Mirco Nanni
KDD Lab,ISTI-CNR, Pisa
mirco.nanni@isti.cnr.it

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 fomulation:
  - 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

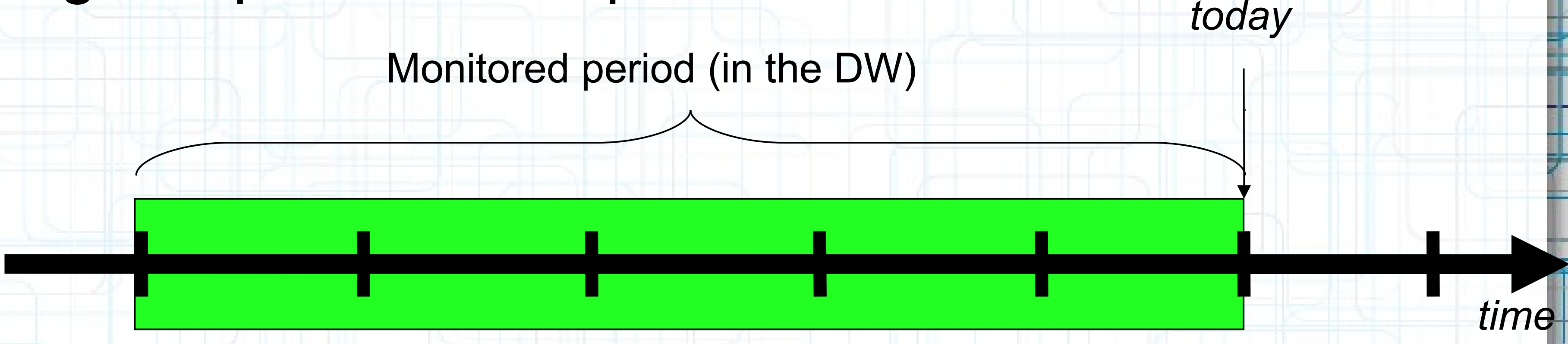
- Most common situation in retail selling
  - Free loyalty cards
  - The churner simply stops using them
- Q.: how to understand whether a customer is actually a churner?
  - Stops purchasing for 1 month?
  - Stops purchasing for 1 year?
  - Visits the shops less than twice a month?
  - Spends less than 50% of what he used to?

“Soft” churn

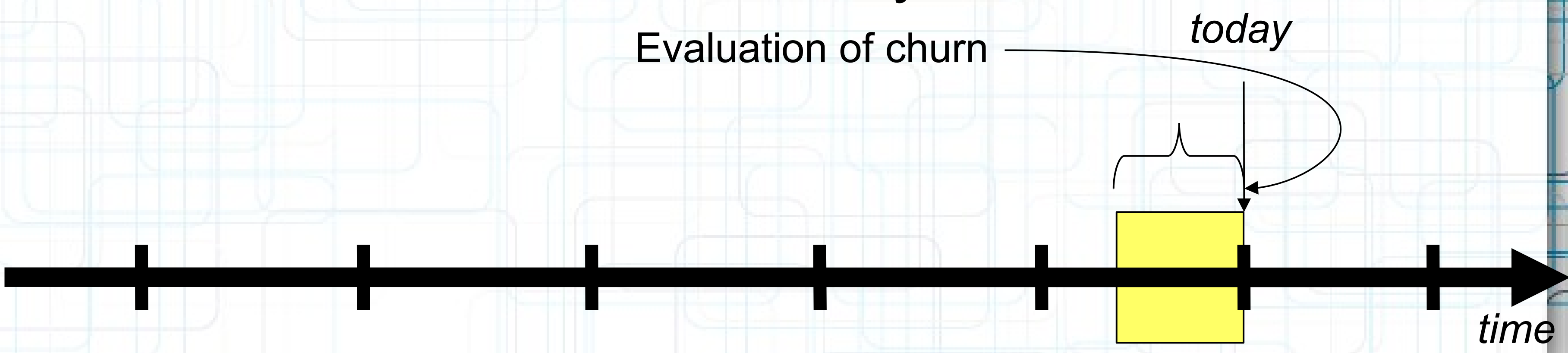
- Alternative notion of churn
  - Switch from a kind of relation with the bussiness to another one
  - Extreme case: from “loyal” to “complete churn”
- In retail selling
  - Loyal customers provide (some) guarantees about future income of the seller
  - Downgrading from “loyal” to “occasional” has large 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

