Customer Relationship management & Data Mining

Dino Pedreschi, Fosca Giannotti, Mirco Nanni
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

Master MAINS 2015

Outline

- Basic elements of CRM
- CaseStudy1- Redemption
- A real case study: Coop DW (10)
- Churn Analysis (30)
- Estimating promotion sales (40)
- Forecasting “out-of-stock” during sales (10)
- Customer profiling(30)
  - Collective profiles
  - Individual profiles
- Discovering Innovators (20)

Customer Relationship Management

CRM is about acquiring and retaining customers

- Customer acquisition (often via targeted marketing)
- Customer retention (minimizing churn)
- Cross-sell and up-sell (extending the relationship)
- Customer win-back (recovering customer lost to competitors)
- Customer support (increasing customer satisfaction)
- Maximizing lifetime value (repeat purchase, subscription services).

CRM as a Business Strategy

- CRM is not just technology, but a strategy, process, and business goal that an organization must embrace on an enterprisewide level

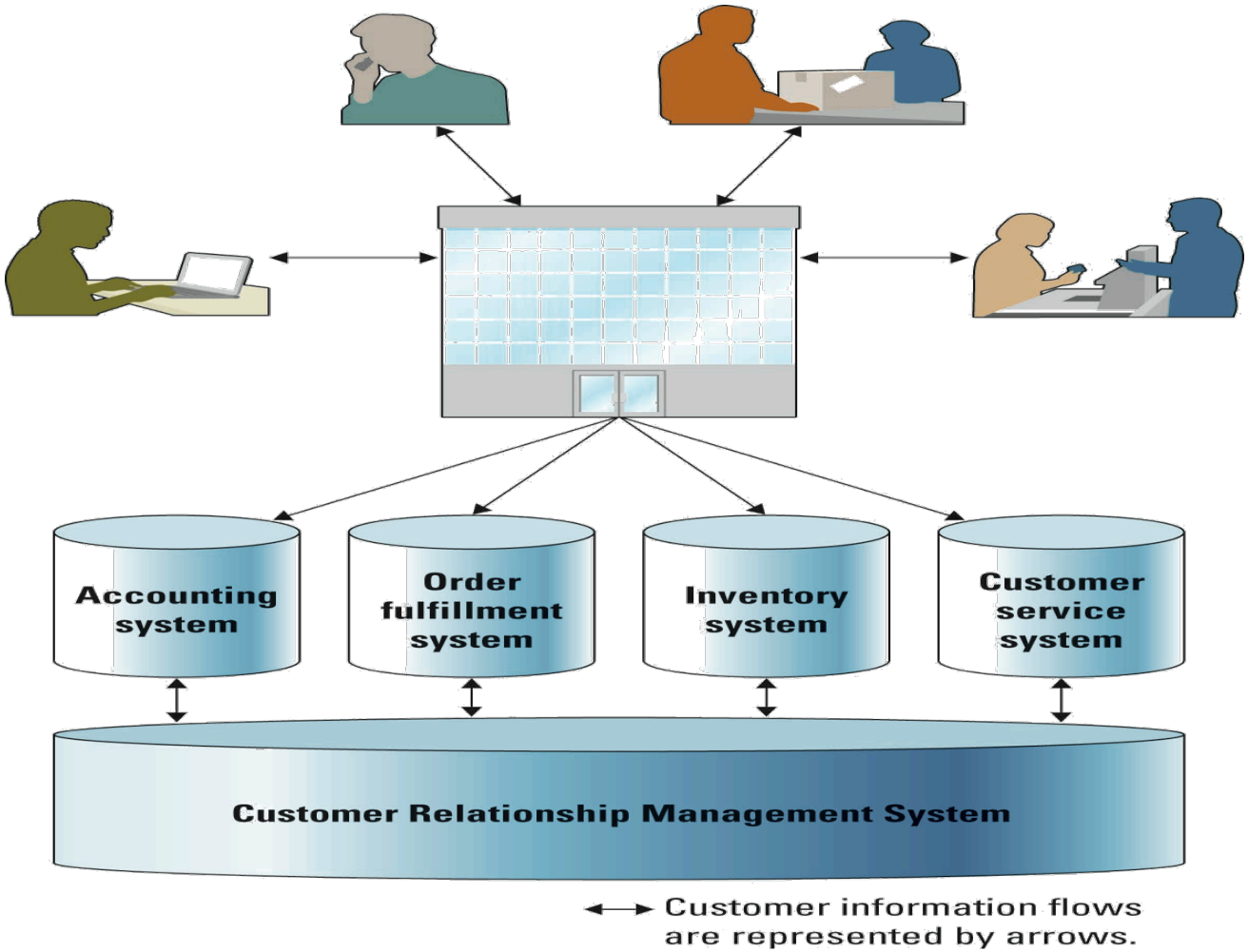
- CRM can enable an organization to:
  - Identify types of customers
  - Design individual customer marketing campaigns
  - Treat each customer as an individual: olistic view
  - Understand customer buying behaviors

Business Benefits of CRM

- Organizations can find their most valuable customers through “RFM” - **R**ecency, **F**requency, and **M**onetary value
  - How recently a customer purchased items (Recency)
  - How frequently a customer purchased items (Frequency)
  - How much a customer spends on each purchase (Monetary Value)

CRM FUNDAMENTALS

- CRM overview



Role of CRM analytics

- Trasform the raw data from each of the operational systems and contact point into a set of customer-specific beahavior patterns, tailored for the specific business goal at hand

Evolution of CRM

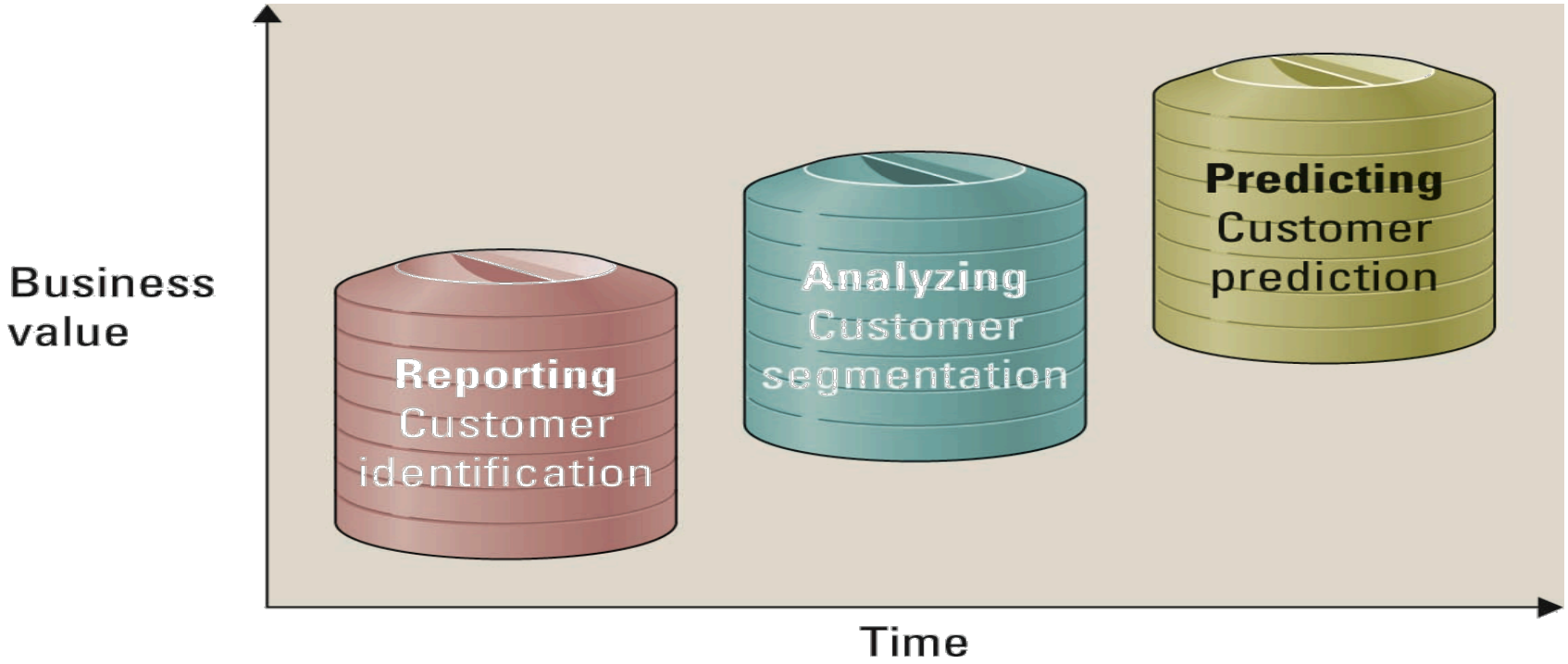
- CRM enables an organization to:
  - Provide better customer service
  - Make call centers more efficient
  - Cross sell products more effectively
  - Help sales staff close deals faster
  - Simplify marketing and sales processes
  - Discover new customers
  - Increase customer revenues

Evolution of CRM

- ***CRM reporting technology*** – help organizations identify their customers across other applications
- ***CRM analysis technologies*** – help organization segment their customers into categories such as best and worst customers
- ***CRM predicting technologies*** – help organizations make predictions regarding customer behavior such as which customers are a risk of leaving

Evolution of CRM

- Three phases in the evolution of CRM include reporting, analyzing, and predicting



Evolution of CRM

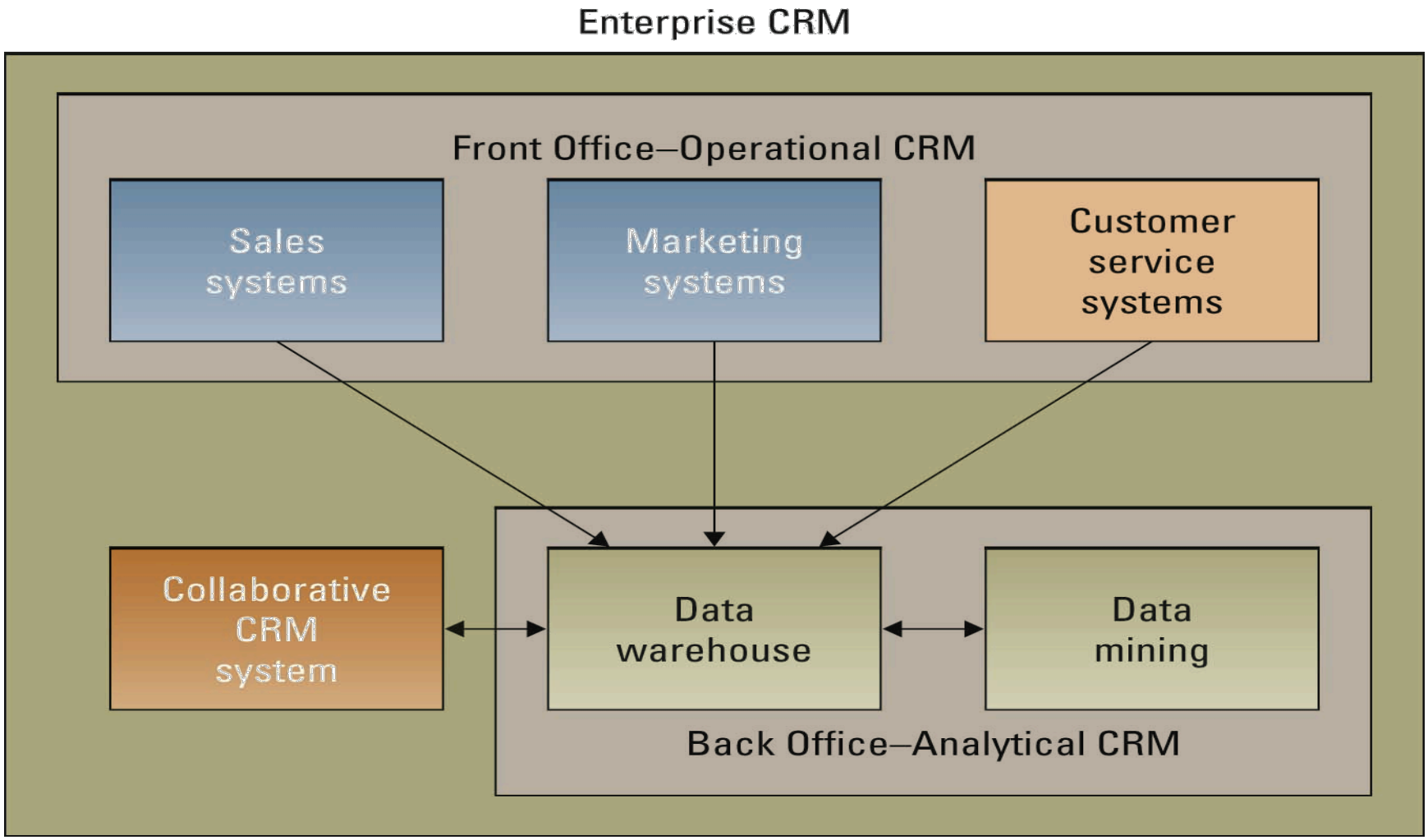
REPORTING "Asking What Happened"	ANALYZING "Asking Why It Happened"	PREDICTING "Asking What Will Happen"
What is the total revenue by customer?	Why did sales not meet forecasts?	What customers are at risk of leaving?
How many units did we manufacture?	Why was production so low?	What products will the customer buy?
Where did we sell the most products?	Why did we not sell as many units as last year?	Who are the best candidates for a mailing?
What were total sales by product?	Who are our customers?	What is the best way to reach the customer?
How many customers did we serve?	Why was customer revenue so high?	What is the lifetime profitability of a customer?
What are our inventory levels?	Why are inventory levels so low?	What transactions might be fradulent?

Operational and Analytical CRM

- ***Operational CRM*** – supports traditional transactional processing for day-to-day front-office operations or systems that deal directly with the customers

- ***Analytical CRM*** – supports back-office operations and strategic analysis and includes all systems that do not deal directly with the customers

Operational and Analytical CRM



USING IT TO DRIVE OPERATIONAL CRM

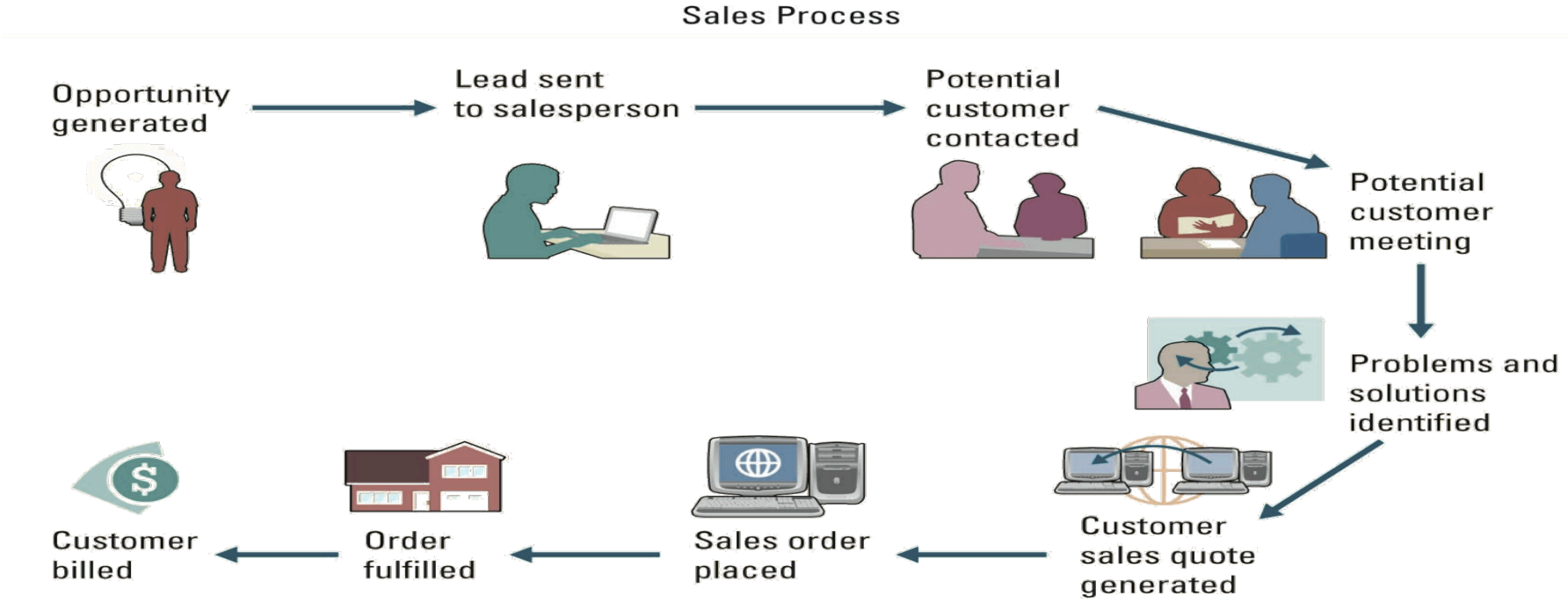
Operational CRM Technologies		
Marketing	**Sales**	**Customer Service**
1. List generator	1. Sales management	1. Contact center
2. Campaign management	2. Contact management	2. Web-based self-service
3. Cross-selling and up-selling	3. Opportunity management	3. Call scripting

Marketing and Operational CRM

- Three marketing operational CRM technologies:
  1. ***List generator*** – compiles customer information from a variety of sources and segment the information for different marketing campaigns
  2. ***Campaign management system*** – guides users through marketing campaigns
  3. Cross-selling and up-selling
     - ***Cross-selling*** – selling *additional* products or services
     - ***Up-selling*** – *increasing* the value of the sale

Sales and Operational CRM

- The sales department was the first to begin developing CRM systems with ***sales force automation*** – a system that automatically tracks all of the steps in the sales process



OPERATIONAL, TACTICAL, AND STRATEGIC BI

- Claudia Imhoff, president of Intelligent Solutions, divides the Spectrum of data mining analysis and business intelligence into three categories:
  - Operational
  - Tactical
  - Strategic

OPERATIONAL, TACTICAL, AND STRATEGIC BI

	Operational BI	Tactical BI	Strategic BI
Business focus	Manage daily operations, integrate BI with operational systems	Conduct short-term analysis to achieve strategic goals	Achieve long-term organizational goals
Primary users	Managers, analysts, operational users	Executives, managers	Executives, managers
Time frame	Intraday	Day(s) to weeks to months	Months to years
Data	Real-time metrics	Historical metrics	Historical metrics

OPERATIONAL, TACTICAL, AND STRATEGIC BI

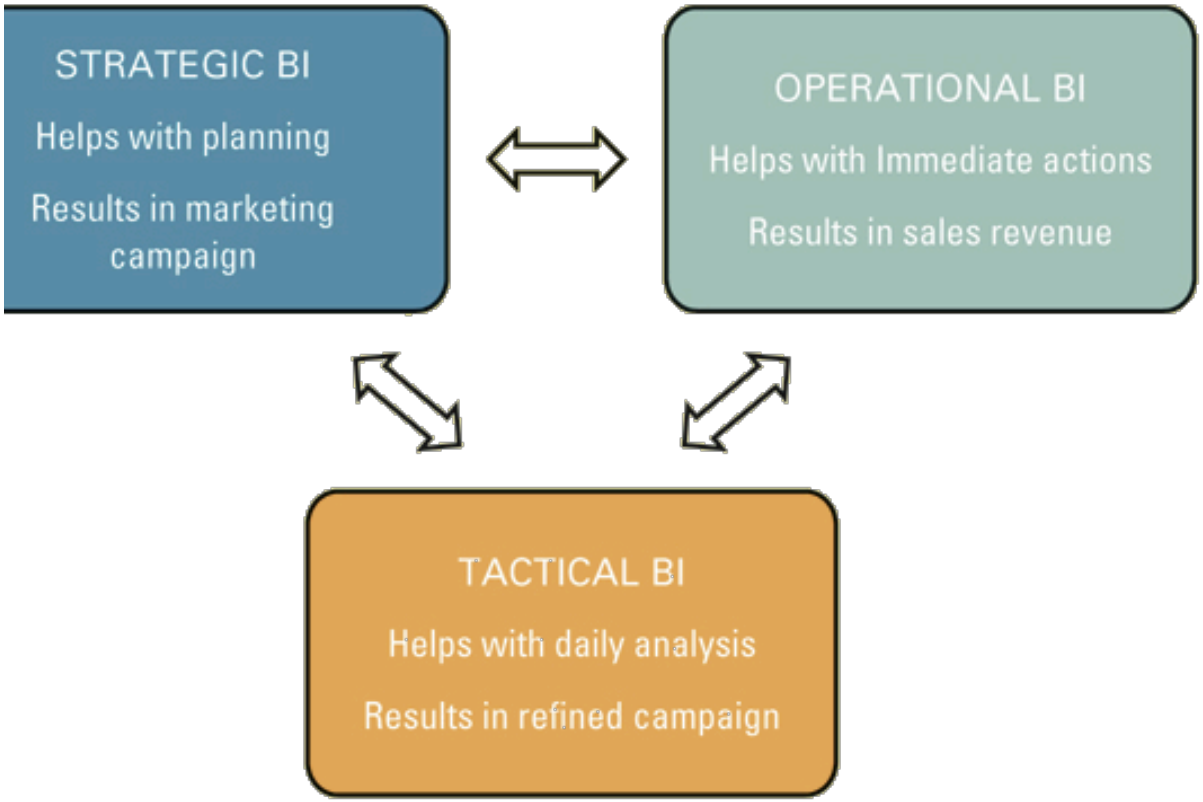


FIGURE 9.1

The Three Fo
Work Towar

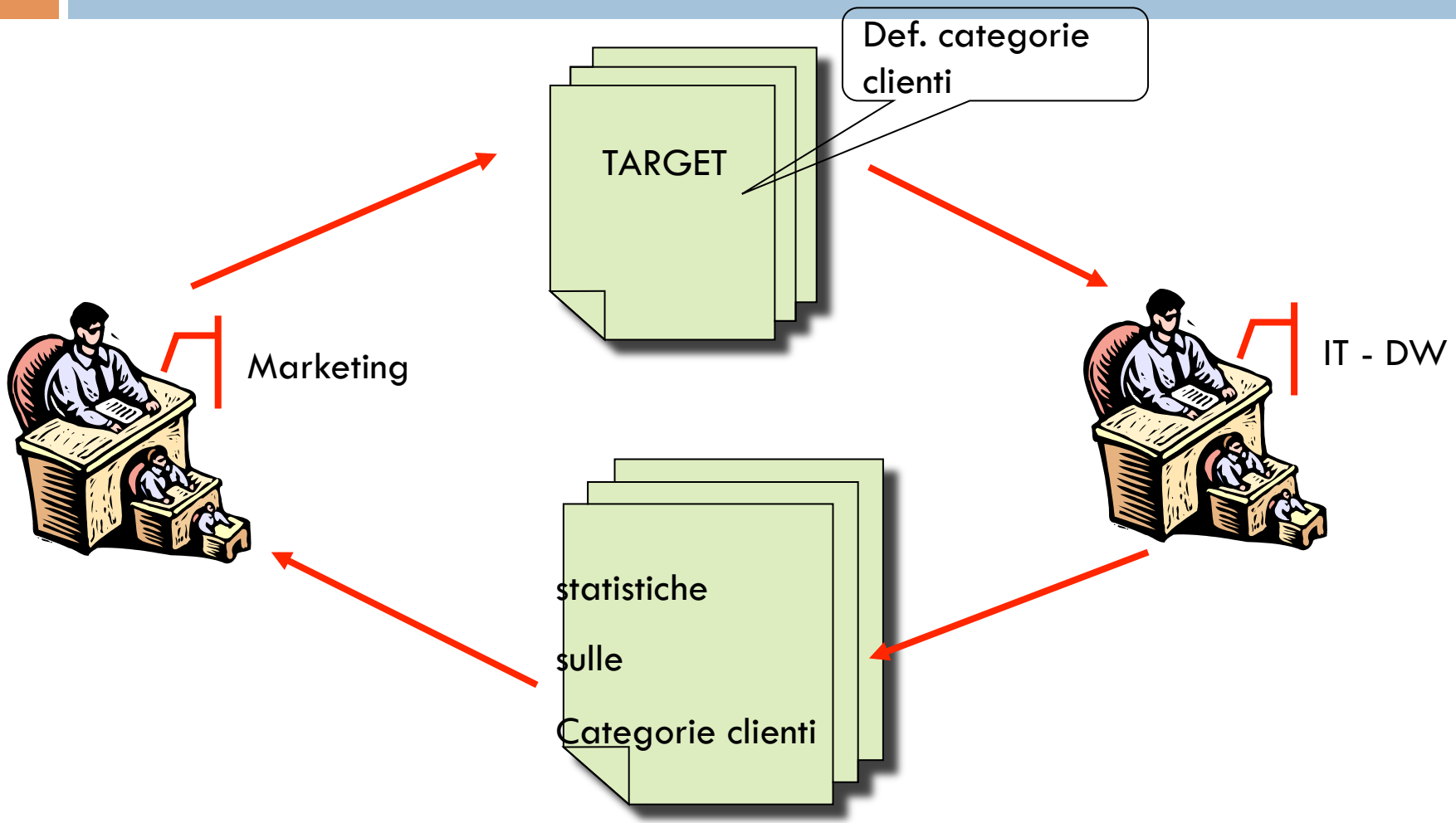
PROMORANK: INTELLIGENT TARGET MARKETING (OPTIMIZATION OF MAIL ADVERTISING CHAMPAIGNS)

KDD LAB

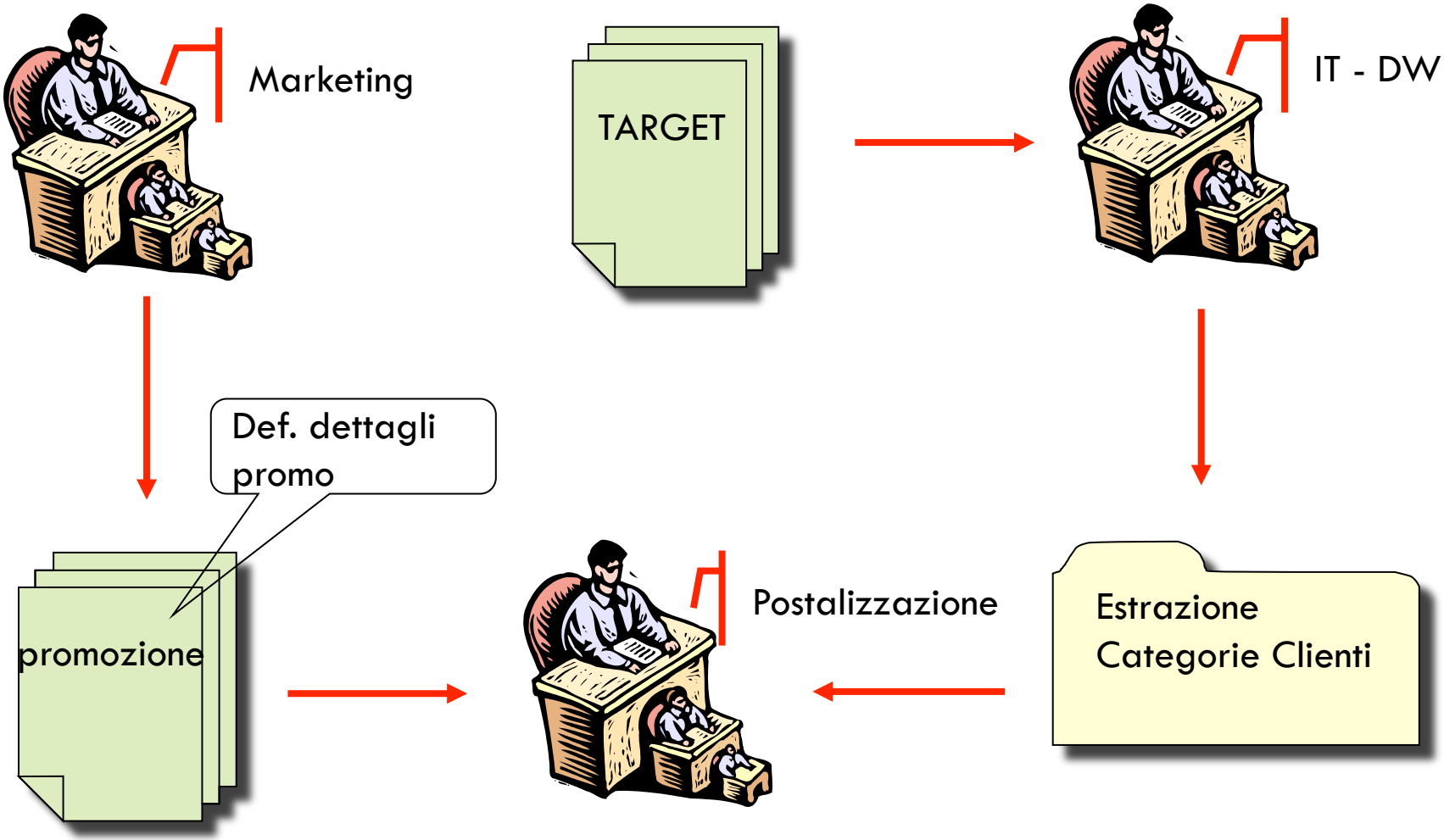
Campagna pubblicitaria postalizzata

- Decion process steps
  - Invent the champaign
  - Select the target
  - Contact the target
  - Deliver the awards
  - keep trace of redempts)
  - Measure the success (evaluating the champaign)
- The actors
  - Ufficio Marketing, Ufficio IT/DW, Postalizzatore, Ufficio IT/ DW , Ufficio Marketing

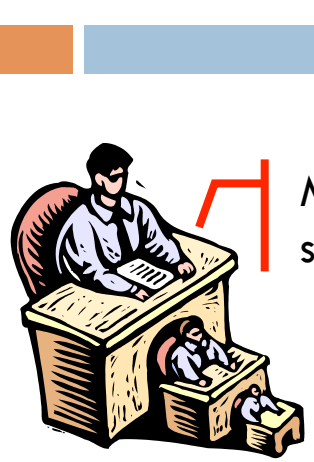
Inventare la campagna



Selezionare i clienti (da postalizzare)



Consegnare i premi e tenere traccia



Marketing
sviluppo

promozione

Def. dettagli
promo

Sistema
automatico
aggiornamento
DATA Mart

Sistema
automatico
attivazione
mercati

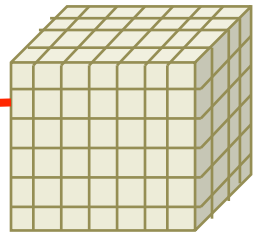
Mercati

DATA MART
PROMOZIONI

Analizzare i risultati

Marketing
analisi

analisi redemption



DATA MART
PROMOZIONI

DOMANDA?
DOVE POSSIAMO MIGLIORARE?

Dove possiamo migliorare?

- OTTIMIZZARE I CLIENTI DA POSTALIZZARE
- Se stimassimo la probabilità di redemption potremmo postare solo quelli a probabilità più alta!!!

- Problema
  - Come stimare la probabilità di redemption?
  - Quale sottoinsieme della popolazione scegliamo?

Ranking dei clienti

- Costruiamo un modello previsionale a partire dallo storico delle precedenti campagna promozionali.
- Ordianiamo i clienti secondo la probabilità fornita dal modello predittivo (della classe di maggioranza).

