Modelli di data mining per la previsione del churn

Università Cattolica del Sacro Cuore
Milano, 24 ottobre 2005

Bruno Scarpa
Università di Pavia

la mia esperienza in azienda

- **Nel 1997, dopo gli studi, ho avuto l'occasione di cominciare un'avventura nel mondo delle aziende...**
  - **a. in RAS mi sono occupato della quotazione dei rischi e della definizione delle tariffe e dei prezzi delle assicurazioni non vita (auto, infortuni, incendio...)**
  - **b. Sono poi passato in omnitel dove ho avuto modo di impostare le attività di data mining come strumenti statistici per il marketing sulla clientela...**
  - **c. Dopo un po' di esperienza sono passato a tin.it dove oltre alle analisi statistiche avevo il compito di curare le azioni di marketing verso i clienti/navigatori/sottoscrittori...**
  - **d. L'ultima tappa della mia esperienza aziendale è stata in 3, azienda che era in start up, dove avevo il compito di curare l'impostazione globale e quindi la definizione di requisiti per la gestione della relazione con i clienti.**

statistica nel marketing

Diversi livelli di coinvolgimento della statistica nel „fare business“

Strumenti:	Reports Tabelle Semplici indicatori	Reports Tabelle Semplici indicatori Semplici modelli	Tabelle Indicatori/tassi Modelli Data mining	Modelli Statistici Data mining Reports Tabelle Indicatori

customer base: approccio strategico

Un unico obiettivo

Aumentare il Customer Lifetime Value attraverso la riduzione del churn e l'aumento dell'ARPU

… attraverso

PROFILING & SEGMENTATION

Identificare i potenziali "churners" → **Ridurre il churn**

Massimizzare la soddisfazione dei clienti → **Ridurre il churn**

Focalizzarsi su target per cross e up sell → **Massimizzare il valore del cliente**

Customer Relationship Management

Identificare

Realizzare

Praticare

tutte le attività necessarie a garantire il processo di attenzione e fidelizzazione dell'individuo verso l'azienda e la sua offerta di prodotti e servizi

e, conseguentemente

la massimizzazione delle opportunità di business attraverso la soddisfazione costante dei bisogni

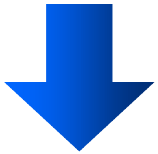
profiling

Identificare

Classificare

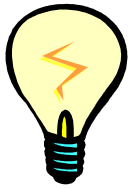
Acquisire

Gestire



tutte le informazioni che consentono la conoscenza e l'analisi del proprio target di riferimento

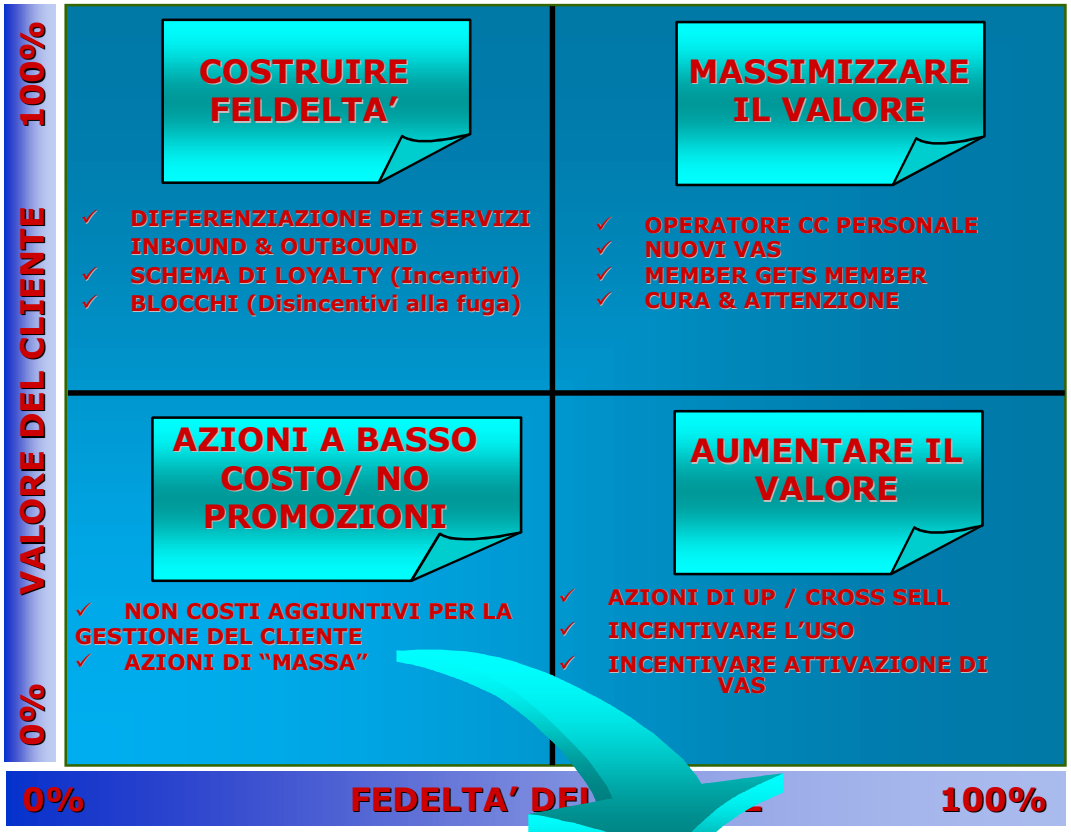
e, conseguentemente



la realizzazione di prodotti e servizi ad elevata probabilità di soddisfazione dei suoi bisogni

sources **analysis** **mgmnt**

customer base: segmentazione di base



...ALLA CONCORRENZA!

L'approccio di crm

OBIETTIVI

Generazione di profitti attraverso l'acquisizione di clienti ad alto valore, la creazione di valore e un'attenta politica di "cost to serve"

- Acquisizione di clienti (conversione di prospect in clienti)
- Aumento del valore dei clienti attraverso azioni di cross-sell e di up-sell
- "trattenere" i clienti (retention)
- Attenzione al "cost to serve"

AZIONI

- Campagne
- Programmi di "Loyalty"
- Personalizzazione dei Servizi
- Gestione dei canali di contatto
- (sviluppo di nuovi prodotti)

ANALISI (PROFILING)

- Analisi di dati sui clienti
- Segmentazione & micro-segmentazione
- Suggerimenti sui bisogni e le preferenze dei clienti

- Analisi delle propensioni sui "contatti"
  - Uso dei canali
  - Ragioni di contatto
- Esperienza del cliente per ogni canale (CC, Portale, chat, Messaging, Dealers)
- Efficienza del Customer Care

DATI "GREZZI"

CUSTOMER DATA

- ID (nome, indirizzo, num. di telefono, piano tariffario...)
- Dati Demografici (età, città, ...)
- Dati Socio-economici (reddito, professione, titolo di studio, classe di valore, ...)
- Dati Psicografici (ambiente, preferenze, hobbies, ...)
- Patterns d'uso (spesa telefonica, servizi usati e frequenza, ...)

- Storia dei contatti del cliente
  - Uso dei canali
  - Motivo di contatto
  - Numero di contatti

DATI SUI LIVELLI DI SERVIZIO DEI CANALI

...