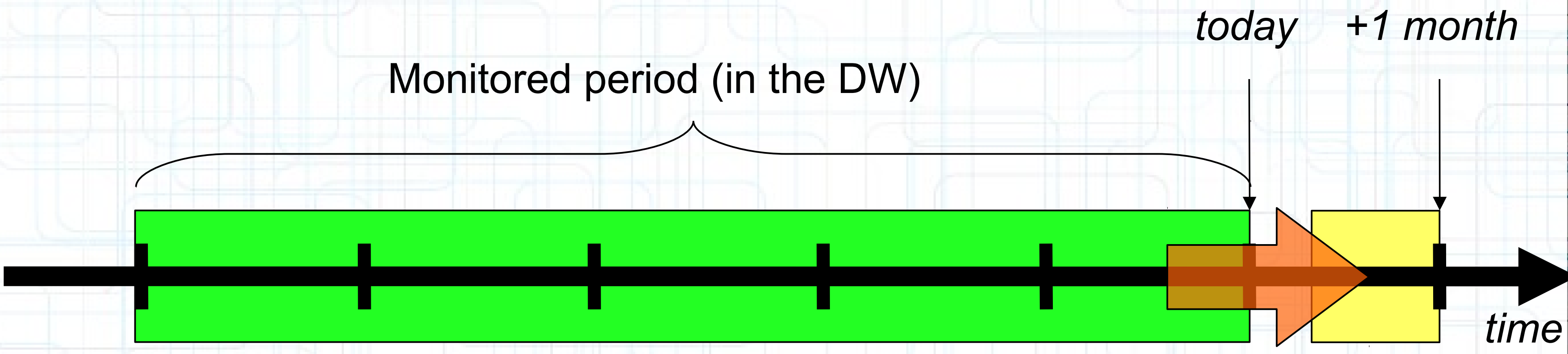


- The status of the customer can only be evaluated on the information we have “today”



Predicting churn

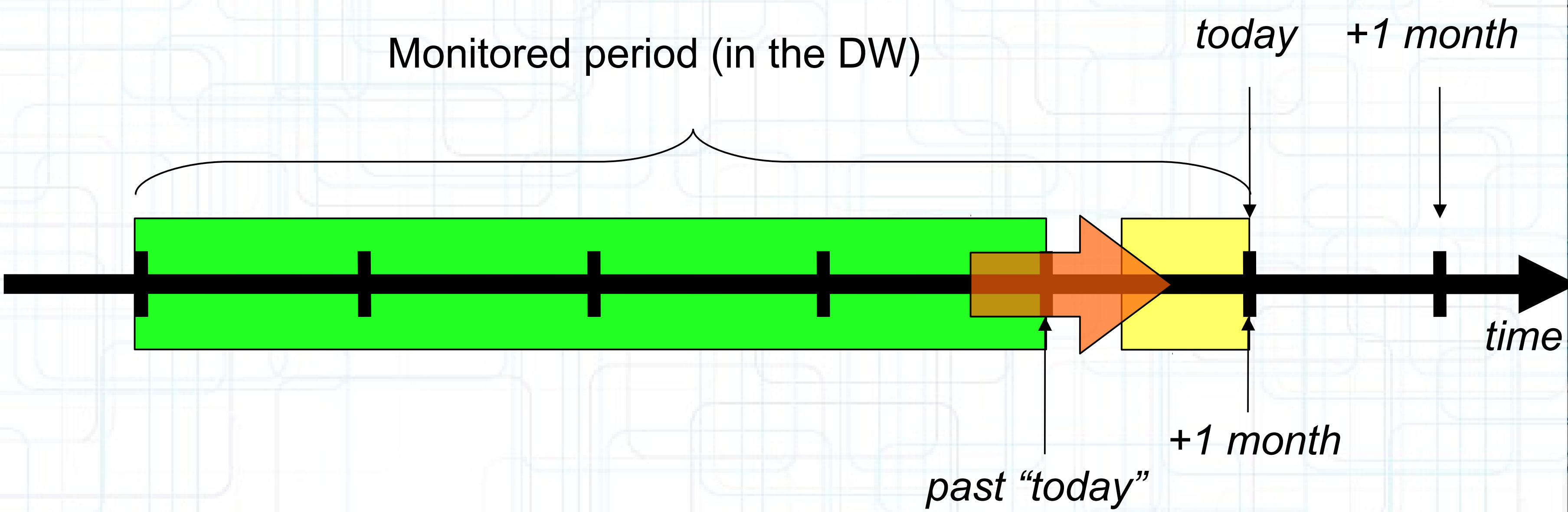
- Objective: predicting the *future* status of the customer, based on his recent history



