



StatSoft®

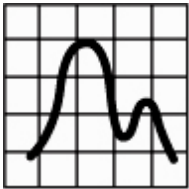
data analysis • data mining • quality control • web-based analytics

Istituti di Credito

e

STATISTICA

Case Study: Credit Scoring



STATISTICA
Soluzioni per Business Intelligence,
Data Mining, Quality Control, e
Web-based Analytics

Tabella dei Contenuti

INTRODUZIONE: COS'È IL “CREDIT SCORING”?	3
CREDIT SCORING: OBIETTIVI DI BUSINESS	3
1. Aspetti di Marketing	3
Obiettivi.....	3
Esempi.....	4
2. Aspetti Applicativi	4
Obiettivi.....	4
Esempi.....	4
3. Aspetti Prestazionali	4
Obiettivi.....	4
Esempi.....	4
4. Gestione del Credito “Cattivo”	4
Obiettivi.....	4
Esempi.....	5
CASE STUDY: CREDIT SCORING SULLA CLIENTELA	5
Descrizione.....	5
File Dati.....	5
ANALISI DEI DATI CON STATISTICA	6
Preparazione dei Dati	6
Selezione delle Caratteristiche	7
Spazio di Lavoro di <i>STATISTICA Data Miner</i>	8
ANALISI DEI RISULTATI	9
Alberi Decisionali - <i>CHAID</i>	9
Matrice di Classificazione – Modello <i>CHAID</i>	10
VALUTAZIONE COMPARATIVA DEI MODELLI	11
Gain Chart.....	11
Lift Chart.....	12
Matrice di Classificazione – <i>Alberi Boosted</i>	13
DEPLOYMENT DEL MODELLO	14
CONCLUSIONI	14

Introduzione: Cos'è il “Credit Scoring”?

Nel settore finanziario, è assai frequente la situazione in cui i clienti richiedono prestiti per finanziare i propri acquisti. Il livello di rischio che corrono gli istituti di credito nel concedere prestiti dipende dalla capacità degli istituti stessi di distinguere tra clienti buoni e clienti cattivi (cioè a cattivo rischio di credito). Una tecnica ampiamente adottata per la soluzione di questo problema è il “Credit Scoring”.

Il Credit Scoring è composto da un insieme di modelli decisionali e relative tecniche sottostanti con la funzione di aiutare i creditori a decidere quando concedere un prestito. Queste tecniche consentono di stabilire chi possa accedere al credito, il livello del prestito da concedere, e quali strategie operative adottare per migliorare la profittabilità dei clienti nei confronti del creditore. Queste tecniche consentono inoltre a stimare il livello di rischio associato al prestito. Il Credit Scoring è in tutto e per tutto una procedura di verifica dipendente dal grado di solvibilità della persona in quanto basata su dati reali.

Un creditore solitamente esegue due tipi di decisioni: per prima cosa, stabilisce se concedere il prestito ad un nuovo richiedente oppure no, e seconda cosa, decide come trattare i clienti a cui è già stato concesso un prestito, inclusa l'opportunità o meno di aumentare i relativi limiti di credito oppure no. In entrambi i casi, indipendentemente dalla tecnica utilizzata, è importante che sia disponibile un grande campione d'informazioni relative ai clienti passati, ai trend comportamentali, e alla successiva storia creditizia personale. Per identificare la connessione tra le caratteristiche dei clienti (reddito annuo, età, numero di anni lavorativi, ecc.) ed il livello di “bontà” della relativa storia creditizia successiva, la maggior parte delle tecniche disponibili richiedono l'impiego di questo tipo di campioni.

Credit Scoring: Obiettivi di Business

L'applicazione dei modelli di scoring nell'attuale contesto di business soddisfa un'ampia gamma di obiettivi. La funzione originale della stima del rischio di default è stata arricchita ed ampliata da una serie di modelli di credit scoring al fine di considerare altri aspetti inerenti alla gestione del rischio di credito: nella fase pre-applicativa (identificazione dei potenziali richiedenti), nella fase applicativa (identificazione dei richiedenti accettabili), e nella fase prestazionale (identificazione del possibile comportamento della clientela attuale). Per questo motivo sono stati sviluppati modelli di scoring da utilizzare in base ai differenti obiettivi. Essi possono essere generalizzati nelle seguenti quattro categorie:

1. Aspetti di Marketing

Obiettivi

- 1.1 Identificazione dei potenziali richiedenti che più facilmente potrebbero rispondere ad attività promozionali; questa operazione ha l'obiettivo di ridurre il costo di acquisizione di nuova clientela e di minimizzare l'insoddisfazione della clientela esistente.
- 1.2 Previsione della probabilità di perdere clienti facoltosi, consentendo in questo modo di formulare strategie efficaci di trattenimento della clientela.

