KDD-Cup 2000 Organizers' Report: Peeling the Onion

Ron Kohavi
Blue Martini Software
2600 Campus Dr.
San Mateo, CA 94403
ronnyk@bluemartini.com

Carla E. Brodley
School of Electrical and Computer Engineering
Purdue University
West Lafayette, In 47907
brodley@ecn.purdue.edu

Brian Frasca
Blue Martini Software
2600 Campus Dr.
San Mateo, CA 94403
brianf@bluemartini.com

Llew Mason
Blue Martini Software
2600 Campus Dr.
San Mateo, CA 94403
lmason@bluemartini.com

Zijian Zheng
Blue Martini Software
2600 Campus Dr.
San Mateo, CA 94403
zijian@bluemartini.com

ABSTRACT

We describe KDD-Cup 2000, the yearly competition in data mining. For the first time the Cup included insight problems in addition to prediction problems, thus posing new challenges in both the knowledge discovery and the evaluation criteria, and highlighting the need to "peel the onion" and drill deeper into the reasons for the initial patterns found. We chronicle the data generation phase starting from the collection at the site through its conversion to a star schema in a warehouse through data cleansing, data obfuscation for privacy protection, and data aggregation. We describe the information given to the participants, including the questions, site structure, the marketing calendar, and the data schema. Finally, we discuss interesting insights, common mistakes, and lessons learned. Three winners were announced and they describe their own experiences and lessons in the pages following this paper.

Keywords

KDD-Cup, e-commerce, competition, data mining, real-world data, insight, data cleansing, peeling the onion, best practices.

1. INTRODUCTION

The KDD-Cup is a yearly competition in data mining that started in 1997. KDD-Cup 2000, the fourth competition, involved multiple problems, following the suggestions of previous organizers [1]. For the first time, the Cup included insight questions in addition to prediction problems.

The domain for the KDD-Cup was e-commerce, considered a "killer domain" for data mining because it contains all the ingredients necessary for successful data mining [2]. The ingredients include (i) wide records (many attributes), (ii) many records (large volume of data), (iii) controlled data collection (e.g., electronic collection), (iv) ability to evaluate results and demonstrate return on investment, and (v) a domain where action can easily be taken (e.g., change the site, offer cross-sells). Blue Martini Software approached several clients using its Customer Interaction System to volunteer their data, and a small dot-com company called Gazelle.com, a legwear and legcare retailer, agreed to volunteer their data, properly sanitized.

After studying the data and consulting with Gazelle.com and retail experts at Blue Martini Software, five questions were defined. Two questions were prediction questions while the remaining three were insight questions. Only a portion of the available data was made available to competitors (about the first two months) while a test-set (the third month) was kept for evaluation, in line with standard best practices of having a separate test set.

To make the problem more realistic, we collected background information from Gazelle.com and made their marketing calendar available to competitors. The events (e.g., a TV advertisement) help explain the changes in the number of visitors over time.

Data was made available to in two formats: original data and aggregated data. While the original data was collected at the page request level, the questions were at the session and customer level. Because most tools do not have sufficiently powerful aggregation capabilities, we used the Blue Martini Customer Interaction System to generate the aggregated data, summarizing session-level and customer-level behavior. Further details about the data and aggregations are provided in Section 4.

The evaluation of the insight questions was done in consultation with Blue Martini's retail experts. We created a standardized scoring mechanism described in Section 3. As we evaluated the submissions whose statistics can be found in Section 5, we found many observations that were "shallow," i.e., they involved patterns that did not lead to deep understanding of the issues. We would like to highlight the need for "peeling the onion" when doing data mining investigations. Results and insights are described in Section 6.

We conclude the paper with lessons learned. Also in this issue are three reports from the winners of the competition.

2. BACKGROUND INFORMATION

It is helpful to know the following background information about the Gazelle.com webstore:

- The home page contained more than 70 images. This made downloads extremely slow for modem-based visitors.
- As with many dot-coms, Gazelle.com's initial goal was to attract customers, even if it meant losing money in the short term. They had many promotions that are relevant for mining, because promotions affected traffic to the site, the type of customers, etc. The important promotions were
  - FREE - Free shipping ($3.95 value). Active from March 20 to April 30 (shipping was normally free if sale was above $40).
  - MARCH1 - $10 off from March 1 to April 1.
  - FRIEND - $10 off from March 1 to April 30.
  - FREEBAG – A free bag from March 30 to April 30.

  Note that both the MARCH1 and FRIEND promotions offered $10 off. They were used for different purposes, and were run with different promotion codes.
- Gazelle.com ran a TV advertisement during a prime-time episode of the popular comedy show, Ally McBeal, on February 28.
- Gazelle.com changed their registration form significantly on February 26, so some customer attributes were only collected prior to this date and some were collected only after this date.

3. THE QUESTIONS AND EVALUATION CRITERIA

There were five independent questions for KDD-Cup 2000. Two of the questions were standard prediction problems with objective evaluation criteria, while the remaining three were subjective "insight" questions.

Question 1

Given a set of page views, will the visitor view another page on the site or will the visitor leave?

This question was motivated by the idea that knowing whether a visitor is likely to leave can help determine the "best" page to display (e.g., special promotions could be shown to encourage the visitor to stay). The evaluation criterion for this question was simply the number of correct predictions on the test set. The winner was the entry with the highest accuracy.

Question 2

Given an initial set of page views, which product brand ("Hanes", "Donna Karen", "American Essentials", or "Other") will the visitor view in the remainder of the session?

This question was motivated by the problem of improving navigation by automatically placing a hyperlink on the current page pointing to a particular brand page. To make the problem more manageable, we restricted the task to predicting one of three most commonly sold brands, or "Other" (defined as not viewing any of the three brands in the remainder of the session). The evaluation criterion was a weighted prediction score where points were awarded as follows:

2 points: If they predicted one of the three specific brands and one of the remaining pages in the session included the predicted brand.

1 point: If they predicted "Other" and none of the remaining pages in the session included a visit to one of the three specific brands.

0 points: All other cases.

The winner was the entry with the highest score.

For the remaining three questions, the competitors were required to submit text and graphs that a business user would be able to understand and find useful. Each submission was limited to 1,000 words and ten graphs.

Question 3

Given a set of purchases over a period of time, characterize visitors who spend more than $12 on an average order at the site.

The motivation for this question was that insight about a website's more valuable customers could be useful for determining marketing directions, product selection, etc.

