KDD-Cup 2000: Question 3

Winner's Report

Salford Systems

Dan Steinberg, Richard Carson, Deepak Agarwal, JunYan, Andre Rupp

dstein@salford-systems.com

Characterizing the high spenders at a site is a classic marketing research and Customer Relationship Management (CRM) task that can be tackled in several rather different ways. Rather than attempting to segment the high spenders through a cluster analysis, as many CRM analysts would do, we chose instead to organize our analysis around these questions:

- Who are the high spenders (demographically) and how do they compare to low spenders, non-spenders, the US population overall and the US internet using population?
- What products did the high spenders purchase?
- Where did the high spenders come from on the internet (referring url), and, where on the Gazelle.com web site did they spend their viewing time?
- When, in calendar time, did high spenders make their purchases?
- Why did the high spenders choose to purchase from Gazelle.com? Were they attracted to the site by banner ads, discounts, or other promotions or did they find the site through their own search?

Responding to the challenge required an extended process of data preparation and exploratory data analysis, combining our data with other information from the US Census, deciding on the specific contrasts to study (high spender vs. low spender, high spender vs. non-spender), and running a series of CART models to separate the high spenders from other groups. This section briefly describes the steps we took to arrive at our final analysis.

Before launching into data preparation and exploration we realized we had to become familiar with the site and the nature of the Gazelle business. We visited the site frequently and sent several non-technical (male and female) staffers on subsidized shopping trips. These visits familiarized us with the organization of the site, the characteristics of the major brands carried, some innovative features of the site such as the option to display a model wearing specific pantyhose products, and some of the difficulties a shopper might encounter in making selections (sizing conventions differed considerably across brands; adding an item to the shopping cart was easy but removing an item was nearly impossible).

A major hurdle was adapting to the fact that the site we were viewing during June and July, 2000, had evolved substantially from the site as it was in February, March and April, the period from which all our data were drawn. We therefore began a process of reverse engineering the earlier site. The Blue Martini Customer Interaction System (CIS) organizes a site into meaningful groupings of pages and serves up pages by combining templates with dynamically generated content. As the page view database provided to us contained template and content information in the form of file paths, it was possible to compare the paths at different points in time and determine (abstractly) what had been added or deleted. To capture the browser query strings we had to visit every page on the site; it was not possible to simply download the entire site as all pages were created dynamically by the server. Tracking the evolution of the site during the study period was also directly relevant in the CRM analysis: the Donna Karan brand was not carried during the first month for which we had data, and self-reported behavioral data was gathered on the registration page only in the second month.

The data preparation stage was largely conducted as part of our efforts in the accuracy challenge of Question 2. The processing of the clickstream database involved data cleansing, feature extraction, and the creation of summary data and various rollups. We created a session database with one record for each visit that summarized the page view and the purchase behavior of that session. In addition we created a visitor database with one record for every unique cookie found in the clickstream data. We also filled in information wherever possible. For example, the Blue Martini server added known demographics to page views whenever a registered user actually logged on but did not if the visitor declined to log in. We decided to fill in known variables if we could match the cookie ID even though, strictly speaking, the cookie ID identifies the visitor's computer rather than the visitor. We then merged all past and future summary information back to the clickstream database so that at every page view we had detailed information available to us regarding the number of past visits, how many pages were viewed in the last session and in the session before that and ever, what was viewed, what was purchased, and the waiting time between visits. We also recorded the same information for the future so that we also could see what would happen in subsequent page views and sessions. In addition to the clickstream data we had preference data available for the 3,000+ visitors who had registered at the site and detailed purchase information for the 1781 orders placed.

Because most of our analysis focused on the contrast between high spenders and low spenders, we leveraged as much as possible from the clickstream data. A key component of the analysis was to determine which differences were genuinely informative and which were artifacts of the data, illusions, or reflections of more interesting differences. We therefore needed to carefully assess each candidate difference by looking at the calendar time for which the data were available, the segments for which no data were missing, and the circumstances regarding the data generation. Once we had decided on admissible predictors, we proceeded with a series of CART trees, selecting those that were most informative from a business decision maker's perspective. We experimented with a broad range of trees, varying control parameters such as priors, and penalties placed on predictors with missing values and penalties placed on nominal predictors having many levels. We also experimented with including and excluding entire categories of predictors such as demographics, and different versions of clickstream aggregates, looking for trees that told the most interesting and defensible stories.