E-Retail Example- A Web-Mining Scenario Using CRISP-DM[1]

Business Understanding

Determine Business Objectives

As more companies make the transition to selling over the Web, an established computer/electronics e-retailer is facing increasing competition from newer sites. Faced with the reality that Web stores are cropping up as fast (or faster!) than customers are migrating to the Web, the company must find ways to remain profitable despite the rising costs of customer acquisition. One proposed solution is to cultivate existing customer relationships in order to maximize the value of each of the company's current customers.

Thus, a study is commissioned with the following objectives:

- Improve cross-sales by making better recommendations.
- Increase customer loyalty with a more personalized service.

Tentatively, the study will be judged a success if:

- Cross-sales increase by 10%.
- Customers spend more time and see more pages on the site per visit.
- The study finishes on time and under budget.

[1] Source: https://www.ibm.com/docs/en/spss-modeler/saas?topic=guide-introduction-crisp-dm

Assess Situation

This is the electronics e-retailer's first attempt at Web mining, and the company has decided to consult a data mining specialist to help in getting started. One of the first tasks the consultant faces is to assess the company's resources for data mining.

- Personnel. It's clear that there is in-house expertise with managing server logs and product and purchase databases, but little experience in data warehousing and data cleaning for analysis. Thus, a database specialist may also be consulted. Since the company hopes the results of the study will become part of a continuing Web-mining process, management must also consider whether any positions created during the current effort will be permanent ones.
- Data. Since this is an established company, there is plenty of Web log and purchase data to draw from. In fact, for this initial study, the company will restrict the analysis to customers who have "registered" on the site. If successful, the program can be expanded.
- Risks. Aside from the monetary outlays for the consultants and the time spent by employees on the study, there is not a great deal of immediate risk in this venture. However, time is always important, so this initial project is scheduled for a single financial quarter.
- Also, there is not a lot of extra cash flow at the moment, so it is imperative that the study come in under budget. If either of these goals should be in danger, the business managers have suggested that the project's scope should be reduced.

Determine Data Mining Goals

With the help of its data mining consultant, the e-retailer has been able to translate the company's business objectives into data mining terms. The goals for the initial study to be completed this quarter are:

- Use historical information about previous purchases to generate a model that links "related" items. When users look at an item description, provide links to other items in the related group (market basket analysis).
- Use Web logs to determine what different customers are trying to find, and then redesign the site to highlight these items. Each different customer "type" will see a different main page for the site (profiling).
- Use Web logs to try to predict where a person is going next, given where he or she came from and has been on your site (sequence analysis).

Produce Project Plan

The overview plan for the study is as shown in the table below.

Table 1. Sample project plan overview

Phase	Time	Resources	Risks
Business understanding	1 week	All analysts	Economic change
Data understanding	3 weeks	All analysts	Data problems, technology problems
Data preparation	5 weeks	Data mining consultant, some database analyst time	Data problems, technology problems
Modeling	2 weeks	Data mining consultant, some database analyst time	Technology problems, inability to find adequate model
Evaluation	1 week	All analysts	Economic change, inability to implement results
Deployment	1 week	Data mining consultant, some database analyst time	Economic change, inability to implement results

Data Understanding

Collect Initial Data

The e-retailer in this example uses several important data sources, including:

- Web logs. The raw access logs contain all of the information on how customers navigate the Web site. References to image files and other non-informative entries in the Web logs will need to be removed as part of the data preparation process.
- Purchase data. When a customer submits an order, all of the information pertinent to that order is saved. The orders in the purchase database need to be mapped to the corresponding sessions in the Web logs.
- Product database. The product attributes may be useful when determining "related" products. The product information needs to be mapped to the corresponding orders.
- Customer database. This database contains extra information collected from registered customers. The records are by no means complete, because many customers do not fill out questionnaires. The customer information needs to be mapped to the corresponding purchases and sessions in the Web logs.

At this moment, the company has no plans to purchase external databases or spend money conducting surveys because its analysts are busy managing the data they currently have. At some point, however, they may want to consider an extended deployment of data mining results, in which case purchasing additional demographic data for unregistered customers may be quite useful. It may also be useful to have demographic information to see how the e-retailer's customer base differs from the average Web shopper.