Question 4

This was the same as Question 1, but the goal was to provide insight, rather than predict accurately.

Question 5

This was the same as Question 2, but the goal was to provide insight, rather than predict accurately.

For Questions 3, 4, and 5, for which no simple objective measure existed, we talked to retail experts at Gazelle.com and Blue Martini Software about the submissions. We then formalized the evaluations by collecting all of the significant insights, weighting them, and creating a combined score based on the insights found, the correctness of the submission, and the presentation of their submission (keeping in mind that *business users* were the target audience). The actual formula used for computing an entrant's score was

$$Score = 3P + 3C + \sum_{i=1}^{N} w_i I_i$$

where P is the entrant's presentation score (0-10), C is the entrant's correctness score (0-10), and for each insight i, w_i is the weight assigned to the insight and I_i is the entrant's score for the insight (0-2). The number of insights and their weights varied for each question.

The presentation score captured the effectiveness of presentation of the entrant's submission. This included factors like:

- How readable and easy to understand was the submission?
- Were there graphs, tables, and figures that business people could understand?
- Was there an effort to distill the important information, or was too much irrelevant information presented?

The correctness score was based on whether the entrant's claims were correct and whether the claims had sufficient data to support them.

For each question, we defined a complete set of insights based on all of the insights provided by every competitor. These insights were weighted to reflect how interesting they would be to a business user (based on conversations with retail experts from Gazelle.com and Blue Martini Software). Many insights were given low weight (and sometimes even zero weight) because they simply correlated with more fundamental insights. For each entry, every insight was awarded an insight score which was either zero (if they didn't discover the insight), one (if they partially described the insight) or two (if they fully described the insight). Due to the large number of insights (over 30 each for Questions 3 and 4), we do not include a list here. A complete list of insights with detailed explanations and weights can be found on the KDD-Cup 2000 home page [4].

4. THE DATA

In this section, we describe what data was collected in the webstore, how we generated the initial star schema for the data warehouse, what types of data cleansing/obfuscating were performed, and which data transformations were applied. Finally, we summarize the final schema and data formats provided for the KDD-Cup.

4.1 Initial Data Collection

Gazelle.com went live with Blue Martini's Customer Interaction System (CIS) on January 30, 2000 with soft-launch to friends and families. On the webstore, an application server in the Blue Martini architecture generates web pages from Java based templates. Among the other things, the architecture logs customer transactions and clickstreams at the application server layer. Since the application server generates the content (e.g., images, products and articles), it has detailed knowledge of the content being served. This is true even when the content is dynamically generated or encrypted for transmission commonly used for checkout. Weblog data is not needed. Application servers use cookies (or URL encoding in the absence of cookies) to keep track of a user's session, so there is no need for "sessionizing" clickstreams as there is for standard weblogs. Since the application server also keeps track of users using login mechanisms or cookies, it is easy to associate individual page views with a particular visitor.

Among the data collected by the Blue Martini application server, the following three categories are related to this KDD-Cup:

- Customer information, which includes customer ID, registration information, and registration form questionnaire responses.
- Order information at two levels of granularity: 1) Order header, which includes date/time, discount, tax, total amount, payment, shipping, status, and session ID; and 2) Order line, which includes quantity, price, product, date/time, assortment, and status.
- Clickstream information at two levels of granularity: 1) Session, which includes starting and ending date/time, cookie, browser, referrer, visit count, and user agent; and 2) Page view, which includes date/time, sequence number, URL, processing time, product, and assortment.

In general, each customer can have multiple sessions. Each session can have multiple page views and multiple orders. Each order can have multiple order lines. Each order line is a purchase record of one product with a quantity of one or more.

4.2 Star Schema Creation

The data collector in the Blue Martini application server is implemented within an On-Line Transaction Processing (OLTP) system. OLTP systems are designed for efficient handling of a large number of small updates and short queries. This is critical for running an e-commerce business, but is not appropriate for analysis, which usually requires full scans of several very large tables and a star schema[1] design [7][8] which business users can understand. For data mining, we need to build a data warehouse using dimensional modeling techniques. Both the data warehouse design and the data transfer from the OLTP system to the data warehouse system are very complex and time-consuming tasks. Because Blue Martini's architecture contains metadata about tables, columns, and their relationships, it can automatically construct the data warehouse from the OLTP system [6].

When preparing the data for the KDD-Cup, we integrated syndicated data from Acxiom into the schemas, which enriched the customer information for analysis by introducing more than fifty new attributes such as Gender, Occupation, Age, Marital Status, Estimated Income, and Home Market Value.

Two star schemas used for generating the KDD-Cup data are the Clickstream star and the Order Lines star. The Clickstream star consists of one fact table: "Clickstream" and six dimension tables: "Customer Profiles", "Acxiom", "Web Sessions", "Products", "Assortments", and "Contents". The Order Lines star consists of the one fact table: "Order Lines" and six dimension tables: "Customer Profiles", "Acxiom", "Order Headers", "Products", "Assortments", and "Promotions".

4.3 Data Cleansing/Obfuscating

To protect customer privacy, we removed attributes containing information about individuals such as Login Name, Password, Credit Card, Customer Name, and Session IP Address. We also removed attributes containing profit-related information such as Product Unit Cost. For attributes that we believe are important for mining this data (solving the KDD-Cup questions), we scrambled the data. For example, the values of the Email attribute were mapped to keep only the domain suffix such as COM, EDU, ORG, and GOV. In addition, we kept "Gazelle.com" for email addresses with the suffix of gazelle.com. All company names were mapped to "COMPANY" and a number, so that it is possible to tell people are from the same company without knowing which company it is. Session Cookie IDs were encoded, so that each Cookie ID appears as a different number, while it is still possible to determine that several sessions are from the same cookie.

Data cleansing is usually a part of the KDD process. We chose to do some initial data cleansing ourselves for three reasons. Firstly, unlike a real data mining project, the participants of the KDD-Cup did not have direct contact with the domain experts. Secondly, data obfuscating must be done before releasing the data, and thirdly, the questions are challenging enough even after this initial data cleansing. To clean the data, we

- Removed Keynote records. Keynote hit the Gazelle.com home page 3 times a minute, 24 hours a day, 7 days a

[1] A star schema is a join of database tables with one central fact table joined to several other tables (called dimensions).

week, generating about 125,000 sessions per month. These records can skew mining results.

