Metodi Quantitativi per Economia, Finanza e Management

Lezione n°10

Regressione Logistica: Le ipotesi del modello, la stima del modello, l'interpretazione del del modello

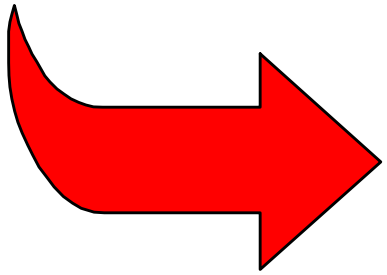
L’impostazione del problema

Redditività = ricavi - costi

- redditività var. continua
- classi di redditività (< 0 ; >= 0)

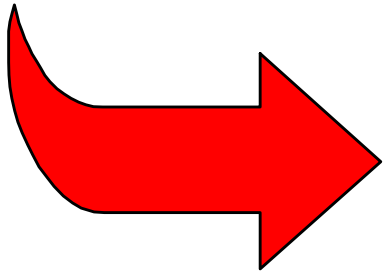
L’impostazione del problema

- Redditività var. dicotomica



Pr (Y=1 | X)

Il modello di regressione lineare è inadeguato quando la variabile risposta è dicotomica, poiché in non garantisce il rispetto del campo di variazione [0,1]



Regressione Logistica

Il modello di regressione logistica

La regressione logistica appartiene alla categoria dei Modelli Lineari Generalizzati.

Consente di prevedere una variabile discreta, che può essere intesa come l'appartenenza a un gruppo, a partire da un insieme di variabili (continue, discrete, dicotomiche).

Generalmente, la variabile dipendente, o variabile risposta, è dicotomica e rappresenta una assenza/presenza o un fallimento/successo.

Esempi:

- Modello di Churn (evento: abbandono)
- Modello di Propensity (evento: acquisto)

Il modello di regressione logistica

Le ipotesi del modello

Y	X_1	X_2	X_3	…	…	…	X_p
y_1	x_{11}	x_{12}	x_{13}	…	…	…	x_{1p}
y_2	x_{21}	x_{22}	x_{23}	…	…	…	x_{2p}
y_3	x_{31}	x_{32}	x_{33}	…	…	…	x_{3p}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
y_n	x_{n1}	x_{n2}	x_{n3}	…	…	…	x_{np}

(nx1) (nxp)

- n unità statistiche
- vettore colonna (nx1) di n misurazioni su una variabile dicotomica (Y)
- matrice (nxp) di n misurazioni su p variabili quantitative ($X_1,\ldots,X_p$)
- la singola osservazione è il vettore riga ($y_i,x_{i1},x_{i2},x_{i3},\ldots,x_{ip}$) i=1,…,n

Il modello di regressione logistica

Le ipotesi del modello

Y, la variabile dipendente dicotomica, indica la presenza o l'assenza di una particolare caratteristica.

Y assume valore 1 con probabilità π e valore 0 con probabilità 1-π.

Y si distribuisce come una variabile casuale **bernoulliana** di parametro π, che descrive l'esito di un esperimento casuale che ha probabilità di risultare in "successo" con probabilità pari a π.

$$Y \sim Bernoulli(\pi)$$

$$Pr(Y) = \pi^{Y}(1-\pi)^{(1-Y)}$$

Il modello di regressione logistica

Le ipotesi del modello

Il modello di regressione lineare è inadeguato quando la variabile risposta è dicotomica, poiché:

1. Non garantisce il rispetto del campo di variazione [0,1]
2. La componente erratica può assumere solo due valori, non può avere una distribuzione normale.
3. La componente erratica viola l'ipotesi di omoschedasticità, la varianza dipende dal particolare valore di Xi

