USER GUIDE

E-Business Technology Institute,
The University of Hong Kong

AUGUST 2005

AlphaMiner User Guide

E-Business Technology Institute, The University of Hong Kong.
August 2005

First Edition (AUGUST 2005 v1.2)

Comments may be addressed to:
E-Business Technology Institute
Room G01-G05,
Technology Innovation & Incubation Building,
The University of Hong Kong,
Pokfulam Road, Hong Kong
Website: http://www.eti.hku.hk
Tel: (852) 2299-0505
Fax: (852) 2299-0500

INTRODUCTION TO ALPHAMINER

AlphaMiner is a general-purpose data mining system that is used to facilitate the implementation of a data mining process. It provides a wide range of functionality to enable the user to perform the following data mining steps:

- Access a variety of data sources
- Explore data in histogram and multi-variables plots
- Manipulate data
- Build different data mining models
- Analyze models
- Deploy models in the enterprise environment

1.1 System Overview

The left column shown in *Figure 1-1* below shows the data mining cases. For easy retrieval, they are grouped and organized under different folders by industry, problem type, business objectives, data mining goal or company names.

Figure 1-2 shows an instance of a data mining case. In order to perform data mining, AlphaMiner provides different kinds of data processing and model construction operations which are shown on the left of the user interface. Operations can be classified into Data Access, Data Exploration, Data Transformation, Modeling, Assessment and Deployment. The yellow dialog box holds the case information. The case workflow can be constructed by dragging and dropping the operators from the menu onto the Case Diagram panel.