Esempi

Stima di risposta. Modelli di scoring che stimano la probabilità che un cliente possa rispondere o meno ad un campagna promozionale diretta relativa ad un nuovo prodotto.

Stima di trattenimento/attrattività. Modelli di scoring che stimano la probabilità che un cliente possa continuare ad usare il prodotto oppure passare ad un nuovo creditore al termine di un periodo di attività.

2. Aspetti Applicativi

Obiettivi

- 2.1 Stabilire se estendere il credito, ed in quale misura.
- 2.2 Prevedere il comportamento futuro di un nuovo richiedente prevedendo la sua probabilità d'insolvenza

Esempi

Stime sui richiedenti. Modelli di scoring che stimano la probabilità che un cliente possa divenire insolvente.

3. Aspetti Prestazionali

Obiettivi

- 3.1 Prevedere il futuro comportamento creditizio dei debitori esistenti al fine d'identificare/isolare i clienti cattivi e dedicare maggiori sforzi ad una loro assistenza, in modo da ridurre la possibilità che questi in futuro divengano un problema per l'istituto.

Esempi

Stime dei comportamenti. Modelli di scoring che stimano i livelli di rischio dei debitori esistenti.

4. Gestione del Credito “Cattivo”

Obiettivi

- 4.1 Selezione di polizze ottimali che consentano di minimizzare i costi di amministrazione o di massimizzare l'ammontare di liquidità recuperabile dai clienti insolvibili.

Esempi

Modelli di scoring per i processi decisionali. Modelli di scoring che consentono di stabilire quando agire sui conti dei possibili insolventi e a quali delle molte tecniche alternative di recupero ricorrere per ottenere il maggior successo possibile.

Quindi, l'obiettivo generale del credit scoring non è solo determinare se il richiedente si rivelerà un buon cliente, ma anche attrarre richiedenti di qualità che possano essere in seguito fidelizzati e controllati sull'intero periodo di profittabilità dei rispettivi portafogli generali.

Case Study: Credit Scoring sulla Clientela

Descrizione

Nelle maggior parte delle applicazioni creditizie, le banche sono interessate a capire se un cliente sarà in grado di restituire un prestito o meno. L'obiettivo di questo studio è modellare o prevedere la probabilità che un cliente che richiede un prestito possa venire categorizzato come buono o cattivo.

In questo case study verrà illustrato come realizzare un modello di credit scoring usando *STATISTICA Data Miner* per identificare gli input o i predittori che discriminano i clienti "rischiosi" dagli altri (sulla base di trend osservati sui clienti precedenti), identificare le tecniche predittive che funzionano bene sui dati di test, e successivamente eseguire il deployment di tali modelli al fine di prevedere il livello di rischio dei nuovi clienti.

File Dati

L'insieme di dati d'esempio usato in questo caso, *CreditScoring.sta*, contiene 1000 casi e 20 variabili (o predittori) contenenti informazioni riguardanti i clienti passati e correnti che hanno richiesto prestiti presso banche Tedesche (sorgente: http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html) per diverse ragioni. L'insieme di dati contiene informazioni relative alla condizione finanziaria dei clienti, ai motivi del prestito, allo stato occupazionale, ecc. Il file dati d'esempio si trova nella cartella dei dati d'esempio di *STATISTICA*.

Per ogni cliente, è disponibile un risultato binario relativo alla rispettiva "solubilità". Questa variabile contiene la voce *Good o Bad* a seconda che il prestito del cliente sia stato restituito oppure no. L'insieme di dati è composto da un 70% di clienti buoni e da un rimanente 30% di clienti cattivi. I clienti insolventi dopo oltre 90 giorni dalla data di pagamento sono considerati come clienti ad alto rischio, mentre i clienti che non hanno mancato neppure un pagamento sono considerati come clienti a basso rischio. Altre misure tradizionalmente considerate per la determinazione dei clienti buoni e dei clienti cattivi sono il numero di mesi trascorsi dalla scadenza del pagamento, la frequenza con cui viene "rimpiungato" il conto corrente, ecc.

