Best Practice

Jesus Mena

CEO of WebMiner Inc. and author of *Data Mining Your Website*, is a noted expert on artificial intelligence for e-business applications and developer of a data-driven e-CRM ASP.

Dr Raymond Pettit

is president of ERP Associates, providing consulting in market research/e-research solutions, CRM/e-CRM analytic integration and statistical training and education. He is on the faculty of Baruch College.

Web-mining case study: An Internet radio website

Jesus Mena and Raymond Pettit

Received: 21 February 2001

It is of the essence of a good case study that it should include an evaluation of activity being studied. The following case is, however, so recent that it is premature to provide such an assessment. We thought the material interesting enough in its own right to print as it stands; the authors have promised to provide a second bite at the cherry in the shape of an evaluation when that can be done — perhaps in time for our next issue, or failing that the next but one.

Abstract

Data mining is the key to customer knowledge, intimacy and profitability. Today, businesses are overwhelmed with data, prompting the thoughtful and consistent use of analytic techniques that can enable understanding and knowledge of the customer in ways that can maximise relationships and the opportunity for profit.

This case study concerned an Internet radio station striving to become more intelligent about their visitors. The exploratory phase combined data-mining analytic techniques and appended third-party demographic and lifestyle information. Key segments of the visitor population could then be isolated by infusing marketing logic and rules that emerged from the results of the analysis. Visitors were identified by heavy listeners; music preferences; and non-listeners.

In the final analysis, the data-mining process informed the radio station that traditional business principles apply to the Internet, and revealed the importance of defining clear business objectives, the critical need for quality data and data preparation, and that long-term survival is dependent on integrating data-driven results with business logic and experience.

Nothing new

Data mining is the key to customer knowledge, intimacy and profitability. For years marketers have been using databases to construct a picture of their customers and have turned to data-mining tools and associated techniques to extract the most valuable information possible to drive marketing decisions. Artificial intelligence (AI) technology (in the form of data mining) has been used to:

- stop ‘churn’ (or customer attrition) by cellular phone companies
- manage financial and risk portfolios by financial services firms
- detect fraud and set pricing by credit card companies

Keywords: personalisation, customer relationship management (CRM), e-business, e-commerce, Internet marketing (e-marketing), data mining

Jesus Mena
Lumina Americas,
EVP, WebMiner,
670 Broadway, Suite 2,
New York, NY 10012,
USA
Tel: +1 212 813 6761
Fax: +1 212 253 9799

- lift response rates by mail cataloguers
- analyse ‘shopping carts’ (market basket analysis) by retailers.

In today’s hyper-competitive marketplace, the creative and strategic use of customer information has become critical to survival. And nowhere is more ‘data’ accumulated than by e-retailers, e-businesses and e-commerce solutions. Remarkably, most of this potential has yet to be realised.

To foster long-term e-customer relationships and repeated online sales, more is needed than the tracking of clickstream behaviour. Yet most solutions and services offered by today’s ad networks, collaborative filtering tools and e-commerce (e-CRM) systems are limited to this very small piece of the total picture. It is as if you, as an online store manager, have to watch customers through a one-way mirror as they wander around your store, *and* you — without any interaction or relationship generation of any kind — must infer who these ‘visitors’ are, where their interests lie and what they eventually may purchase.

This flimsy premise is what most ad networks and e-commerce systems are offering with the tracking of cookies and meta-tags to initiate cross- and up-selling efforts online. In truth, this sort of tracking may help some e-retailers sell low-hanging fruit like CDs or books, but it will not suffice in establishing profitable and long-term customer relationships or in selling big-ticket items like sports utility vehicles or expensive vacation packages.

Customer retention, relationships and profitability

Recent studies have shown that e-retailers, on average, spend US\$100–250 acquiring a new customer. This new acquisition then spends about \$24 and most likely never returns. On average, only 35 per cent of online buyers make a second purchase from the same site. Given these dismal statistics, it is apparent that to survive and be profitable all businesses working the e-channel — online pure plays and offline established bricks and mortars — would benefit by learning what their online customers are like *offline*.

Fostering repeated sales (usage or visits for content providers) requires knowledge about customers’ preferences, rates of consumption, behaviour, attitudes and lifestyle. This generally requires knowing such things as a customer’s age, income, values, preferred lifestyle and stage in life that can allow, for example, a financial site to ‘know’ whether to offer an auto loan, a credit card or a savings account. Thus, a visitor from a zip code with the demographics associated with ‘college campus’ is a prime candidate for a credit card, but not a home loan. This is more than cross-selling or up-selling; this is knowing who your customers are when they enter the online store or environment, anticipating their needs, and providing solutions that make sense to them and can encourage a long-term profitable relationship.

Imagine data mining a website database that included clickstream and server-side (behavioural) information, client-supplied transaction or usage information (stripped of personal identifying information), original

Required knowledge

customer survey results, and appended third-party demographic, psychographic, lifestyle, financial and/or cognographic data. Now, when a visitor enters your online store, you know who they are, where they are from, how much they earn, how many kids they have, what kind of car they may own, and whether they prefer to vacation in Atlantic City or the south of France. Thus you are now in a superior position to offer an appropriate product, service or content piece that makes sense to the customer and allows you to build a bigger share of their value over a longer period of time.

