Churn analysis

Introduction
Context

- Activities and services characterized by:
  - Continuous relationships between consumer/customer and provider
  - Possibility to trace user behavior
  - Competitors

- Business context
  - Telco
  - Supermarkets retail
Problem Definition

- Several synonyms
  - Churn
  - Abandon
  - Desertion

- Typically not announced
  - Explicit (Es.: contract termination)
  - Hidden (Es.: change of supermarket)
General goals

- **Customer retention**
  - Retain old client is less expensive than looking for new ones
  - Provide resilience to the service

- **Goal:**
  - identify the abandoner
  - Implement retention strategies; special offers, new services, special discounts
Challenge: FAST identification

- Often, when churn occurs it is too late
  - Recovering the churner is not possible any more, or
  - Recovering him is not convenient any more

- Foundamental: detect churning **immediately** or even **in advance**

- New problem formulation:
  - Churn Analysis = *churn prediction*
Modalities of Churn

- **Explicit**
  - Typical of customer relationships that involve a contract with costs or other involvement
    - E.g.: Telecom, when a rent is required
    - E.g.: non-free cards

- **Implicit**
  - Typical of customer relationships without contracts and/or free of charge
    - E.g.: Most loyalty cards
Implicit churn

- Most common situation in retail selling
  - Free loyalty cards
  - The churner simply stops using them

- Q.: how to understand whether a customer is actually a churner?
  - Stops purchasing for 1 month?
  - Stops purchasing for 1 year?
  - Visits the shops less than twice a month?
  - Spends less than 50% of what he used to?
“Soft” churn

- Alternative notion of churn
  - Switch from a kind of relation with the business to another one
  - Extreme case: from “loyal” to “complete churn”

- In retail selling
  - Loyal customers provide (some) guarantees about future income of the seller
  - Downgrading from “loyal” to “occasional” has large effects on the company
    - As important as the “hard” churn
Predicting churn

- Customer traces allow to reconstruct his history for a given period in the past

Monitored period (in the DW)

- The status of the customer can only be evaluated on the information we have “today”
Predicting churn

- Objective: predicting the future status of the customer, based on his recent history

- Recent history provides clues about the behaviour he is going to follow
  - Some clues help recognizing the future churners, others do not
  - Some clues are explicitly available in the dataset, others need to be inferred from it
Predicting churn

- How to learn *today* the correlations between present situation and future status?
  - Try to learn them looking at the past correlations
  - The correlations “past → today” learnt will be exploited to make predictions (today → tomorrow)

Monitored period (in the DW)
Schema of Churn prediction applications

- Define/extract working variables
  - Predictive: the *clues* available *today/past*
  - Target variable: *future/today* status

- Build the predictive model
  - Look for correlations between predictive and target variables, to be exploited in the prediction phase

- Apply predictive model
  - The correlations are applied to the present situation (i.e. predictive variables) to estimate the target variable
BICOOP – Churn Analisys
Problem Definition:

Estimate the probability of churn on the base of DW evidences:

- Detailed buying records
- Demographic data

Churn risk definition

- For a client the churn risk appears when a dramatic decrease of her/his expenditure measures:
  - Number of visits
  - Total amount of expenditure value
  - Number of items bought
Predictive analysis

• Collect historical data to build:
  - Demographic and purchase variables, to be used as predictors (green bar)
  - Target variables (red bar)

• Build a predictive model
  - Learned on historical data
  - To be applied for predictions
Si sono estratte dal data warehouse, per il periodo di 9 mesi (Dicembre 2006 – Agosto 2007) le seguenti informazioni:

• Dati anagrafici (sesto, età, professione etc.)
• Dati di spesa
  – Globale
  – Settori specifici: fresco, carni, pesce, ortofrutta
  – Pesata (abbattimento no-food)
• Trend di spesa:
  – Tipologia cliente (per ogni mese)
  – Regressione spese
  – Regressione spesa
  – Regressione battute

(Extract from COOP report)
Data preparation – target variables (red bar)

• Over the last 3 months, the following information were extracted from the data warehouse:
  - Number of purchases
  - Purchases variation w.r.t. “green bar”
    • Total amount spent
    • Number of articles bought
    • Number of visits
The final dataset contains a row for each customer, excluding those without any purchase in the period:
- 517,000 rows
- 47 attributes