Segue una lista completa di variabili utilizzate in questo insieme di dati:

Categoria	Variabili
1. Informazioni Personali di Base	Età, Sesso, Telefono, Lavoratore straniero
2. Informazioni Familiari	Stato civile, Numero di dipendenti
3. Informazioni Residenziali	Anni trascorsi presso l'indirizzo attuale, Tipo di appartamento
4. Condizione Lavorativa	Anni d'impiego, Occupazione
5. Condizione Finanziaria	Maggiori investimenti registrati, Ulteriori prestiti in corso, Stato del conto corrente, Numero di prestiti correnti presso la banca
6. Informazioni di Sicurezza	Valore dei risparmi o degli investimenti
7. Altro	Scopo del credito, Ammontare del credito in Deutsche Marks (DM)

In questo esempio, verrà analizzata la modalità in cui le variabili sopra elencate consentono di discriminare tra *Good* o *Bad Credit Standing*. Nel caso in cui sia possibile distinguere tra questi due gruppi, sarà quindi possibile usare il modello predittivo per classificare o prevedere nuovi casi lì dove siano disponibili le informazioni sopra indicate ma non si ha idea di quale sia la condizione creditizia della singola persona. Questo potrà rivelarsi utile, per esempio, per stabilire se qualificare un soggetto come adatto per ricevere un prestito.

Analisi dei Dati con STATISTICA

Preparazione dei Dati

Con *STATISTICA Data Miner*, è estremamente facile applicare ai dati potenti strumenti di modellazione e giudicare il valore dei modelli risultanti basati sui rispettivi valori predittivi e descrittivi. Ciò non diminuisce il ruolo assolutamente prezioso rappresentato dalla fase di preparazione preliminare dei dati. Dato infatti che le decisioni maggiormente strategiche saranno determinate da questi risultati, qualsiasi errore commesso in questa fase potrebbe comportare gravi perdite d'informazioni. Quindi, è decisamente importante riuscire a pre-elaborare i dati e migliorare l'accuratezza del modello in modo da poter raggiungere la migliore decisione possibile.

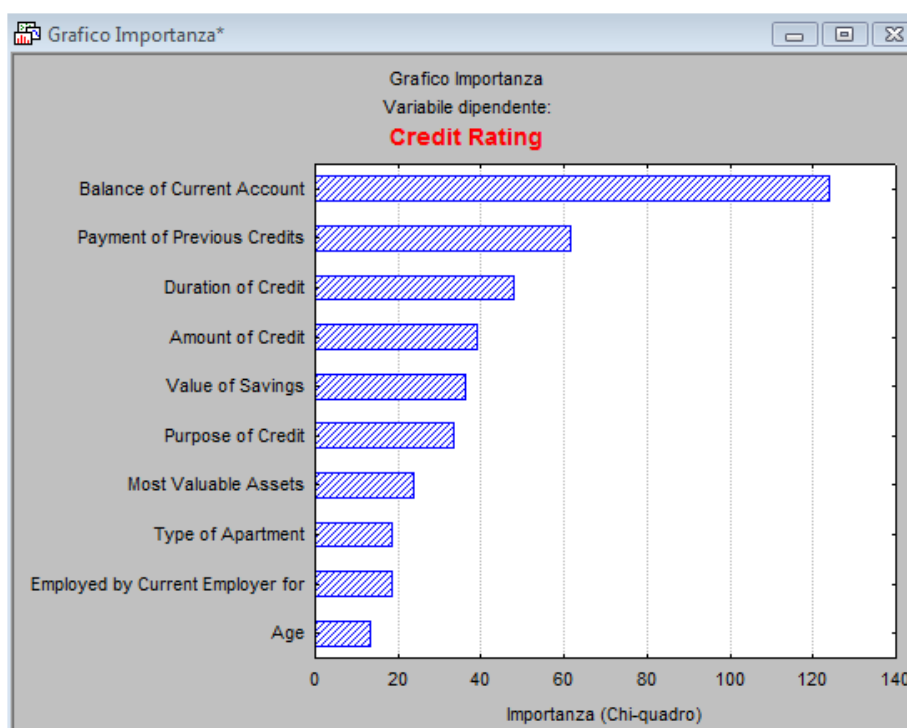
Durante questa fase sono state verificate le seguenti condizioni:

- Suggerimenti forniti dai dati: Statistiche descrittive (osservazione delle distribuzioni, delle medie, dei valori minimo e massimo, dei quartili, ecc.)
- Nei dati non sono presenti outlier
- Nei dati non sono presenti dati mancanti
- Non è richiesta alcuna trasformazione
- Selezione delle Caratteristiche – Variabili ridotte da 20 a 10

Selezione delle Caratteristiche

Per ridurre il livello di complessità del modello, l'insieme di dati può essere trasformato in un insieme di dati di dimensioni inferiori. Lo strumento di *Selezione di Caratteristiche e Screening di Variabili* disponibile in *STATISTICA Data Miner* ha automaticamente individuato i predittori più importanti che discriminano chiaramente tra buoni e cattivi clienti.