Describe Data

There are many records and attributes to process in a Web-mining application. Even though the e-retailer conducting this data mining project has limited the initial study to the approximately 30,000 customers who have registered on the site, there are still millions of records in the Web logs.

Most of the value types in these data sources are symbolic, whether they are dates and times, Web pages accessed, or answers to multiple-choice questions from the registration questionnaire. Some of these variables will be used to create new variables that are numeric, such as number of Web pages visited and time spent at the Web site. The few existing numeric variables in the data sources include the number of each product ordered, the amount spent during a purchase, and product weight and dimension specifications from the product database.

There is little overlap in the coding schemes for the various data sources because the data sources contain very different attributes. The only variables that overlap are "keys," such as the customer IDs and product codes. These variables must have identical coding schemes from data source to data source; otherwise, it would be impossible to merge the data sources. Some additional data preparation will be necessary to recode these key fields for merging.

Explore Data

Although CRISP-DM suggests conducting an initial exploration at this point, data exploration is difficult, if not impossible, on raw Web logs, as our e-retailer has found out. Typically, Web log data must be processed first in the data preparation phase to produce data that can be meaningfully explored. This departure from CRISP-DM underscores the fact that the process can and should be customized for your particular data mining needs. CRISP-DM is cyclical, and data miners typically move back and forth between phases.

Although Web logs must be processed before exploration, the other data sources available to the e-retailer are more amenable to exploration. Using the purchase database for exploration reveals interesting summaries about customers, such as how much they spend, how many items they buy per purchase, and where they come from. Summaries of the customer database will show the distribution of responses to the items on the registration questionnaire.

Exploration is also useful for looking for errors in the data. While most of the data sources are automatically generated, information in the product database was entered by hand. Some quick summaries of listed product dimensions will help to discover typos, such as "119-inch" (instead of "19-inch") monitor.

Verify Data Quality

The verification of data quality is often accomplished during the course of the description and exploration processes. Some of the issues encountered by the e-retailer include:

Missing Data. The known missing data includes the unanswered questionnaires by some of the registered users. Without the extra information provided by the questionnaire, these customers may have to be left out of some of the subsequent models.

Data Errors. Most of the data sources are automatically generated, so this is not a great worry. Typographical errors in the product database can be found during the exploration process.

Measurement Errors. The greatest potential source for measurement error is the questionnaire. If any of the items are ill-advised or poorly worded, they may not provide the information the e-retailer hopes to obtain. Again, during the exploration process, it is important to pay special attention to items that have an unusual distribution of answers.

Data Preparation

Select Data

Many of the e-retailer's decisions about which data to select have already been made in earlier phases of the data mining process.

- Selecting items. The initial study will be limited to the (approximately) 30,000 customers who have registered on the site, so filters need to be set up to exclude purchases and Web logs of nonregistered customers. Other filters should be established to remove calls to image files and other non-informative entries in the Web logs.
- Selecting attributes. The purchase database will contain sensitive information about the e-retailer's customers, so it is important to filter attributes such as the customer name, address, phone number, and credit card numbers.

Clean Data

The e-retailer uses the data cleaning process to address the problems noted in the data quality report.

- Missing data. Customers who did not complete the online questionnaire may have to be left out of some of the models later on. These customers could be asked again to fill out the questionnaire, but this will take time and money that the e-retailer cannot afford to spend. What the e-retailer can do is model the purchasing differences between customers who do and do not answer the questionnaire. If these two sets of customers have similar purchasing habits, the missing questionnaires are less worrisome.
- Data errors. Errors found during the exploration process can be corrected here. For the most part, though, proper data entry is enforced on the Web site before a customer submits a page to the back-end database.
- Measurement errors. Poorly worded items on the questionnaire can greatly affect the quality of the data. As with missing questionnaires, this is a difficult problem because there may not be time or money available to collect answers to a new replacement question. For problematic items, the best solution may be to go back to the selection process and filter these items from further analyses.

Construct Data

The processing of Web logs can create many new attributes. For the events recorded in the logs, the e-retailer will want to create timestamps, identify visitors and sessions, and note the page accessed and the type of activity the event represents. Some of these variables will be used to create more attributes, such as the time between events within a session.

