Data Mining on Promotional Sales

Obiettivi

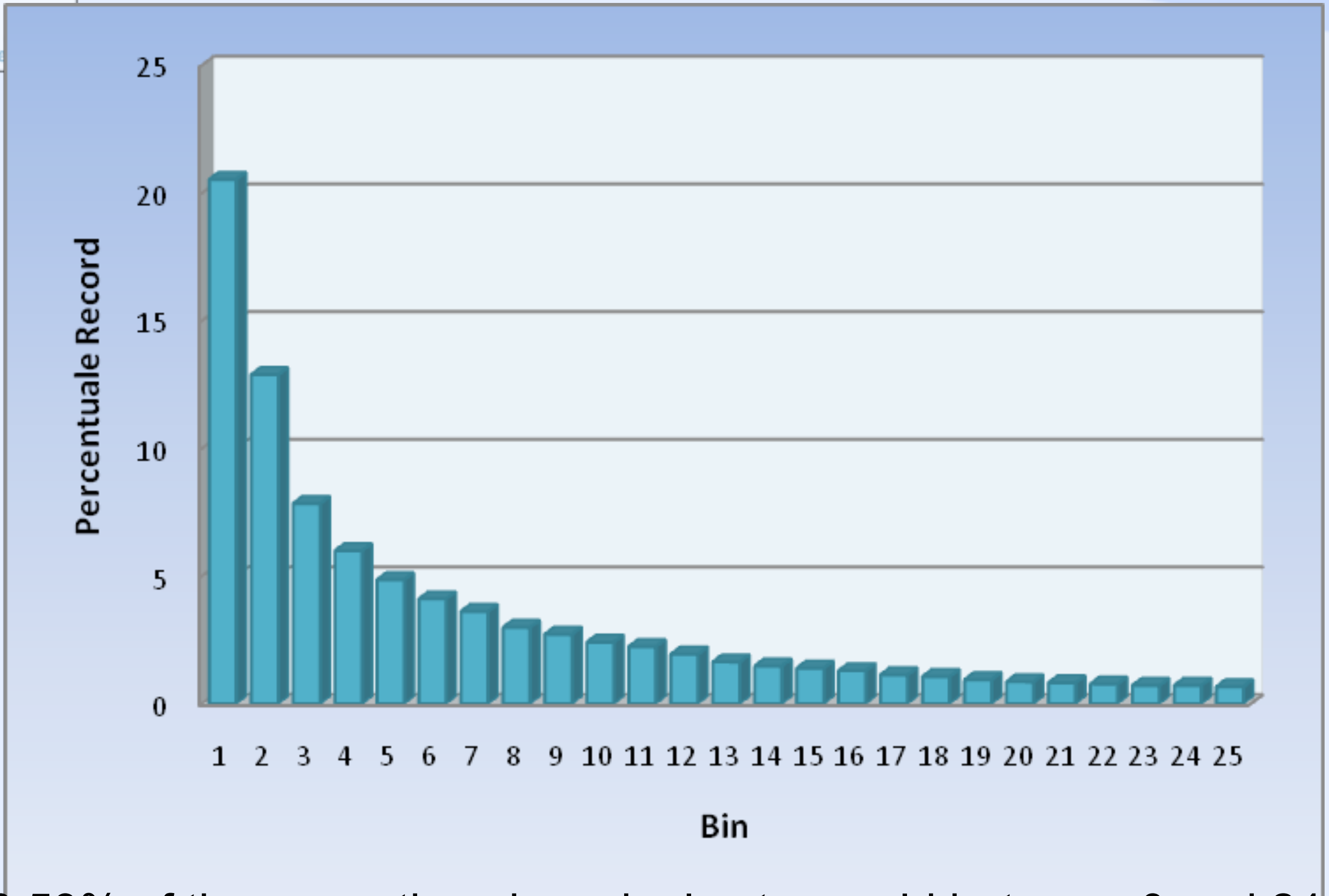
Goal: focusing on the effects of promotions on the sales of a single product, mainly aimed at optimizing its stoking:

- Forecasting sales of promoted products .
- Forecasting “out-of-stock”

Case study on product category = Food,

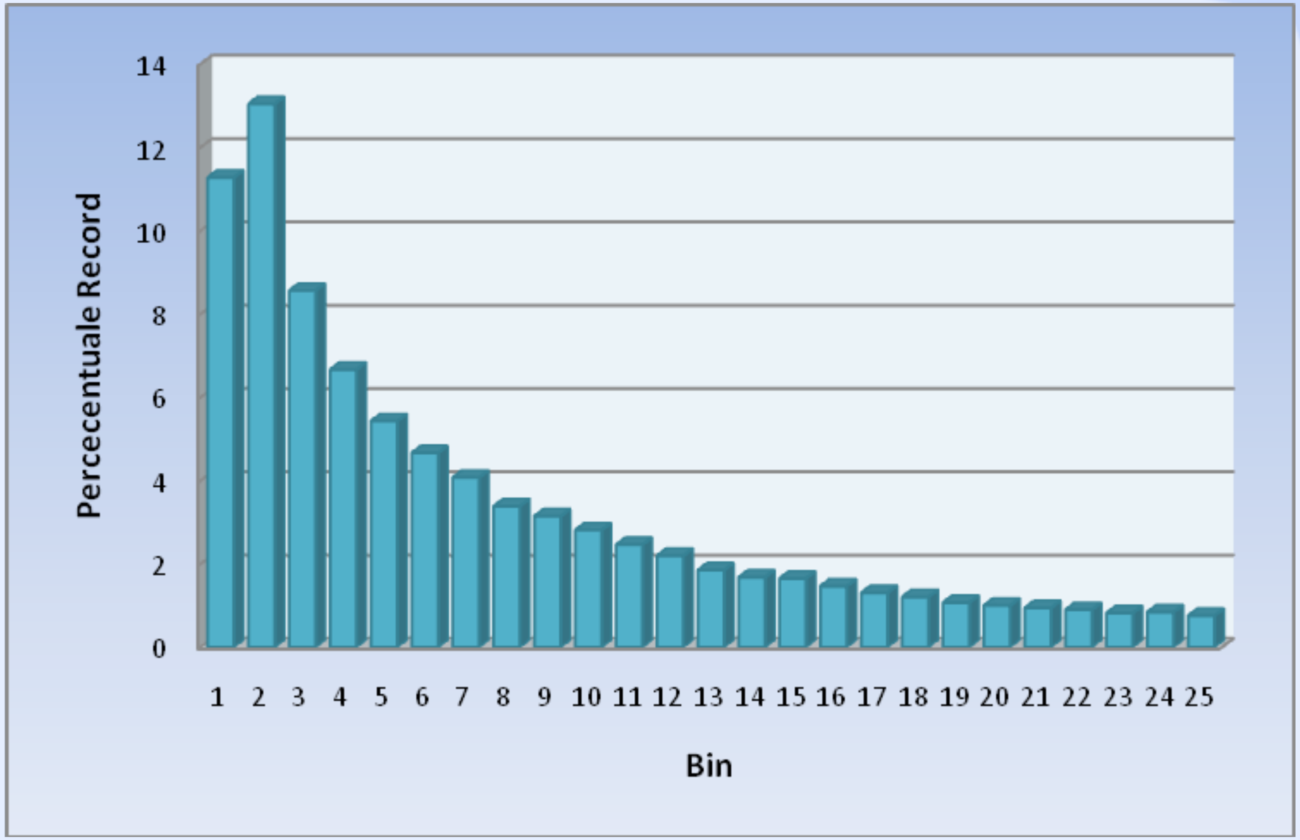
Two months April 2006 & April 2007.