- Removed test users. We used criteria such as with "test" in customer names or purchased using a credit card that were used by more than 15 different users. Note that the test users have very different purchasing and browsing behaviors.
- Removed returned and uncompleted orders. The number of these orders is small, but they may cause confusion.

4.4 Data Transformations

We provided two types of data for the KDD-Cup questions, namely unaggregated and aggregated.

The data transformation for the unaggregated data is very simple. Questions 1, 2, 4, and 5 share the same unaggregated dataset. It is a flat table created by joining the Clickstream star. In this table, each record is a page view. Session attributes are repeated multiple times if the session has multiple page views. Similarly, customer information is also repeated in the table. To define the targets for these four questions, we added three Boolean attributes in the table as follows. Three example sessions are given in Table 1, showing how the sessions were clipped.

- "Question 1 Test Set" indicating whether you will see this page view if the session is in the test set for Questions 1 and 4. This is defined based on a clipping point in half of the randomly selected sessions. For a selected clipping session, we randomly generated a clipping point between one and the session length minus one. No clipping was performed for sessions of length one.
- "Question 2 Test Set" indicating whether you will see this page view if the session is in the test set for Questions 2 and 5. This is defined based on a clipping point in all the sessions. The clipping point is generated in the same way as for Question 1.
- "Session Continues" as the target of Questions 1 and 4.

Session ID	Request Sequence	Question 1 Test Set	Question 2 Test Set	Session Continues
29	1	T	T	F
29	2	T	T	F
29	3	T	F	F
56	1	T	T	T
56	2	T	T	T
56	3	F	F	T
68	1	T	T	F

Table 1: How sessions got clipped.

The unaggregated dataset for Question 3 is also a flat table created by joining the Order Lines star. Each order line is a record in the table. Attribute values for order headers and customers may repeat multiple times. A Boolean attribute "Spend Over $12 Per Order On Average" is added to the table as the target. This attribute is defined at the customer level.

These two unaggregated datasets contain the raw data, providing enough information for those people with data transformation ability to do the data mining. Note that the first dataset does not contain the order information while the second dataset does not contain the clickstream information. Participants could join them together if they thought doing so could help them to solve the questions.

Considering that many researchers, especially those working on data mining algorithms, do not have software readily available to transform (including aggregate) the raw data, we provided an aggregated version of the data. The aggregated data consists of three datasets: one for Questions 1 and 4, one for Questions 2 and 5, and the other for Question 3. These datasets are derived by aggregating the two unaggregated datasets to the level of granularity appropriate for mining. That is, the session level for Questions 1, 2, 4, and 5 and the customer level for Question 3. At the same time, we added new attributes based on examination of existing attributes. For example, we extracted the session browser family names, the browser names, and the top three browser family names. In the two aggregated datasets for Questions 1, 2, 4, and 5, each session is a single record. During the generation of these two datasets, all page views marked "not in the corresponding test sets" in the unaggregated datasets were removed before the aggregation operation. In the aggregated dataset for Question 3, each customer is a single record.

The aggregation operations generated 151 and 153 new attributes for Questions 1 & 4 and Questions 2 & 5, respectively. Examples include the number of views of individual top products which were selected based on the statistics of the datasets, the number of views of assortments, the number of views of different templates, and information about the last page, which includes information appearing on it and its date/time information. For questions 2 and 5, we defined three numeric attributes indicating the number of views of the respective brands (Hanes, Donna Karan, American Essentials) in the remainder of the session. In addition, we also defined a Boolean attribute that was set to true if none of the brands were viewed in the remainder of the session and false otherwise.

When generating the aggregated dataset for Question 3, we joined clickstream data to the order lines data since we believed that clickstream can help to answer Question 3 and it is hard to join them after aggregation. The aggregation for this dataset was carried out at two levels: first to the session level and then to the customer level, generating 434 new attributes in total such as "Average Session Request Count", "First Session First Referrer Top 5", and "Percent of Products Purchased on Sunday".

4.5 Final Data Schema and Formats

The datasets were released in flat files using C5 format (www.rulequest.com), a widely used data format for data mining. There was no training/test split for Question 3 data, as it was a pure insight question. Questions 1 and 2 had training and test datasets. The training datasets contain the target information while the test datasets do not. To avoid leaks (with respect to targets), we did the training/test splits using time. The data we got from Gazelle.com was collected from January 30, 2000 to April 30, 2000 (3 months). We used the data before April 1, 2000 (2 months) for training for all of the questions. Since Questions 1 and 2 share information, their test sets could not overlap. We used the data after April 14, 2000 (half a month) as the test set for Question 1, and the data from April 1, 2000 to April 14, 2000 (half a month) as the test set for Question 2.

Table 2 summarizes the number of attributes and the number of records in the datasets. Questions 4 and 5 do not appear in the table because Question 4 used the same data as Question 1, while Question 5 used the same data as Question 2. It is worth mentioning that for Question 2, we had four target attributes in the training set, and only one dummy target attribute in the test set.

	Training set		Test set	
Question	Attributes	Records	Attributes	Records
1: Unaggregated	217	777,480	215	164,364
2: Unaggegated	217	777,480	215	142,204
3: Unaggregated	232	3,465	-	-
1: Aggregated	296	234,954	296	50,558
2: Aggregated	299	234,954	296	62,913
3: Aggregated	518	1,781	-	-

Table 2: Dataset statistics.

5. SUBMISSION STATISTICS

We received 170 non-disclosure agreements requesting access to the data. Of these, there were 31 participants who submitted an entry for one or more of the questions. The number of entries we received for each question is shown in Figure 1. Since the competition ended, we have received more than 190 additional click-through agreements for access to the data for research or educational purposes.

Each participant was also required to submit a questionnaire answering questions about their efforts. This included questions on the resources they utilized (e.g., the number of people involved and the number of hours spent in each phase of analysis), the software and hardware that they used, and the data mining techniques and methodologies used in both processing and analyzing the data. The statistics presented in this section are based on the answers we received in these questionnaires.

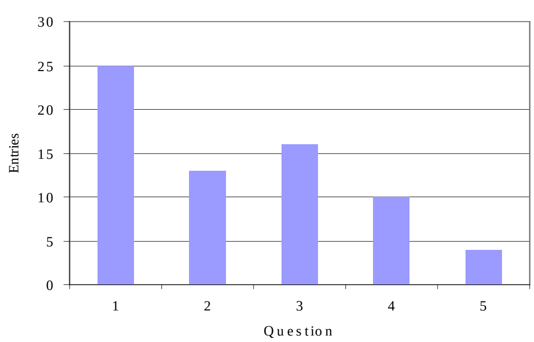


Figure 1: Number of entries for each question.

