****

***Frontespizio***

**Università degli studi di Roma**

**Unitelma Sapienza**

**Dipartimento di Scienze giuridiche ed economiche**

**Corso di Laurea TRIENNALE in SCIENZE DELL’ECONOMIA E DELLA GESTIONE AZIENDALE**

**Tesi di Laurea in**

**[Statistica]**

**Classificazione binaria tramite la regressione logistica: caso di studio sul *cross-selling* bancario**

Relatore:

Ch.mo Prof. **Pasquale Sarnacchiaro**

Candidato:

Americo Costantini

Anno Accademico 2016/2017

INTRODUZIONE

# Obiettivo dell’elaborato

Lo scopo di questo elaborato è la creazione di un modello predittivo volto alla classificazione delle unità di analisi secondo due possibili uscite, la presenza o l’assenza di un attributo. A tal fine si è deciso di utilizzare la regressione logistica. Un approccio scientificamente ottimale non contempla la scelta a priori del modello, bensì l’utilizzo di più tecniche differenti e poi la selezione (o la combinazione) del modello più performante secondo le metriche di *performance* prescelte[[1]](#footnote-1). Nel nostro caso la scelta è stata fatta a priori perché un approccio come quello su descritto va oltre i limiti delle ambizioni dell’elaborato che si sta redigendo, inducendo quindi chi scrive a individuare a priori un singolo metodo con il quale svolgere l’analisi dei dati. Data questa limitazione, la regressione logistica è sembrata la scelta più idonea in quanto tra i metodi di classificazione binaria è quello che più permette un approfondimento di concetti di statistica inferenziale, facendo così scampare il pericolo di una eccessiva robotizzazione (leggi: scarsa interpretabilità) del modello predittivo.

La struttura dello scritto consiste di due capitoli iniziali di presentazione degli strumenti statistici poi utilizzati nel capitolo terzo, quello centrale, di svolgimento di una analisi dei dati originale. Alla luce di ciò si è limitato l’approfondimento di dimostrazioni matematiche delle formule presenti a quanto era strettamente necessario per la comprensione dei concetti al fine del loro utilizzo nell’analisi stessa.

Lo scopo dell’analisi è quello di stimare un modello predittivo sui dati raccolti da una banca durante una campagna promozionale per la sottoscrizione di un nuovo contratto rivolta a un campione dei suoi clienti. L’obiettivo sarà individuare quelle caratteristiche che permettano di definire in anticipo, su tutta la popolazione dei clienti, quali saranno favorevoli alla sottoscrizione del nuovo contratto e discriminarli da quelli contrari, in modo da canalizzare in maniera efficiente lo sforzo dei venditori della banca[[2]](#footnote-2).

# Notazione matematica[[3]](#footnote-3)

Si denoterà con il numero di osservazioni del dataset, vale a dire la dimensione del campione. Con il numero di variabili indipendenti presenti nel dataset, quindi il numero di parametri. Con lettere maiuscole, , verranno indicate genericamente le variabili casuali. Data una matrice che in riga ha le misurazioni per ogni i-esima osservazione e in colonna le misurazioni del j-esimo parametro per tutte le unità osservate, allora si denoterà con la misurazione del j-esimo parametro relativa alla i-esima osservazione, con il vettore contenente le misurazioni per la i-esima unità e con il vettore delle misurazioni del j-esimo parametro. Inoltre con verrà indicata la realizzazione i-esima della variabile casuale . Con si indicherà la probabilità che un evento accada, mentre con si rappresenterà una funzione dipendente da alcuni parametri la cui uscita è la probabilità che l’attributo sia presente. Ad esempio, significherà: la probabilità che la variabile casuale abbia una realizzazione pari a per la i-esima osservazione dati certi valori delle variabili indipendenti, mentre sarà la funzione che, dati certi valori delle variabili indipendenti, avrà come uscita la probabilità che. Con si denoterà il vero coefficiente del j-esimo parametro, ma soprattutto con il relativo coefficiente stimato. Anche starà per la stima del valore della variabile di risposta per la i-esima osservazione. In generale il simbolo del “cappello” (*hat*, in inglese) starà a significare che quel valore è stimato e non osservato. Con si indicherà l’insieme dei coefficienti di un modello.

# Terminologia

Quanto alla terminologia, la variabile dipendente potrà alternativamente essere chiamata variabile di uscita, variabile di risposta, variabile target, mentre quelle indipendenti verranno indistintamente denotate come parametri, regressori, variabili esplicative, variabili di input. La presenza di un attributo dicotomico verrà chiamata “evento”, e l’assenza “non-evento”, oppure “caso” e “non-caso”. La probabilità stimata che una unità sia classificabile come evento potrà essere definita, oltre che con la locuzione estesa, anche come semplice “probabilità stimata”.

CAPITOLO PRIMO

LA REGRESSIONE LOGISTICA

# La funzione di ipotesi nella regressione logistica

### Perché non la regressione lineare

L’intento della classificazione binaria è predire, in base ai valori assunti da variabili indipendenti, l’uscita di una variabile dicotomica che può assumere soltanto due valori:

*Y*=

Un tipico approccio per la predizione è quello di stimare la probabilità che assuma il valore in base ai valori del vettore delle variabili indipendenti, . Come noto, la probabilità che un evento accada può assumere soltanto valori tra e *[[4]](#footnote-4)*, quindi a questo scopo la regressione lineare è inadeguata, in quanto modello volto a predire il valore di una variabile dipendente quantitativa che non può essere sottoposta a una restrizione di intervallo di questo tipo[[5]](#footnote-5).

Inoltre una variabile casuale dicotomica ha una distribuzione di *Bernoulli*, che prevede una varianza pari a . Tale varianza dipende da un valore non costante (perché l’assunto di cui sopra è che , quindi che la probabilità che l’attributo sia presente dipenda dai valori delle variabili indipendenti), e quindi non è costante. Se ne conclude quindi ulteriormente l’inadeguatezza dell’utilizzo della regressione lineare, che regge sull’assunto della omoschedasticità e quindi della omogeneità della varianza per tutte le combinazioni di valori dei regressori[[6]](#footnote-6).

Insomma, l’obiettivo è quello di individuare una funzione che a partire dal generico vettore dei valori delle variabili indipendenti ***X*** fornisca una probabilità che :

[[7]](#footnote-7)

e quindi:

che non rappresenta altro che la media condizionata a di .

### La funzione logistica

Una funzione adatta allo scopo di legare ***X*** a è la cosiddetta funzione logistica[[8]](#footnote-8):

dove è anch’essa una funzione ancora da specificare. Una funzione del genere, detta “funzione logistica”, assume la seguente forma:



La funzione logistica passa sempre per il punto di coordinate , dove interseca l’asse delle ordinate. Tale forma assunta dalla funzione logistica si chiama sigmoide. Una delle più importanti proprietà di questa funzione è che qualunque valore assuma , il valore di sarà sempre compreso nell’intervallo . Se tende a allora il valore di uscita tenderà a , in caso contrario a . Questo è esattamente quello che ci aspettiamo da una funzione la cui uscita deve essere una probabilità. può quindi essere visto come una sintesi dei vari fattori che determinano la probabilità dell’accadimento di un evento[[9]](#footnote-9).

### La trasformazione in relazione lineare

La funzione logistica può assumere una forma molto più interpretabile se passiamo dalla considerazione della probabilità a quella dell’ODDS:

ODDS =

L’ODDS è quindi il rapporto tra probabilità complementari ed esprime quanto le probabilità a favore sono maggiori (o minori) di quelle contro[[10]](#footnote-10).

L’ODDS è una misura che che si muove nell’intervallo . Tuttavia, se all’ODDS applichiamo il logaritmo naturale il risultato è una quantità che va da . Otteniamo così il cosiddetto LOGIT:

Se vediamo la probabilità che un evento accada (leggi: che un attributo sia presente, cioè che ) come funzione di regressori , è matematicamente dimostrabile[[11]](#footnote-11) questa equazione:

= z

A questo punto, se esprimiamo z come una combinazione lineare delle variabili indipendenti abbiamo realizzato la trasformazione della funzione logistica in una funzione lineare, tramite il link del LOGIT[[12]](#footnote-12), e possiamo modellare la probabilità che un evento a due uscite accada come:

dove quindi il LOGIT è una funzione lineare che dipende da variabili indipendenti e una intercetta, quindi parametri.

# La stima e l’interpretazione dei coefficienti

### Analisi campionaria

Poiché l’analisi statistica si rivolge quasi sempre a campioni estratti dalla popolazione e non alla totalità di essa, il valore dei coefficienti potrà essere soltanto stimato, e non calcolato. Se anche sia vera l’assunzione che la vera funzione che spiega il fenomeno è la funzione logistica, lo studioso non potrà altro che stimare a partire da un campione i coefficienti, e quindi predire l’uscita della variabile dipendente soltanto a partire da quelle stime campionarie, sui cui poi verranno applicate le tecniche di inferenza statistica per estendere quanto più possibile le conclusioni all’intera popolazione, secondo un certo grado di incertezza.

### La stima tramite il metodo della massima verosimiglianza

Esistono varie tecniche per stimare i coefficienti di un modello. Nella regressione lineare il più usato è il metodo OLS, *ordinary least squares*, il metodo dei minimi quadrati. Nella regressione logistica, poiché viene violata l’assunzione di omoschedasticità, questa tecnica non si può usare[[13]](#footnote-13). Si usa invece il metodo della massima verosimiglianza. In termini generici, si tratta di creare una funzione con i parametri del modello di studio (nel nostro caso quello logistico), la cui uscita è la probabilità che venga estratto dalla popolazione proprio il campione che stiamo analizzando; poi, tramite tecniche matematiche, risolvere l’equazione e dare un valore ai coefficienti col fine di massimizzare questa probabilità. La massima verosimiglianza è una tecnica ampiamente usata, e invero la stima tramite OLS nella regressione lineare conduce ai medesimi risultati[[14]](#footnote-14). I dettagli matematici della funzione di massima verosimiglianza sono al di là dello scopo di questo elaborato, basti aggiungere che la si può risolvere solo con metodi numerici iterativi. Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti[[15]](#footnote-15).

### Il significato dei coefficienti nel modello di regressione logistica

Ci sono differenti modi di interpretare il significato di un coefficiente in una regressione logistica. Per semplificare, immaginiamo uno scenario con una sola variabile. La funzione sarebbe la seguente:

è il parametro intercetta, ovvero il valore del LOGIT quando tutte le variabili indipendenti assumono valore pari a 0. rappresenta l’aumento del LOGIT per un aumento unitario di , proprio come nella regressione lineare. L’aumento del LOGIT significa generalmente un aumento delle probabilità a favore, anche se rimane una interpretazione non troppo concreta. Se tuttavia facciamo una trasformazione esponenziale, allora l’equazione diventa

Allora un aumento unitario di si risolverebbe così:

e quindi rappresenterebbe l’esponente di *e* tale per cui un aumento unitario di comporterebbe un aumento moltiplicativo e non additivo di sull’ODDS.

Nel mondo reale di solito i fenomeni sono multivariati, e quindi il modello logistico conterrà coefficienti. Proprio come nella regressione lineare, il valore di ogni è stimato come *adjusted*, vale a dire che viene stimato tenendo conto di tutti gli altri parametri e che deve essere interpretato numericamente come il valore che il coefficiente assume mantenendo tutti gli altri coefficienti costanti[[16]](#footnote-16).

# Test di ipotesi sui coefficienti del modello

Nella costruzione di un modello predittivo tramite la regressione logistica l’obiettivo di un test di ipotesi è capire quanto probabile possa essere, sotto certe assunzioni, che un regressore risulti utile alla predizione solo in ragione della casualità con cui è stato estratto il campione analizzato. Si tratta di una fondamentale tecnica di inferenza statistica che ha l’ambizione di estendere, sotto dei gradi di incertezza definiti, le conclusioni campionarie all’intera popolazione[[17]](#footnote-17). Esistono due approcci fondamentali al test di ipotesi dei coefficienti nella regressione logistica: il test rapporto di verosimiglianza e il test di Wald. Entrambi assumono come ipotesi nulla che alcuni coefficienti del modello siano pari a zero; se l’ipotesi viene rigettata, significa che il regressore associato a quel coefficiente è utile al modello.

Vediamoli entrambi nel dettaglio.