Il grafico a barre e lo spreadsheet d'importanza dei predittori forniscono suggerimenti circa quali siano i predittori più efficaci per la previsione della variabile dipendente. Per esempio, di seguito è riportato il grafico a barre relativo all'importanza dei predittori per la previsione della variabile dipendente "Credit Rating".



In questo caso, le variabili *Balance of current account*, *Payment of previous credits*, e *Duration in months* risultano come predittori più importanti.

Questi predittori saranno ulteriormente esaminati attraverso il ricorso ad un'ampia gamma di strumenti di data mining e di algoritmi di apprendimento automatico disponibili in *STATISTICA Data Miner* come ad esempio:

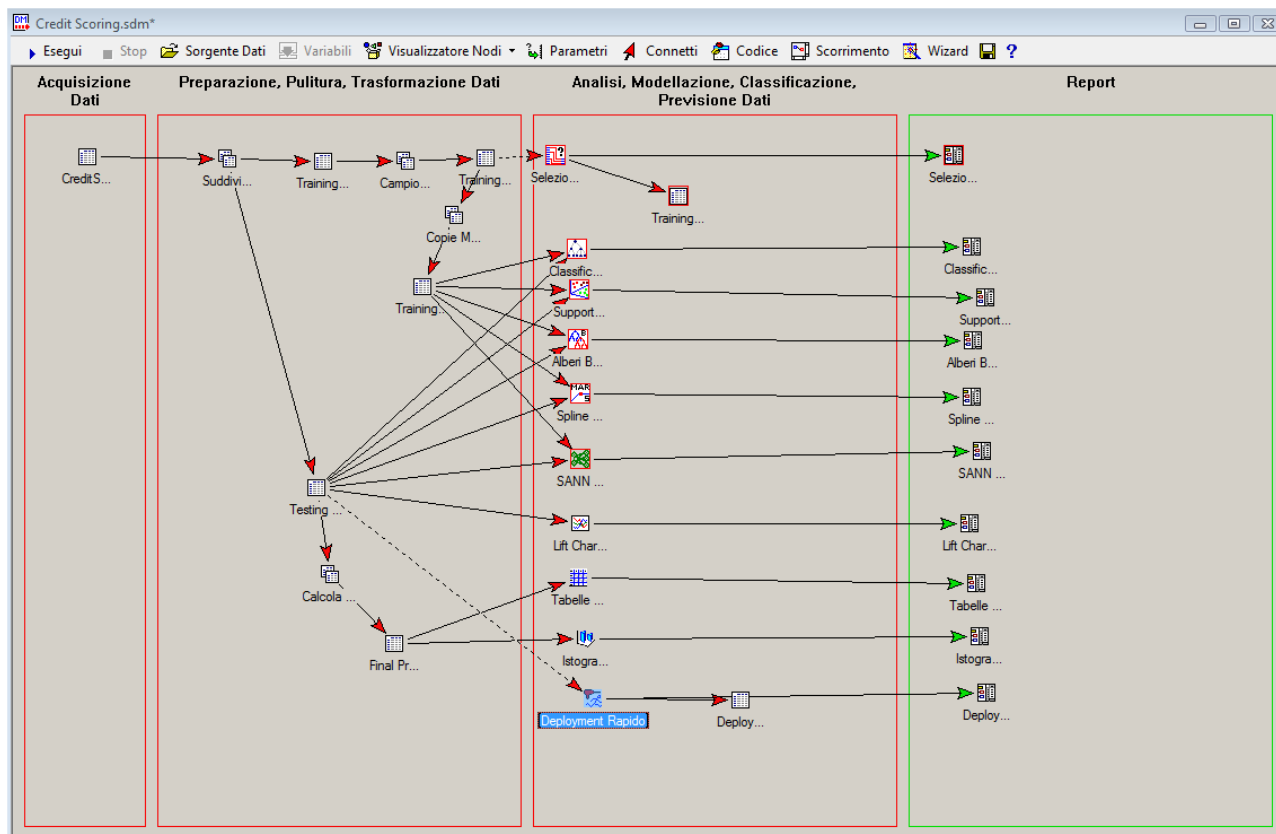
- *Alberi di Classificazione Standard con Deployment*
- *CHAID di Classificazione Standard con Deployment*
- *Alberi di Classificazione Boosted con Deployment*
- *Reti Neurali Automatizzate di STATISTICA con Deployment*
- *Support Vector Machine con Deployment (Classificazione)*
- *Spline MAR per Classificazione con Deployment*