Further attributes can be created as a result of a merge or other data restructuring. For example, when the event-per-row Web logs are "rolled up" so that each row is a session, new attributes recording the total number of actions, total time spent, and total purchases made during the session will be created. When the Web logs are merged with the customer database so that each row is a customer, new attributes recording the number of sessions, total number of actions, total time spent, and total purchases made by each customer will be created.

After constructing new data, the e-retailer goes through an exploration process to make sure that the data creation was performed correctly.

Integrate Data

With multiple data sources, there are many different ways in which the e-retailer can integrate data:

- Adding customer and product attributes to event data. In order to model Web log events using attributes from other databases, any customer ID, product number, and purchase order number associated with each event must be correctly identified and the corresponding attributes merged to the processed Web logs. Note that the merged file replicates customer and product information every time a customer or product is associated with an event.
- Adding purchase and Web log information to customer data. In order to model the value of a customer, their purchases and session information must be picked out of the appropriate databases, totaled, and merged with the customer database. This involves the creation of new attributes as discussed in the constructing data process.

After integrating databases, the e-retailer goes through an exploration process to make sure that the data merge was performed correctly.

Format Data

This case does not require any particular data formatting, because standard data mining algorithms will be used.

Modeling

Select Modeling Technique

The modeling techniques employed by the e-retailer are driven by the company's data mining goals:

- Improved recommendations. At its simplest, this involves clustering purchase orders to determine which products are most often bought together. Customer data, and even visit records, can be added for richer results. The two-step or Kohonen network clustering techniques are suited for this type of modeling. Afterward, the clusters can be profiled using a C5.0 ruleset to determine which recommendations are most appropriate at any point during a customer's visit.
- Improved site navigation. For now, the e-retailer will focus on identifying pages that are often used but require several clicks for the user to find. This entails applying a sequencing algorithm to the Web logs in order to generate the "unique paths" customers take through the Web site, and then specifically looking for sessions that have a lot of page visits without (or before) an action taken. Later, in a more in-depth analysis, clustering techniques can be used to identify different "types" of visits and visitors, and the site content can be organized and presented according to type.

Generate Test Designs

The criteria by which the models are assessed depend on the models under consideration and the data mining goals:

- Improved recommendations. Until the improved recommendations are presented to live customers, there is no purely objective way to assess them. However, the e-retailer may require the rules that generate the recommendations to be simple enough to make sense from a business perspective. Likewise, the rules should be complex enough to generate different recommendations for different customers and sessions.
- Improved site navigation. Given the evidence of what pages customers access on the Web site, the e-retailer can objectively assess the updated site design in terms of ease of access to important pages. However, as with the recommendations, it is difficult to assess in advance how well customers will adjust to the reorganized site. If time and finances allow, some usability testing may be in order.

Build Model

- Improved recommendations. Clusterings are produced for varying levels of data integration, starting with just the purchase database and then including related customer and session information. For each level of integration, clusterings are produced under varying parameter settings for the two-step and Kohonen network algorithms. For each of these clusterings, a few C5.0 rulesets are generated with different parameter settings.
- Improved site navigation. The Sequence modeling node is used to generate customer paths. The algorithm allows the specification of a minimum support criterion, which is useful for focusing on the most common customer paths. Various settings for the parameters are tried.

Assess Model

- Improved recommendations. One of the Kohonen networks and a two-step clustering each give reasonable results, and the e-retailer finds it difficult to choose between them. In time, the company hopes to use both, accepting the recommendations that the two techniques agree on and studying in greater detail the situations in which they differ. With a little effort and applied business knowledge, the e-retailer can develop further rules to resolve

differences between the two techniques.
The e-retailer also finds that the results that include the session information are surprisingly good. There is evidence to suggest that recommendations could be tied to site navigation. A ruleset, defining where the customer is likely to go next, could be used in real time to affect the site content directly as the customer is browsing.

- Improved site navigation. The Sequence model provides the e-retailer with a high level of confidence that certain customer paths can be predicted, producing results that suggest a manageable number of changes to the site design.

Evaluation

Evaluate Results

The overall results of the e-retailer's first experience with data mining are fairly easy to communicate from a business perspective: the study produced what are hoped to be better product recommendations and an improved site design. The improved site design is based on the customer browsing sequences, which show the site features that customers want but require several steps to reach. The evidence that the product recommendations are better is more difficult to convey, because the decision rules can become complicated. To produce the final report, the analysts will try to identify some general trends in the rulesets that can be more easily explained.