### Il test di rapporto di massima verosimiglianza

Il test rapporto di massima verosimiglianza confronta il valore assunto dalla funzione di massima verosimiglianza per due modelli: il modello cosiddetto *full*, esteso, che contiene più parametri, e il modello ridotto, che ne contiene di meno[[18]](#footnote-18). I parametri assenti dal secondo – perché assunti come aventi coefficienti pari a zero – sono quelli su cui si vuole testare l’ipotesi nulla. La statistica test con cui si rigetta o meno l’ipotesi viene costruita come differenza tra i due valori di massima verosimiglianza (la funzione , *likelihood* in inglese), a cui poi si applica il logaritmo naturale moltiplicato per due e cambiato di segno. Per le proprietà dei logaritmi viene che:

dove LRT sta per *likelihood ratio test* e per *likelihood* (verosimiglianza). Si può dimostrare che, sotto l’assunzione di un campione abbastanza grande, questa statistica test ha una distribuzione con gradi di libertà pari al numero di parametri di differenza tra i due modelli[[19]](#footnote-20). Se denominiamo con l’insieme dei parametri presenti nel modello più esteso ma non in quello ridotto, allora la nostra ipotesi nulla è:

:

o anche

:

e cioè che nessuno dei parametri aggiuntivi porti un contributo al modello. Se invece questo dovesse accadere, possiamo aspettarci che ; più piccolo sarà questo valore, più grande sarà allora . Alti valori della variabile casuale comportano *p-value* molto bassi; significa quindi che la probabilità che un alto valore di (e quindi l’evidenza che i coefficienti dei parametri aggiuntivi sono diversi da ) sia dovuto al caso sotto l’ipotesi nulla (e cioè che i coefficienti dei parametri aggiuntivi sono uguali a ) è molto bassa; se è inferiore al livello di significatività del test (usualmente ) allora si avrà il rigetto di : i coefficienti aggiuntivi sono utili al modello.

### Il test di Wald

Il test di Wald invece è costruito dividendo il coefficiente stimato per il suo errore standard stimato. Questa statistica test si può dimostrare abbia una distribuzione normale standardizzata se il campione è abbastanza grande[[20]](#footnote-21). Basterà quindi utilizzare la statistica test [[21]](#footnote-22)per rigettare o meno l’ipotesi nulla che il coefficiente sia diverso da zero.

### Test multipli e correzione di Bonferroni

Per completezza, andrebbe aperta una digressione sull’incremento della probabilità di errore di tipo I nel test di ipotesi quando si effettuano test multipli (come nel caso di più coefficienti da testare). L’errore di tipo I è la probabilità di rigettare l’ipotesi nulla quando questa è vera, ed è pari a . La probabilità di non commettere nessun errore è quindi pari a 0.95. Questa probabilità è chiamata FWER (*family wise error rate*)[[22]](#footnote-23) e si può esprimere come . Matematicamente si esprime come:

[[23]](#footnote-24)

dove è il numero di test eseguiti. Al crescere di questo cresce il FWER.

Esistono procedure di aggiustamento del livello di significatività che tengono conto di questo fenomeno, la più famosa delle quali è il metodo Bonferroni, che consiste nel diminuire in livello di significatività del test dividendo quello stabilito per T[[24]](#footnote-25). Il rischio del metodo Bonferroni è che rendendo molto difficile il rigetto dell’ipotesi nulla possa abbattere la potenza del test[[25]](#footnote-26), cioè la probabilità di rigettare l’ipotesi nulla quando questa è falsa.

# Valutazione della bontà di adattamento ai dati del modello

Esiste una differenza sostanziale tra statistiche test atte a testare la bontà di adattamento ai dati di un modello e misure del potere predittivo dello stesso modello[[26]](#footnote-27). Il potere predittivo di un modello misura quanto il modello predice in maniera corretta la variabile di risposta grazie ai valori delle variabili indipendenti. I test statistici sulla bontà dell’adattamento ai dati invece testano se un modello differente si adatti ai dati meglio di quello appena stimato. Capita che un modello con scarso potere predittivo si adatti bene ai dati (in tal caso sarà un fenomeno con tantissimo rumore e quindi con un termine di errore nella funzione vera molto grande[[27]](#footnote-28)) o che un modello con scarso potere predittivo si adatti bene ai dati (in tal caso siamo di fronte a *overfitting*[[28]](#footnote-29).

Come misure del potere predittivo di una regressione logistica sono utilizzate, tra le altre, differenti versioni di e l’AUROC (*area under the curve*)[[29]](#footnote-30). È fuori dallo scopo di questo elaborato approfondire le argomentazioni a favore di una certa versione dell’ piuttosto che di un’altra (ne sono state individuate almeno 12, di cui le più importanti quella di Mc Fadden e quella di Cox e Snell). Nella nostra ricerca useremo l’AUROC, che trova un consenso molto meno controverso tra gli studiosi. Questo paragrafo vuole trattare invece le statistiche di bontà di adattamento ai dati di un modello.

### I residui

Stimato un modello, si può dire che si adatti bene ai dati se, trovata una misura dello scarto tra valore stimato e valore osservato, la misura complessiva per le unità statistiche analizzate è molto piccola, e se il contributo dell’i-esimo confronto a tale misura rientra nei limiti della variabilità intrinseca del fenomeno, descritta dal termine d’errore del modello[[30]](#footnote-31).

Nella regressione lineare la bontà di adattamento ai dati osservati è una funzione dei residui, definiti come la differenza tra valore reale e valore atteso per ogni osservazione i-esima . Nella regressione logistica il valore stimato della variabile di uscita da cui calcolare il residuo non è associato alla i-esima osservazione, ma al k-esimo gruppo di osservazioni accomunate dalla medesima combinazione di valori delle varabili indipendenti (*covariate pattern*)[[31]](#footnote-32).

Esistono due differenti metodologie per calcolare i residui di un modello di regressione logistica, il residuo di *Pearson* e il residuo di *deviance*.

Sia il numero di *covariate pattern* delle variabili indipendenti osservate per le *n* unità statistiche; sia il numero di unità statistiche presenti in ogni g-esimo covariate pattern ; siano e rispettivamente il numero osservato e il numero stimato di unità statistiche appartenenti al k-esimo *covariate pattern* per cui , cioè:

Allora il singolo k-esimo residuo di *Pearson* si definisce come:

cioè come residuo diviso la sua deviazione standard, e una misura sintetica della bontà di adattamento come somma dei quadrati dei residui:

=

che per campioni sufficientemente grandi e per si distribuisce come una variabile casuale con gradi di libertà (sotto l’ipotesi nulla che il modello stimato sia quello corretto)[[32]](#footnote-33). Perciò valori bassi della statistica , che conducono a *p-value* molto alti, porteranno all’impossibilità di rigettare l’ipotesi nulla e quindi di considerare il modello stimato adattato ai dati.

Per spiegare il residuo di *deviance* utilizzeremo invece un altro approccio[[33]](#footnote-34). Un modello saturo è un modello che contiene tanti parametri quante sono le osservazioni, e non genera residui poiché le sue predizioni sono sempre esatte. La devianza è una statistica test che confronta le performance di un modello saturo con quelle del modello da testare. Si può dimostrare[[34]](#footnote-35) che la devianza è una statistica test risultante dalla differenza tra il valore di massima verosimiglianza del modello saturo e il valore di massima verosimiglianza del modello da testare:

Questo test assume la forma di un test rapporto di massima verosimiglianza, e quindi si potrebbe ragionevolmente concludere che la statistica associata si distribuisca, sotto l’ipotesi nulla, come una variabile casuale con ) gradi di libertà. Data l’ipotesi nulla:

:

dove racchiude tutte le stime dei coefficienti presenti nel modello saturo ma non in quello stimato, allora un *p-value* sotto il livello di significatività (quindi una statistica molto grande, quindi un rapporto tra le verosimiglianze molto vicino a , quindi i coefficienti dei parametri aggiuntivi presenti nel modello saturo pari a ) ci indurrebbe a rigettare l’ipotesi nulla e quindi a concludere che il modello saturo non fornisce un contributo significativamente migliore del nostro modello. La bontà di adattamento infatti testa non il potere predittivo del modello, ma se un modello con più termini di interazione (piuttosto che con una funzione link differente) si adatti meglio ai dati in termini di entità dei residui[[35]](#footnote-36).

Ora, per le due statistiche sulla bontà di adattamento finora viste, la statistica di *Pearson* e la devianza , l’approssimazione della statistica test alla variabile casuale si verifica se e solo se il numero dei gruppi di soggetti accomunati dalla stessa combinazione di valori delle variabili indipendenti è molto inferiore rispetto alla grandezza del campione[[36]](#footnote-37) e se la numerosità degli eventi e dei non-eventi presenti in ogni singolo *covariate pattern* è almeno. Con molte variabili continue presenti nel dataset, questo assunto viene facilmente violato e i test non sono più affidabili.

### La statistica Hosmer – Lemeshow (HL)

Una ulteriore statistica test per la bontà di adattamento ai dati è stata proposta dagli studiosi Hosmer e Lemeshow[[37]](#footnote-39), per ovviare ai limiti di statistiche test che si basato sul numero di *covariate pattern* .

Hosmer e Lemeshow hanno proposto il raggruppamento delle osservazioni in base ai valori stimati della variabile dipendente . In particolare, i valori predetti sono ordinati dal più piccolo al più grande, e poi separati in gruppi di approssimativamente uguale dimensione. Dieci gruppi è la raccomandazione standard.

Per ciascun gruppo, si calcola il numero osservato di eventi () e non-eventi () – presenza o assenza di attributo, così come il numero atteso di eventi (e non-eventi (). Il numero atteso di eventi è solo la somma delle probabilità previste per tutti le osservazioni del gruppo. E il numero atteso di non-eventi è la dimensione del gruppo meno il numero atteso di eventi.

La statistica HL risultante

si distribuisce come La variabile casuale con gradi di libertà e può essere utilizzata per confrontare i conteggi osservati con conteggi attesi. Se il modello da testare fosse un modello saturo, la statistica HL sarebbe nulla; sotto l’ipotesi nulla che il modello è corretto, un basso valore di HL e quindi un alto *p-value* ci indurrebbero a non poter rigettare l’ipotesi nulla che il modello è corretto[[38]](#footnote-40). Per quanto anch’essa oggetto di critiche la statistica HL è ampiamente la più usata e quella su cui c’è maggiore convergenza tra gli studiosi[[39]](#footnote-41).

# Selezione del miglior modello

I metodi di selezione delle variabili per la costruzione di un modello verranno approfonditi nel capitolo secondo di questo elaborato[[40]](#footnote-42). Questi possono applicarsi anche nel caso della regressione logistica. Sono stati introdotti anche dei metodi specifici di selezione delle variabili per la regressione logistica come il *Purposeful selection**algorith*[[41]](#footnote-43) , che rappresenta comunque un ibrido dei modelli di selezione basati sul *p-value* dei coefficienti. Essenziale nella scelta del modello è lo scopo del medesimo: se il fine è la predizione, bisogna avvalersi di misure del potere predittivo; se il fine è l’inferenza, specialmente nel caso in cui si voglia analizzare il rapporto causale tra una variabile di esposizione e una variabile di risposta alla luce dell’influenza di alcuni fattori di rischio, un approccio possibile è la costruzione di un modello gerarchico che includa la variabile di esposizione, alcune variabili di controllo e altre di interazione[[42]](#footnote-44). È al di là dello scopo dell’elaborato approfondire le tecniche di selezione di un modello gerarchico, in quanto nella analisi dei dati il nostro obiettivo sarà la predizione e tutte le variabili verranno trattate come possibili variabili esplicative.

# Diagnostica nella regressione logistica

### Assunzioni del modello

La regressione logistica, al contrario della regressione lineare, regge su molte meno assunzioni che, se violate, tolgono validità al modello. Non viene assunta una dipendenza lineare tra la variabile dipendente e quelle indipendenti; la regressione logistica può spiegare effetti non lineari in ragione del fatto che la funzione di link (il LOGIT) non è lineare. Inoltre non ci sono assunzioni sulla distribuzione delle variabili dipendenti, né sulla omogeneità della varianza della variabile dipendente per ogni livello (combinazione di valori) di quelle indipendenti. Da ultimo, non vengono fatte assunzioni sulla normalità della distribuzione del termine di errore casuale[[43]](#footnote-45). La diagnostica del modello non deve quindi orientarsi a verificare codeste assunzioni, come accade invece nel modello di regressione lineare.

