EPSILON CONFIDENTIAL   EPSILON CONFIDENTIAL   EPSILON CONFIDENTIAL

INFORMATION LISTED BELOW IS AVAILABLE UNDER THE TERMS OF THE CONFIDENTIALITY AGREEMENT

DOCUMENTATION TO ACCOMPANY

KDD-CUP-98

The Second International Knowledge Discovery and Data Mining Tools Competition

Held in Conjunction with KDD-98

The Fourth International Conference on Knowledge Discovery and Data Mining

Sponsored by the

American Association for Artificial Intelligence (AAAI)
Epsilon Data Mining Laboratory
Paralyzed Veterans of America (PVA)

Created: 7/20/98
Last update: 7/22/98
Edited by Dean Abbott: 10/1/14
File name: cup98DOC.txt

Table of Contents:

PROJECT OVERVIEW: A Fund Raising Net Return Prediction Model

BACKGROUND AND OBJECTIVES

The data set for this year's Cup has been generously provided by the Paralyzed Veterans of America (PVA). PVA is a not-for-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. With an in-house database of over 13 million donors, PVA is also one of the largest direct mail fund raisers in the country.

Participants in the '98 CUP will demonstrate the performance of their tool by analyzing the results of one of PVA's recent fund raising appeals. This mailing was sent to a total of 3.5 million PVA donors who were on the PVA database as of June 1997. Everyone included in this mailing had made at least one prior donation to PVA.

The mailing included a gift (or "premium") of personalized name & address labels plus an assortment of 10 note cards and envelopes. All of the donors who received this mailing were acquired by PVA through similar premium-oriented appeals such as this.

One group that is of particular interest to PVA is "Lapsed" donors. These are individuals who made their last donation to PVA 13 to 24 months ago. They represent an important group to PVA, since the longer someone goes without donating, the less likely they will be to give again. Therefore, recapture of these former donors is a critical aspect of PVA's fund raising efforts.

However, PVA has found that there is often an inverse correlation between likelihood to respond and the dollar amount of the gift, so a straight response model (a classification or discrimination task) will most likely net only very low dollar donors. High dollar donors will fall into the lower deciles, which would most likely be suppressed from future mailings. The lost revenue of these suppressed donors would then offset any gains due to the increased response rate of the low dollar donors.

Therefore, to improve the cost-effectiveness of future direct marketing efforts, PVA wishes to develop a model that will help them maximize the net revenue (a regression or estimation task) generated from future renewal mailings to Lapsed donors.

POPULATION

The population for this analysis will be Lapsed PVA donors who received the June '97 renewal mailing (appeal code "97NK"). Therefore, the analysis data set contains a subset of the total universe who received the mailing.

The analysis file includes all 191,779 Lapsed donors who received the mailing, with responders to the mailing marked with a flag in the TARGET_B field. The total dollar amount of each responder's gift is in the TARGET_D field.

The overall response rate for this direct mail promotion is 5.1%. The distribution of the target fields in the learning and validation files is as follows:

```
  Learning Data Set
  Target Variable: Binary Indicator of Response to 97NK
  Mailing
                                Cumulative  Cumulative
  TARGET_B   Frequency   Percent   Frequency    Percent
  -------------------------------------------------------
         0      90569      94.9       90569       94.9
         1       4843       5.1       95412      100.0


  Learning Data Set
  Target Variable: Donation Amount (in $) to 97NK Mailing

  Variable       N          Mean   Minimum       Maximum
  ------------------------------------------------------
  TARGET_D   95412     0.7930732         0   200.0000000
  ------------------------------------------------------



  Validation Data Set
  Target Variable: Binary Indicator of Response to 97NK
  Mailing
                                Cumulative  Cumulative
  TARGET_B   Frequency   Percent   Frequency    Percent
  ------------------------------------------------------
         0      91494      94.9       91494       94.9
         1       4873       5.1       96367      100.0

  Validation Data Set
  Target Variable: Donation Amount (in $) to 97NK Mailing

  Variable       N          Mean   Minimum       Maximum
  ------------------------------------------------------
  TARGET_D   96367     0.7895819         0   500.0000000
  ------------------------------------------------------
```

