UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Handling class imbalance in customer churn prediction

Jonathan Burez [1]

Dirk Van den Poel [2]

May 2008

2008/517

[1] Researcher, Department of Marketing, Ghent University, Tweekerkenstraat 2, B-9000 Gent, Belgium
[2] Corresponding author: Dirk Van den Poel, Professor of Marketing Modeling / Analytical Customer Relationship Management at Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Gent, Belgium. Phone: +32 9 264 89 80, Fax: +32 9 264 42 79. Email: dirk.vandenpoel@UGent.be, URL: www.crm.UGent.be , www.textmining.UGent.be

D/2008/7012/26

HANDLING CLASS IMBALANCE IN CUSTOMER CHURN PREDICTION

Burez, J. & Van den Poel, D.[3]
Ghent University, Faculty of Economics and Business Administration, Belgium

ABSTRACT

Customer churn is often a rare event in service industries, but of great interest and great value. Until recently, however, class imbalance has not received much attention in the context of data mining (Weiss, 2004). In this study, we investigate how we can better handle class imbalance in churn prediction. Using more appropriate evaluation metrics (AUC, lift), we investigated the increase in performance of sampling (both random and advanced under-sampling) and two specific modelling techniques (gradient boosting and weighted random forests) compared to some standard modelling techniques.

AUC and lift prove to be good evaluation metrics. AUC does not depend on a threshold, and is therefore a better overall evaluation metric compared to accuracy. Lift is very much related to accuracy, but has the advantage of being well used in marketing practice (Ling and Li, 1998).

Results show that under-sampling can lead to improved prediction accuracy, especially when evaluated with AUC. Unlike Ling and Li (1998), we find that there is no need to under-sample so that there are as many churners in your training set as non churners. Results show no increase in predictive performance when using the advanced sampling technique CUBE in this study. This is in line with findings of Japkowicz (2000), who noted that using sophisticated sampling techniques did not give any clear advantage. Weighted random forests, as a cost-sensitive learner, performs significantly better compared to random forests, and is therefore advised. It should, however always be compared to logistic regression. Boosting is a very robust classifier, but never outperforms any other technique.

Keywords: rare events, class imbalance, undersampling, oversampling, boosting, random forests, CUBE, customer churn, classifier

[3] Corresponding author: Dirk Van den Poel, Professor of Marketing Modeling / Analytical Customer Relationship Management at Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Gent, Belgium. Phone: +32 9 264 89 80, Fax: +32 9 264 42 79. Email: dirk.vandenpoel@UGent.be, URL: www.crm.UGent.be , www.textmining.UGent.be