In total, the 31 participants spent 6,129 person-hours preparing and analyzing the data. This corresponds to about 200 person-hours per participant. One participant spent more than 900 person-hours on their submission. The number of people involved varied from one to thirteen, although most entries came from teams of two or three people. The breakdown of how the hours were spent on average is shown in Figure 2.

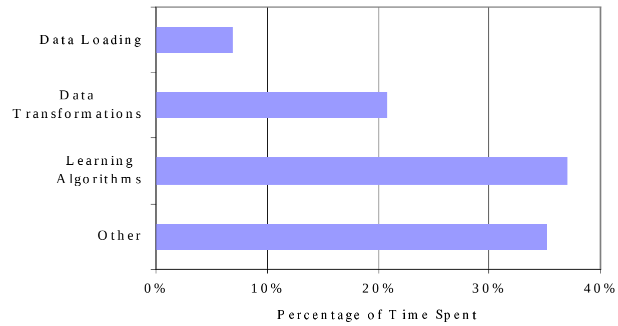


Figure 2: Average time spent on each phase of analysis.

Notice that, in contrast to most studies [5], less than 30% of the time was spent in data loading and transformations. Most likely, this was due to two factors. Firstly, the data was collected within Blue Martini's integrated e-commerce system designed for data mining, and thus was in a form more amenable to analysis [6]. Secondly, as described in Section 4, we spent significant time aiding the contestants by transforming the data and constructing new features for use in analysis.

The breakdown of data mining software origin used by participants is shown in Figure 3. One interesting trend to note is the increase in the use of commercial software for the KDD-Cup: the proportion of entries using commercial or proprietary software has grown from 44% (KDD-Cup 1997) to 52% (KDD-Cup 1998) to 77% (KDD-Cup 2000).

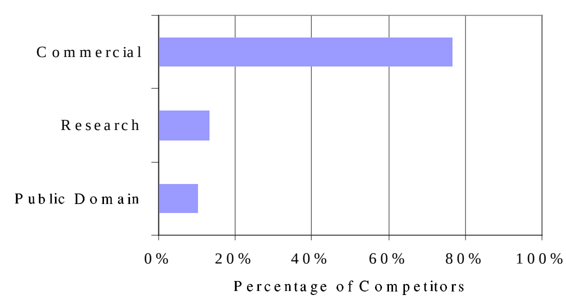


Figure 3: Type of software used by the competitors.

The operating system used by competitors was an even mix of Microsoft Windows (54%) and Unix (46%). Of those competitors using Unix, various flavors of commercial Unix accounted for 65%, while Linux accounted for the remaining 35%. Despite the balance between Microsoft Windows and Unix operating systems, the hardware used was primarily desktop PCs (73%), rather than Unix workstations (27%).

For data access, 32% of competitors used a database, while 68% used flat files. The breakdown of data processing tools used by competitors is shown in Figure 4. From this figure it can be seen that most competitors made use of the data processing tools built into their analysis software rather than developing proprietary data processing tools for the KDD-Cup.

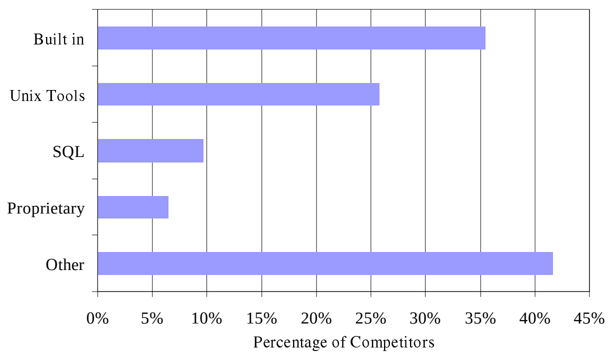


Figure 4: Data processing tools used.

As mentioned in Section 4, we provided both aggregated and unaggregated data. The majority of competitors used the aggregated data (59%) rather than the unaggregated (41%). This suggests that many data mining tools provide only limited support for data aggregation.

Figure 5 shows the top algorithmic techniques used by the competitors. The figure shows both the percentage of competitors who tried that algorithm and the percentage of competitors who submitted a solution to least one question using that algorithm. As can be seen, decision trees were by far the most popular choice, with more than 50% of the competitors submitting a solution to at least one question using decision trees.

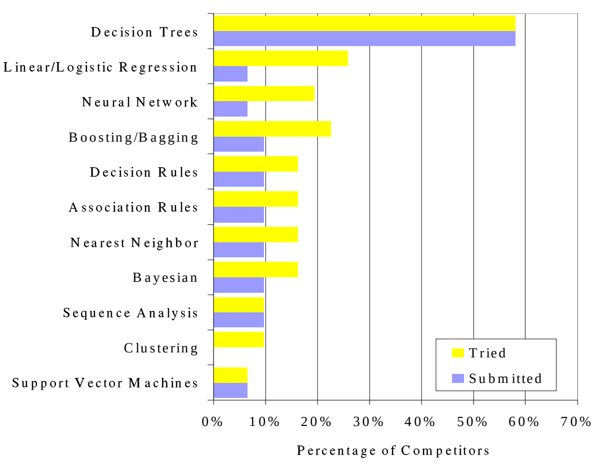


Figure 5: Algorithms tried versus submitted.

6. RESULTS AND INSIGHTS

In this section we present the results for each of the five questions.

Participants discovered many important actionable insights, including which referrers resulted in heavy spending, which pages cause abandonment, and what segments of the population are heavy spenders. Many seemingly interesting insights were obvious once one discovered the underlying cause, which was usually related to time or session length. For example, many participants noted a correlation between being a heavy spender and a visitor's answer as to whether they would like email from Gazelle.com. When this response is plotted against time, it is easy to see that it varies dramatically -- this is because Gazelle changed the default for this field twice. Predicting who would leave the site was made particularly challenging because many sessions were of length one -- in this data web crawlers that viewed a single page in each session accounted for 16% of sessions. Despite this, surprisingly few participants identified which visitors were actually web crawlers rather than real people. In examining the results when shorter sessions were removed, we noted that was possible to predict accurately when the prediction confidence was high.