Il modello di regressione logistica invece assume che le osservazioni siano indipendenti tra di loro[[44]](#footnote-46); questo può non verificarsi in ragione dei dati che si è raccolto (serie storiche piuttosto che campionamento a grappoli[[45]](#footnote-47)); altra assunzione è che tutte le variabili significative siano state incluse (la cosiddetta assunzione di specificazione del modello). La regressione logistica assume inoltre che tra LOGIT e variabili dipendenti ci sia una relazione lineare.

La trattazione seguente è frutto di ricerche che hanno sofferto di una difficoltà fondamentale: non c’è davvero comunanza di vedute tra gli studiosi su quali siano i migliori e più adatti e affidabili strumenti diagnostici nella regressione logistica[[46]](#footnote-48).

### Analisi dei residui

Nella regressione lineare l’analisi dei residui gioca un ruolo importante nella verifica della violazione delle assunzioni del modello; l’idea generale è che il residuo debba essere indipendente dai predittori, vale a dire che nella stima non ci si debba mai sbagliare in maniera sistematica, secondo un qualche trend. Se così fosse, ci sarebbe un effetto sistematico che agisce e che può essere spiegato includendo nel modello qualche regressore non utilizzato, rimanendo così solo il residuo dovuto alle limitazioni della stima campionaria e alla realizzazione del termine di errore casuale (il rumore) dei parametri (per il fatto che nessuna funziona descrive perfettamente un fenomeno).

Nella regressione logistica il problema è che, innanzitutto, di residui si può parlare solo per dati raggruppati, dove , e inoltre essendo il residuo k-esimo dipendente da , il residuo può dipendere dal regressore essendo comunque il modello corretto. Quindi non è chiaro che tipo di pattern rimanente possa essere sospetto oppure no[[47]](#footnote-49). Chi scrive nota una certa carenza di letteratura sull’argomento, essendo anche i manuali più noti sprovvisti della trattazione di questo argomento[[48]](#footnote-50). Questo non stupisce considerando la scarsità di assunzioni del modello di regressione logistica.

### Multicollinearità

È statisticamente dimostrabile che includere in un modello lineare (e anche in un modello lineare generalizzato) regressori linearmente correlati tra loro comporta un aumento dell’errore standard degli stimatori dei coefficienti dei regressori stessi, fenomeno noto come multicollinearità o inflazione della varianza[[49]](#footnote-51). A volte il ricercatore può decidere di includere delle variabili (magari di controllo o comunque variabili che se escluse creerebbero problemi di distorsione della stima[[50]](#footnote-52)) anche se incidono notevolmente sulla varianza dei coefficienti stimati (e quindi sui test di ipotesi e sulla stima intervallare della variabile di risposta); tuttavia una diagnostica su un modello deve garantire il governo di questo fenomeno e una scelta consapevole sull’inclusione o meno di variabili fortemente correlate ad altre.

Poiché la multicollinearità può riguardare più di due variabili, non basta una matrice degli scatterplot per individuare la correlazione; si pensi a = - .

Serve quindi uno strumento diagnostico adatto a modelli multivariati; nel nostro caso proponiamo il CNIs (*condition indices*) e il VDPs (*variance decomposition proportions*)[[51]](#footnote-53). I due indici sono stati creati per il modello di regressione lineare, e sono stati poi adattati anche alla regressione logistica[[52]](#footnote-54). Spiegare come vengono calcolati i due indici richiederebbe tecniche di algebra matriciale, che è al di là degli scopi di questo elaborato.

Come regola d’uso si può procedere come segue: data la tabella con CNIs e VDPs:



Sulla prima riga ci sono i vari CNIs. Se il CNI più grande è considerevolmente maggiore di , allora se almeno due valori della VDP legata a quel CNI sono maggiori di , allora c’è un problema di multicollinearità[[53]](#footnote-55). Esso si può affrontare togliendo la variabile con VDP più alto e riprocessando il modello per poi verificare in sede diagnostica se il problema di multicollinearità sussiste ancora.

Un metodo alternativo potrebbe essere quello di combinare due variabili correlate in un'unica variabile che contenga entrambe le informazioni (ad esempio combinare altezza e peso nell’indice di peso corporeo[[54]](#footnote-56)), mantenendo così l’informazione ma al tempo stesso non soffrendo di nessun problema di inflazione della varianza degli stimatori dei coefficienti.

### Osservazioni influenti

Nella regressione lineare vi sono misure atte a diagnosticare il comportamento anomalo di alcune osservazioni, per capire se alcune di esse sono molto lontane dal resto dei *datapoints* (gli *outlier*) o se alcune di esse esercitano una leva significativa sulla retta di regressione[[55]](#footnote-57). In un modello multivariato l’ispezione grafica è purtroppo insufficiente come strumento diagnostico, servono invece indici sintetici. Misure che si basano sull’analisi dei residui risentono dell’assunto di fondo che il numero di *covariate pattern* sia molto minore del numero di osservazioni. Poiché noi non vogliamo sottostare a questo assunto[[56]](#footnote-58) tratteremo solo misure diagnostiche che non lo prevedono.

Tecnicamente una osservazione è influente se e esercita alta leva se la sua rimozione dal modello già stimato comporta un cambiamento significativo del valore stimato di alcuni o tutti i coefficienti[[57]](#footnote-59).

I *dfbetas* o *delta-betas* sono una misura del cambiamento nel valore stimato di tutti i coefficienti qualora venga rimossa una certa osservazione. Tale misura coincide con la differenza tra le due stime diviso l’errore standard dello stimatore iniziale. Sia la stima di escludendo dal campione la i-esima osservazione, allora:

Ad ogni i-esima osservazione vengono quindi associati tutti i coefficienti e se ne può verificare il cambiamento nella stima qualora quella certa osservazione venga rimossa. La distribuzione dei *dfbetas* è ignota quindi non si può stabilire *tramite p-value* quando sia davvero abbastanza grande da dover considerare una osservazione come influente; come regola d’suo si stabilisce che un valore maggiore di due sia altamente indicativo di una osservazione influente su almeno un coefficiente[[58]](#footnote-60). Individuare una osservazione come influente non determina necessariamente la decisione di rimuoverla dal dataset e dal modello; i valori anomali che assume potrebbero dipendere da un errore di misurazione tanto quanto da caratteristiche proprie del fenomeno di cui bisogna comunque tenere conto.

### Errata funzione di link – mal specificazione del modello

Una assunzione fondamentale della regressione logistica è il rapporto lineare tra predittori e LOGIT, la cosiddetta funzione di link. Un modello può essere inadeguato perché manca questo assunto, o perché non sono state incluse variabili significative nel modello. Una tecnica per diagnosticare se la funzione di link è adeguata è quella di includere nel modello predittori quadrati dei predittori già inseriti; se il coefficiente è significativamente diverso da zero, allora la funzione di link è mal specificata. È importante che una tecnica del genere la si può applicare anche a dati non raggruppati, dove non è molto minore di [[59]](#footnote-61). Si potrebbe tentare anche la strada dell’ispezione grafica – sempre utile, che tuttavia soffrirebbe di due limitazioni: la prima è che in presenza di una analisi multivariata con l’inclusione di molti regressori, si dovrebbero creare e interpretare molti grafici e quindi verrebbe di molto compromesso il vantaggio dell’immediatezza visiva; il secondo è che un modello multivariato si basa sul concetto già menzionato dell’*adjustment*: il rapporto tra variabile esplicativa e variabile di risposta cambia a seconda di quali altre variabili si tengono in conto nello stesso modello, poiché ogni coefficiente stimato va interpretato come indicativo della relazione tra le due variabili mantenendo costanti i valori di tutti gli altri coefficienti.

### Riepilogo

In sintesi, una volta stimato un modello, le tecniche diagnostiche di cui bisogna avvalersi per sincerarsi che il modello sia adeguato e che si adatti ai dati sono le seguenti (alcuni punti sono stati trattati in paragrafi precedenti e a sé stanti per la loro rilevanza):

* Test per la bontà di adattamento;
* Analisi dei residui;
* Diagnosi di multicollinearità;
* Diagnosi di osservazioni particolarmente influenti;
* Correzione (eventuale) di Bonferroni per test multipli;
* Correttezza della funzione di link.

CAPITOLO secondo

I MODELLI PREDITTIVI E LA CLASSIFICAZIONE BINARIA

Questo capitolo tratta alcuni concetti di base dell’apprendimento statistico volti a fondare concettualmente il caso di studio del capitolo terzo, che per l’appunto consiste nella creazione di un modello di regressione logistica binaria. Per questo molti temi di alto valore ma altrettanta alta complessità, legati o ad algoritmi di tipo differente (KNN, reti neurali, alberi decisionali, regressioni non-lineari) piuttosto che a casi di analisi sostanzialmente differenti (ad esempio poche osservazioni e migliaia di variabili, come nel caso della genomica) sono stati ignorati. Lo stesso dicasi per una trattazione matematica di tipo dimostrativo dei concetti esposti.

# Il senso dei modelli di apprendimento statistico

L’apprendimento statistico (detto anche *machine learning*, *predictive modeling*, *data mining*) è un insieme di tecniche statistiche supportate da strumenti automatici di calcolo che ha l’obiettivo di sviluppare modelli in grado di fornire predizioni accurate[[60]](#footnote-62).

Supponendo di aver raccolto dei dati e di avere a disposizione un campione da analizzare, lo studioso conoscerà una serie di caratteristiche delle unità oggetto di studio; queste caratteristiche osservate e misurate sono le cosiddette variabili indipendenti, o esplicative, le informazioni cioè grazie alle quali vogliamo conoscere qualcosa sul fenomeno in oggetto. Spesso (non sempre, come nel caso dei modelli non-supervisionati di cui si tratterà in seguito) questo fenomeno si concretizza in un attributo che viene misurato tramite una variabile dipendente, o di risposta, che può essere qualitativa o quantitativa. L’obiettivo dell’apprendimento statistico è quello di individuare grazie ai dati raccolti una funzione che permetta in futuro, partendo dai medesimi dati di input, di predire in futuro l’attributo prima che si debba (magari con alti costi) misurarlo. Ad esempio, una azienda di *e-commerce* che subisce frodi da parte dei suoi clienti, e che quindi può associare ad ogni unità (il cliente) un attributo (frode sì o no) può ambire ad utilizzare questi dati e le caratteristiche note del cliente (ad esempio età, numero di transazioni precedenti, genere, occupazione, ecc.) per poter calcolare la propensione alla frode dei futuri clienti prima che essi facciano una transazione. Oppure una banca che ha crediti in sofferenza e vuole utilizzare questa esperienza “costosa” per individuare quelle caratteristiche che permettano di individuare in anticipo un cliente che avrà difficoltà a pagare. Piuttosto che una compagnia assicuratrice che deve calcolare il giusto premio che un cliente deve pagare alla luce della sua storia assicurativa.

# La stima di [[61]](#footnote-63)

Il primo passo dell’apprendimento statistico è quindi quello di assumere che esista una funzione che descrive un fenomeno della forma:

dove è la variabile di risposta che ci interessa predire e è la matrice dei predittori grazie ai quali vogliamo predire il fenomeno. invece rappresenta il termine di errore casuale, che è una componente fondamentale della funzione. Infatti il termine di errore rappresenta il contributo di variabili non misurabili ma utili a predire il fenomeno, e rappresenta la variabilità non sistematica (e quindi per natura imprevedibile) di un fenomeno. Ai fini dell’apprendimento statistico è importante assumere che il termine di errore abbia media nulla (cioè che ci si sbagli in ogni predizione, ma in maniera non sistematica, come una retta di regressione che trafigge a mezza altezza una nuvola di punti) e che sia indipendente (non predicibile) dai predittori[[62]](#footnote-65). Questi ultimi infatti rappresentano e spiegano la variabilità sistematica del fenomeno, e non dovrebbero essere associati a quella non sistematica. Dato questo, il rapporto tra variabile indipendente e variabili dipendenti si può riscrivere come:

dove diventa quindi il valore atteso della variabile casuale , una sorta di media condizionata di a .

Bene, il secondo passo dell’apprendimento statistico è quello di provare a stimare tramite , cioè la funzione stimata grazie all’analisi campionaria. Sarà questo l’oggetto della nostra attenzione lungo tutto l’elaborato.

