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

Extracted and adapted from

https://kemlg.upc.edu
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 ≡ “occasional”
    - 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:

<table>
<thead>
<tr>
<th>Modality</th>
<th>Absolute Frequency</th>
<th>Relative Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y = 0</td>
<td>1457</td>
<td>59,71</td>
</tr>
<tr>
<td>Y = 1</td>
<td>1013</td>
<td>40,29</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Sex</th>
<th>Y = 0</th>
<th>Y = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>61,04%</td>
<td>38,96%</td>
</tr>
<tr>
<td>Male</td>
<td>57,88%</td>
<td>42,12%</td>
</tr>
</tbody>
</table>

No differences

<table>
<thead>
<tr>
<th>Area</th>
<th>Y = 0</th>
<th>Y = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>55,40%</td>
<td>44,60%</td>
</tr>
<tr>
<td>Center</td>
<td>58,22%</td>
<td>41,78%</td>
</tr>
<tr>
<td>South</td>
<td>62,73%</td>
<td>37,27%</td>
</tr>
</tbody>
</table>

% Loyal customers decrease from North to South
### Exploratory Data Analysis (4)

- Conditional distribution of sociodemographic variables regarding variable Y:

<table>
<thead>
<tr>
<th>Age</th>
<th>Y = 0</th>
<th>Y = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-35</td>
<td>68.80%</td>
<td>31.20%</td>
</tr>
<tr>
<td>36-50</td>
<td>53.44%</td>
<td>46.56%</td>
</tr>
<tr>
<td>51-89</td>
<td>60.42%</td>
<td>39.58%</td>
</tr>
</tbody>
</table>

% Loyal customers increase with age

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Y = 0</th>
<th>Y = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (&lt;15)</td>
<td>60.39%</td>
<td>39.61%</td>
</tr>
<tr>
<td>Medium (≥15 &amp; &lt;30)</td>
<td>56.95%</td>
<td>43.05%</td>
</tr>
<tr>
<td>Large (≥30 &amp; &lt;60)</td>
<td>62.11%</td>
<td>37.89%</td>
</tr>
</tbody>
</table>

% Loyal customers decrease in large agencies
Exploratory Data Analysis (5)

- Contingency Table of variables *instalment* and Y:

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

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
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:

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

- A customer will be “valuable” \( \iff \)
  \[ p(Y=1) > 0.5 \iff t + t_a \cdot A + t_b \cdot B + t_c \cdot C > 0 \]
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

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

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

![Proportion Correctly Classified](image)

**Figure 10.3** Evolution of the classification accuracy for the classification tree as the number of leaves increases.
Decision Tree Models (3)

Classification Rules extracted from the Decision Tree:

<table>
<thead>
<tr>
<th>IF</th>
<th>2659000 &lt;= FIRST_AMOUNT_SPENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43.8%</td>
</tr>
<tr>
<td>IF FIRST_AMOUNT_SPENT &lt; 515000 AND INSTALMENT_EQUALS 1</td>
<td></td>
</tr>
<tr>
<td>THEN</td>
<td>N : 55</td>
</tr>
<tr>
<td>1</td>
<td>89.1%</td>
</tr>
<tr>
<td>0</td>
<td>10.9%</td>
</tr>
<tr>
<td>IF 375000 &lt;= FIRST_AMOUNT_SPENT &lt; 2659000 AND INSTALMENT_EQUALS 0</td>
<td></td>
</tr>
<tr>
<td>THEN</td>
<td>N : 709</td>
</tr>
<tr>
<td>1</td>
<td>18.6%</td>
</tr>
<tr>
<td>0</td>
<td>81.4%</td>
</tr>
<tr>
<td>IF NORTH_EQUALS 0 AND NUMBER_OF_PRODUCTS &lt; 2.5 AND 515000 &lt;= FIRST_AMOUNT_SPENT AND INSTALMENT_EQUALS 1</td>
<td></td>
</tr>
<tr>
<td>THEN</td>
<td>N : 99</td>
</tr>
<tr>
<td>1</td>
<td>47.5%</td>
</tr>
<tr>
<td>0</td>
<td>52.5%</td>
</tr>
<tr>
<td>IF NORTH_EQUALS 1 AND NUMBER_OF_PRODUCTS &lt; 2.5 AND 515000 &lt;= FIRST_AMOUNT_SPENT AND INSTALMENT_EQUALS 1</td>
<td></td>
</tr>
<tr>
<td>THEN</td>
<td>N : 42</td>
</tr>
<tr>
<td>1</td>
<td>73.8%</td>
</tr>
<tr>
<td>0</td>
<td>26.2%</td>
</tr>
<tr>
<td>IF 2.5 &lt;= NUMBER_OF_PRODUCTS &lt; 5.5 AND 515000 &lt;= FIRST_AMOUNT_SPENT AND INSTALMENT_EQUALS 1</td>
<td></td>
</tr>
<tr>
<td>THEN</td>
<td>N : 178</td>
</tr>
<tr>
<td>1</td>
<td>78.7%</td>
</tr>
<tr>
<td>0</td>
<td>21.3%</td>
</tr>
<tr>
<td>IF 5.5 &lt;= NUMBER_OF_PRODUCTS AND 515000 &lt;= FIRST_AMOUNT_SPENT AND INSTALMENT_EQUALS 1</td>
<td></td>
</tr>
<tr>
<td>THEN</td>
<td>N : 3</td>
</tr>
</tbody>
</table>
| (continued overleaf)
Decision Tree Models (4)