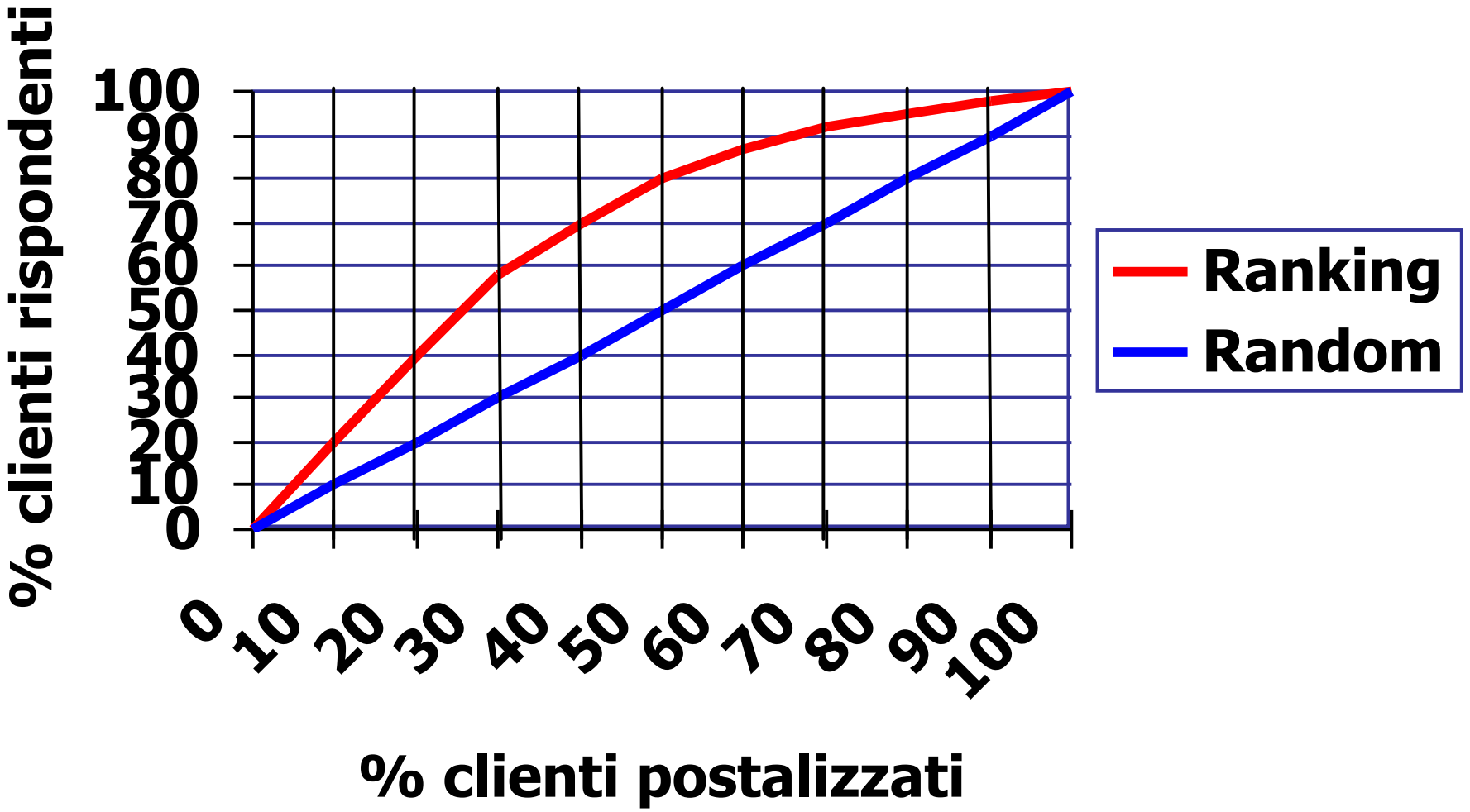
Selezione dei clienti da postalizzare

- Una volta ottenuto il ranking, occorre un criterio per scegliere:
  - La porzione di clienti da postalizzare per raggiungere un rapporto ottimale fra
    - costo di postalizzazione e
    - raggiungimento di clienti ad alta probabilità di redemption
  - La modulazione di postalizzazione fra le varie categorie di clienti definite per la promo
    - costanti, saltuari, inattivi, …

Come ci si inserisce nel processo decisionale delle promozioni

- Nella preparazione della definizione della Promozione
- Per ogni **gruppo** di clienti della promozione è disponibile un meccanismo per l'analisi di previsione della redemption e di ottimizzazione della postalizzazione
- Meccanismo di base:
  - LIFT CHART

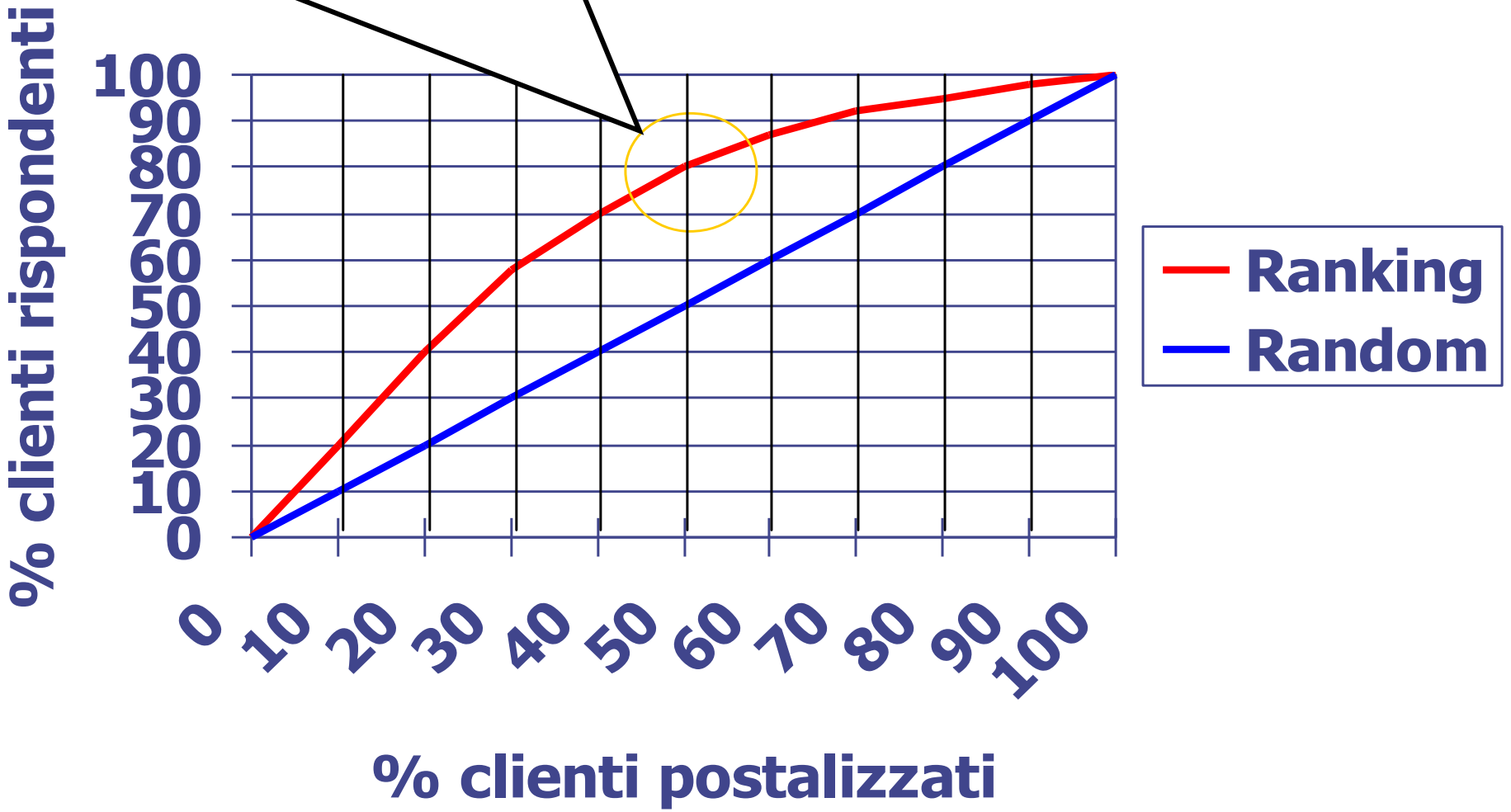
Lift Chart


% clienti rispondenti
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
% clienti postalizzati
Ranking
Random

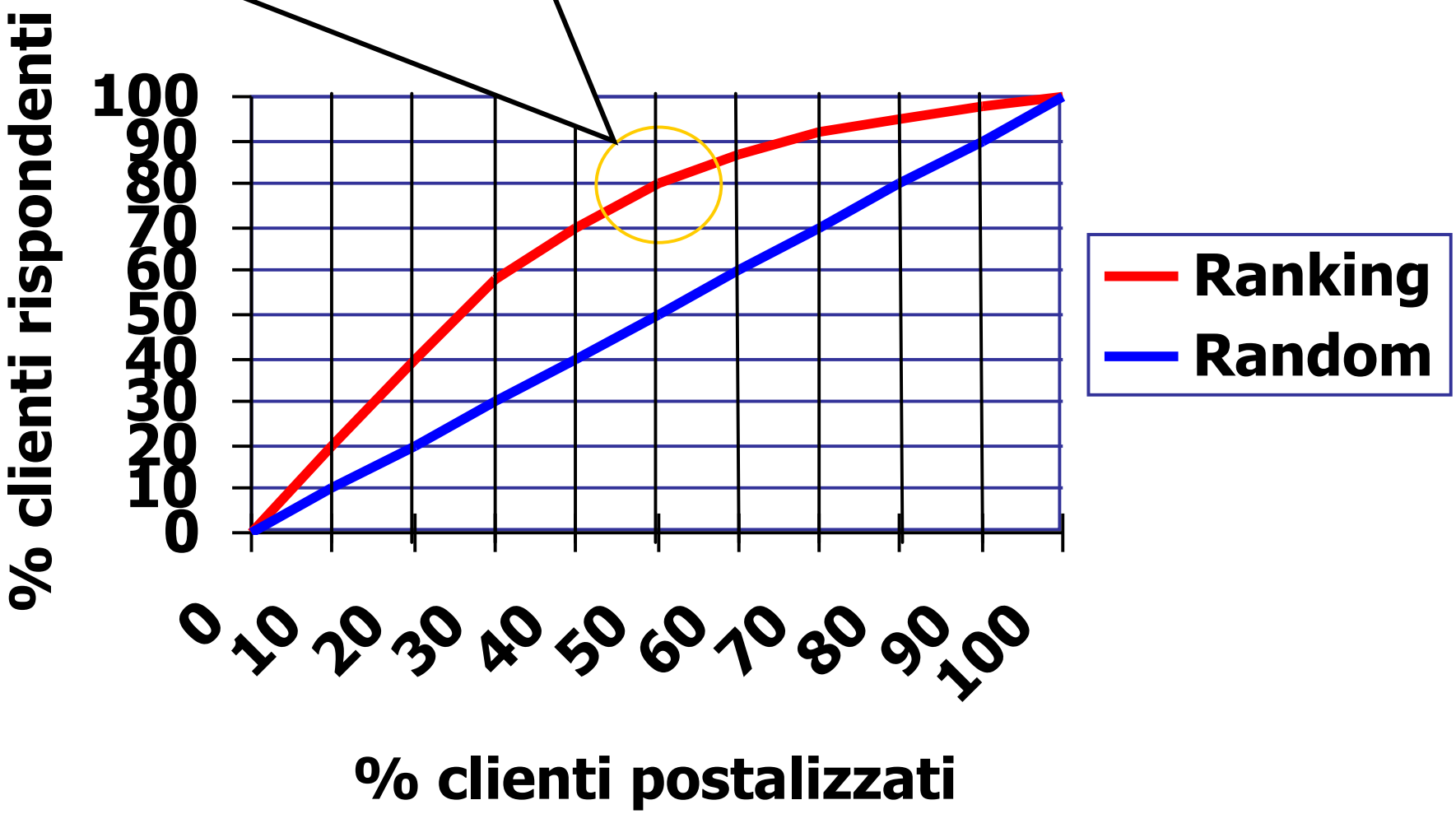
LIFT CHART

- Asse **X**: percentuali di clienti postalizzati (rispetto al totale del gruppo)
- Asse **Y**: percentuale dei clienti rispondenti che sono raggiunti dalla postalizzazione
- Linea **BLU**: andamento di Y in funzione di X, rispetto ad una scelta **casuale** dei clienti
- Linea **ROSSA**: andamento di Y in funzione di X, rispetto al ranking dei clienti col modello di data mining

Postalizzando il primo 50% dei clienti secondo il ranking si **stima** di raggiungere l'80% dei clienti che redimeranno.



Con la metà dei costi di postalizzazione si **stima** di raggiungere l' 80% dei clienti che redimeranno.



Leggere il Lift Chart (1)

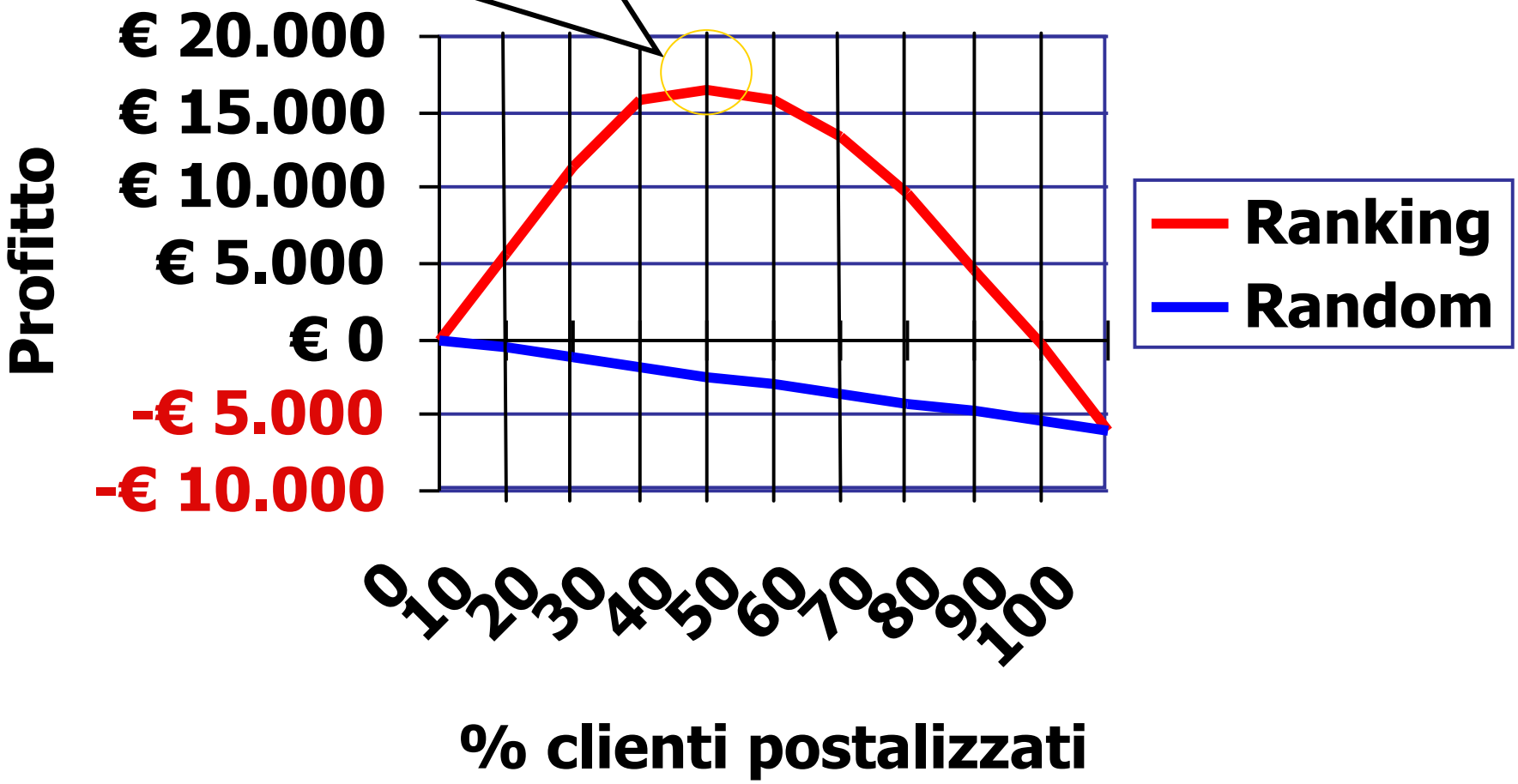
- Il Lift Chart rappresenta un aiuto grafico per ragionare sul rapporto ottimale fra costi di postalizzazione e percentuale di redemption
  - a fronte di sostanziali riduzioni di postalizzati (=budget) permette di ridurre di poco il numero di redenti
  - a parità di budget, permette di incrementare il numero di promozioni oppure di allargare la numerosità delle classi di clienti.

Leggere il Lift Chart (2)

- A partire dal Lift Chart è possibile costruire modelli economici della postalizzazione. **A titolo di esempio:**
  - C = costo unitario di postalizzazione, es. 2,30€
  - B = beneficio unitario di redenzione, es. 6,00€
  - N = numero postalizzabili, es. 30.000
  - T = numero rispondenti postalizzando tutti (stima sulla base dello storico di promozioni simili), es. 10.500 (pari al 35% di 30.000)
  - Profitto = Beneficio – Costo
    - Postalizzando una percentuale P
    - Beneficio = B x T x Lift(P) / 100
    - Costo = C x N x P / 100

Postalizzando il primo 40% dei clienti secondo il ranking si **stima** di massimizzare il beneficio

C=2,30€ B=6,00€ N=30.000 T=10.500.



Le nuove funzionalità per l'ufficio marketing

- Nuova funzionalità per il decisore:
  - accedere al meccanismo di analisi previsionale mediante lift-chart separato per ogni gruppo di clienti
  - modulare la scelta del sottoinsieme di clienti da postalizzare in base:
    - Al ragionamento sul lift-chart, combinato con
    - L'obiettivo di dirigere la promozione in modo preferenziale verso determinati gruppi di clienti (fedeli vs. occasionali, etc.)
  - verificare le conseguenze delle scelte di postalizzazione operate in termini complessivi (copertura, risparmio, etc.), ed eventualmente modificarle

Ma dov'è il **data mining**?!?

- Risposta: **dietro le quinte**!
- Il ranking dei clienti rispetto alla probabilità di redemption è il risultato dello sviluppo di una serie di modelli predittivi che classificano i clienti come rispondenti o meno in base allo storico delle promozioni desumibile dal venduto nel datamart dei Fidelizzati

Dietro le quinte

- Il lift-chart della scheda promo e gli elenchi dei clienti da postalizzare sono calcolati, ad ogni richiesta
  e marketing
  odelli
  he risiedono sul server (di progetto o di DW)
- I modelli predittivi sono riaggiornati periodicamente, ad ogni richiesta dell'utente IT-DW, sulla base dei
  i dal contenuto attuale del DW, mediante
  data mining

On-line

a cura dell'ufficio marketing

Off-line

a cura dell'ufficio IT/DW

La Grande Distribuzione

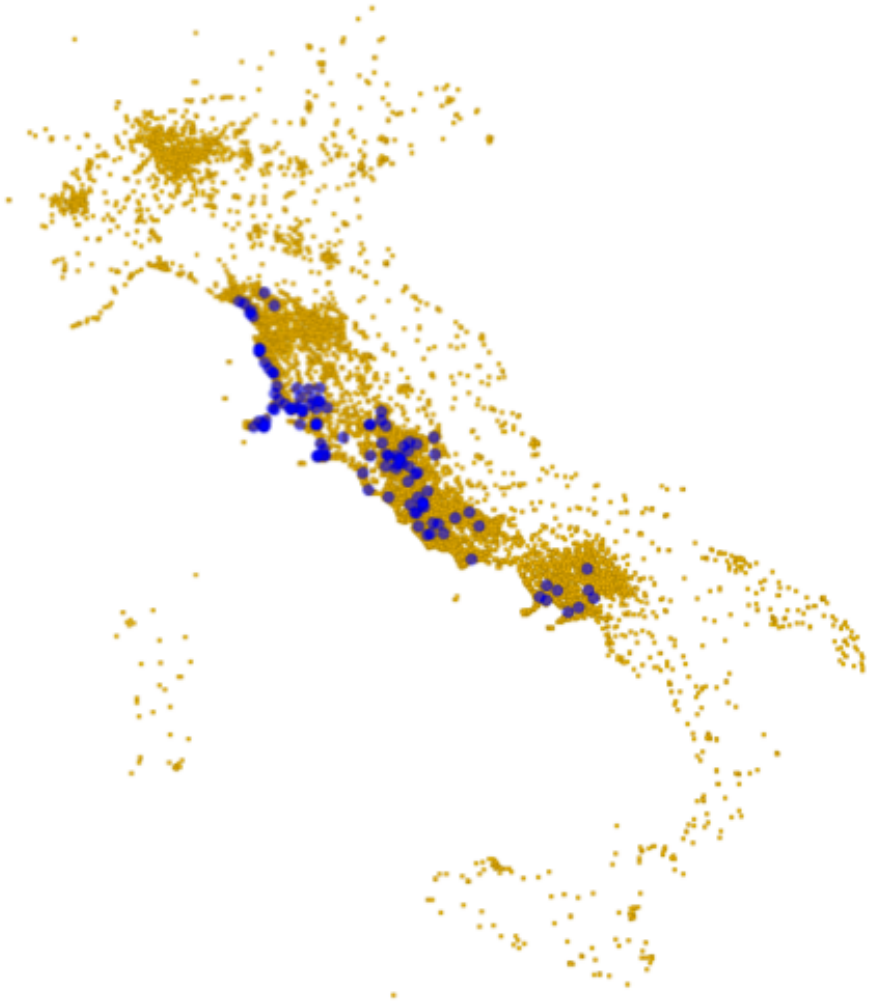
Qualche numero

- 115 Negozi in 4 regioni (Toscana, Lazio, Umbria e Campania)
- 3 Canali (+ e-commerce)
- 1 canale *franchising*
- Fatturato di c.1.100+ M€ (alimentari e non)
- Quarta Cooperativa in fatturato tra le «nove grandi» (Prima Unicoop Firenze con c.2.300+ M€)
- Totale fatturato 9 Grandi Cooperative: 11.400+ M€ (*market share* c.16,5%)

Qualche numero del DW

- Oltre 7 anni di venduto (dal 1/1/2007 fino ad Aprile 2014)
- ~1 M di clienti attivi e riconoscibili
- 138 negozi (Toscana, Umbria, Lazio, Campania)
- ~450K diversi tipi di prodotti
- ~280M di scontrini
- ~280G di scansioni

La distribuzione dei clienti e dei negozi



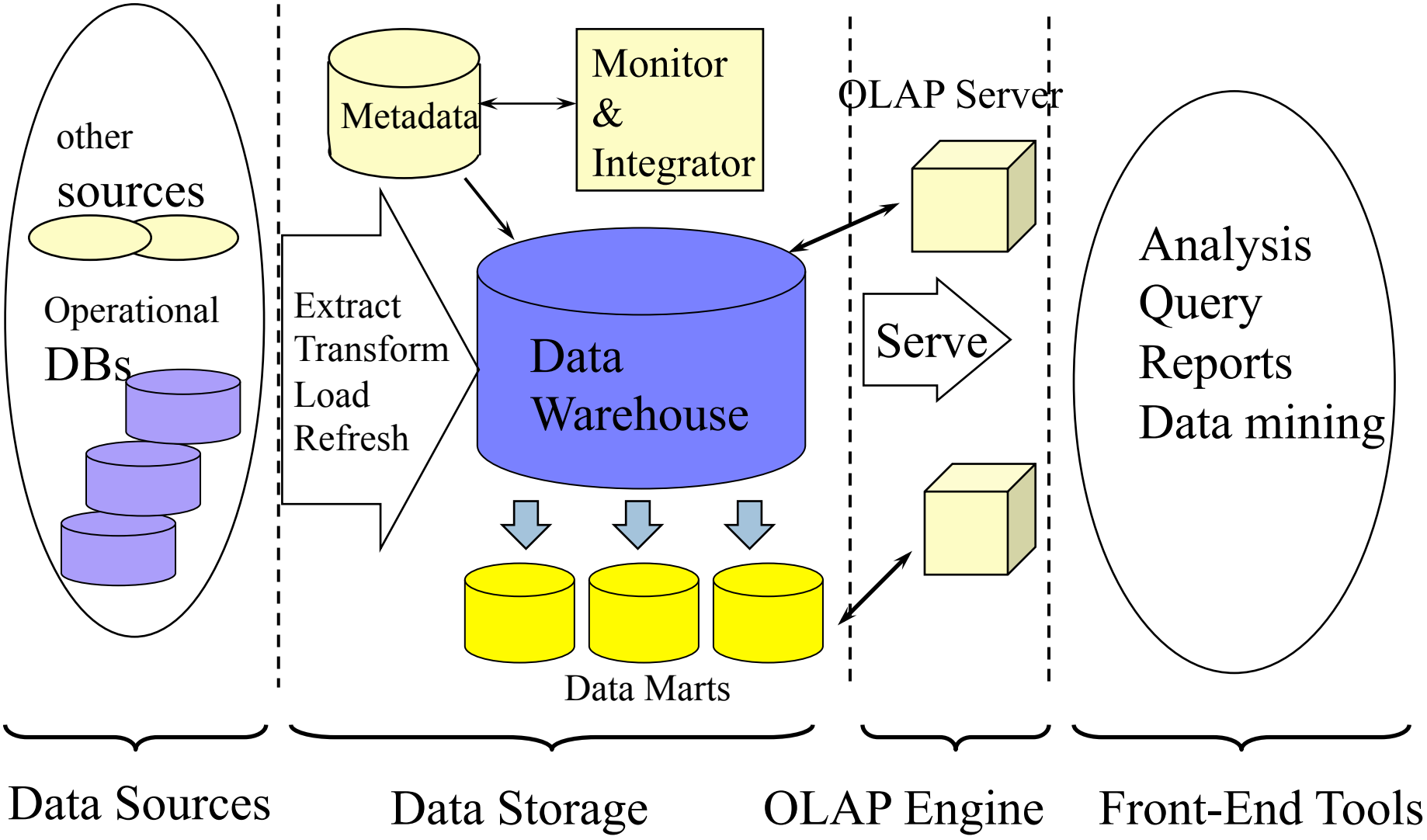
TOP BUSINESS CHALLENGES

Le domande più importanti relative al *Pricing*

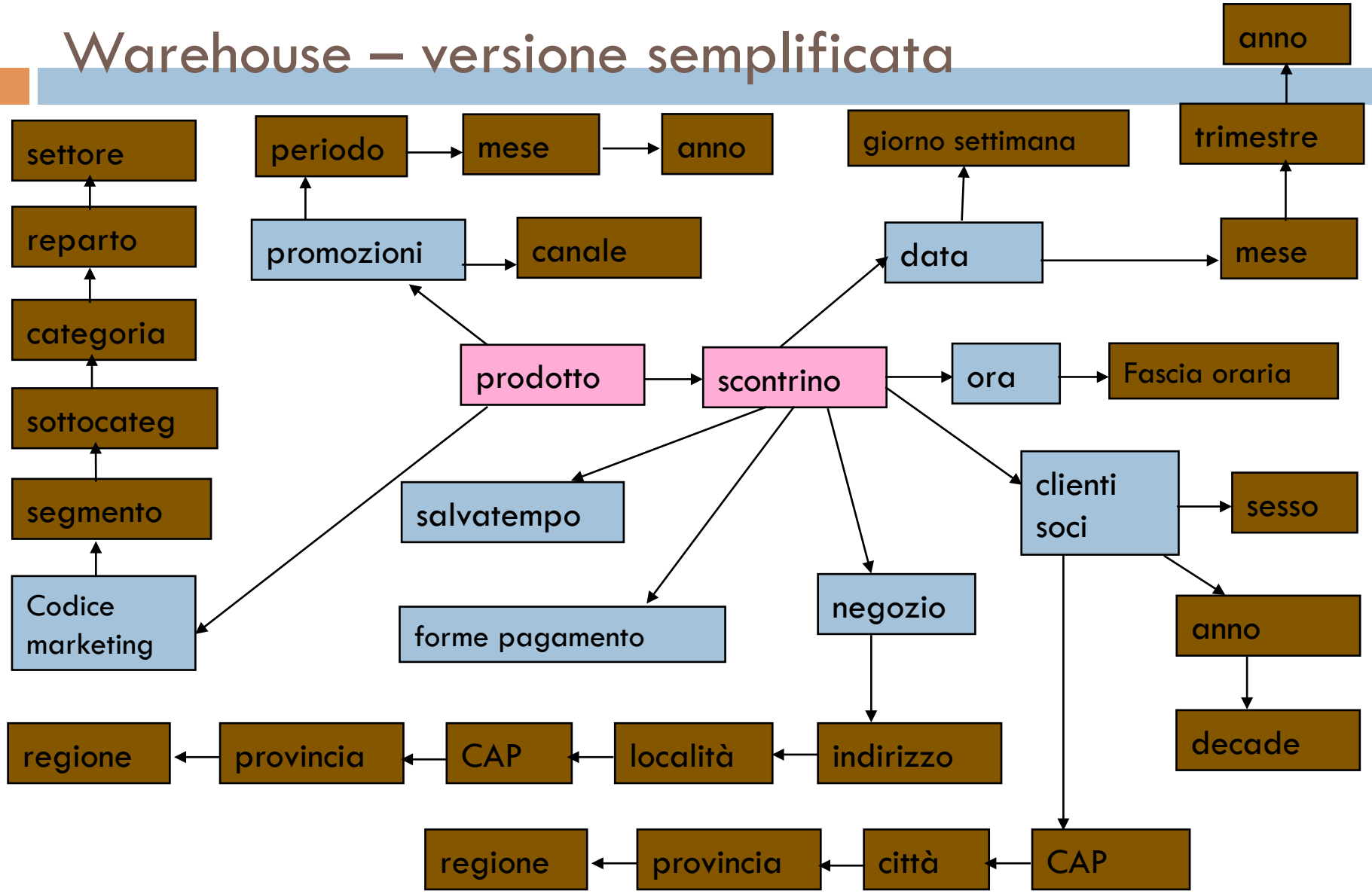
- **Quali sono le categorie/articoli è meglio investire?**
- **Qual è il posizionamento più corretto rispetto alla concorrenza** (*i*.e.: qual è l'investimento minimo necessario?)
- **Quali e quante *price zones*?**
- **Quante variazioni di prezzo è necessario gestire a PV per sostenere una strategia di prezzo «raffinata»? Sono numeriche sostenibili per i nostri PV?**
- **Ci sono risorse o no nel *pricing*?**