L'approccio di crm

OBIETTIVI

Generazione di profitti attraverso l'acquisizione di clienti ad alto valore, la creazione di valore e un'attenta politica di "cost to serve"

- Acquisizione di clienti (conversione di prospect in clienti)
- Aumento del valore dei clienti attraverso azioni di cross-sell e di up-sell
- "trattenere" i clienti (retention)
- Attenzione al "cost to serve"

AZIONI

- Campagne
- Programmi di "Loyalty"
- Personalizzazione dei Servizi
- Gestione dei canali di contatto
- (sviluppo di nuovi prodotti)

ANALISI (PROFILING)

- Analisi di dati sui clienti
- Segmentazione & micro-segmentazione
- Suggerimenti sui bisogni e le preferenze dei clienti

- Analisi delle propensioni sui "contatti"
  - Uso dei canali
  - Ragioni di contatto
- Esperienza del cliente per ogni canale (CC, Portale,chat, Messaging, Dealers)
- Efficienza del Customer Care

DATI "GREZZI"

CUSTOMER DATA

- ID (nome, indirizzo, num. di telefono, piano tariffario...)
- Dati Demografici (età, città, ...)
- Dati Socio-economici (reddito, professione, titolo di studio, classe di valore, ...)
- Dati Psicografici (ambiente, preferenze, hobbies, ...)
- Patterns d'uso (spesa telefonica, servizi usati e frequenza, ...)

- Storia dei contatti del cliente
  - Uso dei canali
  - Motivo di contatto
  - Numero di contatti

DATI SUI LIVELLI DI SERVIZIO DEI CANALI

...

***costa poco rilevare dati in modalità automatica**

***costa poco immagazzinare dati in data-base sempre più grandi**

Contesti rilevanti:

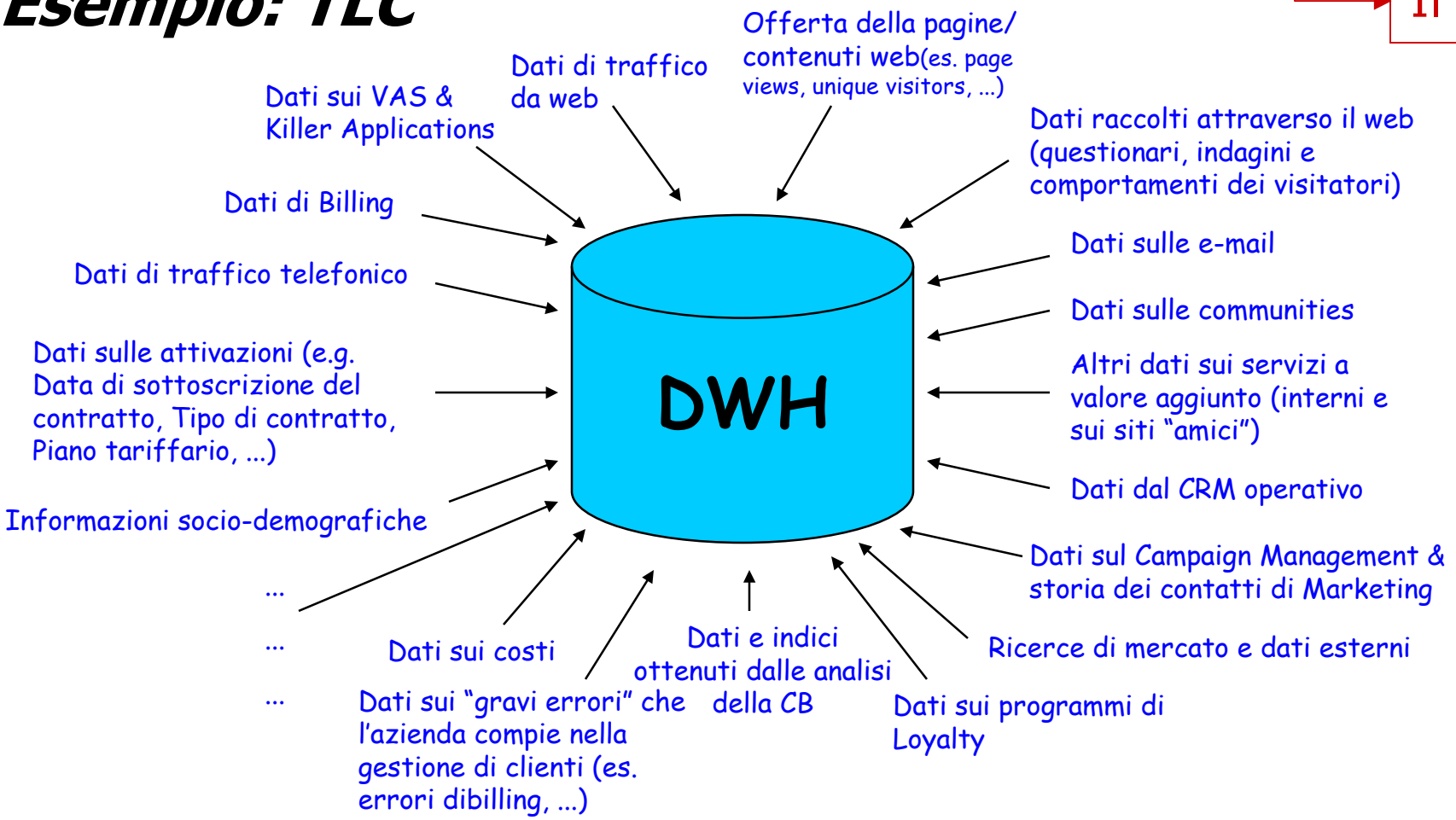
* data-base aziendali (customer-base, CRM, ...) soprattutto per telefoniche, banche e assicurazioni, grande distribuzione (cfr carte fedeltà)

* ambito scientifico: microarrays, radiotelescopi, fisica delle alte energie

* tecnologie varie: telerilevazione, riconoscimento vocale, OCR, etc.

* dati non strutturati
  - text-mining (motori di ricerca web)

Esempio: TLC


DWH
Dati sui VAS & Killer Applications
Dati di traffico da web
Offerta della pagine/ contenuti web(es. page views, unique visitors, ...)
Dati raccolti attraverso il web (questionari, indagini e comportamenti dei visitatori)
Dati di Billing
Dati di traffico telefonico
Dati sulle attivazioni (e.g. Data di sottoscrizione del contratto, Tipo di contratto, Piano tariffario, ...)
Informazioni socio-demografiche
...
...
...
Dati sulle e-mail
Dati sulle communities
Altri dati sui servizi a valore aggiunto (interni e sui siti "amici")
Dati dal CRM operativo
Dati sul Campaign Management & storia dei contatti di Marketing
Ricerce di mercato e dati esterni
Dati sui programmi di Loyalty
Dati e indici ottenuti dalle analisi della CB
Dati sui costi
Dati sui "gravi errori" che l'azienda compie nella gestione di clienti (es. errori dibilling, ...)

*il data-base disponibile è enorme!

è quindi opportuno:

- tener conto degli obiettivi delle analisi

- non concentrarsi solo su da dove e come raccogliere informazioni

- utilizzare estrazioni di parti del data-base diverse a seconda degli obiettivi

Ad esempio:

* Nel datamart per la previsione della disattivazione, è più utile tenere l'informazione sugli errori di fatturazione rispetto ai dettagli socio-demografici

* Campioni casuali di clienti possono essere molto utili

L'approccio di crm

OBIETTIVI

Generazione di profitti attraverso l'acquisizione di clienti ad alto valore, la creazione di valore e un'attenta politica di "cost to serve"

- Acquisizione di clienti (conversione di prospect in clienti)
- Aumento del valore dei clienti attraverso azioni di cross-sell e di up-sell
- "trattenere" i clienti (retention)
- Attenzione al "cost to serve"

AZIONI

- Campagne
- Programmi di "Loyalty"
- Personalizzazione dei Servizi
- Gestione dei canali di contatto
- (sviluppo di nuovi prodotti)

ANALISI (PROFILING)

- Analisi di dati sui clienti
- Segmentazione & micro-segmentazione
- Suggerimenti sui bisogni e le preferenze dei clienti

- Analisi delle propensioni sui "contatti"
  - Uso dei canali
  - Ragioni di contatto
- Esperienza del cliente per ogni canale (CC, Portale,chat, Messaging, Dealers)
- Efficienza del Customer Care

DATI "GREZZI"

CUSTOMER DATA

- ID (nome, indirizzo, num. di telefono, piano tariffario...)
- Dati Demografici (età, città, ...)
- Dati Socio-economici (reddito, professione, titolo di studio, classe di valore, ...)
- Dati Psicografici (ambiente, preferenze, hobbies, ...)
- Patterns d'uso (spesa telefonica, servizi usati e frequenza, ...)