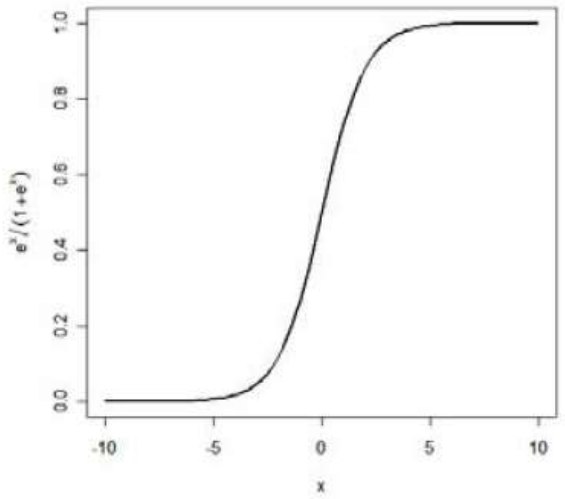
Il modello di regressione logistica

Le ipotesi del modello

Nell'ambito della regressione logistica si ipotizza che π: Pr(Y=1 I X) sia definito dalla seguente forma funzionale:

$$\Pr(Y_i = 1 \mid \underline{X}_i) = \pi_i = \frac{\exp(\underline{X}_i^T \underline{\beta})}{1 + \exp(\underline{X}_i^T \underline{\beta})}$$

Funzione Logistica



Il modello di regressione logistica

Le ipotesi del modello

Il modello logistico gode di alcune importanti proprietà:

1. Rispetta il vincolo che il valore stimato di: $\Pr(Y_i = 1 \mid \underline{X}_i) = \pi_i$ sia compreso nell'intervallo [0,1];
1. La forma ad «esse» della funzione logistica garantisce un avvicinamento graduale ai valori estremi 0 e 1;
2. La funzione logit di: π_i è esprimibile come combinazione lineare delle variabili indipendenti X_1,.., X_k:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

Il modello di regressione logistica

Le ipotesi del modello

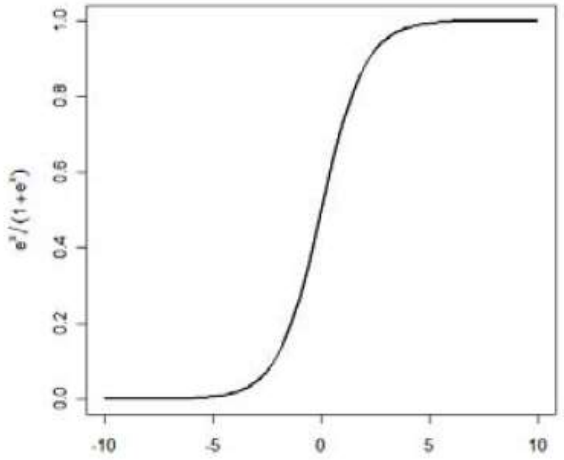
Posto:

$$\Pr(Y_i = 1 \mid \underline{X}_i) = \pi_i$$

Si dimostra che

$$\pi_i = \frac{\exp(\underline{X}_i^T \underline{\beta})}{1 + \exp(\underline{X}_i^T \underline{\beta})}$$

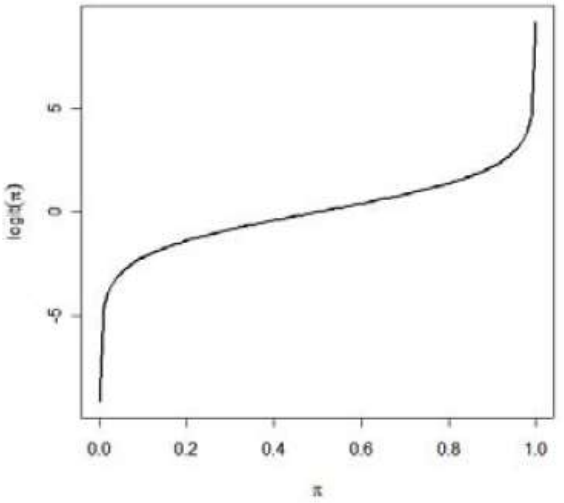
LOGISTICA



equivale a

$$\log\left(\frac{\pi_i}{1 - \pi_{i\,i}}\right) = \underline{X}_i^T \underline{\beta}$$

LOGIT



Il modello di regressione logistica

La stima del modello

Analogamente al modello di regressione lineare, la relazione tra la variabile dipendente e le indipendneti è nota a meno del valore dei parametri:

$$\mathrm{logit}\ (\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

E' necessario un metodo che permetta di ottenere delle "buone" stime dei parametri sulla base delle osservazioni campionarie disponibili.