Sales volume distribution in hypermarkets with equal-size binning (0-25-50-75-)



- 20.53% of the promotions in a single store sold between 0 and 24 items (the leftmost bar in the figure),
- the 12.89% sold between 25 and 49 items,
- Many promotions with almost 0 sold items

Sales volume distribution of promotions with al least 5 items sold



- 11.% of the promotions in a single store sold between 0 and 24 items (the leftmost bar in the figure),
- the 13.79% sold between 25 and 49 items,
- The tale is less flat

Data Preparation

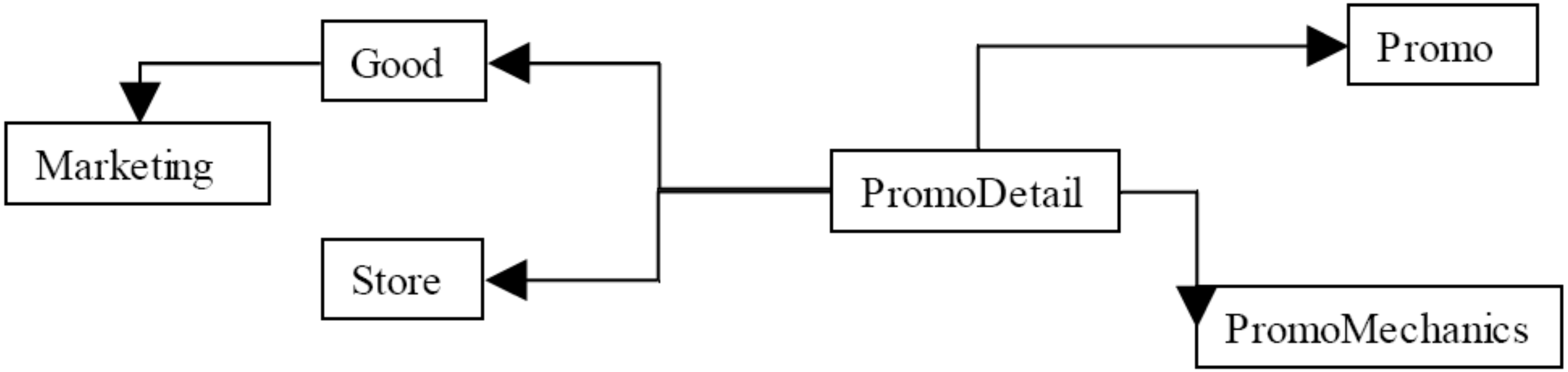


Figure 3: structure of relevant portion of the data warehouse

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).

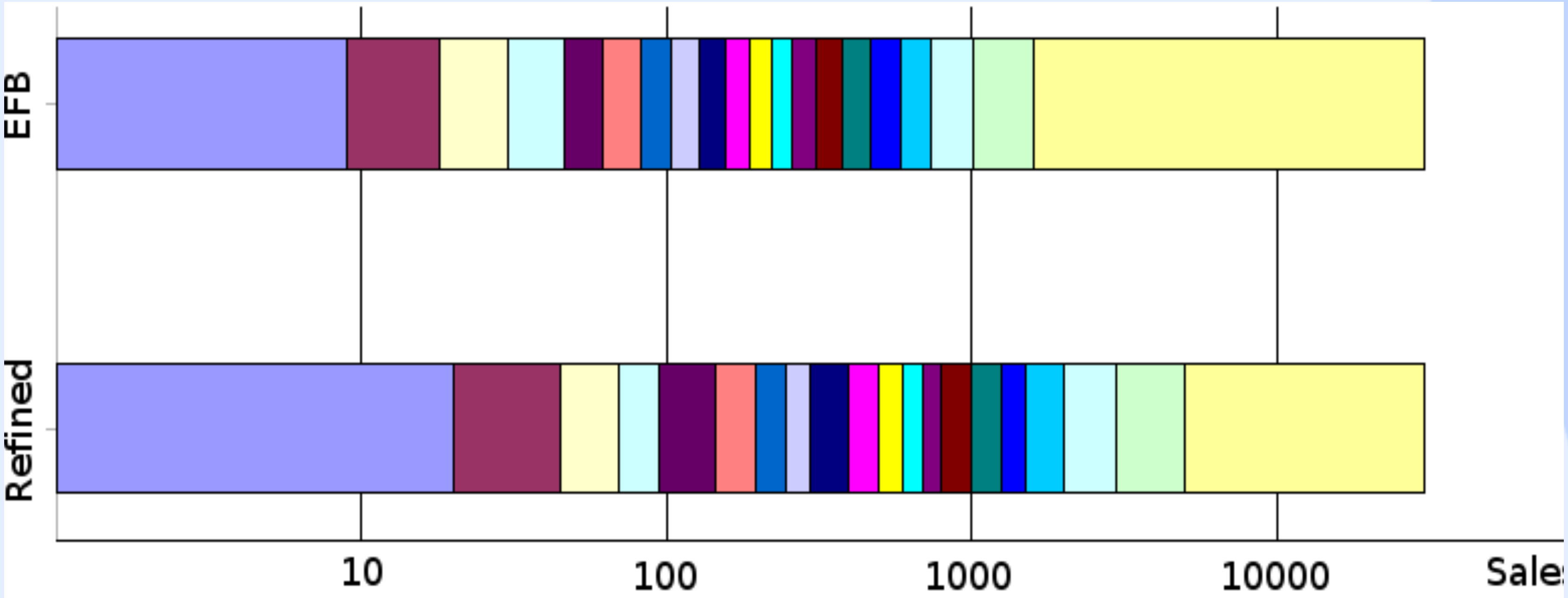
The target variables used to train the models are:
(1)the sales amount of the promoted item and
(2) The number of out of stock that occurred during the promotion.

