Chapter 7.22
Mobile Phone Customer Type Discrimination via Stochastic Gradient Boosting

Dan Steinberg
Salford Systems, USA

Mikhaylo Golovnya
Salford Systems, USA

Nicholas Scott Cardell
Salford Systems, USA

ABSTRACT

Mobile phone customers face many choices regarding handset hardware, add-on services, and features to subscribe to from their service providers. Mobile phone companies are now increasingly interested in the drivers of migration to third generation (3G) hardware and services. Using real world data provided to the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2006 Data Mining Competition we explore the effectiveness of Friedman's stochastic gradient boosting (Multiple Additive Regression Trees [MART]) for the rapid development of a high performance predictive model.

INTRODUCTION

The PAKDD 2006 Data Mining Competition required the analysis of real world data from an industry that includes people of all ages and walks of life. In economically developed regions it is increasingly common for elementary school children to have their own mobile phones, and mobile communication is now preferred over fixed lines for undeveloped regions. Evolving 3G technologies offer a considerable expansion of the communication services routinely supported by mobile phone networks to include multi-player games, video conferencing, and enhanced Web browsing. Excitement over 3G technology has

waxed and waned since 2000 as the early promises were not fulfilled, but 3G is now becoming a fixture of the global mobile marketplace. Thus, a competition focused on analysis of 3G mobile phone customers is both topical and readily understood by data analysts, modelers, and business decision makers from all industries.

This Salford Systems report is organized as follows. In the first sections we offer our understanding of the competitive challenge, the data available, and how we framed the modeling objectives. The competition organizers have provided their own description of the nature of the modeling challenge and the data, but we believe that our perspective on these topics is somewhat different and is thus needed to explain our strategy. In the second section we provide a summary of the key descriptive statistics that gave us our initial picture of the nature of the data and its adequacy for modeling purposes. The third section describes our modeling methods and reports our results and performance based on the labeled data. The fourth section delves further into the results to examine specific findings at the predictor level. Finally, the last section summarizes our results and offers conclusions.

THE MODELING CONTEXT

The data provided for the PAKDD 2006 modeling competition consisted of summary data for each of 18,000 customers of an Asian mobile phone service provider. The data included customer demographics, a calling plan indicator, 6-month summaries of calling behavior, handset characteristics, summaries of billing amounts and late payment patterns, and other communication-related behavior, including Web, e-mail, and game usage. The training data came in the form of a flat file containing 252 columns, with 15,000 rows drawn from second generation (2G) customers and 3,000 rows drawn from 3G customers. In addition, a further 6,000 rows of prediction set data were provided in the same format, but with the 2G/3G flag suppressed. Essentially, the competition required the development of a classification model learned from the training set able to predict the 2G/3G class membership of the customers in the prediction set. However, some fine points regarding the competition require elaboration.

In predictive modeling of a binary (2 class) outcome, a number of performance criteria have been discussed extensively in the literature. For example, Caruana (2004) discusses cross-entropy (likelihood), the area under the receiver operating characteristic (ROC) curve, and classification accuracy in the context of the KDD2004 competition, and lift in a specified percentile was used as one performance criterion in the Duke/NCR Teradata 2002 churn modeling competition. In the PAKDD 2006 competition, the stated performance measure was classification accuracy, a metric that by itself appears to take no account of the ability of a model to properly rank order data from most probable to least probable 3G class membership. This competition had an important wrinkle, however. Classification accuracy was to be measured for the 3G class only. The competition organizers wanted to rule out the degenerate solution (all customers are 3G) as uninteresting, and also rule out what they termed "manipulated" solutions. A successful manipulated solution can be extracted from a model "that has strong rank ordering performance" by assigning the least probable customer to the 2G class and all others to the 3G class. This "solution" would have a high probability of yielding a perfect score on the 3G class, because even a moderately good model should be able to successfully identify a single 2G customer to place in the 2G class. Such a solution would presumably be disqualified as manipulated.

Less obviously manipulated solutions are possible, however. Given a good rank ordering of the customers by the probability of being 3G, a decision rule that assigns relatively few records to the 2G class should exhibit a high classifica-