The average donation amount (in $) among the responsers is:

```
  Learning Data Set
  Target Variable: Donation Amount (in $) to 97NK
  Mailing

      N          Mean       Minimum       Maximum
  -----------------------------------------------
   4843    15.6243444     1.0000000   200.0000000
  -----------------------------------------------


  Validation Data Set
  Target Variable: Donation Amount (in $) to 97NK
  Mailing

      N          Mean       Minimum       Maximum
  -----------------------------------------------
   4873    15.6145372     0.3200000   500.0000000
  -----------------------------------------------
```

COST MATRIX

The package cost (including the mail cost) is $0.68 per piece mailed.

ANALYSIS TIME FRAME AND REFERENCE DATE

The 97NK mailing was sent out on June 1997. All information included in the file (excluding the giving history date fields) is reflective of behavior prior to 6/97. This date may be used as the reference date in generating the "number of months since" or "time since" or "elapsed time" variables. The participants could also find the reference date information in the filed ADATE_2. This filed contains the dates the 97NK promotion was mailed.

EVALUATION RULES

Once again, the objective of the analysis will be to maximize the net revenue generated from this mailing - a censored regression or estimation problem. The response variable is, thus, continuous (for the lack of a better common term.) Alhough we are releasing both the binary and the continuous versions of the target variable (TARGET_B and TARGET_D respectively), the program committee will use the predicted value of the donation (dollar) amount (for the target variable TARGET_D) in evaluating the results. So, returning the predicted value of the binary target variable TARGET_B and its associated probability/strength will not be sufficient.

The typical outcome of predictive modeling in database marketing is an estimate of the expected response/return per customer in the database. A marketer will mail to a customer so long as the expected return from an order exceeds the cost invested in generating the order, i.e., the cost of promotion. For our purpose, the package cost (including the mail cost) is $0.68 per piece mailed.

KDD-CUP committee will evaluate the results based solely on the net revenue generated on the hold-out or validation sample.

The measure we will use is:

> Sum (the actual donation amount - $0.68) over all records for which the expected revenue (or predicted value of the donation) is over $0.68.

This is a direct measure of profit. The winner will be the participant with the highest actual sum. The results will be rounded to the nearest 10 dollars.

SUMMARY STATISTICS (MIN & MAX)

Summary statistics are provided for the numeric variables only.

