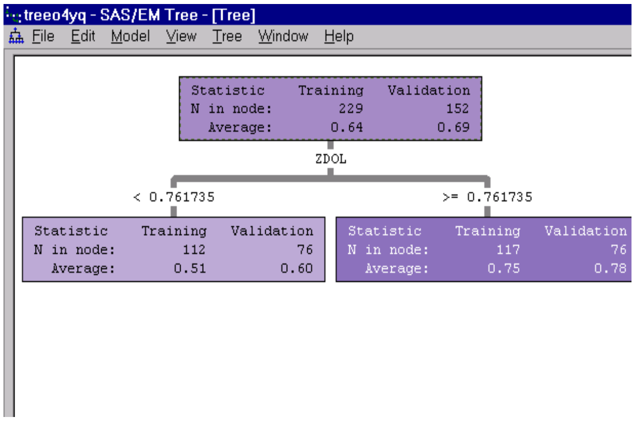
```
Statistic      Training   Validation
N in node:          229          152
  Average:         0.66         0.66
```

```
Statistic      Training   Validation
N in node:          151          108
  Average:         0.79         0.74
```

```
Statistic      Training   Validation
N in node:           78           44
  Average:         0.39         0.47
```

Finally, Figure 5 presents still another decision tree, based on the same training set as Model 1, but allowing for the possibility of three-way splits in the data. Though the data miner tool allowed for three-way splits, none were actually chosen for the tree. Nevertheless, the tree differs significantly from Model 1, which used the same training data set. It splits not on Specialty but ZDOL. Reassuringly, providers with high ZDOL tend to have high overpayments.

Figure 5. Decision Tree for Model 3, which uses the same training set as Model 1 but allows 3-way splits.


treeo4yq - SAS/EM Tree - [Tree]
File Edit Model View Tree Window Help
ZDOL
< 0.761735
>= 0.761735

Statistic	Training	Validation
N in node:	229	152
Average:	0.64	0.69

< 0.761735

Statistic	Training	Validation
N in node:	112	76
Average:	0.51	0.60

>= 0.761735

Statistic	Training	Validation
N in node:	117	76
Average:	0.75	0.78

WHY DO THE RESULTS DIFFER?

The small sample sizes within some of the specialties contributed, in large part, to the apparent instability of the tree. Unless all of the RATIO_PDs within a small specialty are homogeneous, the chance draw of which providers go in to the training data set will have a big impact on the average RATIO_PD in that specialty. Hence, the way the specialty sorts in a split between ‘good’ and ‘bad’ specialties may vary wildly between each selection of a training data set.

To understand the change when 3-way splits are allowed, you need to understand something about the stopping rules in trees. When the input is nominal (as it is for specialty), there are $k^L - 1$ possible splits, where k is the number of different splits allowed from a single node. In our case, L is 32, the number of different specialties in our data set. When k is 3, the number of possible different splits on specialty is much larger than it is when k is 2. The stopping rules employ a type of multiple-comparison adjustment called a Kass-adjustment which adjusts the required p-value that a split’s reduction in squared error must achieve in order for the rule to be a viable candidate. This adjustment reduces the p-value as the number of possible splits increases, essentially raising the bar that an input must meet to be selected as the number of possible rules that could be made from that input increase. Kass-adjustments work more against nominal inputs than they do against ordinal or interval inputs, where the number of possible rules increases more slowly as a function of k. In short, by making 3-way splits possible, we actually made it harder for Specialty to be selected as an input.

To put this variety of results in perspective, you should realize that classical stepwise regression techniques have the same difficulty. If you create multiple training data sets out of this master set, and allow the same inputs into stepwise regression, you see the same kind of changes from one training set to the next. The point is that before the advent of easy data re-sampling techniques, we rarely carried out such comparisons. We just ran the regression once on the full data set and accepted the results. This gave us a feeling of confidence in its results that was probably unwarranted.

CONCLUSION

The overall impression gained from these analyses is that dollars per patient, when adjusted by specialty in some way, can give valuable insight in identifying providers for audit. While this variable has not been a part of the standard screening tools, it will now receive more attention. Interestingly, the traditional peer-comparison measures based on Ratio I and Ratio II show little predictive value within the archive.

We are not saying that Ratio I and Ratio II are not valuable items to examine. Once a provider is selected for audit, the procedure code specific z-scores for these variables can be very helpful in focusing the audit on particular types of services. Furthermore, we caution again that this is not a random sample of providers. Many of these providers had high values of Ratio I and/or II for at least some procedure codes. Indeed, that was why many of them were selected. It is quite possible that for general screening of providers, this information could still have excellent predictive power.

The second conclusion we reached is data mining, like coal mining, is hard and dirty work. Construction of the database was time-consuming because we had to go back to construct the peer comparison benchmarks. Running SAS Enterprise Miner™ is relatively simple, but making sense of results requires practice and the patience to compare many possible runs. Perhaps in much larger data sets, or with less collinearity in the input variables, the results would be more stable. However, in this particular application, we found that looking at multiple runs was a sobering experience.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact:

Noel McKetty
First Coast Service Options
532 Riverside Ave
Jacksonville, FL 32202
(904) 791-6626
noel.mcketty@fcso.com

OR

Donna Mohr
University of North Florida
Mathematics and Statistics
4567 St. Johns Bluff Road
Jacksonville, FL 32224
(904) 565-9053
dmohr@unf.edu