- Recent history provides clues about the behaviour he is going to follow
  - Some clues help recognizing the future churners, other 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)



Schema of Churn prediction applications

- Define/extract working variables
  - Predictive: the *clues* available *today/past*
  - Traget 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

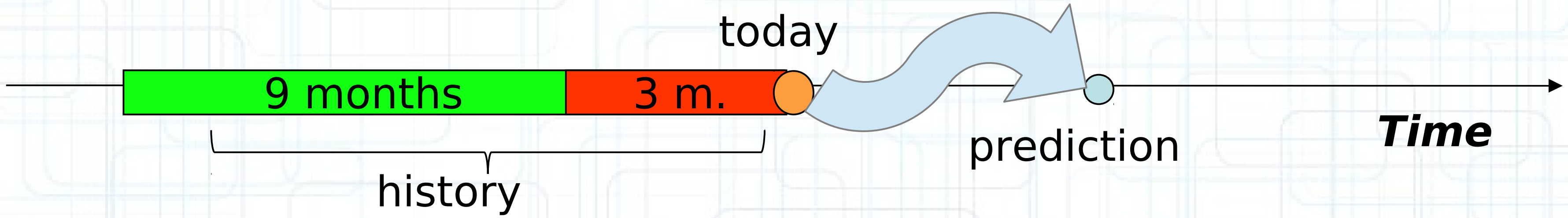
Churn risk definition

- 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 variabls (red bar)

- Build a predictive model
  - Learned on historical data
  - To be apply for predictions

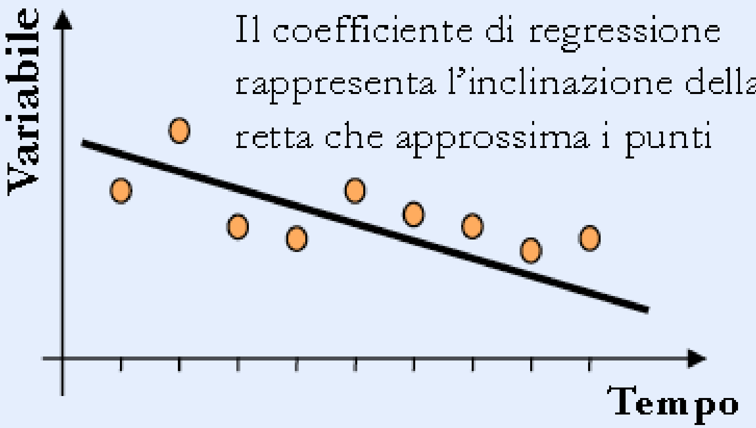


Data preparation – predictive variables (red bar)

Si sono estratte dal data warehouse, per il periodo di 9 mesi (Dicembre 2006 – Agosto 2007) le seguenti informazioni:

- Dati anagrafici (sesso, 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

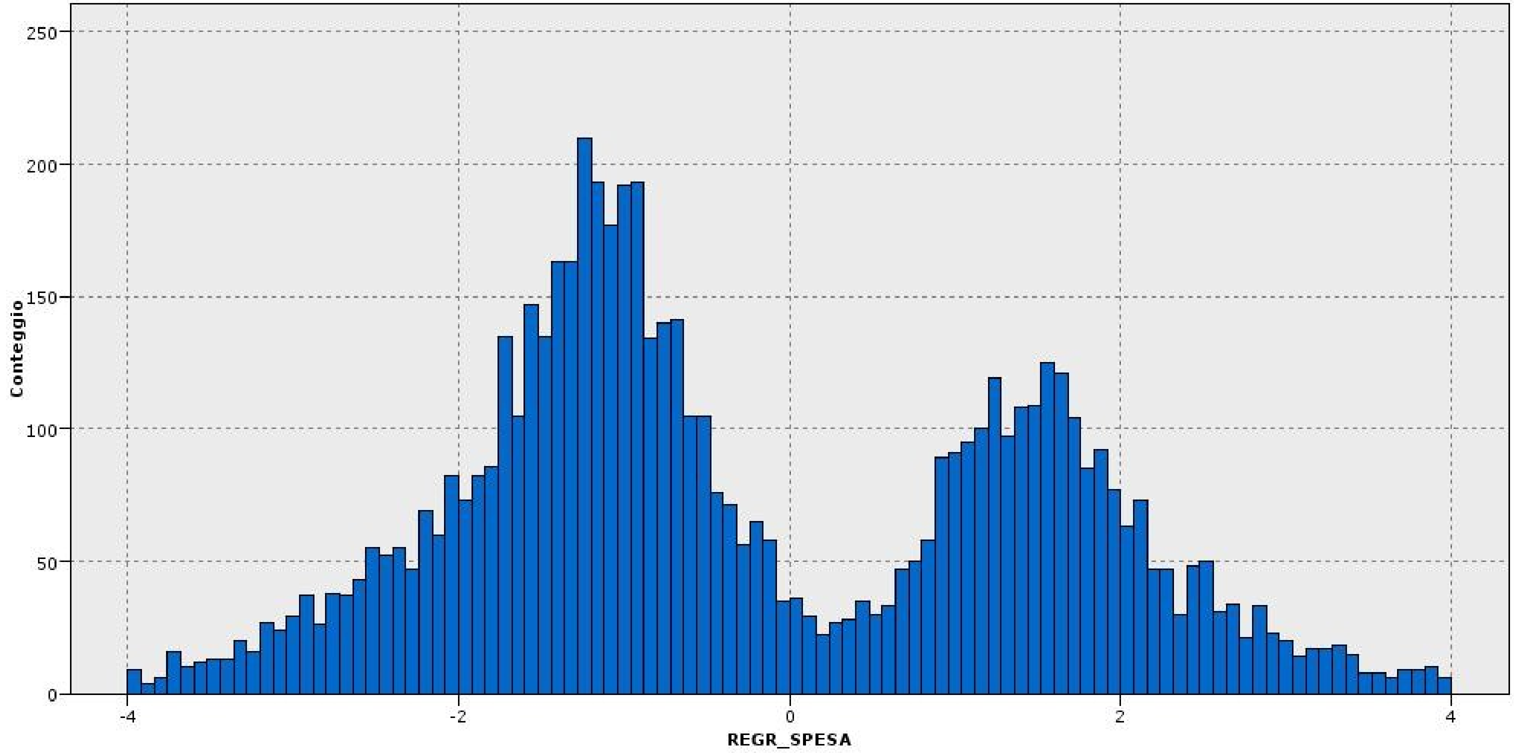
Dataset

- The final dataset contains a row for each customer, excluding those without any purchse in the period
  - 517.000 rows
  - 47 attributes

Predittori Anagrafici	Predittori di spesa	Predittori di trend	Variabili target
CLIENTE_ID	DATA_ULTIMA_SPESA	TIPOLOGIA_01	T_NUM_SPESE
SESSO	NUM_SPESE	TIPOLOGIA_02	T_RAPP_SPESE
STATO_CIVILE	SPESA_TOT	TIPOLOGIA_03	T_RAPP_SPESA
PROFESSIONE	SPESA_TOT_PESATA	TIPOLOGIA_04	T_RAPP_BATTUTE
TITOLO_STUDIO	SPESA_MEDIA	TIPOLOGIA_05	
PROVINCIA	SPESA_MEDIA_PESATA	TIPOLOGIA_06	
REGIONE	BATTUTE	TIPOLOGIA_07	
ANNO_SOCIO	FRESCHI_TOT	TIPOLOGIA_08	
FASCIA_ANNO_SOCIO	FRESCHI_SPESE	TIPOLOGIA_09	
FL_INVIO_RIVISTA	CARNI_TOT	TIPOLOGIA_MEDIA	
COD_NEGOZIO	CARNI_SPESE	TIPOLOGIA_ZERI	
ETA	PESCE_TOT	REGR_NUM_SPESE	
ETA_FASCIA	PESCE_SPESE	REGR_SPESA	
	ORTOFRUTTA_TOT	REGR_SPESA_PESATA	
	ORTOFRUTTA_SPESE	REGR_BATTUTE	