For Question 1 (Given a set of page views will the visitor view another page on the site or will the visitor leave), the accuracy values ranged from 77.06% to 59.56% with a mean of 73.14%. The difference between the top two performers was only 0.10%, which translates into 50 sessions. In fact, the difference in accuracy of the top five participants was statistically insignificant (a 95% confidence interval corresponds to $\pm$ 0.37%). Despite this result, if we restrict the evaluation to predicting sessions with five or more page views the results are far more significant (the difference between first and second place was 1.5% and a 95% confidence interval corresponds to $\pm$ 0.79%). Figure 6 shows that the gains charts for the top two participants track the optimal gain for 10% of these longer sessions, which account for 43% of the target. The optimal gain is shown by the leftmost curve on the graph.

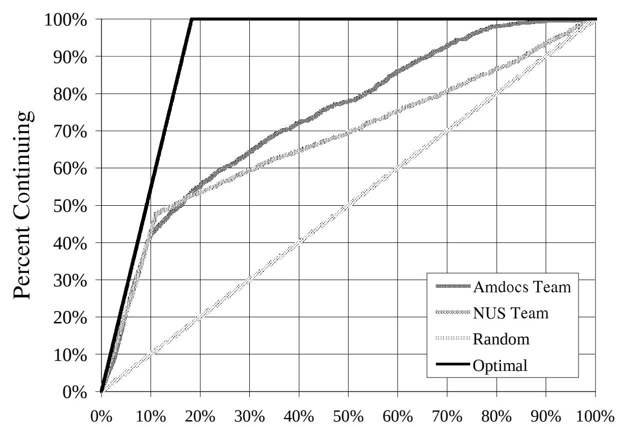


Figure 6: Cumulative gains chart for sessions with five or more page views.

Question 4 was the insight question corresponding to Question 1. Some of the key insights found were that web crawlers and gazelle testers leave and that the length of stay depends on the referrer site (users from Mycoupons had longer sessions, whereas users from ShopNow tended to leave quickly). Participants noted that a returning user's probability of continuing was double that of a first time visitor. Viewing some specific products caused users to leave the site. This is an example of an actionable insight, in that the web site might consider removing those products. Another actionable insight is that 32% of customers left after entering the replenishment section of the site. Many "discoveries" were explained by noticing that the probability of leaving decreases with the number of pages viewed in the session. For example, the insight that "viewing many different products in a session implies low abandonment" is explained by this fact.

For Question 2 (Given a set of page views which product brand will the visitor view in the remainder of the session), the scores ranged from 60956 to 60697 with a mean of 60814.8. Like Question 1, we found the difference between the top participants to be statistically insignificant. However, like Question 1, we observed very good lift curves when we restricted our evaluation to sessions with five or more page views. One of the best predictors was the referrer URL: Fashionmall and Winnie-Cooper are good referrers for Hanes and Donna Karan, whereas Mycoupons, Tripod, and Deal-finder are good referrers for American Essential. When we look more closely at this result we see that the American Essentials brand primarily contains socks, a low priced item which often falls under the $10 coupon price. Very few participants realized that the Donna Karan brand was only available starting February 26.

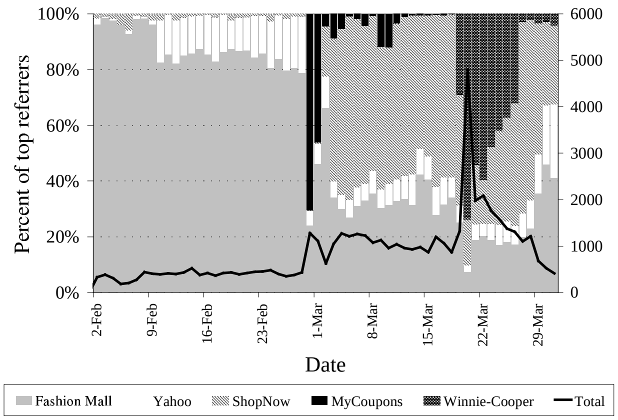


Figure 7: Top referrers by date.

For Question 3 (Characterize visitors who spend more than $12 on an average order at the site) many interesting insights were simply related to time. For example, noting that significant activity began on February 29th, when the TV ad, Friends promotion and hard launch occurred. Another example is that the referring site traffic changed dramatically over time (see Figure 7). Some of the deeper insights that arose from this observation were related to the conversion rate. While the overall conversion rate for the site was only 0.8%, Mycoupons had an 8.2% conversion rate, but generated low spenders. On the other hand, the conversion rate from Fashionmall and ShopNow was only 0.07% even though they brought 35,000 visitors. Some of the other factors correlating with heavy purchasers were:

- They were not an AOL customer (the Gazelle.com site displayed badly within the AOL browser window).
- They came to the site after seeing a print advertisement.
- They had either a very high or very low income.
- They were living in the Northeastern U.S.

7. LESSONS LEARNED

The KDD-Cup is a great tool for highlighting both to the data mining research community and to the users of data mining the issues faced by the participants and by the organizers. We now describe the main lessons learned.

The most important lesson is that humans are an important part of the KDD process, even if the only interesting measurement is accuracy or score (Questions 1 and 2). The fully automated tools can never match the human insight that allowed the winners to create multi-stage models (see the KDD-Cup 2000: Winner's Reports in this issue), identify crawlers, local testers, and construct additional features. The importance of human understanding was also apparent in the choice of algorithms tried versus submitted: Decision Trees were used the most often and submitted the most often while Neural Networks, Logistic Regression, and Clustering had the worst try-to-submit ratios. Many participants who thought they found an interesting result did not spend the time to "peel the onion" and find the true underlying causes. For the insight questions, the iterative process is even more important because many of the initial correlations are obvious and not interesting to the business users (e.g., those who purchase an item that costs over the heavy-spender threshold of $12 are indeed heavy spenders). Many insights that seemed interesting had extremely low support. For example, several participants claimed that all purchasers who came from Shopnow were heavy spenders. While the statement was true, the support was six people! With the human involvement required, it takes time to derive useful insight---hundreds of hours.

The changes to the site created interesting effects and biases. Those that ignored the documentation about special marketing events did not do well. Time is a crucial attribute and changes to the site and products needs to be taken into account. In one case, a competitor claimed that the problem was "too real" and that we should simplify it. Our questions were hard, but they represent real-world problems in a real-world setting. The results showed significant lift, especially on longer sessions, and many insights were extremely interesting and actionable. For many one-click sessions, it was impossible to predict well, but for when the confidence was high, especially on longer sessions, predictions were very good.