- Storia dei contatti del cliente
  - Uso dei canali
  - Motivo di contatto
  - Numero di contatti

DATI SUI LIVELLI DI SERVIZIO DEI CANALI

...

TLC: alcuni problemi

☐acquisizione della clientela

- **prospect**

  →Raggiungere una *soglia minima* di clienti ASAP

  →Trovare e attrarre i clienti *giusti*: quanto spendere per ciascun diverso cliente potenziale?

- **Scoprire frodi di sottoscrizione**

  →determinare domande di sottoscrizione fraudolente

☐profittabilità dei clienti

- **valore del cliente**
- **„dormienti" e share of wallet**
- **monitoraggio e management del rischio**

  →Determinare e ottimizzare i parametri di rischio

TLC: alcuni problemi

- **fedeltà**
  - **approccio predittivo**
    - →modelli di previsione del churn
  - **attrito e retention**
    - →Modellazione e determinazione e delle principali cause
  - **azioni: programmi di loyalty/campagne/up sell-cross sell**
  - **relazione col cliente**
    - →Personalizzazione dell'attenzione e del contatto

- **Customer profiling**
  - **chi sono i clienti?**
  - **cosa ciascun cliente vuole?**
  - **come contattare ogni cliente?**

- **valutazione delle azioni**
  - **Spesso non è possibile effettuare esperimenti caso-controllo**
    - →I clienti sono autoselezionati
  - **Valutare a posteriori alcune azioni**
    - →Stima gli effetti delle azioni condizionatamente all'effetto di tutte le altre variabili

L'approccio di crm

OBIETTIVI

Generazione di profitti attraverso l'acquisizione di clienti ad alto valore, la creazione di valore e un'attenta politica di "cost to serve"

- Acquisizione di clienti (conversione di prospect in clienti)
- Aumento del valore dei clienti attraverso azioni di cross-sell e di up-sell
- "trattenere" i clienti (retention)
- Attenzione al "cost to serve"

AZIONI

- Campagne
- Programmi di "Loyalty"
- Personalizzazione dei Servizi
- Gestione dei canali di contatto
- (sviluppo di nuovi prodotti)

ANALISI (PROFILING)

- Analisi di dati sui clienti
- Segmentazione & micro-segmentazione
- Suggerimenti sui bisogni e le preferenze dei clienti

- Analisi delle propensioni sui "contatti"
  - Uso dei canali
  - Ragioni di contatto
- Esperienza del cliente per ogni canale (CC, Portale,chat, Messaging, Dealers)
- Efficienza del Customer Care

DATI "GREZZI"

CUSTOMER DATA

- ID (nome, indirizzo, num. di telefono, piano tariffario...)
- Dati Demografici (età, città, ...)
- Dati Socio-economici (reddito, professione, titolo di studio, classe di valore, ...)
- Dati Psicografici (ambiente, preferenze, hobbies, ...)
- Patterns d'uso (spesa telefonica, servizi usati e frequenza, ...)

- Storia dei contatti del cliente
  - Uso dei canali
  - Motivo di contatto
  - Numero di contatti

DATI SUI LIVELLI DI SERVIZIO DEI CANALI

...

modelli statistici e data mining

- Utilizzo di tecniche e metodologie statistiche di vario tipo e di diverso livello di complessità

- L'approccio è graduale: si parte da soluzioni e metodologie semplici e poi man mano ci si sposta su modelli e strumenti più sofisticati (KISS = Keep It Simple. Sam!)

- È preferibile non affidarsi a soluzioni automatiche (black box) che propongono "schiaccia il bottone e il computer farà tutto da solo" (la proposta tipica dei tools in vendita)

- Data mining: Insieme di tecniche statistiche (e non) per la stima di modelli non-lineari per grosse quantità di dati, ma caratterizzate da ridotta complessità computazionale.

Definizione abbastanza condivisa:

'Data mining' rappresenta l'attività di elaborazione in forma grafica o numerica di grandi raccolte o di flussi continui di dati con lo scopo di estrarre informazione utile a chi detiene i dati stessi.

***ma di fatto ognuno la vive in modo diverso**

***soprattutto in aree disciplinari diverse**

Voci autorevoli:

Data mining is fundamentally an applied discipline (...)

data mining requires an understanding of both statistical and computational issues. (p. xxviii)

The most fundamental difference between classical statistical applications and data mining is the size of the data. (p. 19)

[da Hand, Mannila & Smith, 2001]

Aspetti salienti:

* la dimensione dei dati lievita
(qui n.righe ~ $10^3/10^6$,
n.colonne ~ $10^2/10^3$)

* ambito osservazionale

* ma non esiste un "piano campionamento";
semplicemente i dati "esistono"

* dati raccolti per esigenze gestionali o
simili, non per scopi di analisi

* i dati sono sporchi, anzi luridi

* campioni o censimenti?

Osservazioni sparse:

* La dimensione dei dati è importante:

  "every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it" (J.Friedman, 1997)

* tutti i valori-p sono ultra-significativi

* tutti i modelli sono "sbagliati"
  --> gestire conflitto/compromesso tra distorsione e varianza

* ma abbiamo *n* grande quanto si vuole, finalmente! possiamo fare a pezzi il campione e usarli per ruoli diversi (tipicamente: apprendere e verificare)

customer base: il churn

Modellare la disattivazione:

costruire, validare, interpretare un modello che descriva il comportamento degli utenti in termini di disattivazione in relazione ad altre variabili note

Perché?

- ✓**Per descrivere il fenomeno**
- ✓**Per prevedere i potenziali futuri disattivi**
- ✓**Per predisporre azioni**
- ✓**Per verificare l'efficacia di operazioni di marketing/Customer Operation**

customer base: il churn

Le fonti

Aziendali: ("DWH", database operazionali, …)

- **Per tutti i clienti**
- **Informazioni su**
  - **traffico**
  - **servizi opzionali**
  - **comportamento del cliente**
  - **reclami-rapporti con customer care**
  - **azioni di marketing/customer care**
  - **fatture/ricariche**
  - **dati demografico/anagrafici**

Altre fonti: Ricerche di mercato

- **Per un piccolo campione "casuale" di clienti**
- **Informazioni su**
  - **comportamenti**
  - **stili di vita**
  - **motivi della disattivazione**
  - **tempi della scelta**
- **ottenuti tramite interviste**

il churn: gli obiettivi

- **Determinare un indicatore di propensione alla disattivazione per ogni login**
- **Prevedere i potenziali futuri disattivi**
- **Capire i motivi fondamentali che portano alla disattivazione e i comportamenti che la precedono**
- **Individuare possibili azioni volte alla *retention* del cliente**
- **Verificare l'efficacia di operazioni di Marketing/Customer Operation**

il churn: gli obiettivi

Passi principali

- **Identificazione della popolazione**
- **Determinazione e reperimento delle variabili**
- **Definizione del target**
- **Stima del modello**
- **Verifica dell'accuratezza**
- **Utilizzo del modello**

Selezionare la popolazione

Prima di estrarre i dati da DWH è necessario definire in maniera precisa la popolazione da analizzare.