MINING TASK: MULTICLASS over ORDINAL CLASSES

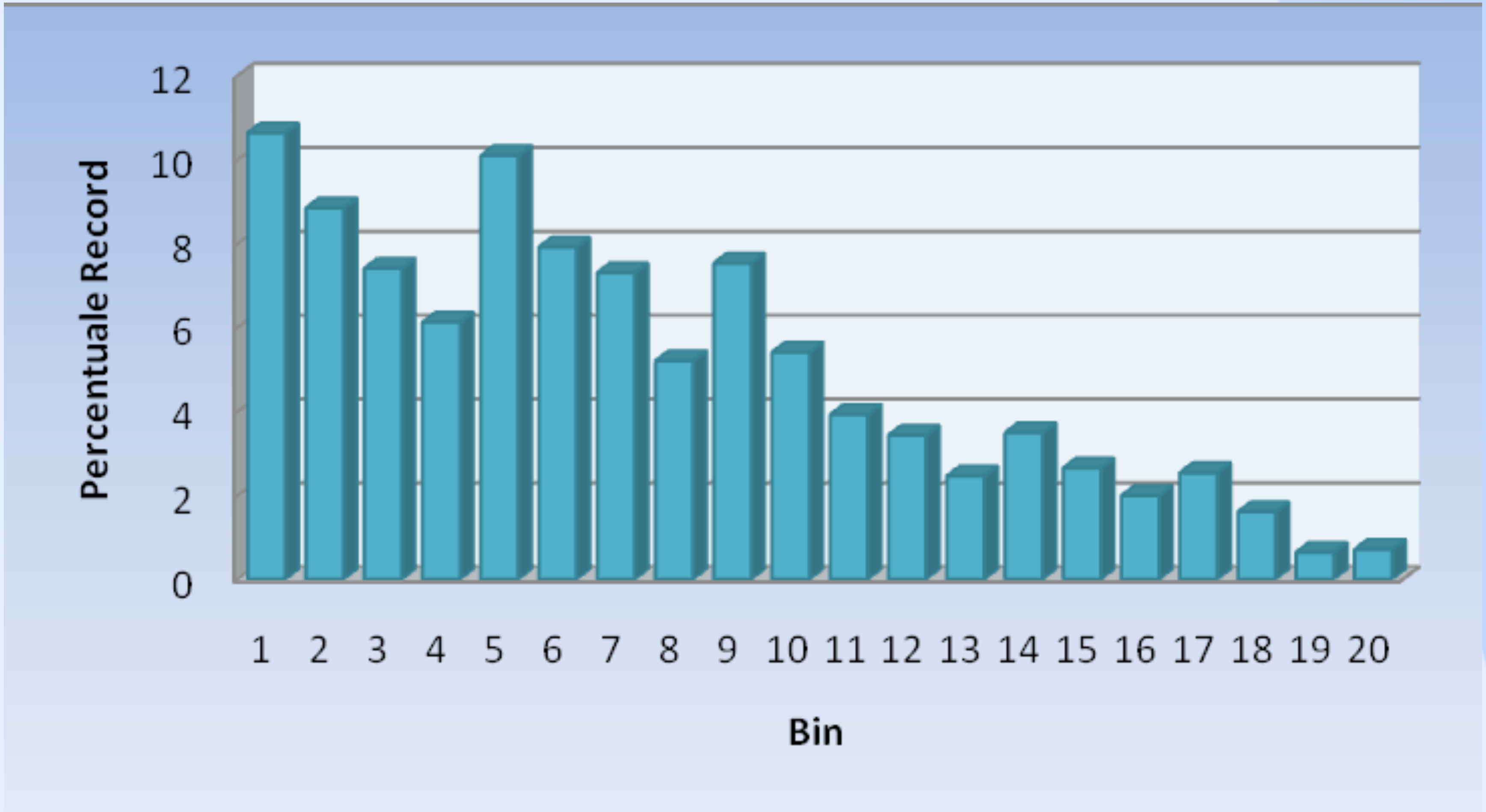
Case1: Predicting interval in volume sale

- 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

Manual discretization



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

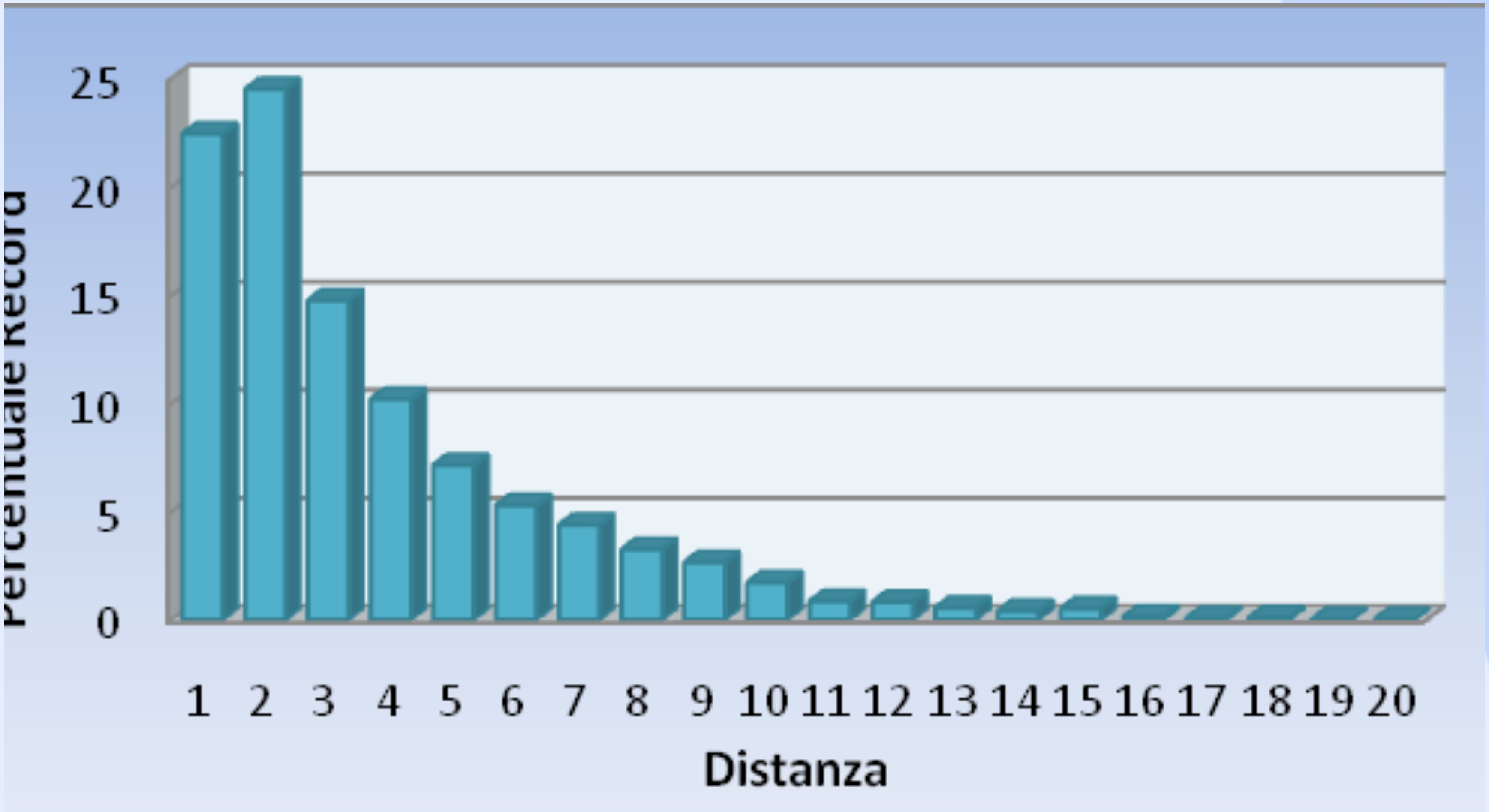
Confronto di $C-VEND_ART_PROMO_TILEN_String con VEND_ART_PROMO_TILEN_String

'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	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
_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:

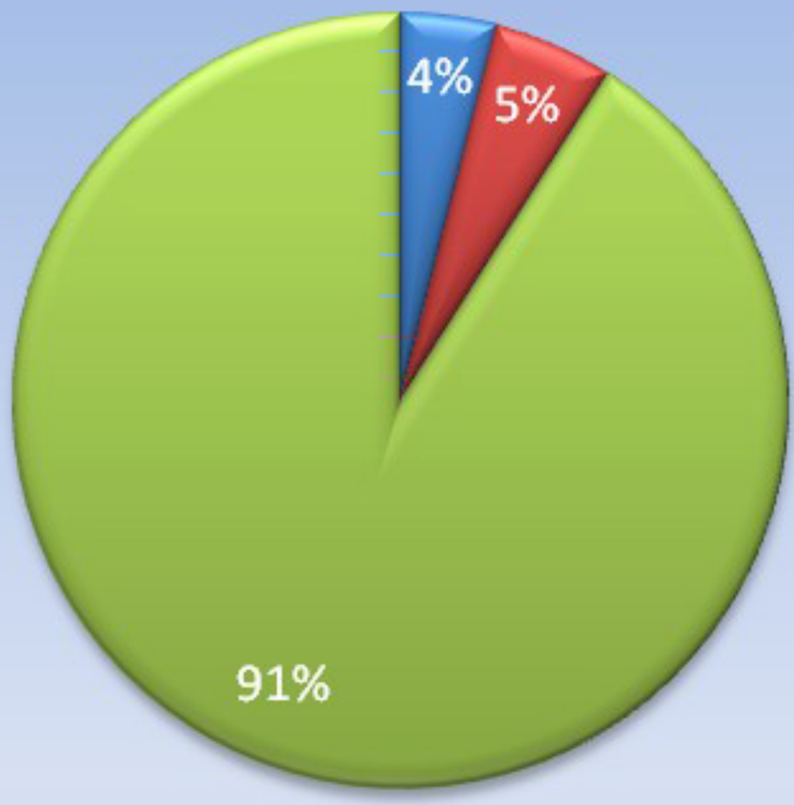
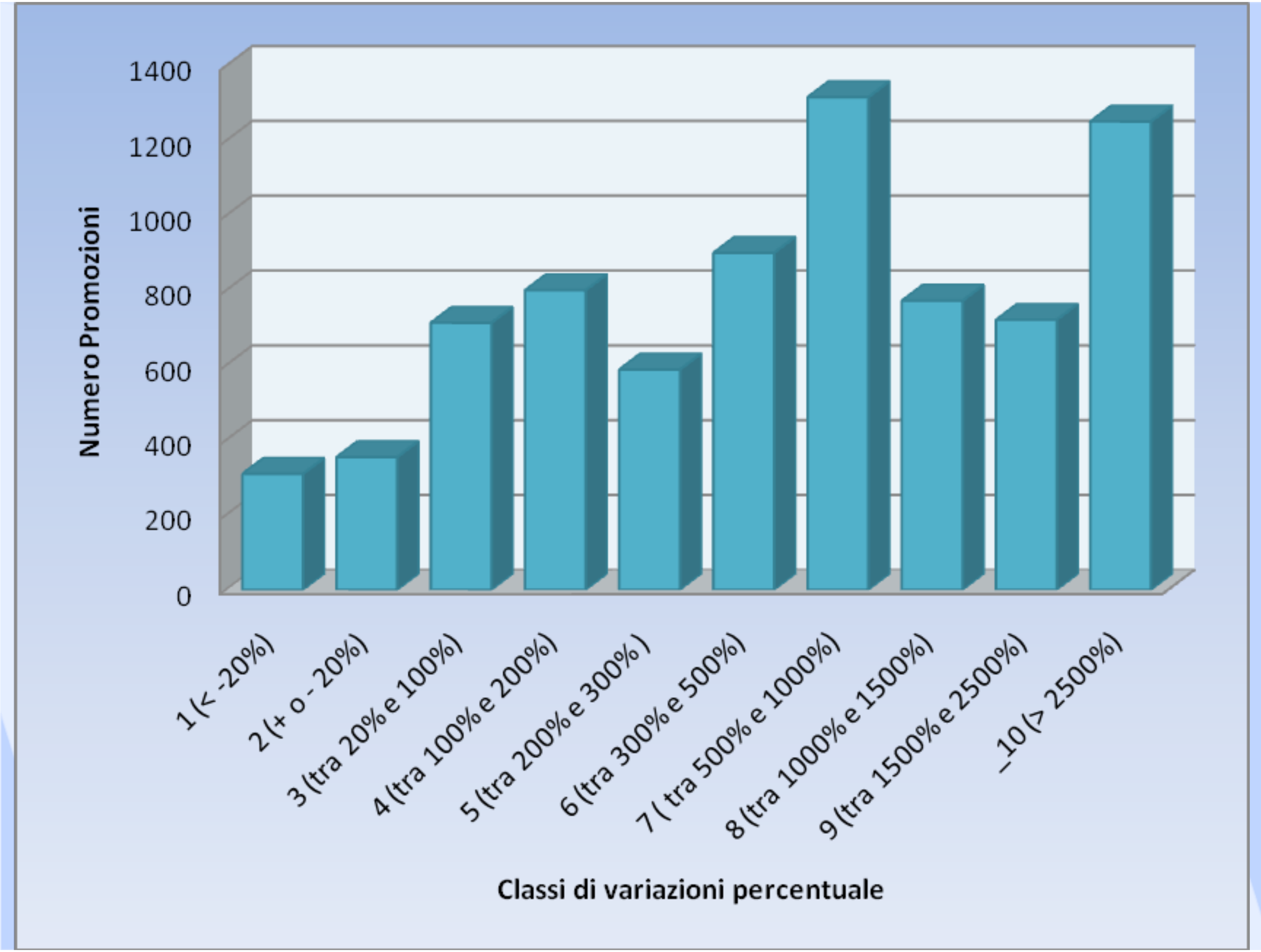


Figure 7: Percentage variation of sales under promotion

Minore (< -20%) Uguale (+ o - 20 %) Maggiore (> +20%)

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%)	



1400
1200
1000
800
600
400
200
0
Numero Promozioni
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

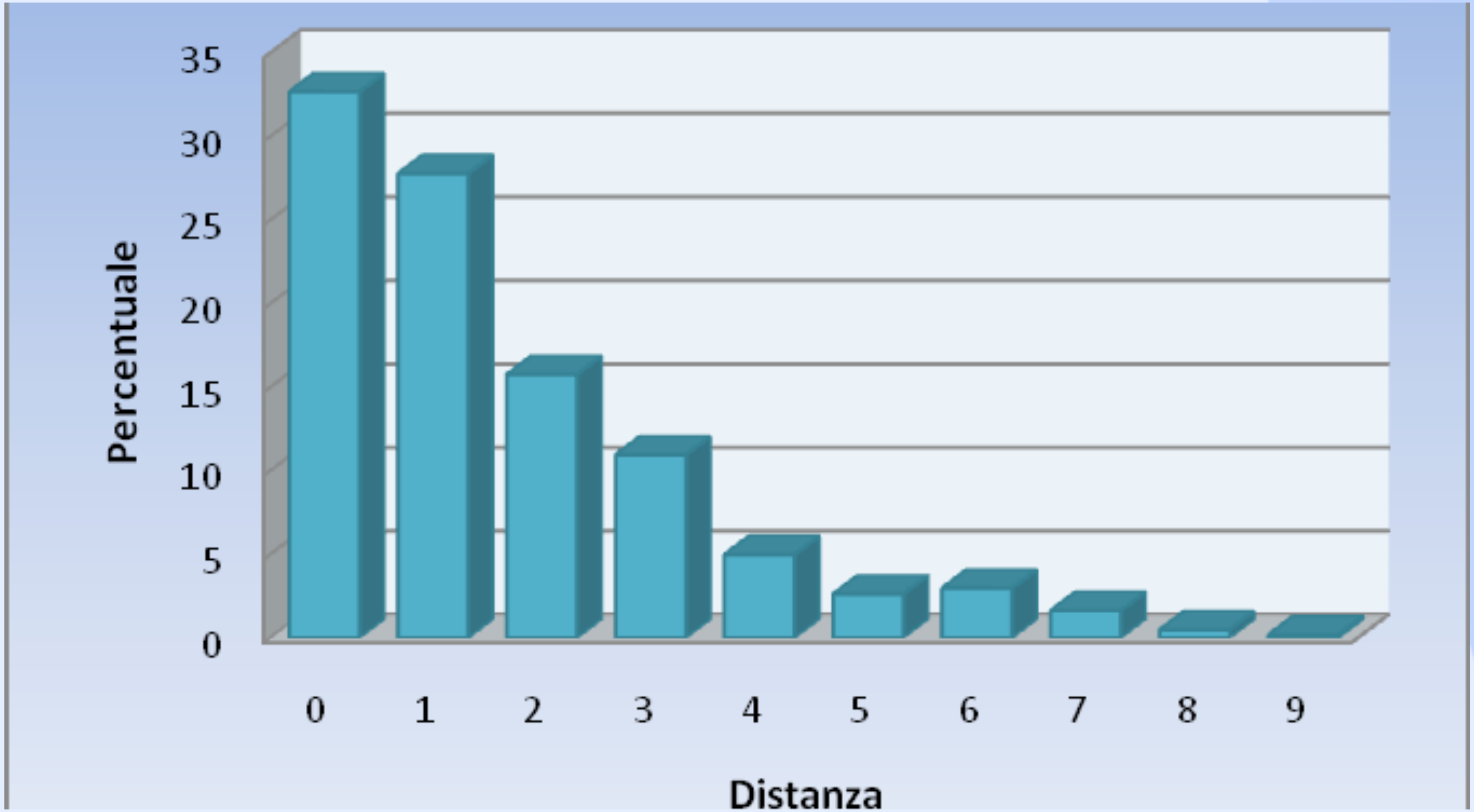
Confronto di $C-VariazionePercentualiS con VariazionePercentualiS