### Errore riducibile e irriducibile

Nello stimare tramite sicuramente si commetterà un errore, perché una stima perfetta partendo da dati campionari limitati è molto complessa. Questo errore è comprimibile in funzione della qualità dell’analisi svolta, ma anche immaginando di poter fornire una stima perfetta per rimarrà una parte di errore “irriducibile”, perché è anche funzione di . Questo comporta un limite alla accuratezza della predizione[[63]](#footnote-66), anche se questo limite è ignoto dato che nella ricerca è ignota.

### Perché stimare

Gli obiettivi nell’apprendimento statistico sono principalmente due: la predizione e l’inferenza[[64]](#footnote-67).

Nel caso della predizione si è interessati esclusivamente alla uscita di e, fintantoché le metriche di performance predittiva sono buone, non interessa capire come la funzione ha operato per generare quella uscita. Insomma, sta bene trattare come una *black box*.

Qualora invece il ricercatore sia interessato a capire quali variabili indipendenti sono associate alla variabile dipendente, qual è l’intensità e la direzione di questa associazione, e come cambierebbe il valore della variabile di risposta al variare del valore di una variabile esplicativa, allora l’obiettivo principale è quello dell’inferenza statistica, e il ricercatore sarà molto interessato a capire come funziona il modello. Questo, come vedremo, indirizzerà la scelta verso una tipologia di modello predittivo piuttosto che di un altro[[65]](#footnote-68).

### Come stimare : metodi parametrici e non-parametrici

Esistono due metodologie di fondo differenti per stimare .

La prima, cosiddetta dei metodi parametrici, parte con l’assunzione che abbia una certa forma, per poi concentrarsi sulla stima dei parametri della funzione e sulla verifica ex-post delle assunzioni relative a quella forma. L’esempio più immediato è quello della regressione lineare: data una variabile quantitativa e alcuni regressori, si assume che la relazione che li lega sia:

e quindi

In tal caso poi si dovrà adottare un certo metodo di stima dei parametri ignoti (nel caso della regressione lineare il metodo dei minimi quadrati) e si avrà così la stima di .

Questo approccio facilita di molto l’individuazione di un modello, ma si assume il rischio di sbagliare dal principio la supposizione sulla reale forma della funzione che descrive il fenomeno.

Un approccio differente è quello dei modelli non-parametrici. Un modello non-parametrico non fa alcuna assunzione sulla forma della funzione che regola il fenomeno, ma si lascia – per così dire – guidare completamente dai dati – e per questo per funzionare bene ha bisogno di una quantità notevole di osservazioni da processare[[66]](#footnote-70). Un ulteriore svantaggio è che, adattandosi molto ai dati processati, i modelli non-parametrici tendono a soffrire di più di un concetto chiave che incontreremo nel proseguo della trattazione, ovvero l’*overfitting*.

### Misurare l’adattamento ai dati del modello

Ogni modello ha le sue metriche di performance predittive piuttosto che le statistiche test[[67]](#footnote-71) per valutare la bontà di adattamento ai dati. La regressione lineare ha l’errore quadratico medio e la statistica F[[68]](#footnote-72), la regressione logistica l’AUROC e la devianza piuttosto che la statistica HL, la KNN l’errore quadratico medio piuttosto che la sensitività[[69]](#footnote-73) a seconda che la si usi per fini di regressione o classificazione (sulla cui distinzione si approfondirà successivamente[[70]](#footnote-74)), tuttavia ciò che qui essenzialmente si vuole discutere è che al fine di valutare qualunque modello di apprendimento statistico è necessario distinguere tra dati di training e dati di test[[71]](#footnote-75). Il dataset di training è il campione disponibile che viene utilizzato per l’addestramento del modello, cioè per la stima della funzione affinché quanto più possibile . Tuttavia proprio perché con qui si intende il vettore osservato della variabile di risposta nel campione di training, qualunque modello addestrato viene stimato per minimizzare la funzione di costo sul dataset di training. In verità poi ogni modello imparando dal “passato” (il dataset di training per l’appunto) dovrà essere applicato al mondo reale delle osservazioni di cui non si conosce il valore della variabile di risposta (il “futuro”), e solo su questi dati (comunemente detti “di test”[[72]](#footnote-76)) si potranno accertare le performance del modello. Sfortunatamente questi dati non sono disponibili al momento dell’addestramento del modello, e vedremo nel corso della trattazione dei metodi di ricampionamento per stimare le performance del modello su dati di test. Nascendo tuttavia per minimizzare la funzione di costo sulla base dei dati di training, è ovvio e fisiologico che le performance del modello sui dati di test siano peggiori[[73]](#footnote-77); diverso invece è il fenomeno dell’*overfitting*, ovvero dell’eccessivo adattamento del modello ai dati di training; per “eccessivo” si intende un adattamento del modello anche al rumore presente nei dati, il tentativo di spiegare la varianza non sistematica e quindi di comprimere l’errore irriducibile. Quando ciò accade, il modello performa in maniera eccellente nel dataset di training ma in maniera povera in quello di test, nel quale non riconosce più quei *pattern* non sistematici che avevano determinato le performance di training; soprattutto si riconosce *overfitting* quando un modello più semplice performa meglio di uno più complesso che si è troppo adattato ai dati di training[[74]](#footnote-78).

In generale il rischio di *overfitting* è più alto laddove è più alta la flessibilità del modello, intesa come capacità del medesimo di generare più funzioni di forma differente[[75]](#footnote-79) a seconda dei dati su cui si addestra. Ad esempio la regressione lineare o la regressione logistica sono modelli abbastanza rigidi, ed è difficile che possano correre il rischio di adattarsi troppo ai dati, come può accadere invece a una KNN con un o a modelli non lineari complessi.

# Gli algoritmi dell’apprendimento statistico

### Algoritmi supervisionati e non-supervisionati[[76]](#footnote-80)

Gli algoritmi di apprendimento statistico sono supervisionati quando nel dataset di training ad ogni osservazione è associata una misurazione della variabile di risposta che andrà poi predetta dal modello per le osservazioni di test. In casi del genere esiste una uscita attesa e il modello stimato può essere valutato in termini di quanto si avvicina a quella uscita attesa.

Gli algoritmi non-supervisionati invece non hanno lo scopo di predire il valore di una variabile di uscita, giacché quest’ultima non viene osservata, o semplicemente non esiste. Ad esempio una analisi che voglia individuare differenti segmenti di clienti accomunati da caratteristiche simili non include una variabile di risposta, ma si limita a individuare dei *pattern* significativi nei dati di input – è un caso di *cluster analysis*.

Oltre queste due macro-categorie esistono in realtà molte sottocategorie di algoritmi in base alla presenza della variabile di risposta: per dovere di menzione si citano gli algoritmo semi-supervisionati, dove la misurazione della variabile dipendente è parziale e presente solo per alcune osservazioni, gli algoritmi di rinforzo in cui la variabile di uscita è assente ma va stimata tramite tentativi ed errori con la supervisione di una ricompensa (così una macchina può imparare a giocare a scacchi).

### Regressione e classificazione

Se la variabile di interesse è quantitativa, allora il modello di apprendimento statistico assume la forma di una regressione (lineare o meno, questo non interessa). In questo caso il modello verrà valutato per quanto la quantità stimata riesce ad avvicinarsi a quella osservata (tipicamente tramite l’errore quadratico medio). Un modello di classificazione invece mira ad assegnare ogni unità statistica a una certa classe; la classificazione sarà binaria (caso più frequente) se i gruppi sono due, altrimenti sarà multinomiale. Un modello di classificazione può avere l’obiettivo di predire se un cittadino americano voterà per i democratici, per i repubblicani o per il candidato indipendente. Anche un modello non-supervisionato come la *cluster analysis* esegue una classificazione, benché non si conoscano a priori i gruppi sotto cui sussumere le osservazioni. Per i modelli di classificazione le performance genericamente si valutano su quante unità sono state classificate correttamente, anche se nel paragrafo del presente capitolo vedremo che sarà necessario considerare anche altri aspetti per una valutazione più approfondita e attendibile delle performance di un classificatore.

# Distorsione e varianza della stima del modello

Data come stima della funzione , la varianza dello stimatore si riferisce a quanto cambierebbe se venisse addestrata su un campione differente da quello utilizzato. In generale la varianza di uno stimatore – che ricordiamo essere una variabile casuale – è:

cioè quanto la stima oscilla rispetto al valore atteso dello stimatore[[77]](#footnote-81). Se parlassimo dello stimatore media campionaria, la varianza (che sappiamo essere ) la si potrebbe interpretare come la media degli scarti quadratici tra tante medie campionarie (calcolate a partire da tanti campioni) e la media di queste medie (tecnicamente il “valore atteso”, che sappiamo coincidere con la media della popolazione ). Se media degli scarti è piccola, chiaramente tutti i campioni tendono a generare medie vicine a quella della popolazione.

Uno stimatore invece è corretto (o *unbiased*) se dato il parametro da stimare , allora:

cioè, empiricamente e riprendendo l’esempio della media campionaria, se su tanti campioni la media delle medie coincide con quella della popolazione (anche se tale media delle medie potrebbe essere frutto di stime molto inferiori e molto superiori rispetto al vero parametro, e quindi soffrire di molta varianza).

Tornando a , sia la stima della variabile di uscita per la generica osservazione di test , allora lo stimatore è tanto più variabile se a differenti dataset di training con cui è stata stimata corrispondono molto diverse tra loro, mentre lo stimatore è corretto se a differenti dataset di training con cui è stata stimata corrispondono tante la cui media è identica a .

Per variabili dipendenti quantitative l’indicatore di prossimità dello stimatore al parametro è lo scarto quadratico medio (*mean squared error*, MSE), che per si può dimostrare[[78]](#footnote-82) essere uguale a:

Quindi posto che rimane la componente di errore irriducibile, ogni modello dovrà minimizzare queste altre due quantità (varianza e *bias*) sempre positive.

Minimizzare queste due quantità tuttavia è problematico: generalmente un modello più rigido ha un *bias* maggiore mentre uno più flessibile tende ad averne meno, ma ad avere una varianza più grande; di solito al crescere della flessibilità del modello il *bias* comincia a decrescere più velocemente di come cresca la varianza, sino al punto ottimale del *trade-off* (chiaramente ignoto) dopo il quale il *bias* non decresce quasi più mentre la varianza sale drammaticamente[[79]](#footnote-83). Trovare una soluzione vicina a quella ottimale è fondamentale per avere un buon modello, e questo rimane uno dei punti di attenzione più critici per lo studioso.

# Metodi di ricampionamento: la cross-validation

Uno dei problemi principali dell’apprendimento statistico è la difficoltà di stimare le performance (il tasso di errore, *error rate[[80]](#footnote-84)*) in fase di test di un algoritmo addestrato su dati di training. L’importanza di questo punto non risiede solo nella comunicazione “pubblica” della stima delle performance del modello, ma anche e soprattutto nella scelta “relativa” del miglior modello tra tutti quelli che si sono addestrati sui dati e che competono per essere scelti. Per affrontare questa sfida si adotta una prassi di ricampionamento chiamata *cross-validation* che consiste nel partizionare casualmente il campione a disposizione in più set, per poi utilizzarli in fasi diverse della creazione del modello (queste fasi sono tipicamente l’addestramento di più modelli, la selezione del migliore e la stima delle performance).

### Il paradigma train-validate-test[[81]](#footnote-85)

Questo approccio consiste nella divisione del campione in un set di training, uno di validazione, uno di test, con una proporzione tipica di [[82]](#footnote-86). Con una partizione del genere possiamo raggiungere più obiettivi: utilizzare il dataset di training per addestrare vari modelli (quindi calcolare i coefficienti), quello di validazione per selezionare il modello con l’*error rate* più basso, e quello di test per stimare il *test error rate* e avere quindi una idea di come l’algoritmo predittivo potrà performare su dati della medesima distribuzione di quelli utilizzati per l’analisi[[83]](#footnote-87).

