Customer-Base Analysis Using
Repeated Cross-Sectional Summary (RCSS) Data

Kinshuk Jerath Peter S. Fader Bruce G.S. Hardie

June 2007

Abstract

A number of researchers have developed models of repeat buying behavior that can be used as a basis for computing quantities of managerial interest such as customer lifetime value. The Pareto/NBD is an example of such a model.

These models typically require that the analyst have access to individual-customer-level data. However, this is not always practical. While many reporting systems are able to create simple data summaries for a fixed period of time (e.g., a quarterly histogram of number of purchases), the process of extracting raw individual-level data can be a time-consuming task. The mere process of physically transferring raw customer-level data can be risky, as a number of privacy-related news stories attest. And in many countries, data protection laws can complicate the process of transferring raw data to the analyst, especially when the analyst is located in another country.

This paper explores the possibility of estimating these models using repeated cross-sectional summaries of the transaction data (e.g., four quarterly histograms), focusing on the case of the Pareto/NBD model. Such summaries are easy to create, are easy to distribute, and are free of any privacy concerns. In particular, we examine i) how much “information” is lost when fitting the model to these repeated cross-sectional data summaries data instead of the raw individual-customer-level data, and ii) the number of cross-sections required to minimize information loss.

We carry out a comprehensive simulation study covering a vast spectrum of market scenarios characterized by various levels of customer base penetration and purchase frequency. Across most of these datasets, our results consistently establish that the model fit (and parameter values) associated with the use of RCSS data can closely match the model fit (and parameter values) associated with individual-customer-level data.