Esempio
I clienti utilizzati per costruire un modello di *churn* per il prodotto "pippo" sono i clienti che hanno data di attivazione precedente il 1/12/2004 ed aventi data di disattivazione superiore al 31/1/2005 o ancora attive nel mese di Febbraio 2005.

Per ciascuna di queste login sono stati estratti da DWH (input al sistema di data mining) i dati (le variabili x,y,z...) di Luglio 2004, Agosto 2004, Settembre 2004, Ottobre 2004, Novembre 2004.

il churn: data mining

L'oggetto della previsione

La variabile target va definita con precisione in termini di status dei clienti e date degli eventi considerati (disattivazione, attivazione, sospensione…)

Esclusione variabili leaker

Alcune delle variabili presenti nei dati in ingresso sono strettamente correlate con l'oggetto della previsione per diversi motivi.

Le variabili che "trasudano" informazione sulla variabile target (*leakers*) devono venire identificate ed escluse dall'insieme di dati a disposizione (ad es. data di disattivazione, status della login, flag varii)

Schede postpagate

Selezione della Popolazione

Si analizza la customer base dei clienti post-pagati al 31 ottobre 2004, attivati almeno 4 mesi prima (prima di luglio 2004) confrontando coloro che si sono disattivati nel mese di novembre 2004 rispetto agli altri. Per queste login si è considerato il traffico fino a settembre 2004.

L'oggetto della previsione

La variabile target è un indicatore (variabile dummy) dell'evento disattivazione nel mese di novembre 2004.

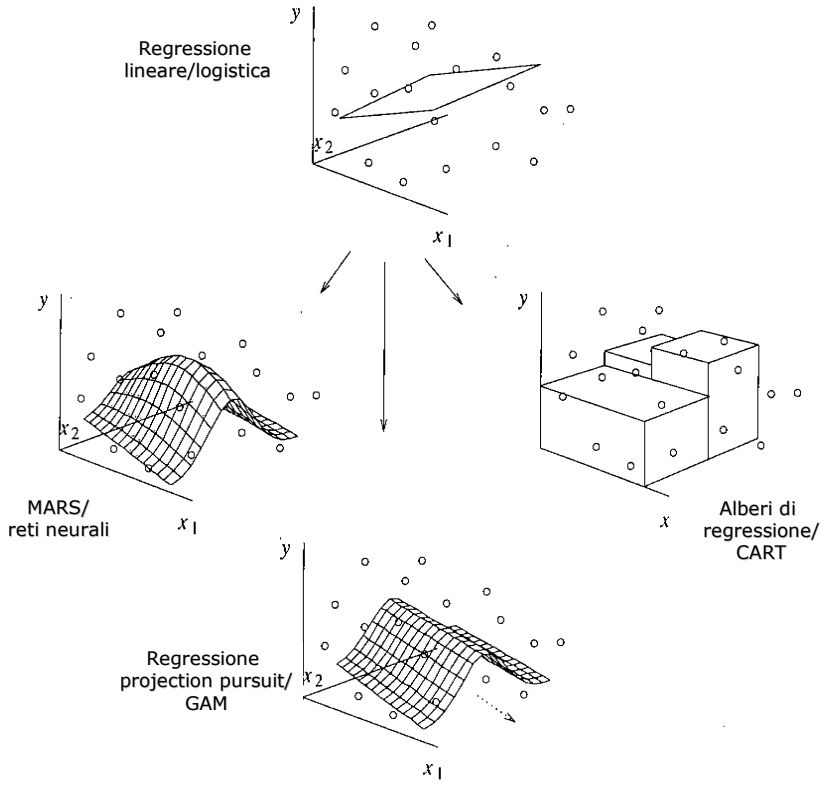
il churn: data mining

Modelli per il churn

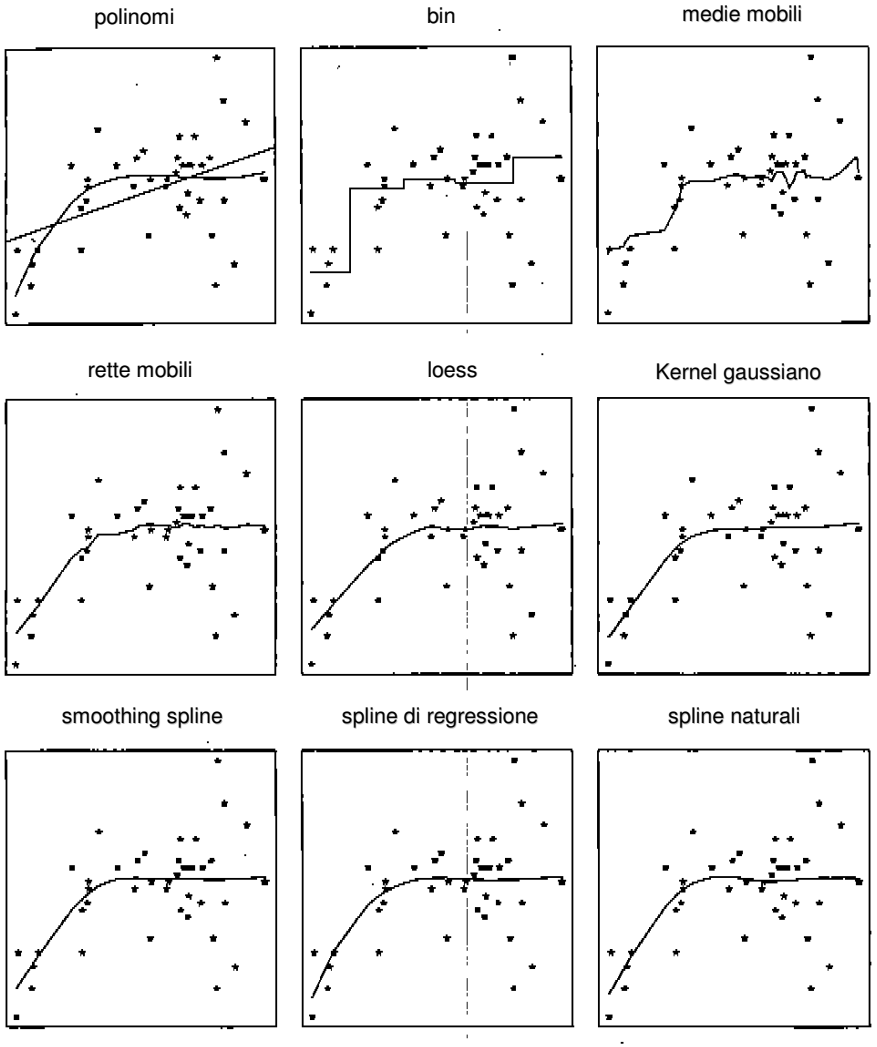
Il modello più semplice (lineare) non è sufficiente a descrivere i dati

✓bisogna ipotizzare modelli più complessi

✓lasciarsi guidare dalle osservazioni per costruire le relazioni tra variabili e disattivazione



Smoothers monodimensionali



- ✓ **Stimatori nonparametrici basati sulle serie o su regressioni (polinomiali, regressione di Fourier, splines di regressione, filtraggio)**
- ✓ **Stimatori nonparametrici kernel (Nadaraya-Watson, medie localmente pesate, regressione locale, loess)**
- ✓ **Smoothing Splines (penalizzazione)**
- ✓ **Stimatori nonparametrici basati sui vicini più prossimi - Nearest neighbor (medie mobili, mediane, stimatori di Tukey)**

GAM

Generalized Additive Models

Idea di base:
usare stimatori non parametrici unidimensionali come blocchi per la costruzione di una classe ristretta di modelli non parametrici per la regressione multipla

