Authors: Dan Steinberg, Ph.D., Nicholas Scott Cardell, Ph.D., Mikhail Golovnya, M.Sc.
Salford Systems, San Diego CA
April 2005

Default Models based on "Small" Data Sets: Leveraging Multi-tree methods for reliable scorecards

Credit risk scoring models are typically based on large to huge databases. Even modest finance companies in out out of the way countries can have 1 million or more borrowers and global finnacial services companies enjoy tens of millions of accounts. However, not all areas of risk modeling are well data-endowed, particularly when the sample size of interest regards a relatively rare event. One example we have encountered frequently is small company corporate borrowing where even substantial lenders may have only a few hundred defaults. For such cases, the limiting data factor is the number of defaults and when there are few defaults it does not matter how many goods the lender has in their portfolio. In this paper we discuss tree-based ensemble methods which combine the results of possibly thousands of trees to generate surprisingly high performance generalizable models. In particular, we focus on new methods recently developed separately by two of the original authors of the CART(R) classification tree: Leo Breiman and Jerome Friedman. What we observe is that by hitting the data set repeatedly with differently constructed trees it is possible to extract far more information from the small data set than is possible with either conventional logistic regression models or single decision trees. Examples are drawn from our consulting practice and cover: a motor vehicle scorecard, an SME scorecard, and a model to reverse engineer agency commercial bank ratings in Europe.

Outline:

Introduction: General Problem statement
- Specific Problem of our clients
  - Only a few hundred defaults
  - Only 60 records total for bank ratings model
- Further complicated by pervasive missing values

Overview of New Tools
- TreeNet(tm)/MART(tm) Friedman's stochastic gradient boosting
- RandomForests(tm) Breiman's partially randomly generated trees

Rationale
- Why do these multi-tree methpods work at all?
- How do we test in a limited data environment?
  - Cross-Validation and Out-of-Bag (OOB) testing

Exemplary Results
- Internally generated ROC and other performance measures
- insights into the data generation process derived from

dependcy graphs and tree-based cluster results
Post deployment results from one of our scorecards

Biography of first Author:

Dan Steinberg is the CEO and founder of Salford Systems, a data mining software provider best known for the CART(R) decision tree. He has been a member of
the data mining, statistics, credit risk scoring,
and analytical consultation communities for more than 20 years. His experience in the field includes statistical modeling at AT&T Bell Laboratories, Assistant Professor of Economics at the University of California, San Diego, and numerous modeling engagements with Fortune 100 clients. He obtained his Ph.D. in Economics from Harvard University, and he has received honors from multiple organizations including the US Department of Labor, SAS User's Group International, the American
Marketing Association, Japan's Deming Committee, the Association for Computing Machinery, and the Duke University/NCR Teradata Center for CRM. The latter two awards marked Salford Systems first place wins in the KDDCup 2000 data mining competition and the Duke/NCR Churn modeling competition. Dr. Steinberg has published
in statistics, econometrics, computer science, and marketing journals, and in addition to guiding the development of Salford Systems' data mining tools he has also led
the teams developing comprehensive logistic regression, survival analysis and other advanced data analysis software. He has been a featured data mining issues
speaker for the American Marketing Association, American Statistical Association and the
Direct Marketing Association.