

Biased Sampling - Solution For Lower Incidence Rate:

M Muthu Mangai*

Senior Consultant, Genpact Analytics

Genpact India,
#99 Surya Park , Electronic City,
Ring Road,
Bangalore- 560100,
India

February 2008

* Corresponding author. Tel.: +91 080-4119-8086

E-mail addresses:

Muthu.mangai@gecis.ge.com (Muthu Mangai)

ABSTRACT:

Given the lower incidence rate, use of decision tree techniques like “Classification & Regressing 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 is identifying the best weight. Best weight for Bankruptcy (BKO) profiling problem in hand, occurred when the incidence rate was around 50% where entropy reaches its maximum.

Key Words:

CART, Credit risk, Operational risk, Decision tree, Optimal Weight, Maximize separation

Paper Type:

Application

Introduction:

Lower incidence rate are quite low when look 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: 1 Bad is as bad as 10 bad's). 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, question of best weight arises. In this paper we take the problem of profiling BKO for Personal Loan product of well established financial organization in US and try to understand the impact of weight on the profiling exercise. BKO Loss rate for this portfolio stands at 2% which perfectly fits the attempted exercise.

Decision Tree & CART Technique:

Unlike econometric modeling where objective is 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 tool that produces readable description of trends in the underlying relationships and favors prediction. It is represented by a set of rules that are 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 concept of Node "purity" or homogeneity is crucial in developing a decision tree. One of the ways of defining a leaf or node purity lead to the leading algorithms for constructing decision trees *Classification and Regression Trees (CART)*. Homogeneity can also be achieved by segmenting of data by pure subjective business logic. Difference between subjective business logic and scientific tried and tested

algorithm like CART would be efficiency in getting homogeneity and level of homogeneity. Subjective business logic has the disadvantage of large and instable tree. Instable and different runs would produce different results.

In case of the CART Technique, the objective is to classify the population into classes. In our case, it means we are exactly able to identify the profile of the BKO vs Non BKO's. By this we mean, at end nodes we have 100% BKO or 100% Non BKO's. A classification tree requires the dependant or objective variable to be categorical and independent variable to be ordered continuous or unordered categorical. Under the CART Technique, a tree is created by repeated 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.

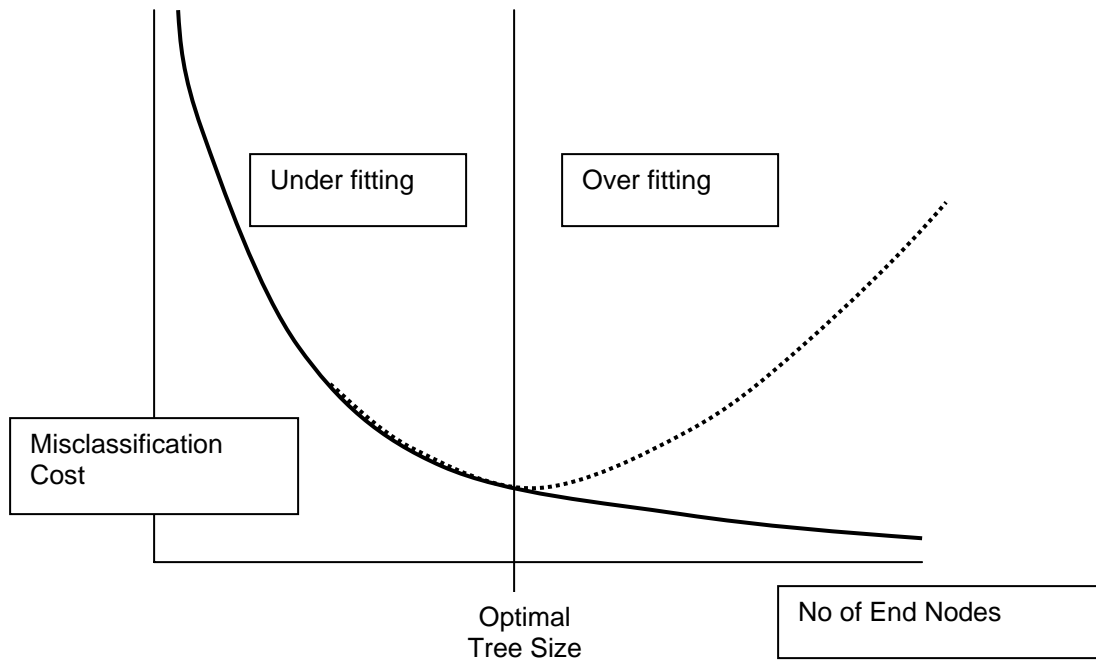
$$\Delta I(s,t) = I(t) - p_{r_l}(t_l) - p_{r_r}(t_r)$$