In questo modo si mitiga anche il rischio di *overfitting*[[84]](#footnote-88); potendo infatti stimare più *test error rate* per più modelli, quelli che si adattano eccessivamente ai dati (presumibilmente i più complessi[[85]](#footnote-89) e flessibili) performeranno scarsamente sul set di validazione e quindi verranno scartati. Tuttavia questo approccio non è immune da controindicazioni: la prima è che la divisione del set, specialmente per set di dati di piccola dimensione, riduce ancor di più il numero di osservazioni su cui addestrare il modello, e quindi l’accuratezza della stima di potrebbe risentirne; la seconda è che, sempre su campioni non grandi, la casualità della partizione del dataset comporta una alta variabilità di , che per partizioni differenti potrebbe generare *test error rate* molto diversi tra loro[[86]](#footnote-90).

### K-fold cross-validation

Una seconda metodologia di *cross-validation* è la cosiddetta *k-fold cross-validation*, che consiste nel partizionare il campione in set differenti (di solito –ma non necessariamente[[87]](#footnote-91) – o [[88]](#footnote-92) e di uguale proporzione), per poi addestrare il modello volte su set e validarlo (quindi stimare il *test error rate*) sul k-esimo *fold* rimanente. Si avranno così, per ogni modello addestrato, *test error rate* da cui derivare una stima unica tramite la media[[89]](#footnote-93).

Rispetto al paradigma *train-validate-test* questo approccio è computazionalmente più dispendioso, ma ha alcuni vantaggi.

Il primo è che il modello viene addestrato sempre su set, che di solito per è maggiore del previsto dal paradigma *train-validate-test*, comportando questo una stima più accurata del *test error rate*. Inoltre addestrandosi il modello alternativamente su tutte le osservazioni del dataset, la variabilità della stima si riduce, perché non più soggetta alla casualità della partizione dei dati. Tuttavia il compromesso tra *bias* e varianza della stima si sbilancia se , la cosiddetta *leave-one-out-cross-validation*, poiché mediando la stima di *test error rate* calcolati addestrando un modello su pressoché gli stessi dati (cambia una osservazione per volta) avremo quantità molto correlate tra loro che genereranno una stima sì quasi corretta (perché desunta da osservazioni) ma altamente variabile, e quindi non ottimale.

### L’addestramento finale del modello e la nested k-fold cross-validation

La stima del *test error rate* è un presupposto importante per la selezione del miglior modello tra una lista di modelli che competono per essere scelti come quello migliore[[90]](#footnote-94). Si è visto come la *k-fold cross-validation* rappresenti il miglior compromesso tra varianza e *bias* del modello, tuttavia lascia una domanda insoluta: dato un modello vincente sugli altri, che valore assumono i suoi coefficienti, visto che il modello è stato addestrato volte? Esso andrà riaddestrato su tutto il dataset di training[[91]](#footnote-95) per poterlo poi applicare a un dataset di test da cui ottenere una stima definitiva (e relativa all’unico modello vincente) dell’*error rate*. Ma la classica *k-fold cross-validation* non prevede una partizione del dataset in modo tale che ne rimanga uno indipendente di test (infatti il campione si divide in parti, e la k-esima viene usata per la validazione cioè per la scelta del miglior modello); questa soluzione è invece prevista dalla *nested k-fold cross-validation*: immaginando un campione di osservazioni, si può pensare che l’ formerà il dataset di training-validazione eseguendo una *k-fold cross-validation*; il modello vincente verrà poi addestrato su tutto il dataset di training -validazione, e alla fine se ne stimeranno le performance applicandolo al restante dei dati. È evidente che un approccio del genere – che prevede una maggiore partizione dei dati – ha efficacia massima e minori controindicazioni per campioni di grande dimensione.

# La selezione delle variabili

I metodi di *cross-validation* servono per confrontare modelli diversi e stimare il *test error rate* di ognuno di loro per scegliere il modello che performa meglio. Tuttavia per confrontare modelli essi vanno costruiti; e la costruzione comporta come passo fondamentale la scelta di quali variabili includere nel modello. Esistono varie tecniche, alcune delle quali a forte componente automatica, altre che lasciano molto più spazio all’intuito e alla *expertise* del ricercatore[[92]](#footnote-96).

### La necessità di scegliere le variabili

La selezione delle variabili è un procedimento cardine nella costruzione di un modello predittivo. Gli obiettivi di questo procedimento possono essere due: nel caso di un modello volto alla predittività, l’obiettivo è quello di trovare la miglior combinazione possibile di variabili in grado di predire bene il fenomeno; nel caso di un modello votato all’inferenza, lo scopo è quello di esplorare la relazione tra una variabile dipendente e più variabili esplicative. Se in questo secondo caso la necessità di filtrare il set di variabili di input è chiara, nel primo caso invece ci si potrebbe porre la domanda: perché non limitarsi a includere tutte le variabili nel modello?

Innanzitutto un principio molto diffuso e apprezzato in statistica è il principio della parsimonia[[93]](#footnote-97): è preferibile, a parità – o quasi – di risultato ottenuto arrivarci con la minore complessità possibile. Vale a dire: col minor numero possibile di parametri inclusi nel modello. La parsimonia gioca anche a favore dell’interpretabilità di un modello: abbiamo visto nel corso del capitolo che a volte avere un modello interpretabile ricopre un ruolo fondamentale nella ricerca, e non basta una *black box* enormemente predittiva per soddisfare le esigenze di studio[[94]](#footnote-98). A tal fine la presenza di una variabile inutile potrebbe comportare (alla luce della stima dei coefficienti basata – almeno nei modelli lineari e lineari generalizzati – sull’*adjustment*) a interpretare in maniera completamente erronea la relazione tra le variabili di interesse.

Inoltre per campione di grande dimensione e per set di variabili di grande numerosità possono sopravvenire problemi di carichi computazionali nella esecuzione dell’algoritmo; in tali occasioni ridurre il numero di variabili da includere nel modello potrebbe persino fare la differenza tra permetterne l’applicabilità oppure no[[95]](#footnote-99).

Uno dei principali fattori di dissuasione all’inclusione indiscriminata delle variabili è la multicollinearità, già vista nel capitolo primo: avere variabili indipendenti correlate tra loro all’interno del modello aumenta l’errore standard dei coefficienti e quindi va ad inficiare pesantemente l’accuratezza della stima di a causa della inflazione della varianza.

Non va ignorato, persino nell’era dei *big data*, che in qualche circostanza fare le rilevazioni (raccogliere quindi le misurazioni delle variabili) sia estremamente costoso; in tal caso accertarsi che alcune informazioni sono inutili al modello può permettere al ricercatore di interrompere il sostenimento di quel costo.

### L’analisi esplorativa

L’analisi esplorativa è una tecnica di esplorazione statistica e grafica dei dati volta a indagare la distribuzione delle variabili e possibili relazioni tra le stesse, e a individuare in maniera intuitiva la forza e natura di eventuali associazioni esistenti tra loro[[96]](#footnote-100). È una fase idealmente preliminare alla costruzione di un modello (o di test statistici), in grado di guidare il ricercatore verso la creazione di una formulazione statistica consapevole e più adatta ai suoi scopi.

Strumenti tipici dell’esplorazione dei dati sono i grafici multipli e i grafici condizionati, la matrice degli scatterplot, tutte soluzioni in grado di rappresentare più dimensioni senza l’ausilio della grafica tridimensionale[[97]](#footnote-101).

Tra le tecniche esplorative propedeutiche alla selezione delle variabili di un modello di classificazione binaria citeremo il *weight of evidence* e l’*information value*, una tecnica di ordinamento delle variabili in base al potere predittivo che esse manifestano in una relazione univariata con la variabile target[[98]](#footnote-102).

Data una variabile categorica che si distribuisce su livelli, e sia il numero di unità presenti nel k-esimo livello, il numero di eventi nel medesimo livello, il numero di eventi nel campione; allora il *weight of evidence* di quel k-esimo livello è pari a :

e l’*information value* di quella variabile, l’indice sintetico del suo potere predittivo incondizionato, è:

Insomma si tratta di applicare il logaritmo naturale alla differenza tra la frequenza relativa di una classe e la frequenza relativa degli eventi in quella classe.

Per variabili quantitative è necessario eseguire un raggruppamento, chiaramente soggetto a una qualche arbitrarietà del ricercatore, e uno degli svantaggi dell’IV è proprio la sua sensibilità a come sono raggruppate le variabili. Il secondo svantaggio dell’IV è che presenta la forza della relazione di una variabile di input con la variabile target senza tenere in considerazione l’azione delle altre variabili, cosa che potrebbe sensibilmente cambiare la natura della relazione tra le altre due.

### I metodi stepwise[[99]](#footnote-104)

I metodi *stepwise* sono metodi di selezione delle variabili basati su criteri oggettivi, tipicamente una misura del potere predittivo dell’intero modello oppure il *p-value* di un coefficiente.

La *best-subset-selection* è una procedura guidata da un software di calcolo che esamina tutte le possibili combinazioni delle variabili selezionando poi (chiaramente tramite *cross-validation*) quello con la migliore performance. È un metodo che misconosce il contributo dell’esperto in materia nella selezione delle variabili, ed è soprattutto computazionalmente insostenibile per molto grande.

La *forward-stepwise-selection* è una procedura che, partendo da un modello senza variabili, misura le performance di modelli composti da una sola variabile, e seleziona il migliore; poi misura le performance di modelli composti da due variabili, includendo volta per volta un secondo predittore, e seleziona quello con le migliori performance. Procede iterativamente fino all’aver provato l’inclusione di tutte le variabili. È un approccio computazionalmente meno pesante ma non meno guidato, e inoltre incapace di riconoscere la miglior combinazione possibile se questa non include la prima variabile selezionata per il modello[[100]](#footnote-105).

Da ultimo la *backward-stepwise-selection*, che parte dal modello che include tutte le variabili e ne toglie iterativamente una alla volta, partendo da quella che abbatte di meno l’indicatore di performance o da quella con il *p-value* più alto[[101]](#footnote-106), fino alla selezione del miglior modello.

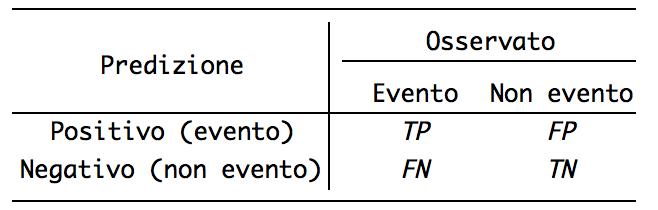
Restando inteso che un approccio ibrido è possibile e che il ricercatore – seppur secondo linee guida prestabilite, ha molta libertà nel seguire un suo metodo di selezione delle variabili (anche avvalendosi dello spirito critico di chi ha competenza settoriale)[[102]](#footnote-107), va evitato assolutamente un errore comune tra gli studiosi: la selezione delle variabili va fatta valutando performance e testando ipotesi solo sul dataset di training, e non anche su quello di test. In quest’ultimo caso infatti il modello avrebbe un vantaggio illecito nell’essere selezionato tramite indicatori di performance ottenuti utilizzando anche le informazioni di un dataset che dovrebbe rimanere completamente indipendente, pena una sovrastima del *test error rate[[103]](#footnote-108)*.

# Misurazione delle performance dei modelli di classificazione binaria

Le performance predittive di un modello di classificazione binaria si distinguono a seconda che l’uscita del modello predittivo sia una mera classe di appartenenza dell’unità di analisi (quindi nella classificazione binaria “evento” oppure “non-evento”, o ) oppure una probabilità (o comunque uno *score* ordinato di propensione) di appartenere alla classe “evento”. Nel primo caso gli strumenti di valutazione delle performance tipici sono la matrice di confusione e i suoi indicatori, nel secondo caso la curva ROC e le sue declinazioni.

### La matrice di confusione e i suoi indicatori

Dato un set di dati su cui si testa il modello, per il quale è disponibile la misurazione della variabile target che assume solo due valori (“evento” e “non-evento”, di solito codificati come e ) e la previsione del modello associata a ogni unità statistica del campione, anch’essa della forma o , è possibile costruire una matrice di questo tipo, denominata matrice di confusione[[104]](#footnote-109):



dove TP sta per *true positive* – cioè gli eventi correttamente predetti come eventi, FN sta per *false negative* – cioè gli eventi erroneamente predetti come non-eventi,

TN sta per *true negative* - cioè i non-eventi correttamente predetti come non-eventi, FP sta per *false positive* – cioè i non-eventi erroneamente predetti come non-eventi.

È evidente che:

dove “positivo” sta per “predetto come evento” e “negativo” sta per “predetto come non-evento”.