PRICE STRUCTURE

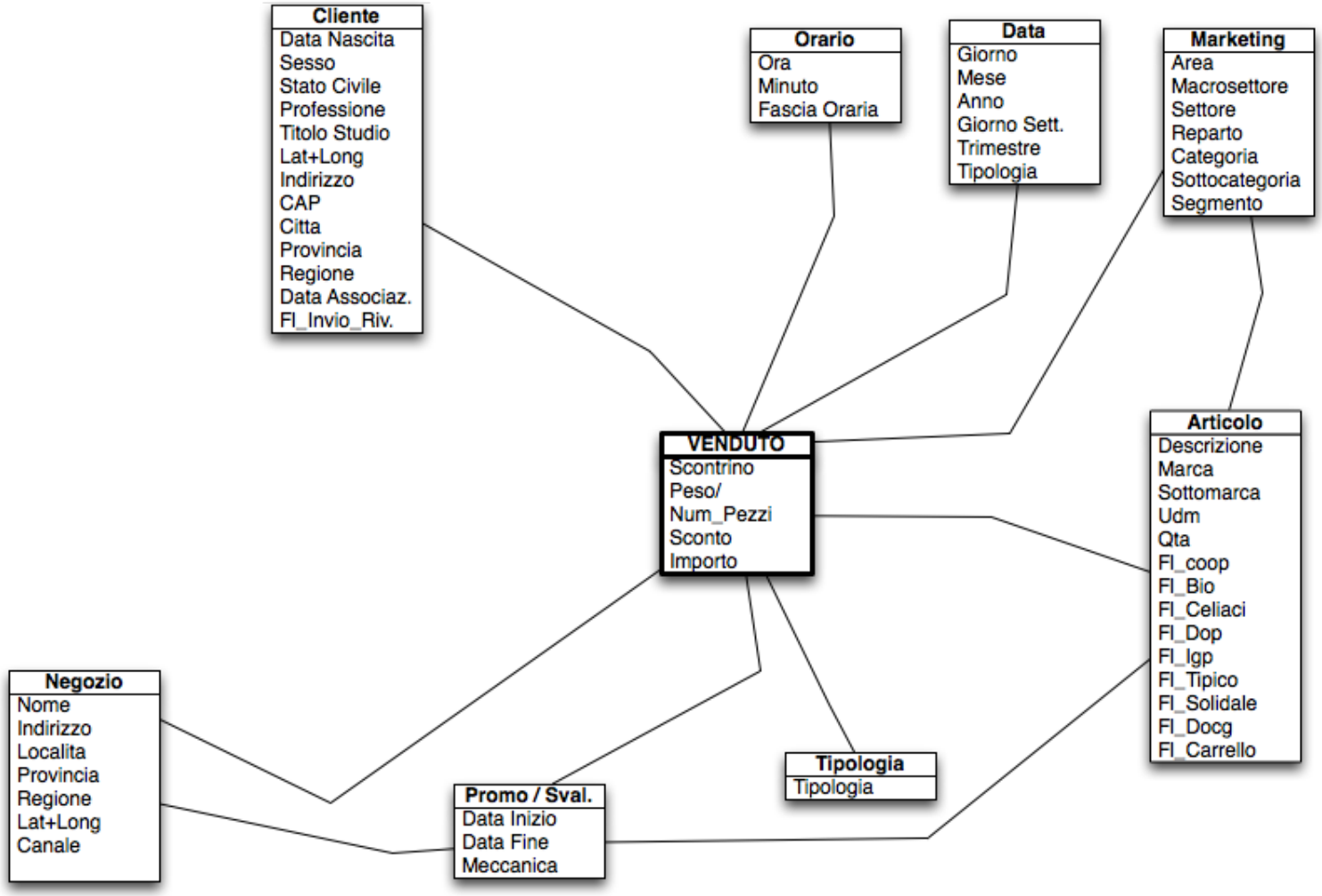
Tecnologia per BI



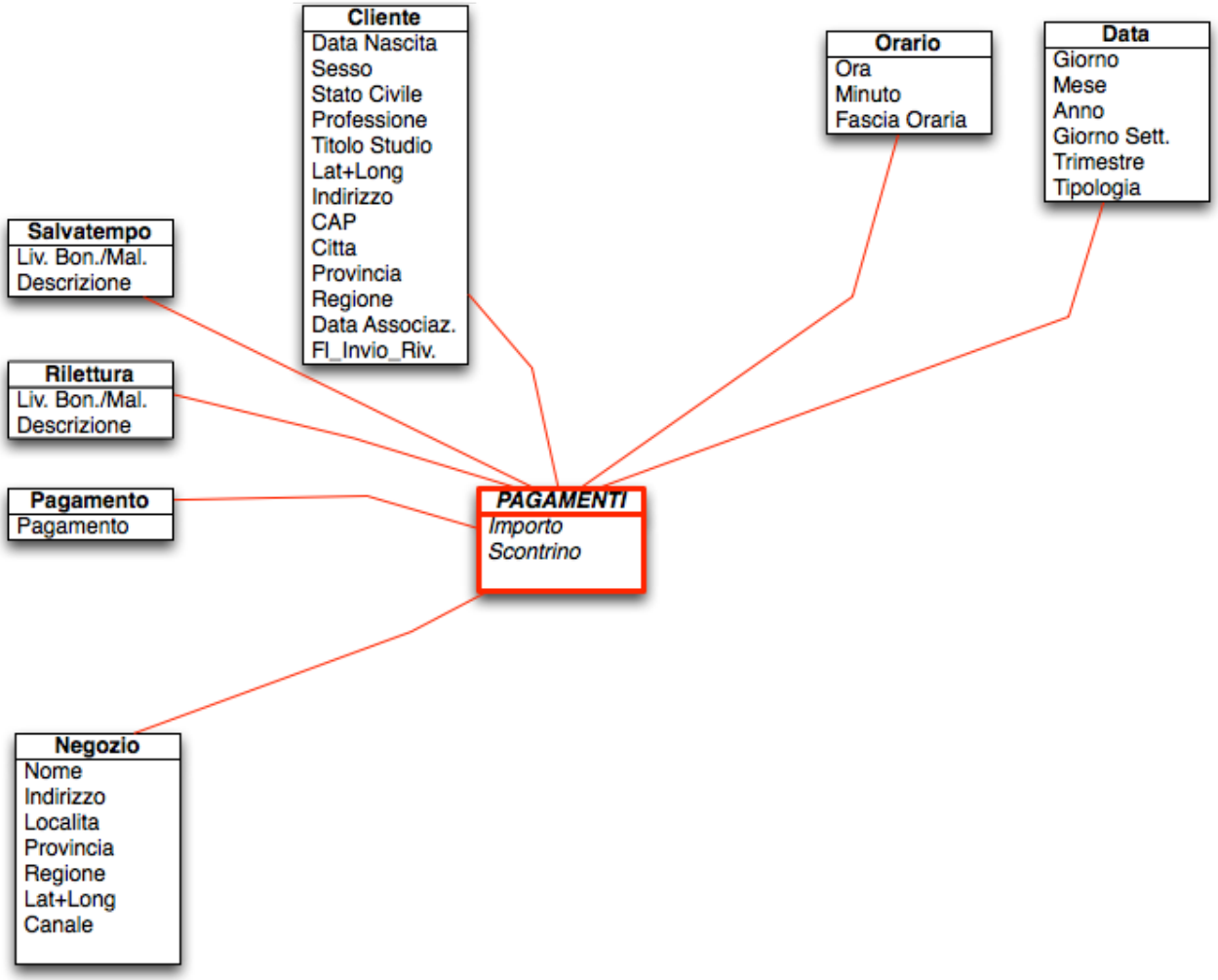
Schema dei fatti e delle dimensioni del Data Warehouse – versione semplificata



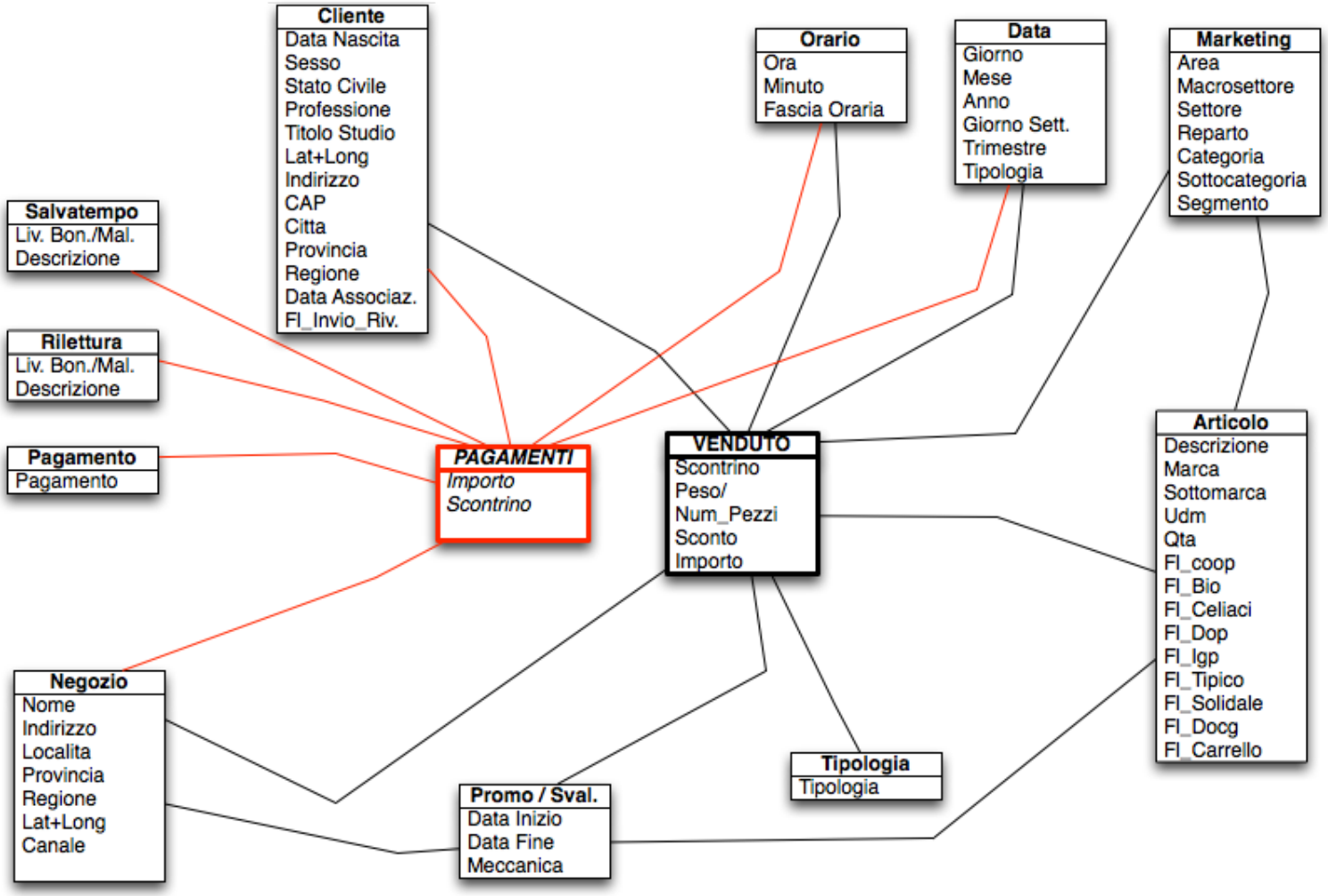
Il Data Warehouse (1)



Il Data Warehouse (2)

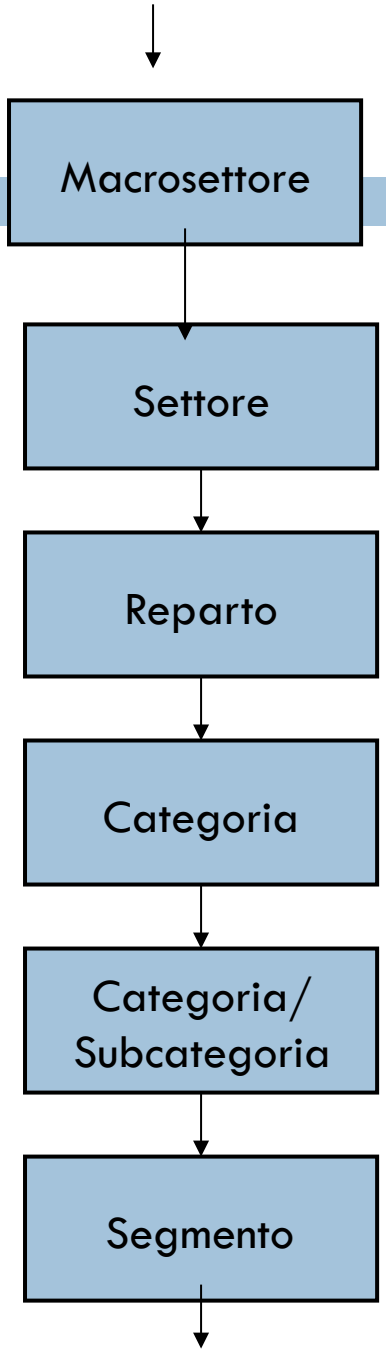


Il Data Warehouse (3)



La gerarchia Marketing

- **Area** – 3 valori
- **Macrosettore** – 4 valori
- **Settore** – 13 valori
- **Reparto** – 76 valori
- **Categoria** – 529 valori
- **Sottocategoria** – 2665 valori
- **Segmento** – 7656 valori
- **Articolo** – 571092 valori



La gerarchia Marketing - Esempi

I primi 3 livelli

Area	Macrosettore	Settore
Alimentari	Freschi	Freschi Freschissimi
	Generi vari	Grocery Alimentari
Altro	Altro	Confezionato per vendita Erogazione carburanti Iniziative speciali Non definito
Non alimentari	No Food	Casa Chimica Multimedia Persona Salute Stagionali e Brico

Esempio: coche decaffeinate

Area	Alimentari
Macrosettore	Generi vari
Settore	Grocery alimentari
Reparto	Liquidi
Categoria	Bibite
Sottocategoria	Cole
Segmento	Decaffeinata

COCA-COLA SENZA CAFFEINA LATTINA ML.330X6
COCA COLA SENZA CAFFEINA BOTTIGLIA PET ML.500
COCA COLA SENZA CAFFEINA LATTINA ML.330
COCA COLA SENZA CAFFEINA PET LT.1,5
COCA-COLA SENZA CAFFEINA LATTINA ML.330
COCA COLA SENZA CAFFEINA PET ML.500X4
COCA COLA SENZA CAFFEINA LATTINA ML.330 X6
COCA COLA SENZA CAFFEINA PET LT. 1,5 + ML. 250 OMAGGIO
BIBITA COCA COLA SENZA CAFFEINA CLUSTER LT.1,5 X 6
COLA SENZA CAFFEINA COOP NO COLOR. NUOVA RICETTA BOTTIGLIA PET LT
COLA SENZA CAFFEINA COOP PET. LT. 1,5
COLA SENZA CAFFEINA HAPPYHAND PET LT 1,5
COLA SENZA CAFFEINA PERLA PET 1,5 L
BIBITA PEPSI BOOM JUNIOR PET ML.330X4
BIBITA PEPSI BOOM PET LT.1,5
PEPSI BOOM PET LT.2
BIBITA PEPSI BOOM LATTINA ML. 330

Churn analysis

Introduction

Challenge: FAST identification

- Churn – defezione - abbandono
- In alcuni casi, nel momento in cui la defezione si manifesta, è troppo tardi per intervenire
  - non è più possibile recuperare il defezionante, o
  - non è più conveniente recuperarlo
- Fondamentale identificare il defezionante *immediatamente*, o addirittura *in anticipo*
- Nuova formulazione del problema:
  - Churn Analysis = *Previsione dell'abbandono*

Interruzione del rapporto: modalità

- Interruzione esplicita
  - Tipica dei rapporti che richiedono un contratto o impegni da parte del fruitore
    - Es.: Telefonia, nei casi in cui è previsto un canone
    - Es.: Tesseramenti rinnovabili non gratuiti
- Interruzione implicita
  - Tipica dei rapporti non formalizzati o privi di costi per il fruitore
    - Es.: Tessere sconto e carte fedeltà

Interruzione implicita

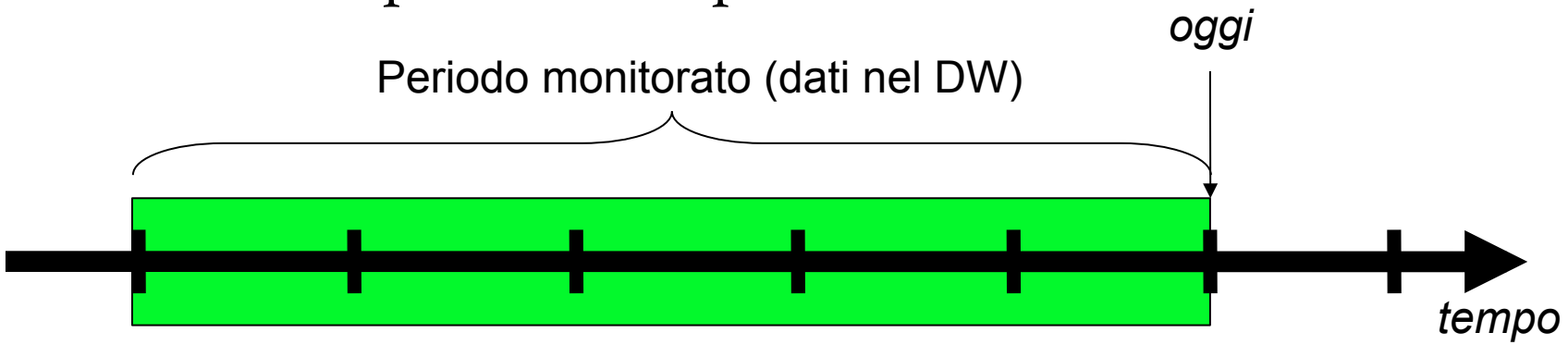
- E' la situazione più comune nel settore delle vendite al dettaglio
  - Carte fedeltà che non richiedono rinnovi né costi
  - Il defezionante semplicemente non la usa più
- Domanda: è sempre facile capire quando il cliente/fruitore ha abbandonato?
  - Non fa acquisti per un mese?
  - Non fa acquisti per un anno?
  - Visita il punto vendita meno di 2 volte al mese?
  - Spende meno del 50% di quanto faceva 3 mesi fa?

Abbandono “soft”

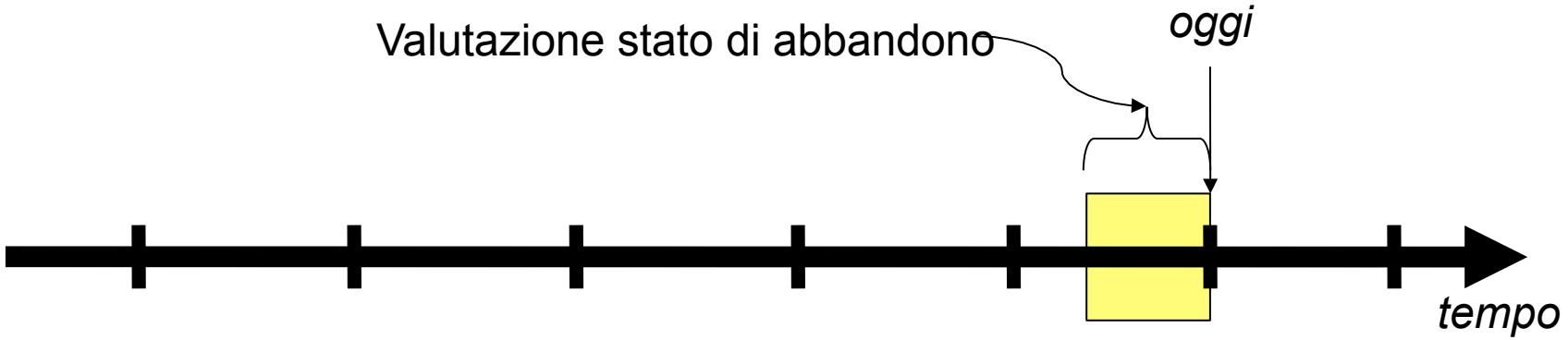
- Nozione alternativa di abbandono:
  - Passaggio da un tipo di rapporto ad uno diverso
  - Caso estremo: da “fedele” a “abbandono totale”
- Situazione naturale nella vendita al dettaglio
  - Il segmento “fedele” fornisce (parziali) garanzie su un indotto minimo dell'attività
  - Il degrado del cliente “fedele” a cliente “saltuario” ha effettivi negativi sulla gestione aziendale
    - Valgono le stesse motivazioni dell'abbandono “hard”

Previsione dell'abbandono

- Il tracciamento del cliente ci consente di ricostruire la sua "storia" per un certo periodo

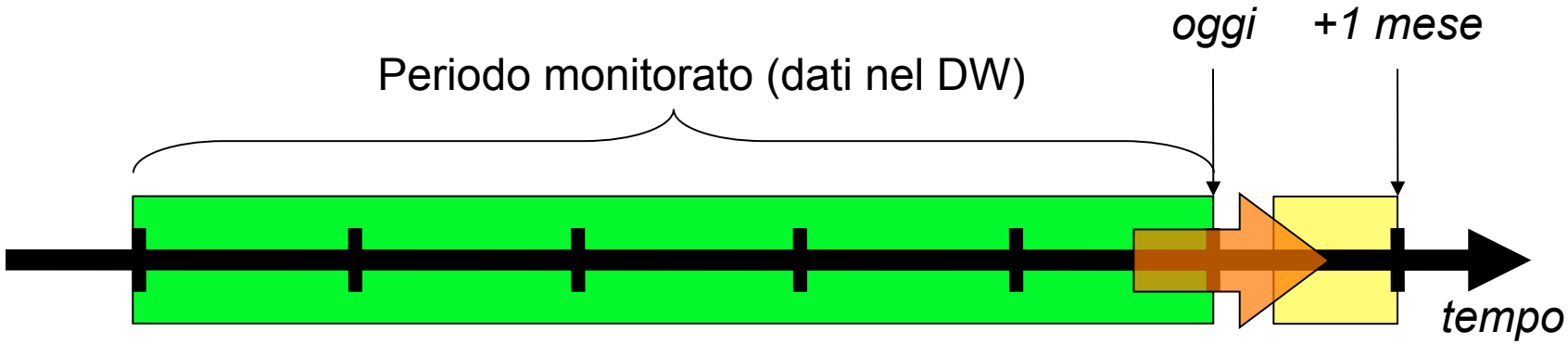


- La nozione di abbandono adottata sarà valutata su un segmento recente di tale storia



Previsione dell'abbandono

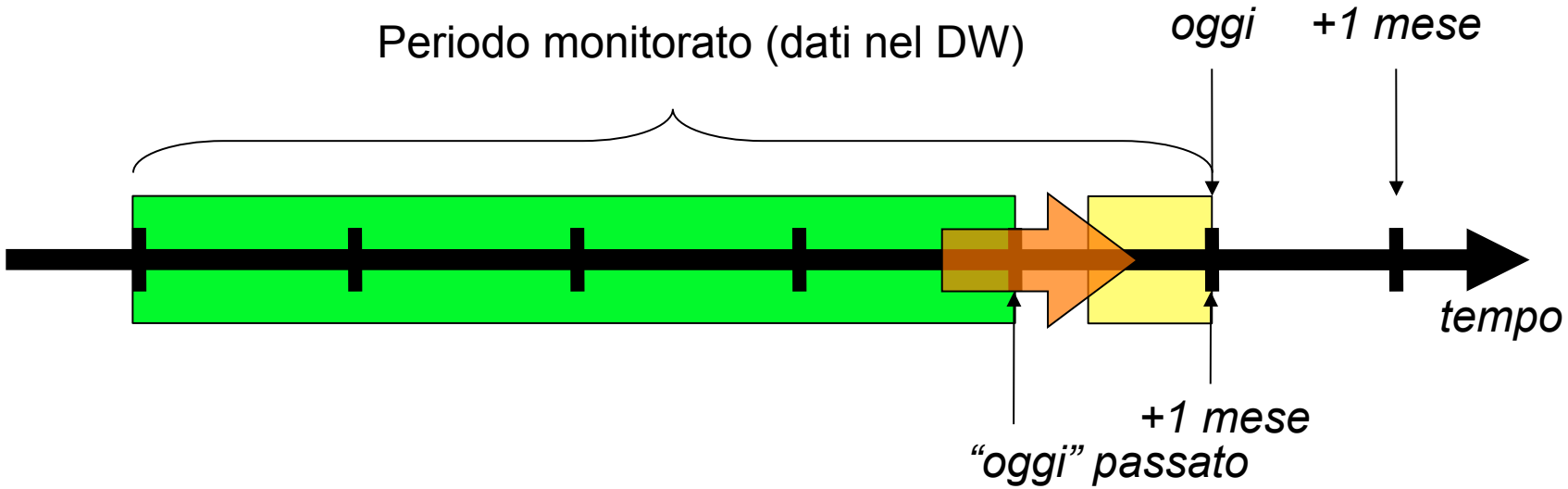
- Obiettivo: previsione dello stato di abbandono *futuro,* conoscendo la storia recente del cliente



- La storia recente fornisce indizi sul comportamento che il cliente si presta a tenere
  - alcuni indizi permettono di discriminare i futuri defezionanti, altri no
  - alcuni indizi sono espliciti nei dati a disposizione, altri vanno derivati da essi

Previsione dell'abbandono

- Come determinare *oggi* le correlazioni tra situazione attuale e stato futuro?
  - Cerchiamo queste correlazioni nel *passato*
  - Le relazioni "passato → oggi" verranno sfruttate per predire il futuro dall'oggi



Schema dell'applicazione

- Definizione/estrazione delle variabili di lavoro
  - Variabili predittive: gli *indizi* disponibili *oggi/passato*
  - Variabile target: lo stato di abbandono *futuro/oggi*
- Estrazione del modello predittivo
  - Ricerca di correlazioni tra variabili predittive e variabile target, da sfruttare in fase di predizione
- Applicazione del modello predittivo
  - Le relazioni variabile predittiva → target vengono applicate alla situazione odierna (in termini di variabili predittive) per stimare la var. target

BICOOP – Churn Analisys

Definizione del concetto di abbandono e creazione di modelli previsionali

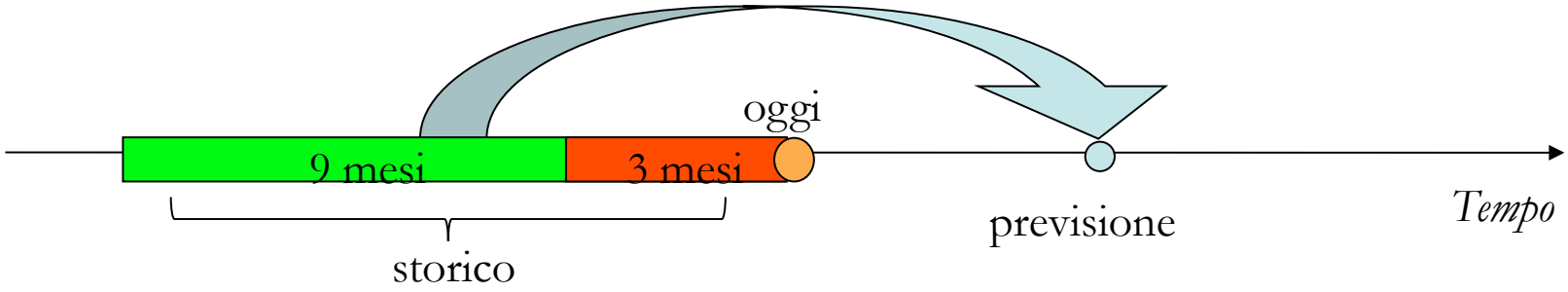
Problem Definition:

- 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

Analisi previsionale

- Raccolta dei dati storici per l'estrazione di:
  - Variabili di vendita e anagrafiche, i predittori (periodo verde)
  - Variabili obiettivo (periodo rosso)

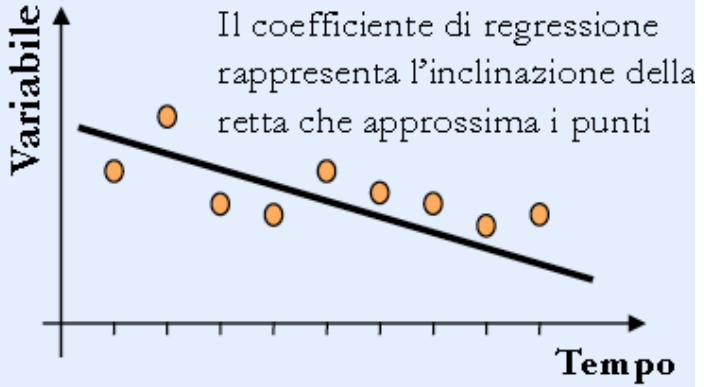
- Costruzione di un modello predittivo
  - Addestrato in modo opportuno su dati storici
  - Utilizzabile per ottenere informazioni previsionali



Data preparation

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



Il coefficiente di regressione rappresenta l'inclinazione della retta che approssima i punti

Preparazione dati – target (periodo rosso)

- Si sono estratte dal data warehouse, per il periodo di 3 mesi (Settembre 2007 – Novembre 2007) le seguenti informazioni:
  - Numero di spese
  - Variazione di spesa rispetto al periodo verde
    - Volume di spesa
    - Battute di cassa
    - Numero di visite

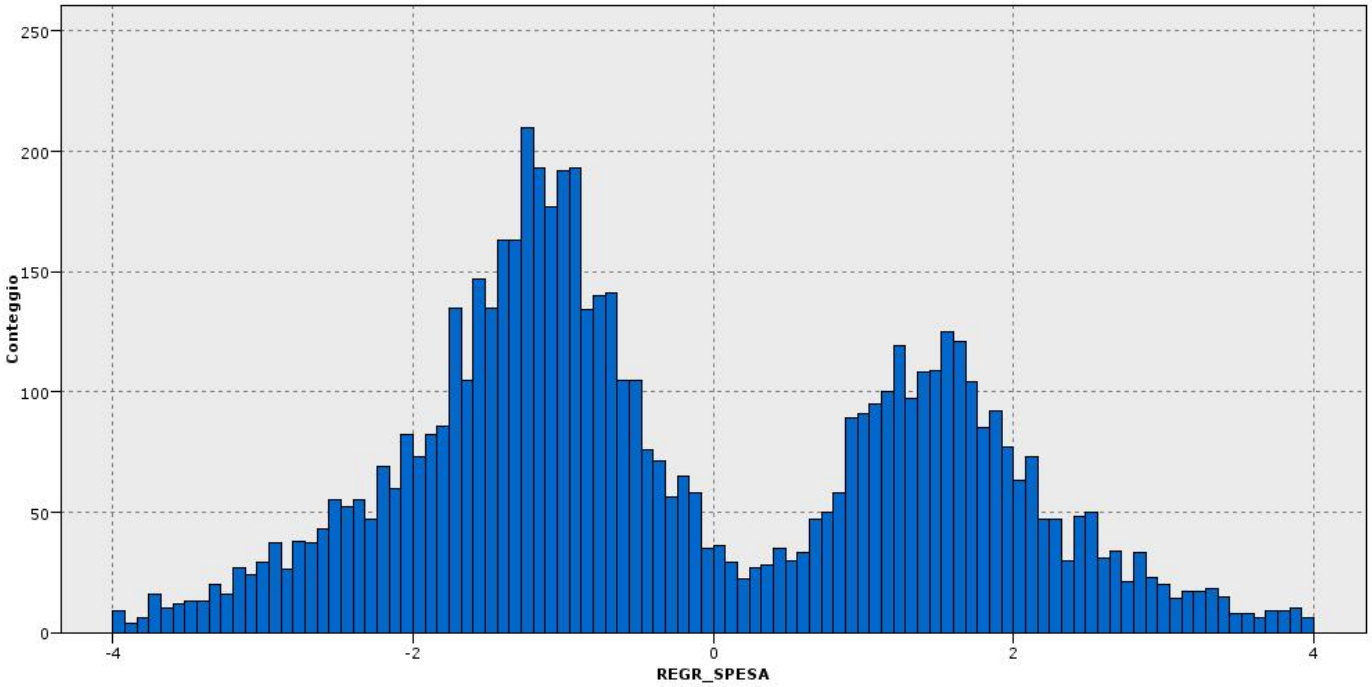
Dataset

- Il dataset così ottenuto presenta una riga per ogni cliente che ha effettuato almeno una spesa nei nove mesi di osservazione. In tutto abbiamo ottenuto:
  - 517.000 righe
  - 47 attibuti

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	

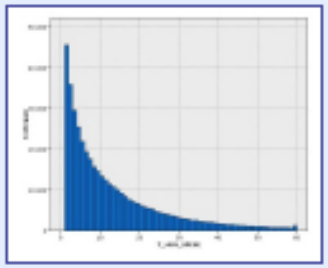
Data Exploration

- Distribuzione trend di spesa

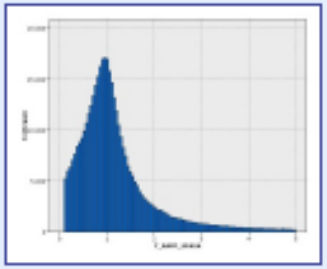


Trend dei clienti con spesa totale > 400€

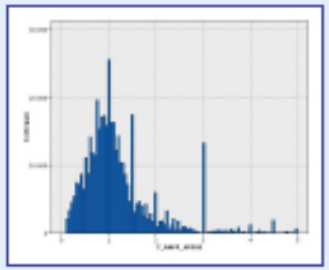
Funzioni Obiettivo



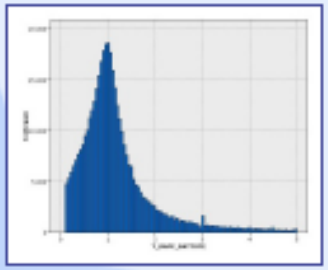
NUM_SPESE: spese del cliente nel periodo target



