Intelligent Decision Support Systems

(Case Study 1 – CUSTOMER RELATIONSHIP MANAGEMENT [CRM] / LOYALTY ANALYSIS)

Miquel Sànchez i Marrè

miquel@cs.upc.edu

http://www.cs.upc.edu/~miquel

Course 2016/2017

https://kemlg.upc.edu

Knowledge Engineering and Machine Learning Group
UNIVERSITAT POLITÈCNICA DE CATALUNYA

CASE STUDY 1 – CUSTOMER RELATIONSHIP MANAGEMENT (CRM) / LOYALTY ANALYSIS

Extracted and adapted from

P. Giudici. *Applied Data Mining*, John Wiley, 2003.

https://kemlg.upc.edu

Knowledge Engineering and Machine Learning Group
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Contents

- Problem Analysis
- IDSS Goal definition
- IDSS Development steps
  - Data Description and Data Management
  - Data Analysis
  - Model Selection
  - Model Building
  - Model Validation and Model Comparison
- Conclusions

Problem Analysis

- Italian company selling products bought by mailing system
- *Study of the buying behaviour* of the company customers, and *discovering of factors* which make a customer to be *loyal* or to be an *ocasional* buyer
- Priority issues of companies:
  - Get the most of their *loyal customers*
  - Need to characterize and distinguish the *loyal customers* from *ocasional customers* to focus their marketing efforts on the right audience (*loyal customers*).
- How to *minimize* the costs (marketing, etc.), to get a *maximum* benefit

IDSS Goal Definition

- Main objective: to classify customers into homogeneous groups, which characterize different objective profiles
- To solve the problem of the company, it seems that a series of sub-objectives must be met:
  - **Characterization of the relevant information** related to customer *loyalty*
  - **Identification of loyal customers** using different models and comparing them
  - **Obtaining one/several predictive model/s** allowing the company to easily differentiate new (and old) customers as loyal or not loyal

Data Management (1)

DATA COLLECTION AND PLANNING

- Product sale data ordered by mail in Italy
- Customers in the DB between 1992 and 1996
  - 210.085 customers
- Stratified sample by time intervals
  - 2470 customers

DATA DEPURATION AND FILTERING

- 3 Databases: customers, buying orders in local agencies, buying orders at the central agency
  - Different record structure and type
- Building up an specific DB for marketing (datamart)

Data Management (2)

- Marketing status
- Client active?
- Client in debt?
- Total number of orders
- Date of first order
- Date of last order
- Total amount ordered
- Total amount paid
- Current balance
- Payments delayed?
- Time lag between 1st & 2nd order
- Amount of current instalment
- Residual number of instalments
- Dimension of the shop
- Age
- Area of residence
- Sex
- 1st payment with instalments?
- First amount spent
- Number of products at 1st order

Exploratory Data Analysis (1)

- Statistical Descriptive Analysis
- New Variable Creation (response variables)
  - Y ≡ “Loyalty of a customer”
    - Y = 0, Number of orders equal to 1 ≡ “ocasional”
    - Y = 1, Number of orders higher than 1 ≡ “loyal”
  - Other possible variables:
    - Total amount paid
    - Number of products at first order
- Distribution of the response variable:

Modality	Absolute Frequency	Relative Frequency (%)
Y = 0	1457	59,71
Y = 1	1013	40,29

Exploratory Data Analysis (2)

- More than 19 observations have a *missing value for the* variable Y
  - Treatment: remove them
- Possible explicative variables
  - Behaviour variables related to the 1st contact (1st order)
  - Sociodemographic variables
- Just a few missing values in the explicative variables
  - Treatment: mean/median/mode substitution

Exploratory Data Analysis (3)

- Conditional distribution of sociodemographic variables regarding variable Y:

Sex	Y = 0	Y = 1
Female	61,04%	38,96%
Male	57,88%	42,12%

No differences

Area	Y = 0	Y = 1
North	55,40%	44,60%
Center	58,22%	41,78%
South	62,73%	37,27%

% Loyal customers decrease from North to South

Exploratory Data Analysis (4)

- Conditional distribution of sociodemographic variables regarding variable Y:

Age	Y = 0	Y = 1
15-35	68,80%	31,20%
36-50	53,44%	46,56%
51-89	60,42%	39,58%

% Loyal customers increase with age

Dimension	Y = 0	Y = 1
Small (<15)	60,39%	39,61%
Medium (≥15 & <30)	56,95%	43,05%
Large (≥30 & <60)	62,11%	37,89%

% Loyal customers decrease in large agencies

Exploratory Data Analysis (5)

- Contingency Table of variables *instalment* and Y:

	Instal. = 0	Instal. = 1	Total
Y = 0	1239 50,16% 85,04%F 68,04%C	218 8,83% 14,96%F 33,59%C	1457 58,99%
Y = 1	582 23,56% 57,45%F 31,96%C	431 17,45% 42,55%F 66,41%C	1013 41,01%
Total	1821 73,72%	649 26,28%	2470 100%