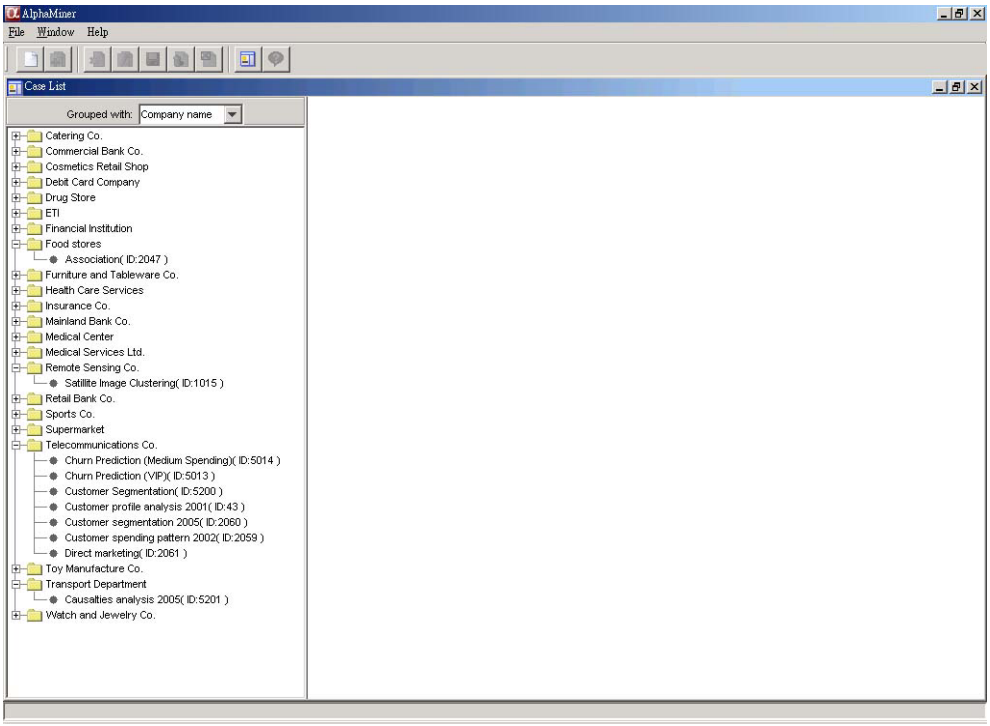


Figure 1-1 Data mining cases

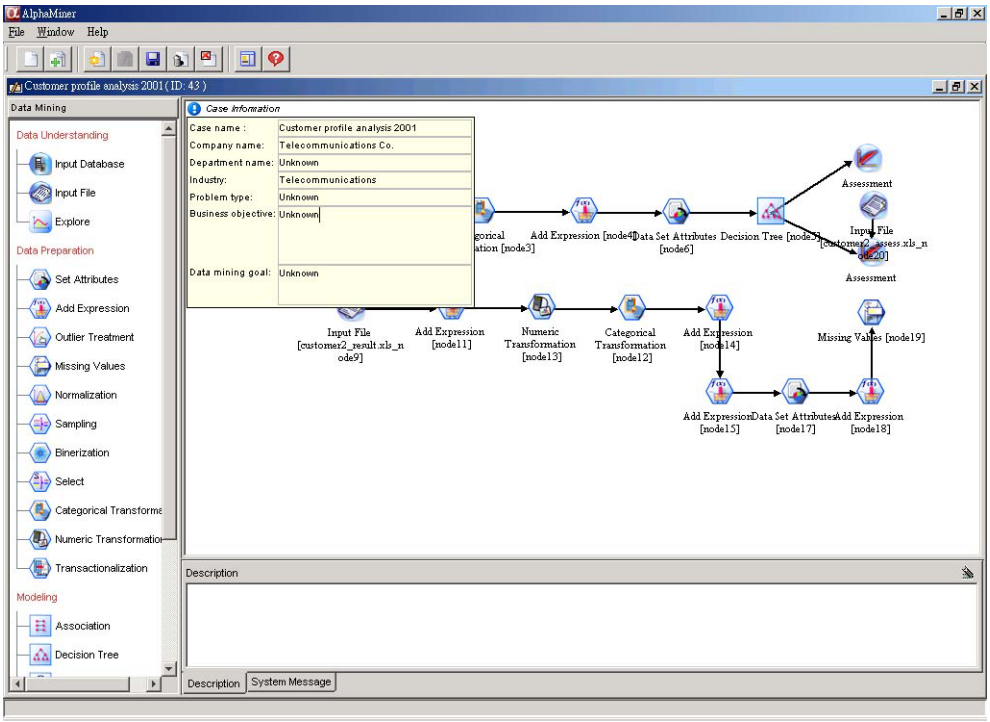


Figure 1-2 A workflow-based data mining case

1.2 Data Access

Two kinds of data input are supported by AlphaMiner as shown in *Table 1-1*:

Operator Name	Description
Input File	Imports dataset in flat files
Input Database	Imports data from ODBC compliant databases

Table 1-1 Data access operators

Input File (see *Figure 1-3*) enables the user to select the following type of files as its input data format: Attribute-Relation, Comma Separated Value, Excel, MS IIS Log, NCSA Common Log, and NCSA Combined Log.

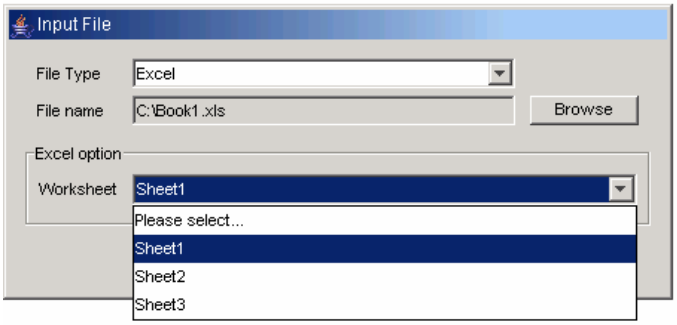


Figure 1-3 Input File

Input Database (see *Figure 1-4*) allows the user to select database as input data source through JDBC-ODBC Bridge.