Data Exploration

- Distribution of expenditure trends

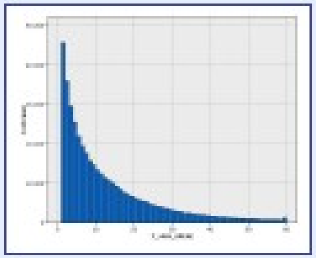


Customers with total purchases > 400€

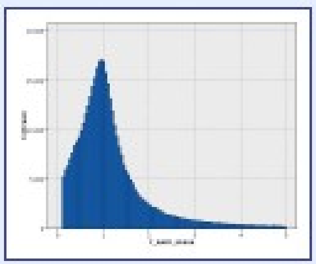
Target variables

Funzioni Obiettivo

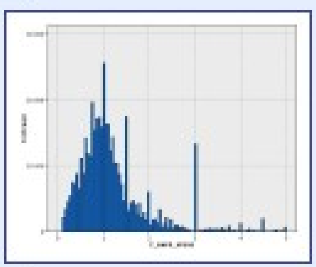
(Extract from COOP report)



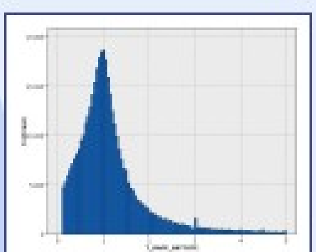
NUM_SPESE: spese del cliente nel periodo target



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



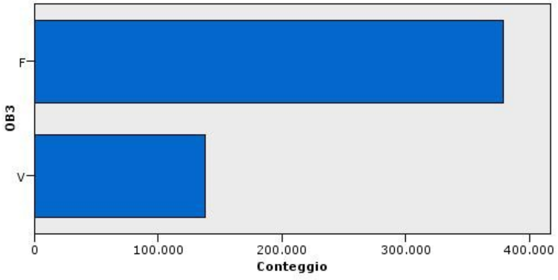
RAPP_BATTUTE: rapporto fra le battute di cassa del periodo target e quelle del periodo d'osservazione

Normalized measures

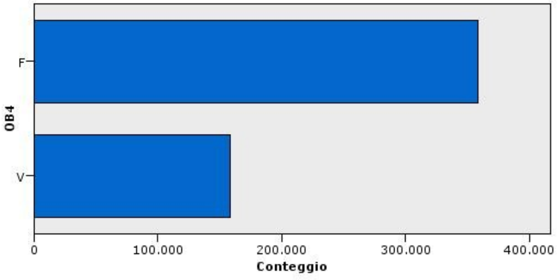
Discretizing target vars.

- Choose an alarm threshold
- Result: three churn binary variables
- Chosen threshold: 0.5   i.e. 50% decay
- Distributions obtained

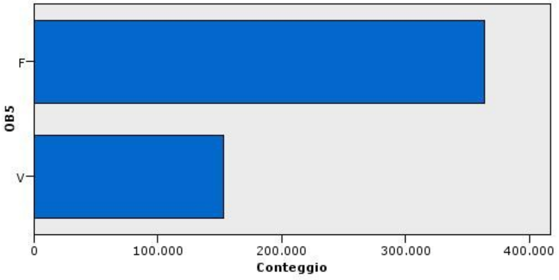
  (F= low risk, V = high risk)



OB1: RAPP_SPESE



OB2: RAPP_SPESA



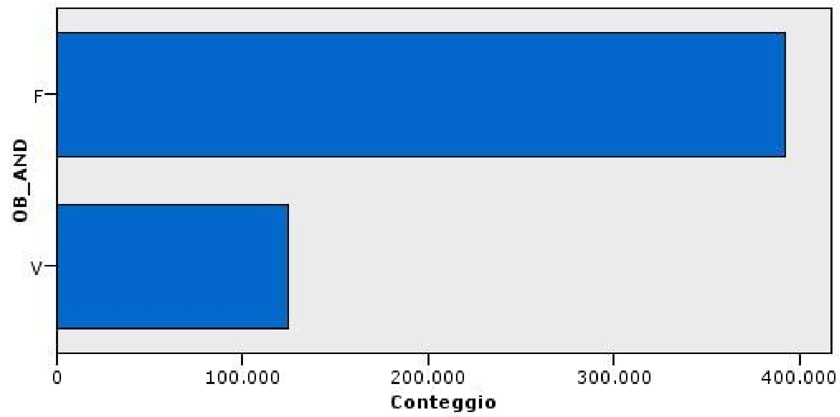
OB3: RAPP_BATTUTE

Discretizing target vars.

