

**Data Analysis & Data Mining**

**Module code: SD3331**

**Institution**: University of East London

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

[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]

The dataset description states that the dataset was provided by a non profit organization called Paralyzed Veterans of America. This organization provides services and programs as well as organises donations for US military veterans with spinal cord injuries or disease. It is also considered to be *“one of the largest direct mail fund raisers in the United States of America”* [1]

Everyone instance of the dataset is someone who had made at least one prior donation to PVA. The mailing for donation was sent to everyone that was on the PVA database on June of 1997.

**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 Strengths**

The dataset has a very consistent structure, having in mind its size (96K instances). An application parsed all of the lines successfully without stopping because of inconsistency or corruption.

Also, its redundancy of instances is an advantage, as someone can have a randomly selected sample without any missing or noisy values. This means that dataset pre-processing can be fairly easy and quick for someone that has a basic knowledge on database queries.

**Dataset Weaknesses**

The dataset contains nearly every attribute that PVA keeps in its databases, thus is very generic. It was difficult out of 781 attributes to select the correct 22 for our data mining purposes.

**Dataset pre-processing: 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.

**Dataset Pre-processing: 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,

RECINHSE,

RECP3,

DOMAIN,

AGE,

HOMEOWNR,

INCOME,

GENDER,

WEALTH1,

HIT,

WEALTH2,

ADATE\_2,

RFA\_2,

RAMNTALL,

NGIFTALL,

AVGGIFT,

TARGET\_B

FROM

kdd98

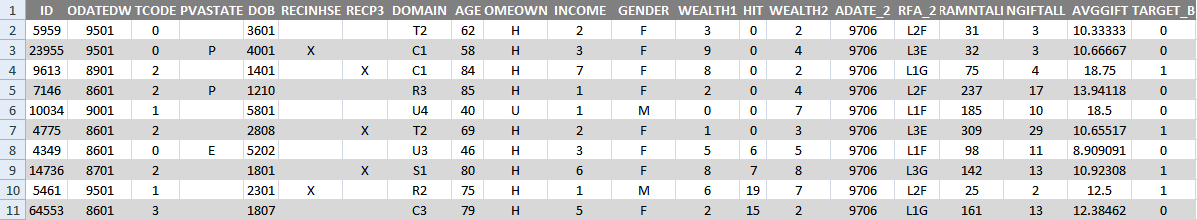
WHERE ODATEDW <> '' AND DOB <> '' AND DOMAIN <> '' AND AGE <> '' AND HOMEOWNR <> '' AND INCOME <> ''

ORDER BY (SELECT RAND())

LIMIT 2000);

So, because of the redundancy of instances, the task of coping with missing values was solved by inserting restriction parameters to the selection query. At the moment there is no attribute that was supposed to have a value but the value is missing, except from the attributes that a missing value means something, like for example RECP3, where absence of value means that the instance does not belong to a particular group.

To make the dataset structure more comprehensive, a snapshot of 10 random records can be found below.



Analysis of Dataset Structure

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

**Number of Attributes: 20**

**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) |
| 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 |

Data Mining

All data mining techniques are applied using WEKA (Waikato Environment for Knowledge Analysis), a free open source tool for data mining purposes.

At first, an **ADTree (Alternating Decision Tree)**[2],[3] is created.

The reason this particular technique is used is because it combines normal boosting and normal decision trees. In application, this means that prediction performance is increased, as a result of boosting, as well as the trees can be merged together, which is not a characteristic of normal boosting.

Boosting actually is a set of meta-algorithms that actually use many weak classifiers in together to create one strong classifier[4]. Boosting in alternating decision trees adds three nodes to the tree for each iteration and then the boosting algorithm determines the correct place for the “splitter node” after analyzing all nodes that have already been created.

After the tree is created, it can be traversed in order to arrive at predictions. For this procedure and in order to gain prediction values, the overall sum of all the prediction nodes that are crossed in the transversal are taken into account.

So, all the weak hypotheses are used in boosting in order to arrive at a single and easily-understood representation.

The results can be seen below:

=== Run information ===

Scheme:weka.classifiers.trees.ADTree -B 10 -E -3

Relation: QueryResult-weka.filters.unsupervised.attribute.Remove-R20

Instances: 2000

Attributes: 19

ODATEDW

TCODE

PVASTATE

DOB

RECINHSE

RECP3

DOMAIN

AGE

HOMEOWNR

INCOME

GENDER

WEALTH1

HIT

WEALTH2

ADATE\_2

RFA\_2

RAMNTALL

NGIFTALL

TARGET\_B

Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

Alternating decision tree:

