The Hong Kong Polytechnic University
Department of Computing
COMP 578 Data Mining and Data Warehousing
Semester 1, 03-04
Assignment 2 Part II (Group assignment)
Due: December 23, 2003

For this assignment, you are to get into groups of either 3 or 4 to work on a real world business problem for a Bank. You are given a data set that contains information on customers of an insurance company. The data consists of 86 attributes and includes product usage data and socio-demographic data.

Data Characteristics

Database contains data of 32,562 customers of a bank. The data are stored in four tables, CUSTOMER, ITEM, TRANSACTION and TRANSACTION_ITEM. The CUSTOMER table contains demographic data of the customers. The ITEM table contains descriptions of the item that a customer may purchase. The TRANSACTION table contains the dates in which transactions are made by the customers. And the TRANSACTION_ITEM table contains a list of items purchased during a transaction. The details of these tables are given below.

Table: CUSTOMER
Description: Each record in this table represents a customer.
No. of records: 32,562

	Attribute	**Description**
1	CUSTOMER_ID	Customer ID
2	AGE	Age
3	WORKCLASS	Work class of the customer (e.g., private, federal government, etc.)
4	EDUCATION	Education level (e.g., preschool, master, etc.)
5	MARITAL_STATUS	Marital status
6	OCCUPATION	Occupation
7	RELATIONSHIP	Relationship (e.g., not in family, own child, etc.)
8	RACE	Race
9	SEX	Sex
10	CAPITAL_GAIN	Gain in capital
11	CAPITAL_LOSS	Loss in capital
12	HOURS_PER_WEEK	No. of working hours per week
13	NATIVE_COUNTRY	Native country
14	SALARY	Salary

Table: ITEM
Description: Each record in this table represents an item.
No. of records: 10

	Attribute	Description
1	ITEM_ID	Item ID
2	DESCRIPTION	Description
3	PRICE	Unit price

Table: TRANSACTION
Description: Each record in this table represents the transaction made by a customer in a purchase.
No. of records: 222,566

	Attribute	Description
1	TRANSACTION_ID	Transaction ID
2	CUSTOMER_ID	Customer ID
3	TRANSACTION_DATE	Transaction Date

Table: TRANSACTION_ITEM
Description: Each record in this table represents an item a customer purchased in a transaction.
No. of records: 454,693

	Attribute	Description
1	TRANSACTION_ID	Transaction ID
2	ITEM_ID	Item ID
3	QUANTITY	Quantity

Data File
The file, data.mdb, contains the dataset described above.

What to do

You are to use the technique you learned in this course to mine the dataset. In the report, you should give details of the process you took to mine the data and what you have discovered. You are to convince the management of the bank the effectiveness of data mining, both objectively and subjectively.

Please submit the report and state, at the end of it, the role each team member played in the assignment.

To mine the data, you are free to develop your own data mining software, or use SPSS's Clementine. You can also consider using WEKA, a public domain data mining software for this assignment. The details of WEKA is attached.

WEKA

Weka is a collection of machine learning algorithms for solving real-world data mining problems. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License.

Installation instruction (assumes WEKA will be installed on PC):

Go to WEKA Homepage http://www.cs.waikato.ac.nz/ml/weka/index.html and download the GUI version of Weka (weka-3-2-3.exe) , save it in a temp directory (e.g. d:\temp)

Make sure that you have Java1.2 in your computer, otherwise download Java 2 Standard Edition from http://java.sun.com .

Execute weka-3-2-3.exe, install WEKA in a directory you like (e.g. d:\Weka-3-2)

Using WEKA

Go to *Start ® Programs ® WEKA ® Weka-3-2* and then the WEKA application GUI chooser will be opened. Next, choose *Explorer* and then you can try to analyse sample data in ARFF format. Some sample dataset can be found under "data" subdirectory of WEKA folder (e.g. d:\Weka-3-2\data).

References

Tutorial
Package Documentation } Can be founded under *Start ® Programs ® WEKA ® Weka-3-2*
Troubleshooting http://www.cs.waikato.ac.nz/~ml/weka/tips_and_tricks.html
ARFF data format http://www.cs.waikato.ac.nz/~ml/weka/arff.html
The development version also has a user guide for the Explorer GUI. A tutorial for the Experimenter is available from http://www.cs.waikato.ac.nz/~ml/weka/Experiments.pdf