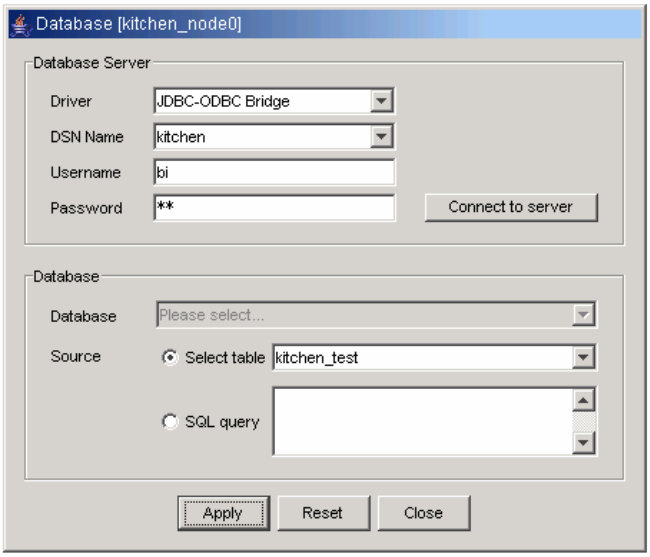


Figure 1-4 Input Database

1.3 Data Exploration

Data exploration function is implemented in AlphaMiner as shown in *Table 1-2*.

Operator Name	Description
Explore	Provides information on data distribution and statistics, and visualizes data in attribute specific histogram and cross attribute scatter plot

Table 1-2 Data exploration operator

Explore provides the following statistics exploration of data:

- Number and percentage of missing values
- Number of distinct values
- Number of unique values
- Minimum value (for numeric attributes)
- Maximum value (for numeric attributes)
- Mean (for numeric attributes)
- Standard deviation (for numeric attributes)
- Categorical values and respective counts (for categorical attributes)

In addition, single or cross attribute distribution can be reviewed in histograms or plots (*Figure 1-5*).

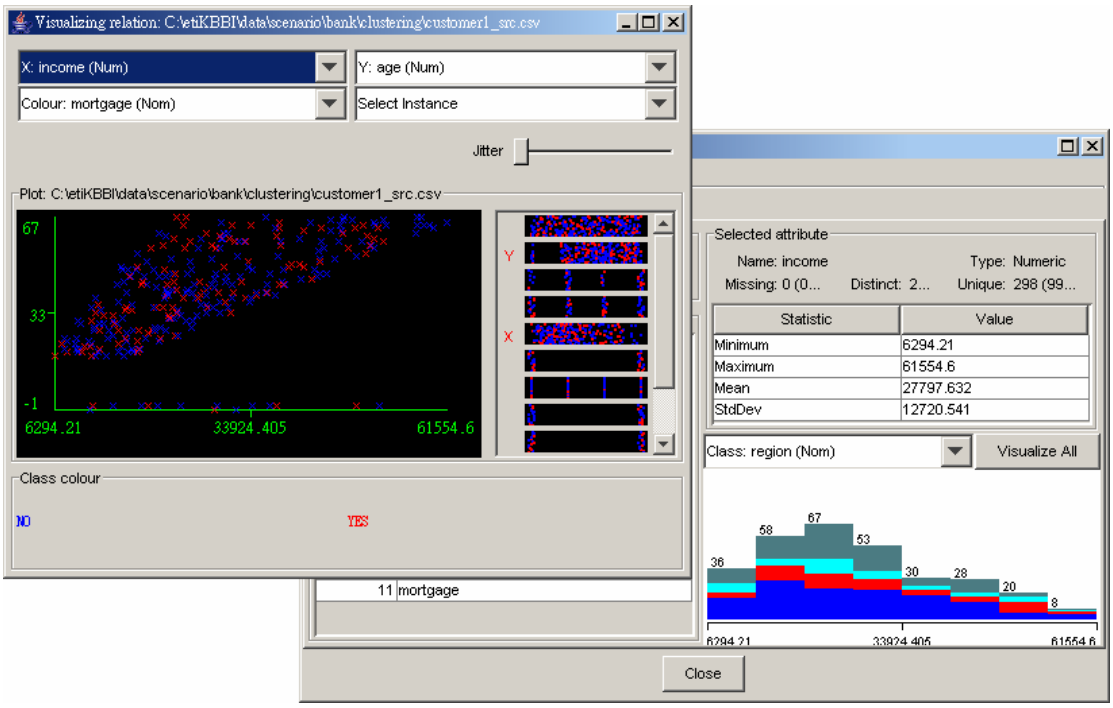


Figure 1-5 Explore

1.4 Data Transformation

The following data transformations are implemented in the AlphaMiner system (see *Table 1-3*):