Il modello di regressione logistica

La stima del modello

Si dimostra che gli stimatori ottenuti mediante il metodo dei Minimi Quadrati non godono delle proprietà ottimali garantite nel caso della regressione lineare.

Viene utilizzato il metodo più generale della Massima Verosimiglianza, che si basa sulla massimizzazione della probabilità di osservare l'insieme di dati campionari disponibili in funzione di β.

- Le equazioni di verosimiglianza non sono lineari nei parametri e non ammettono (salvo casi particolari) soluzione esplicita.

- E' necessario ricorrere a metodi numerici iterativi per approssimare la soluzione (Algoritmo di Newton-Raphson o di Scoring's Fisher)

Il modello di regressione logistica

La stima del modello

Gli stimatori di massima verosimiglianza godono di proprietà ottimali in presenza di campioni numericamente grandi:

- asintoticamente corretti (le stime sono non distorte, si avvicinano al valore vero)
- asintoticamente efficienti (con standard error delle stime sono piccoli almeno come quelli di ogni altro metodo di stima)
- asintoticamente normali (è possibile usare la distribuzione normale o chi quadro per calcolare gli intervalli di confidenza)

Il modello di regressione logistica

La stima del modello

Test per valutare la significatività congiunta dei coefficienti ("Testing Global Null Hypothesis: BETA=0")

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

- Likelihood Ratio
- Score
- Wald

Queste statistiche hanno distribuzione Chi-quadro con *n* gradi di libertà dove n corrisponde al numero di coefficienti stimati delle variabili indipendenti.

Se il p-value piccolo (rifiuto H_0), quindi il modello ha buona capacità esplicativa.

N.B. Equivalenti al Test F della regressione lineare

Il modello di regressione logistica

La stima del modello

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2192.4978	7	<.0001
Score	1399.0552	7	<.0001
Wald	876.2357	7	<.0001

Il modello di regressione logistica

La stima del modello

Indicatori sintetici di bontà del Modello

- Likelihhod ratio test → OK p-value con valori piccoli
  → E' l'analogo del test F nella reg. lin.

- Wald Chi_square test → OK p-value con valori piccoli
  → E' l'analogo del test t nella reg. lin.

- Akaike Criterion → OK valori piccoli

- Schwart Criterion → OK valori piccoli

Il modello di regressione logistica

La valutazione del modello

Si definiscono PAIRS il numero di coppie di osservazioni (i,h con i≠h) che in un caso hanno Y=1 e nell'altro Y=0.

La coppia di osservazioni (i,h con i≠h) per la quale $Y_i = 1$ e $Y_h = 0$ è:

- concordante se 
- tied se 
- discordante se 

Tanto maggiore è il numero dei CONCORDANT (e quindi tanto minore è il numero dei DISCORDANT), tanto più il modello rappresenterà adeguatamente il fenomeno indagato.

Il modello di regressione logistica

La valutazione del modello

Le statistiche seguenti sono calcolate sulla base del numero di coppie CONCORDANT, DISCORDANT e TIED.

$$Tau-a=\frac{C-D}{N}$$

$$Gamma=\frac{C-D}{C+D}$$

$$Somer'sD=\frac{C-D}{C+D+T}$$

$$c=0.5*(1+Somer'sD)$$

Association of Predicted Probabilities and Observed			
Percent Concordant	89.6	Somers' D	0.796
Percent Discordant	10.0	Gamma	0.800
Percent Tied	0.4	Tau-a	0.146
Pairs	643691936	c	0.898

Indicando con:

- C è il numero di coppie concordanti,
- D il numero di coppie discordanti,
- T il numero di ties
- N il numero totale di coppie

Tanto più questi indicatori sono elevati, tanto più il modello è “corretto”. Queste misure variano tra 0 ed 1. Valori più grandi corrispondono a più forte associazione tra valori predetti e valori osservati.