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



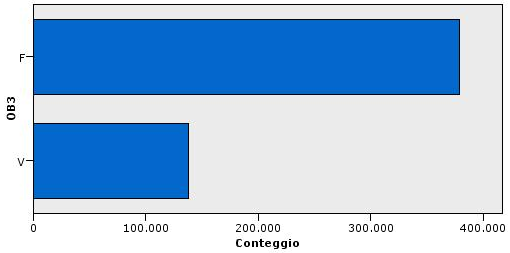
RAPP_SPESA: rapporto tra la spesa del periodo target e quella del periodo d'osservazione



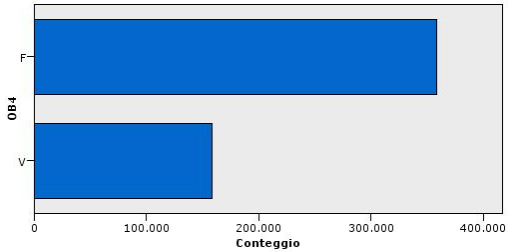
RAPP_BATTUTE: rapporto fra le battute di cassa del periodo target e quelle del periodo d'osservazione

F. Obiettivo – Thresholds

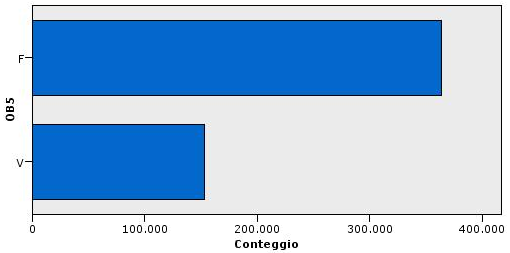
- Scelta una soglia di allarme per indicare un possibile cliente defezionario i rapporti si trasformano in tre indicatori di abbandono
- Abbiamo scelto come soglia una diminuzione sulle 3 misure del 50%
- Otteniamo le seguenti distribuzioni: F (Basso rischio
, V (Alto rischio Abbandoni)



OB1: RAPP_SPESE



OB2: RAPP_SPESA

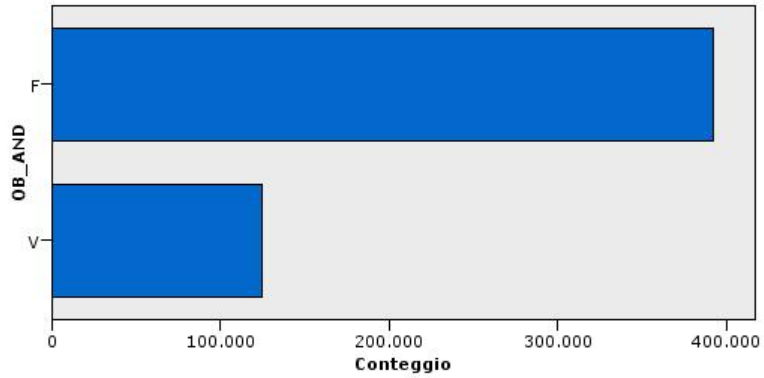


OB3: RAPP_BATTUTE

F. Obiettivo – Sintesi

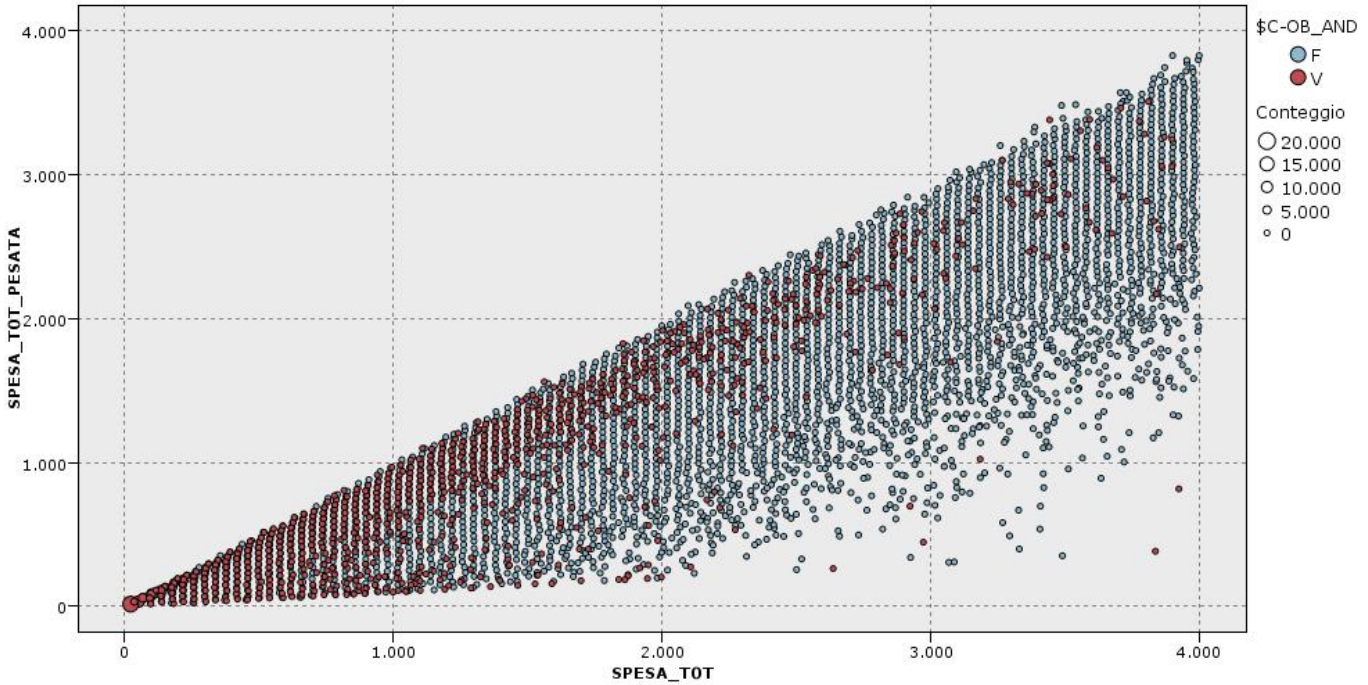
Per la funzione obiettivo finale si è deciso di considerare come potenziali defezionari tutti i clienti che superato la soglia di allarme, in ognuno dei tre indicatori OB1, OB2, OB3:

OB_AND: OB1 and OB2 and OB3



Modello previsionale e Risultati

- Distribuzione spesa totale vs. spesa pesata



Modello previsionale e Risultati

- Esempio di regole associative:

se REGIONE = TOSCANA
e NUM_SPESE <= 128
e TIPOLOGIA_01 = 7
e TIPOLOGIA_09 = 0
e TIPOLOGIA_ZERI > 2
e REGR_BATTUTE <= -0,98
allora V (confidenza 82,8%)

se DATA_ULTIMA_SPESA > 183
e NUM_SPESE <= 21
e TIPOLOGIA_ZERI > 1
e REGR_NUM_SPESE <= -0,02
e REGR_BATTUTE <= -0,98
allora V (confidenza 92%)

Modello previsionale
Risultati Globali

- Correttezza generale del modello:
  - 81.06% sul training set (70% del dataset, 360.000 righe)
  - 80.94% sul test set (30% del dataset, 155.000 righe)
- Matrici di confusione:

Valori Predetti

Valori Reali

Training Set	F	V
F	256.608	17.920
V	50.540	36.466

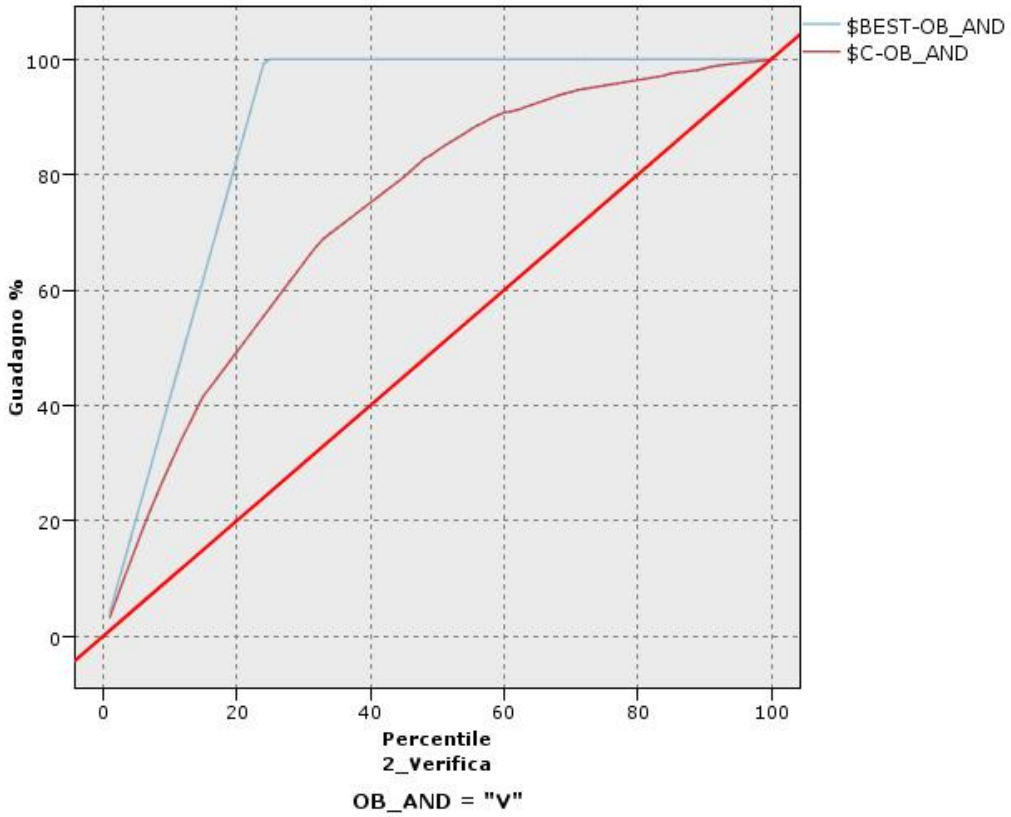
Test Set	F	V
F	110.029	7.767
V	21.855	15.734

66.9%

Con un guadagno netto del **42.8%**

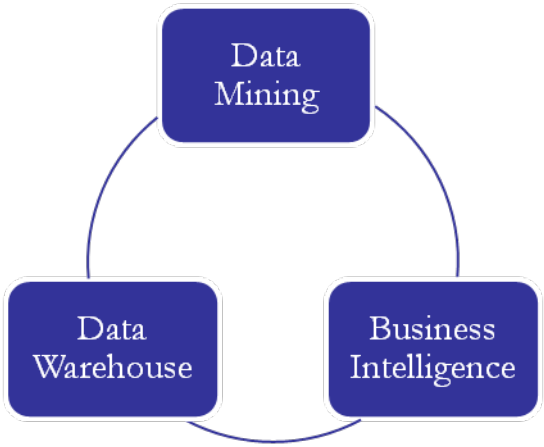
Modello previsionale
Risultati Globali

- Lift chart



Scenario d'uso – Esempio

- Creare un ambiente aperto e dinamico nel quale i dati forniti dal data warehouse vengono elaborati e trasformati in modelli di tipo previsionale.



- I modelli previsionali possono essere usati per arricchire il data warehouse, innestando un circolo virtuoso di informazioni utilizzabili anche direttamente in ambienti di Business Intelligence.

Conclusioni

- Per concludere:
  - Sono stati utilizzati dati provenienti dal data warehouse, risparmiando tempo e ottenendo dati di buona qualità
  - Abbiamo usato tecniche di mining avanzate per generare modelli predittivi, principalmente regole associative e alberi di decisione.
  - I risultati ottenuti sono soddisfacenti e si intravedono buone prospettive di miglioramento
- Possibili sviluppi futuri
  - Sperimentazione di altri tipi di analisi: sub group analysis, market segmentation, clustering ect.
  - Consolidamento e validazione dei risultati ottenuti
  - Incrementare la collaborazione con gli esperti del dominio per una migliore taratura del problema, delle definizioni usate e delle funzioni obiettivo
  - Integrazione dei dati previsionali forniti dai modelli predittivi all'interno della struttura di business intelligence aziendale

Analisi della defezione nella grande distribuzione tramite indicatori geografici e socio demografici/1

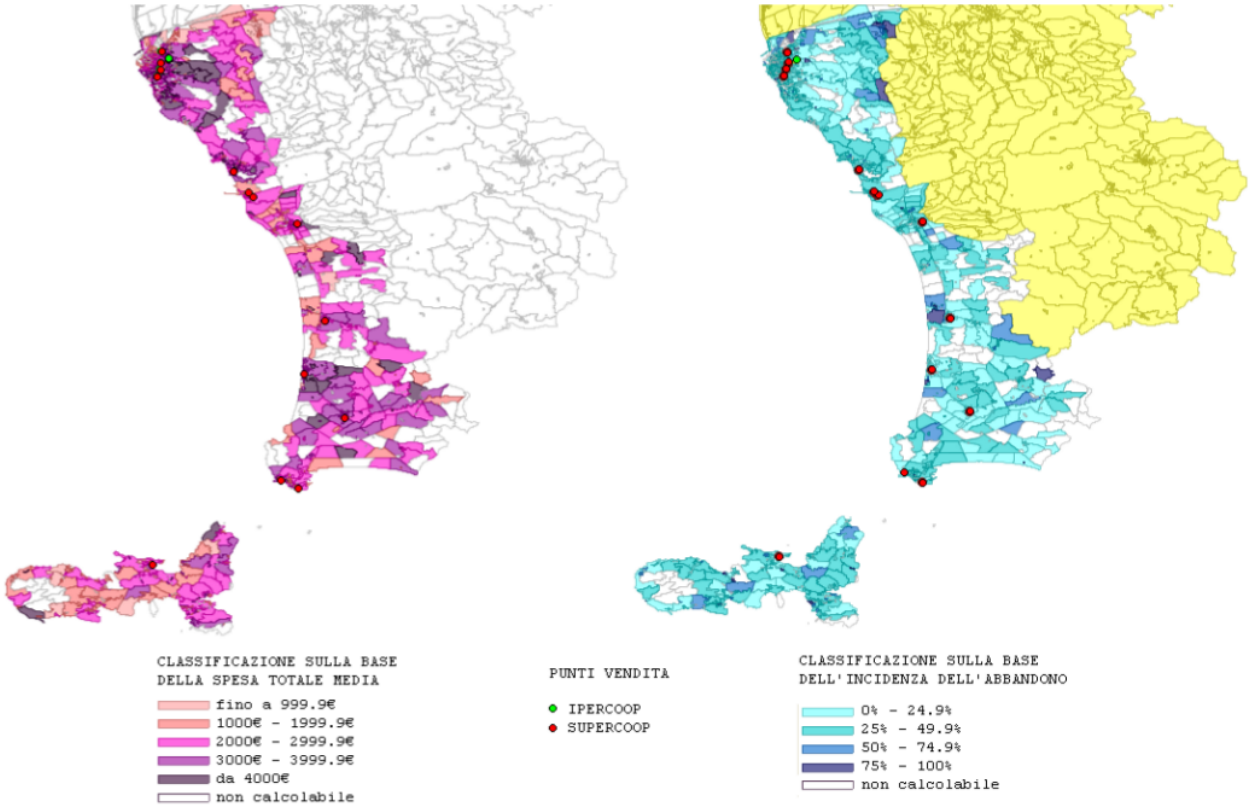


Figura 5.3: Provincia di Livorno con colorazione tematica in base all'incidenza del fenomeno dell'abbandono e alla media degli acquisti totali.

Analisi della defezione nella grande distribuzione tramite indicatori geografici e socio demografici/2

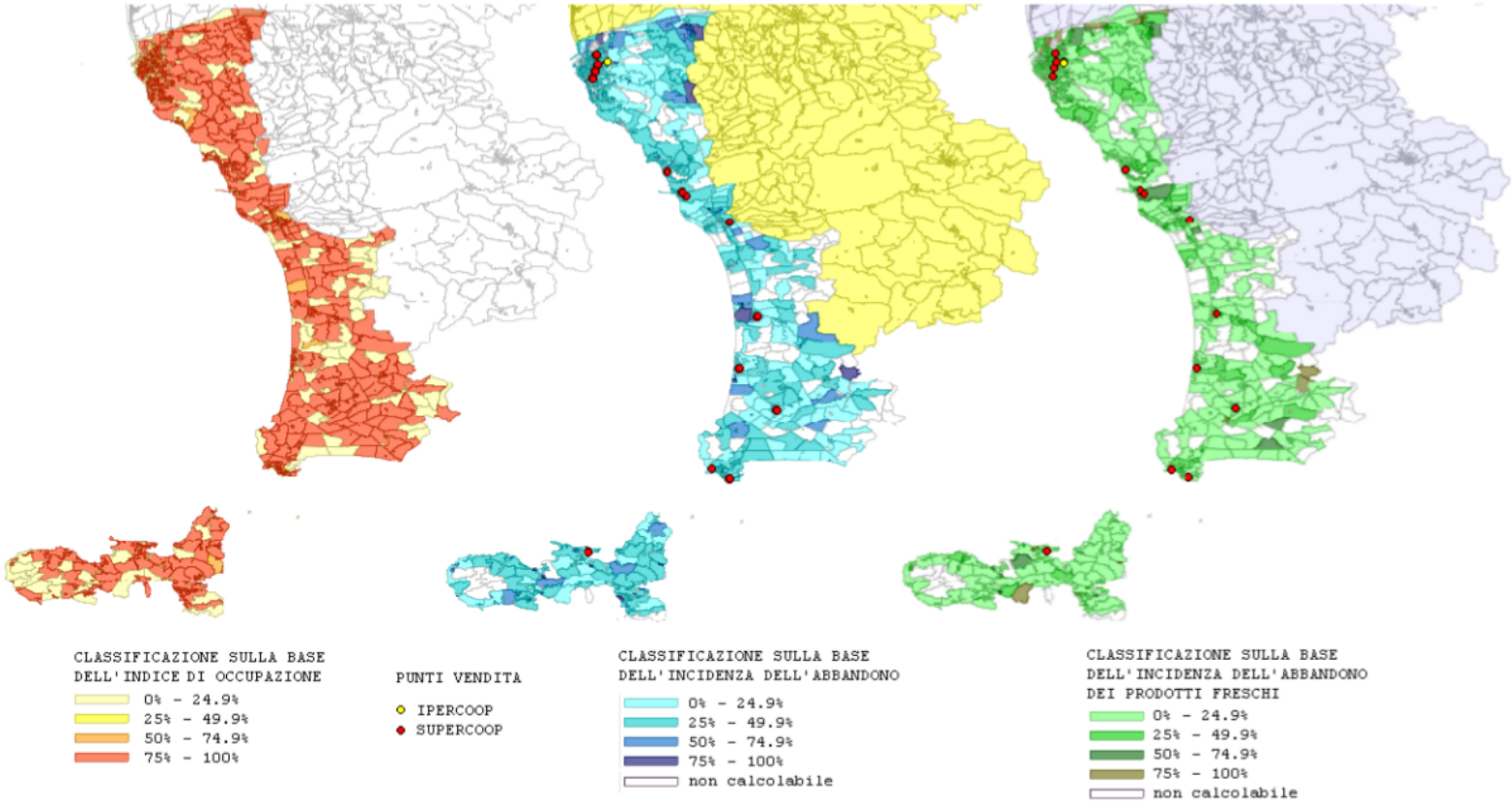


Figura 5.14: Provincia di Livorno con colorazione tematica in base al livello di occupazione e all'incidenza dell'abbandono.

Analisi Spazio-Temporale della defezione nella grande distribuzione / Eventi individuali e fenomeni collettivi nella grande distribuzione. Analisi spazio-temporale dei dati di vendita

- Varie tipologie di eventi individuali
  - Abbandono
  - Fidelizzazione ad un prodotto
  - Ecc.
- Analisi spaziale e temporale alla ricerca di grandi gruppi di eventi co-localizzati e contemporanei
- Metodi:
  - Clustering density-based
  - SatScan (metodo statistico usato in epidemiologia)

Analisi Spazio-Temporale della defezione nella grande distribuzione / Eventi individuali e fenomeni collettivi nella grande distribuzione. Analisi spazio-temporale dei dati di vendita

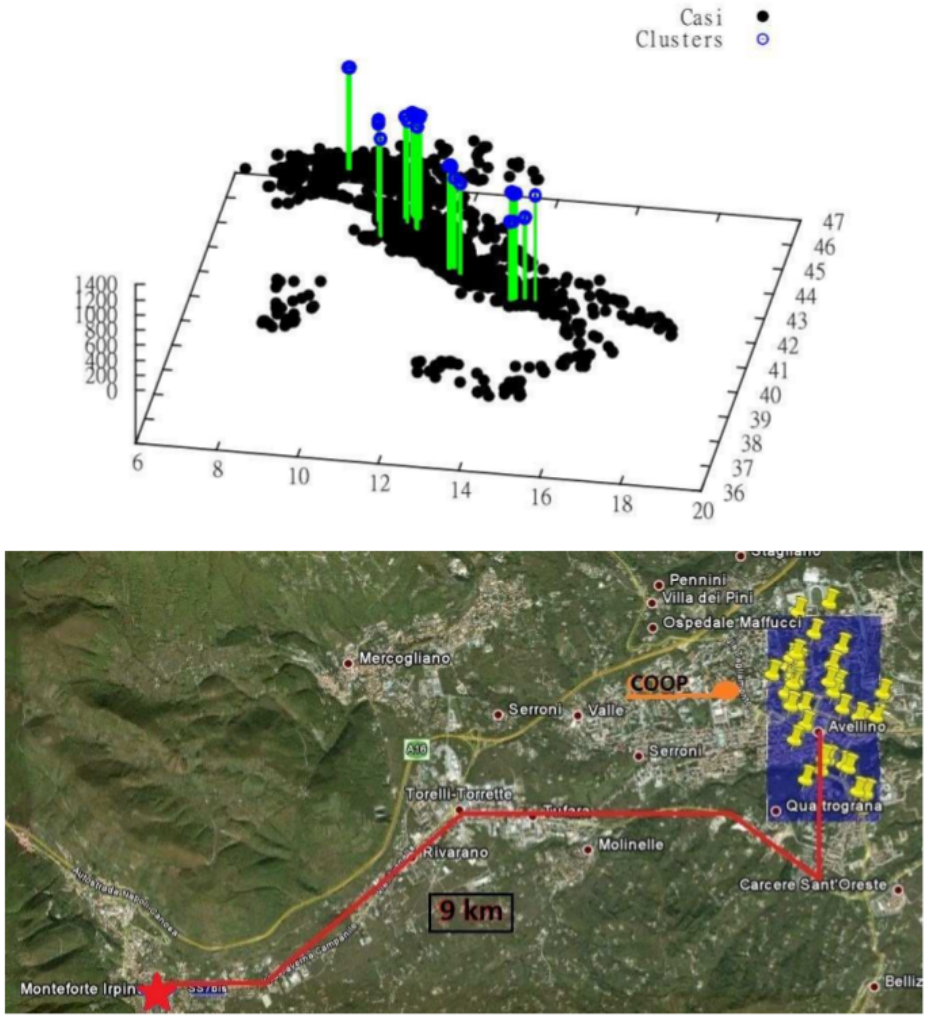


Fig.42 – Apertura di un nuovo centro commerciale a Monteforte Irpino.

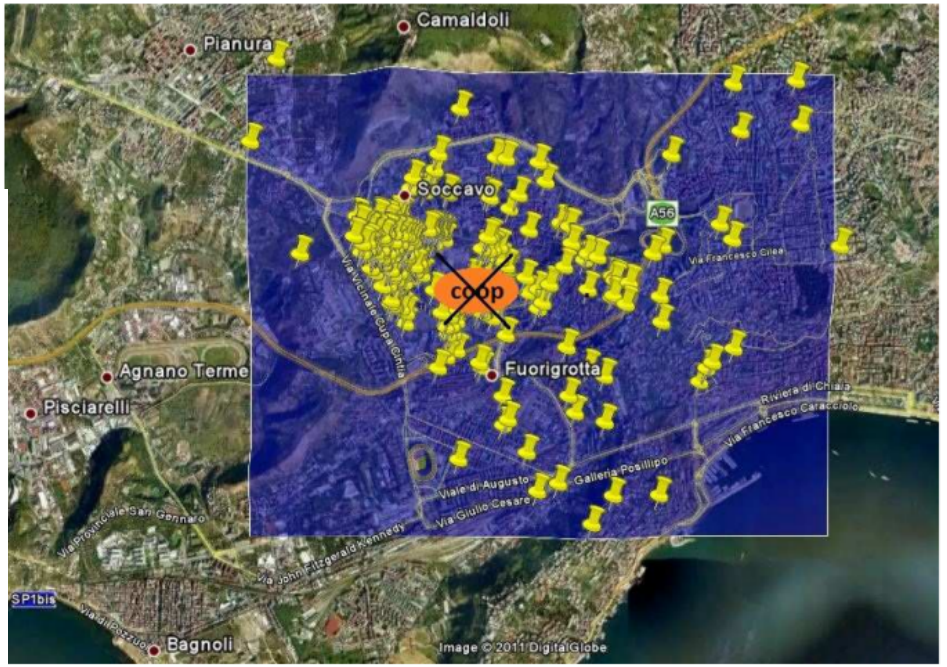
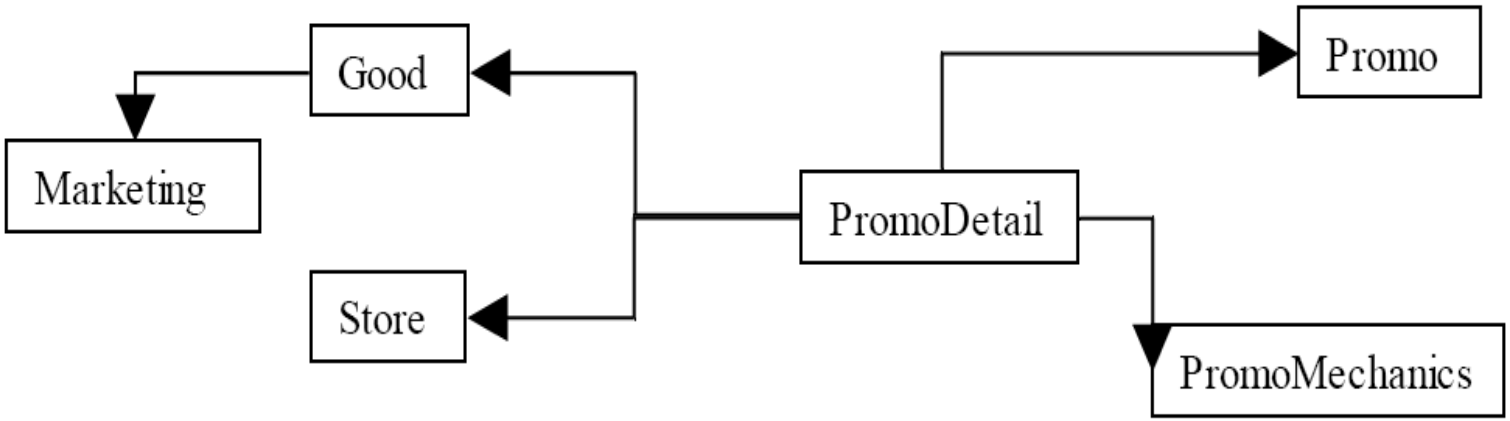


Fig.38 – Chiusura del supermercato Unicoop di Soccavo.

Dat



Data Mining on Promotional Sales

Study the impact of promo on sales

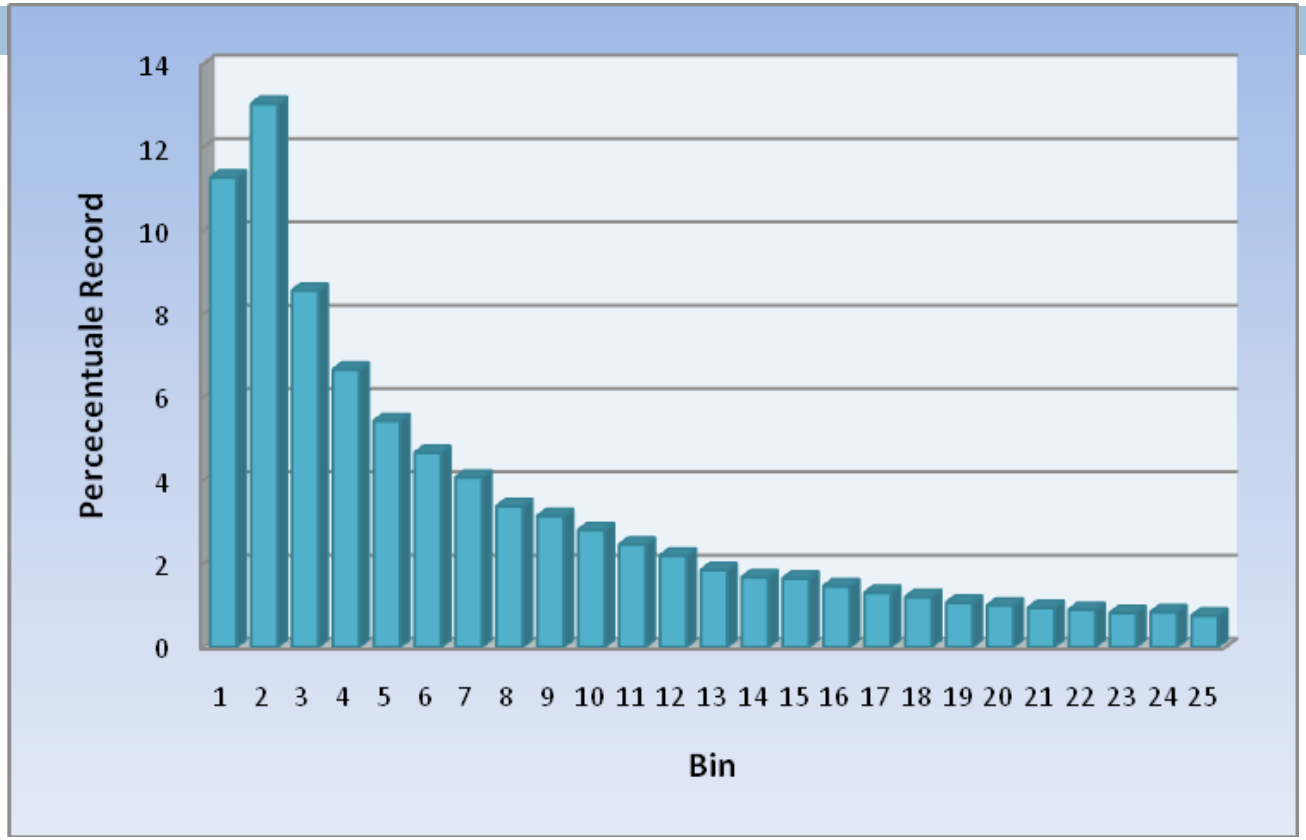
Goal: focusing on the effects of promotions on the sales of a single product, mainly aimed at optimizing its stoking:

- **Forecasting sales of promoted products .**
- **Forecasting "out-of-stock"**

Case study on product category = Food,

Two months April 2006 & April 2007.

Sales volume distribution of promotions with al least 5 items sold



- 11.% of the promotions in a single store sold between 5 and 24 items (the leftmost bar in the figure),
- the 13.79% sold between 25 and 49 items,
- The tale is less flat

Model building

Predictors:

- Product details
- Promo details
- Volume of sales in the periods befors the promotion

Target Variable:

- Number of sales for the promoted item
- **Variation w.r.t the month before the promotion**

Mining Table

Field nme	Description
Vend_Art_3_1	Sales of the article from 3 months to 1 month before the promotion
Vend_Seg_3_1	Sales of the segment from 3 months to 1 month before the promotion
Vend_Art_1_0	Sales of the article in the last month before the promotion
Vend_Seg_1_0	Sales of the segment in the last month before the promotion
Giorni_Promozione	

data sales of 16 months in 134 stores (522,541,764 records).