Customer which make payments with instalments is very probable to be a loyal customer

Exploratory Data Analysis (6)

- Variables Boxplots:
  - First amount spent
  - Number of products at 1st order

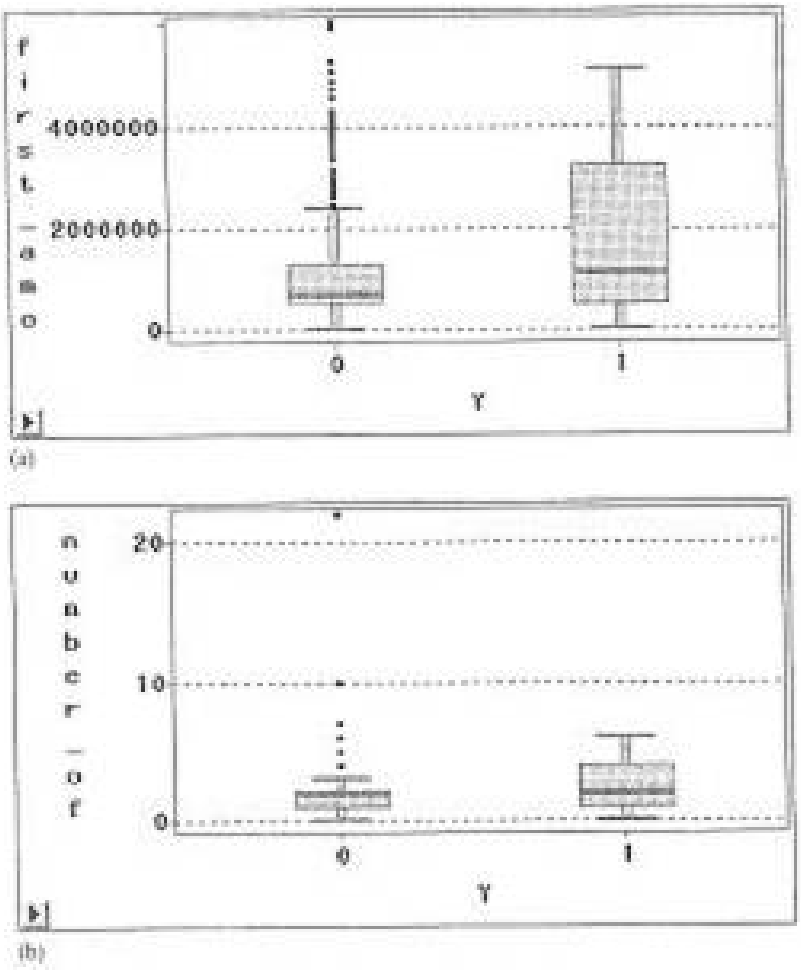


Figure 10.1 Conditional distribution of (a) the amount spent and (b) the number of products with respect to the levels of Y.

Exploratory Data Analysis (7)

- Variable/Attribute Relevance
  - Bivariant charts
  - Contingency Tables
  - Boxplots
  - Feature Selection and Feature Weighting
- Variable selection

Planning and Model Selection (1)

- Data Transformations
  - Binarization of qualitative modalities of variables *Age*, *Dimension*, *Area* => 9 binary variables
  - Variable *Sex*, is already binary
  - Variable *Total amount* and *Number of orders* are quantitatives
- Hypotheses
  - 12 explicative variables and one binary response variable (Y ≡ customer loyalty)

Planning and Model Selection (2)

- Data matrix:

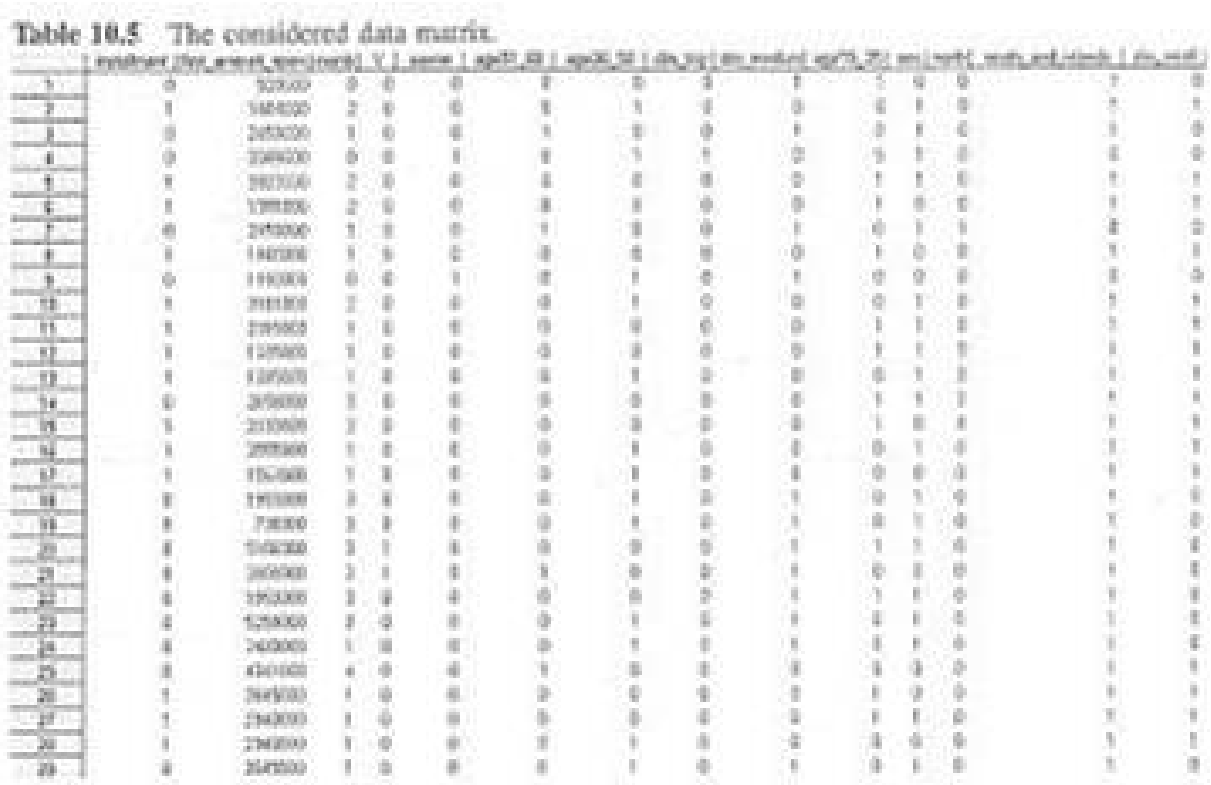


Table 10.5 The considered data matrix.

Planning and Model Selection (3)

- Models
  - Logistic Regression Model
  - Connexionist Model (Artifical Neural Networks, ANNs)
  - Decision Tree Model
  - Case-Based Reasoning Model (CBR)

Logistic Regression

- Results:

Table 10.6 The selected logistic regression model.

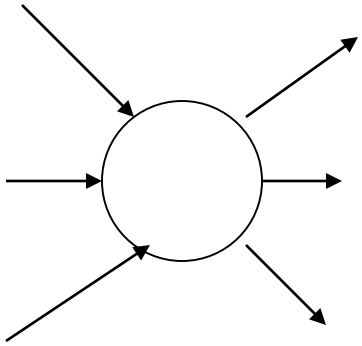
	Estimates	Stderr	Wald	Pr>Chi-square	Odds ratio
Intercept	0.3028	0.1248	108.93	<.0001	.
age15_35	-0.5440	0.1367	15.84	<.0001	0.580
installment	1.6107	0.1371	137.98	<.0001	5.006
number_of_products	0.3043	0.0465	42.78	<.0001	1.356

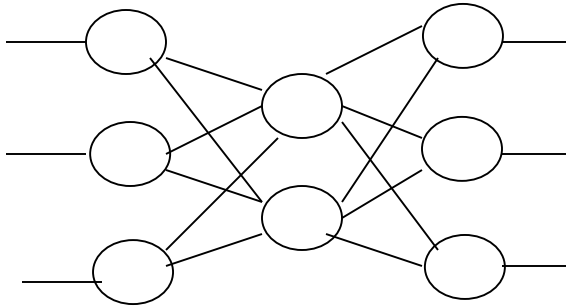
- Model: $p(Y=1) \equiv t + t_a*A + t_b*B + t_c*C$ is significative

- A customer will be "valuable" $\Leftrightarrow$

$$p(Y=1) > 0{,}5 \Leftrightarrow t + t_a*A + t_b*B + t_c*C > 0$$

Connexionist Models (ANN, RBF)





- Training step
- Decodification/classifying step
- Perceptrons, Backpropagation and Kohonen Maps (SOMs).

Radial Basis Function Network (1)

- Network Description:
  - One RBF with a hidden node
  - 13 explicative variables ≡ 13 input nodes
  - Input combination function: a Gaussian Radial Function with equal heights and equal weights
  - Activation function for hidden *node* is the *identity function*
  - Activation function for the *output node* is the *softmax* function (normalized output of Y probability)
- Network parameters trained by means of minimization of error rate in the classification process

Radial Basis Function Network (2)

- Error rate evolution in the classification
  - 7 iterations makes the error stable