Operator Name	Description
 Set Attributes	Chooses attributes and target attribute used in a dataset, and maps numeric into categorical attributes or vice versa
 Add Expression	Appends a new attribute by applying a mathematical expression involving attributes and numeric constants to a dataset
 Outlier Treatment	Treats and remove outliers of attributes
 Missing Values	Replaces missing attribute values by a given value
 Normalization	Performs Linear Normalization or Standard Normal Distribution on a dataset
 Sampling	Produces a sub sample of a dataset
 Binerization	Transforms a categorical attribute into a set of binary attributes representing its categories
 Select	Selects a subset of a dataset by range or by value
 Categorical Transformation	Maps categories of categorical attributes onto other categories or numeric values
 Numeric Transformation	Transforms numeric into categorical attribute simply using every value as category, or discretization
 Transactionalization tion	Transforms a non-transactional dataset into transactional format

Table 1-3 Data transformation operators

Outlier Treatment (see *Figure 1-6*) replaces outliers of a numeric attribute as missing values. For categorical attribute, it treats the outlier as missing values automatically according to the definition in the meta-data. All categories not defined in the meta-data will be considered as outliers and will be removed.

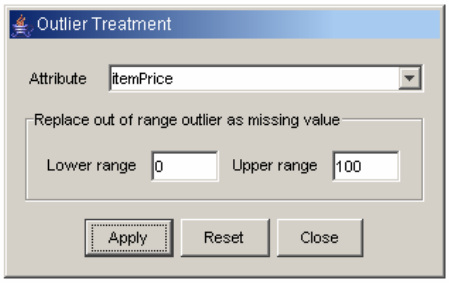


Figure 1-6 Outlier Treatment

Sampling (see *Figure 1-7*) produces a sub sample from the dataset by random sampling with a chosen sample size, selecting one sample of every n instance (1-in-N sampling) or picking the specified number of records (First-N sampling).

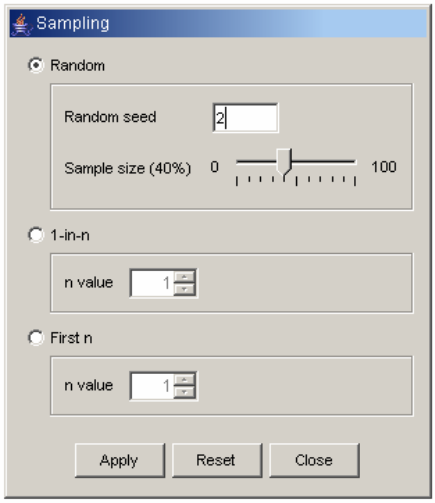


Figure 1-7 Sampling

Numeric Transformation (see *Figure 1-8*) transforms numeric attribute into categorical attribute by creating a new category using each numeric value. On the other hand, the user can use discretization to decompose the value range of the source attribute into an ordered set of intervals. It can be performed either by setting the interval width or number of intervals (bins).

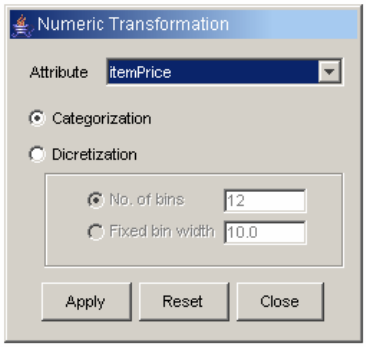


Figure 1-8 Numeric Transformation

1.5 Modeling

Four modeling algorithms are implemented in the AlphaMiner system as shown in *Table 1-4*:

Operator Name	Description
Association	Identifies relationships or affinities between items and/or between features (also know as basket analysis)
Clustering	Partitions a dataset into clusters such that the data in each cluster shares some common traits
Decision Tree	Constructs a decision model to perform classification and prediction for a subject of interest. Rule sets are produced to illustrate how decisions are made using the model
Logistic Regression	Similar to Decision Tree modeling technique. Uses regression technique to perform classification and prediction for a subject of interest

Table 1-4 Modeling operators

Association (see *Figure 1-9 and Figure 1-10*) identifies the relationships or affinities between items and/or between features. These relationships are then expressed as a collection of association rules. The user can set the generated rules by specifying the minimum and maximum numbers of items each rule contains.

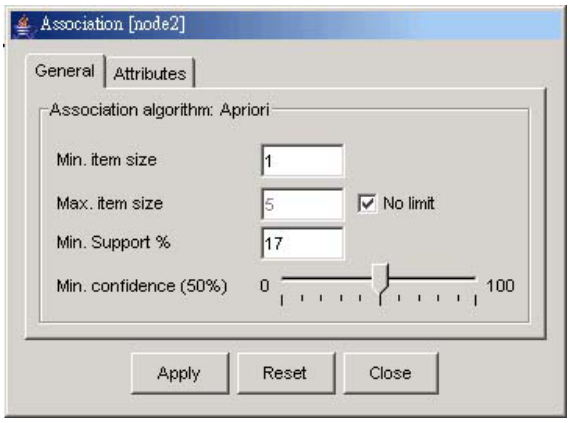
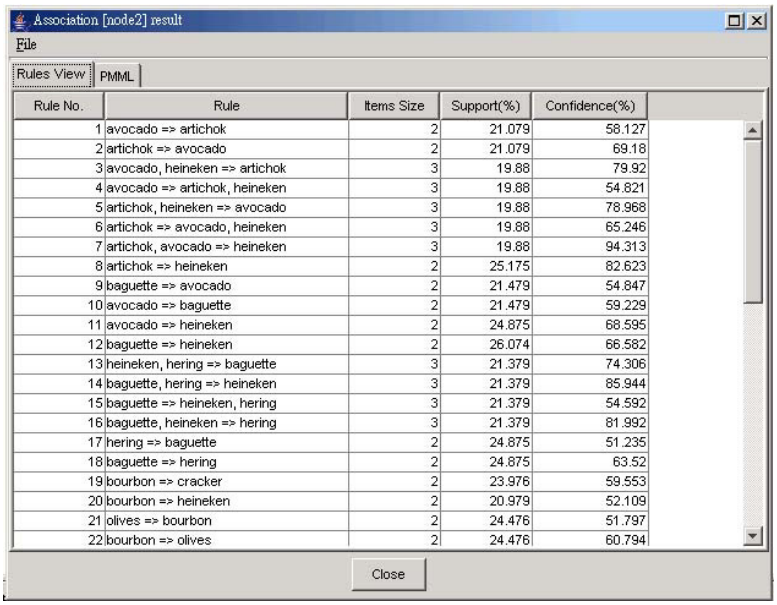


Figure 1-9 Association



Rule No.	Rule	Items Size	Support(%)	Confidence(%)
1	avocado => artichok	2	21.079	58.127
2	artichok => avocado	2	21.079	69.18
3	avocado, heineken => artichok	3	19.88	79.92
4	avocado => artichok, heineken	3	19.88	54.821
5	artichok, heineken => avocado	3	19.88	78.968
6	artichok => avocado, heineken	3	19.88	65.246
7	artichok, avocado => heineken	3	19.88	94.313
8	artichok => heineken	2	25.175	82.623
9	baguette => avocado	2	21.479	54.847
10	avocado => baguette	2	21.479	59.229
11	avocado => heineken	2	24.875	68.595
12	baguette => heineken	2	26.074	66.582
13	heineken, hering => baguette	3	21.379	74.306
14	baguette, hering => heineken	3	21.379	85.944
15	baguette => heineken, hering	3	21.379	54.592
16	baguette, heineken => hering	3	21.379	81.992
17	hering => baguette	2	24.875	51.235
18	baguette => hering	2	24.875	63.52
19	bourbon => cracker	2	23.976	59.553
20	bourbon => heineken	2	20.979	52.109
21	olives => bourbon	2	24.476	51.797
22	bourbon => olives	2	24.476	60.794