Target variables

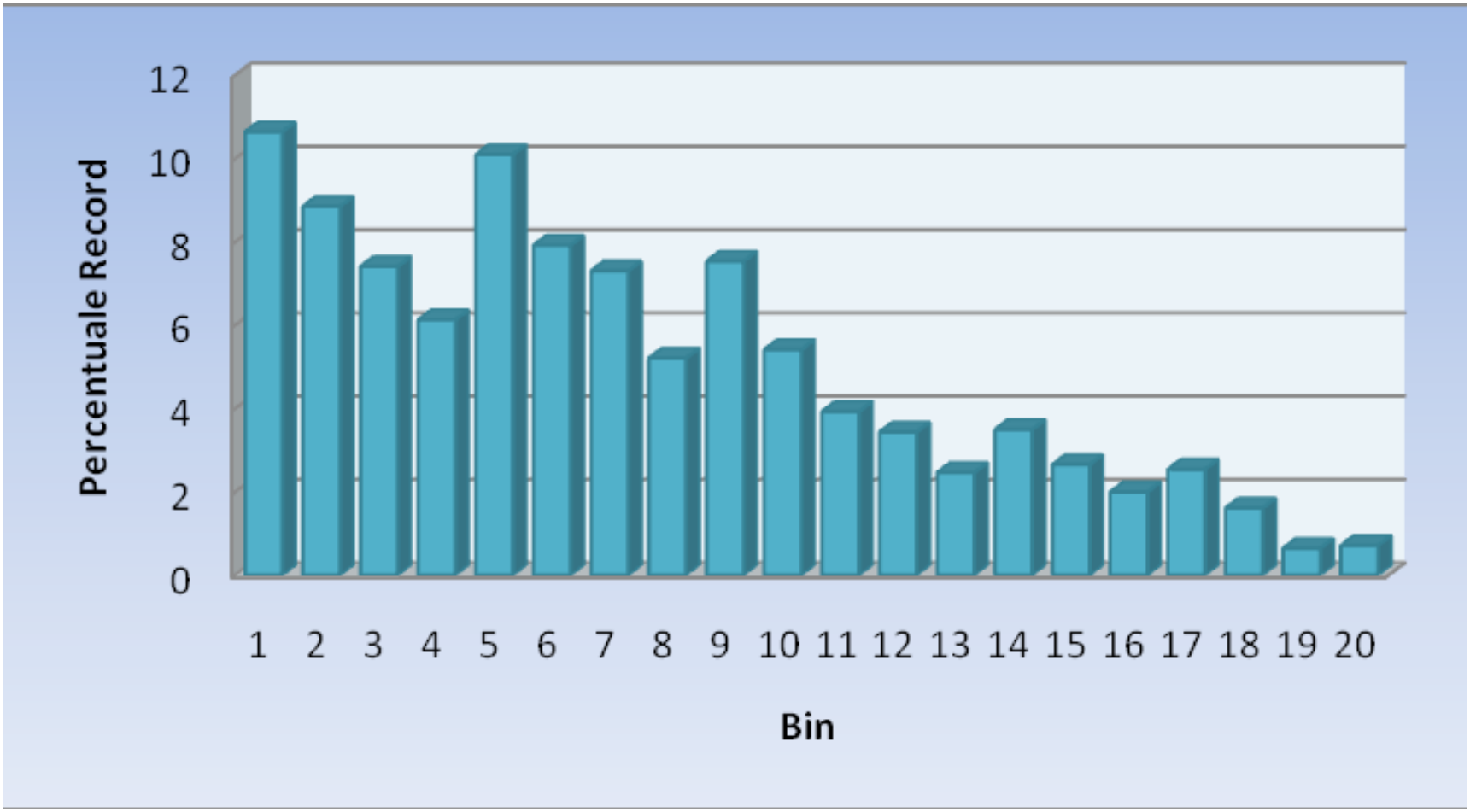
- Case1- The **sales amount** of the promoted item and
- Case2- The **sales variations** of the promoted item and
- Case3 - The number of out of stock that occurred during the promotion.

- MINING TASK: MULTICLASS over ORDINAL CLASSES

Case1: Predicting **volume sale**

- Volume is continuous, range 0 –105.650, 80% less than 500 item
  - Discretize (how many classes)
  - Multiclass predictor
- Equal size binning:
  - 10 binwith => 965 bin (classes) 18% in first 3 classes
  - 100 binwith => 249 bin (classes) 64% in first 3 classes
- Equal frequency binning
  - 20 bins => ..refined

Distribution of sales volume discretized in 20 bins – refined discretization



Results evaluation

- Accuracy 55,1% on the training set, which drops to 22,45% over test set

isultati per campo di output VEND_ART_PROMO_TILEN_String

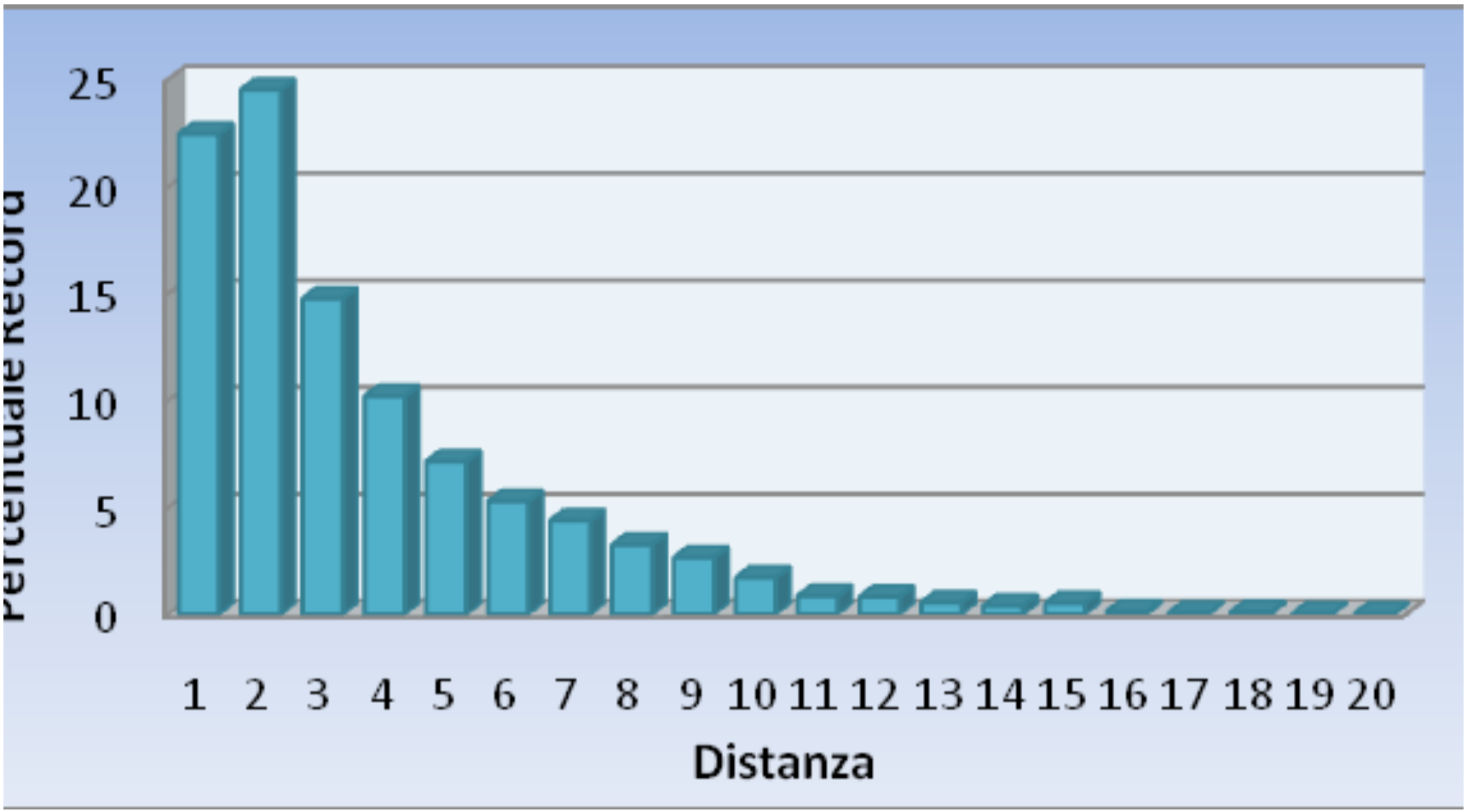
Confronto di $C-VEND_ART_PROMO_TILEN_String con VEND_ART_PROMO_TILEN_String

'Partizione'	1_Addestramento	
Corretto	3.386	55,1%
Sbagliato	2.759	44,9%
Totale	6.145	

Matrice coincidenza per $C-VEND_ART_PROMO_TILEN_String (le righe mostrano i valori effettivi)

'Partizione' = 1_Addestramento	1	2	3	4	5	6	7	8	9	_10	_11	_12	_13	_14	_15	_16	_17	_18	_1
1	545	44	25	6	3	6	5	1	5	4	2	0	0	2	0	0	0	0	
2	111	318	47	12	21	13	6	1	16	5	1	0	0	0	1	0	1	0	
3	47	51	270	16	23	9	5	0	9	9	2	0	0	0	2	1	0	0	
4	50	32	44	179	40	6	9	4	3	1	2	3	1	1	0	0	0	0	
5	45	22	43	26	401	32	17	11	16	7	8	1	1	4	2	0	1	0	
6	34	18	22	19	62	243	24	11	15	6	3	1	1	3	0	1	0	1	
7	25	10	15	12	38	43	255	11	20	6	3	5	1	2	2	0	3	0	
8	19	11	12	16	21	12	39	141	20	7	5	3	0	4	0	1	1	0	
9	27	9	12	9	29	34	32	38	247	15	6	2	3	3	2	0	1	0	
_10	14	11	8	8	17	20	18	14	40	146	3	15	4	6	3	1	1	1	
_11	9	3	5	2	11	11	14	12	25	25	94	12	2	9	2	0	5	1	
_12	5	5	6	6	10	9	11	7	16	20	10	91	5	7	1	2	2	1	
_13	6	1	3	2	6	6	5	7	6	7	9	12	47	8	7	4	2	1	
_14	7	2	4	5	5	2	7	8	9	13	12	9	6	119	3	2	3	2	
_15	3	3	4	5	8	2	8	8	7	10	8	6	1	18	78	4	6	0	
_16	3	0	2	1	6	3	4	4	1	5	3	5	7	13	11	43	11	2	
_17	3	0	3	4	6	3	4	2	5	4	6	7	4	11	9	1	80	2	
_18	1	0	2	1	4	3	0	3	2	0	2	2	3	7	11	7	12	36	
_19	1	0	1	1	1	0	1	2	0	1	1	0	1	0	3	2	5	0	1
_20	0	0	0	0	0	0	0	0	0	3	2	0	0	0	1	1	1	0	

Class displacement distribution of predicted class vs. real class



NOTE: it holds for ordinal classifiers

Rule	Support	Confidence	Confidence with error ≤ 1	Confidence with error ≤ 2
if CATEGORIA = ZUCCHERO E DOLCIFICANTI e FL_VOLANTINO = No e VEND_ART_1_0 > 37 **then** class = 2	47	23%	82%	93%
if CATEGORIA = 'ALIMENTI INFANZIA' e VEND_ART_1_0 > 275 **then** class = 3	138	50%	82%	97%
if CATEGORIA = CONSERVE DI FRUTTA e MESE = 8 **then** class = 5	113	24%	65%	86%
if CATEGORIA = YOGURT e DESCRIZ. = TAGLIO PREZZO e MESE = 9 e VEND_ART_1_0 > 54 e VEND_SEG_1_0 <= 4487 **then** class = 6	110	35%	65%	82%
if CATEGORIA = 'PASTA FRESCA' e MESE = 10 e VEND_ART_1_0 > 51 **then** class = 7	42	38%	57%	78%
if FL_COOP = Si e CATEGORIA = BISCOTTI e FL_VOLANTINO = Si e VEND_ART_1_0 <= 275 **then** class = 8	52	25%	61%	78%

Table 6 - Classification rules with support and confidence, including limited tolerance to errors

Rule 1: if more than 37 articles were sold in the last month before the promotion (vent_art_1_0
37) in the category “sugar” (categoria = zucchero e dolcificanti),

- and the promotion was not advertised in the advertising leaflets,
- the promoted item will sell the same or just a slightly higher amount than before the promotion (class = 2).

Case2: New Target Variable:

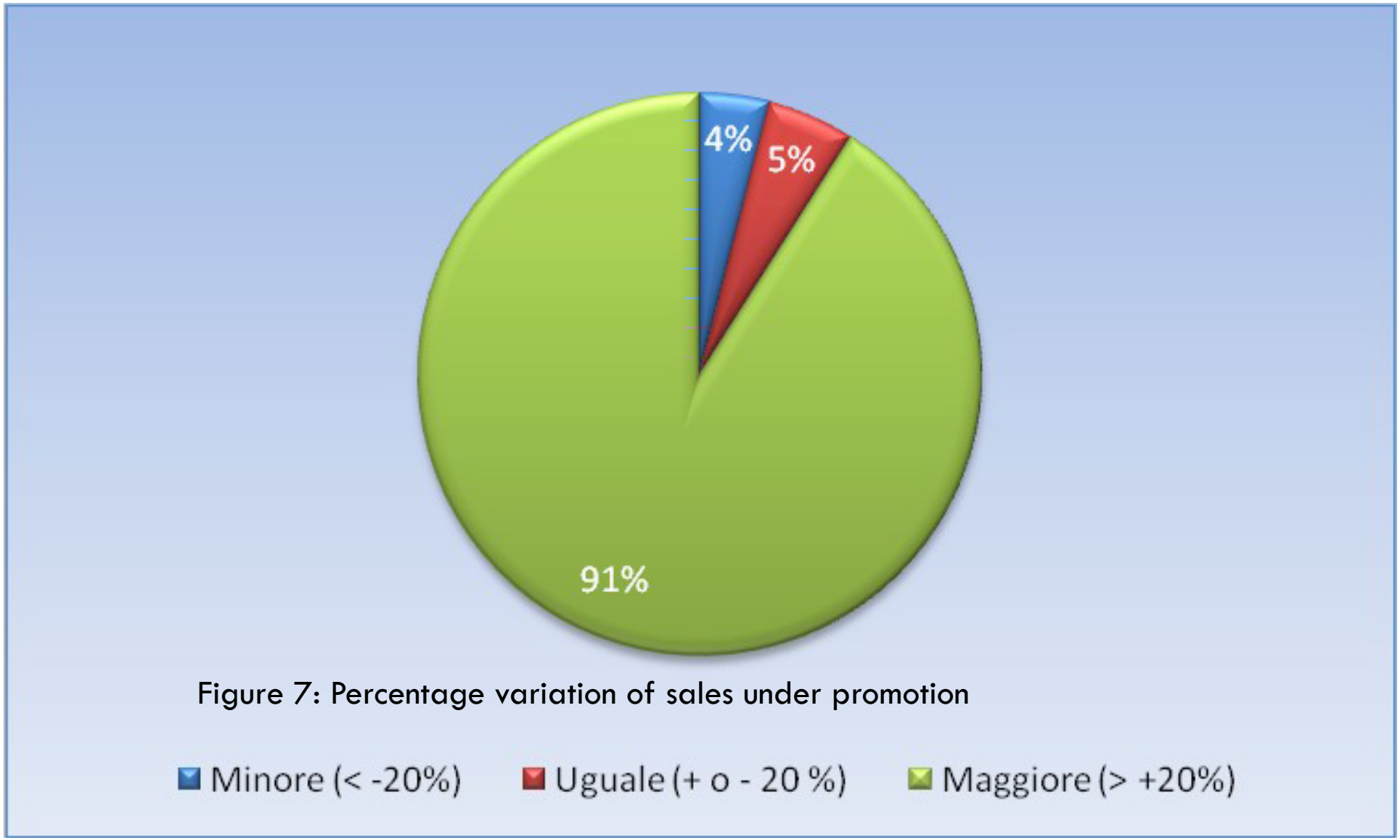
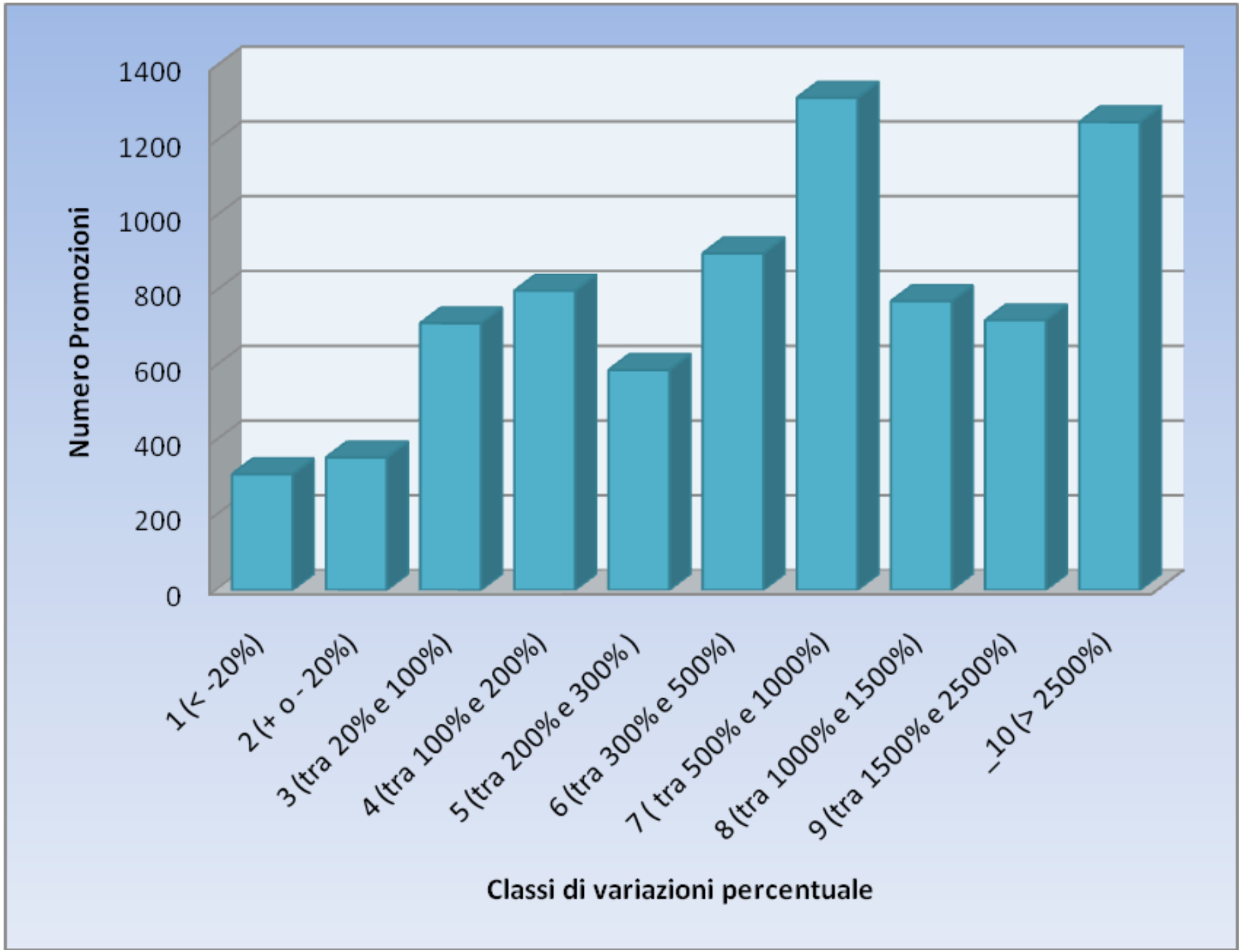


Figure 7: Percentage variation of sales under promotion

Case2: Variation of sales

Class	Meaning	
1	Drop of sales (sales variation ≤ -20/%)	
2	Drop of sales (sales variation ≤ -20/%)	
3	Small increase 1 (variation between + 20% and +100%)	
4	Small increase 2 (variation between +100% and 200%)	
5	Small increase 3 (variation between +200% and 300%)	
6	Large increase 1 (variation between +300% and 500%)	
7	Large increase 2 (variation between +500% and 1000%)	
8	Large increase 3 (variation between +1000% and 1500%)	
9	Extreme increase 1 (variation between +1500% and 2500%)	
10	Extreme increase 2 (variation ≥ 2500%)	


Numero Promozioni
1400
1200
1000
800
600
400
200
0
1 (< -20%)
2 (+ o - 20%)
3 (tra 20% e 100%)
4 (tra 100% e 200%)
5 (tra 200% e 300%)
6 (tra 300% e 500%)
7 (tra 500% e 1000%)
8 (tra 1000% e 1500%)
9 (tra 1500% e 2500%)
_10 (> 2500%)
Classi di variazioni percentuale

Results evaluation

- Accuracy reaches the 49.99% on the training set and 32.67% on the test set

Risultati per campo di output VariazionePercentualiS

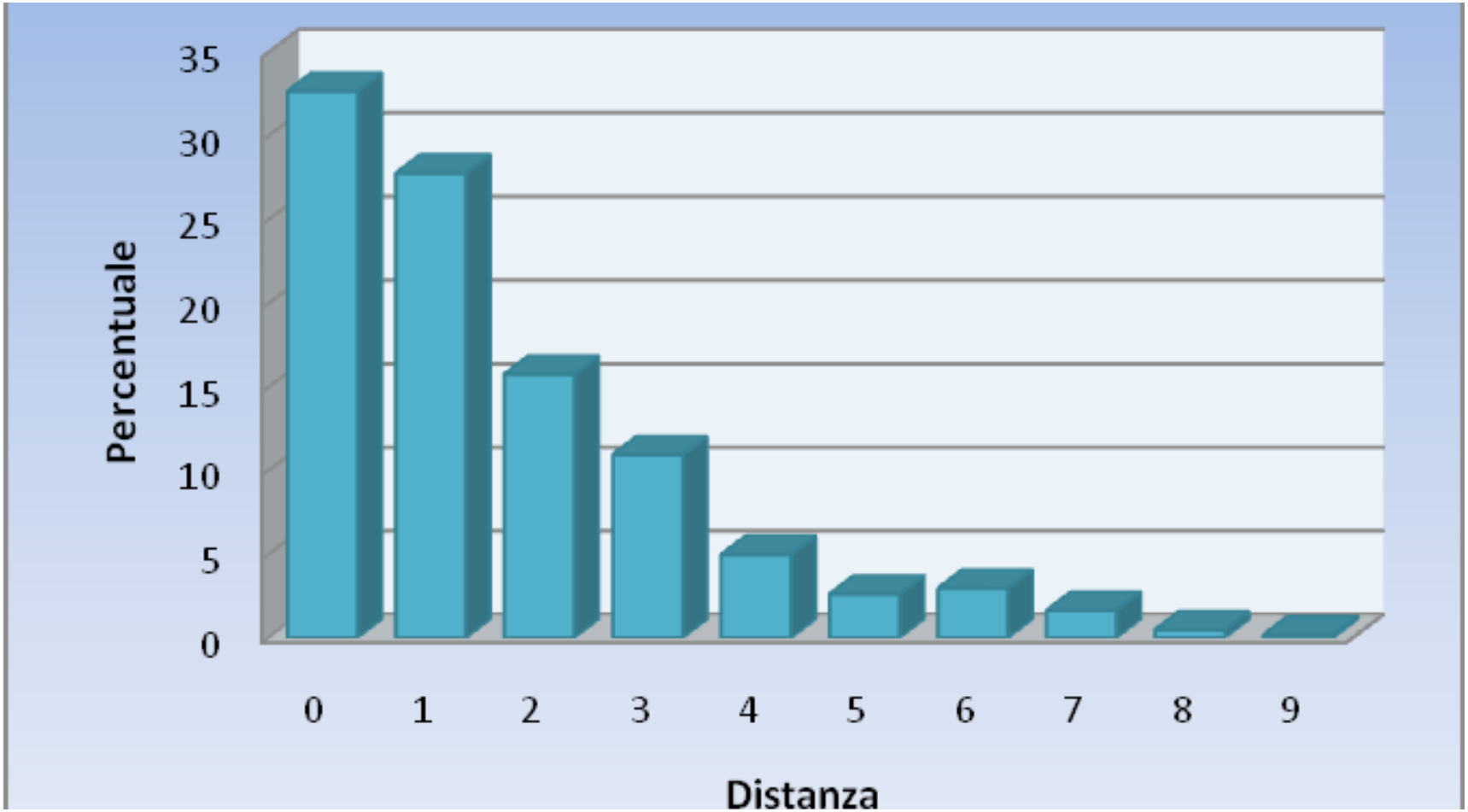
Confronto di $C-VariazionePercentualiS con VariazionePercentualiS

Corretto	2.702	49,99%
Sbagliato	2.703	50,01%
Totale	5.405	

Matrice coincidenza per $C-VariazionePercentualiS (le righe mostrano i valori effettivi)

	1	2	3	4	5	6	7	8	9	_10
1	75	16	53	14	7	9	25	0	7	10
2	11	83	69	25	3	9	28	4	7	16
3	6	17	281	63	11	23	57	1	10	35
4	2	9	88	265	21	32	77	5	9	55
5	2	3	61	57	100	51	80	6	8	34
6	2	5	64	39	16	250	147	8	10	59
7	6	5	64	41	16	53	572	29	28	92
8	2	8	32	10	3	13	164	180	27	103
9	1	1	19	14	0	11	132	26	159	162
_10	0	2	24	5	1	4	71	9	39	737

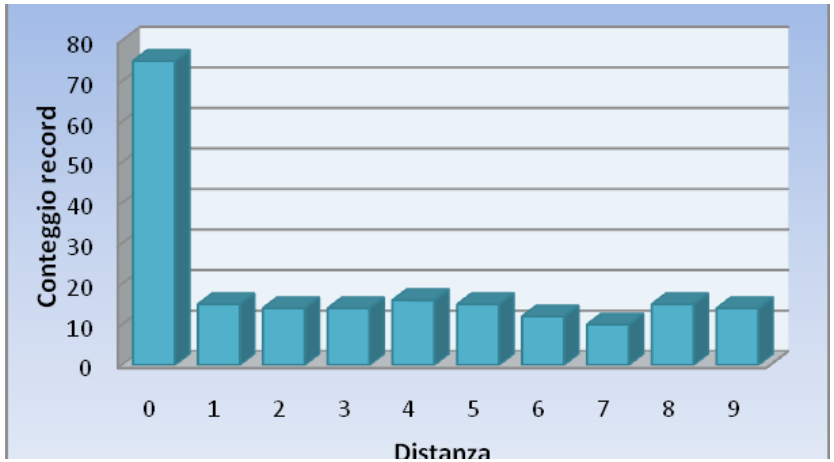
Class displacement distribution of predicted class vs. real class

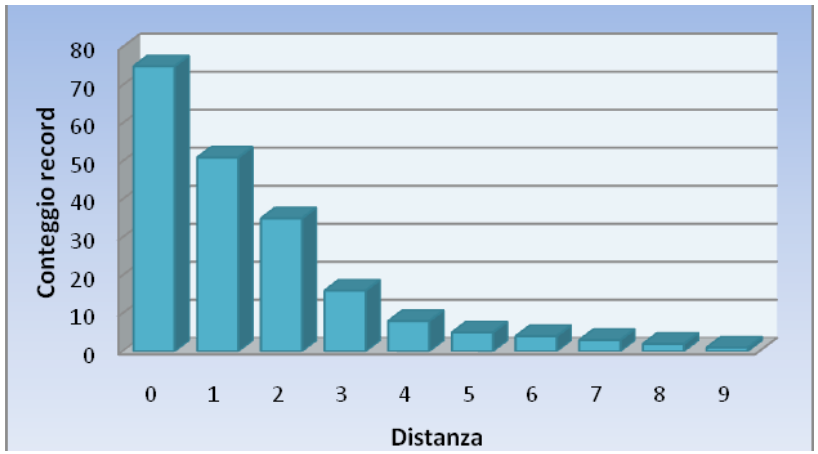


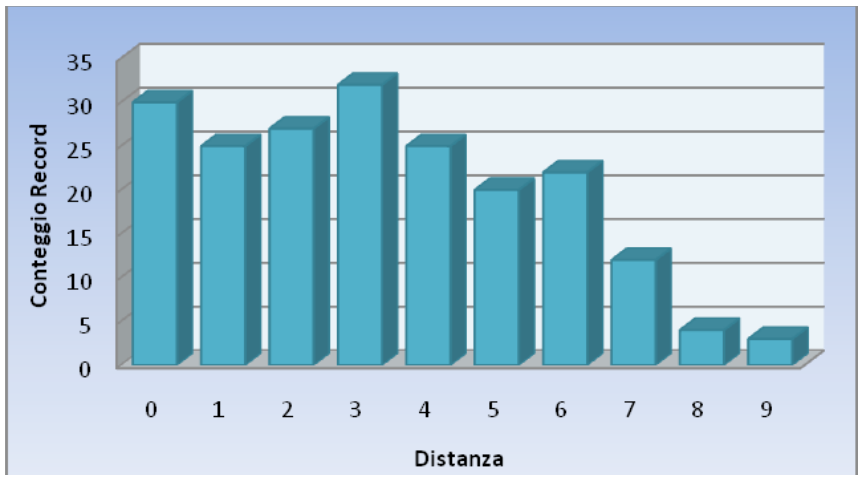
NOTE: it holds for ordinal classifiers

Evaluating ordinal (multiclass) classifier

Evaluating Ordinal Classifiers with distance matrix







Weights Vector-based approach

Distance	Weights 1	Weights 2	Weights 3
0	1	1	1
1	0	0,7	0,7
2	0	0,5	0,5
3	0	0,2	0,2
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	-0,2
8	0	0	-0,3
9	0	0	-0,5

Table 7 - Three sample weight vectors over 10 classes

$$Accuracy^{vector} = \frac{\sum_{i=1}^{N} \left(freq[i] \cdot weights[i] \right)}{\sum_{i=1}^{N} freq[i]}$$

$$Accuracy^{matrix} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (mat\_confusion[i,j] \cdot mat\_weights[i,j])}{\sum_{i=1}^{N} \sum_{j=1}^{N} mat\_confusion[i,j]}$$

		Predicted Class									
		1	2	3	4	5	6	7	8	9	10
	1	1,00	0,85	0,70	0,50	0,40	0,00	0,00	-0,50	-0,75	-1,00
	2	0,85	1,00	0,85	0,70	0,50	0,30	0,00	0,00	-0,50	-0,75
	3	0,70	0,85	1,00	0,80	0,65	0,40	0,20	0,00	0,00	-0,50
	4	0,50	0,70	0,80	1,00	0,80	0,65	0,30	0,10	0,00	0,00
	5	0,40	0,50	0,65	0,80	1,00	0,80	0,65	0,20	0,00	0,00
	6	0,00	0,30	0,40	0,65	0,80	1,00	0,75	0,60	0,20	0,00
	7	0,00	0,00	0,20	0,30	0,65	0,75	1,00	0,75	0,60	0,15
	8	-0,50	0,00	0,00	0,10	0,20	0,60	0,75	1,00	0,70	0,55
	9	-0,75	-0,50	0,00	0,00	0,00	0,20	0,60	0,70	1,00	0,70
	10	-1,00	-0,75	-0,50	0,00	0,00	0,00	0,15	0,55	0,70	1,00

Table 14: Matrix of weights 1

Traditional accuracy	Matrix-based accuracy Weights 1	Matrix-based accurac Weights 2
37.50 %	70.38%	66.50%

Table 16: Accuracies for the sales prediction model, using weights in Tables 15 and 1

Coming back

Our target vaiable

Class	Meaning	
1	Drop of sales (sales variation ≤ -20/%)	
2	Drop of sales (sales variation ≤ -20/%)	
3	Small increase 1 (variation between + 20% and +100%)	
4	Small increase 2 (variation between +100% and 200%)	
5	Small increase 3 (variation between +200% and 300%)	
6	Large increase 1 (variation between +300% and 500%)	
7	Large increase 2 (variation between +500% and 1000%)	
8	Large increase 3 (variation between +1000% and 1500%)	
9	Extreme increase 1 (variation between +1500% and 2500%)	
10	Extreme increase 2 (variation ≥ 2500%)	

Analisi dei risultati previsionali

Accuratezza del modello

Distribuzione della distanza tra classe predetta e classe effettiva

Distanza	Proporzione	%	Conteggio
0.000		32,67	761
1.000		27,78	647
2.000		15,76	367
3.000		10,95	255
4.000		4,98	116
5.000		2,62	61
6.000		2,96	69
7.000		1,63	38
8.000		0,47	11
9.000		0,17	4

Nel 75% dei casi la predizione è corretta a meno di una distanza di 2 classi

Ipotesi di utilizzo

Ipotesi:

- Si mette in promozione di un articolo il cui venduto nei 15 gg precedenti è pari a 30 pezzi;
- Supponiamo che il classificatore preveda classe 5 (venderà tra il 200% e il 300% in più).