Il modello di regressione logistica

La stima del modello

Test per valutare la significatività dei singoli coefficienti

$$H_0 : \beta_j = 0$$

– Wald Chi-square: il quadrato del rapporto tra stima e standard error

Il coefficiente è significativamente diverso da zero se il corrispondente p-value è piccolo (ossia, rifiuto l'ipotesi di coefficiente nullo) → il regressore a cui il coefficiente è associato è rilevante per la spiegazione del fenomeno

N.B. Equivalente al Test t della regressione lineare

Il modello di regressione logistica

La stima del modello

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**Standardized Estimate**
Intercept	1	-1.2530	0.1147	119.3602	<.0001	
PAG_ORD	1	0.000070	5.295E-6	175.1845	<.0001	1.1035
TOT_ORD	1	0.5151	0.0432	142.1610	<.0001	0.6494
PAG_MES	1	0.000120	8.608E-6	194.9225	<.0001	0.6074
SUD	1	-0.8965	0.1038	74.6650	<.0001	-0.2381
CEN	1	-0.2745	0.1294	4.5039	0.0338	-0.0571
SESSO	1	0.2729	0.1005	7.3780	0.0066	0.0695
LISTA	1	-0.00293	0.0553	0.0028	0.9577	-0.00134

Il modello di regressione logistica

La stima del modello

In presenza di regressori quantitativi, i coefficienti standardizzati possono essere utili per valutare l'importanza relativa delle variabili, capire quali sono quelle che pesano di più nel modello.

Relativamente all'esempio sopra riportato:

- la variabile maggiormente influente nel modello è PAG_ORD (Standardized estimated: 1.1035),
- segue TOT_ORD (Standardized estimated: 0.6494),
- segue PAG_MES (Standardized estimated: 0.6074), etc.

Il modello di regressione logistica

La stima del modello

Analogamente al modello di regressione lineare, la relazione tra la variabile dipendente e le indipendeneti è nota a meno del valore dei parametri:

$$\mathrm{logit}\ (\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

Ai fini della formulazione di un modello di tipo lineare è stato necessario:

1.trasformare le probabilità in odds π/(1- π) per rimuovere il limite superiore (Sup=1)
2.applicare la funzione logaritmica agli odds per rimuovere il limite inferiore (Inf=0)

Il modello di regressione logistica

L'interpretazione del modello

Nelle scommesse si dice che un certo evento è dato 5 a 2 che vuol dire 5/2 è l'odds: il rapporto tra il numero atteso di volte che un evento accada e il numero atteso di volte che un evento non accada.

C'è una semplice relazione tra l'odds e la probabilità:

$$O = \frac{\pi}{1-\pi}$$

$$\pi = \frac{O}{1+O}$$

dove π è la probabilità dell'evento e O è l'odds.

Il modello di regressione logistica

L'interpretazione del modello

Un odds inferiore a 1 corrisponde a una probabilità inferiore a 0.5. Il limite inferiore è 0 come per la probabilità ma non ha limiti superiori.

Probabilità dell'evento	odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Il modello di regressione logistica

L'interpretazione del modello

Nella regressione logistica un coefficiente di 0.2 ci dice che il logit di Y (il log dell'odds) aumenta di 0.2 in corrispondenza al possesso dell'attributo X. Ma cosa significa un aumento di 0.2 del logit?

Dato che la relazione tra probabilità e regressore non è lineare, risulta più facile parlare in termini di odds. I coefficienti stimati, a parte per il segno, non sono interpretabili, l'odds ratio (l'exp del coeff.) sì.

Esempio (Modello di Churn):

Sesso	Estimate	CHURN RATE	Odds Ratio Estimate
0 (femmina)		1.98%	
1(maschio)	0.2103	2.52%	1.23
TOTAL		2.24%	

I maschi hanno un churn rate più alto delle femmine.

L'odds previsto dell'abbandono per i maschi è 1.234 volte quello delle femmine (è 23% più alto).