Figure 1-10 Rules generated by Association

Decision Tree (see *Figure 1-11 and Figure 1-12*) provides prediction and classification for a subject of interest. Rule sets produced provide insights on how decisions are made using the model. AlphaMiner Decision Tree provides different settings, such as pruning, minimum leaf node size, for building the tree.

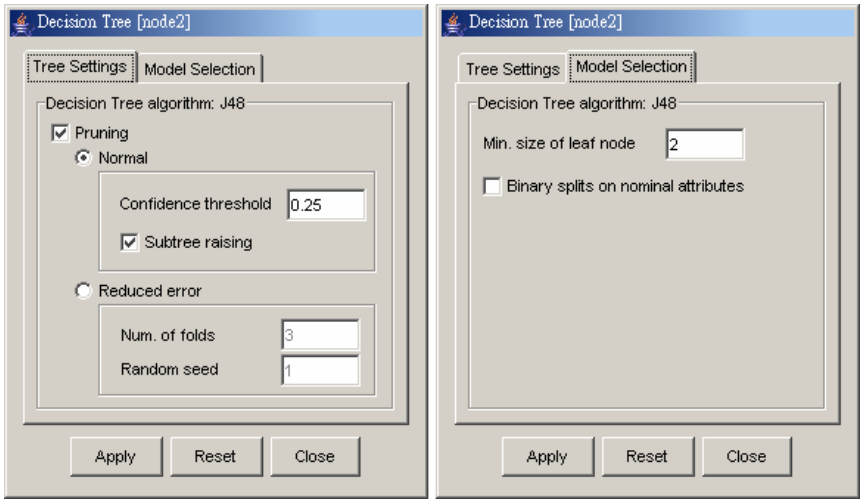
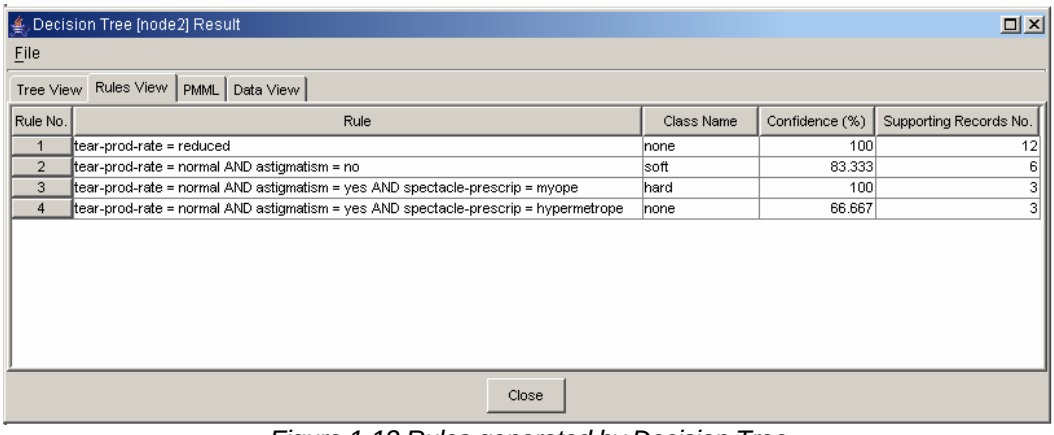


Figure 1-11 Decision Tree - Settings



Rule No.	Rule	Class Name	Confidence (%)	Supporting Records No.
1	tear-prod-rate = reduced	none	100	12
2	tear-prod-rate = normal AND astigmatism = no	soft	83.333	6
3	tear-prod-rate = normal AND astigmatism = yes AND spectacle-prescrip = myope	hard	100	3
4	tear-prod-rate = normal AND astigmatism = yes AND spectacle-prescrip = hypermetrope	none	66.667	3

Figure 1-12 Rules generated by Decision Tree

1.6 Assessment

Assessment offers two methods to evaluate the accuracy and performance of decision tree model or logistic regression model (see *Table 1-5*).