Risultati:

- Al 32% (circa 1/3) di possibilità il prodotto venderà tra 90 e 120 pezzi (predizione corretta),
- Al 60% il prodotto venderà tra 60 e 180 pezzi (scarto di una classe),
- Al 75% il prodotto venderà tra 36 e 330 pezzi (scarto di due classi).

Rottura di Stock

Due differenti tipologie di rottura di stock:

- A livello di **magazzino**; nella quale il magazzino si trova in difetto di merci durante il periodo di promozione ed è quindi impossibilitato a rifornire i negozi.
- A livello di **negozio**; in cui il negozio rimane sprovvisto di merci nell'arco di una singola giornata di promozione, a causa di probabili rifornimenti insufficienti.

CASE2: OUT OF STOCK

Two cases

- Out of stock in the warehouse
  - We would need stocking data

- Out of shop the on a specific day of promotion

Capturing out-of stock in a day

Definizione del modello

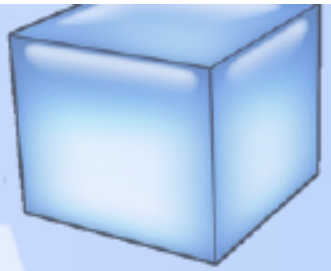
- Divisione di una giornata di vendita in quattro fasce orarie : Mattina, Pranzo, Pomeriggio, Sera (come definito nel DW)
- Rilevazione brusche cadute nei volumi di vendita

)90%- (

Mattina	Pranzo	Pomeriggio	Sera
40	30	2	1

- Come si vede nell'esempio si verifica rottura di stock tra le fasce Pranzo-Pomeriggio, si ha infatti un brusco calo delle vendite.

Definizione del modello



- Non è considerata rottura di stock (Le vendite riprendono).

Mattina	Pranzo	Pomeriggio	Sera
40	2	10	10

- Non è considerata rottura di stock (Forte variazione percentuale ma bassi volumi di vendita).

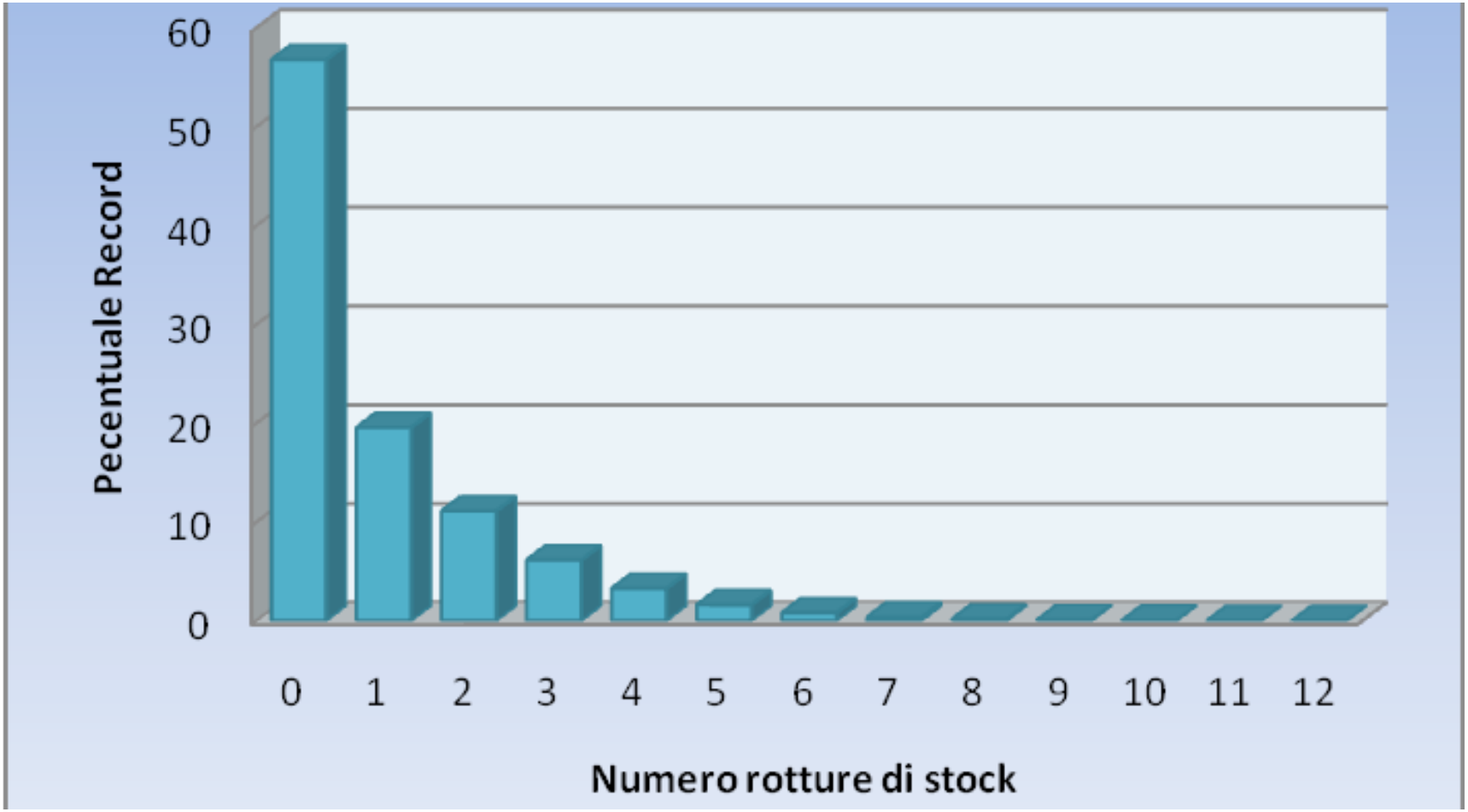
Mattina	Pranzo	Pomeriggio	Sera
2	1	1	0

- Si considerano casi di decrescita graduale.

Mattina	Pranzo	Pomeriggio	Sera
25	4	0	0

- Cond1: decrease of adjacent sales 90%
- Cond2: no further (no adjacent) increase
- Cond3: minimum number BEFORE the out-stock
- Cond4: if no number before reduce threshold 75%

Distribution of out of stocks in the Super stores

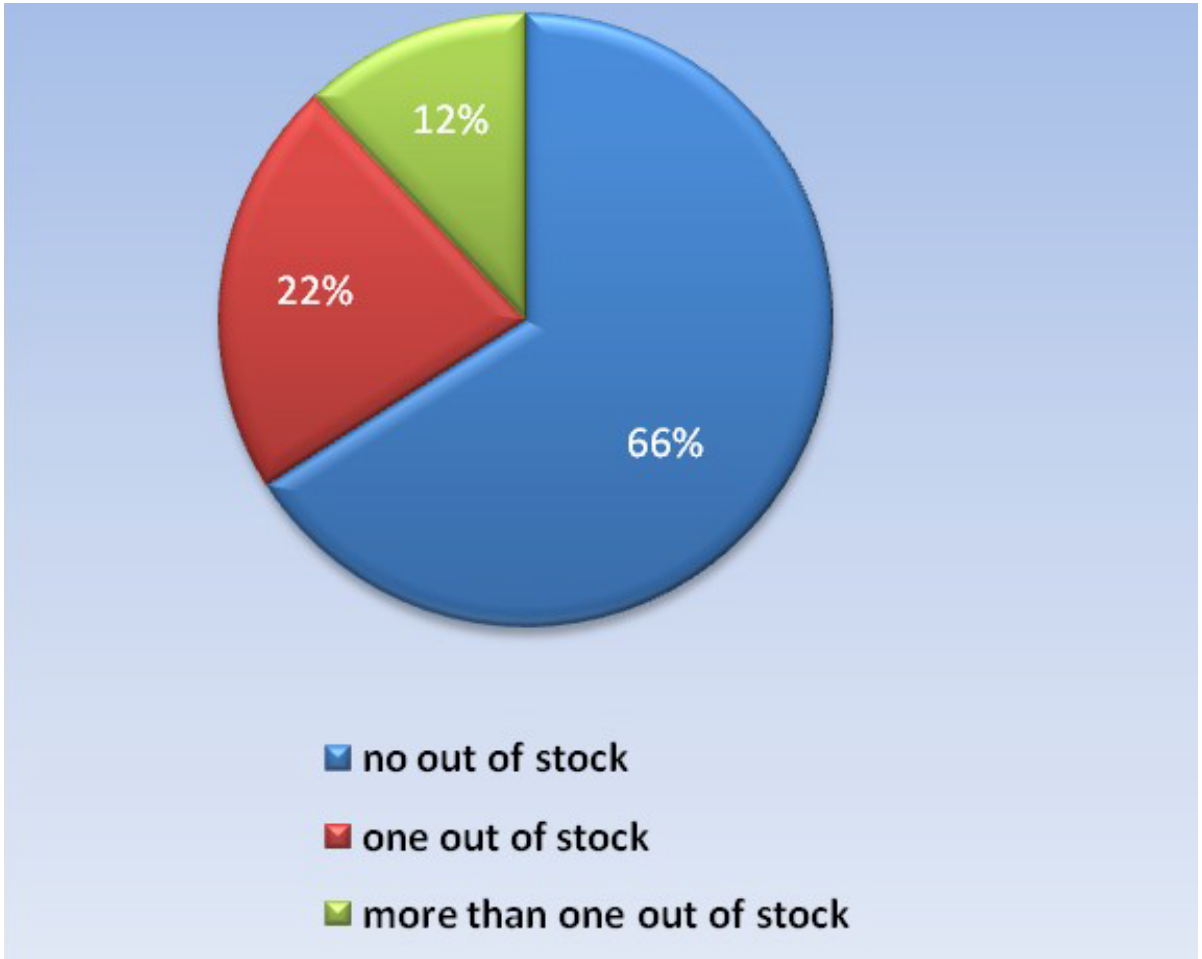


Analisi dati Super

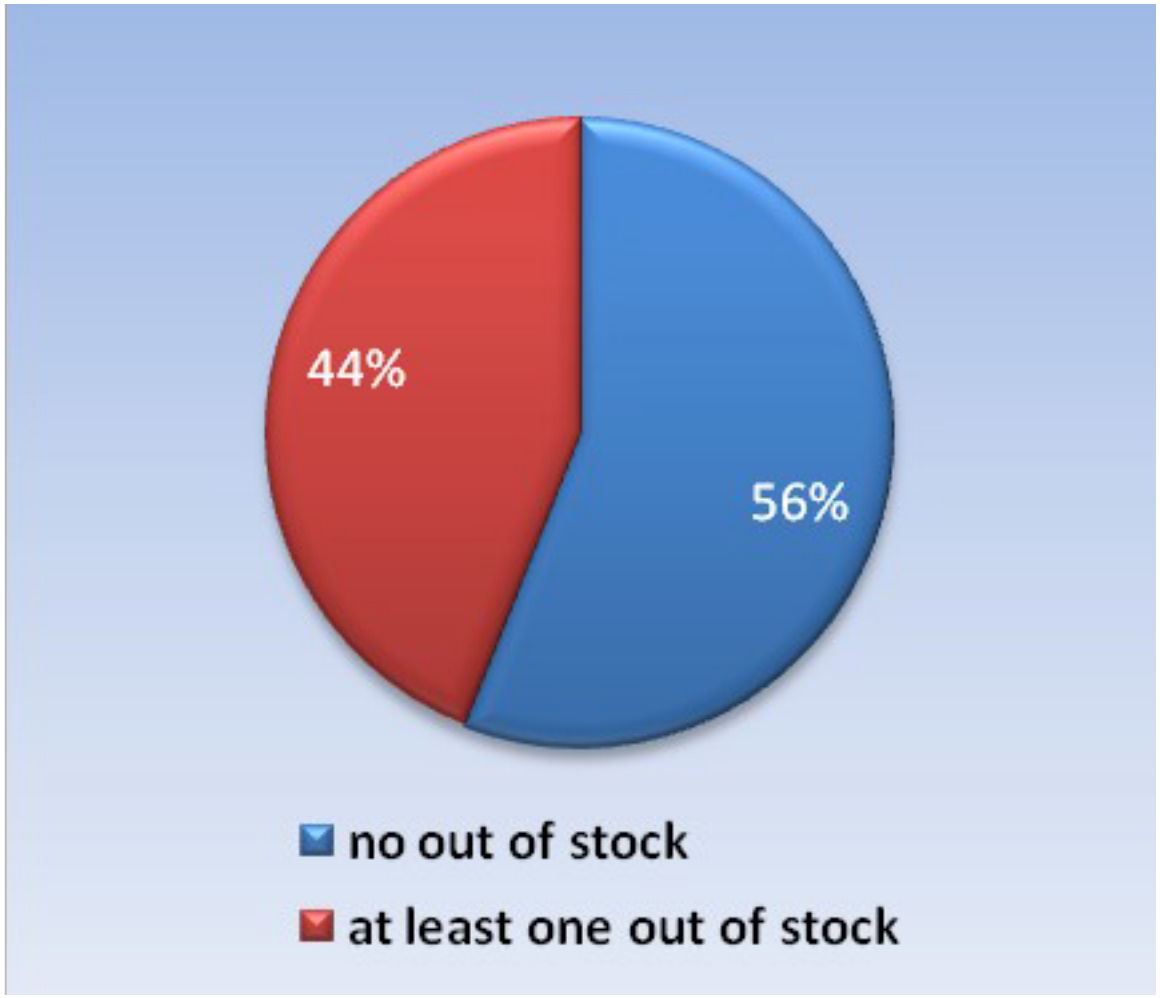
Distribuzione del numero di giorni in cui il prodotto in promozione va in rottura di stock

Valore	Proporzione	%	Conteggio
0		56,78	147736
1		19,56	50891
2		11,12	28924
3		6,16	16019
4		3,27	8520
5		1,62	4211
6		0,8	2076
7		0,35	915
8		0,17	441
9		0,08	211
10		0,04	105
11		0,02	59
12		0,01	27
13		0,01	29
14		0,01	17
15		0,0	3

Target variable: out-stocks (3 value)



Target variable out-stocks (2 value)



Es. Rule

if FL_VOLANTINO = Si

and CATEGORIA = ALIMENTI INFANZIA

and VEND_ART_3_1 > 142

and VEND_ART_1_0 > 96

then class = 1

support = 677 confidence= 65%

Rule 2

if MESE = 12
and CATEGORIA =
GELATI then class = 0
supp= 379 conf= 95%

Accuracy

Risultati per campo di output Rottura_Stok_Binario

Confronto di $C-Rottura_Stok_Binario con Rottura_Stok_Binario

Corretto	177.209	71,61%
Sbagliato	70.238	28,39%
Totale	247.447	

Matrice coincidenza per $C-Rottura_Stok_Binario (le righe mostrano i valori effettivi)

	0	1
0	109.412	31.424
1	38.814	67.797

Se considero sbagliati solo i “falsi negativi” l’accuratezza del modello passa all’ 84,31%

Esempio Regola: Caffè

Valore	Proporzione	%	Conteggio
0		56,78	147736
1		43,22	112461

Distribuzione binaria rottura di stock

se PRES_MKT = LEADER
e VendSeg_1_0 > 479
e CATEGORIA = CAFFE'
allora 1

Valore	Proporzione	%	Conteggio
0		41,3	1156
1		58,7	1643

Distribuzione binaria rottura di stock ristretto ai prodotti della categoria caffè

Supporto: 1643 Confidenza 58,7%

Esempio Regola: Caffè

se PRES_MKT = LEADER
e VendSeg_1_0 > 479
e CATEGORIA = CAFFE'
allora 1

Valore	Proporzione	%	Conteggio
0		32,07	339
1		67,93	718

Supporto: 718 Confidenza 67,93%

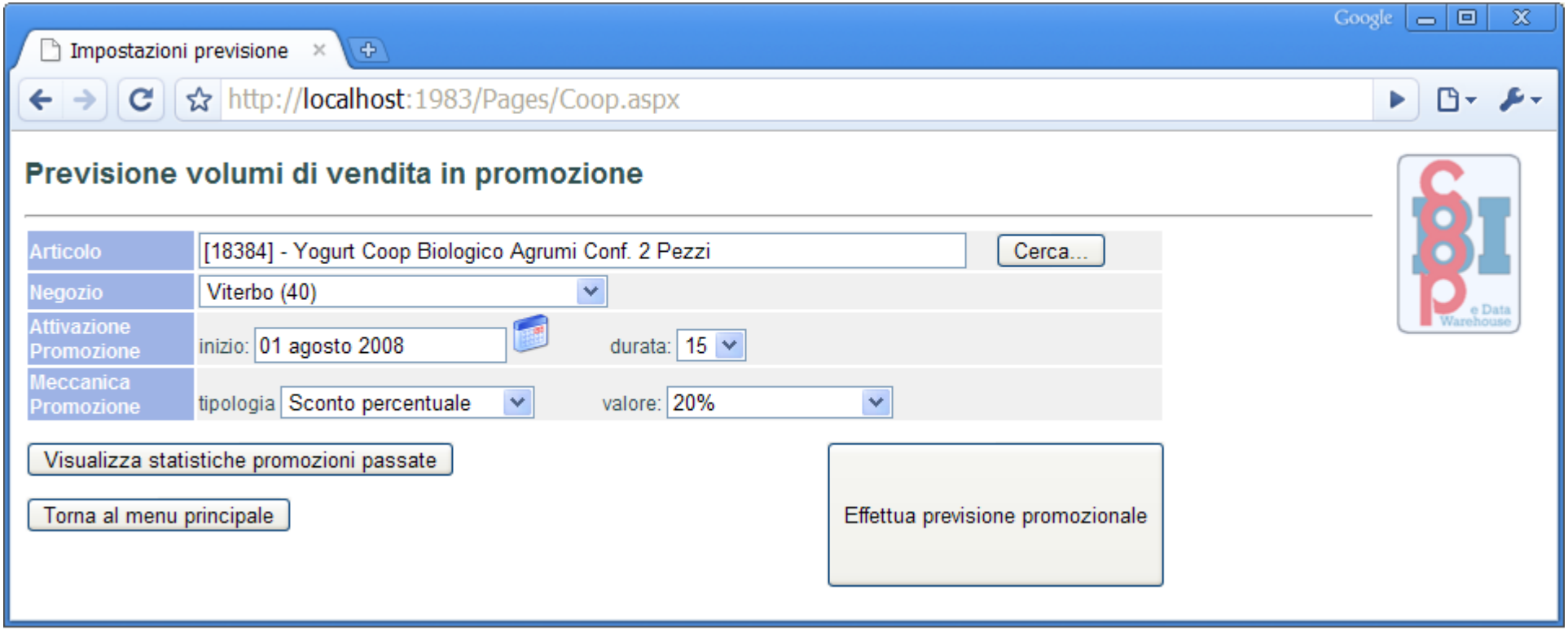
se PRES_MKT = LEADER
e VendSeg_1_0 > 479
e CATEGORIA = CAFFE'
allora 1

Valore	Proporzione	%	Conteggio
0		13,93	78
1		86,07	482

Supporto: 560 Confidenza 86%

Deployment

- I modelli predittivi consentono di “arricchire” i dati storici con dati previsionali
- Interfaccia uniforme verso l'utente finale



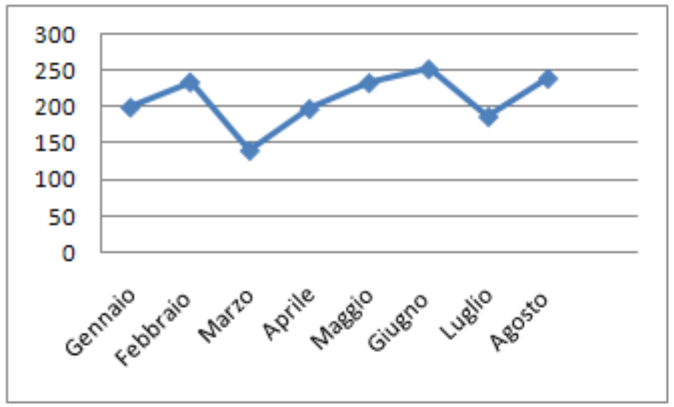
Risultati previsione

http://localhost:1983/Pages/Coop.aspx

Previsione volumi di vendita in promozione

Dettagli previsione	Articolo: [18384] - Yogurt Coop Biologico Agrumi Conf. 2 Pezzi Negozio: Viterbo (40) Data inizio: 01 Settembre 2008 - Durata: 15 giorni Meccanica: Sconto 10%

Andamento vendite



Gennaio	200
Febbraio	235
Marzo	140
Aprile	198
Maggio	234
Giugno	253
Luglio	187
Agosto	240

Provisioni

Vendita	Variazione percentuale	Rischio rottura di stock
da 250 a 300 pezzi	+100% a +200% (da 240 a 360 pezzi)	SI

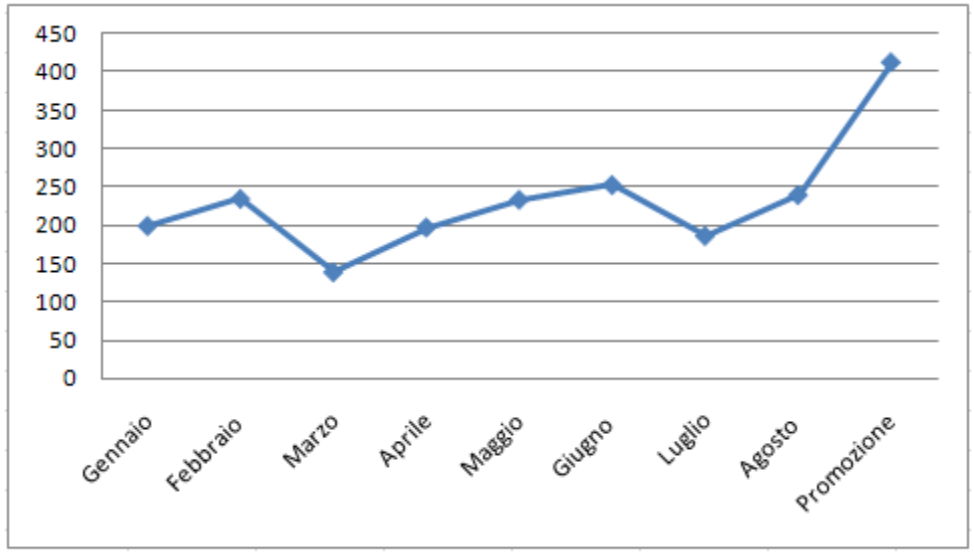
Indietro

Statistiche Promozioni

http://localhost:1983/Pages/Coop.aspx

Statistiche promozioni

Negozio	Viterbo (40)
Periodo ricerca	da: 01 agosto 2008 a: 30 settembre 2008
Articolo in promozione	[18384] - Yogurt Coop Biologico Agrumi Conf. 2 Pezzi
Dettagli promozione	Codice promodettaglio - 18384 Codice promoId - 235 Data inizio promozione - 01 settembre 2008 Data inizio promozione - 15 settembre 2008 Negozio - Viterbo Codice Negozio - 40



Confronta con segmento

Rotture di stock in promozione
2

Gennaio	200
Febbraio	235
Marzo	140
Aprile	198
Maggio	234
Giugno	253
Luglio	187
Agosto	240
Promozione	412

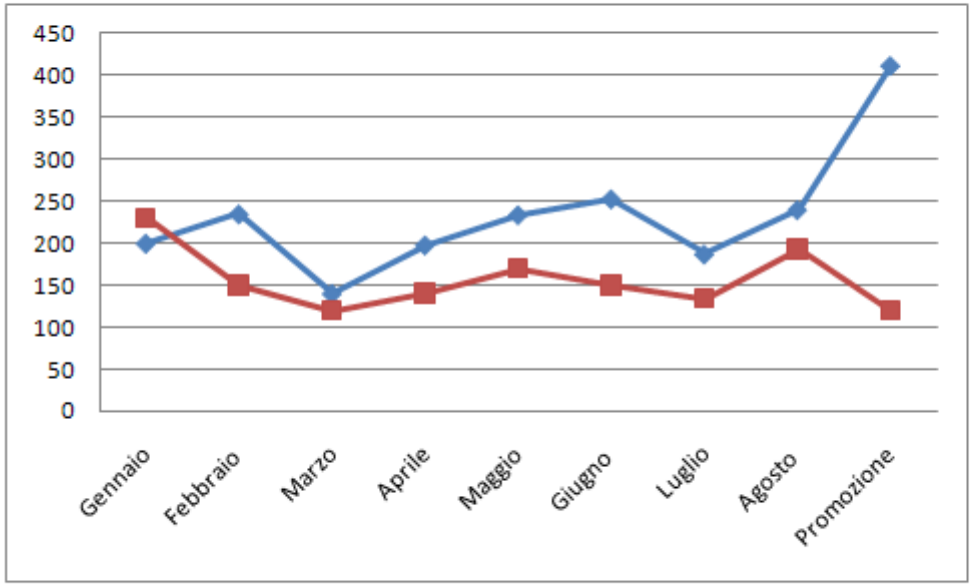
Torna al menu principale

Confronto prodotti

http://localhost:1983/Pages/Coop.aspx

Confronto con segmento

Negozio	Viterbo (40)
Periodo ricerca	da: 01 agosto 2008 a: 30 settembre 2008
Dettagli promozione selezionata	Codice promodettaglio - 18384 Codice promoId - 235 Data inizio promozione - 01 settembre 2008 Data inizio promozione - 15 settembre 2008 Negozio - Viterbo Codice Negozio - 40
Confronta con	[19472] - Yogurt Yomo Agrumi Conf. 2 Pezzi



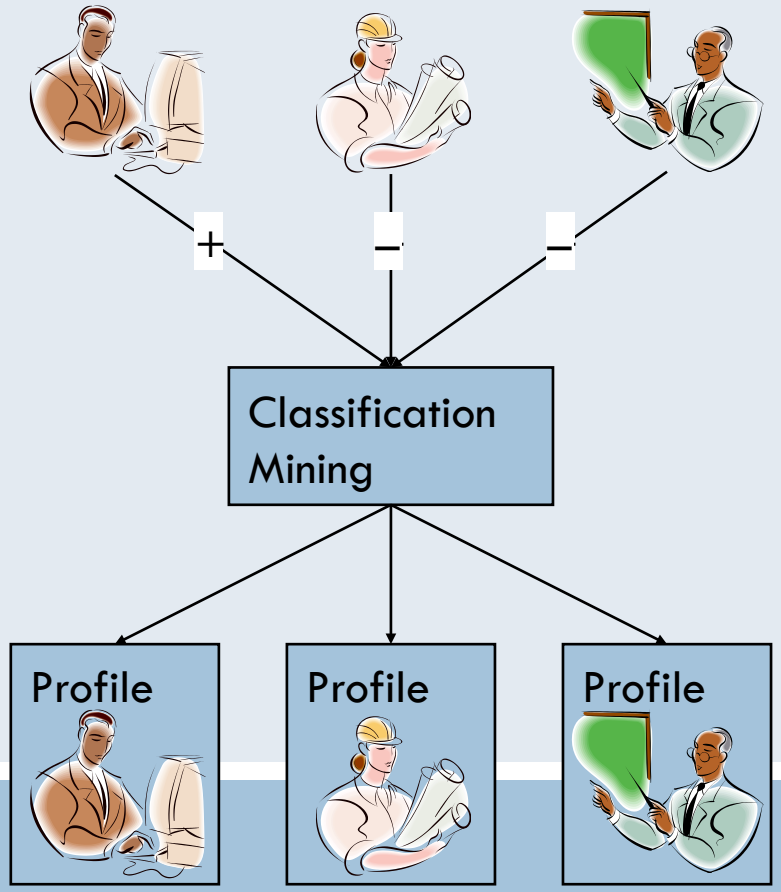
Gennaio	200	230
Febbraio	235	150
Marzo	140	120
Aprile	198	140
Maggio	234	170
Giugno	253	150
Luglio	187	124
Agosto	240	193
Promozione	412	120

Indietro

Conclusioni

- Buoni risultati da affiancare con report statistici e personale con esperienza nel settore
- Possibilità di raffinamento del modello venendo incontro ad esigenze più specifiche (singoli negozi, singole categorie)
- Miglioramento della qualità dei dati nel datawarehouse
  - Es. ruolo ed esposizione sono valorizzati con "non disponibile" nel 74% dei casi, ma non sono gli unici…
  - *potrebbe* portare ad un significativo aumento della qualità dei risultati

Profiling



Profiling

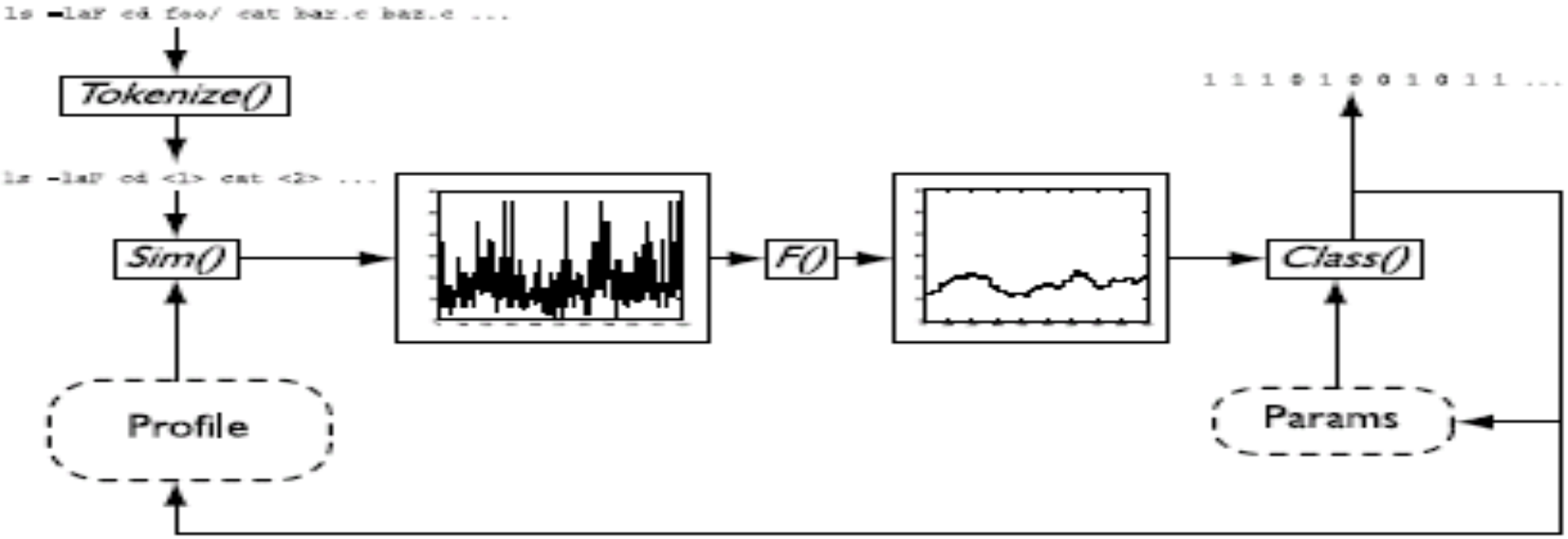
- Determine individual behavior
  - What is normal for the individual
  - What separates one individual from another
- Gives profile of individual behavior
- How do we do this?

profiles

- Cluster users
- Develop profiles for clusters
  - E.g., differential profiling
- Old customers: Do they match profile for their cluster?
  - Allows wider range of acceptable behavior
- New customer: Do they match *any* profile?

