A comparison between statistical and Data Mining methods for credit scoring in case of limited available data

Hassan Sabzevari
Department of Risk Management
Karafarin Bank, Tehran, Iran
Tel: (+98) 21 88660418-20
Fax: (+98) 21 88664600
Email: h.sabzevari@karafarinbank.com

Mehdi Soleymani
Department of Sciences
Shahid Beheshti University
Tehran, Iran
Tel: (+98) 21 88660418
Fax: (+98) 21 88664600
Email: soleymani_mehdi@yahoo.com

Eman Norbakhsh
Department of Risk Management
Karafarin Bank, Tehran, Iran
Tel: (+98) 21 88660418-20
Fax: (+98) 21 88664600
Email: i.noorbakhsh@karafarinbank.com

Abstract
Credit scoring is a method used to estimate the probability that a loan applicant or existing borrower will default or become delinquent. There are two types of methods used for scoring: Traditional statistics models like Probit and Logistic regression and Data Mining models such as Classification and Regression Trees (CART), Multivariate Adaptive Regression Splines (MARS) and Bootstrap Aggregating. In this paper, we have examined the performance of different models in credit scoring on real data of a bank and compared two approaches above. We found out that Bootstrap Aggregating (Bagging) model in Data Mining approach and the Logit regression in traditional statistical credit scoring performs better than other methods. The real-world data used for scoring models contains not only many observations, but also a large number of features. Some of the features may be irrelevant or redundant due to their high inter-correlation. With many irrelevant and redundant features, most of the classification algorithms suffer from extensive computation time, possible decrease in model accuracy and decrease in scoring interpretation. We also present an empirical study on the machine learning feature selection methods, which provide an automatic technique for reducing the feature space. The study illustrates how these methods help to improve the performance of scoring models.

Keywords
Credit Scoring, Probit and Logit Regression, CART, Neural Network, MARS, Bagging, Feature selection.