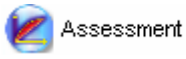
Operator name	Description
	Assesses models by two measurements (Confusion Matrix and Evaluation chart) to evaluate the accuracy and performance of models built by Decision Tree and Logistic Regression operators

Table 1-5 Assessment operator

Assessment (see *Figure 1-13*) is very useful for comparing and evaluating decision tree and logistic regression models built from the same training dataset. The process of assessment involves applying the model against a testing dataset and comparing predicted values with observed values of the dataset. Confusion matrix and evaluation chart are provided to evaluate the accuracy and performance of models.

- Confusion Matrix (see *Figure 1-13*)

  The confusion matrix represents a cross tabulation of the actual and predicted values, based on the principle that identifies the nature of the errors as well as their quantity. This enables users to evaluate the performance of a model in terms of the relative severity of misclassifications. In addition, the figures for correct classifications and misclassifications are evaluated using metrics such as: Recall, Precision, F-Measure, TP Rate, and FP Rate.

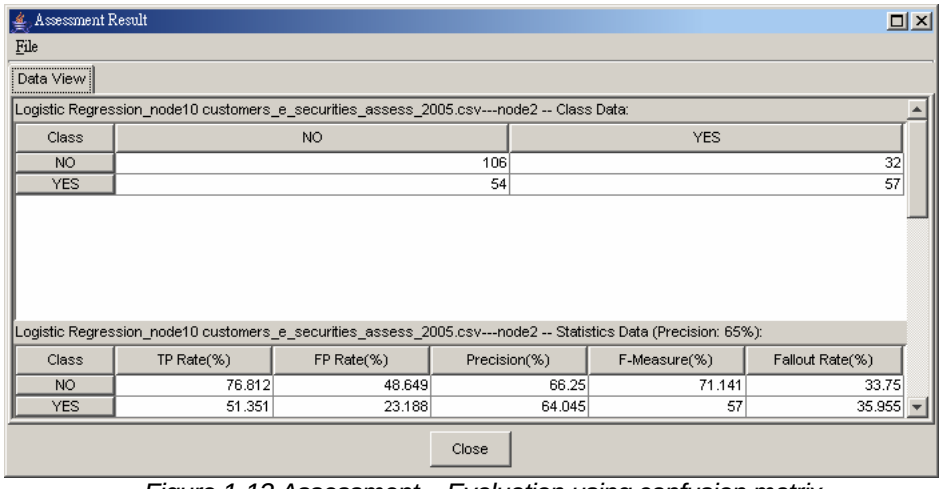


Figure 1-13 Assessment – Evaluation using confusion matrix

- Evaluation Chart (see *Figure 1-14*)

  Cumulative Gain Chart and Lift Chart are provided to assess models wherein there are only two output levels in target variables: a positive response and a negative response. Each chart uses the predicted values from a model to compute lift measurements, which is a way of measuring the model's performance using a completely random approach.