La novità e l'abbondanza di tecniche e algoritmi disponibili nella fase di modellazione rendono questa fase la parte più interessante del processo di data mining. I metodi di classificazione sono le tecniche di data mining più comunemente utilizzati ed applicati nell'ambito del credit scoring, qualora si desideri prevedere il livello di rischio associabile ai clienti che richiedono un prestito. In più, è buona cosa praticare una serie di esperimenti applicando un numero di metodi differenti durante la fase di modellazione dei dati. Differenti tecniche potrebbero fare luce su un problema o confermare precedenti conclusioni.

STATISTICA Data Miner rappresenta un insieme completo e user-friendly di strumenti di data mining progettati per consentire agli utenti di analizzare facilmente e rapidamente i propri dati allo scopo di scoprire trend nascosti, di spiegare tendenze note, e di prevedere il futuro. Dall'interrogazione dei database e dalla segmentazione, fino alla generazione di report e grafici finali, questa piattaforma offre una facilità d'uso senza alcun sacrificio in termini di potenza e completezza. Inoltre, *STATISTICA Data Miner* offre la più grande selezione di algoritmi attualmente disponibili sul mercato per le operazioni di classificazione, di previsione, di analisi dei gruppi, e di modellazione, così come un'intuitiva interfaccia basata su icone. Sono disponibili semplici tecniche quali *C&RT* e *CHAID* così come tecniche più avanzate quali *Reti Neurali*, *Alberi Boosted*, *Foreste Casuali*, *Support Vector Machine*, *Spline MAR*, ecc.

Spazio di Lavoro di *STATISTICA Data Miner*

Lo spazio di lavoro di *Data Miner* funziona sulla base di un flusso di analisi; tutti gli strumenti di *STATISTICA Data Miner* sono disponibili in forma di icona gestibili attraverso semplici operazioni di trascina-e-incolla.



Di seguito sono riportate le descrizioni delle diverse fasi di preparazione dei dati e di flusso delle analisi:

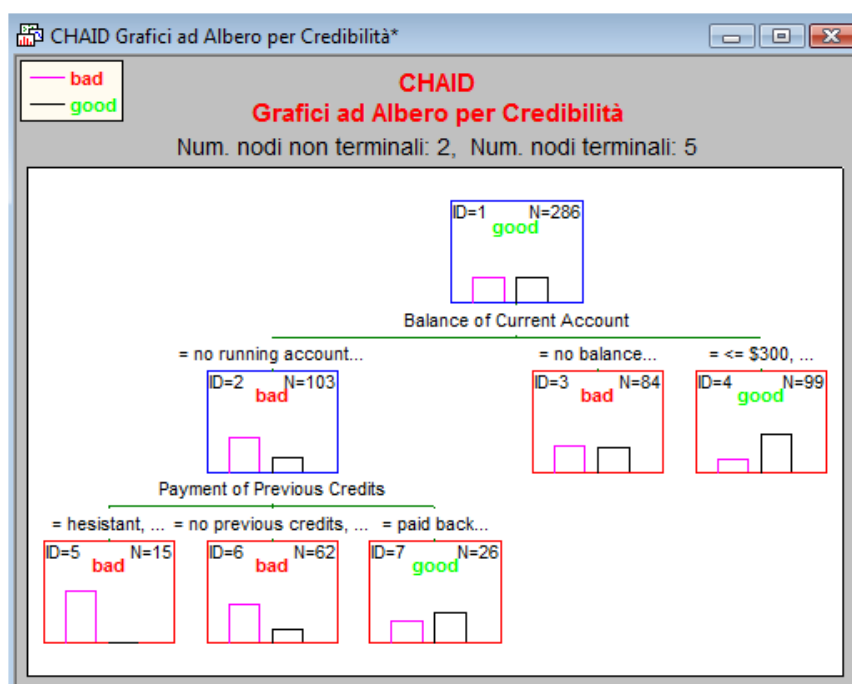
1. Suddivisione del file dati originale in due sottoinsiemi: il 34% dei casi viene trattenuto per scopi di verifica mentre il 66% dei casi viene utilizzato per la costruzione dei modelli.
2. Applicazione del Campionamento Casuale Stratificato per l'estrazione di proporzioni uguali di osservazioni corrispondenti a clienti ad alto rischio e a basso rischio.
3. Applicazione dello strumento di Selezione delle Caratteristiche per ordinare per importanza le migliori variabili predittrici da utilizzare per distinguere i clienti cattivi da quelli buoni.
4. Riduzione del numero di possibili predittori da 20 a 10 sulla base dei risultati della Selezione delle Caratteristiche.
5. Utilizzo di differenti Modelli Predittivi (algoritmi di Apprendimento Automatico) da utilizzare per individuare e comprendere eventuali relazioni esistenti.
6. Uso di strumenti comparativi quali le Lift Chart, Gain Chart, le tabelle incrociate, ecc., per individuare il miglior modello da applicare per il medesimo obiettivo di analisi.
7. Applicazione del modello all'Insieme di Test (campione conservato a parte) per la validazione della capacità/accuratezza predittiva.

