**IGI PUBLISHING** **ITJ3623**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-pub.com

This paper appears in the publication, **International Journal of Data Warehousing and Mining**, Volume 3, Issue 2
edited by **David Taniar © 2007, IGI Global**

# Classification Of 3G Mobile Phone Customers

*Ankur Jain, Inductis India Pvt. Ltd., India*

*Lalit Wangikar, Inductis India Pvt. Ltd., India*

*Martin Ahrens, Inductis India Pvt. Ltd., India*

*Ranjan Rao, Inductis India Pvt. Ltd., India*

*Suddha Sattwa Kundu, Inductis India Pvt. Ltd., India*

*Sutirtha Ghosh, Inductis India Pvt. Ltd., India*

## ABSTRACT

*In this article we discuss how we have predicted the third generation (3G) customers using logistic regression analysis and statistical tools like Classification and Regression Tree (CART), Multivariate Adaptive Regression Splines (MARS), and other variables derived from the raw variables. The basic idea reflected in this paper is that the performance of logistic regression using raw variables standalone can be improved upon, by the use for various functions of the raw variables and dummies representing potential segments of the population.*

*Keywords:* *3G technology; CART; contingency table; customer look-alike modeling; Hosmer-Lemeshow; logistic regression; MARS; missing imputation; model performance; outlier treatment; predictive modeling; population segments; SAS; sensitivity analysis; statistical modeling*

## INTRODUCTION

An Asian telecommunication operator which has successfully launched a 3G mobile telecommunications network would like to make use of existing customer usage and demographic data to identify which customers are likely to switch to using their 3G network.

The objective of this competition was to develop a prioritization mechanism that will accurately predict as many current 3G customers as possible from the "holdout" sample provided. It also involved identifying the profiles of 3G customers that can be used in identifying potential 3G customers among the existing second generation (2G) base.

The competition organizers were provided with a sample of 24,000 mobile phone subscribers, out of which customer

type was provided for 18,000 subscribers, 15,000 being 2G and the rest 3G. Around 250 variables describing call and usage-related information was provided for all of the 18,000 subscribers. A holdout sample of another 6,000 subscribers was provided with the same set of variables, but without the 2G/3G flag. The task was to accurately predict as many 3G customers as possible from the holdout sample.

The organization of the article is as follows: We discuss the methodology approach taken and the modeling techniques used to develop the logistic model. Then we discuss the model results and the cutoff we have selected to generate the predictions. Finally, we discuss an alternative approach that we have tried.

## METHODOLOGY APPROACH

### MODELING METHODOLOGY

The modeling approach used for determining the 3G customers is a combination of logistic regression, CART, MARS, and other derived variables. The CART and MARS are modeling tools of Salford Systems. This combination is an improvement over the logistic regression model with raw variables only. The potential segments of the population are identified by CART, and potential splines for various important variables obtained by MARS are used along with the other variables. Logistic regression is used as the dependent variable is dichotomous (reference Hosmer W. David, Stanley Lemeshow: Applied Logistic Regression, Wiley, New York (1989) Chapter 1 Pages 8-10, Chapter 2 Pages 25-29). In addition, we have selected specific segments of some of the raw variables, which have very high or low event rates.

The variables obtained from CART are indicators of potential segments of the population. By potential segments, we mean segments of population with very high or low event rates. These indicators are used in the logistic model as independent variables. MARS, on the other hand, generates splines from variables, thereby capturing important segments of a variable. These splines, termed as basis functions, are then used in the logistic model. Some variables have a very high or low event rate in a particular range. We have analyzed these ranges and created segments to be used as independent variables in the model. The CART, MARS, and other derived variables, when included in the model, show a higher predictive power than what is obtained from the raw variables standalone.

Once the CART and MARS variables have been included, a stepwise logistic regression is used to reach an optimum model. The stepwise regression is used for the sake of parsimony as the number of variables (raw, CART, MARS, and derived variables combined) is large, thereby creating a scope of overfitting. Moreover, by using a stepwise procedure it is ensured that the variables in the model are all significant at the desired level. The variance inflation factors of each of the variables entering the model are scrutinized in order to prevent multi-collinearity.

## MODELING TECHNIQUES USED:

### MISSING IMPUTATION AND OUTLIER TREATMENT

In order to prepare the population for building the model, missing values had to be imputed and outliers had to be smoothed out. Missing imputation is done on variables which have less than 70% missing values. Variables with more than 70% missing values are omitted. The respective medians