Has this been done? Intrusion Detection *(Lane&Brodley)*

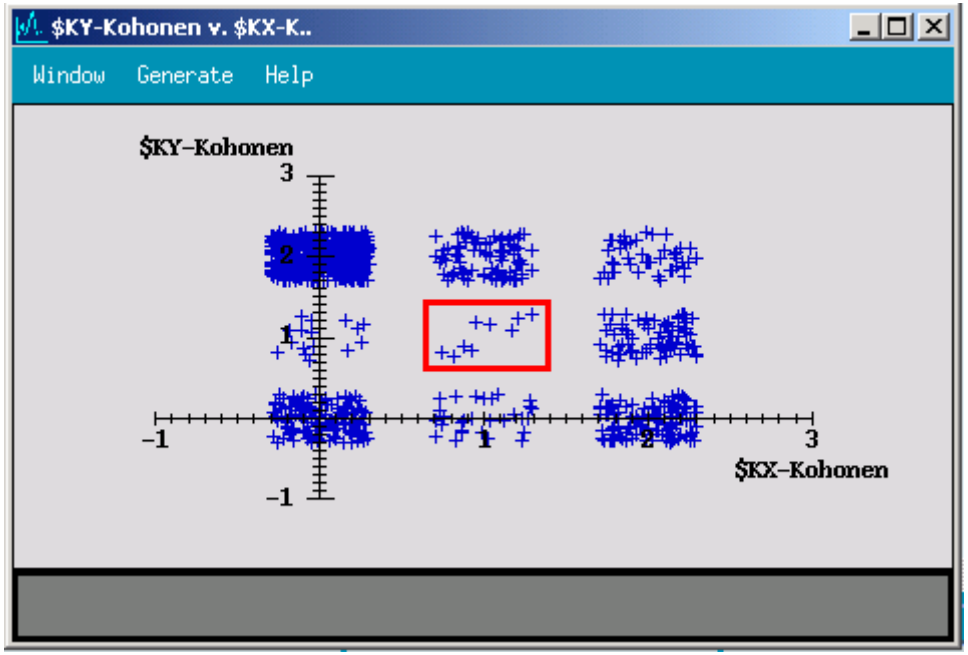
- Profiled computer users based on command sequences
  - Command
  - Some (but not all) argument information
  - Sequence information



Profiling

- ***Clustering and Associations***

  - Group behavior using a clustering algorithm
  - Find groups of events using the association algorithms
  - Identify outliers and investigate



PROGETTO "COOL PATTERNS"

ANALISI DELLE VENDITE NELLA GRANDE DISTRIBUZIONE

Analisi dei Dati ed Estrazione di Conoscenza 2004/2005

Federico Colla

Evaluation – Obj 1

Regole associative – Articoli

- [10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]
  - Support: 0,01 Confidence: 92,5 Lift: 4237,263
- [TELO 100X150 460 GR/MQ TU] [OSPITE 40X60 460 GR/MQ TU] → [ASCIUGAMANO 60X110 460 GR TU]
  - Support: 0,01 Confidence: 91,4 Lift: 965,993
- [BOCC.CANI POLLO/TACCH.KG1.23] [BOC/NI GATTO VITELLO SIM.KG415] → [BOCC.GATTI CONIGLIO SIMBA G415]
  - Support: 0,01 Confidence: 91,4 Lift: 390,042
- [PIATTO FRUTTA MAZIME B.CO CM21] [PIATTO F.DO MAXIME B.CO CM.17] → [PIATTO P.NO MAXIME B.CO CM.25]
  - Support: 0,01 Confidence: 90 Lift: 3052,386
- [LENZUOLO PIANO 150X280 RIGHE] [LENZUOLO ANGOLI 90X200 TU] → [FEDERA 50X80 STAMPA RIGHE]
  - Support: 0,01 Confidence: 87,8 Lift: 809,222

Evaluation – Obj 1

Regole associative – Articoli

- [10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]
  - Support: 0,01 Confidence: 92,5 Lift: 4237,263
- [TELO 100X150 460 GR/MQ TU] [OSPITE 40X60 460 GR/MQ TU] → [ASCIUGAMANO 60X110 460 GR TU]
  - Support: 0,01 Confidence: 91,4 Lift: 965,993
- [BOCC.CANI POLLO/TACCH.KG1.23] [BOC/NI GATTO VITELLO SIM.KG415] → [BOCC.GATTI CONIGLIO SIMBA G415]
  - Support: 0,01 Confidence: 91,4 Lift: 390,042
- [PIATTO FRUTTA MAZIME B.CO CM21] [PIATTO F.DO MAXIME B.CO CM.17] → [PIATTO P.NO MAXIME B.CO CM.25]
  - Support: 0,01 Confidence: 90 Lift: 3052,386
- [LENZUOLO PIANO 150X280 RIGHE] [LENZUOLO ANGOLI 90X200 TU] → [FEDERA 50X80 STAMPA RIGHE]
  - Support: 0,01 Confidence: 87,8 Lift: 809,222

Business Understanding:

Data mining goals

- Primo obiettivo – Analisi delle vendite
  - Regole associative a singola dimensione e multilivello.
  - Pattern sequenziali di prodotti venduti nel tempo, a diversi livelli di astrazione.

- Secondo obiettivo – Estrazione profilo clienti
  - Per ogni regola associativa/pattern sequenziale interessante
    - Crea un albero di decisione che classifica i clienti rispetto ad un attributo target il quale è positivo se il cliente soddisfa la regola associativa/pattern sequenziale, e negativo altrimenti.
    - L'analisi dell'albero risultante permette l'estrazione del profilo di interesse.

Modeling –

Per ogni regola "interessante:

- Il dataset finale è una tabella che contiene i dati di *tutti* e *soli* i clienti che hanno effettuato acquisti nel trimestre.
- La variabile target: ogni cliente ha associato un attributo binario (supporta o non supporta).
- Attributi predittori:
- A partire dall'albero di decisione ottenuto sono state generate le regole per la classificazione delle due classi.
- Per la creazione delle regole sono stati impostati livelli di confidenza minimi del 95%.

Attributo	Tipo
sesso	flag
stato_civile	insieme discreto
professione	insieme discreto
titolo_studio	insieme discreto
age	intervallo

Evaluation – Obj 2

Regole associative – Articoli

- La regola
  - [GOURM.GOLD DADINI GELLEE G85X8] [GOURMET PERLE FIL.C/ MANZO G85]→[GOURMET PERLE FIL.CONIGLIO G85]
  - Support: 0,01 Confidence: 87,8 Lift: 492,757
- è supportata da 41 clienti. Il classificatore ottenuto ha una accuratezza del 96,07% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Casalinghe, non sposate, di età tra i 57 e i 63 anni, che hanno la terza media inferiore come titolo di studio.
  - Ragazze single ragioniere tra i 26 e i 29 anni che lavorano come impiegate
  - Uomini pensionati aventi età minore di 51 anni

Evaluation – Obj 2

Regole associative – Articoli

- La regola
  - [APER.CAMPARI MIXX PEACH ML275] [APERIT.CAMPARI MIXX LIME ML275] [APERITIVO CAMP.GRADI 6,5 ML275] → [CAMPARI MIXX ORANGE ML275]
  - Support: 0,01 Confidence: 83,3 Lift: 1314,55
- è supportata da 27 clienti. Il classificatore ottenuto ha una accuratezza del 97,6% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Ingegneri maschi aventi 26-27 anni
  - Ragazze dai 26 ai 30 anni che hanno un lavoro autonomo e sono diplomate
  - Impiegati single dai 26 ai 53 anni

Evaluation – Obj 2

Regole associative – Subcategorie

- La regola
  - [USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-ACCESSORI USA E GETTA] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA]→[USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
  - Support: 0,1 Confidence: 84,7 Lift: 12,767
- è supportata da 158 clienti. Il classificatore ottenuto ha una accuratezza del 88,13% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Uomini celibi che lavorano per enti pubblici aventi 43-45 anni
  - Liberi professionisti aventi titolo di studio media inferiore di 59-60 anni
  - Militari di carriera sposati di aventi 38-41 anni

Evaluation – Obj 2

Regole associative – Subcategorie

- La regola
  - [SNACK SALATI-POP CORN/CEREALI] [SNACK SALATI-ESTRUSI] [BIBITE-ARANCIATE] → [BIBITE-COLE]
  - Support: 0,1 Confidence: 82,2 Lift: 10,34
- è supportata da 155 clienti. Il classificatore ottenuto ha una accuratezza del 86,73% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Ragazze disoccupate di 30-34 anni aventi un diploma magistrale
  - Vedovi di 57-60 anni liberi professionisti
  - Uomini/donne sposati di 32-40 anni e impiegati

INDVIDUAL PURCHAISING PROFILES

Obiettivo di analisi: realizzare un modello di segmentazione dei clienti basato sulla sistematicità degli acquisti. Tale sistematicità viene calcolata utilizzando due componenti: livello di regolarità dei prodotti nel carrello e livello di regolarità della componente spazio-temporale.

Data Preparation

- Filtro sulla dimensionalità dei dati
  - Si considera il solo venduto nell'anno 2012
  - Si considerano i soli negozi della provincia di Livorno
  - Si considerano solo i clienti frequenti (che abbiano fatto almeno una spesa al mese)

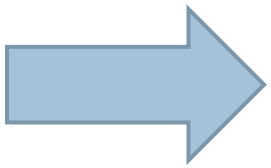
Data Preparation (cnt)

- Dimensioni del dataset dopo il filtro
  - 71.172.672 scansioni
  - 56.448 clienti
  - 84.362 articoli distinti
  - 23 negozi (1 Iper, 9 Super e 13 GestIn)

Data Preparation (cnt)

- Per il calcolo del BRI (indice di regolarità del carrello) bisogna applicare aPriori, quindi i dati vanno trasformati, per ogni cliente, da formato relazionale a formato transazionale

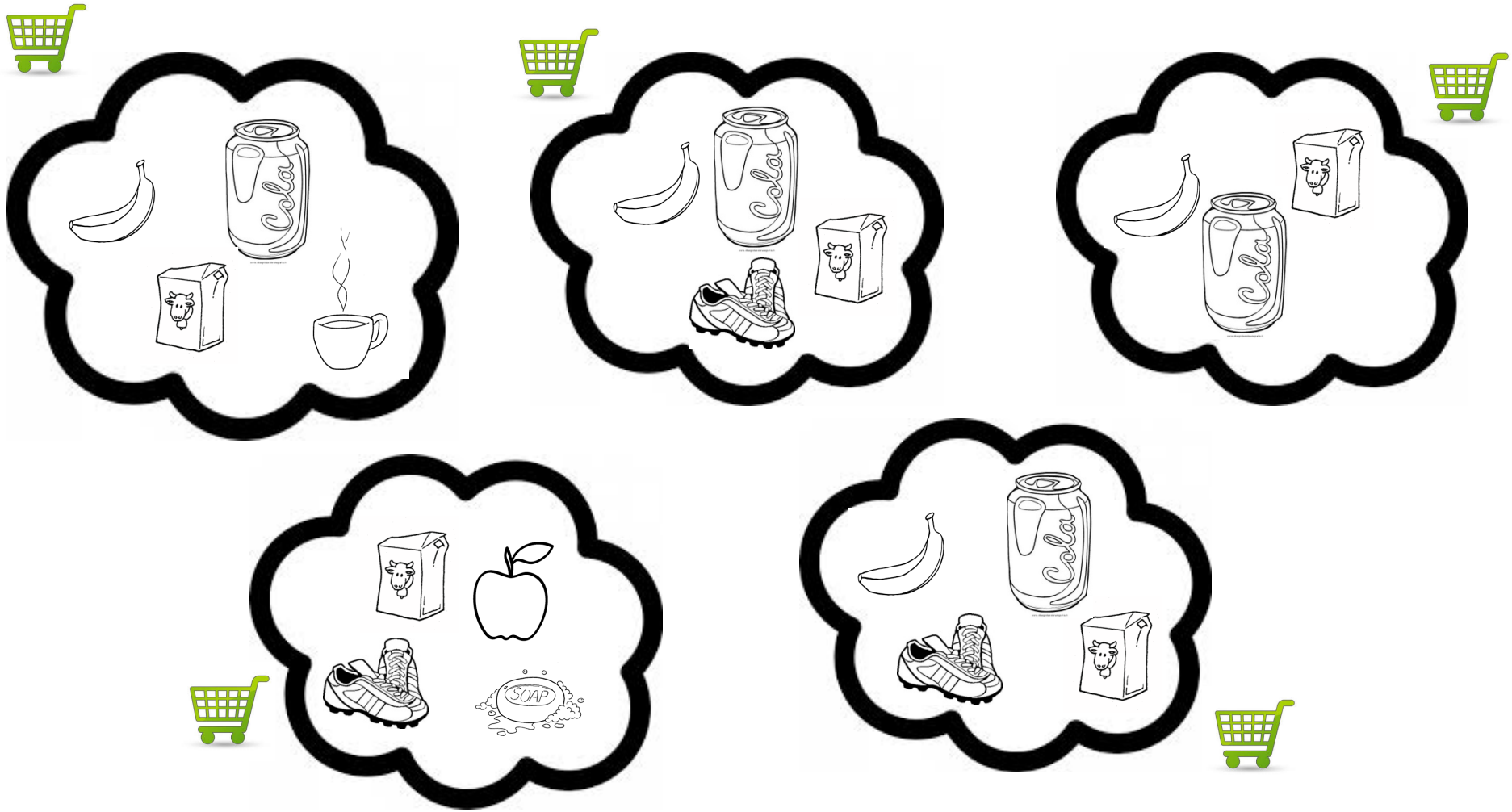
Scontrino	Prodotto
1	A
1	B
1	C
1	D
2	A
2	D
2	F



Scontrino	Lista Prodotti
1	A B C D
2	A D F

II Basket Regularity Index (BRI)

Starting from a set of baskets



II Basket Regularity Index (cnt)

Performing Frequent Pattern Mining (min supp=3)


Sup=5


Sup=4


Sup=4


Sup=4


Sup=4


Sup=3

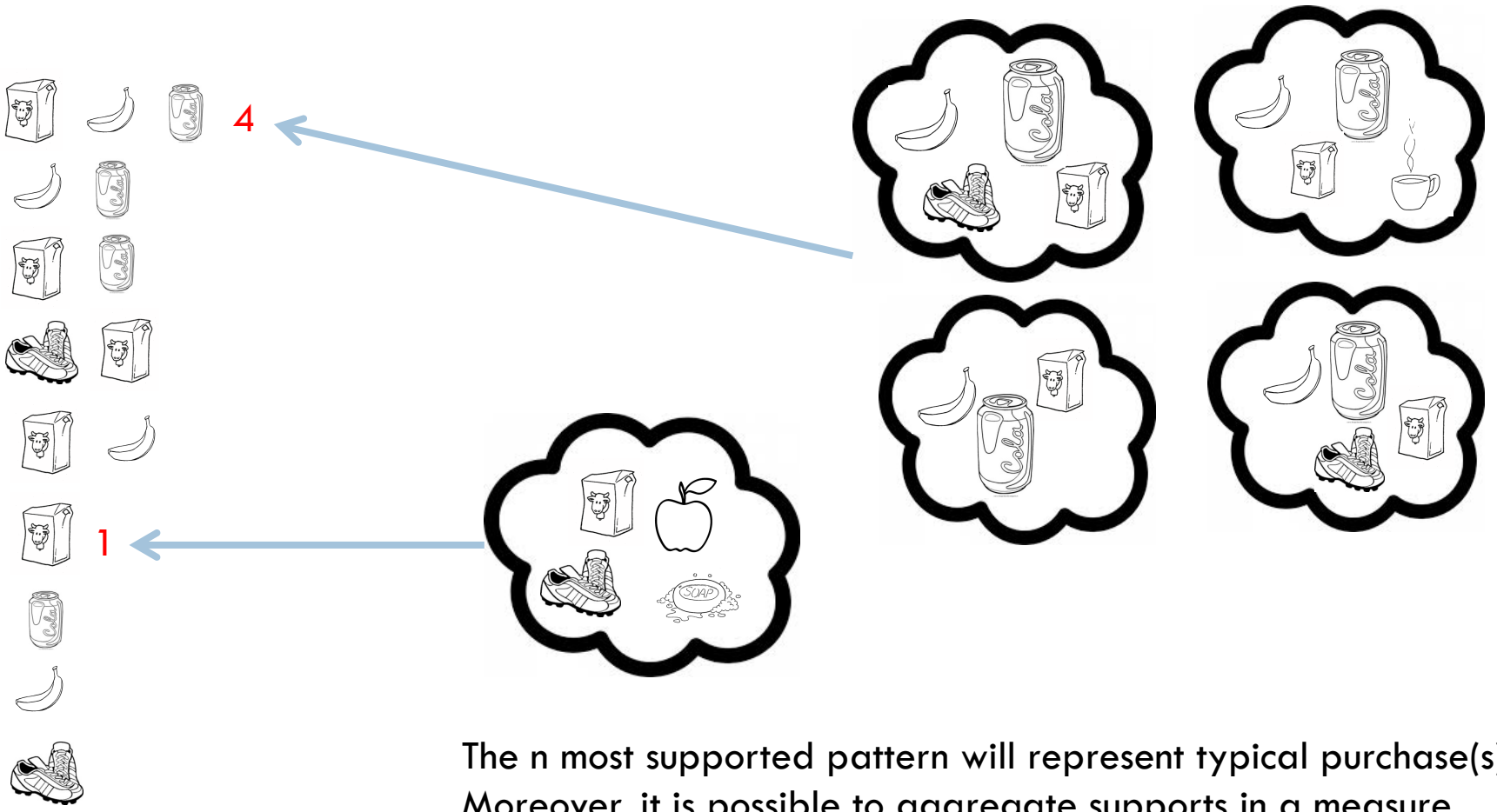

Sup=3


Sup=4


Sup=4

II Basket Regularity Index (BRI)

Each basket will support the longest pattern contained



The n most supported pattern will represent typical purchase(s). Moreover, it is possible to aggregate supports in a measure describing the Basket Regularity Index (BRI) of the customer

II Basket Regularity Index (BRI)

The BRI is the information Entropy calculated among all the supported patterns

 4

 1

$$BRI = -\left(\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log\frac{1}{5}\right) = -\left(0.8 \cdot -0.321928 + 0.2 \cdot -2.321928\right) = 0.721928$$

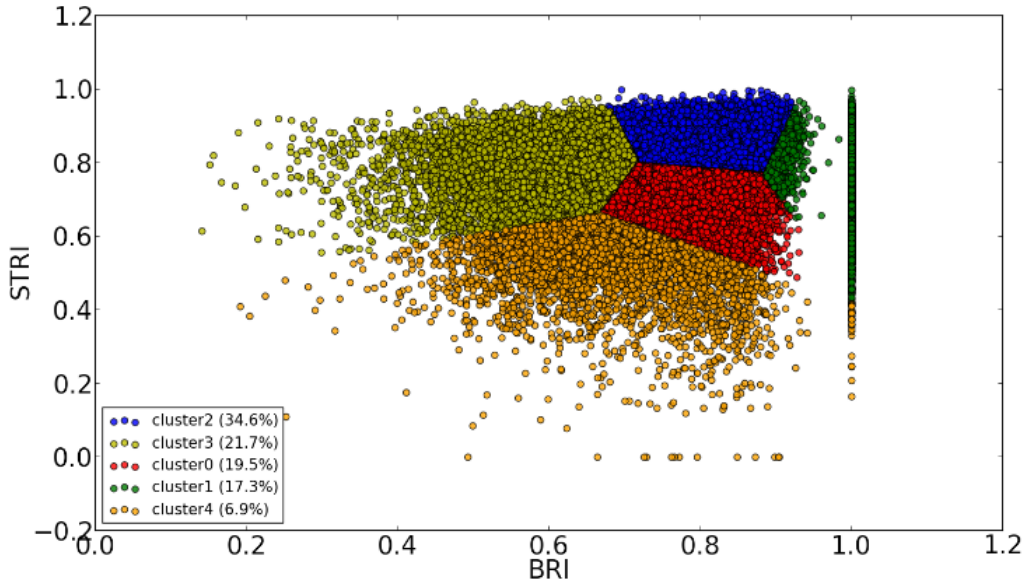
Lo Spatio-Temporal Regularity Index (STRI)

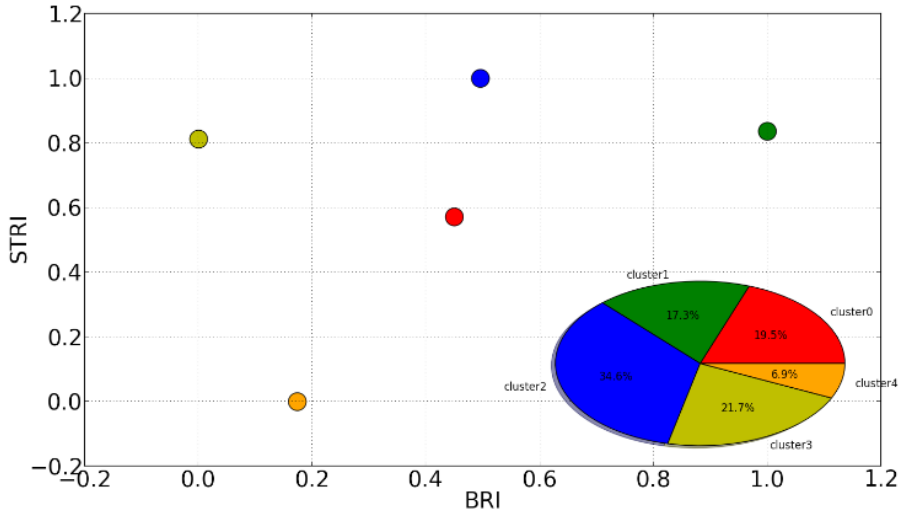
- Si salta il passaggio con aPriori, in quanto le informazioni sono più concise
- Si costruiscono triple contenenti informazioni su:
  - Negozio dove si è fatta la spesa
  - Tipo di giorno in cui si è fatta la spesa (feriale-festivo)
  - Fascia Oraria in cui si è fatta la spesa (mattina presto, tarda mattinata, primo pomeriggio, secondo pomeriggio, sera)
- Per ogni cliente, si calcola l'entropia su tali triple

Analisi: segmentazione

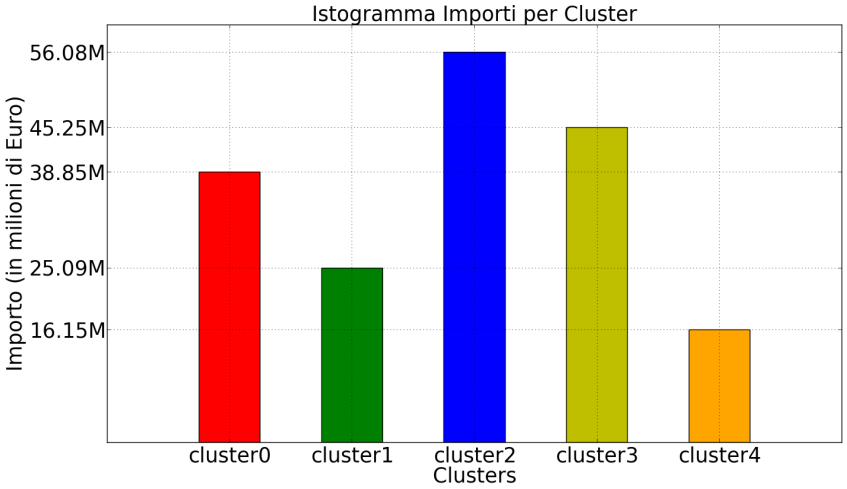
- Ad ogni cliente, quindi, sarà associata una coppia di indici (BRI e STRI)
- Queste coppie di indici possono essere riportate su un piano e sulle stesse possono essere applicati degli algoritmi di clustering
- In questo modo, la clientela verrà segmentata in base alla loro propensione ad essere sistematici per una delle due componenti principali del processo di spesa (cosa compro e quando/dove compro)

Analisi: segmentazione (cnt)

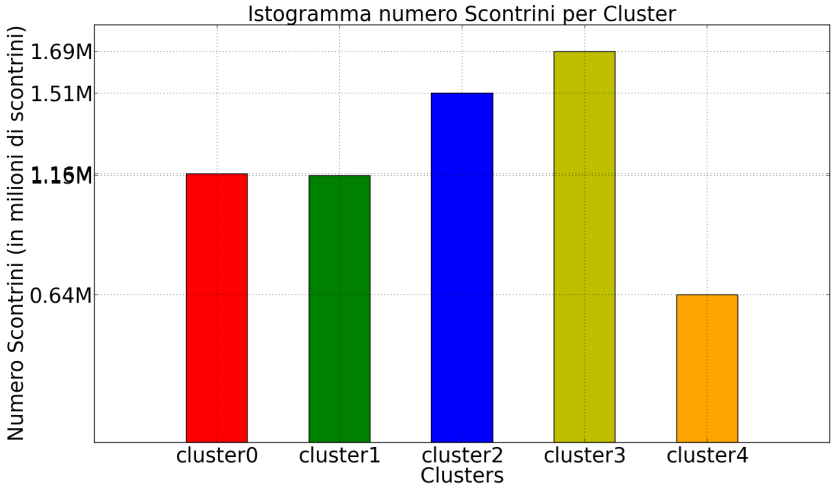




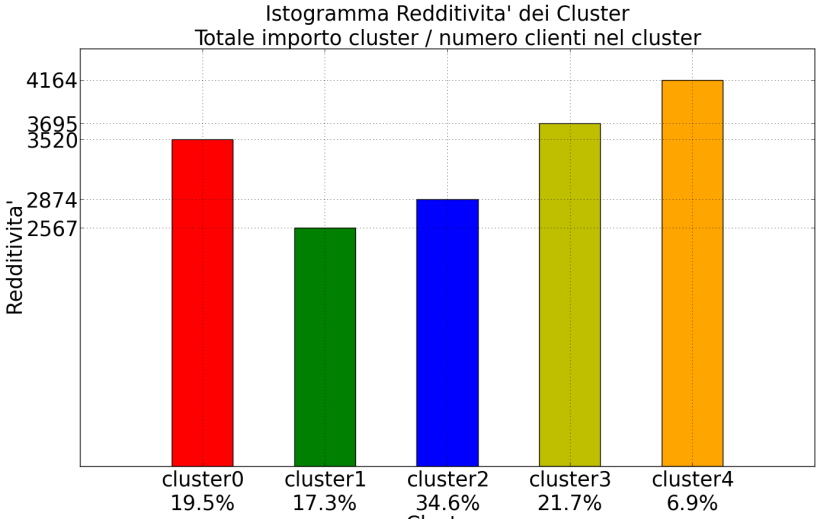
Analisi: segmentazione (cnt)



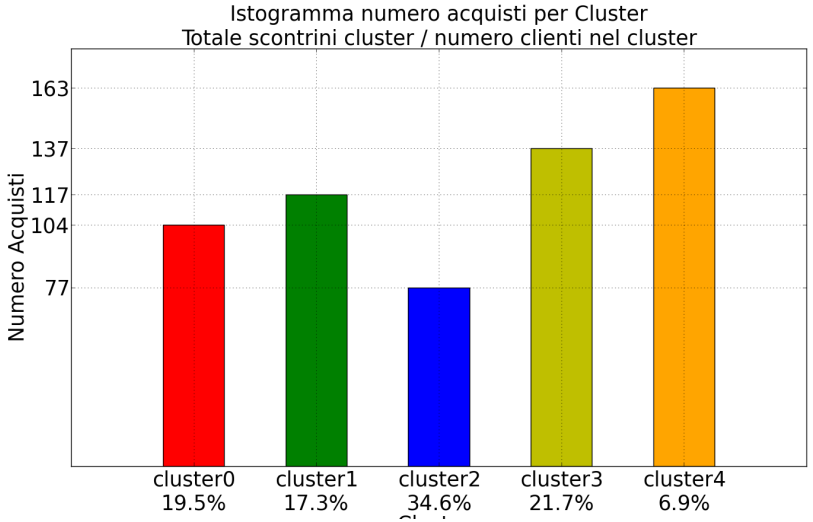
Tot importi



Tot scontrini

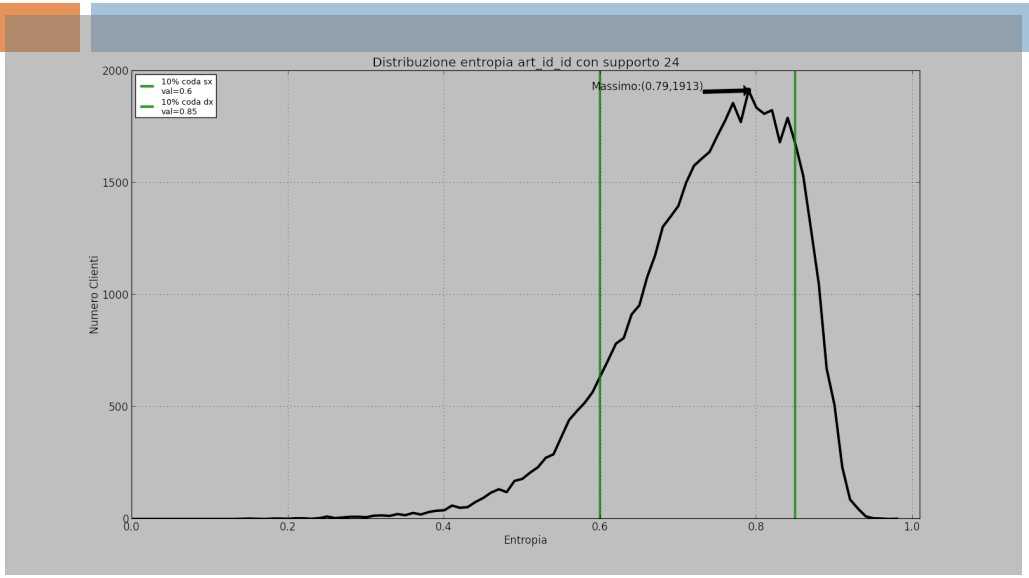


Importi medi



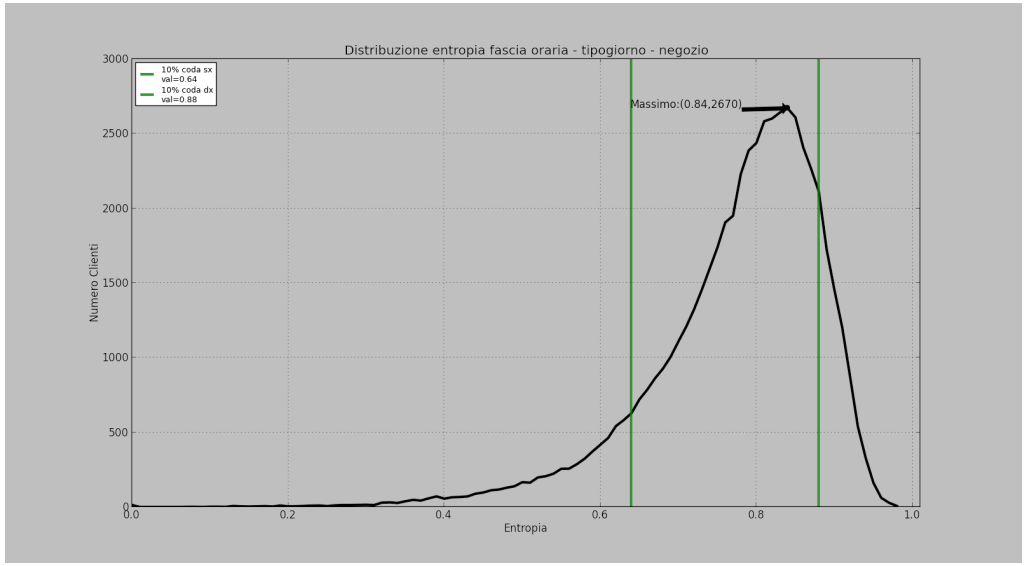
Scontrini medi

Analisi degli indici



BRI

STRI



Il 60% dei bi-sistematici compra

- ZUCCHINE SCURE IT 14-21 I^ SF
- PESCHE GIALLE IT AA I^ SF
- POMODORI OBLUNGO VERDE IT 35-40 I^ SF
- POMODORI CILIEGINO IT I^ VH G 500
- POMODORI ROSSO GRAPPOLO IT I^ SF
- BANANE COOP ES 19+ I^ SF
- UOVA FRESCHE ALLEVAM A TERRA M COOP POLPA LEGNO X6
- PREZZEMOLO G. 70 CA IT MAZZI SFUSO
- PP-ZUCCHERO SEMOLATO ITALIA ZUCCHERI SCATOLA KG 1
- MELONE SEMIRETATO IT 800-1200 I^ SF
- SCONTO FIDELITY 1000 PUNTI
- RIVISTA 'NUOVO CONSUMO'

- UOVA EXTRA FRESCHE ALLEVAM A TERRA L COOP POLPA LEGNO X6
- POM.OBL.ROSSO IT30-35 I^ SF
- POM.TONDO LISCIO IT 67-82 I^SF
- ZUCCHINE CHIARE FIORE IT I^ SF
- FINOCCHI IT 8-10 I^ SF

SHOPPER COMPOSTABILE BIOFLEX ST.COOP

"IT'S A LONG WAY TO THE TOP…" PREDICTING SUCCESS VIA INNOVATORS' ADOPTIONS

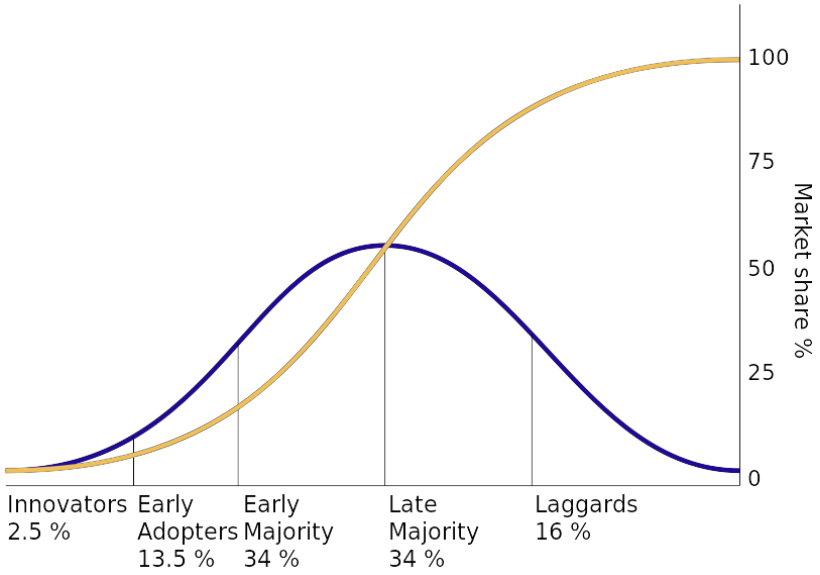
Obiettivo di analisi: esistono consumatori "speciali" che hanno una propensione ad adottare nuovi prodotti/tecnologie prima di altri? Tra questi, esiste un sottogruppo con un "sesto senso" per prodotti che avranno successo?