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

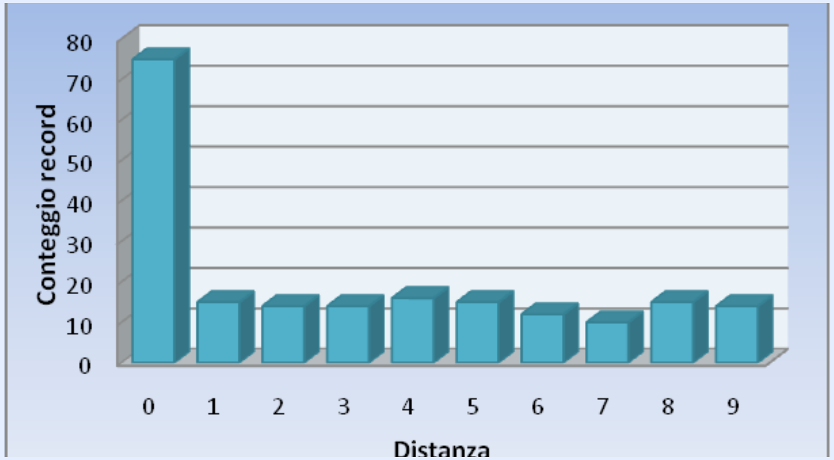
Class displacement distribution of predicted class vs. real class

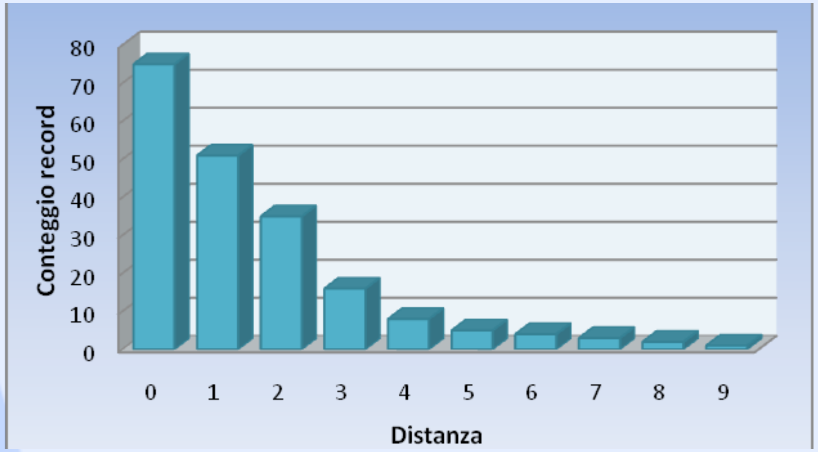


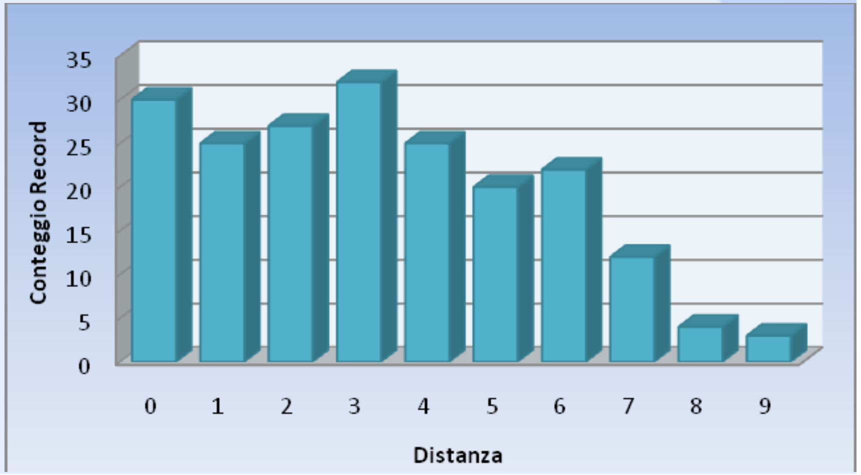
NOTE: it holds for ordinal classifiers

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

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

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

		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 16

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).