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

[Diagram of CBR Models]

1. **Retrieve**
   - new case
   - retrieved cases
   - best case

2. **Adapt**
   - adapted solution

3. **Eval**
   - evaluated solution (fail/success)

4. **Learn**
   - case to store

[CASE LIBRARY]

[DOMAIN KNOWLEDGE]
K-NN classification model

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

<table>
<thead>
<tr>
<th>K</th>
<th>Misclassification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>732</td>
<td>0.41</td>
</tr>
<tr>
<td>100</td>
<td>0.328</td>
</tr>
<tr>
<td>10</td>
<td>0.316</td>
</tr>
</tbody>
</table>
Model Validation (1)

GENERAL

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

PARTICULAR

- Validations (started)
  - Crossed
- Confusion Matrices
  - On validation set
- Misclassification Rate Table
- ROC Curves
- Gini Index
Functional Architecture of the IDSS

LOYAL CUSTOMER IDENTIFICATION TASK

- Extract
- Final DB
  - Data Descriptive Model
    - Clustering technique
  - Expert-based Model
    - Rule-based Reasoning
  - Expert Knowledge
- Cluster 1 Loyal
- Cluster 2 Occasional
- Final DB1 + “kind-of-customer”
- Combine
- Consensus DB

NEW CUSTOMERS’ PROGNOSIS TASK

- Extract
- Labelled Test DB1
  - Data Discriminatant Model
    - Decision Tree Induction
- Labelled Test DB2
  - Case Base
  - Data Discriminatant Model
    - K-Nearest Neighbour
- Training DB (70%)
- Test DB (30%) + New Customers
Model Validation (2)

- Confusion Matrix

<table>
<thead>
<tr>
<th>LOGISTIC REGRESSION</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>48,02</td>
</tr>
<tr>
<td>1</td>
<td>22,92</td>
</tr>
</tbody>
</table>

**Error Type I** False Negatives

**Error Type II** False Positives

<table>
<thead>
<tr>
<th>CART CLASSIFICATION TREE</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>43,52</td>
</tr>
<tr>
<td>1</td>
<td>14,32</td>
</tr>
</tbody>
</table>
Model Validation (3)

- Confusion Matrix

<table>
<thead>
<tr>
<th>RBF neural network</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Observed</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K-NN MBR</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Observed</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Model Validation (4)

- Misclassification Rate Table

<table>
<thead>
<tr>
<th>MODEL</th>
<th>MISCLAS. RATE</th>
<th>VALIDATION MISCLAS. RATE</th>
<th>TEST MISCLAS. RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART Tree</td>
<td>0.2593856655</td>
<td>0.2974079127</td>
<td>0.2909836066</td>
</tr>
<tr>
<td>K-NN MBR</td>
<td>0.2894197952</td>
<td>0.2974079127</td>
<td>0.3155737705</td>
</tr>
<tr>
<td>Regression</td>
<td>0.3071672355</td>
<td>0.3383356071</td>
<td>0.3770491803</td>
</tr>
<tr>
<td>RBF</td>
<td>0.3051194539</td>
<td>0.3246930423</td>
<td>0.3360655738</td>
</tr>
</tbody>
</table>
Model Validation (5)

- ROC Curves
  - **Sensitivity** \( (1 - \text{probability(error type I)}) \) versus \( 1 - \text{specificity} \) \( \text{(probability(error type II)} \):

![ROC Curves Diagram](image.png)
Model Validation (6)

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

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>RBF</th>
<th>CART Tree</th>
<th>K-NN MBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gini Index</td>
<td>0.4375</td>
<td>0.4230</td>
<td>0.4445</td>
<td>0.5673</td>
</tr>
</tbody>
</table>
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 Sàncchez i Marrè  
(miquel@cs.upc.edu)

http://kemlg.upc.edu/