Il primo indicatore[[105]](#footnote-110) di performance di un modello tramite matrice di confusione che potrebbe venire in mente è il rapporto tra predizioni corrette e predizioni totali, l’*accuracy[[106]](#footnote-111)*:

che come massimo valore può assumere , il che implicherebbe un modello perfetto. Purtroppo alti valori di accuracy non sempre sono segno di un ottimo modello. Infatti data la *prevalence*[[107]](#footnote-112)

cioè la proporzione di eventi nel campione, e il *null error rate*:

cioè il tasso di errore se il modello classificasse tutte le unità nella categoria maggioritaria, allora: se la *prevalence* è particolarmente bassa, come ad esempio per malattie rare (dove può anche essere dello ), e se l’algoritmo predice sempre non-evento, l’*accuracy* del modello sarà e pur tuttavia sarà completamente inutile[[108]](#footnote-113).

Per questo ha senso utilizzare delle metriche condizionate di accuratezza, e cioè:

TPR sta per *true positive rate*, e misura la probabilità che un evento venga classificato come tale dal modello in termini di proporzione di eventi classificati come positivi.

TNR sta per *true negative rate*, e misura la probabilità che un non-evento venga classificato come tale dal modello in termini di proporzione di non-eventi classificati come negativi.

Due valori alti, prossimi all’unità, di questi due indicatori rappresentano una vista molto più robusta delle performance di un modello[[109]](#footnote-114).

Tuttavia in molti settori di ricerca etichettare una unità come positiva ha un costo molto alto anche se poi si rivela un non-evento[[110]](#footnote-115); le performance di un modello sotto questo aspetto si misurano con l’indicatore di *precision*:

cioè la probabilità che un positivo sia un evento[[111]](#footnote-117). Se questo indicatore è prossimo a (cioè se i *false positives* sono irrilevanti) allora il modello comporta bassi costi in termini di *precision*.

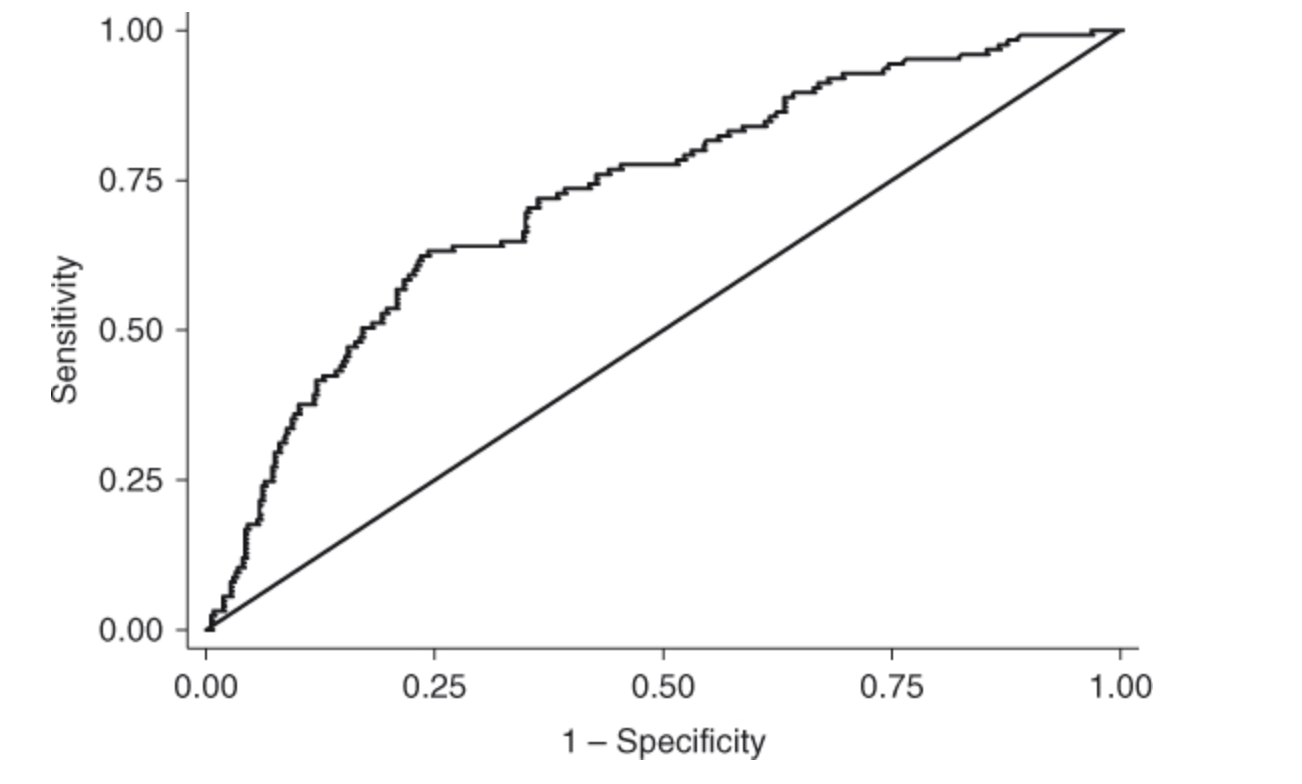
Da un punto di vista probabilistico si può considerare la *prevalence* come la *precision* di un classificatore casuale. Un confronto tra *prevalence* e *precision* ci direbbe quindi quanto il modello migliora le performance classificatorie del puro caso. A tal fine si può usare il *positive diagnostic likelihood ratio*, PDLR[[112]](#footnote-118), che confronta non le probabilità ma gli ODDS:

che si può dimostrare[[113]](#footnote-119) essere identico a:

cioè al rapporto tra la probabilità che un evento venga classificato come positivo e la probabilità che un non-evento venga classificato come positivo. Un modello con un PDLR alto, diciamo sopra , ha buone performance[[114]](#footnote-121).

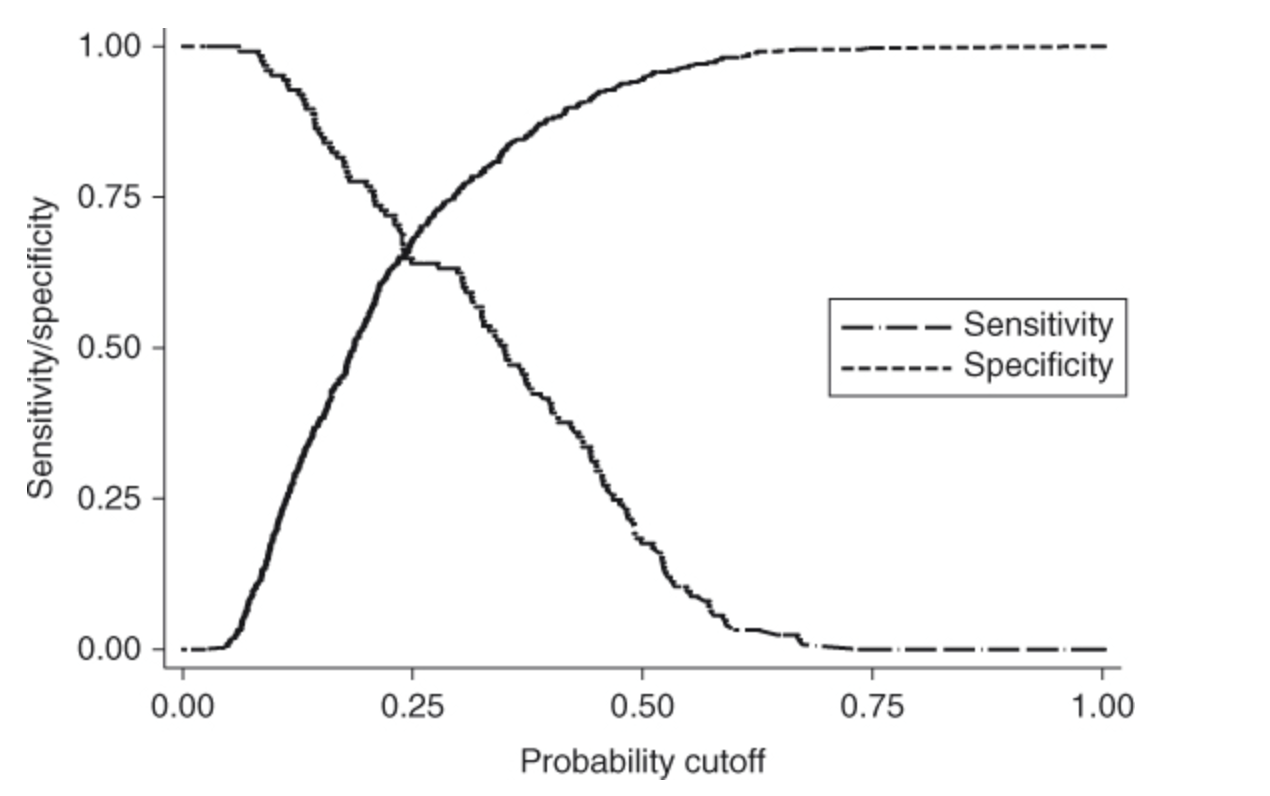
### La curva ROC

La curva ROC (*receiver operating characteristic*), curva nata nel contesto della teoria dei segnali[[115]](#footnote-122), è un strumento utile a rappresentare le performance di un classificatore che assegna a ogni unità una probabilità (o comunque un punteggio ordinato) di appartenere alla classe “evento”. Poiché tuttavia il fine è la classificazione binaria, bisognerà in qualche modo stabilire una soglia di probabilità o *score* superata la quale si considera l’unità come appartenente alla classe “evento”, e sotto la quale si considera l’unità come appartenente alla classe “non-evento”[[116]](#footnote-123). La curva ROC rappresenta sull’asse delle ascisse il FPR e su quello delle ordinate il TPR; ogni *datapoint* rappresenta una possibile soglia di *cut-off*, della forma:

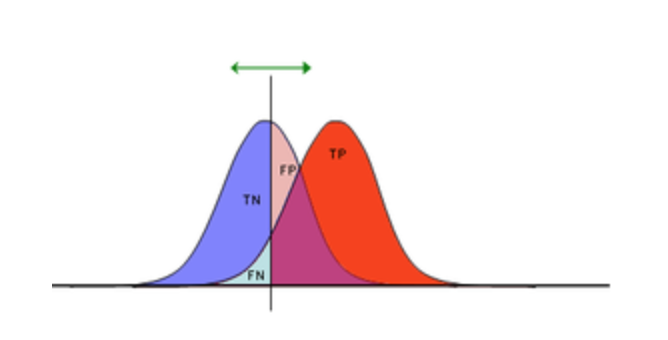


La retta rappresenta le performance di un classificatore casuale[[117]](#footnote-124); più la curva va in alto a sinistra più il classificatore performa bene (con TPR alto e FPR basso). L’area sotto la curva ROC (AUROC) è un indicatore sintetico delle performance predittive del modello[[118]](#footnote-125): se siamo di fronte a un classificatore perfetto, se siamo di fronte a un classificatore casuale. In genere si considera un modello come sufficientemente performante se . Peraltro si può dimostrare[[119]](#footnote-127) che l’AUROC è interpretabile come la probabilità che a un evento casualmente considerato venga assegnata una probabilità stimata di essere evento più alta di quella assegnata a un non-evento casualmente considerato.

