Description and Documentation for the Cooperative Database Company Dataset

Version 1.0

By Richard J. Courtheoux, President,
Marketing Analysis Applications, Inc.

The Direct Marketing Educational Foundation has received a dataset for classroom teaching from a cooperative database company. A cooperative database takes transaction records from hundreds of companies and builds a household based view of purchasing over time among the contributing companies. This data is then made available for various marketing purposes. The participant companies benefited by:

- Being able to rent names for their targeted new customer acquisition programs. Because the cooperative database company has visibility to a household's purchases across many companies, very effective selections of names can be made.
- Overlaying information summaries or model scores from the database onto their own customer files in order to do a better job of targeting customer marketing communications. Having some form of information about what a household is buying in the marketplace improves a company's ability to predict the household's future purchasing behavior.
- Doing certain forms of more detailed market research to understand trends that are affecting their business. For example, companies can examine market trends by merchandise category in order to develop merchandising strategies.

A relatively small number of similar cooperative database companies provide these capabilities to direct marketers.

The company's core challenge is to find as many ways as possible to monetize the information in the database. There are substantial fixed costs associated with updating, maintaining and providing access to this large database, so additional sales volume tends to have high margin. Conversely, participating companies have a fixed cost associated with providing data to the cooperative database in order to gain rights to use it, so the more ways they can use the database the more leverage they obtain on their fixed costs. The larger the number of companies that derive a positive benefit from being in the cooperative data the larger and more valuable the database becomes.

The cooperative database company has to efficiently manage the large volume of data provided by the hundreds of participating companies. The entire database consists of nearly 500 million order records and the associated 1.2 billion merchandise line items records for those orders. Besides the data volume, there are several technical challenges associated with the nature of the data and its intended applications.

- The data are coming from hundreds of operational systems at the participant companies, each of which has its own record formats and coding schemes. The cooperative database company has to accept these records, audit them and reformat them into a common format for its database.

- Name and address matching needs to be performed to create a household centric view of customer purchasing. Rules were developed to decide when names and addresses are similar enough to regard them as coming from the same household. Mechanisms were also developed to handle household address changes.
- Each participant company categorizes merchandise in its own unique ways. For the product level information to be meaningful across companies the cooperative database company had to develop its own merchandise classification system and map each participant company's categories to it.
- In order to make so much detailed data usable, the cooperative database company had to develop robust statistical models. The single most important models predict the likelihood that a customer will respond to a prospecting (i.e., new customer acquisition) marketing contact.

The DMEF dataset consists of a small portion of the overall database. It was developed as follows:

- The participant companies that contributed data for all months in the period from **January 2005 through December 2007** were identified. While companies were both joining and leaving the cooperative database for various reasons during this 3 year period, there were **207** companies that were consistent participants throughout the period. There are many analytical complexities that arise when trying to work with data where the base of participating companies keeps changing, so for teaching purposes it seemed more appropriate to focus on only data for this set of consistent participants.
- All order records from January 2005 through December 2007 from the 207 companies were extracted. **13,382,011** order records were extracted.
- The **35,536,676** line item records associated with these orders were extracted.
- The extracted order and line item records came from **2,451,988** households. Records with the ZIPCode of each of those households were extracted.

To make the product information less abstract, reference tables with product area and major category descriptions are provided. For some purposes it may be useful to know which participant companies are part of the same corporation, so a reference table of those relationships is provided. More detailed information about the various files is provided below.

Using the extract data sample students could be asked to address such marketing issues as:

- Developing and comparing predictive models for housefile marketing selections. Using data from some of the larger companies in the extract it is possible to create predictive housefile models. Since there are many companies in the database students could be asked to compare models developed for multiple companies to see how similar or dissimilar they are in terms of their predictive variables and parameters. Comparisons could also be done between predictions generated from a Recency/Frequency/Monetary cell approach versus a multivariate statistical model.
- Building new customer acquisition selection systems. Students could be asked to focus on a particular participant company and find ways to use database information to predict which households will become new customers of the company. For example, students could be asked to use January 2005 through June 2007 data in order to identify new customers of a particular company during the July – December, 2007 period. When a successful predictive mechanism (model) is found for one company the students could be asked to test whether that same approach can be successfully applied to other companies.