<table>
<thead>
<tr>
<th>Predittori Anagrafici</th>
<th>Predittori di spesa</th>
<th>Predittori di trend</th>
<th>Variabili target</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIENTE_ID</td>
<td>DATA_ULTIMA_SPESA</td>
<td>TIPOLOGIA_01</td>
<td>T_NUM_SPESA</td>
</tr>
<tr>
<td>SESSO</td>
<td>NUM_SPESA</td>
<td>TIPOLOGIA_02</td>
<td>T_RAPP_SPESA</td>
</tr>
<tr>
<td>STATO_CIVILE</td>
<td>SPESA_TOT</td>
<td>TIPOLOGIA_03</td>
<td>T_RAPP_SPESA</td>
</tr>
<tr>
<td>PROFESSIONE</td>
<td>SPESA_TOT_PESATA</td>
<td>TIPOLOGIA_04</td>
<td>T_RAPP_BATTUTE</td>
</tr>
<tr>
<td>TITOLO_STUDIO</td>
<td>SPESA_MEDIA</td>
<td>TIPOLOGIA_05</td>
<td></td>
</tr>
<tr>
<td>PROVINCIA</td>
<td>SPESA_MEDIA_PESATA</td>
<td>TIPOLOGIA_06</td>
<td></td>
</tr>
<tr>
<td>REGIONE</td>
<td>BATTUTE</td>
<td>TIPOLOGIA_07</td>
<td></td>
</tr>
<tr>
<td>ANNO_SOCIO</td>
<td>FRESCHI_TOT</td>
<td>TIPOLOGIA_08</td>
<td></td>
</tr>
<tr>
<td>FASCIA_ANNO_SOCIO</td>
<td>FRESCHI_SPESE</td>
<td>TIPOLOGIA_09</td>
<td></td>
</tr>
<tr>
<td>FL_INVIO_RIVISTA</td>
<td>CARNI_TOT</td>
<td>TIPOLOGIA_MEDIA</td>
<td></td>
</tr>
<tr>
<td>COD_NEGOZIO</td>
<td>CARNI_SPESE</td>
<td>TIPOLOGIA_ZERI</td>
<td></td>
</tr>
<tr>
<td>ETA</td>
<td>PESCE_TOT</td>
<td>REGR_NUM_SPESA</td>
<td></td>
</tr>
<tr>
<td>ETA_FASCIA</td>
<td>PESCE_SPESE</td>
<td>REGR_SPESA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ORTOFRUTTA_TOT</td>
<td>REGR_SPESA_PESATA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ORTOFRUTTA_SPESE</td>
<td>REGR_BATTUTE</td>
<td></td>
</tr>
</tbody>
</table>
Data Exploration

- Distribution of expenditure trends

Customers with total purchases $> 400\text{€}$
Target variables

**Funzioni Obiettivo**

**NUM_SPESE**: spese del cliente nel periodo target

**RAPP_SPESE**: rapporto tra il numero delle spese del periodo target e quello del periodo d'osservazione

**RAPP_SPESA**: rapporto tra la spesa del periodo target e quella del periodo d'osservazione

**RAPP_BATTUTE**: rapporto fra le battute di cassa del periodo target e quelle del periodo d'osservazione
Discretizing target vars.

- Choose an alarm threshold
- Result: three churn binary variables
- Chosen threshold: 0.5 i.e. 50% decay
- Distributions obtained
  (F = low risk, V = high risk)

OB1: RAPP_SPESE

OB2: RAPP_SPESA

OB3: RAPP_BATTUTE
Discretizing target vars.

Combined variable:
- Alarm = alarm in all 3 variables seen before

**OB_AND: OB1 and OB2 and OB3**
Results

- Churn distribution w.r.t. expenditure & weighted expenditure
Sample classification rules

if REGIONE = TOSCANA
& NUM_SPESE <= 128
& TIPOLOGIA_01 = 7
& TIPOLOGIA_09 = 0
& TIPOLOGIA_ZERI > 2
& REGR_BATTUTE <= -0.98
then V (conf. 82.8%)

if DATA_ULTIMA_SPESA > 183
& NUM_SPESE <= 21
& TIPOLOGIA_ZERI > 1
& REGR_NUM_SPESE <= -0.02
& REGR_BATTUTE <= -0.98
then V (conf. 92%)
Performances of classifier

• Accuracy:
  - 81.06% on training set (70% of the dataset, 360,000 rows)
  - 80.94% on test set (30% of the dataset, 155,000 rows)

• Confusion matrices

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th></th>
<th>Test Set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>V</td>
<td>F</td>
<td>V</td>
</tr>
<tr>
<td>Real values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>256.608</td>
<td>17.920</td>
<td>110.029</td>
<td>7.767</td>
</tr>
<tr>
<td>V</td>
<td>50.540</td>
<td>36.466</td>
<td>21.855</td>
<td>15.734</td>
</tr>
</tbody>
</table>

Gain: 42.8%
Performances of classifier

• Lift chart