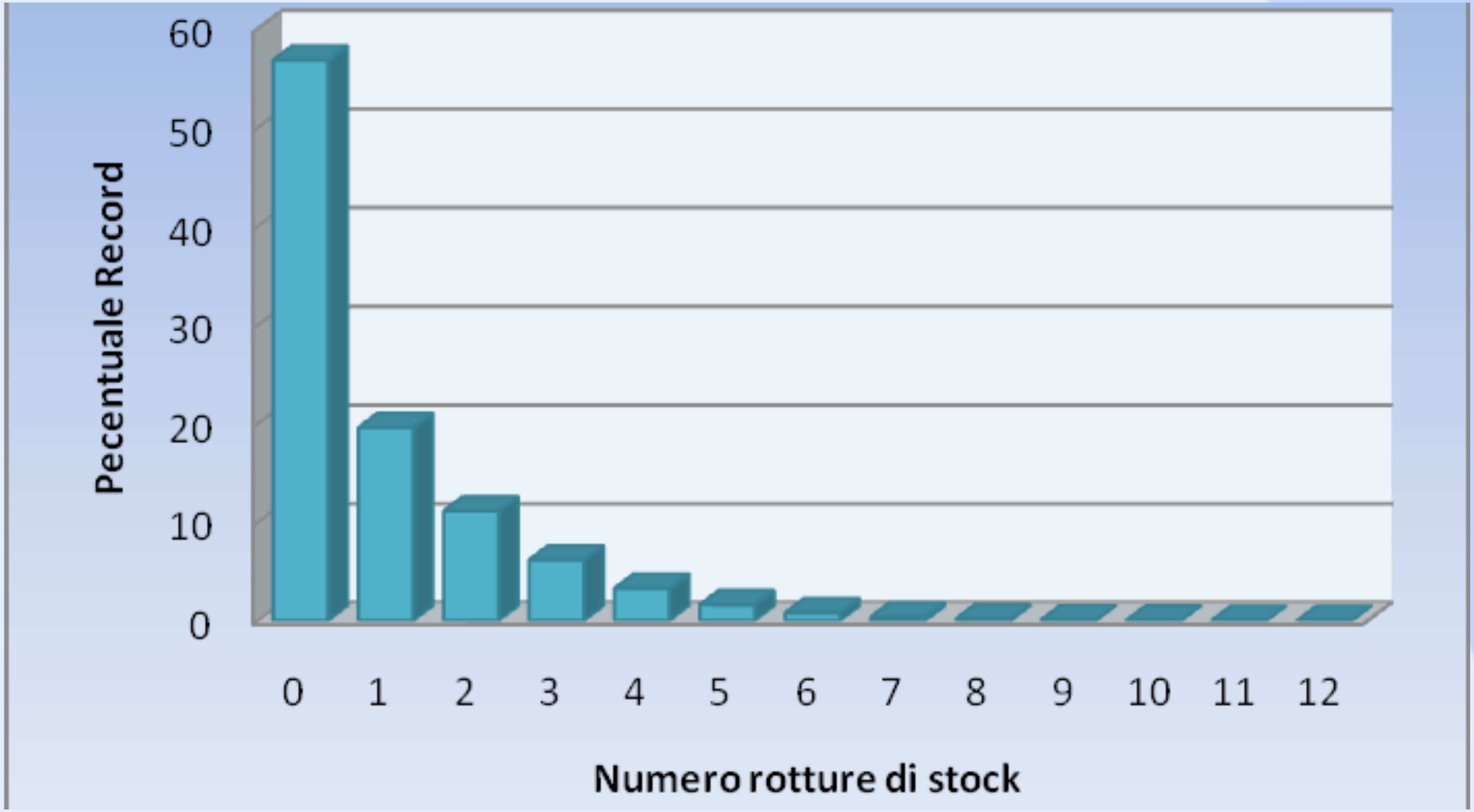
Mattina	Pranzo	Pomeriggio	Sera
2	1	1	0

- 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

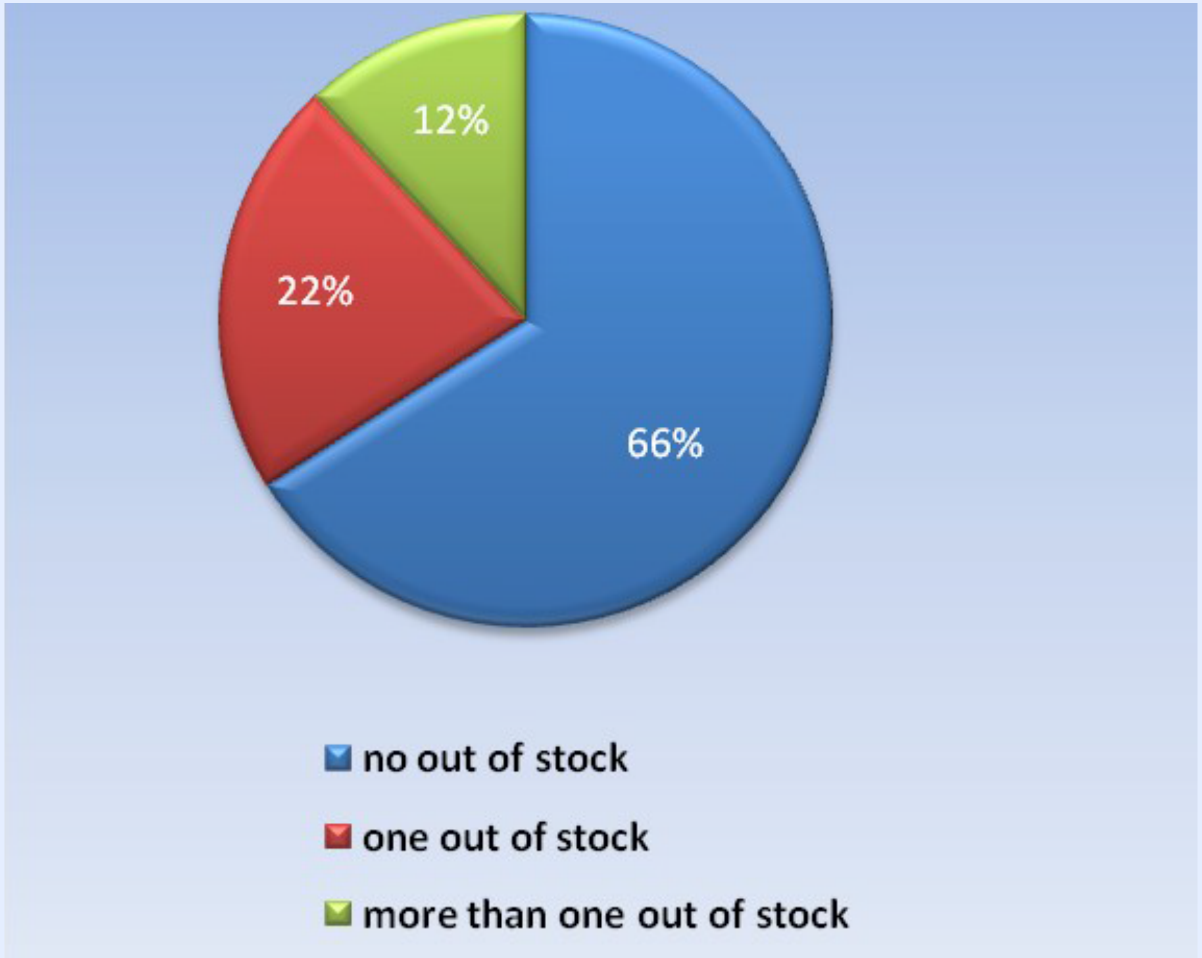


Analisi dati Super

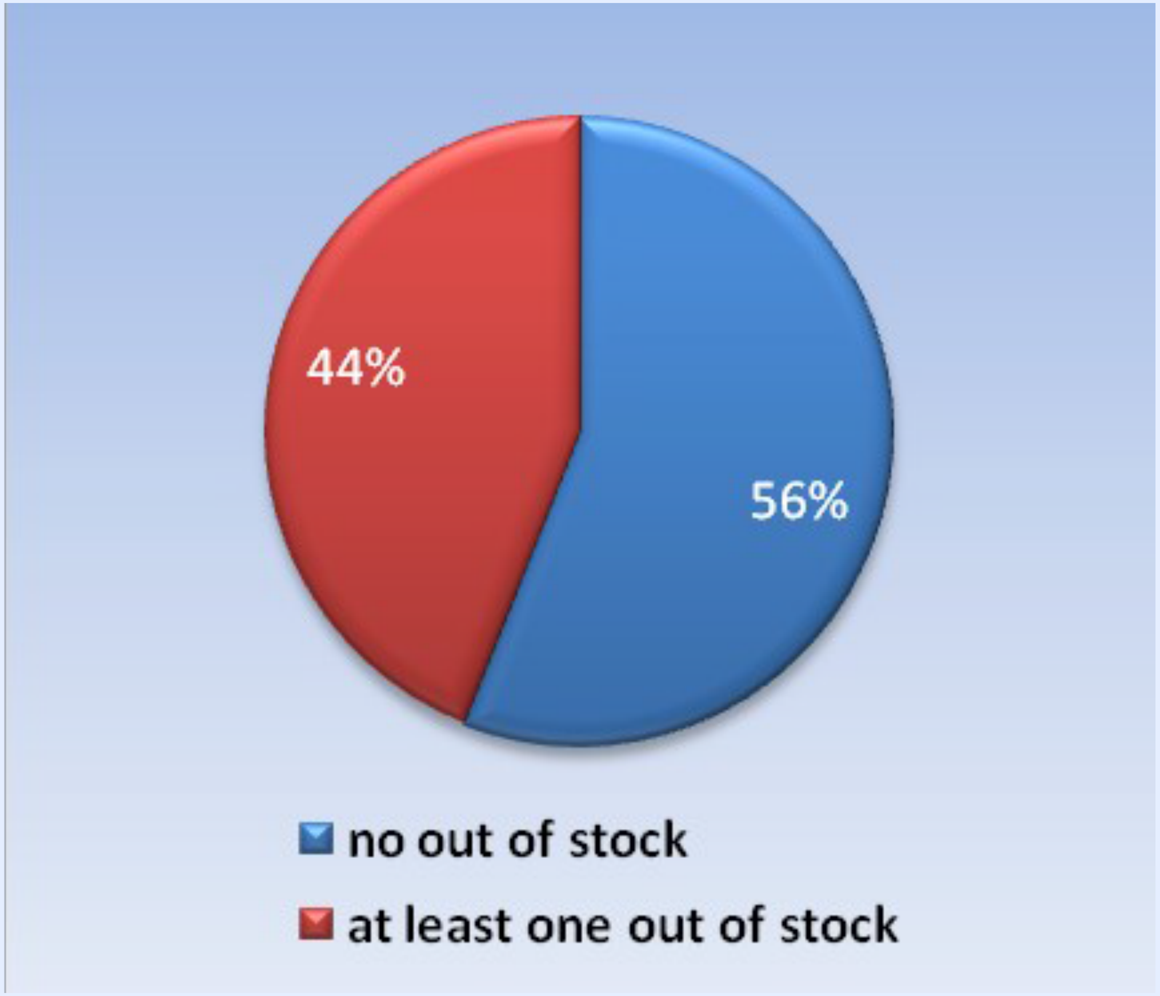