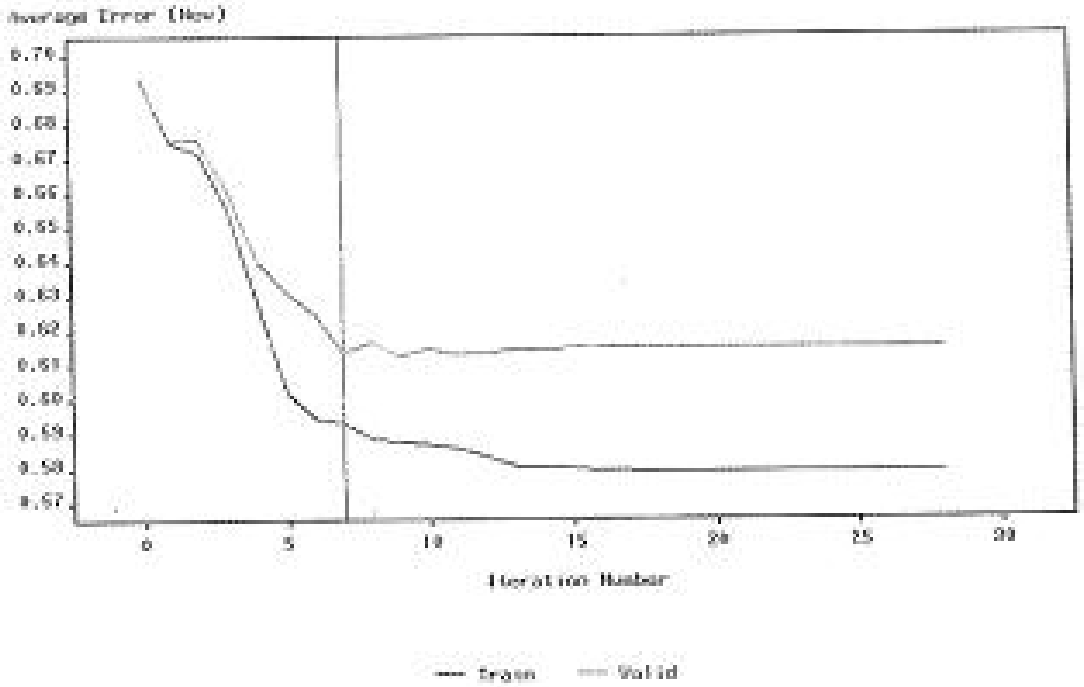
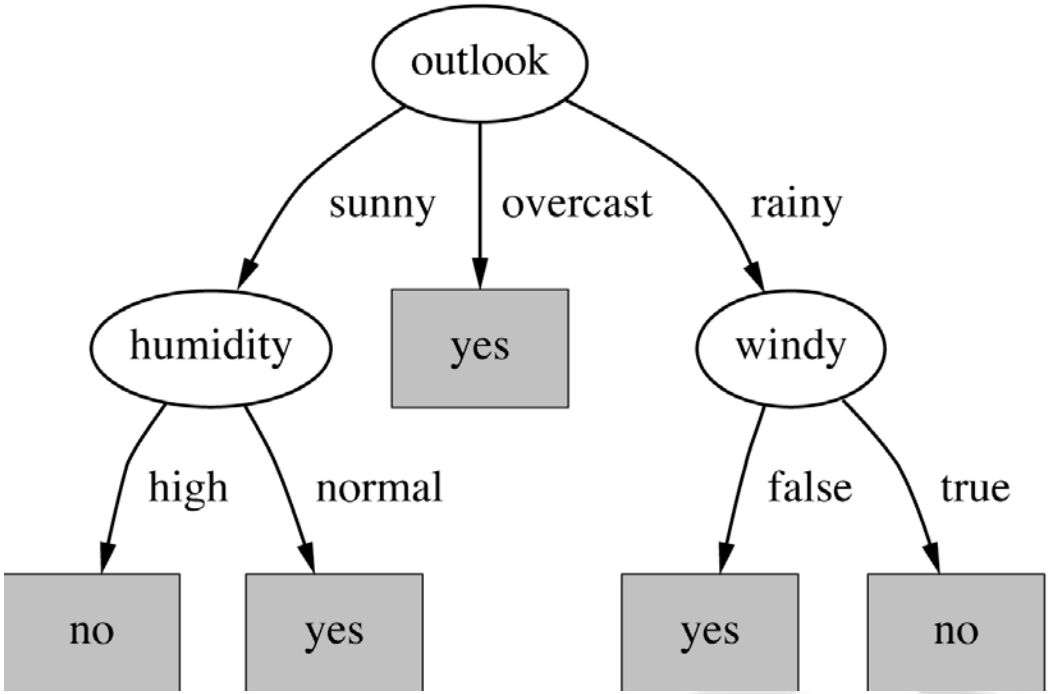


Figure 10.2 Evolution of the misclassification rate for the RBF network.

- Adjusted Weights with higher values:
  - Age15_35
  - Instalment
  - Number_of_products

Decision Tree Models



Decision tree for the weather data.