Choice of split :
$$\Delta I(s^*,t) = \max_{s \in S} \Delta I(s,t)$$

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

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, it indicates relationship too specific to development data sets is captured and relationships are estimated with greater accuracy. Most likely that 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

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

Problems of Lower Incidence Rate and Biased Sampling as Solution:

When profiling bankruptcy or defaulting behavior, we face the problem of lower incidence rate. As expected for Personal Loan product, BKO Rate is less than 2%. With this type of low incidence rate, we would be failing to get good separation. Biased sampling is a quick solution applied to handle these types of rare events. 1 BKO customers is not equal to a good customer, as the money we might lose from BKO customer might be significantly larger than the money we gain from a good customer. Example: Avg Loss from a BKO customer is 1000\$ and Average revenue from good customer is 50\$. Based on this example, to compensate for the loss incurred from 1 BKO customer, we would need 20 good customers. Hence we can attach weight of 20 to BKO's. This approach is more driven by business understanding and has its own

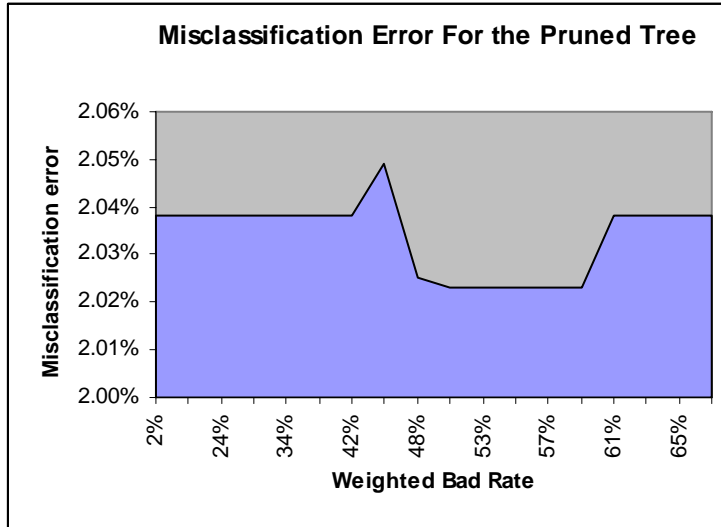
limitation as it's based only on averages which might not hold as the population is being segmented.

Alternative approach applied in this profiling exercise is 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 purpose. 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. Tree with minimum misclassification error is identified. Weight applied in this tree is the best weight which can be used for profiling exercise.

Results:

Weights were selected at 5 point break from 10-95 to control the number of iterations. To ensure there is no subjectivity in tree building exercise was carried out using auto build option available in the tool. Pruning was carried out using “Minimal Cost Complexity Pruning Algorithm” available in the tool using 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 incidence rate is low it is observed that pruning exercise cut back the full tree.

In the BKO Profiling exercise, for no weight and weights from 10-30 and 75-95(with 5 point break) tool pruned back the tree to mother node. From 35-70 weight, which gave bad/incidence rate of from 42%-59% gave a pruned version of the tree. Lowest misclassification error happened at multiple weights 50-70 which gave incidence rate of 51-59%. Though these weights resulted in different maximal tree, pruned tree was the same. Hence the misclassification errors are same across these trees.

When incidence rate is 50% entropy or impurity is maximum. Looks like to get a pruned tree with lowest misclassification error, 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 interpreting as general rule. Never the less, given the advance tools with very low processing time it might be a worth while option to adopt an iterative approach to decide on the best weight.