Analisi dei Risultati

In questa sede verranno illustrati i risultati dell'analisi per meglio comprendere le caratteristiche dei clienti buoni e dei clienti cattivi. Per prima cosa si osservino i risultati degli alberi decisionali *CHAID*.

Alberi Decisionali - CHAID

Gli alberi decisionali sono strumenti potenti e molto popolari nel campo della classificazione e della previsione. Il fatto che gli alberi decisionali possano essere facilmente leggibili grazie al loro accattivante aspetto grafico li rende particolarmente facili da interpretare.



Si noti che i risultati ottenuti con il proprio computer potrebbero essere differenti rispetto a quelli riportati in questa sede in virtù dei differenti risultati del campionamento e della suddivisione randomizzata del campione originario in dati di test e di addestramento. Tuttavia, in generale, i risultati dovrebbero essere gli stessi in termini di suddivisione delle variabili principali e di tipi di suddivisioni riportate nel grafico precedente.

Dall'osservazione dell'albero qui riportato, è possibile verificare come l'algoritmo *CHAID* abbia creato un albero con 5 nodi terminali (evidenziati in rosso), come risultato di 2 condizioni *se-allora* applicate a livello di suddivisione. I nodi terminali (o foglie terminali, come talvolta vengono chiamati) sono quelli dopo i quali non è possibile applicare alcun'altra suddivisione ai fini del miglioramento dell'accuratezza predittiva della soluzione finale (dati i parametri correnti selezionati per guidare il processo di costruzione degli alberi). L'albero parte con il nodo decisionale iniziale (anche noto come nodo radice) contenente 286 casi nell'insieme di addestramento con proporzioni pressoché identiche di clienti classificati come "buoni" e come "cattivi" e ottenute attraverso l'uso dello strumento di Campionamento Casuale Stratificato. La legenda che identifica quali barre negli istogrammi di nodo corrispondono alle due categorie è posizionata nell'angolo in alto a sinistra del grafico.

L'interpretazione dell'albero è piuttosto semplice. Il nodo più a destra è il risultato della prima suddivisione, e contiene 99 istanze per la maggior parte associate a clienti *buoni*. Dato che ulteriori suddivisioni da questo punto in poi non aiuterebbero a migliorare l'accuratezza predittiva del modello (in base alle opzioni impostate), questo nodo diverrà "terminale". Il nodo più a sinistra contiene 103 istanze, e viene suddiviso ulteriormente sulla base del predittore *Payment of previous Credits*, producendo altri tre nodi.