Modello additivo lineare

Modello lineare:

$$\mathbf{Y} = \alpha + \sum_{j=1}^{p} \beta_j \mathbf{X}_j + \varepsilon$$

Modello additivo:

$$\mathbf{Y} = \alpha + \sum_{j=1}^{p} f_j(\mathbf{X}_j) + \varepsilon$$

✓**Le f_j sono funzioni arbitrarie, una per ogni variabile predittiva**

✓**Gli ε_i sono variabili aleatorie di errore e vengono assunti indipendenti tra loro, dalle $\mathbf{X}_j$ con $E(\varepsilon_i)=0$ e $\mathrm{var}(\varepsilon_i)=\sigma^2$**

✓ **inoltre per l'identificabilità si assume che** $E\{f_j(\mathbf{X}_j)\} = 0$

GAM

Modello additivo logistico

GLM logistico:

$$\log \frac{p(y_i|x_{i1},\dots x_{ip})}{1-p(y_i|x_{i1},\dots x_{ip})} = \beta_0 + x_{i1}\beta_1 + \dots x_{ip}\beta_p$$

GAM logistico:

$$\log \frac{p(y_i|x_{i1},\dots x_{ip})}{1-p(y_i|x_{i1},\dots x_{ip})} = \beta_0 + f_1(x_{i1}) + \dots f_p(x_{ip})$$

- ✓ Le f_j sono funzioni arbitrarie, una per ogni variabile predittiva
- ✓ Le Y_i sono variabili aleatorie Binomiali e vengono assunte indipendenti tra loro
- ✓ inoltre per l'identificabilità si assume che $E\{f_j(X_j)\}=0$

GAM

Algoritmo di backfitting

1. **Inizializzazione:** $\alpha = \frac{1}{n}\sum_{i=1}^{n} y_i$

$$f_j = f_j^{(0)}, j = 1, \ldots, p$$

2. **Ciclo: per** $i=1, 2\ldots,$ $j=1,\ldots, p$

$$f_j^{(i)} = S_j\left(\mathbf{Y} - \alpha - \sum_{k \neq j} f_k^{(i-1)} \middle| \mathbf{X}_k\right)$$

3. **Fino a: ciascuna funzione** $f_j^{(i)}$ **è uguale alla funzione** $f_j^{(i-1)}$.

La convergenza non è assicurata in generale, ma per casi particolari (anche se molto frequenti).

GAM

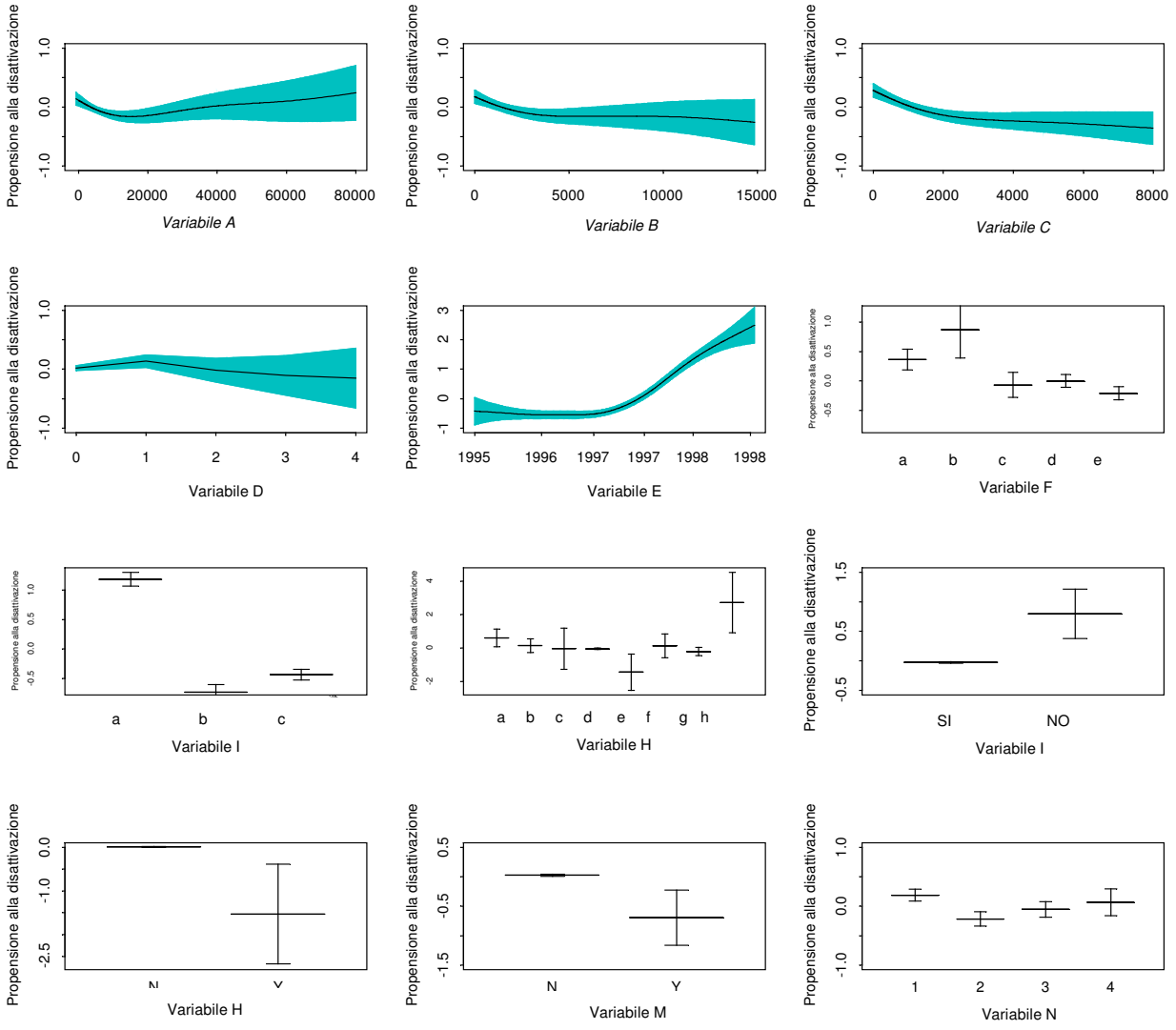
✓**Per il nostro problema di prevedere il churn sui post pagati, si stima un modello GAM con le scelte seguenti:**

- **funzione legame logistica**
- **per le variabili continue, stimatore univariato spline cubiche**
- **selezione dei parametri di "lisciamento" attraverso ispezione grafica/tuning manuale**
- **stima con algoritmo di backfitting**

✓**Le variabili risultate non significative effettuando test statistici asintotici approssimati sono state escluse**

✓**Calcolo della stima di una misura di propensione al churn per ciascun cliente utilizzando il modello stimato e determinazione di eventuali classi di rischio.**

GAM


Propensione alla disattivazione
Variabile A
Variabile B
Variabile C
Variabile D
Variabile E
Variabile F
Variabile I
Variabile H
Variabile I
Variabile H
Variabile M
Variabile N
SI
NO
N
Y

il churn: esempio

TLC - Mobile

È necessario un diverso approccio tra prepagate e post-pagate perché

- Per il post-pagato
  - Il cliente **CHIEDE** di essere disattivato via raccomandata
- Per il pre-pagato
  - Il cliente **VIENE** disattivato quando non ricarica per 12 mesi

I clienti prepagati decidono di andare alla concorrenza molto prima della disattivazione "tecnica"

il churn: esempio

Churn: prepagato-postpagato

Per il post-pagato, per disattivare il servizio è necessario inviare una raccomandata

C'è un chiaro evento: l'azienda sa quando l'utilizzatore vuole disattivare

Il prepagato non viene disattivato. Esce dalla Customer base dopo 12 mesi consecutivi senza ricarica

Non c'è evidenza di quando il cliente decide di abbandonare

il churn: esempio

L'oggetto della previsione: prepagato

- Identificazione di un **segnale** del churn effettivo

  Tale segnale dovrebbe essere

  - "intuitivo" e "semplice" da calcolare
  - "legato" alla decisione del cliente di andarsene
  - accurato e autoesplicativo

  Il "segnale" viene individuato sulla base di

  - Traffico outgoing
  - Traffico incoming

L’oggetto della previsione

La variabile target viene definita con precisione attraverso un semplice **segnale** che si basa sul pattern di utilizzo del servizio.