Decision Tree Models (1)

- Prediction of Y value according to the explicative variables, through a discriminant process
- Methods
  - CART
  - ID3
  - C4.5
  - ...
- Used Methods
  - CART with entropy criterium
  - CART with Gini's impurity criterium (best model)

Decision Tree Models (2)

- Classification accuracy versus number of tree leaves:

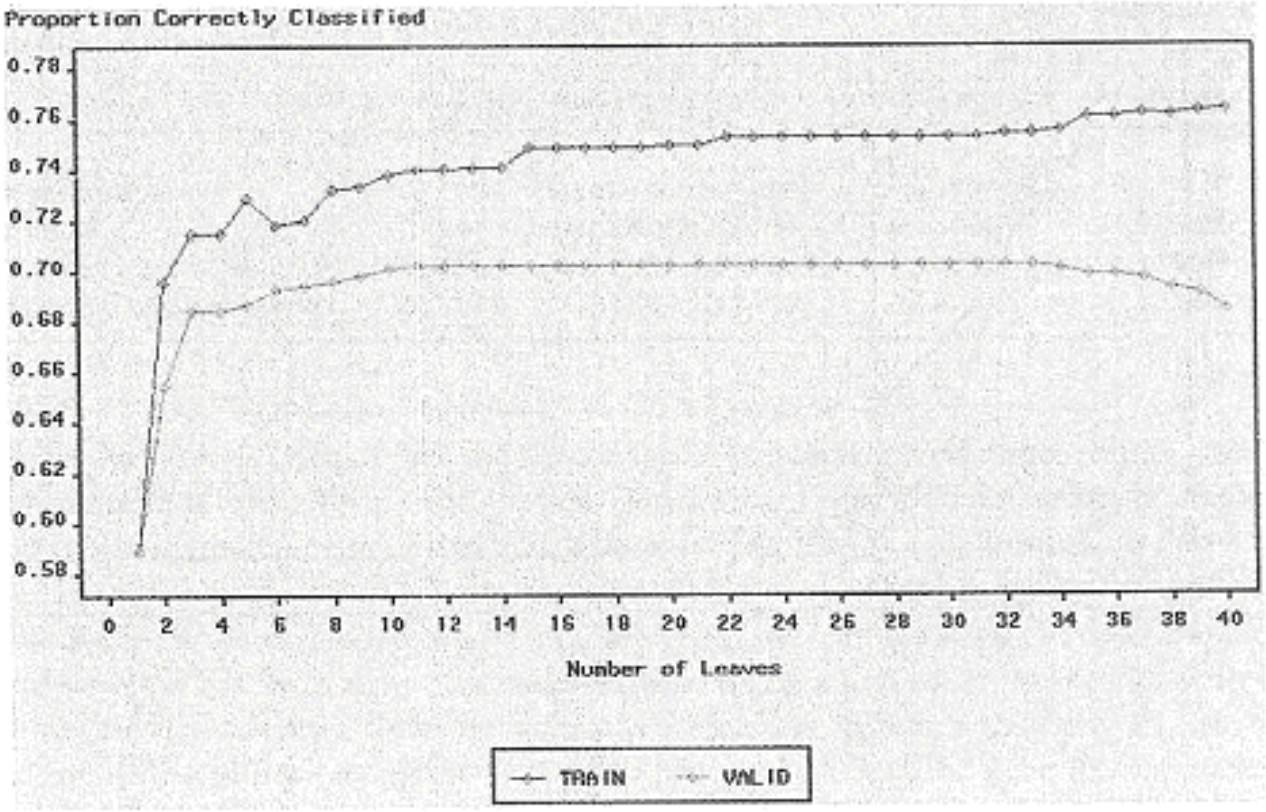


Figure 10.3 Evolution of the classification accuracy for the classification tree as the number of leaves increases.

Decision Tree Models (3)

Table 10.7 The rules for the classification tree.

```
IF    2659000 <=FIRST_AMOUNT_SPENT
AND INSTALMENT EQUALS 0
THEN
 N : 226
 1 : 56.2%
```

- Classification Rules extracted from the Decision Tree:

Table 10.7 *(continued)*

```
IF    2659000 <=FIRST_AMOUNT_SPENT
 0 : 43.8%

IF FIRST_AMOUNT_SPENT <   515000
AND INSTALMENT EQUALS 1
THEN
 N : 55
 1 : 89.1%
 0 : 10.9%

IF  375000 <=FIRST_AMOUNT_SPENT <   2659000
AND INSTALMENT EQUALS 0
THEN
 N : 709
 1 : 18.6%
 0 : 81.4%

IF NORTH EQUALS 0
AND NUMBER_OF_PRODUCTS < 2.5
AND    515000 <=FIRST_AMOUNT_SPENT
AND INSTALMENT EQUALS 1
THEN
 N : 99
 1 : 47.5%
 0 : 52.5%

IF NORTH EQUALS 1
AND NUMBER_OF_PRODUCTS < 2.5
AND    515000 <=FIRST_AMOUNT_SPENT
AND INSTALMENT EQUALS 1
THEN
 N : 42
 1 : 73.8%
 0 : 26.2%

IF    2.5 <=NUMBER_OF_PRODUCTS < 5.5
AND    515000 <=FIRST_AMOUNT_SPENT
AND INSTALMENT EQUALS 1
THEN
 N : 178
 1 : 78.7%
 0 : 21.3%

IF    5.5 <=NUMBER_OF_PRODUCTS
AND  515000 <=FIRST_AMOUNT_SPENT
AND INSTALMENT EQUALS 1
THEN
 N : 3
```