The data was collected through the Blue Martini application server, and avoids the use of standard weblogs and allows correlating purchases to clickstreams. The data collector also saves information about the products shown in addition to URLs, making information more stable across site changes. Such data was rich and easier to work with, and the addition of Acxiom attributes certainly helped in deriving insights. Even with all these advantages over weblogs, identifying crawlers and test users remains a hard problem

For future organizers of the KDD-Cup, we offer some suggestions. Before volunteering to organize the KDD-Cup, make sure you understand the amount of effort involved. We estimated that we spent a total of 800 hours on getting the data, cleansing it, obfuscating it, transforming it, setting the web pages, working on the legal agreements, and evaluating the results. Plan on spending significant time in thinking about data obfuscation and identifying "leaks" in the data (giveaway attributes that predict the target because they're downstream in the collection process). For example, our system stored the session length, an attribute that we had to recompute after clipping the data, or else it would giveaway the target. We were very careful about removing leaks this time, having seen problems in previous years, but we still had to re-release the data twice in the initial phase due to mistakes in randomization and cookie obfuscation. We spent significant time writing the introductory material, giving the background knowledge, explaining the columns, yet we still had to develop a FAQ, which had 67 questions at the end of the competition. We gave the participants two question periods, one right after we released the data, and one before submission. We believe this was useful to get people started and also allowed us to plan our time better. The evaluation took a very long time, especially creating the weighted list of insights and validating the insights. We asked the participants to write a report for business users, but after reading the reports we suspect that many of the authors have never talked to a business user. On the bright side, we learned many things about the data that we did not know and we saw some excellent methods to present results and make them more accessible to a wider audience.

8. ACKNOWLEDGMENTS

We thank Gazelle.com for providing the data. We thank the Acxiom Corporation for providing the syndicated data overlay. Catharine Harding and Vahe Katros, our retail experts at Blue Martini Software, were helpful in reviewing submissions and explaining some of the patterns. Sean MacArthur from Purdue University helped write the scoring code.

REFERENCES

[1] Ismail Parsa, KDD-Cup-97: A Knowledge Discovery and Data Mining Tools Competition talk. Newport beach, CA, USA. http://www-aig.jpl.nasa.gov/public/kdd97/.

[2] Ron Kohavi and Foster Provost, Applications of Data Mining to E-commerce (editorial), Special issue of the International Journal on Data Mining and Knowledge Discovery, Jan 2001. http://xxx.lanl.gov/abs/cs.LG/0010006.

[3] ACM *SIGKDD Explorations* homepage, the information for authors link: http://www.acm.org/sigkdd/explorations/.

[4] *KDD-Cup 2000* homepage. Carla Brodley and Ron Kohavi. http://www.ecn.purdue.edu/KDDCUP/.

[5] Gregory Piatetsky-Shapiro, Ron Brachman, Tom Khabaza, Willi Kloesgen, and Evangelos Simoudis, An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[6] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng. Integrating E-Commerce and Data Mining: Architecture and Challenges. *WEBKDD'2000 workshop: Web Mining for E-Commerce – Challenges and Opportunities*, 2000.

[7] Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite, *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, John Wiley & Sons, 1998.

[8] Ralph Kimball and Richard Merz, *The Data Webhouse Tookkit: Building the Web-Enabled Data Warehouse*, John Wiley & Sons, 2000.

About the authors:

Ron Kohavi is the Director of Data Mining at Blue Martini Software. Prior to joining Blue Martini, he managed the MineSet project, Silicon Graphics' award-winning product for data mining and visualization. Kohavi received his Ph.D. in Machine Learning from Stanford University, where he led the MLC++ project, the Machine Learning library in C++. He received his BA from the Technion, Israel.

Carla E. Brodley is an Associate Professor in the School of Electrical and Computer Engineering at Purdue University. She received her bachelors degree from McGill University in 1985 and her PhD in computer science from the University of Massachusetts in 1994. Her research interests include machine learning, computer vision, and content-based image retrieval. She has applied techniques from these areas to problems from a variety of fields including remote sensing, medical images and computer security.

Brian Frasca is a Principal Engineer at Blue Martini Software. He received his M.S. from Stanford specializing in databases. He joined Blue Martini in April 1999 where he has designed the decision support data warehouse and has developed data mining transformations.

Llew Mason is the Manager of Data Mining Analytics at Blue Martini Software. He joined Blue Martini Software in September 1999 after completing his Ph.D. in Systems Engineering from the Australian National University, Canberra, Australia. His research interests include machine learning, computational learning theory, large margins analysis and methods for combining classifiers.

Zijian Zheng received his Ph.D. in computer science from the University of Sydney in 1996. He joined Blue Martini Software as a Senior Software Engineer in May 1999. He is an active researcher and developer in the areas of data mining and machine learning.

KDD-Cup 2000: Question 1

Winner's Report

Amdocs

Aron Inger, Nurit Vatnik, Saharon Rosset, Einat Neumann
{aroni, nuritv, saharonr, einatn} @Amdocs.com

1. THE KDD-CUP QUESTION

The training database contained 234,954 sessions, some complete and some clipped. We were asked to build a prediction model to describe which sessions are going to continue and which are about to terminate, and use this model to predict the outcomes on a test dataset (where the information on whether the sessions were clipped or not was withheld). The prediction problem was particularly difficult for several reasons, most notably:

1. The skewed class distribution (over 80% of the sessions were complete) together with the scarcity of powerful predictive attributes meant it was very difficult to improve over the "majority" model.
2. The test data was from a different period, so it was difficult to tell which of the patterns discovered in the training data would persist.

It should be noted, however, that both these issues are very relevant when dealing with real applications and handling them correctly may be an important ingredient in success.

2. METHODOLOGY AND ALGORITHMS

Our analysis approach is based on the combination of automated and manual analysis, integrating the power of the machine with expert knowledge and human insight. In this case, we can divide the analysis and modeling into alternating stages of manual and automated knowledge discovery. The main tools we used were the proprietary Amdocs Business Insight (BI) tool and SAS (mainly for data manipulation and simple reporting).