Selezione della Popolazione

Si analizza la customer base dei clienti **prepagati** al 31 gennaio 2005 che si fosse attivata almeno 6 mesi prima (prima di agosto 2004) confrontando coloro che hanno mostrato il “segnale” per la prima volta nel mese di gennaio 2005 rispetto agli altri. Per questi record si sono considerati i dati fino a novembre 2004.

Predisposizione data set

- **Divisione casuale a metà** (circa) dell'insieme dei dati a disposizione. Creazione dei dataset "TOP" (che verrà utilizzato per la **stima**) e "BOTTOM" (che verrà utilizzato per la **validazione**).

- Dal file TOP, selezione di tutti i clienti disattivati.

- **Selezione casuale** (senza ripetizione) di un insieme di clienti attivi di numerosità (circa) uguale al numero di disattivi nel TOP. Unione di questi clienti ai disattivi appena estratti dal file TOP.

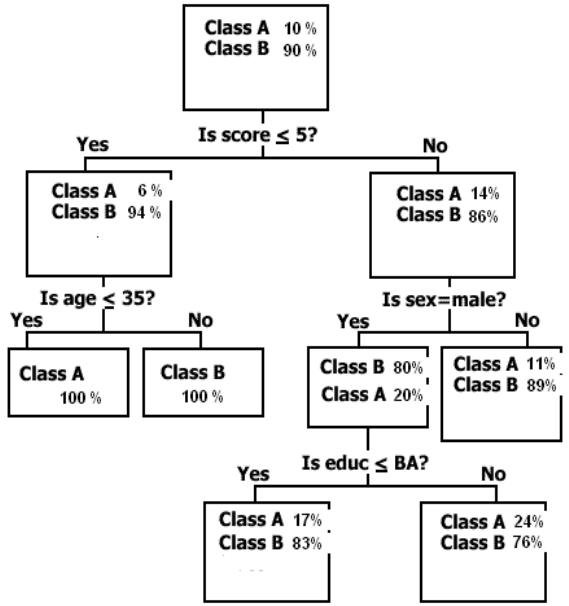
- **Divisione del dataset** ottenuto in due parti uguali selezionate casualmente, l'una servirà per **stimare** il modello e l'altra per **selezionare** il modello migliore.

CART: Alberi di classificazione

- Alberi che crescono in maniera ricorsiva
- Un nodo terminale g è diviso in due parti (figli di destra e di sinistra, gL e gR) in maniera da aumentare maggiormente il criterio di divisione (split)

$$D_g - D_{gL} - D_{gR}$$

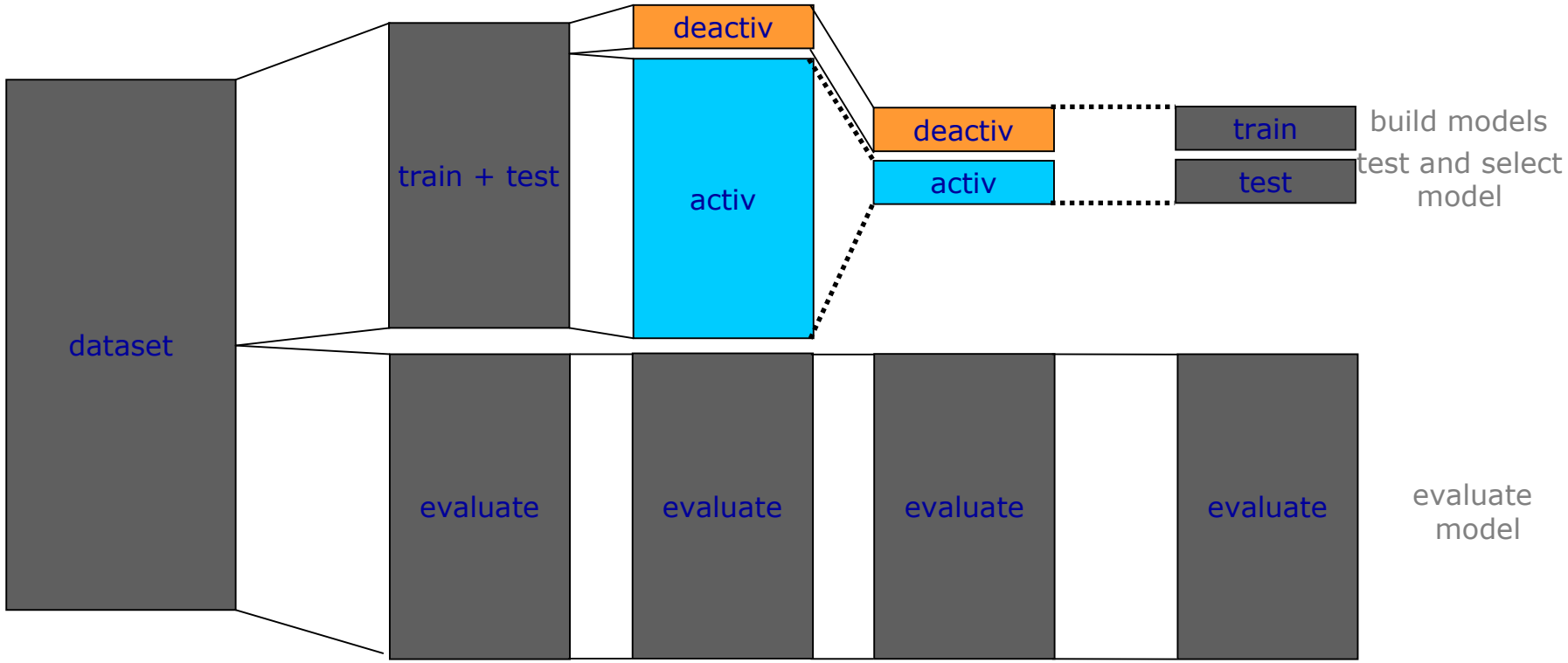
- dove D è una misura della bontà di adattamento (*goodness of fit*)
- Tipicamente gli *split* vengono definiti come partizioni univariate di ogni singola variabile di input
- Una volta costruito l'albero più grande viene generalmente "potato" (*pruned*) seguendo un criterio (generalmente basato su una funzione di costo)



- Principali Vantaggi:
  - Facile da capire e da interpretare
  - Considera facilmente osservazioni mancanti attraverso la creazione di variabili fittizie
- Principali Svantaggi:
  - Enfatizza le interazioni
  - La superficie di previsione non è liscia

churn: data mining

Predisposizione data set



CART: stima del modello

✓Si stima un modello CART (Classification and Regression Tree)
✓Si predispone un cammino di stima guidato, per cui le variabili da inserire nel modello e l'ordine di entrata viene definito e deciso a priori sulla base di

- Conoscenza del business
- *Actionability*
- Modelli di data mining stimati in precedenza
- Analisi preliminari e stime univariate