Alla fine le "regole decisionali" potranno essere generate proprio sulla base di questa serie di suddivisioni. Una sua formulazione alternativa può quindi essere:

IF Balance of current account => no running account, no balance
AND Payment of previous credits => hesitant, problematic running accounts
THEN Credibility = "bad"

Tradotto in Italiano

SE Stato del conto corrente => nessun conto aperto, in rosso
E Restituzione di prestiti precedenti => esitante, nessun prestito concesso in precedenza
ALLORA Credibilità = "cattiva"

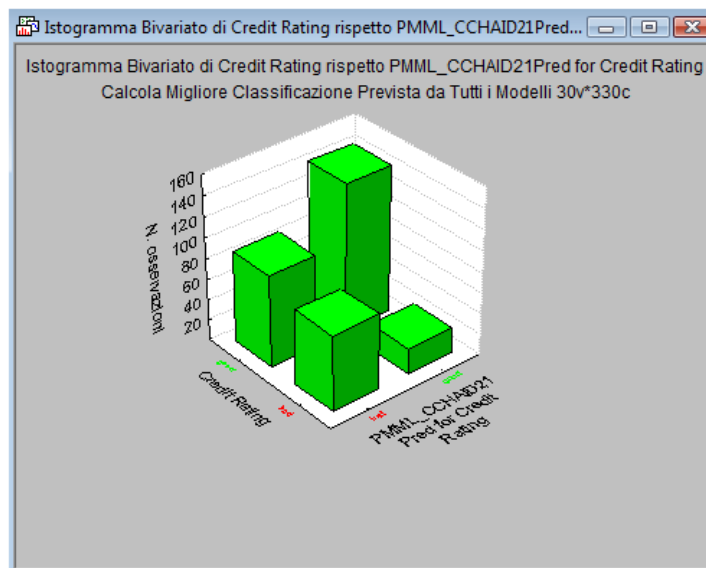
Matrice di Classificazione – Modello *CHAID*

Attraverso la *Matrice di classificazione* è possibile confrontare le classificazioni osservate con quelle previste, allo scopo di ottenere un riepilogo relativo alla specifica accuratezza di classificazione (tassi di errata classificazione) in corrispondenza delle differenti categorie di output.

Dati: Tabella Sunto Frequenze (Calcola Migliore Classificazione Prevista ...)

Tabella Sunto Frequenze (Calcola Migliore Classificazione Prevista da Tutti i Modelli 30v*330c)
 Marcate le celle con cont. > 10
 (Riassunti marginali non marcati)

	Credit Rating	PMML_CCHAIID 21Pred for Credit Rating bad	PMML_CCHAIID 21Pred for Credit Rating good	Totali Riga	Percent Correct
valori previsti					
1	bad	73	24	97	75,26%
2	good	89	144	233	61,80%
3	Tutti	162	168	330	65,76%



La matrice di classificazione riporta il numero di casi correttamente classificati (presenti cioè sulla diagonale maggiore della matrice) e di quelli classificati in modo errato ed associati ad altre categorie.

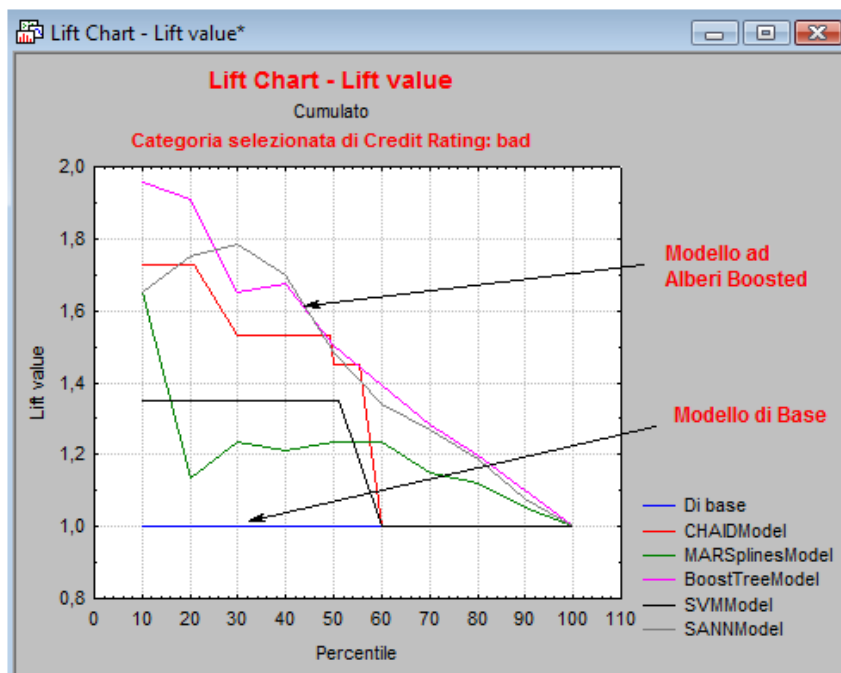
In questo caso, il modello generale può prevedere correttamente se lo stato creditizio di un cliente è da considerarsi buono o cattivo con un'accuratezza del 65,76% $(73 + 144)/(73 + 24 + 89 + 144)$. Si noti che l'obiettivo principale è ridurre la proporzione di prestiti destinati a dimostrarsi cattivi e classificati come buoni. La percentuale di previsioni corrette per la categoria "bad" è 75,26%.