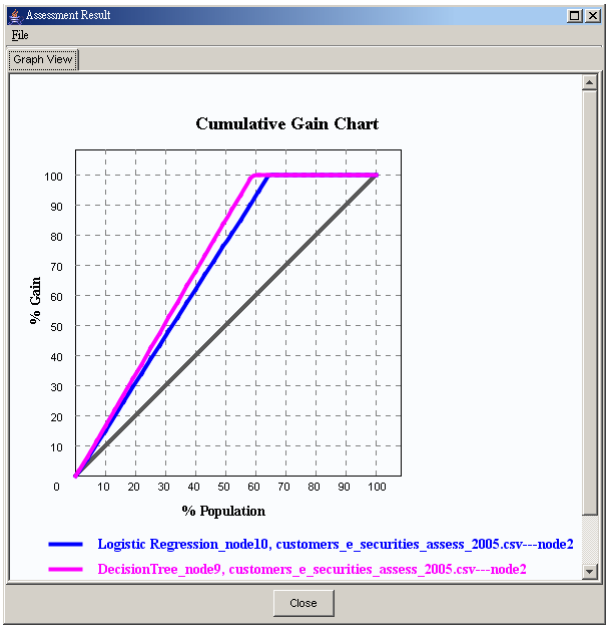


Figure 1-14 Assessment – Cumulative Gain Chart

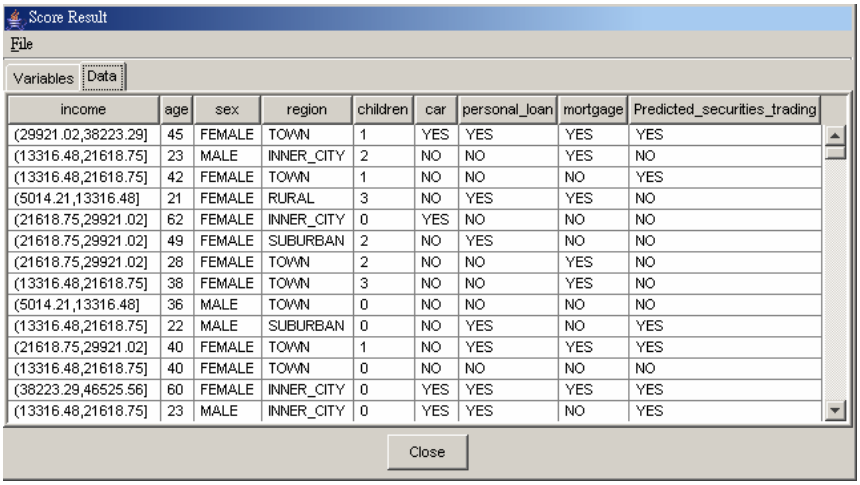
1.7 Deployment

The following deployment operation is implemented in the AlphaMiner system (see *Table 1-6*):

Operator Names	Description
Score	Provides prediction and classification for new dataset based on constructed Decision Tree or Logistic Regression model

Table 1-6 Deployment Operator List

Score (see *Figure 1-15*) performs prediction or classification for a newly external dataset. The dataset does not contain a target variable, since it is the attribute that will be generated in the score process, whose values are predicted or classified according to the constructed models.



income	age	sex	region	children	car	personal_loan	mortgage	Predicted_securities_trading
(29921.02,38223.29]	45	FEMALE	TOWN	1	YES	YES	YES	YES
(13316.48,21618.75]	23	MALE	INNER_CITY	2	NO	NO	YES	NO
(13316.48,21618.75]	42	FEMALE	TOWN	1	NO	NO	NO	YES
(5014.21,13316.48]	21	FEMALE	RURAL	3	NO	YES	YES	NO
(21618.75,29921.02]	62	FEMALE	INNER_CITY	0	YES	NO	NO	NO
(21618.75,29921.02]	49	FEMALE	SUBURBAN	2	NO	YES	NO	NO
(21618.75,29921.02]	28	FEMALE	TOWN	2	NO	NO	YES	NO
(13316.48,21618.75]	38	FEMALE	TOWN	3	NO	NO	YES	NO
(5014.21,13316.48]	36	MALE	TOWN	0	NO	NO	NO	NO
(13316.48,21618.75]	22	MALE	SUBURBAN	0	NO	YES	NO	YES
(21618.75,29921.02]	40	FEMALE	TOWN	1	NO	YES	YES	YES
(13316.48,21618.75]	40	FEMALE	TOWN	0	NO	NO	NO	NO
(38223.29,46525.56]	60	FEMALE	INNER_CITY	0	YES	YES	YES	YES
(13316.48,21618.75]	23	MALE	INNER_CITY	0	YES	YES	NO	YES

Figure 1-15 Score – scored dataset