✓Si utilizza come regola di *split* l'indice di Gini

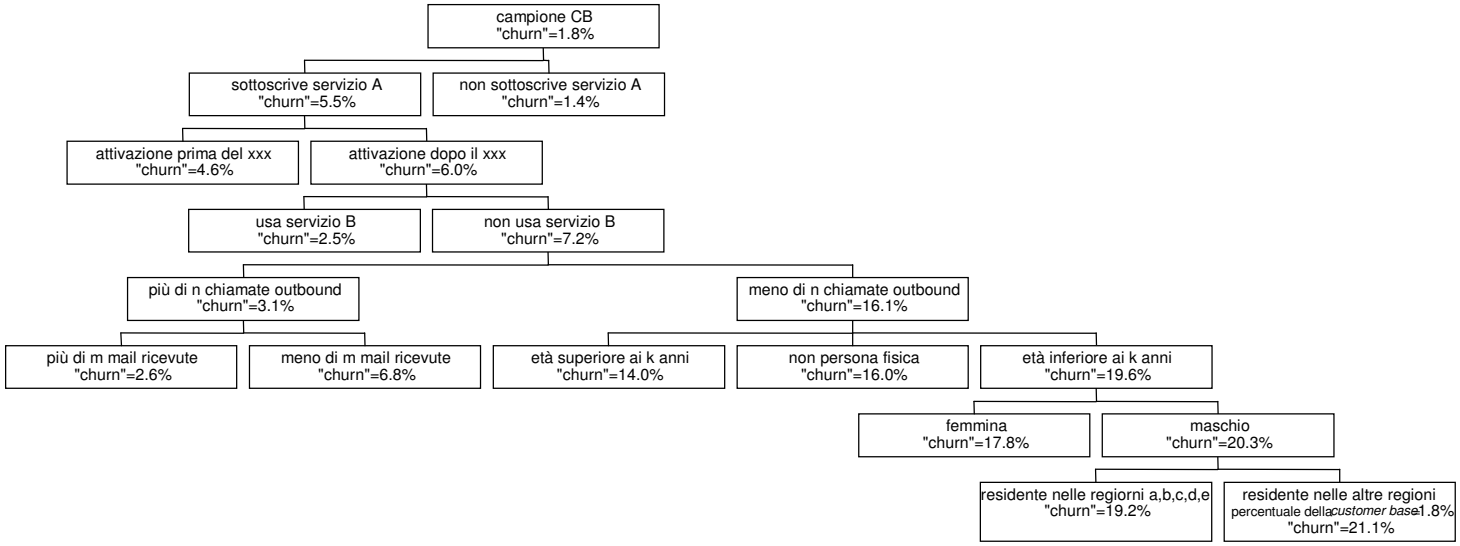
✓Le variabili risultate non importanti all'entrata per qualche ramo vengono eliminate solo nel ramo di riferimento

✓Non è necessaria una analisi di *pruning* globale per eliminare variabili

✓Calcolo della stima di una misura di propensione al churn per ciascun nodo nel dataset "eval" e determinazione dell'ordine dei nodi rispetto alla propensione alla disattivazione.

CART: previsione

Ad ogni nodo e foglia dell'albero è associato un diverso indice di propensione al churn (segnale). (I colori delle "foglie" dell'albero indicano classi di rischio: ■ bassa, ■ media, ■ alta)



CART: previsione

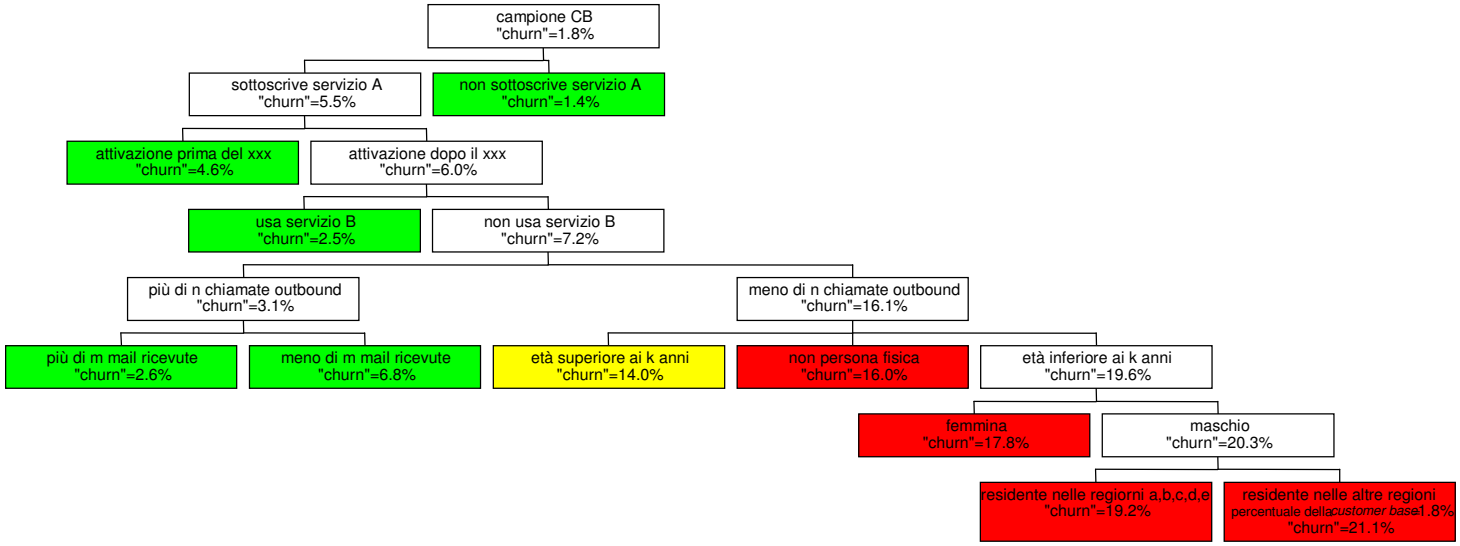
✓Determinazione delle soglie per la scelta di tre classi di rischio sulla base di numerosità dei nodi nel dataset “eval” e del livello di rischio dei nodi.

✓Per ciascun cliente si determina la foglia nell’albero a cui appartiene e si definisce Propensione al churn per quel cliente il valore della propensione nella foglia di riferimento.

✓Classificazione dei clienti nelle tre classi di rischio

CART: previsione

Ad ogni nodo e foglia dell'albero è associato un diverso indice di propensione al churn (segnale). (I colori delle "foglie" dell'albero indicano classi di rischio: ■ bassa, ■ media, ■ alta)



CART: la valutazione dei modelli

Il modello è stato stimato per poter essere utilizzato per fare previsione:

deve essere valido per qualsiasi altra situazione analoga.

(PCR)

Misure di accuratezza

- ✓Matrice di "confusione"
- ✓Lift

Misure globali: gli errori

Omissione: percentuale di clienti previsti ad alto rischio sul totale dei clienti effettivamente disattivati (cioè quanti di quelli effettivamente disattivati erano nella classe più a rischio il mese prima). L'errore di omissione viene indicato anche come "falsi negativi"

Commissione: percentuale di clienti disattivati sul totale dei clienti nella classe più a rischio (cioè quanti di quelli considerati a rischio sono stati effettivamente disattivati il mese successivo). L'errore di commissione è indicato anche con il termine "falsi positivi".

CART: la valutazione dei modelli

accuratezza del modello

omissione:

	alto Rischio	medio Rischio	basso Rischio	
SC = Y	27.94%	29.46%	42.60 %	(= 100%)
SC = N	3.57%	6.83%	89.60%	(= 100%)

comissione:

	SC = Y	SC = N	
alto Rischio	49.33%	50.67%	(= 100%)
medio Rischio	34.92%	65.08%	(= 100%)
basso Rischio	5.58%	94.42%	(= 100%)

Misure locali: il lift

I record (i clienti) vengono ordinati per propensione al churn decrescente, in modo da avere gli elementi ritenuti più a rischio nella prima parte della lista.