- Ranking the Models. Because several of the initial models seemed to make business sense, ranking within that group was based on statistical criteria, ease of interpretation, and diversity. Thus, the model gave different recommendations for different situations.
- New Questions. The most important question to come out of the study is, How can the e-retailer find out more about his or her customers? The information in the customer database plays an important role in forming the clusters for recommendations. While special rules are available for making recommendations to customers whose information is missing, the recommendations are more general in nature than those that can be made to registered customers.

Review Process

As a result of reviewing the process of the inital data mining project, the e-retailer has developed a greater appreciation of the interrelations between steps in the process. Initially reluctant to "backtrack" in the CRISP-DM process, the e-retailer now sees that the cyclic nature of the process increases its power. The process review has also led the e-retailer to understand that:

- A return to the exploration process is always warranted when something unusual appears in another phase of the CRISP-DM process.
- Data preparation, especially of Web logs, requires patience, since it can take a very long time.
- It is vital to stay focused on the business problem at hand, because once the data are ready for analysis, it's all too easy to start constructing models without regard to the bigger picture.
- Once the modeling phase is over, business understanding is even more important in deciding how to implement results and determine what further studies are warranted.

Deployment

Plan Deployment

A successful deployment of the e-retailer's data mining results requires that the right information reaches the right people.

- Decision makers. Decision makers need to be informed of the recommendations and proposed changes to the site, and provided with short explanations of how these changes will help. Assuming that they accept the results of the study, the people who will implement the changes need to be notified.
- Web developers. People who maintain the Web site will have to incorporate the new recommendations and organization of site content. Inform them of what changes could happen because of future studies, so they can lay the groundwork now. Getting the team prepared for on-the-fly site construction based upon real-time sequence analysis might be helpful later.
- Database experts. The people who maintain the customer, purchase, and product databases should be kept apprised of how the information from the databases is being used and what attributes may be added to the databases in future projects.

Above all, the project team needs to keep in touch with each of these groups to coordinate the deployment of results and planning for future projects.

Plan Monitoring and Maintenance

The immediate task for monitoring is to determine whether the new site organization and improved recommendations actually work. That is, are users able to take more direct routes to the pages that they're looking for? Have cross-sales of recommended items increased? After a few weeks of monitoring, the e-retailer will be able to determine the success of the study.

What can be handled automatically is the inclusion of new registered users. When customers register with the site, the current rulesets can be applied to their information to determine what recommendations they should be given.

Deciding when to update the rulesets for determining recommendations is a trickier task. Updating the rulesets is not an automatic process because cluster creation requires human input regarding the appropriateness of a given cluster solution.

As future projects generate more complex models, the need for and amount of monitoring will almost surely increase. When possible, the bulk of the monitoring should be automatic with regularly scheduled reports available for review. Alternatively, the creation of models that provide predictions on the fly may be a direction the company would like to take. This requires more sophistication from the team than the first data mining project.

Produce Final Report

The greatest deviation from the original project plan is also an interesting lead for further data mining work. The original plan called for finding out how to have customers spend more time and see more pages on the site per visit.

As it turns out, having a happy customer is not simply a matter of keeping them online. Frequency distributions of time spent per session, split on whether the session resulted in a purchase, found that the session times for most sessions resulting in purchases fall between the session times for two clusters of nonpurchase sessions.

Now that this is known, the issue is to find out whether these customers who spend a long time on the site without purchasing are just browsing or simply can't find what they're looking for. The next step is to find out how to deliver what they're looking for in order to encourage purchases.

Review Project

- Project member interviews. The e-retailer finds that project members most closely associated with the study from start to finish are for the most part enthusiastic about the results and look forward to future projects. The database group seems cautiously optimistic; while they appreciate the usefulness of the study, they point out the added burden on database resources. A consultant was available during the study, but going forward, another employee dedicated to database maintenance will be necessary as the scope of the project expands.
- Customer interviews. Customer feedback has been largely positive so far. One issue that was not well thought out was the impact of the site design change on established customers. After a few years, the registered customers developed certain expectations about how the site is organized. Feedback from registered users is not quite as positive as from nonregistered customers, and a few greatly dislike the changes. The e-retailer needs to stay aware of this issue and carefully consider whether a change will bring in enough new customers to risk losing existing ones.