Variable	Learning Data Set		Validation Data Set	
	Minimum	Maximum	Minimum	Maximum
DOB	0	9710.00	0	9705.00
AGE	1.0000000	98.0000000	1.0000000	98.0000000
NUMCHLD	1.0000000	7.0000000	1.0000000	7.0000000
INCOME	1.0000000	7.0000000	1.0000000	7.0000000
WEALTH1	0	9.0000000	0	9.0000000
HIT	0	241.0000000	0	242.0000000
MBCRAFT	0	6.0000000	0	6.0000000
MBGARDEN	0	4.0000000	0	3.0000000
MBBOOKS	0	9.0000000	0	9.0000000
MBCOLECT	0	6.0000000	0	6.0000000
MAGFAML	0	9.0000000	0	9.0000000
MAGFEM	0	5.0000000	0	4.0000000

```
MAGMALE               0     4.0000000             0     4.0000000
PUBGARDN              0     5.0000000             0     6.0000000
PUBCULIN              0     6.0000000             0     4.0000000
PUBHLTH               0     9.0000000             0     9.0000000
PUBDOITY              0     8.0000000             0     9.0000000
PUBNEWFN              0     9.0000000             0     9.0000000
PUBPHOTO              0     2.0000000             0     2.0000000
PUBOPP                0     9.0000000             0     9.0000000
MALEMILI              0    99.0000000             0    99.0000000
MALEVET               0    99.0000000             0    99.0000000
VIETVETS              0    99.0000000             0    99.0000000
WWIIVETS              0    99.0000000             0    99.0000000
LOCALGOV              0    99.0000000             0    76.0000000
STATEGOV              0    99.0000000             0    99.0000000
FEDGOV                0    87.0000000             0    99.0000000
WEALTH2               0     9.0000000             0     9.0000000
CARDPROM      1.0000000    61.0000000             0    62.0000000
MAXADATE        9608.00       9702.00       9607.00       9702.00
NUMPROM       4.0000000   195.0000000     4.0000000   189.0000000
CARDPM12              0    19.0000000             0    21.0000000
NUMPRM12      1.0000000    78.0000000     1.0000000    76.0000000
RAMNT_3       2.0000000    50.0000000     2.0000000   200.0000000
RAMNT_4       1.0000000   100.0000000     1.0000000   100.0000000
RAMNT_5       4.0000000    50.0000000     5.0000000    30.0000000
RAMNT_6       1.0000000   100.0000000     1.0000000   100.0000000
RAMNT_7       1.0000000   250.0000000     1.0000000   203.0000000
RAMNT_8       1.0000000   500.0000000     0.3200000       3713.31
RAMNT_9       1.0000000       1000.00     1.0000000   300.0000000
RAMNT_10      0.3000000   500.0000000     1.0000000      10000.00
RAMNT_11      1.0000000   300.0000000     1.0000000       1000.00
RAMNT_12      1.0000000   300.0000000     1.0000000   500.0000000
RAMNT_13      0.1000000   500.0000000     1.0000000   300.0000000
RAMNT_14      1.0000000   200.0000000     1.0000000   600.0000000
RAMNT_15      1.0000000   300.0000000     1.0000000   500.0000000
RAMNT_16      0.5000000   500.0000000     0.5000000   205.0000000
RAMNT_17      1.0000000   500.0000000     1.0000000   500.0000000
RAMNT_18      1.0000000       1000.00     0.3200000   300.0000000
RAMNT_19      1.0000000   970.0000000     1.0000000   250.0000000
RAMNT_20      0.5000000   250.0000000     1.0000000   200.0000000
RAMNT_21      1.0000000   300.0000000     1.0000000       1000.00
RAMNT_22      0.2900000   300.0000000     1.0000000   500.0000000
RAMNT_23      0.3000000   200.0000000     1.0000000   300.0000000
RAMNT_24      1.0000000   225.0000000     0.5000000   250.0000000
RAMNTALL     13.0000000       9485.00    13.0000000      10253.00
NGIFTALL      1.0000000   237.0000000     1.0000000   126.0000000
CARDGIFT              0    41.0000000             0    45.0000000
MINRAMNT              0       1000.00             0   436.0000000
MINRDATE        7506.00       9702.00       8010.00       9702.00
MAXRAMNT      5.0000000       5000.00     5.0000000      10000.00
MAXRDATE        7510.00       9702.00       8011.00       9702.00
LASTGIFT              0       1000.00             0      10000.00
LASTDATE        9503.00       9702.00       9503.00       9702.00
FISTDATE              0       9603.00             0       9603.00
NEXTDATE        7211.00       9702.00       7312.00       9702.00
TIMELAG               0       1088.00             0       1060.00
AVGGIFT       1.2857143       1000.00     1.5789474   650.0000000
CONTROLN      1.0000000     191779.00     3.0000000     191776.00
TARGET_B              0     1.0000000             0     1.0000000
TARGET_D              0   200.0000000             0   500.0000000
HPHONE_D              0     1.0000000             0     1.0000000
CLUSTER2      1.0000000    62.0000000     1.0000000    62.0000000
```

Time Frame and Date Fields

This mailing was mailed to a total of 3.5 million PVA donors who were on the PVA database as of June 1997. All information contained in the analysis dataset reflects the donor status prior to 6/97 (except the gift receipt dates, which will follow the promotion dates.) This date could be used as the "end date" or "rerefence date" in the calculation of "number of months since" variables.

KDD-CUP-98 Program Committee

o Vasant Dhar, New York University, New York, NY
o Tom Fawcett, Bell Atlantic, New York, NY
o Georges Grinstein, University of Massachusetts, Lowell, MA
o Ismail Parsa, Epsilon, Burlington, MA
o Gregory Piatetsky-Shapiro, Knowledge Stream Partners, Boston, MA
o Foster Provost, Bell Atlantic, New York, NY
o Kyusoek Shim, Bell Laboratories, Murray Hill, NJ