

**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.

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 1000 or less rows, randomly selected out of 96K rows) to shorten calculation time.

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

After the sampling, we will examine our sampled and smaller dataset for data anomalies such as noisy values and missing values. We will then decide how to cope with these values and which is the best solution for our case.

Some numerical values may be normalized in order to reduce calculation errors and improve accuracy and also any nominal values left in the dataset will be transformed to numeric.

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