Valutazione Comparativa dei Modelli

È sempre buona cosa eseguire esperimenti con un numero differente di modelli piuttosto che fare affidamento su un singolo modello da utilizzare per il deployment finale. Differenti tecniche potrebbero infatti fare nuova luce su un problema o confermare precedenti conclusioni. Gli Alberi Boosted ad esempio forniscono una percentuale di previsioni corrette per la categoria "bad" pari al 75,25%.

Gain Chart

La gain chart fornisce un riassunto visivo sull'utilità dell'informazione fornita da uno o più modelli statistici utilizzati per la previsione della variabile dipendente categoriale. In particolare, il grafico



Se si considera il 43-imo decile, si può concludere come si sia in presenza di un campione contenente all'incirca 1,6 volte il numero di clienti "bad" se confrontato con quanto avviene con il modello di base. In altre parole, i valori di lift ottenuti con il modello ad *Alberi Boosted con Deployment* è approssimativamente pari a 1,6.

Matrice di Classificazione - Alberi Boosted

Come per l'analisi *CHAID*, è possibile osservare la matrice di classificazione prodotta dal modello ad *Alberi Boosted*:

Dati: Tabella Sunto Frequenze (Calcola Migliore Classificazione P...				
Tabella Sunto Frequenze (Calcola Migliore Classificazione P... Marcate le celle con cont. > 10 (Riassunti marginali non marcati)				
	Credit Rating	PMML_CBTre s23Pred for Credit Rating bad	PMML_CBTre s23Pred for Credit Rating good	Percent Correct
valori osservati				
1	bad	65	32	67,01%
2	good	71	162	69,53%
3	Tutti	136	194	68,79%

La matrice di classificazione calcolata per i dati di test riporta un numero di casi correttamente classificati (sulla diagonale maggiore) e quelli classificati in modo errata sulla diagonale secondaria.

In questo caso, il modello generale potrebbe prevedere correttamente se lo stato creditizio del cliente si rivelerà buono o cattivo con un'accuratezza del 68,79%. Il principale obiettivo è ridurre la proporzione di clienti classificabili come cattivi. La percentuale di previsioni corrette per la categoria "bad" quando si utilizza un modello ad *Alberi Boosted* è pari al 67,01%.

Deployment del Modello

La fase finale implica l'utilizzo del modello migliore e la sua applicazione su nuovi dati con l'obiettivo di prevedere se un cliente sarà buono o cattivo. In questo caso, si eseguirà il deployment del modello *Alberi Boosted di Classificazione*, che è certamente tra i modelli che forniscono una più alta accuratezza predittiva sull'insieme di test. *STATISTICA* fornisce tutti gli strumenti necessari per eseguire il deployment dei modelli predittivi. Sarà semplicemente necessario salvare il codice PMML di deployment relativo al modello, e quindi caricarlo attraverso il nodo *Deployment Rapido* nello spazio di lavoro di *STATISTICA Data Miner* per prevedere (classificare) il rischio di credito associabile ai nuovi richiedenti.

Conclusioni

L'obiettivo di questo esempio è dimostrare quanto facile è l'addestramento e l'utilizzo dei modelli predittivi quando l'utente dispone di tutti gli strumenti necessari per capire cosa fare in ogni fase del processo di costruzione del modello. *STATISTICA* fornisce inoltre numerosi strumenti da applicare nella fase di *Preparazione/Pulitura dei Dati*. Le tecniche disponibili in *STATISTICA Data Miner* rappresentano alcune delle tecniche predittive più avanzate attualmente disponibili sul mercato. *STATISTICA Data Miner* offre una selezione molto ampia di grafici e diagrammi da poter combinare con tutte le altre funzionalità del programma, che consente all'analista di usare tecniche di "data mining visuale", o semplicemente di combinare tecniche esclusivamente visive (metodi grafici) da integrare con pochi clic nel progetto. Una volta finalizzato il modello, le soluzioni calcolate attraverso *STATISTICA Data Miner* potranno essere sottoposte a deployment sotto forma di progetti completi accessibili pressoché immediatamente.