(continued overleaf)

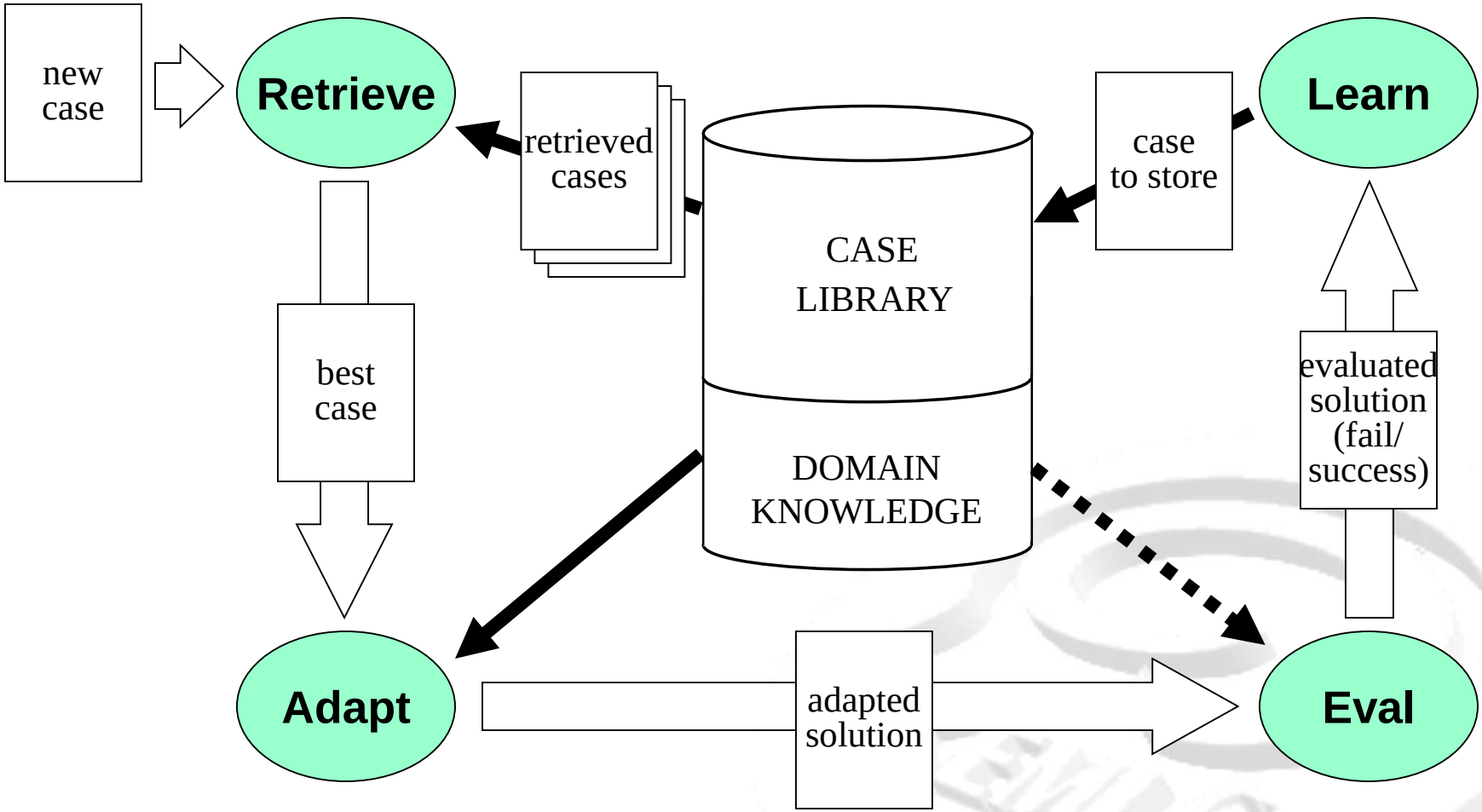
Table 10.7 *(continued)*

```
IF    2659000 <=FIRST_AMOUNT_SPENT
 1 : 0.0%
 0 : 100.0%

IF FIRST_AMOUNT_SPENT < 105000
AND NORTH EQUALS 1
AND INSTALMENT EQUALS 0
THEN
 N : 7
 1 : 0.0%
 0 : 100.0%

IF    105000 <=FIRST_AMOUNT_SPENT < 375000
AND NORTH EQUALS 1
AND INSTALMENT EQUALS 0
THEN
 N : 59
 1 : 72.9%
 0 : 27.1%

IF AGE36_50 EQUALS 1
AND NORTH EQUALS 0
AND FIRST_AMOUNT_SPENT < 375000
AND INSTALMENT EQUALS 0
THEN
 N : 47
 1 : 25.5%
 0 : 74.5%

IF AGE36_50 EQUALS 0
AND NORTH EQUALS 0
AND FIRST_AMOUNT_SPENT < 375000
AND INSTALMENT EQUALS 0
THEN
 N : 40
 1 : 52.5%
 0 : 47.5%
```

