Biased Sampling: Solution for Lower Incidence Rate

M Muthu Mangai

Copyright 2008 by the Society of Actuaries.
All rights reserved by the Society of Actuaries. Permission is granted to make brief excerpts for a published review. Permission is also granted to make limited numbers of copies of items in this monograph for personal, internal, classroom or other instructional use, on condition that the foregoing copyright notice is used so as to give reasonable notice of the Society's copyright. This consent for free limited copying without prior consent of the Society does not extend to making copies for general distribution, for advertising or promotional purposes, for inclusion in new collective works or for resale.

Abstract

Given the lower incidence rate, use of decision tree techniques like Classification and Regression Tree (CART) in understanding credit or operational risk becomes quite challenging. A commonly adopted solution is biased sampling approach, where more weights are attached to bad customers to artificially hike the incidence or bad rate. While adopting this type of biased sampling approach, question of the best weight arises. This paper adopts an iterative approach in identifying the best weight. The best weight for bankruptcy (BKO) profiling problem in hand occurred when the incidence rate was around 50 percent where entropy reaches its maximum.

1. Introduction

Incidence rates are quite low when looking at credit risk (bankruptcy (BKO)/age loss rates) or operational risk (any type of fraud loss rates). This becomes a real challenge when we attempt to model this behavior using decision tree techniques like Classification and Regression Tree (CART). A general solution adopted is biased sampling approach. We introduce bias in the sample either by sampling down goods or by assigning weights to bad (example: one bad is as bad as 10 bads). Weights are assigned only for getting good separation in relation to the target. After the completion of the profiling exercise for validation and estimating the lift from the profiling exercises, bias in the sample has to be removed. While adopting this weights approach, the question of best weight arises. In this paper we take the problem of profiling BKO for personal loan products of well established financial organizations in the United States and try to understand the impact of weight on the profiling exercise. BKO loss rate for this portfolio stands at 2 percent, which perfectly fits the attempted exercise.

2. Decision Tree & CART Technique

Unlike econometric modeling, where the objective has largely to do with proving a hypothesis or relationship, data mining aims at extracting hidden and predictive information from a large database. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Wide use of data mining tools is largely driven by massive data availability and data mining algorithms.

Decision tree is one of the most commonly used data mining tools that produces readable description of trends in the underlying relationships and favors prediction. It is represented by a set of rules that is nothing but conditional probability. The population is segmented into smaller groups called terminal nodes or leaves. These terminal nodes, defined in terms of input variables, are expected to be homogeneous with respect to a target variable. Homogeneity among groups favors prediction of behavior with greater certainty; hence the concept of node “purity” or homogeneity is crucial in developing a decision tree. One of the ways of defining a leaf or node purity leads to the leading algorithms for constructing decision trees, Classification and Regression Tree (CART).

Homogeneity can also be achieved by the segmenting of data by pure subjective business logic. The difference between subjective business logic and a scientific tried and tested algorithm like CART would be the efficiency in getting homogeneity and the level of homogeneity. Subjective business logic has the disadvantage of a large and unstable tree. Unstable and different runs would produce different results.

In the case of the CART technique, the objective is to divide the population into classes. In our case, it means that we are exactly able to identify the profile of the BKOs versus non-BKOs. By this we mean that at end nodes we have 100 percent BKOs or 100 percent non-BKOs. A classification tree requires the dependant or objective variable to be categorical and the independent variable to be ordered continuous or unordered categorical. Under the CART technique, a tree is created by repeatedly partitioning data, and at a given stage only binary splits

are applied. It uses the concept of information gain or entropy reduction for selecting the optimal split.

Its goodness of split is defined as maximum reduction in impurity.