Tuttavia per stabilire la soglia di *cut-off* desiderata ci sono grafici più interpretabili[[120]](#footnote-128); ad esempio quello che rappresenta in ascisse le soglie di *cut-off* e in ordinata TPR e TNR:



piuttosto che le distribuzioni delle probabilità stimate per degli eventi e non-eventi:



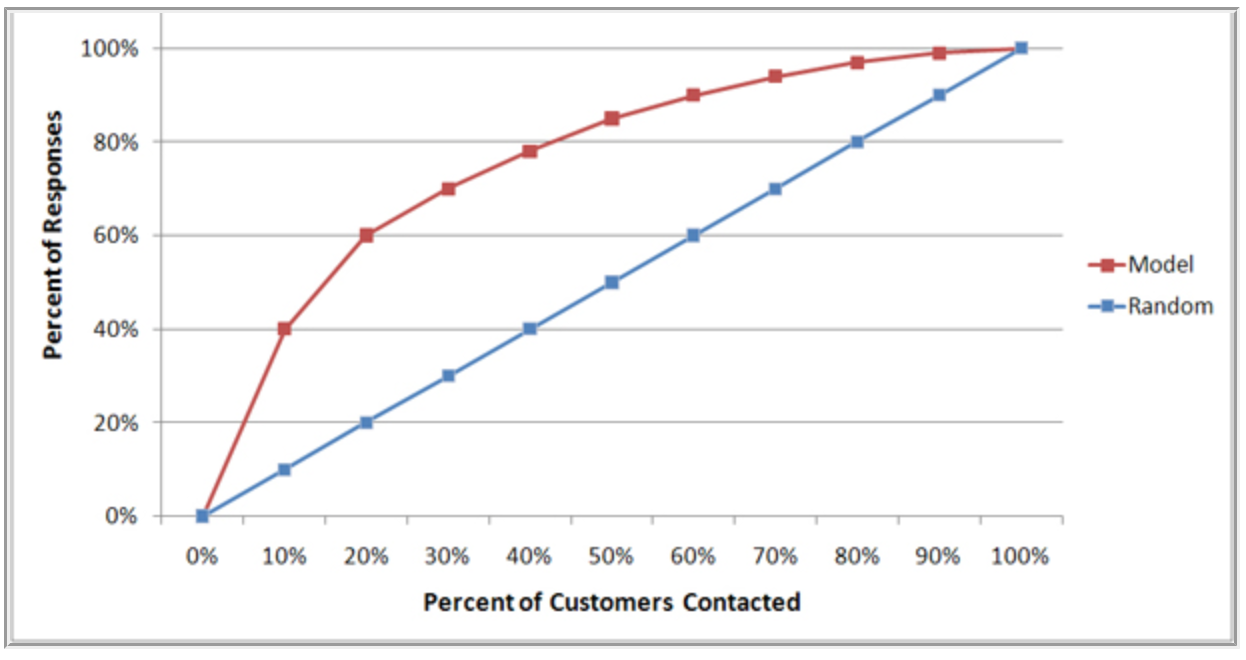
dove all’ascisse ci sono le probabilità stimate e alle ordinate la densità.

Da ultimo si può pensare di rappresentare sulle ascisse le varie soglie di *cut-off* e sull’ordinata un qualsivoglia indice sintetico di performance desumibile dalla derivante matrice di confusione[[121]](#footnote-129).

### I grafici di guadagno – gain charts

Un’ultima vista che qui si vuole presentare sulle performance di un classificatore, sempre nell’ipotesi che l’uscita del modello sia una probabilità o uno *score* ordinato, è il grafico di guadagno. In pratica si divide il campione in quantili (di solito ) ordinati per punteggio assegnato dal modello, dal punteggio più alto (maggior probabilità di essere un evento) al più basso. Per ogni quantile si calcola la percentuale di eventi presente, confrontandola con quella complessiva (cioè la *prevalence*), vale a dire la percentuale di eventi che si troverebbe in un quantile estratto casualmente. Si graficano le due percentuali ponendo sulle ascisse i quantili ordinati e sulle ordinate la percentuale di eventi.

Altrimenti è possibile sviluppare un grafico della percentuale cumulata di eventi nei quantili; in tal caso la *baseline* sarà la proporzione dei quantili stessa: ogni di clienti estratti casualmente ci si aspetta di trovare il degli eventi; il guadagno è lo spazio che si crea tra la spezzata del tasso di eventi cumulato nei quantili ordinati per probabilità stimata e la *baseline[[122]](#footnote-130)*. Ad esempio se il dei clienti di una banca risponde favorevolmente a un outbound per una *promozione*, ma nel decile con le probabilità stimate di essere un evento più alte (si legga: favorevole alla promozione) questa percentuale diventa , allora c’è stato un guadagno:



Anche del grafico di guadagno si può calcolare l’area sottostante come indice sintetico, l’AULIFT[[123]](#footnote-131).