Decision Tree Models (4)

- Discriminant Variables used:
  - Age15_35
  - Instalment
  - Number_of_products
  - First amount spent
  - Geographic area

CBR Models



K-NN classification model

- Simplification of the general CBR model
- Only similar instances are retrieved (k)
- Misclassification rate table according to number k :

K	Misclassification rate
732	0,41
100	0,328
10	0,316

Model Validation (1)

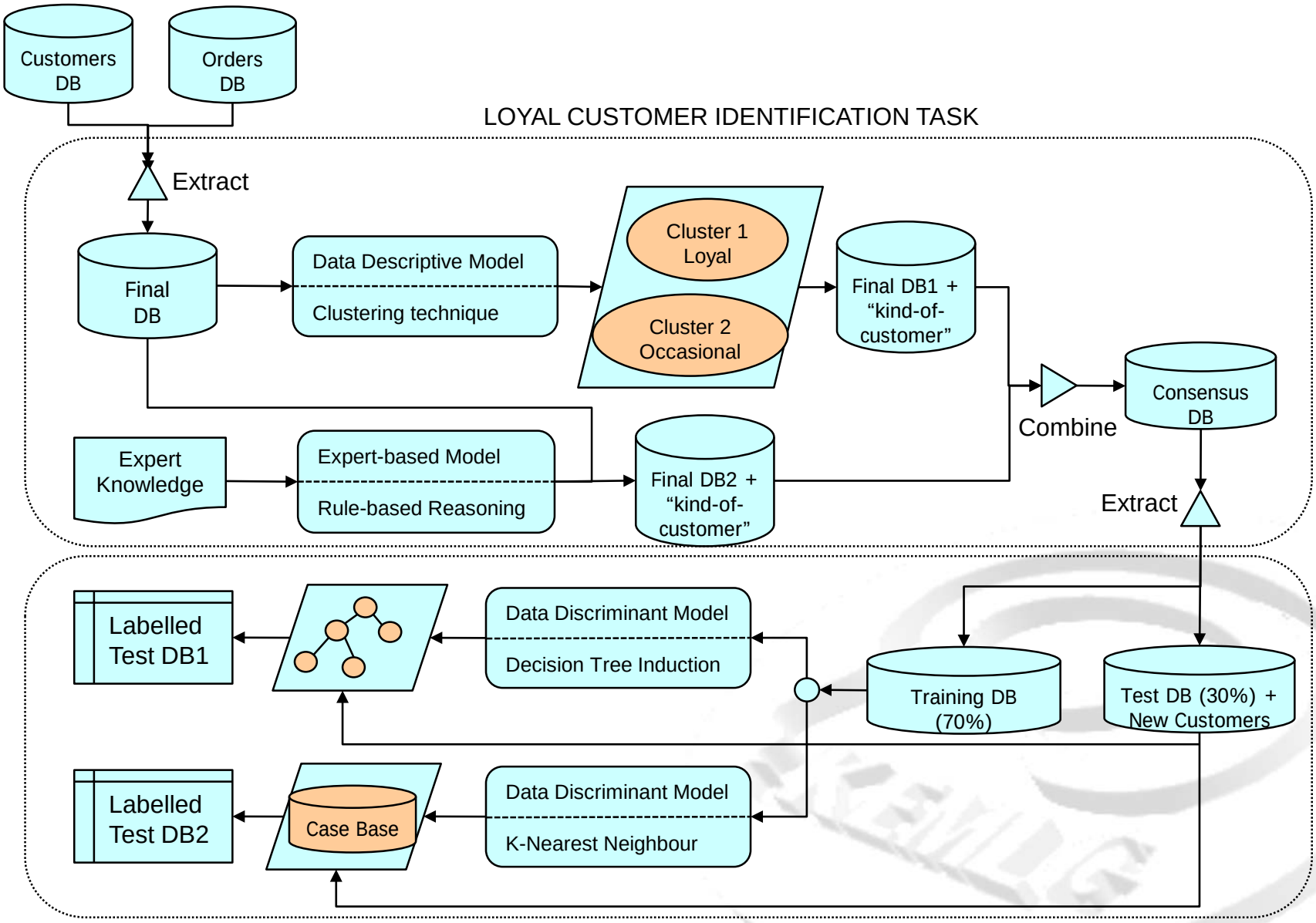
GENERAL

- Precision and fiability of obtained models
- Scalability/Generalization of the system
- Interpretability, flexibility and user friendly system

PARTICULAR

- Validations (startified)
  - Crossed
- Confusion Matrices
  - On validation set
- Misclassification Rate Table
- ROC Curves
- Gini Index