Who listens to music online?

This case study involves a project with an Internet radio station focused on a major ethnic segment in the USA. Specifically, it describes efforts to identify what segments of the 'listening' population spend the most time at the website and to what channels they attend. The site is available in two languages and has 18 distinct channels that offer a variety of music styles. Advertising banners and links surround a surprisingly full Web page of information on music, relevant CDs, and such things as contests, news and weather reports. Revenue for the station is generated by banners, a share of sales of CDs, media players and other related products (terms negotiated separately with each company), and through some online community (or interest-oriented) outlets located elsewhere on the site. Listening to music, however, is free, although limited to the selections offered through one of the 18 listening channels.

Initially, the client was interested in addressing the following business questions.

Business questions

- What are our most popular channels?
- Do different segments of listeners attend to different channels?
- What segments of visitors are most 'sticky'; that is, who comes to the site and remains the longest amount of time?

To answer the first question, a simple frequency analysis was proposed, making use of a measurement called 'session viewing time' (although in this case it should be termed 'session listening time'). This is available from the user log files and easily obtainable via tools available in virtually every Web server/database system. It basically indicates, in minutes, the amount of time each Web page was 'viewed'. The most popular channels, then, would have the highest aggregate (and average) viewing times over the time period the study covered.

The efforts needed to answer the remaining two questions required a bit more analytic work and planning. Indeed, the answers can be found through the use of data-mining algorithms and techniques that can handle a large amount of data to produce identifiable segments. In this case, the author's company suggested the use of a classification tree algorithm in order to differentiate segments and provide an indication of which segments had the 'highest' listening times.

Ultimately, these results would be used to target and market better to different segments of the site's visitors. Personalised ads, content and

even channel choices could be dynamically delivered, via the Web, to different segments (visitors). This particular Web solution process allows the identification of returning visitors via 'cookie' technology.

Alternatively, dynamically delivered content can be triggered by having the visitor input the zip code (postal code) of the area in which they live. The impetus to 'give up' your zip code is determined individually by each client, but can be done in exchange for weather information, entry to some form of lottery, or relevant information about, in this case, popular music.

The final goal of the entire exercise is to increase 'stickiness'; that is, to encourage more people to spend more time on the site (thus increasing the likelihood of revenue-generating activities, such as buying a CD or clicking on a banner of interest). By better 'knowing' who the visitor is, the Internet radio station hoped to be 'smarter' in providing relevant content and choices that would encourage 'stickiness'.

Exploratory analysis

As with any Web-mining project, the first step was to request adequate data to explore in preparation for analysis. In this case, the author's company asked for two weeks' worth of the most current unique session data, to include:

- zip code
- session time
- channel identification number for each session.

Data preparation and exploration constitute the major portion of any data- or web-mining project. Unfortunately, data collection in many locations is sloppy, incomplete, or based on data warehouse needs rather than analytic needs. In this case, the company was fortunate to receive rather quickly the requested data in a format that could be easily accessed.

Exploratory analysis was carried out using SPSS software to gain an understanding of the data distribution, spot any globally detectable incidents, and check for possible mistakes or inaccuracies in the data. Work began with a standard flat file of approximately 59,000 'sessions'. If you could imagine an Excel (or any) spreadsheet, this would be equivalent to 59,000 rows. A look at the data showed that 956 individual visitors generated these 59,000 'sessions'. Some were visitors only once, while others returned multiple times, or clicked on a new radio 'channel' while on the site, thus initiating a new listening session.

A look at a breakdown of the zip codes showed a very typical random pattern. It was thus concluded that no problem existed concerning the likelihood of a 'biased' sample in this regard. If a biased sample were found, it would indicate that the models would be biased as well, and would not deliver as accurate a picture as possible to the client.

A frequency breakdown of the music channels revealed an interesting pattern. The most frequent channel 'choice' was '0', which meant 'no choice'. Apparently, upon arrival to the site, music was launched (no choice needed to be made). Thus the highest proportion of all visitors

**Knowing the visitor =
smarter content**

studied merely went about their business on the site while the music was playing! It was not clear to the researchers (and no explanation was available from the client) how to determine what was playing, so the company was forced to treat this 'choice' as a separate category for this round.

As would be expected, only a handful of stations (channels) stood out as preferred by listeners (or as choices they made to access a channel). The 'no choice' category and three other stations accounted for about 80 per cent of all choices. In addition to the obvious glitch with the 'no choice' choice (which it was suggested should be dealt with by the website), the implications for data mining are that these three stations will dominate the analysis. Put another way, the best predictor of listening time will be one of these three channels.

Data distribution

A slightly more difficult problem was highlighted upon studying the data distribution of listening sessions (time in minutes). A highly disproportionate part of the total sample listened from zero to eight minutes (zero being less than a minute, and eight minutes approximately one or two songs). In fact, of the 59,000 individual sessions tracked, 99 per cent lasted less than 120 minutes. To make matters worse, of the remaining 1 per cent (590 or so sessions), 40 were extremely high (up to 21 *hours*).