Combined variable:

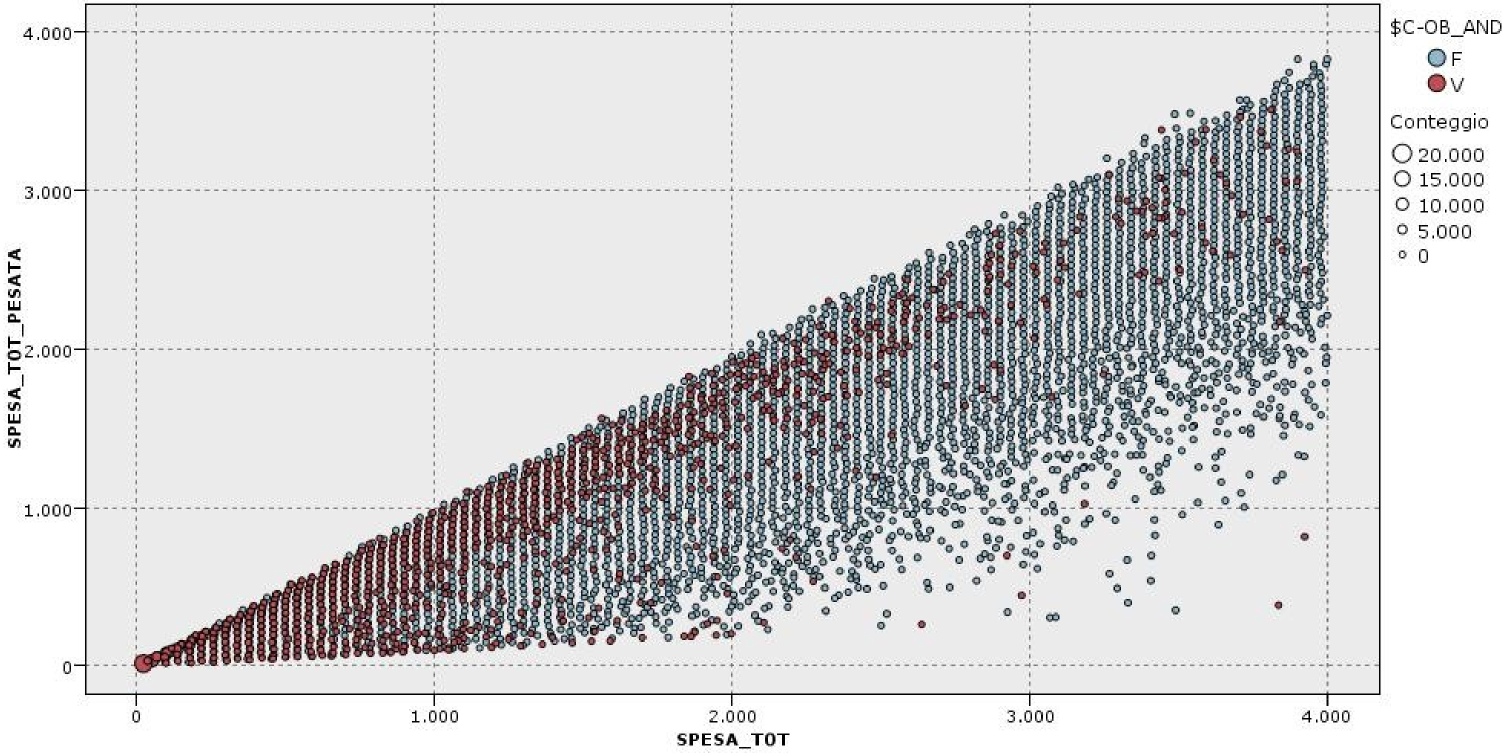
- Alarm = alarm in all 3 variables seen before

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

Predictions

Real values

Training Set	F	V
F	256.608	17.920
V	50.540	36.466

Test Set	F	V
F	110.029	7.767
V	21.855	15.734

66.9%

Gain: **42.8%**

Performances of classifier

- Lift chart