Il modello di regressione logistica

L'interpretazione del modello

Variabile indipendente (es. M=1; F=0)

Variabile risposta (SI=1; NO=0)	x = 1	x = 0
y = 1	π(1)	π(0)
y = 0	1- π(1)	1- π(0)

ODDS RATIO

$$\psi = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{ODDS_1}{ODDS_0}$$

E' una misura di associazione; approssima il Rischio Relativo, ossia quanto più probabile è per la variabile risposta essere presente tra i soggetti con x=1 che tra quelli con x diverso da 0.

Il modello di regressione logistica

L'interpretazione del modello

Nel caso di variabili continue l'interpretazione del parametro è analoga.

Il coefficiente esprime il cambiamento di logit in corrispondenza di un aumento unitario di X.

$$\beta_1 = logit(Pr(Y = 1 \mid X = x + 1) - logit(Pr(Y = 1 \mid X = x)$$

Il modello di regressione logistica

L'interpretazione del modello

Odds Ratio Estimates	
Effect	**Point Estimate**
PAG_ORD	1.000
TOT_ORD	1.674
PAG_MES	1.000
SUD	0.408
CEN	0.760
SESSO	1.314
LISTA	0.997

Il modello di regressione logistica

La valutazione del modello

Analogamente a quanto visto per la regressione lineare, anche per la logistica il problema della multicollinearità può causa effetti indesiderati sulla stabilità delle stime.

I metodi di gestione della problematica sono analoghi a quelli trattati nel modello di regressione lineare.

Il modello di regressione logistica

La valutazione del modello

Analogamente alla regressione lineare è possibile avvalersi di vari metodi di selezione automatica delle variabili.

Anche in questo caso gli algoritmi operano secondo le logiche di:

- Stepwise
- Forward
- Backward

Il modello di regressione logistica

L'utilizzo modello

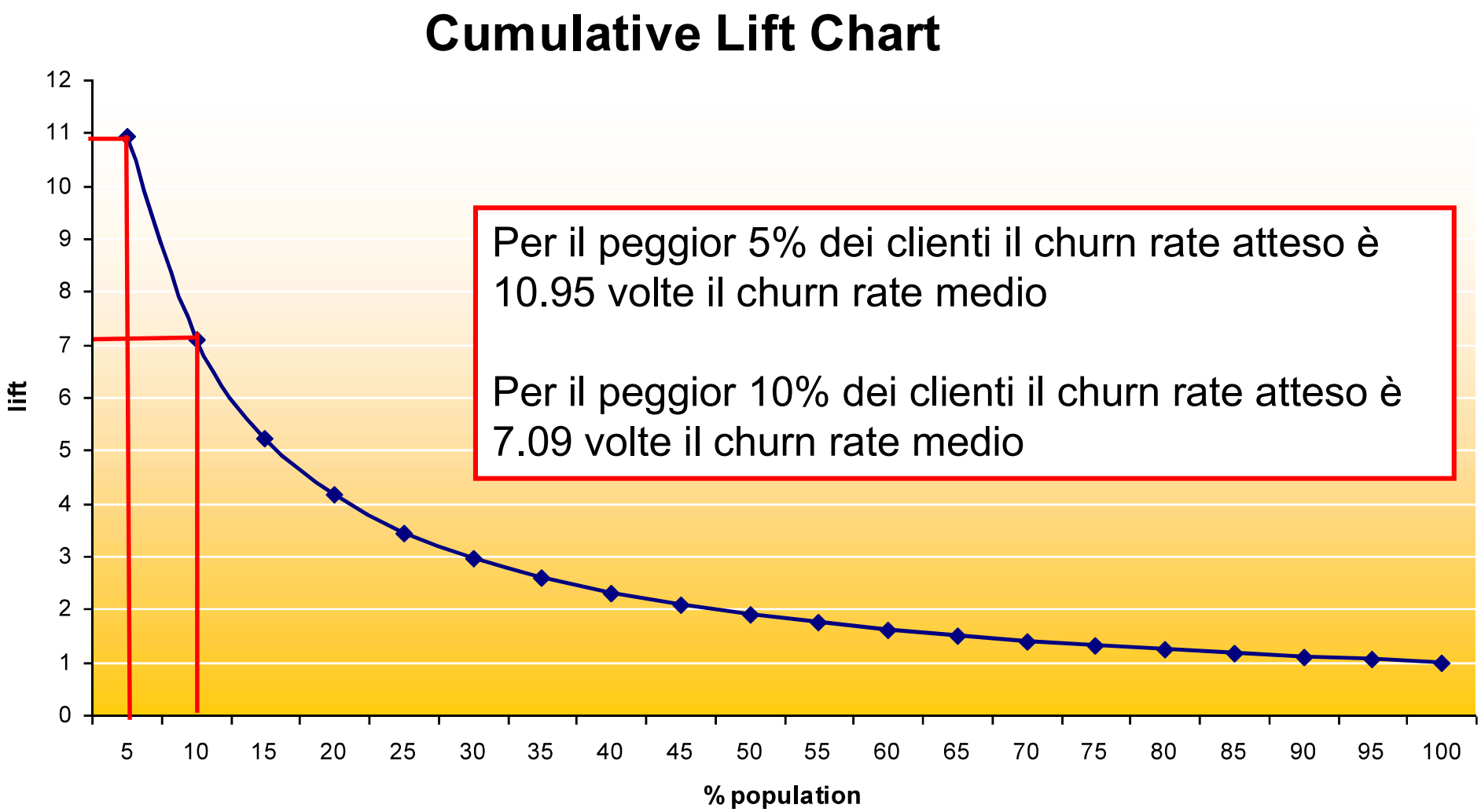
Tutte le osservazioni sono suddivise in ventili in base alla probabilità prevista di risposta.

