

**Data Analysis & Data Mining**

**Module code: SD3331**

**Institution**: University of East London (UEL)

Alexandros Akrivopoulos-Hughes

**ID:** u1235441

**Module Leader: Dr Sin Wee Lee**

Index

[1 Introduction of the dataset 4](#_Toc341098571)

[2 Analysis of dataset structure 5](#_Toc341098573)

[3 Data Mining](#_Toc341098574) 8

[3.1 Technique 1](#_Toc341098575) 9

[3.2 Technique 2](#_Toc341098576) 38

[3.3 Technique 3](#_Toc341098578) 41

[4 Results and Knowledge](#_Toc341098579) 43

[5 Referencing](#_Toc341098580) 44

Introduction of the Dataset

**Background and Objectives**

The dataset used for the data mining process is the **KDD’s Cup 1998 Data.** This was a competition that many different algorithms and techniques were used by the competitors. The competition’s task was a regression problem where the goal was to *“estimate the return from a direct mailing in order to maximize donation profits.”*[1]

According to the dataset description, the dataset was provided by the Paralyzed Veterans of America, a non profit organization that provides programs and services for US veterans with spinal cord injuries or disease. PVA is one of the largest direct mail fund raisers in the United States of America. [1]

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.

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.

PVA also 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 (classification or discrimination task) will most likely net only very low dollar donors.

PVA also 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.

**Evaluation Rules (briefly)**

The objective of the analysis will be to maximise the net revenue generated from this mailing - a censored regression or the lack of a better common term. However, this assignment will not deal with the construction of new models to forecast or predict as it is not a competitor in KDD. Probably some rules, correlations and tendencies will be the outcome of this work.

**Data Sources, Order, Structure, Type and Format of the variables in the datasets**

The table below contains information about the used dataset files as well as the dataset description files.

|  |  |
| --- | --- |
| **cup98LRN.zip** | PKZIP compressed raw LEARNING dataset.  Internal name: cup98LRN.txt  File size: 36,468,735 bytes zipped. 117,167,952 bytes unzipped.  Number of Records: 95412.  Number of Fields: 481. |
| **cup98VAL.zip** | PKZIP compressed raw VALIDATION dataset.  Internal name: cup98VAL.txt  File size: 36,763,018 bytes zipped. 117,943,347 bytes unzipped.  Number of Records: 96367.  Number of Fields: 479. |
| **cup98DIC.txt** | Data dictionary to accompany the analysis dataset. |
| **cup98DOC.txt** | This file, an overview and pointer to more detailed information about the competition |

The dataset includes:

* 24 months of detailed PVA promotion and giving history
* A summary of the promotions sent to the donors over the most recent 12 months prior to the mailing
* Summary variables reflecting each donor’s lifetime giving history
* Overlay demographics, including a mix of household and area level data
* All other available data from the PVA database

The name of the variables in the learning and validation datasets is included in each file as the top (header) record. We have the following data types available:

* Num: numeric
* Char: string/character

The data dictionary file is **cup98DIC.txt**. There, all fields are explained in detail.

**Dataset Preprocessing: Sampling**

The processing will proceed by first importing all comma separated values into a *MySQL database* using a *Java SE Application* written for this purpose, so that we can easily select, transform and process the attributes (columns) that are mandatory and important in our data mining process. Also, as the sample is fairly big, some sampling will occur (probably we will use only 2000 rows, randomly selected out of 96K rows) to shorten calculation time.

*After sampling, the dataset’s structure is the following:*

**Number of Attributes: 23**

**Number of Instances: 2000**

**Attributes of the dataset:**

|  |  |
| --- | --- |
| ID | ID of instance in database. Not used in data mining process |
| ODATEDW | Origin Date. Date of donor's first gift to PVA YYMM format (Year/Month). |
| TCODE | Donor title code |
| PVASTATE | EPVA State or PVA State  Indicates whether the donor lives in a state served by the organization's EPVA chapter  P = PVA State  E = EPVA State (Northeastern US) |
| DOB | Date of birth (YYMM, Year/Month format.) |
| RECINHSE | In House File Flag  Blank means that is not an in house records  X means that the Donor has given to PVA’s In House programme |
| RECP3 | P3 File Flag  Blank means that is not a P3 Record  X means that the Donor has given to PVA’s P3 programme |
| DOMAIN | DOMAIN/Cluster code. A nominal or symbolic field.  It could be broken down by bytes as explained below.  1st byte = Urbanicity level of the donor's neighborhood  U=Urban  C=City  S=Suburban  T=Town  R=Rural  2nd byte = Socio-Economic status of the neighborhood  1 = Highest SES  2 = Average SES  3 = Lowest SES (except for Urban communities, where  1 = Highest SES, 2 = Above average SES,  3 = Below average SES, 4 = Lowest SES.) |
| AGE | Overlay Age  0 = missing |
| HOMEOWNR | Home Owner Flag  H = Home owner  U = Unknown |
| INCOME | Household Income |
| GENDER | Gender  M = Male  F = Female  U = Unknown  J = Joint Account, unknown gender |
| WEALTH1 | Wealth Rating |
| HIT | MOR Flag # HIT (Mail Order Response)  Indicates total number of known times the donor has responded to a mail order offer other than PVA's. |
| WEALTH2 | Wealth Rating  Wealth rating uses median family income and population statistics from each area to index relative wealth within each state The segments are denoted 0-9, with 9 being the highest income group and zero being the lowest. Each rating has a different meaning within each state. |
| ADATE\_2 | Date the 97NK promotion was mailed |
| RFA\_2 | Donor's RFA status as of 97NK promotion date |
| RAMNTALL | Dollar amount of lifetime gifts to date |
| NGIFTALL | Number of lifetime gifts to date |
| AVGGIFT | Average dollar amount of gifts to date |
| *The following two attributes are for validation purposes only* | |
| TARGET\_B | Target Variable: Binary Indicator for Response to 97NK Mailing |
| TARGET\_D | Target Variable: Donation Amount (in $) associated with the Response to 97NK Mailing |

**Dataset Preprocessing: Data Cleaning, Data Integration and Transformation**

The sampled dataset will be examined further for data anomalies such as noisy values as well as missing values. After those problems are solved using the best solution for each case, numerical value normalization will occur in order to reduce calculation errors and improve accuracy.

Also, any nominal values left in the dataset will be transformed to numeric.

It should be noted that for the construction of the sample, important attributes for the data mining process such as ODATEDW, DOB, DOMAIN, AGE, HOMEOWNR, INCOME do not contain missing values. This was feasible by entering restriction parameters to the selection query.

The selection query can be seen below:

CREATE TABLE sample AS (SELECT

ID,

ODATEDW,

TCODE,

PVASTATE,

DOB,

RECINHSE,

RECP3,

DOMAIN,

AGE,

HOMEOWNR,

INCOME,

GENDER,

WEALTH1,

HIT,

WEALTH2,

ADATE\_2,

RFA\_2,

RAMNTALL,

NGIFTALL,

AVGGIFT,

TARGET\_B,

TARGET\_D

FROM

kdd98

WHERE ODATEDW <> '' AND DOB <> '' AND DOMAIN <> '' AND AGE <> '' AND HOMEOWNR <> '' AND INCOME <> ''

ORDER BY (SELECT RAND())

LIMIT 2000);

The selection query can be seen below:

Analysis of Dataset Structure

Analyze the final dataset here

Data Mining

Technique 1 - WEKA

Technique2 - WEKA

Technique 3 - WEKA

Results and Knowledge Acquired

Referencing

[1] kdd.ics.uci.edi/databases/kddcup98/kddcup98.html