: -1.447

| (1)WEALTH1 = 4: -1.759

| (1)WEALTH1 != 4: 0.033

| | (3)WEALTH2 = 9: -0.7

| | | (7)DOB = 2801: 1.066

| | | (7)DOB != 2801: -0.411

| | | | (8)DOB = 5307: 1.251

| | | | (8)DOB != 5307: -1.285

| | (3)WEALTH2 != 9: 0.04

| (2)NGIFTALL = 1: -0.599

| (2)NGIFTALL != 1: 0.031

| | (5)ODATEDW = 9401: -0.375

| | (5)ODATEDW != 9401: 0.05

| | | (6)RAMNTALL = 30: -1.281

| | | (6)RAMNTALL != 30: 0.018

| | | (9)RFA\_2 = L1F: -0.239

| | | | (10)INCOME = 4: -1.31

| | | | (10)INCOME != 4: 0.092

| | | (9)RFA\_2 != L1F: 0.073

| (4)AGE = 66: -1.358

| (4)AGE != 66: 0.009

Legend: -ve = 0, +ve = 1

Tree size (total number of nodes): 31

Leaves (number of predictor nodes): 21

Time taken to build model: 0.5 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 652 95.8824 %

Incorrectly Classified Instances 28 4.1176 %

Kappa statistic 0

Mean absolute error 0.1959

Root mean squared error 0.253

Relative absolute error 207.051 %

Root relative squared error 126.8552 %

Total Number of Instances 680

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1 1 0.959 1 0.979 0.49 0

0 0 0 0 0 0.49 1

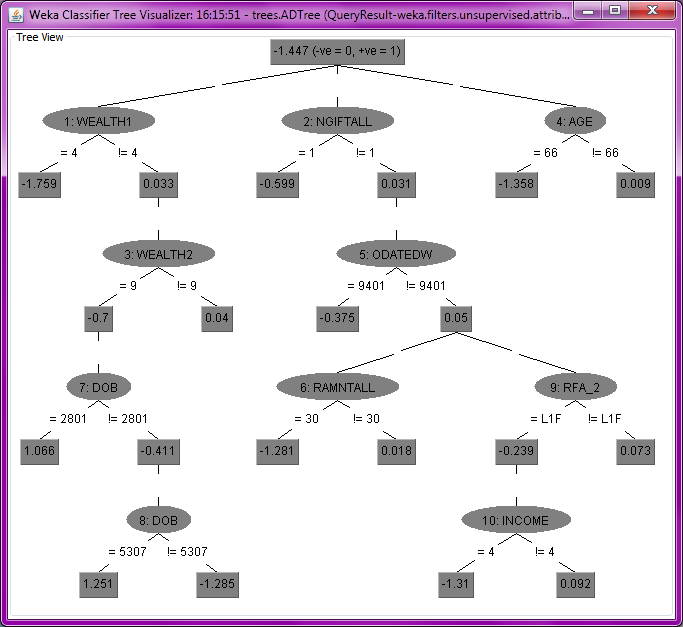
Weighted Avg. 0.959 0.959 0.919 0.959 0.939 0.49

=== Confusion Matrix ===

a b <-- classified as

652 0 | a = 0

28 0 | b = 1



1 Alternating Decision Tree Visualization

Results and Knowledge Acquired

The above alternating decision tree can now classify, using an instance’s attributes whether this instance is a target (response and donation to the mailing) or not.

The scoring is performed using the tree’s rules, summing all the values of the prediction nodes through which the instance of the sample passes. If the final score is positive, the instance can be classified as target.

This means that for future instances it may be feasible to predict whether the instance will be a target or not, by classifying the instance using this decision tree. Of course, it should be taken into account that the sample must be representative of the dataset and the dataset’s data is valid.

Commenting on the resulting tree, it is obvious that the major attributes for classification are wealth, other donations to the organization so far and age.

It is interesting that instances that are not on the highest levels of wealth but instances that belong to medium wealth classes are usually responders. From other wealth classes, the main trend is that people of the highest class of wealth2 field which measures area wealth, are those instances which responders belong to.

Also, PVA’s main concern that people who only buy one gift don’t tend to buy another one in the future, can be confirmed from the decision tree. And for those who are donating more than one gift, if they haven’t donated for a while, again PVA’s concern is verified, as they aren’t scored as responders easily, as more info must be taken into account, as the total dollars given, donor status as well as income.

It is also interesting to realize that people of a specific age group (50-70) tend to be donors. This may be explained by various socio-political reasons.

Referencing

[1] UCI Machine Learning Repository <http://kdd.ics.uci.edi/databases/kddcup98/kddcup98.html> (Retrieved 02/12/2012)

[2] Yoav Freund and Llew Mason, The Alternating Decision Tree Algorithm, <http://perun.pmf.uns.ac.rs/radovanovic/dmsem/cd/install/Weka/doc/classifiers-papers/trees/ADTree/atrees.pdf> (Retrieved 02/12/2012)

[3] Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, Eibe Frank and Mark Hall, Multiclass Alternating Decision Trees, <http://www.cs.waikato.ac.nz/~bernhard/papers/ecml2002.pdf> (Retrieved 02/12/2012)

[4] Robert E. Schapire, The Strength of Weak Learnability, <http://www.cs.princeton.edu/~schapire/papers/strengthofweak.pdf> (Retrieved 03/12/2012)