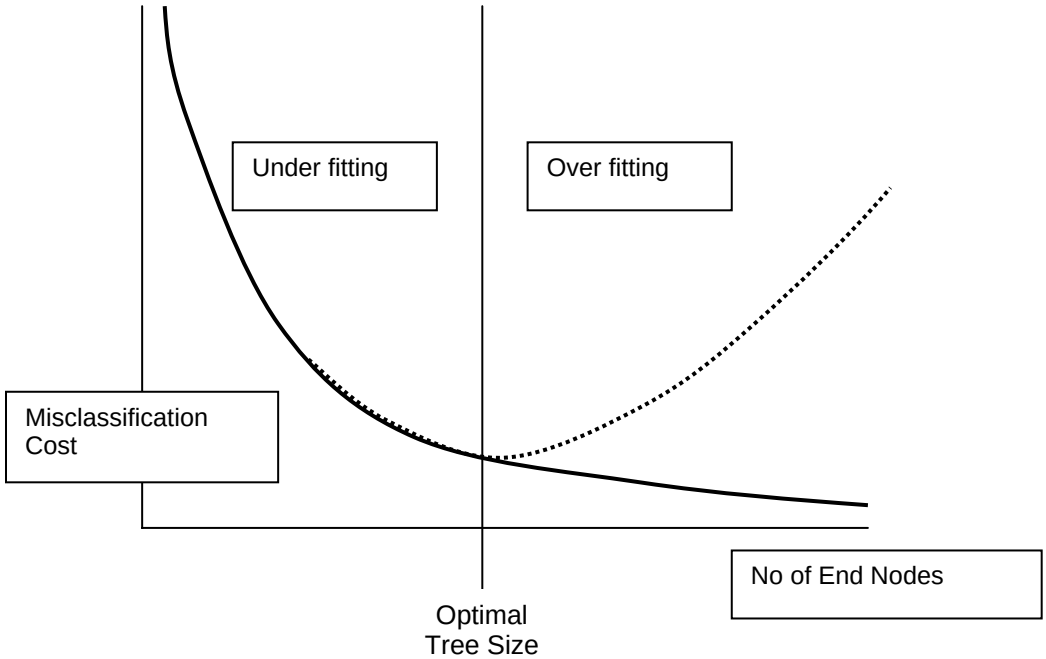
$$\Delta I(s,t)=I(t)- pr_l(t_l) - pr_l(t_r)$$

Choice of split:

$$\Delta I(s^*,t)=\max s\in S\ \Delta I(s,t)$$

The CART algorithm first defines a candidate set of S that would be comprised of all potential binary splits at each end node. After creating the candidate set, it selects the split that gives the largest decrease in impurity. The next question that arises is when we stop growing the tree. This is addressed by pruning methodology, where the maximal tree, which is the largest tree, is grown. As the next steps, the algorithm gets rid of the overgrown tree that's not supported by test data.

The maximal tree will always fit the learning dataset with higher accuracy, but it fails to estimate the performance of the tree on an independent set of data. Decision cost or misclassification cost reduces as tree size (no of end nodes) increases on the training datasets. As the size of the tree increases, relationships are estimated with greater accuracy and indicates it's too specific to development data. Most likely this estimated relationship might not hold on validation dataset; therefore we can expect the misclassification cost to increase in the validation data set for the same tree after a point. This point where the misclassification starts to increase can be interpreted as a point where we are accounting for sample specific trends.



One of the scientific ways of arriving at the optimal tree is "Minimal Cost Complexity Pruning Algorithm." Under this methodology, misclassification cost can be redefined as:

$$R\alpha(T)=R(T) + \alpha\ |\Gamma| \text{ , where } |\Gamma| = \text{no of terminal nodes in T}$$

3. Problems of Lower Incidence Rate and Biased Sampling as Solution

When profiling bankruptcy or defaulting behavior, we face the problem of lower incidence rates. As expected for a personal loan product, the BKO rate is less than 2 percent. With this type of low incidence rate, we would fail to get good separation. Biased sampling is a quick solution applied to handle these types of rare events. One BKO customer is not equal to a good customer, as the money we might lose from a BKO customer might be significantly larger than the money we gain from a good customer. Example: the average loss from a BKO customer is $1,000 and the average revenue from a good customer is $50. Based on this example, to compensate for the loss incurred from one BKO customer, we would need 20 good customers. Hence we can attach the weight of 20 to BKOs. This approach is more driven by business understanding and has its own limitation as it's based only on averages that might not hold as the population is being segmented.