Capitolo terzo

analisi dei dati per il cross-selling bancario

BIBLIOGRAFIA

1. Tale affermazione risulterà evidente dopo la trattazione del capitolo secondo. [↑](#footnote-ref-1)
2. Per lo svolgimento dell’analisi si rimanda, per l’appunto, al capitolo terzo. [↑](#footnote-ref-2)
3. La notazione qui utilizzata prende spunto da una delle opere più note di apprendimento statistico, JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *An introduction to statistical learning with applications in R*, New York, 2014, pp. 9-12. [↑](#footnote-ref-3)
4. BORRA S. – DI CIACCIO A., *Statistica. Metodologia per le scienze economiche e sociali*, Milano, 2014, p. 171. [↑](#footnote-ref-4)
5. LUBISCO A. – MIGNANI S. – PILLATI M., *Analisi statistica multivariata. La regressione logistica* (Internet), disponibile a **http://www2.stat.unibo.it/mignani/Didattica/analisideidati/logistica.pdf** (data di ultima consultazione: Marzo 2016), p.4. [↑](#footnote-ref-5)
6. *Ivi*, p.5. [↑](#footnote-ref-6)
7. *Ivi*, p.3 [↑](#footnote-ref-7)
8. Per la scelta di rappresentare la funzione logistica in questa forma si è preso spunto da KLEINBAUM D. – KLEIN M., *Logistic regression. A self learning text*, New York, 2010, p.5. [↑](#footnote-ref-8)
9. In epidemiologia la si potrebbe definire come la combinazione dei fattori di rischio sulla probabilità di insorgenza di una malattia, cfr. KLEINBAUM D. – KLEIN M., *Op. cit.,* pp. 6-7. [↑](#footnote-ref-9)
10. BOTTARELLI E. – OSTANELLO F., *Epidemiologia. Teoria ed esempi di medicina veterinaria*, Milano, 2011, p. 65 ss. [↑](#footnote-ref-10)
11. KLEINBAUM D. – KLEIN M., *Op. cit.,* p. 18. [↑](#footnote-ref-11)
12. La funzione di link è una funzione che trasforma in modo tale da legare la variabile dipendente a una combinazione lineare dei predittori. Cfr. AGRESTI A., *Foundations of linear and generalized linear models* (e-book), New York, 2015, par. 1.1. [↑](#footnote-ref-12)
13. LUBISCO A. – MIGNANI S. – PILLATI M., *Op. cit.*, p. 11. [↑](#footnote-ref-13)
14. HASTIE T. – TIBSHIRANI R. – FRIEDMAN J., *The elements of statistical learning*, New York, 2009, p. 31. [↑](#footnote-ref-14)
15. LUBISCO A. – MIGNANI S. – PILLATI M., *Op. cit.*, p. 15. Asintoticamente (cioè a dimensione del campione crescente) uno stimatore è corretto se il suo valore atteso (empiricamente, la media di varie stime basate su vari campioni) è uguale al parametro da stimare ed è efficiente se (essendo corretto) ha il minor errore standard possibile; Cfr. BORRA S. – DI CIACCIO A., *Op. cit.,* p. 267 ss. [↑](#footnote-ref-15)
16. CAFFO B., *Regression models for data science in R*, Baltimore, 2015, p. 53 ss, dove ci sono anche curiosi esempi del *Simpon’s Paradox*, cioè il cambio di segno del coefficiente di un parametro tenendo conto di altre variabili oppure escludendole.Test di ipotesi sui coefficienti del modelloanti. deve essere interpretato numericamente come il valore che assume non troppo c [↑](#footnote-ref-16)
17. CAFFO B., *Statistical inference for data science*, Baltimore, 2015, p. 79 ss. [↑](#footnote-ref-17)
18. Cfr KLEINBAUM D – KLEIN MITCHEL, *Op. cit.,* p. 135. [↑](#footnote-ref-18)
19. LUBISCO A. – MIGNANI S. – PILLATI M., *Op. cit.*, p. 23. [↑](#footnote-ref-20)
20. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.,* p. 139. [↑](#footnote-ref-21)
21. CAFFO B., *Statistical inference for data science*, cit., p. 81. [↑](#footnote-ref-22)
22. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.,* p. 280. [↑](#footnote-ref-23)
23. *Ibidem*. [↑](#footnote-ref-24)
24. Sul fatto che la correzione di Bonferroni si possa usare anche nella regressione multipla Cfr. MUNFROM D. – PERRETT J. – SCHAFFER J. – PICCONE A. – ROOZEBOOM M., *Bonferroni adjustments in tests for regression coefficients* (Internet), 2006, disponibile su **https://www.amstat.org/meetings/jsm/2008/onlineprogram/index.cfm?fuseaction=abstract\_details&abstractid=301702** (data di ultima consultazione: Marzo 2016), p.1 ss. [↑](#footnote-ref-25)
25. ESPA G. – MICCIOLO R. – CANAL L., *Ricerca con R. Metodi di inferenza statistica*, Milano, 2013, p. 145. [↑](#footnote-ref-26)
26. ALLISON P., *Measures of Fit for Logistic Regression* (Internet), 2014, disponibile su **http://statisticalhorizons.com/wp-content/uploads/GOFForLogisticRegression-Paper.pdf** (data di ultima consultazione: Marzo 2016), p.1 ss. Buona parte di questo paragrafo si ispira a questo lavoro del Professor Allison della University of Pennsylvaina. [↑](#footnote-ref-27)
27. HOSMER D. – LEMESHOW S., *Applied logistic regression* (e-book), New York, 2013, par. 5.2.4: “*the ability of a fitted model to discriminate between the two outcomes is more a function of the difference between the groups and magnitudes of the slope coefficients than the logistic model itself*”. [↑](#footnote-ref-28)
28. Vedi paragrafo del capitolo secondo. [↑](#footnote-ref-29)
29. Vedi paragrafo 7.2 del capitolo secondo. [↑](#footnote-ref-30)
30. LUBISCO A. – MIGNANI S. – PILLATI M., *Op. cit.*, p. 20. [↑](#footnote-ref-31)
31. HOSMER D. – LEMESHOW S., *Applied logistic regression*, cit., par. 5.2.1. [↑](#footnote-ref-32)
32. Lo sviluppo della formula per il residuo di *Pearson* deve molto a LUBISCO A. – MIGNANI S. – PILLATI M., *Op. cit.*, p. 20 ss. [↑](#footnote-ref-33)
33. Cfr. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.,* p. 305 ss. [↑](#footnote-ref-34)
34. HOSMER D. – LEMESHOW S., *Applied logistic regression*, cit., par. 5.2.1. [↑](#footnote-ref-35)
35. ALLISON P., *Op. cit.*, p. 4. [↑](#footnote-ref-36)
36. Cfr. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.,* pp. 312-318, da cui si prende largamente spunto perla trattazione di questo sottoparagrafo. [↑](#footnote-ref-37)
37. HOSMER D. – LEMESHOW S., *A goodness-of-fit test for the multiple logistic regression model*, in “Communications in Statistics”, 1980, pp. 1043-1069. [↑](#footnote-ref-39)
38. ALLISON P., Logistic regression using SAS: theory and application (ebook), Cary, 2012, par. 3.6 [↑](#footnote-ref-40)
39. ALLISON P., Why I don’t trust Lemeshow Test for Logistic Regression (Internet), disponibile su **http://statisticalhorizons.com/hosmer-lemeshow,** (data di ultima consultazione: Marzo 2016).. [↑](#footnote-ref-41)
40. Vedi paragrafo 6 del capitolo secondo. [↑](#footnote-ref-42)
41. HOSMER D. – BURSAC Z. – GAUSS C. – WILLIAMS D., *Purposeful selection of variables in logistic regression* (Internet), Amherst, 2008, disponibile su **http://www.readcube.com/articles/10.1186%2F1751-0473-3-17** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-43)
42. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.,* p 244 ss. [↑](#footnote-ref-44)
43. Cfr. CURINI L., *Regression with a binary dipendent variable: logistic regression diagnostic* (Internet), Milano, 2014, disponibile su **http://www.sociol.unimi.it/docenti/curini/Multivariate%20PHD%202015/Logit%20&%20Probit%20Diagnostic/PhD%202015%20Probit%20Logit%20Diagnostic.pdf** (data di ultima consultazione: Marzo 2016), p. 1. [↑](#footnote-ref-45)
44. Cosa che, se non accade, può generare problemi di *overdispersion*, cfr. Penn State University, *Analysis of discrete data* (Internet), disponibile su **https://onlinecourses.science.psu.edu/stat504/node/162** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-46)
45. Cfr. CURINI L., *Op. cit.*, p. 4. [↑](#footnote-ref-47)
46. KABACOFF R., *R in action. Data analysis and graphics with R* (ebook), New York, 2015, par. 13.1. [↑](#footnote-ref-48)
47. PARDOE I. – COOK D., *A Graphical Method for Assessing the Fit of a Logistic Regression Model* (Internet), 2002, disponibile su **http://www.amstat.org/sections/SRMS/Proceedings/y2001/Proceed/00276.pdf** (data di ultima consultazione: Marzo 2016), pp. 2-3. [↑](#footnote-ref-49)
48. Come ALLISON P., *Logistic regression using SAS: theory and application,* cit., KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, HOSMER D. – LEMESHOW S., *Applied logistic regression*, cit. [↑](#footnote-ref-50)
49. CAFFO B., *Regression models for data science in R*, cit., pp. 96-97. [↑](#footnote-ref-51)
50. *Ivi*, p.100: “*Omitting variables results in bias in the coefficients of interest - unless the regressors are uncorrelated with the omitted ones*”. nDp. 35.erale68.amata la su (Internet), disponibile su isponibile su di apprendimento statistico, [↑](#footnote-ref-52)
51. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, p. 271. [↑](#footnote-ref-53)
52. PARDOE I. – COOK D., *Op. cit.*, p1. [↑](#footnote-ref-54)
53. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, p. 271. [↑](#footnote-ref-55)
54. *Ivi*, p. 273. [↑](#footnote-ref-56)
55. CURINI L., *Op. cit.*, p. 5. [↑](#footnote-ref-57)
56. Il dataset utilizzato nel capitolo terzo contiene molte variabili quantitative. [↑](#footnote-ref-58)
57. KLEINBAUM D – KLEIN MITCHEL, *Op. cit.*, p. 275. [↑](#footnote-ref-59)
58. KOVAL J. et alia, *Assessing the impact of potentially influential observations in weighted logistic regression* (Internet**),** disponibile su **http://www.statcan.gc.ca/pub/12-002-x/2015001/article/14147-eng.htm** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-60)
59. Penn State University, *Analysis of discrete data* (Internet), disponibile su **https://onlinecourses.science.psu.edu/stat504/node/161** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-61)
60. KUHN M. – JOHNSON K., *Applied predictive modeling*, New York, 2013, p. 2. [↑](#footnote-ref-62)
61. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., pp. 16-42. [↑](#footnote-ref-63)
62. *Ivi*, p. 18. [↑](#footnote-ref-65)
63. si può dimostrare, cfr. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., p.19, che questo limite coincide con la varianza di . [↑](#footnote-ref-66)
64. KUHN M. – JOHNSON K., *Op. cit.*, New York, 2013, p.4 [↑](#footnote-ref-67)
65. SHMUELI G., *To explain or to predict?* (Internet), disponibile su **https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-68)
66. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., p. 23. Un esempio di modello non-parametrico è il KNN (*K-nearest neighbors*), che data una osservazione di cu deve predire la variabile di risposta associata, cerca nel campione disponibile le k osservazioni “più simili” e da quelle deduce la distribuzione condizionata di dato . Cfr. HASTIE T – TIBSHIRANI R – FRIEDMANJ, *Op. cit.*, p 14 ss. [↑](#footnote-ref-70)
67. Vedi il paragrafo 4 del capitolo primo per la distinzione tra i due concetti. [↑](#footnote-ref-71)
68. BORRA S. – DI CIACCIO A., *Op. cit*. (ebook), capitolo 19 p. 15. [↑](#footnote-ref-72)
69. Vedi più oltre, paragrafo 7.1. [↑](#footnote-ref-73)
70. Vedi paragrafo del presente capitolo. [↑](#footnote-ref-74)
71. Quella che in MITCHELL T., *Machine learning*, New York, 1997, p.5, viene chiamata la la *training experience*. [↑](#footnote-ref-75)
72. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit.*, p. 34. [↑](#footnote-ref-76)
73. *Ivi*, p. 47. [↑](#footnote-ref-77)
74. BABAYAK M., *What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models* (Internet), disponibile su **http://people.duke.edu/~mababyak/papers/babyakregression.pdf** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-78)
75. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit.*, p. 24. [↑](#footnote-ref-79)
76. Cfr. BISHOP C., *Pattern recognition and machine learning*, New York, 2006, p.3. [↑](#footnote-ref-80)
77. BORRA S. – DI CIACCIO A., *Op. cit.*, p.268. [↑](#footnote-ref-81)
78. KUHN M. – JOHNSON K., *Op. cit.*, p. 97. [↑](#footnote-ref-82)
79. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit.*, p. 35. [↑](#footnote-ref-83)
80. Denominazione generica che si ritrova in tutte le opere sull’argomento citate in questo elaborato. [↑](#footnote-ref-84)
81. CLIMBER R. – McCOLLOUGH B., *Fundamentals of predictive analytics with JMP*, Cary, 2013. Capitolo 10. [↑](#footnote-ref-85)
82. HASTIE T – TIBSHIRANI R – FRIEDMANJ., *Op. cit.*, p. 222. [↑](#footnote-ref-86)
83. *Ibidem*. [↑](#footnote-ref-87)
84. MITCHELL T., *Op. cit.*, p. 111. [↑](#footnote-ref-88)
85. NG. A, *Preventing overfitting of cross. Validation data* (Internet), disponibile su **http://ai.stanford.edu/~ang/papers/cv-final.pdf** (data di ultima consultazione: Marzo 2016), p.2 [↑](#footnote-ref-89)
86. Per una dimostrazione grafica con dati simulati cfr. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., p. 178. [↑](#footnote-ref-90)
87. KUHN M. – JOHNSON K., *Op. cit*., p. 68. [↑](#footnote-ref-91)
88. *Ivi*, p. 70. [↑](#footnote-ref-92)
89. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., p. 181. [↑](#footnote-ref-93)
90. Per questo in JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., p.33, si riduce la *cross-validation* a un problema di minimo locale, anche se la stima non è accurata. [↑](#footnote-ref-94)
91. *Ivi*, p. 249. [↑](#footnote-ref-95)
92. Cfr. KLEINBAUM D. – KLEIN M., *Op.cit.*, p. 259, in cui si sottolinea che alla fine la composizione delle variabili deve avere sempre senso per lo studioso della disciplina, approccio che verrà seguito in questo elaborato. [↑](#footnote-ref-96)
93. Cfr. CRAWLEY M., *Statistics. An introduction using R*, New York, 2014.p.8 e CAFFO B., *Regression models for data science in R*, cit., p.100. [↑](#footnote-ref-97)
94. PENG R. – MATSUI E., *The art of data science. A guide for anyone who works with data*, Baltimore, 2015, pp. 112-113. [↑](#footnote-ref-98)
95. Esistono anche tecniche avanzate di *dimensionality reduction* che mirano a ridurre il numero di variabili non riducendo la quantità di informazione utile al modello, cfr. GHODSI A., *Dimensionality reduction, a short tutorial* (Internet), disponibile su **http://www.math.uwaterloo.ca/~aghodsib/courses/f06stat890/readings/tutorial\_stat890.pdf** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-99)
96. Cfr. PENG R, *Exploratory data analysis with R*, Baltimore, 2015, pp. 1-2. [↑](#footnote-ref-100)
97. *Ivi*, pp. 41-58. Se ne farà massicciamente uso nel capitolo terzo. [↑](#footnote-ref-101)
98. ANDERSON R., *The credit scoring toolkit. Theory and practice for retail credit risk management and decision automation*, Oxford, 2007, p. 183. [↑](#footnote-ref-102)
99. Per la trattazione dei metodi stepwise cfr. JAMES J. – WITTEN D. – HASTIE T. – TIBSHIRANI R., *Op. cit*., pp. 207-210, da cui si trae notevole spunto. [↑](#footnote-ref-104)
100. Può ben darsi che venga selezionata come prima variabile perché da sola dà il contributo maggiore, ma che la miglior combinazione sia . [↑](#footnote-ref-105)
101. Resta inteso che il test di ipotesi su ogni singolo coefficiente è una prassi fondamentale della selezione delle variabili, ma che anch’esso risente degli effetti dell’*adjustment* e degli effetti sul FWER dei test multipli, cfr. HASTIE T – TIBSHIRANI R – FRIEDMANJ., *Op. cit.*, p. 60. [↑](#footnote-ref-106)
102. Cfr. GUYON I. – ELISSEFF A., *An introduction to variable and feature selection* (Internet), in “Journal of machine learning”, 2003, pp. 1157-1182 disponibile su **http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-107)
103. HASTIE T – TIBSHIRANI R – FRIEDMANJ., *Op. cit.*, pp. 245-246. L’analisi esplorativa iniziale invece può essere fatta su tutto il dataset, dato che non produce misure di performance del modello che potrebbero essere *biased*. [↑](#footnote-ref-108)
104. KUHN M. – JOHNSON K., *Op. cit*., p. 254. [↑](#footnote-ref-109)
105. Per una completa lista di indicatori legati alla matrice di confusione cfr. **http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/** e **https://en.wikipedia.org/wiki/Sensitivity\_and\_specificity**. [↑](#footnote-ref-110)
106. Si usa la denominazione inglese di questi indicatori perché ampiamente pù diffusa della traduzione italiana. [↑](#footnote-ref-111)
107. Che in statistica bayesiana è la probabilità a priori. [↑](#footnote-ref-112)
108. Vedi esempi in vedi CAFFO B., *Statistical inference for data science*, cit., p. 25 e KUHN M. – JOHNSON K., *Op. cit*., p. 255. [↑](#footnote-ref-113)
109. Alcuni (cfr. KUHN M. – JOHNSON K., *Op. cit*., p. 257) usano anche il *J index*, che si calcola sottraendo una unità alla somma di TPR e TNR. [↑](#footnote-ref-114)
110. Pensiamo ad un processo di business dove nel caso in cui un modello contro le frodi etichetti una transazione elettronica come positiva il cliente subirà ritardi nella definizione dell’acquisto, anche se poi si riveli un non-evento; oppure a un test di screening per una malattia grave in cui il modello etichetta un paziente come positivo inducendolo a un sovraccarico emotivo notevole nell’attesa degliesami anche se poi si rivelerà come non-evento. [↑](#footnote-ref-115)
111. In statistica bayesiana la probabilità a posteriori. [↑](#footnote-ref-117)
112. CAFFO B., *Statistical inference for data science*, cit., p. 19. [↑](#footnote-ref-118)
113. Ibidem. [↑](#footnote-ref-119)
114. Il PDLR permette anche di valutare nella giusta luce una *precision* bassa dovuta a una *prevalence* estremamente bassa, in quanto mette in rapporto le due grandezze, cfr. KUHN M. – JOHNSON K., *Op. cit.*, p. 258. [↑](#footnote-ref-121)
115. ANDERSON R., *Op. cit.*, p. 206. [↑](#footnote-ref-122)
116. KUHN M. – JOHNSON K., *Op. cit.*, p. 262. [↑](#footnote-ref-123)
117. ANDERSON R., *Op. cit.*, p. 207. Una dimostrazione molto semplice, per cui non si fanno riferimenti bibliografici in quanto sviluppata dall’autore, è la seguente. Sia la numerosità del campione, quella di un sottocampione estratto casualmente come insieme delle unità a cui casualmente sono state assegnate le probabilità più alte di essere un evento. Siano gli eventi presenti nel campione, e il numero di eventi, allora: , ma anche . [↑](#footnote-ref-124)
118. *Ibidem*. [↑](#footnote-ref-125)
119. *Ibidem*. [↑](#footnote-ref-127)
120. HOSMER D. – LEMESHOW S., *Op. cit.*, paragrafo 5.2.4. già vista nel capitolo primotti.lo. plotone italiana. trae notevole spunti. ndimento statistico, [↑](#footnote-ref-128)
121. Ad esempio l’ score oppure il *Fowlkes–Mallows index* *score*, cfr. **https://en.wikipedia.org/wiki/F1\_score**. [↑](#footnote-ref-129)
122. KUHN M. – JOHNSON K., *Op. cit*., p. 265. [↑](#footnote-ref-130)
123. VUK M. – CURK T., *ROC curve, lift chart and calibration plot* (Internet), disponibile su **http://mrvar.fdv.uni-lj.si/pub/mz/mz3.1/vuk.pdf** (data di ultima consultazione: Marzo 2016). [↑](#footnote-ref-131)