As with every analysis, the data must be firstly organized and understood (the data was generally clean so very little auditing was required). In this step we examined the data using SAS and the basic visualizations of the BI tool. The first major observation we made was that 60% of the data were one-click sessions. These were special since the one-click full sessions were mostly bots and crawlers, which always have a full session length of only one click, and are very easy to identify (they announced themselves). The remaining one-click sessions must still be identified as either clipped or not clipped. We made a decision at this point to separate the one-click session from all the rest, a decision that had proven itself to be very useful later. In this first stage we also investigated the need to generate additional transformations to those supplied in the aggregated data files. Except for a few gaps we used the aggregated data as given.

The next stage was to run the automated knowledge discovery algorithm of the BI tool on the data, and analyze its results. At Amdocs we place great emphasis on the algorithms being "white box" and giving interpretable, understandable results. The user can try to understand and interpret the rules by changing them, adding/deleting conditions and creating new rules based on the insights generated from these rules. This also helps the user identify biases and problems in the data. Analyzing the knowledge discovery results gave us a few key insights about the data and about the prediction model we should be building.

For the one-click sessions, where 87% of the sessions were complete, there were almost no patterns describing sessions that are about to continue. With the exception of five small segments (see Figure 1) for which we were able to predict continuance, we concluded that all other one-click sessions should be automatically classified as complete.

Most of the multi-click sessions (65%) were complete and produced many more useful patterns. We were able to identify an interesting set of patterns from hints in the output of the automated algorithms. Sessions whose referring site was "file://c:…" kept appearing in small but significant patterns. We identified these patterns as originating from internal testing scripts. Since there were a small number of scripts, with many sessions generated by each, we were able to identify with almost perfect certainty for each of these sessions whether it was complete or clipped. In this way about 1% of the data could be marked as 'continue' with near perfect accuracy. The discovery of these patterns illustrates the power of human-machine synergy.

3. THE FINAL MODEL

The final model was built by combining automatically generated prediction models with user knowledge acquired through the entire process, as described in Figure 1

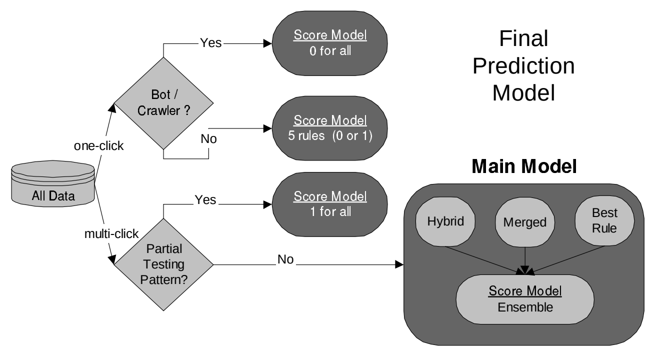


Figure 1 Final Prediction Model

For the one-click sessions, it involved separating the bots and then using a manually crafted set of rules on the remainder of the sessions, for predicting "continue" (e.g.: all identified customers who had personal details). For the multi-click data we manually modeled the testing patterns discussed above, and on the remaining data, built an ensemble prediction model based on three different models. The three models all use a decision tree as their basis. "Best Rule" model chooses the most accurate rule satisfied by each record (and gives the expected accuracy of this rule as a score). "Merged" model performs a logistic regression algorithm on the rule-set, and defines the score as a weighted combination of the rules a record satisfies. "Hybrid" model is similar to the "Merged" model, but also uses raw field values in the logistic regression. The three scores are averaged to create a more robust final score.

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.

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.

KDD-Cup 2000: Question 4

Winner's Report

e-steam, inc.

Rafal Kustra, Jorge A. Picazo, Bogdan E. Popescu

1. PROBLEM DESCRIPTION

The KDD-Cup 2000 domain of analysis consisted of clickstream and purchase data from Gazelle.com, a legwear and legcare retailer. Question 4 was related to characterization of killer pages. In other words, given a clickstream, the task consisted in characterizing under which conditions a visitor was more likely to leave the site.

1.1 Some caveats

The knowledge discovery task for this question was particularly difficult. The reason is that one of the datasets provided to perform such task (*the aggregated dataset*) was constructed based on a random clipping mechanism. This clipping mechanism introduced some bias; bias that if ignored, would lead to wrong and misleading conclusions [1]. Therefore, it was imperative to make adjustments for eliminating such bias.

2. SOLUTION AND METHODS

e-steam's winning submission was a result of three important factors: Building of new features that take into account transactional information; adoption of a weighting scheme that corrected the bias induced by the way the aggregated dataset was generated, and proprietary prediction and visualization technology.

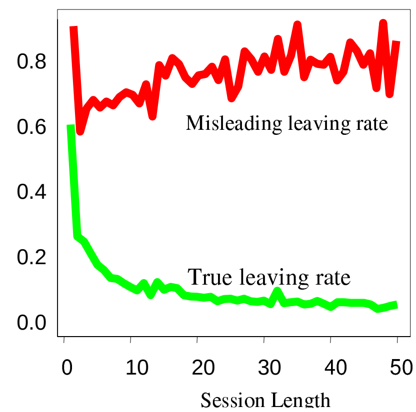
2.1 Building of new features

The participants of this contest were given two datasets. An aggregated one that contained information already prepared and apparently ready for datamining algorithms (we mention "apparently" because bias removal is necessary before this dataset can be analyzed as will be seen later) and a transactional dataset that contained the full clickstream history, but was not at a mining level. However, we were able to use additional information in this later dataset to enrich the aggregated one. This was done by building new features that took into consideration some of the dynamics of the clickstream process, and that undoubtedly helped to have a better understanding of the "leaving" behavior.

Some of such features were: "Last minus median processing time" (which is the processing time of the last "observed" page minus the median of the processing times of the pages generated up to the last "observed" one), and "Number of product views seen in the last 5 pages" among others.

2.2 Bias removal

As mentioned earlier, the aggregated dataset for performing the knowledge discovery task asked in this question had some caveats. Since the clickstream sessions were artificially clipped to make the problem suitable for the classification task for the KDD-Cup question 1, some bias was introduced. To correct for this, it was imperative that the data generation process was taken into account. We have approached this problem from survival analysis point of view and introduced weights to account for the biased distribution [1]. An alternative adjustment consists in taking into account only the unclipped sessions. Regardless of the classifier used, one of these adjustments is absolutely necessary in order to obtain correct insights about the conditions under which users are more likely to leave Gazelle.com Web site. The adjustment is far more important than only for academic curiosity: many trends obtained without this correction are actually a reverse of the true ones that manifest themselves after the adjustment is done. One important example of it was the number of views seen so far: before the correction, the inference seemed to indicate that the more views the user saw so far, the more likely was he or she to leave the site, where in fact the opposite is true: to certain degree at least, users get "hooked" if they stay on beyond the home page, and the page after, and tend to explore the site further. Moreover, the difference is even bigger from the quantitative point of view.