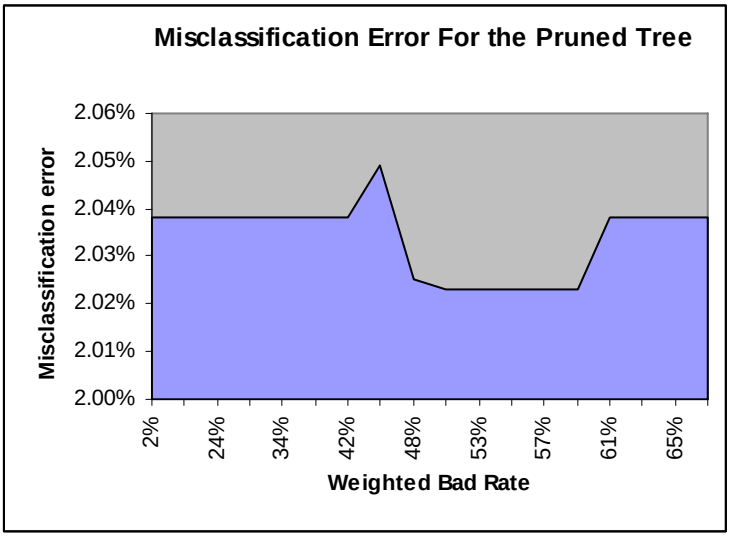
An alternative approach applied in this profiling exercise is the iteration approach, where different iterations can be carried out with different weights. As explained earlier, introduction of bias in the data is purely for profiling purposes. Results after introducing the bias will not represent the actual portfolio numbers. Hence, to understand the performance of the segmentation, bias in the data has to be removed. Performance of a CART can be estimated by the total misclassification error. This measure can be used for comparing different trees obtained by applying different weights. A tree with minimum misclassification error is identified. Weight applied in this tree is the best weight that can be used for the profiling exercise.

4. Results

Weights were selected at five point breaks from 10-95 to control the number of iterations.

To ensure there is no subjectivity in tree building, the exercise was carried out using the auto build option available in the tool. Pruning was carried out using "Minimal Cost Complexity Pruning Algorithm" available in the tool using the validation dataset. Results of the iterative exercise are summarized in the table.

Weights	Actual Bad Rate	Weighted Bad Rate	Misclassification Error For the Pruned Tree
1 (No Weight)	2.0%	2.0%	2.038071%
10	2.0%	17.2%	2.038071%
15	2.0%	23.8%	2.038071%
20	2.0%	29.4%	2.038071%
25	2.0%	34.2%	2.038071%
30	2.0%	38.4%	2.038071%
35	2.0%	42.1%	2.038067%
40	2.0%	45.4%	2.049212%
45	2.0%	48.4%	2.025282%
50	2.0%	51.0%	2.023001%
55	2.0%	53.4%	2.023001%
60	2.0%	55.5%	2.023001%
65	2.0%	57.5%	2.023001%
70	2.0%	59.3%	2.023001%
75	2.0%	60.9%	2.038071%
80	2.0%	62.5%	2.038071%
90	2.0%	65.2%	2.038071%
95	2.0%	66.4%	2.038071%



When the incidence rate is low it is observed that the pruning exercise cut back the full tree. In the BKO profiling exercise, for no weight and weights from 10-30 and 75-95 (with five point break), the tool pruned back the tree to mother node. From 35-70 weight, which gave bad/incidence rate of from 42 percent to 59 percent, it gave a pruned version of the tree. The lowest misclassification error happened at multiple weights 50-70, which gave an incidence rate of 51 to 59 percent. Though these weights resulted in a different maximal tree, the pruned tree was the same. Hence the misclassification errors are the same across these trees.

When the incidence rate is 50 percent, entropy or impurity is maximum. It appears that in order to get a pruned tree with the lowest misclassification error, a higher level of entropy in the dataset is required. In this exercise the best pruned tree is obtained when entropy is around the maximum. This observation needs to be validated on various profiling problems like fraud before being interpreted as a general rule. Nevertheless, given the advance tools with very low processing time, it might be a worthwhile option to adopt an iterative approach to decide on the best weight.