Functional Architecture of the IDSS



Model Validation (2)

- Confusion Matrix

LOGISTIC REGRESSION		Predicted	
		0	1
Observed	0	48,02	10,91
	1	22,92	18,14

Error Type II
False Positives

Error Type I
False Negatives

CART CLASSIFICATION TREE		Predicted	
		0	1
Observed	0	43,52	15,42
	1	14,32	26,74

Model Validation (3)

- Confusion Matrix

RBF neural network		Predicted	
		0	1
Observed	0	47,34	11,60
	1	20,87	20,19

K-NN MBR		Predicted	
		0	1
Observed	0	41,34	17,60
	1	12,14	28,92

Model Validation (4)

- Misclassification Rate Table

MODEL	MISCLAS. RATE	VALIDATION MISCLAS. RATE	TEST MISCLAS. RATE
CART Tree	0,2593856655	0,2974079127	0,2909836066
K-NN MBR	0,2894197952	0,2974079127	0,3155737705
Regression	0,3071672355	0,3383356071	0,3770491803
RBF	0,3051194539	0,3246930423	0,3360655738

Model Validation (5)

- ROC Curves
  - *Sensitivity* (1 – probability(error type I)) versus *1- specificity* (probability(error type II):

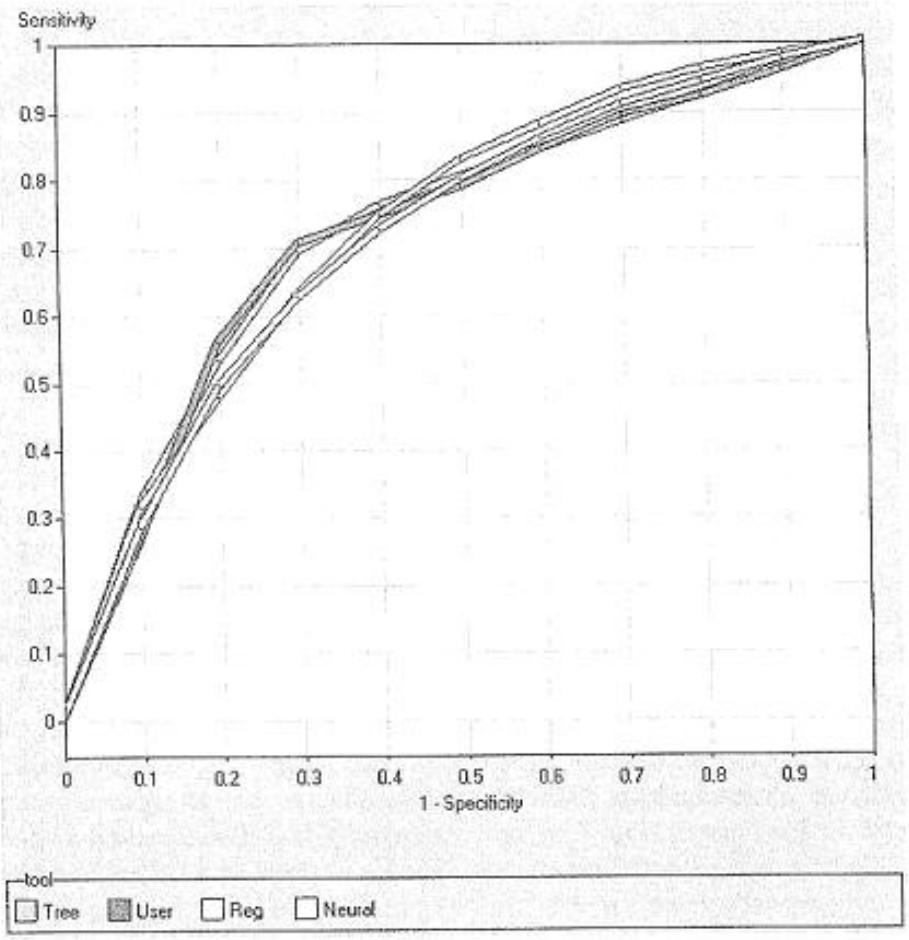


Figure 10.5 ROC curves for the considered models. The curve called user is the MBR model

Model Validation (6)

- Gini Index of performance
  - Area between the ROC curve and the 45º bisector

	Logistic Regression	RBF	CART Tree	K-NN MBR
Gini Index	0,4375	0,4230	0,4445	0,5673

Conclusions

- Goal: To distinguish two types of customers: loyal and ocasional
  - Y variable formulation
- Data treatment
  - Creation of a unique DB (datamart)
  - Missing values management
- Exploratory Analysis based bivariant analysis between Y and other variables
  - Variable selection
- Models selection
  - Discriminant Models
- Models used
  - Logistic Regression
  - CART
  - K-NN MBR
  - RBF
- Model Comparison
  - CART and K-NN seem to be the best ones
- Model Interpretability

Miquel Sànchez i Marrè
(miquel@cs.upc.edu)

http://kemlg.upc.edu/