Si suddivide l'insieme così ottenuto in quantili e si calcola quanti disattivati reali si trovano nel primo quantile.

Il rapporto fra la percentuale di disattivati reali nel primo quantile rispetto alla percentuale di disattivati su tutta la popolazione considerata è detto **lift**.

Il *lift* misura quindi di quanto nel sottogruppo selezionato si prevede meglio la disattivazione rispetto a quello che si farebbe nella popolazione globale.

Più in generale tale misura è definita per un selezionato sottogruppo di una popolazione più vasta come la proporzione di disattivi nel sottogruppo diviso la proporzione di disattivi in tutta la popolazione.

CART: la valutazione dei modelli

Ad esempio:

Se l'insieme totale di login esaminate è costituito da 100.000 elementi di cui 3.000 disattivi, il tasso di churn per questa popolazione sarà di 3.000/100.000=0,03.

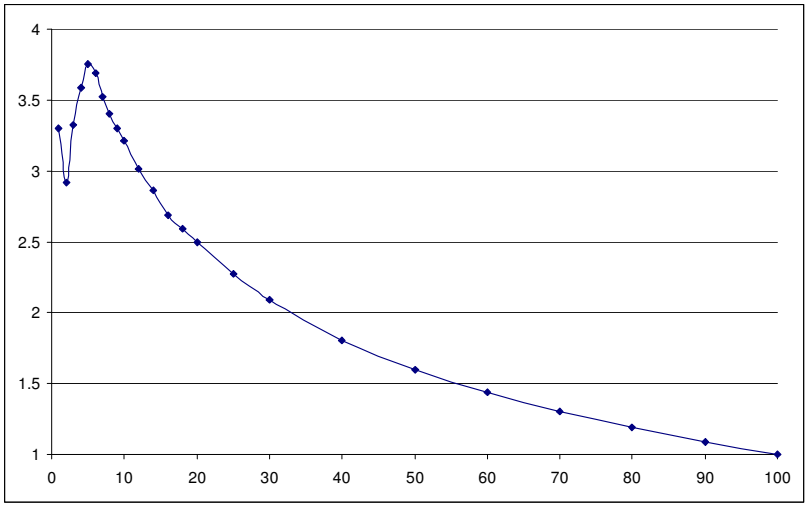
Se ordinando i risultati del sistema di previsione per propensione al churn decrescente si individuano, fra le prime 1000 login di questa lista, 150 disattivati, il tasso di churn sarà di 150/1000=0.15.

Ciò significa che il lift (per il primo 1%=1000/100.000 della popolazione) del sistema di previsione utilizzato è pari a 0.15/0.03=5.

Nota

Nelle telecomunicazioni in Italia il tasso di churn in un mese è molto basso (si aggira attorno all'1%-5%).
In questo caso, anche un metodo particolarmente accurato (es. lift=6) ha comunque un numero molto elevato di falsi positivi, cioè un errore di commissione particolarmente elevato (infatti per es. sui 100 clienti più a rischio secondo il sistema previsionale utilizzato, solo 6 saranno effettivamente disattivati).

Lift



- **Lift:** la funzione descrive, per ogni percentile della distribuzione (ordinata per propensione al churn stimata), il rapporto tra la percentuale di disattivati nel sottogruppo rispetto alla stessa percentuale sull'intera popolazione
- è una misura di quanto meglio si stimi il churn col modello, rispetto all'utilizzo di una strategia di "nessun modello"
- Il lift globale dell'intero gruppo di clienti che si prevede disattivino (circa il 15% della customer base) è di circa 2.74

il churn: previsione

data mining puro

Soluzione a ***black box*** in cui il software (IT) seleziona le variabili e determina il modello in maniera completamente automatica

Esempio:
Modello precedente

✓Calo del traffico

non actionable!

data mining guidato

Soluzione in cui l'analista **guida** le analisi nella scelta, almeno parziale, delle variabili, utilizzando i modelli di data mining come strumenti di analisi

Esempio:
Modello attuale

✓caratteristiche di traffico (es. Alto usage nelle ore di picco)

✓Uso dei servizi X e Y

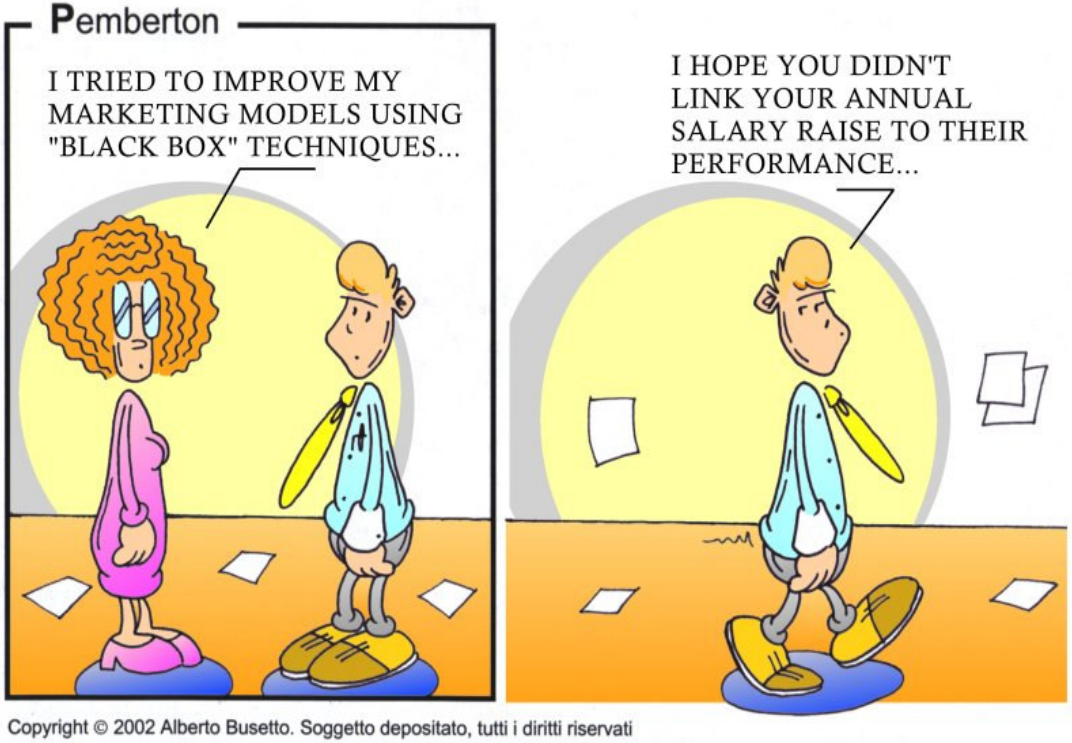
✓calo nel traffico

✓Reclami

Azioni di Marketing e CRM

* Le aziende necessitano di buoni statistici!
* Ci sono parecchi e interessanti problematiche
* I databases sono enormi, ma c'è bisogno di persone che sappiano trarre informazioni dai dati, non solo buoni software con algoritmi efficienti

- Insigth tools (SAS, Oracle, Clementine…)
  - Problemi non standard: Non sempre la soluzione è già nel tool
  - Software commerciale si propone come la soluzione dei problemi „schiacciando un bottone"
  - Non sempre partire con enormi dataset significa dover analizzare tutti i dati. Non sempre sono necessari algoritmi e modelli molto veloci (in parallelo...)



Bruno Scarpa
bruno.scarpa@unipv.it