decili	target	popolazione	target cumulato	popolazione cumulata	redemption	redemption cumulata	lift	lift cumulata	%catturati	%catturati cumulata
5	1028	4191	1028	4191	24.53%	24.53%	10.95	10.95	54.76%	54.76%
10	303	4191	1331	8382	7.22%	15.88%	3.23	7.09	16.13%	70.88%
15	144	4191	1475	12573	3.44%	11.73%	1.54	5.24	7.68%	78.57%
20	85	4191	1560	16764	2.02%	9.30%	0.90	4.15	4.51%	83.08%
25	62	4191	1622	20955	1.48%	7.74%	0.66	3.46	3.31%	86.39%
30	50	4191	1672	25146	1.18%	6.65%	0.53	2.97	2.64%	89.03%
35	35	4191	1707	29337	0.84%	5.82%	0.38	2.60	1.88%	90.91%
40	29	4191	1736	33528	0.69%	5.18%	0.31	2.31	1.54%	92.46%
45	25	4191	1761	37719	0.60%	4.67%	0.27	2.08	1.33%	93.79%
50	23	4191	1784	41910	0.55%	4.26%	0.24	1.90	1.22%	95.01%
55	17	4191	1801	46101	0.41%	3.91%	0.18	1.74	0.92%	95.93%
60	16	4191	1817	50292	0.37%	3.61%	0.17	1.61	0.83%	96.76%
65	13	4191	1830	54483	0.31%	3.36%	0.14	1.50	0.69%	97.46%
70	11	4191	1840	58674	0.25%	3.14%	0.11	1.40	0.57%	98.02%
75	6	4191	1847	62865	0.15%	2.94%	0.07	1.31	0.33%	98.35%
80	6	4191	1853	67056	0.15%	2.76%	0.07	1.23	0.33%	98.68%
85	6	4191	1859	71247	0.15%	2.61%	0.07	1.16	0.33%	99.01%
90	6	4191	1865	75438	0.15%	2.47%	0.07	1.10	0.33%	99.34%
95	6	4191	1871	79629	0.15%	2.35%	0.07	1.05	0.33%	99.67%
100	6	4191	1878	83820	0.15%	2.24%	0.07	1.00	0.33%	100.00%

Il Lift value è ottenuto come rapporto tra la percentuale di positivi contenuti nel ventile e la percentuale di positivi contenuti nella popolazione totale.

Il modello di regressione logistica

L'utilizzo modello



Il modello di regressione logistica

L'utilizzo modello