Distribuzione del numero di giorni in cui il prodotto in promozione va in rottura di stock

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%

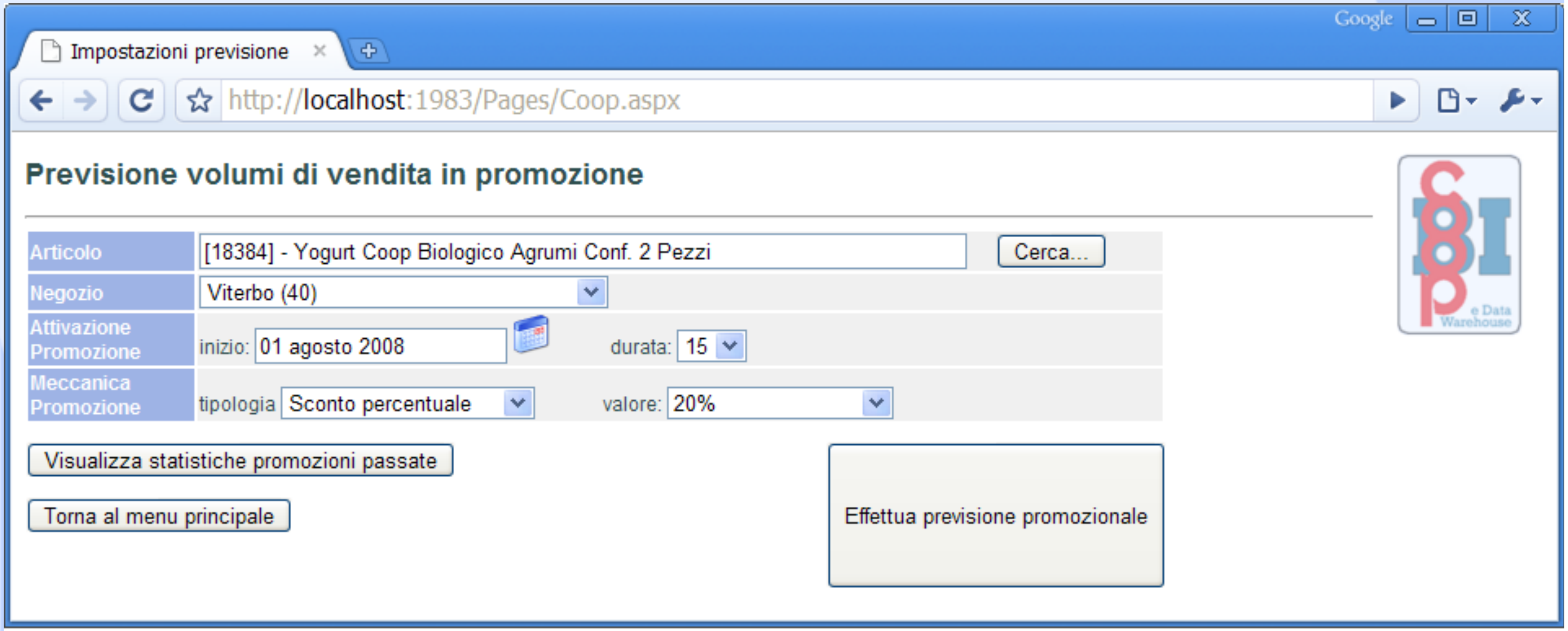
se PRES_MKT = LEADER
e VendSeg_1_0 > 479
e CATEGORIA = CAFFE'
allora 1