2.3 Prediction and visualization technology

e-steam owns proprietary technology for data mining prediction problems, including regression and classification tasks. The theoretical underpinnings of our technology are related to boosting proposals in the literature. One of the important aspects of our data mining technology, besides its scalability, accuracy, speed and robustness in dealing with real data, was the ability to interpret the resulting model ("*Achieving accuracy without compromising interpretability*" e-steam benchmarking white paper in progress). The highly transparent nature of our predictive technology was crucial in discovering important trends that formed the basis of our business report. Using our boosting technology together with classical statistical tools such as logistic regression, coupled with the bias removal through the weighting adjustment, we have built a predictive model to classify clickstreams as full or clipped, based on the cumulative characteristics of the Web session, user information (if available) and some key session attributes derived by us. These attributes were designed to improve the classification accuracy of the model and were validated through the modeling process. The model-building, always an iterative process, was greatly helped with our standard interpretation tools which show the importance of the attributes, their relation to the class probabilities, 3D displays showing the co-dependence of class structure on pair of attributes and many other informative displays.

References

[1] Picazo, J. Popescu, B. and Kustra, R. (2000) KDD-Cup 2000 Paradox: Discovering the right knowledge. *White e-steam paper sent to KDD-Cup 2000 organizers.*

KDD-Cup 2000: Question 5

Winner's Report

Amdocs

Einat Neumann, Nurit Vatnik, Saharon Rosset
Miri Duenias, Isabel Sassoon, Aron Inger
{einatn, nuritv, saharonr, mirid, isabels, aroni} @Amdocs.com

1. THE KDD-CUP QUESTION

Based on data of 234,954 sessions our task was to characterize which product brand visitors will view in the remainder of the session, if any.

2. METHODOLOGY AND ALGORITHMS

This is a knowledge discovery problem and it was important to use techniques that reveal highly interpretable phenomena. The primary tool we used is the Amdocs Business Insight (BI) Tool. It supports both Knowledge Discovery and Predictive Modeling where for this task the former was much more prominent. In addition we used the Splus tool for Exploratory Data Analysis (EDA) purposes.

Following is a discussion of the analysis and results. It is divided into steps, although the actual analysis was an iterative process.

2.1 Preprocessing and the Preliminary Step

The data was generally clean so little auditing was required. We used the aggregated data and performed only a few basic transformations (such as date transformations).

While working on Questions 1 and 4 ("Which sessions will continue and which will terminate?") we discovered that bots and crawler agent sessions always terminate after the first click. Needless to say that these sessions do not view any brand. Thus, the preliminary step was to exclude these sessions from the study. From here on, our analysis is presented given that the session is conducted by a legitimate browser (not a bot or crawler).

2.2 Initial Approach

Our initial approach was to conduct three two-class analyses using a Classification Tree, where we compared each brand-class to the non-brand class separately. We built the trees and then conducted Rule Induction, converting them into Classification Rules, using the BI tool. Classification Tree and Rule Induction are both White-Box methods, which we find very useful. The results generated by the tool are interpretable and intuitive, presented to the user via an interactive GUI and providing him/her with a clear insight about the problem.

Browsing the three two-class analyses results proved effective and helped us discover that most non-brand rules were similar. By reviewing the rules produced by our tool, we discovered that essential differences lie between brand-free and brand-prone sessions. These differences overwhelm those between the specific brands. Based on these findings, we decided to separate the study into two steps. In the first step we differentiate between sessions in which none of the specific brands were viewed and those in which at least one was viewed – a two-class analysis. In the second step, given a surfer viewed one of the three product brands, we differentiate between sessions according to which of the three is viewed – a three-class analysis. The general guidelines for discriminating between sessions are illustrated in Figure 1.

2.3 Two Step Analysis

2.3.1 Step I: Who will view a top brand?

The top brands were viewed in only 11,531 sessions (~5%). Different views and templates can be used to distinguish the brand-prone sessions from the brand-less ones. Many indicators are based on templates or product categories viewed along the session. In addition, we can characterize brand-hunters by looking at factors regarding the referring site, the session's length and loyalty measures. Thus, both session level and user level factors take part in predicting one's browsing course

2.3.2 Step II: Which brand will be viewed?

Among the sessions where a top brand was viewed, ~9% included multi-brand viewing. Since the aim was to distinguish between the brands, these sessions were excluded from this analysis step. It could have been interesting and informative to examine these sessions separately, but we didn't due to lack of time.

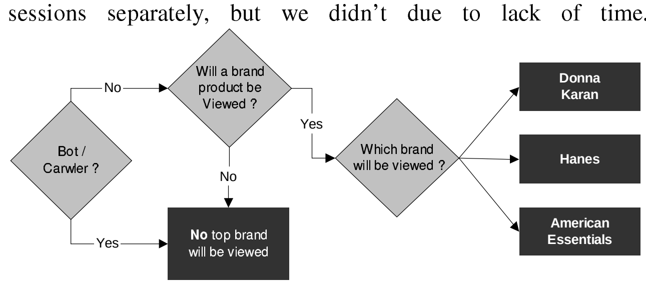


Figure 2. Discriminating between sessions

The brands that were analyzed are the following: Donna Karan (DK), Hanes (H) and American Essentials (AE). Given that only one of the brands was viewed: 40% viewed H, 36% DK and 24% AE. Overall we found that a session's affinity to a certain brand is highly unique. For example, browsing men's product templates is a strong indicator for an AE product view, while pantyhose products indicate the opposite.

2.4 Further Exploration

When using the Classification Tree algorithm we faced the difficulty in exposing time dependent patterns. Such patterns were discovered by the aid of Exploratory Data Analysis (EDA) techniques and basic graphical analysis, rather than data-mining algorithms. We detected significant behavior changes over time in the data. For instance, DK products weren't viewed at all up until February 26th (it turns out DK was not a part of the site until then). This type of exploration enabled us not only to expose time related patterns of brand viewing but also to associate the extent of viewing brands to specific events (such as the Ally Mcbeal ad). It also allowed us to examine the purchases (using Question 3 data) in the context of brand-viewing and vice versa.