- Identifying affinities among participant companies. That is, which companies appeal to many of the same households and are, therefore, possible competitors. Or, they may be different titles owned by the same parent company which is cross-marketing to customers of each title. Students could be asked to apply this information to the problem of finding households to target for new customer acquisition efforts.
- Tracking and comparing customer sales value over time. Students could be asked to isolate groups of company customers in Spring 2005 based on different characteristics (e.g., purchase channel, purchase amount) and see how the sales value to the company builds up over time. Comparisons could be made among the results for different companies in the database.
- Switching channel loyalty over time. Students could examine how initial channel of purchase is predictive of channel usage for subsequent purchases. The period 2005 – 2007 covered by the sales data was one in which the Internet channel continued to mature and became an ever larger part of the direct sales world. Consistency of household channel usage across participating companies could also be examined.
- Purchasing by season effects. The purchase information in the database could be examined to see which customers are unusually likely to purchase in particular seasons. Similarly, students could be asked to identify which companies or product categories seem to have particularly strong seasonality.

Depending on the level of student computer skills and the available software, there may be a need to provide some data aggregations in order to make the exercises suggested above accessible to marketing students.

The company which provided this dataset wishes to remain anonymous. The actual codes used by this company for Household ID's, company identifiers, products and order numbers have all been changed. Because the teaching extract represents only a small portion of the actual database it should be impossible to identify any actual household or to identify the participant companies.

Request for Assistance

Version 1.0 of this dataset and associated documentation is being made available as an initial prototype for testing and comment. The intention is that potential faculty users of the dataset will identify enhancements that will make the dataset more usable for their courses. These might include changes to the sample specifications, specifications of summary data fields to be provided, requests for improvements in the documentation or additional suggested student exercises. Of course, if there are errors these should be identified so they can be corrected. Suggestions for enhancements should be sent to the author at the e-mail address provided above.

Detailed Data Description

The 3 main household data files are provided in CSV format. The first field in all 3 files is a Household ID which can be used to link the files together; all 3 files are sorted by Household ID.

Household File: DMEF3YrBase.csv

The file contains 2 fields: HH_ID and ZIPCode.

Order File: DMEFOrders3Dataset2.csv

The variables in this file are:

- HH_ID is the household identifier number.
- CompanyID is the identifier number of the participating company that contributed the record to the database.
- OrderNum is the order number. It is used to link order and line item records.
- OrderDate is the date (YYYYMMDD) that the order was placed.
- DollarAmount is the dollar amount of the order.
- PaymentType gives the form of payment used. Possible values are 'AX' (American Express), 'D' (Discover), 'H' (house credit), 'MC' (Mastercard), 'O' (other) and 'V' (Visa).
- Channel indicates the ordering channel used for the purchase. Possible values are 'I' (Internet), 'C' (catalog) and 'O' (other).

Line Item File: DMEFLines3Dataset2.csv

The variables in this file are:

- HH_ID is the household identifier.
- CompanyID is the identifier number of the participating company that contributed the record to the database.
- OrderNum is the order number. It is used to link order and line item records.
- OrderDate is the date (YYYYMMDD) that the order was placed.
- ProductArea is the high level classification of the product purchased (see spreadsheet ProductCodes.xls for code values).
- MajorCategory is a more detailed classification of the product purchased (see spreadsheet ProductCodes.xls for code values).
- Dollars is the sales dollars associated with that line.
- Quantity is the number of units of the item that were purchased
- Channel indicates the ordering channel used for the purchase. Possible values are 'I' (Internet), 'C' (catalog) and 'O' (other).

There are 2 reference spreadsheets that provide additional information:

- ProductCodes.xls has the descriptions of ProductArea and MajorCategory codes.
- CorporateLinks.xls has a listing of corporate linkages. The first column contains CompanyID codes found in the order and line item records and the second column contains corporation identifiers. When 2 or more companies are owned by the same

corporation they will have the same associated corporation identifier. All CompanyID codes not found in the spreadsheet pertain to companies where there were no related companies in the dataset.