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



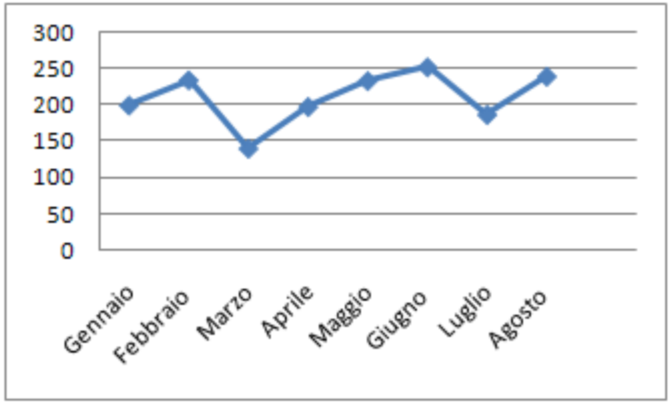
Risultati previsione

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%

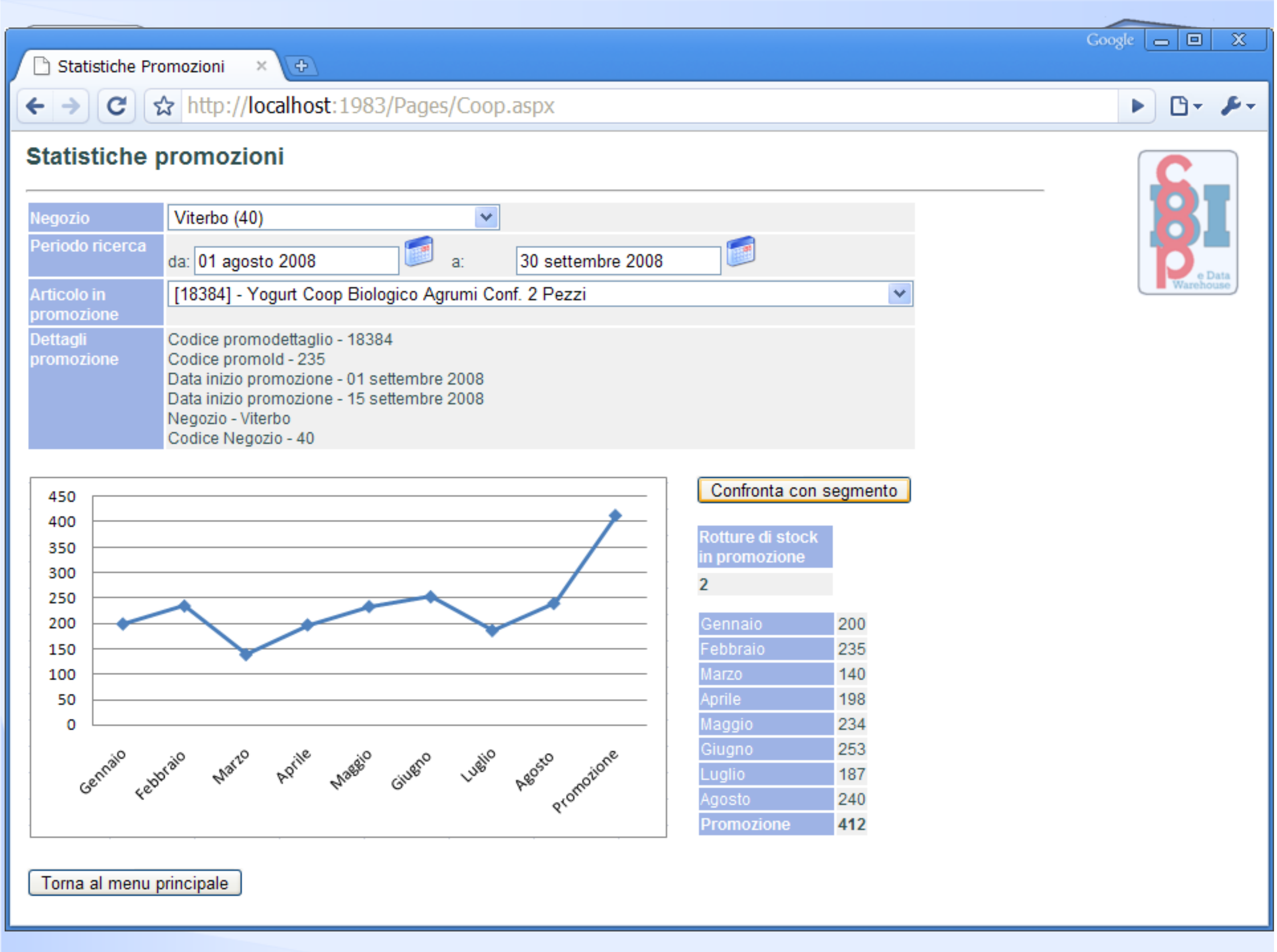
Andamento vendite

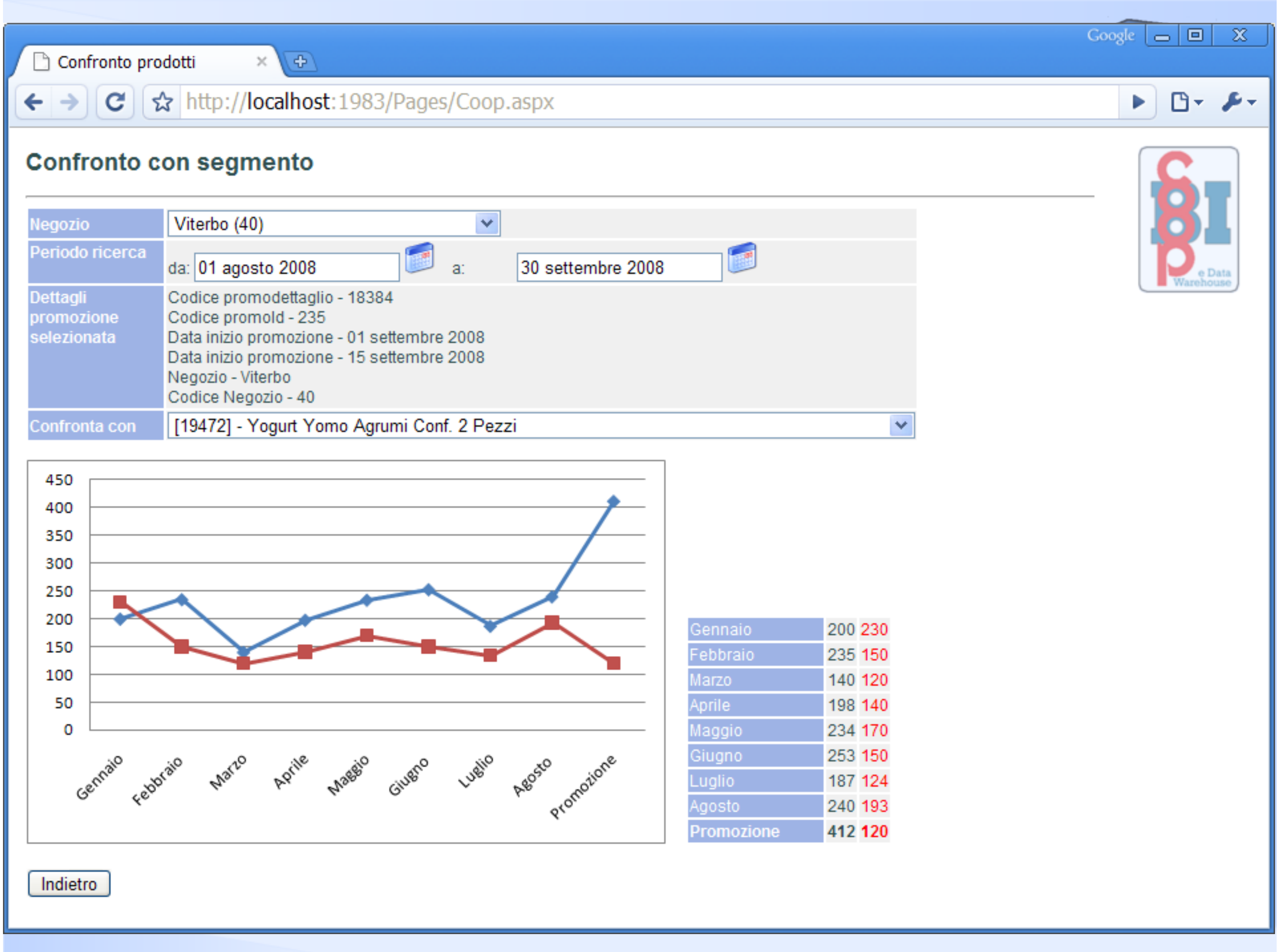


Gennaio	200
Febbraio	235
Marzo	140
Aprile	198
Maggio	234
Giugno	253
Luglio	187
Agosto	240

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 promold - 235
Data inizio promozione - 01 settembre 2008
Data inizio promozione - 15 settembre 2008
Negozio - Viterbo
Codice Negozio - 40
450
400
350
300
250
200
150
100
50
0
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Promozione
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
e Data Warehouse


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
450
400
350
300
250
200
150
100
50
0
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Promozione
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
e Data Warehouse

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