DOMANDA

- **E**sistono consumatori “speciali” che hanno una propensione ad adottare nuovi prodotti/tecnologie prima di altri? Tra questi, esiste un sottogruppo con un “sesto senso” per prodotti che avranno successo?

Punto di partenza

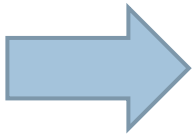
- Rogers identifica le caratteristiche distintive di ogni tipo di adopter
- Si assume che la distribuzione temporale degli Adopters segua una distribuzione normale
- Gli innovatori sono i primi 2.5% tra gli adopters
- ...ma è realmente così?



Data Preparation

- Per studiare le adozioni, occorre modellare i dati con serie temporali

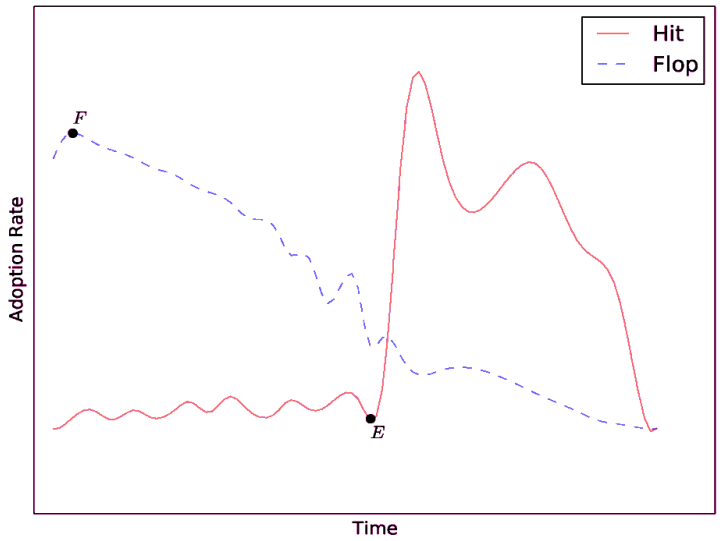
Prodotto	Cliente	Unità temporale
1	A	U1
1	B	U2
1	C	U1
2	C	U2
2	D	U2
2	A	U4
2	E	U4



Prodotto	Serie Temporale
1	<U1,2> , <U2,1>
2	<U1,1>,<U2,2>,<U4, 2>

Hit and Flop: definizioni

- Un **Hit** è un bene il cui trend di adozione cresce lentamente nel tempo fino a raggiungere un punto di esplosione (E nella figura), che segna l'inizio di un forte aumento del numero di adozioni.
- Un **Flop** è un bene il cui trend di adozione non cresce in maniera considerevole nel tempo, oppure raggiunge subito un punto di massimo (F in figura) per poi decrescere rapidamente.



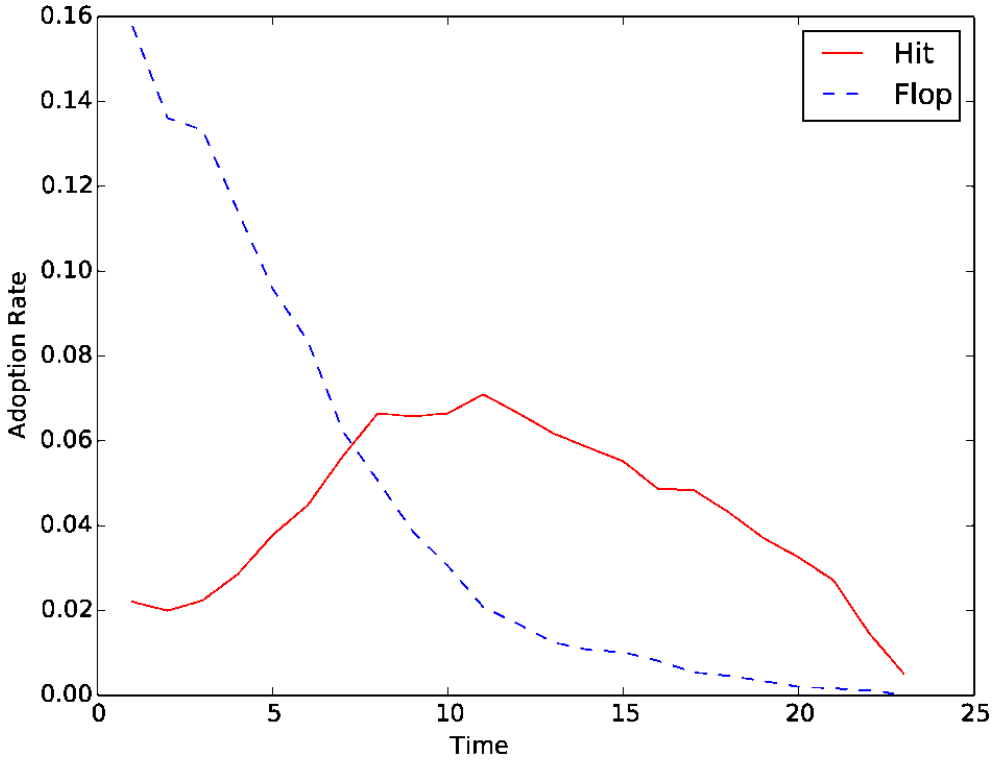
- Non si considerano i volumi di vendita, ma i trend relativi
- La definizione di "successo" non corrisponde a quello in senso comune (un nuovo iPhone cadrebbe nella nostra definizione di Flop)

Il Metodo

- Si estraggono i profili dei beni Hit e Flop
- Si individuano gli innovatori
- Si calcolano gli indicatori di successo/insuccesso per tutti gli innovatori
- Si raffinano e si consolidano gli insiemi di indicatori positivi/negativi
- Si definisce una tecnica di predizione rule-based utilizzando gli indicatori calcolati in precedenza

Sperimentazione

- Una volta costruite le serie temporali per tutti i prodotti nuovi, queste vengono clusterizzate



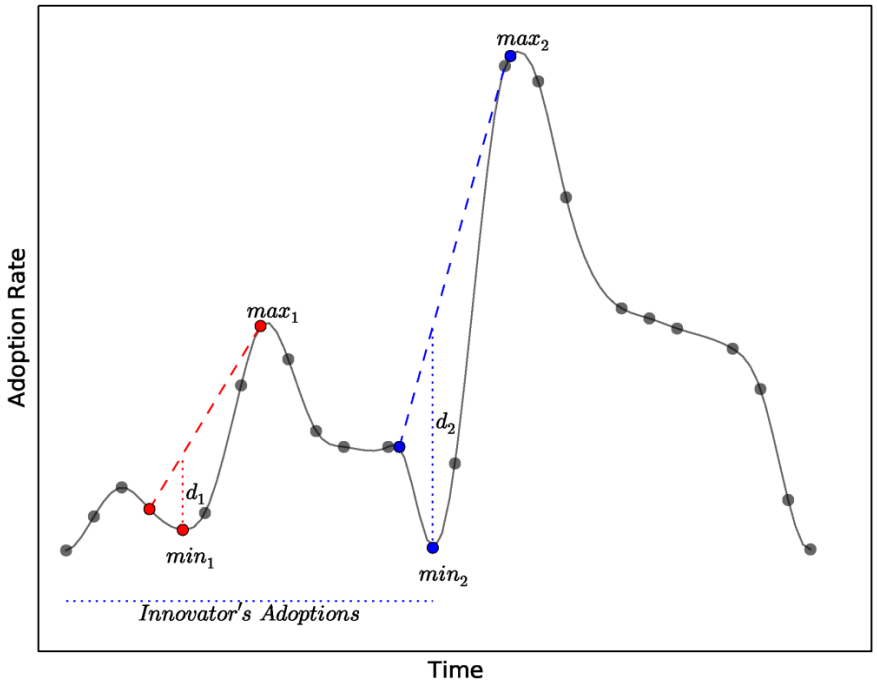
Clustering: K-means

K=2 con SSE

Unità temporale: settimana

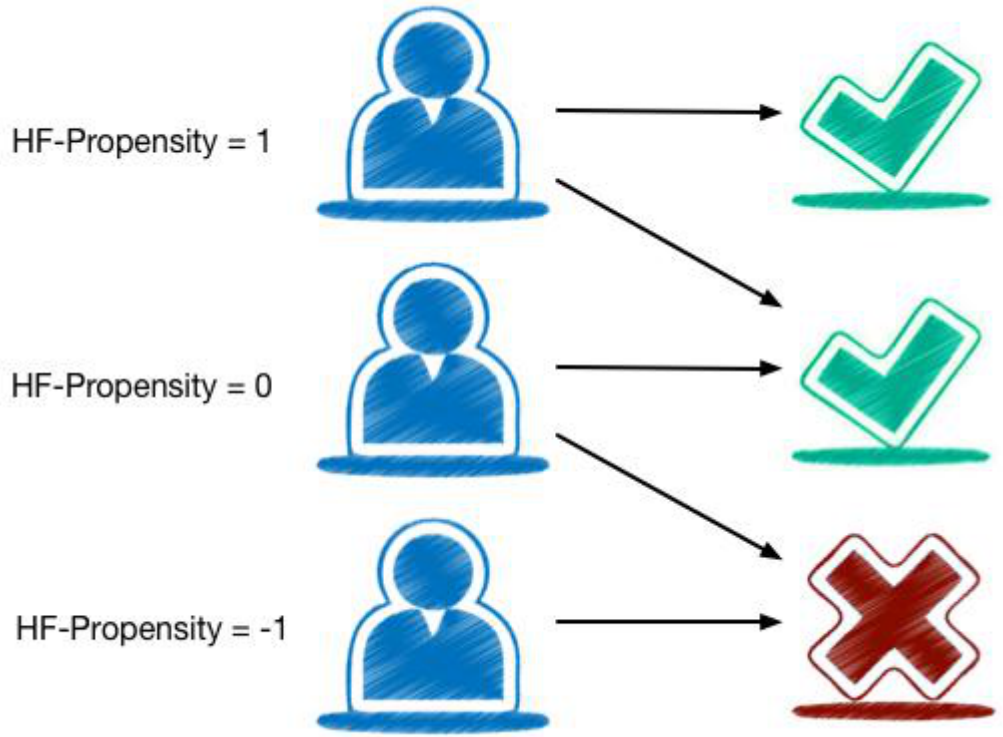
Identificazione Innovatori

- Si identificano i minimi locali (min_1 e min_2) e i massimi locali (max_1 e max_2)
- Si calcolano le distanze d_1 e d_2
- Si sceglie min_2 come punto discriminante, in quanto $d_2 > d_1$
- Si definiscono innovatori tutti i clienti che hanno adottato il bene prima di min_2



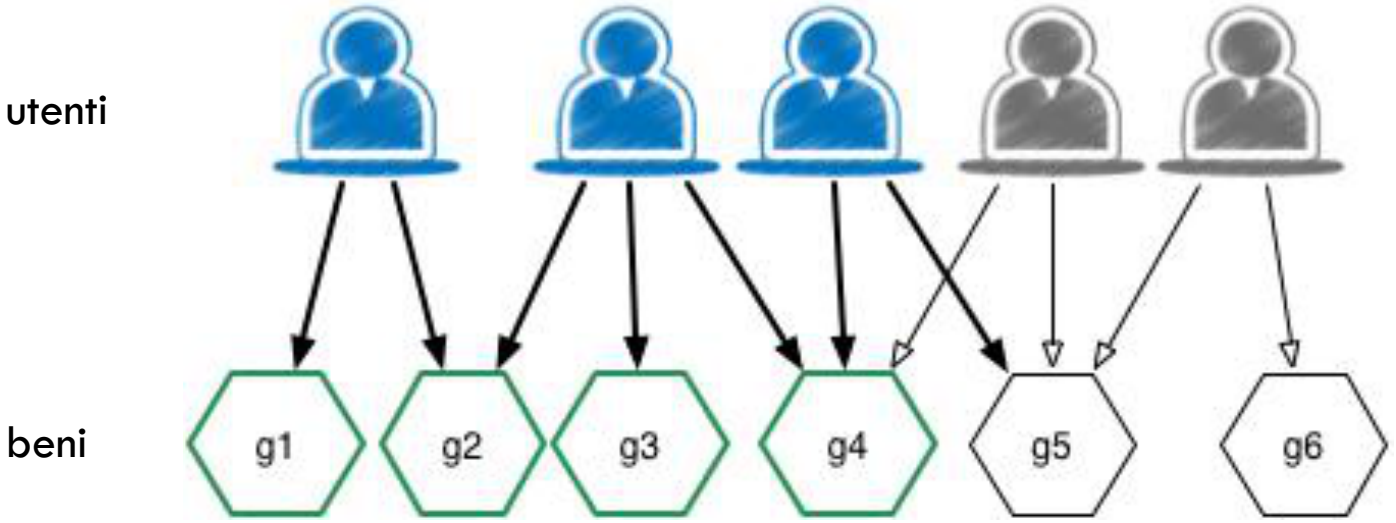
Propensione degli innovatori per i beni Hit e Flop

Si misura, per ogni cliente, la sua propensione ad adottare beni Hit o Flop, assegnando un +1 se ha adottato un Hit da innovatore, un -1 se ha adottato un Flop, nessun incremento/ decremento se ha adottato un Hit dopo l'esplosione



Identificazione degli Hitters e dei Floppers

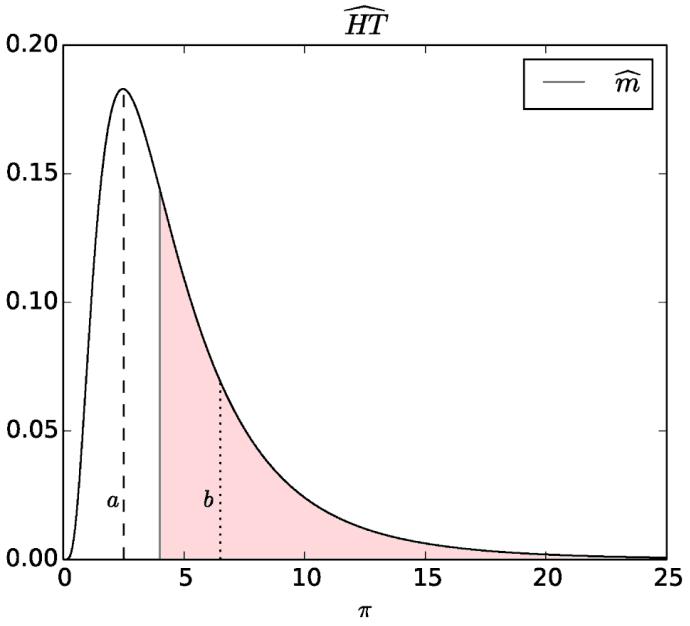
- Bisogna ridurre l'insieme di clienti in modo da filtrare solo gli indicatori forti per gli Hit ed i Flop
- Si costruiscono due grafi bi-partiti <cliente,prodotto> (uno per beni Hit ed uno per beni Flop), dove un arco tra un nodo cliente ed un nodo prodotto viene tracciato se quel cliente ha adottato quel prodotto
- Si risolvono i due problemi di Copertura Minima Pesata (con HF-propensity) dei prodotti
- I clienti selezionati coprono collettivamente almeno una frazione α dell'insieme dei beni ed almeno una frazione β degli archi entranti (supporto) dei beni

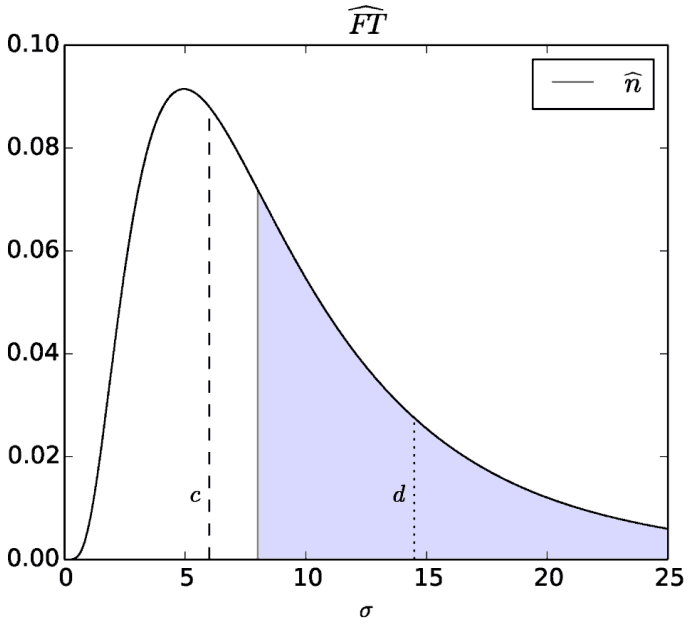


Esempio con
$\alpha = \beta = 0.6$

Modello predittivo

- Si tracciano le distribuzioni del numero di Hitter (sx) e Flopper (dx) che hanno adottato i prodotti
- Si tracciano le mediane delle funzioni
- Per ogni prodotto g, si calcolano $\pi(g)$ (#Hitters che hanno adottato g) e $\sigma(g)$ (#Floppers che hanno adottato g)





$\pi(g) = a$ $\sigma(g) = d$
Flop

$\pi(g) = b$ $\sigma(g) = c$ **Hit**

$\pi(g) = b$ $\sigma(g) = d$
Flop (calcolo distanze)

$\pi(g) = a$ $\sigma(g) = c$ **?**

Analisi sperimentale

- 5.605 articoli
- 620.026 clienti
- 11.204.984 clienti
- Periodo di un anno

- 3.749 Hits
- 1.856 Flops

Risultati

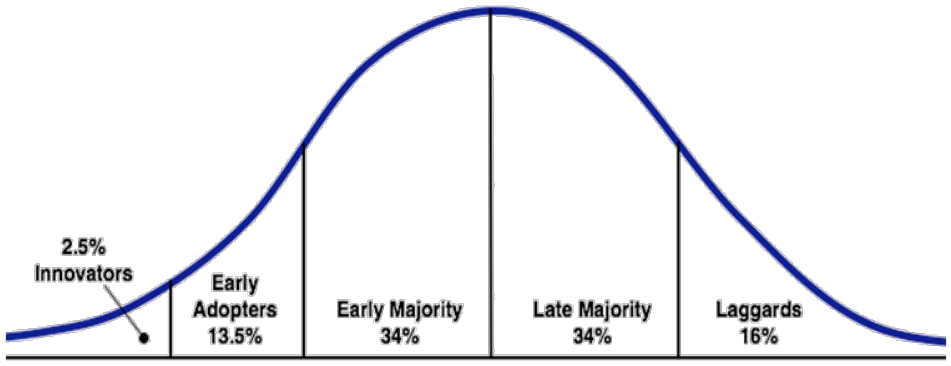
- Il modello viene confrontato con
  - Null Model (tutti i clienti comprano in un istante random)
  - ER-H&F: la fase di identificazione degli innovatori è sostituita considerando la soglia di Rogers (2.5%)
  - ER: la predizione viene fatta utilizzando soltanto gli innovatori estratti secondo la soglia del 2.5%

Coop	H&F		ER-H&F		ER		Null Model	
	avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
Precision	0.781	0.91	0.825	0.211	0.00	0.00	0.547	0.010
NPV	0.316	0.121	0.384	0.057	0.292	0.00	0.051	0.028
Recall	0.586	0.292	0.031	0.017	0.00	0.00	0.818	0.043
Specificity	0.522	0.367	0.983	0.019	1.00	0.00	0.361	0.024

Adopters: Innovators

- **Diffusion of Innovations**
  [Rogers 1962]

- Five "category" of **Adopters** based on the time of first adoptions:
  - Each one has its own semantics;
  - Temporal distribution
    Assumed to be a Gaussian;
  - Categories proportion is univocally determined
    (i.e. Innovators are <u>always</u> the first 2.5%)



Goods: Hits and Flops

- Retail market products,
- Music Artists,
- Business and stores…

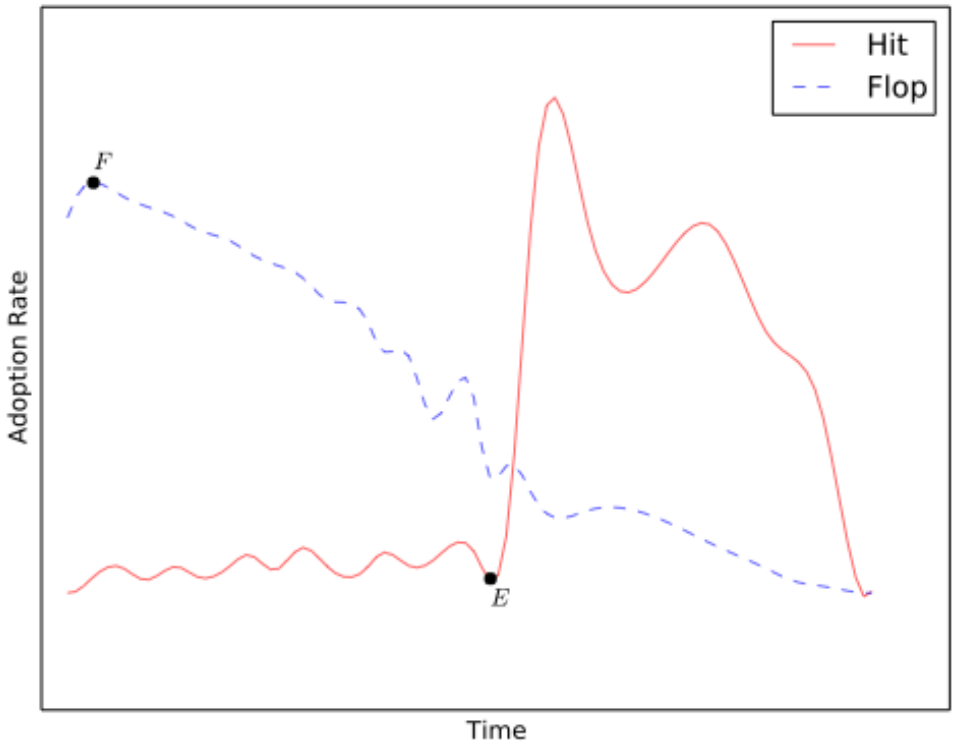
What is a successful (Hit) good?

And an unsuccessful (Flop) one?

Hits and Flops share the same set of adopters or not?

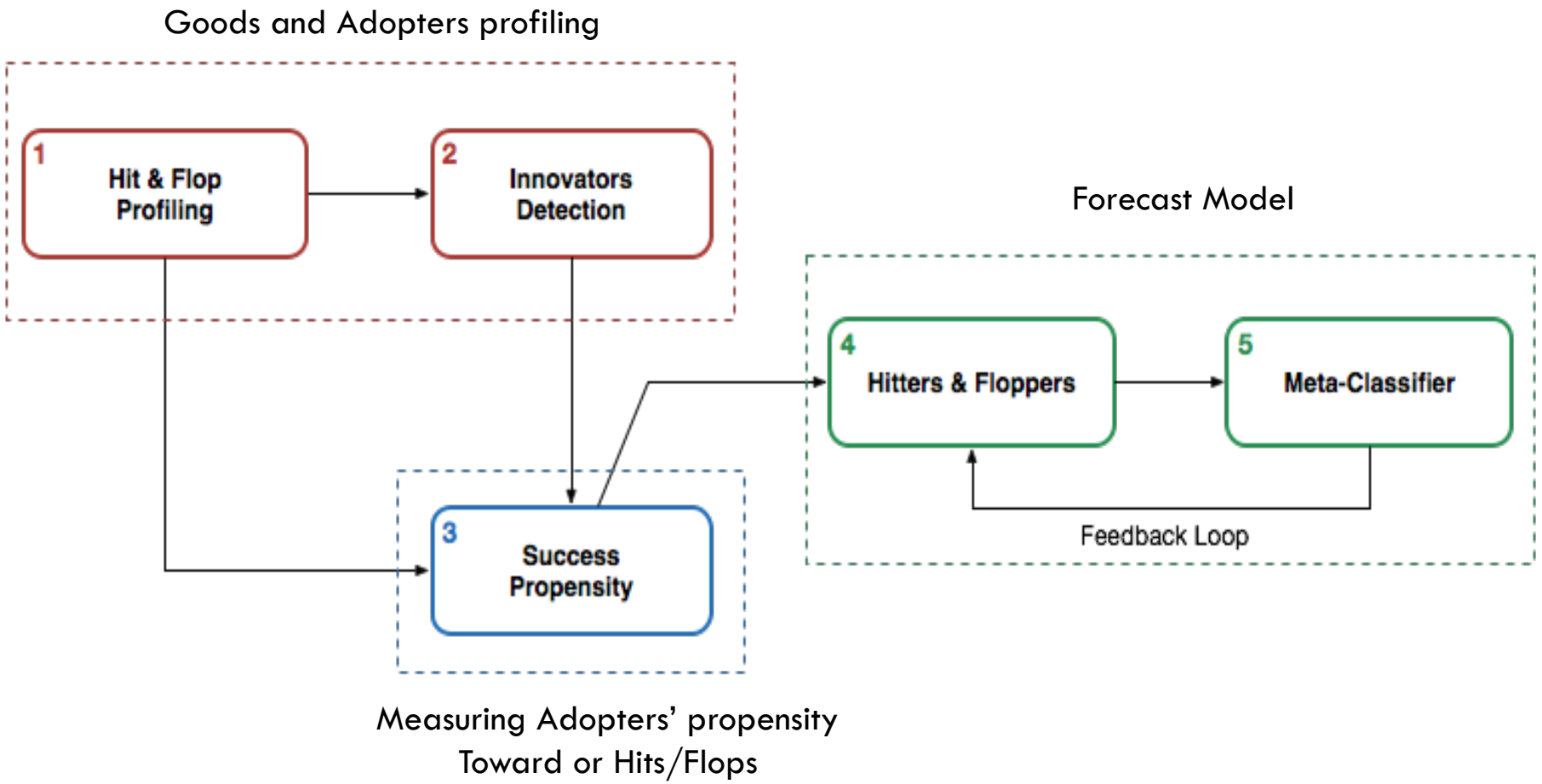
Hits & Flops: qualitative definitions

- **Hit**
  - A good whose trend slowly increases trough time until reaching an explosion point that marks the start of a sharp rising of its adoptions.
- **Flop**
  - A good whose adoption trend does not increase considerably over time or even reaches an early maximum only to sharply decrease.



Given a **partial observation** of the **adoptions** of a **novel good** can we decide if it will became an **Hit** or a **Flop**?

Hit&Flop: Workflow



Forecast Evaluation*

COOP	H&F	ER-H&F	ER	NM
PPV	**.781**(.09)	.825(.21)	0(0)	.547(.01)
NPV	.316(.12)	.384(.06)	.292(0)	.05(.03)
Recall	**.586**(.29)	.03(.01)	0(0)	.818(.04)
Specificity	.522(.38)	.982(.02)	1(0)	.361(.02)

Last.fm	H&F	ER-H&F	ER	NM
PPV	**.766**(.03)	.290(.37)	0(0)	.644(0)
NPV	**.471**(.04)	.047(.39)	.351(0)	.026(.04)
Recall	.520(.04)	.006(.01)	0(0)	.990(.02)
Specificity	.727(.06)	.970(.02)	1(0)	.007(.01)

Yelp	H&F	ER-H&F	ER	NM
PPV	**.990**(.01)	1(0)	0(0)	.488(.04)
NPV	**.631**(.17)	.341(.11)	.306(0)	.099(.08)
Recall	**.897**(.09)	.654(.11)	0(0)	.933(.01)
Specificity	**.906**(.10)	1(0)	1(0)	.007(.01)

Datasets

Dataset	Goods	Adopters	Adoptions	Period	Obs. window
COOP	5605	620026	11204984	1 year	4 weeks
Last.fm	1806	50837	882845	2 years	2 months
Yelp	2499	141936	427894	10 years	30 months

Competitors

H&F: Hits&Flops

ER-H&F: Hits&Flops with Roger's Innovators

ER: Rogers's Innovators

NM: Hits&Flops on Null Model (avg. 100 models)

Results in a nutshell

- H&F guarantee the most stable predictive performances in terms of PPV and Recall
- ER is not able to provide useful classification (2.5% fixed innovator threshold)
- ER–H&F suffer the constrains imposed by ER

*Results after a 10-fold cross validation