While it may be interesting to speculate that these 'extreme' folks are a highly sought-after target segment, their existence wreaks havoc with any sort of statistical procedure. In addition, their listening times may be based on the fact that they open the radio station and just leave it on all day while they do other things online. The reasoning becomes then that these are probably not valuable customers since they are not even looking at the site; they merely have the music playing and the website minimised on their desktop.

A number of classification attempts were made using a variety of algorithms with the data, including the so-called 'heavy listeners'. These proved to be extremely disappointing, as the extremely large discrepancy was pulling the model apart and not allowing a very useful segmentation to take place. As a result, a determination with the client to exclude any session more than two hours in length was made. The two hours cut-off point was not a magic number, just a reasonable one the client felt it was possible to justify. Remembering their initial objective was to focus on keeping people on their site longer, not to have the site minimised and out of the way, the client agreed to have the researchers look at a slightly reduced dataset (59,000–590 $\hat{=}$ 58,410 sessions).

A final part of the analysis was to append so-called 'lifestyle' data to the client-supplied data to describe and understand segments better. In the USA, this sort of demographic, behavioural and attitudinal data are available and can be appended to research data collected when matched by a zip code (postal code) or an actual physical address. In this case, a highly regarded lifestyle cluster database was used and appended to the session database. Thus the researchers were able to run a classification tree algorithm and study the segments that resulted informed by the lifestyle 'picture' obtained from 40 distinct descriptive clusters contained in the third-party database.

Analysis of 'stickiness'

Classification tree procedures were run with 'number of minutes' as the dependent variable and the lifestyle variables and 'channel' as the independent variables. Some interesting patterns emerged:

Patterns

- the segments with the highest 'listening times' were south-western US college students, East Coast immigrant families, and south-western ethnics
- the 'rock' station was highly preferred by these segments (based on amount of time listened)
- the second most-listened station varied for the three segments
- a similar pattern held for other segments, although not all segments showed the significant lift demonstrated with the 'heavy listeners'
- certain segments, such as 'highly affluent East or West Coasters' were not representative of a significant target market for this particular online offering.

Thus it was concluded that — based on the results of the classification tree analysis — the researchers could fashion unique scoring 'rules' that would maximise the impact of offering relevant banners (online advertisements) and suggesting/promoting certain listening channels over others that would most likely prompt a response. These 'rules' could be tied to the zip code of a visitor and 'triggered' when that information was supplied.

Finally, the company sought to balance — with the client's intense participation — the data-driven rules and the business/marketing decisions that needed to be made. Without going into too much detail, this is a key part of arriving at the optimum solution — turning the results of analysis into workable procedures and tangible delivery of content product, or sales opportunities. As such, this could make up another whole case study!

Findings, recommendations and conclusion

Unfortunately, this case study is so current that the company does not have a report on how successful these efforts have been. Currently, it is in the 'rules' stage and in collaboration with the client to produce the offerings that make most sense to drive stickiness to (now) identified visitors.

The results of the classification models, however, are clear: lift (or improvement) in amount of listening time is evident in more than half of the identified clusters in the tree segments. Thus the concerted focus on studying and understanding these clusters from both the data and the marketing side is the essence of the work to be done. Evaluation of success can be measured from the benchmarks established in this analysis. Regular tracking of the changes in the amount of listening time across segments becomes a vital component of this ongoing process.

As mentioned earlier, the author's company recommended to the client to avoid the so-called 'no choice' option, or at the very least find a way to 'identify' what type of music was playing for future analysis. Even at this

early stage of the solution process, it was felt that the initial station 'choice' should be initiated by a pop-up window on the website *for first-time visitors* to offer the choice of music they are most likely to prefer, based on their data-driven profile. Once done, that visitor would be 'cookied' (cyber-identified) and not have to re-enter a zip code on every return visit.

Overall, this experience with the Internet radio station confirmed many things the company has dealt with in the past on projects of this nature.

- The business problem and objective must be clearly understood and stated up front. Web/data mining is not magic — it works best when solid goals and objectives are in place.
- Data quality and preparation remain the most important part of the total solution process. If the data are of poor quality and/or are not explored and prepared adequately, precious few times will data mining result in a useful solution.
- The balance between data-driven and business rules is an extremely important — yet oft neglected — part of the whole solution process. Software solutions tend to encourage one or the other; yet both voices are necessary to arrive at a sensible and useful solution. It is still the case that many important things in the Internet and e-commerce arena cannot be automated, or at least not automated well.
- Most of the problems faced by online retailers, e-services and content sites can be helped by the use of thoughtful implementation of data/ Web-mining techniques, coupled with a tie-down to reality by working closely with the client's own business specialists and experts.
- As with so many facets of business nowadays, the Internet space has become an extraordinary place for learning to occur, as well as an opportunity for creative ideas, solutions and services to grow and thrive.

Data mining your website involves an extraordinary effort to integrate seemingly disparate sources of customer knowledge (e-customer knowledge) to drive improvements and profitability. However, the imperative is clear: it must be done to ensure any kind of e-success as the online sales and marketing channel strives to mature and define itself.