KDD-Cup 2000: Question 2

Winner's Report

Salford Systems

Dan Steinberg, N. Scott Cardell, Mykhaylo Golovnya

dstein@salford-systems.com

The 2nd KDDCup question focused on three popular legware makers, Donna Karan (DK), Hanes (H), and American Essentials (AE), from among hundreds of products and 25 brands on the Gazelle.com web site. The question required modelers to predict which of these brands was most likely to be viewed in the latter portion of a session given that session's initial page views. Available data included page view history from previous visits, self-reported demographic and preference information from the site's registration page, zipcode demographics for registrants provided by Acxiom, and the page sequences viewed. This new site was in a mild state of flux over the study period, with products and brands being added and deleted, and the appearance of the home page changing. In addition several marketing campaigns were launched to attract visitors to the site during this period, including popular product give-aways. The question asked modelers to provide predicted probabilities for each of the three brands DK, H, and AE as well as "Other", and a value-maximizing forecast. The evaluation criterion considered correctly predicting DK, H, or AE to be twice as valuable as correctly predicting "Other."

A lengthy series of steps was required to respond to the question, including: (1) a data cleansing phase to eliminate crawlers, human testers, anomalies, and other potentially misleading data; (2) construction of web visit histories and summaries for repeat visitors tracked by cookies (including the lengths of previous sessions, view counts of the three core brands, what the visitor did first and second in each session); (3) construction of "lagged variables" tracking behavior over the last six page views and selected counts over all page views in the current session; (4) additional feature extraction such as categorization of the 2200+ referring web sites into 50 groups; (5) choice of analytic tools; (6) definition of the target variable; (7) a probability analysis of the impact of "clipping" on the scoring data set; and (8) conversion of model outputs to optimal scores and predictions.

Because web mining is in its infancy, little prior knowledge exists to guide model development or choice of tools. In addition, most of the data, such as page descriptions, browser type, or visitor's operating system, are nominal. Further, the non-uniform misclassification costs induced by favoring correct brand predictions needed to be reflected in the model development. Given this context we elected to use CART® decision trees to develop a fully non-parametric model guided by misclassification costs. We partitioned the training data into learn, test, and validate (L,T,V) portions, making sure that all visits of a case history (cookie) were assigned consistently to one partition. Trees were grown on the learn partition, the best pruned sub-tree was determined by performance on the test partition, and the results were checked for agreement with the validation partition.

The target variable for our CART trees was an eight-class indicator representing whether the remainder of the session contained none of the three core brands (O), exactly one (D, H, or A), exactly two (DK+AE, DK+H, or H+AE), or all three brands (DK+H+AE). The models were trained on all learn sample pages from first to last to maximize the amount of data available; thus, a person with T page views contributed T observations to the training data. From the perspective of each page view we forecast which of the eight outcomes was most likely in the remainder of that session. As we moved through a session the information available to us increased and the forecast was suitably revised.

Two additional fine points needed to be taken into account to obtain the final model and convert results to scores. First, the frequency distribution of the target variable's eight classes was not expected to be the same in the training and scoring databases since the training data consisted of complete and uncensored sessions whereas the scoring data was subject to a powerful censoring process described below. Second, the objective of the question was to maximize not simple predictive accuracy, but a value function reflecting the extra benefit of predicting a core brand correctly.

To deal with the target variable distribution we used CART priors to effectively reweight the data. The priors were estimated from the training data by simulating the expected clipping process and observing the target variable distribution on the last surviving page. To reflect the valuation function, a cost matrix was used with the following costs: 1 for misclassifying other (O) as any brand, 2 for misclassifying any brand, and .001 for confusing overlapping outcomes such as AE and H+AE with each other. Preliminary trees grown on sessions of all lengths suggested that the data structure varied considerably by the rank of the page view (first, second, etc). We therefore developed separate models for the first page view, the second page view, the third and fourth page views pooled, and all views from the fifth on.

Our goal in developing the CART trees was to generate homogenous groups of records (terminal nodes) which could then be optimally scored. It is at the scoring stage that the probability analysis of the censoring mechanism becomes critical. Recall that our training process used every page view of every session. In the scoring data we knew that every session of length one would be kept intact, and every other session would be randomly clipped to a shorter length. This meant that a session that was actually of length $T>1$ in the scoring data would be clipped to length S with probability $1/(T-1)$ for $S=1,\ldots,T-1$ (and thus all sessions of length 2 would be clipped to length 1). Looking at the terminal nodes in each CART tree we weighted training data cases by the appropriate clipping probability and calculated revised within-node probabilities for each of the possible outcomes. With probabilities p_O, p_{AE}, p_{DK}, p_H estimated for the four outcomes O, AE, DK, H (the probabilities sum to greater than 1), we predicted O if $p_O > 2*\max(p_{AE},p_{DK},p_H)$ and the most probable brand otherwise. Thus "Other" was predicted only if its probability was more than twice that of the highest probability brand; otherwise, the highest probability brand was predicted.