UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

Corso di Laurea in Statistica, Popolazione e Società

TESI DI LAUREA

La valutazione dell'efficacia di segmentazioni binarie per l'analisi del rischio di insolvenza

Relatore: Ch.mo Prof. Luigi Fabbris

Laureanda: Cristina Mannino

Matr. N. 498807

ANNO ACCADEMICO 2006-2007

Ringraziamenti

Se a più di 10 anni dal Diploma di Statistica mi sono ritrovata "sui libri", lo devo alla sete di conoscenza trasmessa dalle persone incontrate durante gli studi e nel corso delle esperienze professionali.
Pertanto, ringrazio il Professor Luigi Fabbris, che mi ha da sempre appassionato alla ricerca statistica e mi accompagna ora perla seconda volta in questa avventura.
Ringrazio CRIF, l'azienda per la quale lavoro. In particolare ringrazio il Dottor Davide Capuzzo, mio responsabile, per le opportunità che mi ha offerto e continua a propormi, e il Professor Jean Marie Bouroche, con cui ho avuto la fortuna di collaborare e che tanto mi ha insegnato.

Un abbraccio di cuore lo dedico alla mia famiglia, unica e grande!
Un pensiero speciale ad Alessandro.

Indice dei contenuti

INTRODUZIONE

L'esigenza di disporre di sistemi di valutazione idonei ad apprezzare la "credibilità" delle controparti è da sempre al centro delle problematiche che gli istituti bancari e finanziari si trovano a dover affrontare nella propria quotidiana attività di erogazione e di gestione del credito.

Il progressivo sviluppo dei mercati e la rapida diffusione di nuovi canali di accesso al credito se, da un lato, continuano ad offrire nuove opportunità di accrescimento della clientela di base, dall'altro, implicano più accentuate sfide competitive in contesti in cui i profili di rischio sono più complessi, più interconnessi e quindi più difficili da valutare e da controllare.

Muovendo dalla maturata la convinzione che i rischi vadano affrontati secondo una prospettiva nuova, più rigorosa e strutturata, la presente ricerca si propone di delineare l'ambito del problema della valutazione del rischio di insolvenza all'interno del mondo bancario e finanziario (Capitolo 1), di illustrare le caratteristiche del campione di dati utilizzati nelle analisi di segmentazione (Capitolo 2), di descrivere quindi le caratteristiche generali dei metodi di segmentazione che permettono di formulare una risposta nella fase di esplorazione di gruppi con forte interazione e di individuazione delle determinanti del rischio (Capitolo 3) e, infine, di presentare e di valutare i principali risultati ottenuti nella stima del rischio di insolvenza (Capitoli 4 e 5) attraverso l'applicazione dell'analisi di segmentazione ad un campione rappresentativo della clientela privata del sistema bancario. Il Capitolo 6 ripercorrere le soluzioni individuate in una riflessione conclusiva suggerendo nel contempo alcune possibili linee guida e gli approfondimenti che meritano di essere sviluppati nella trattazione di analoghe problematiche.

CAPITOLO 1

Il problema della valutazione del rischio di insolvenza

Dopo una breve introduzione al problema della valutazione del rischio di insolvenza (Paragrafo 1.1), nel presente capitolo si presenta l'evoluzione dell'approccio valutativo dalla dimensione di singolo prodotto di credito verso la dimensione complessiva di cliente (Paragrafo 1.2), le principali funzionalità dei modelli di valutazione del rischio all'interno dei processi decisionali di erogazione e di gestione del credito (Paragrafo 1.3) e le argomentazioni che conducono a ritenere la determinazione di gruppi caratterizzati da forti interazioni come il passaggio cruciale che le analisi esplorative preliminari finalizzate allo sviluppo di sistemi di valutazione del rischio devono risolvere (Paragrafo 1.4).

1.1 Il problema di "dar credito"

Nel 1955 la rivista "The Lending Banker" pubblicò il seguente commento di L.C.Mather, direttore della Midland Bank:

"The art of banking is surely to know when to accept the risk.
But first the able banker must be able to appreciate and asses that risk"

Dal punto di vista dell'istituto di credito, il problema del rischio di insolvenza è, in sostanza, l'incognita relativa al comportamento di pagamento futuro da parte della controparte finanziaria cui è stato concesso un credito.

Il dover capire quale sia il rischio di insolvenza e, più in particolare,

quali siano le effettive componenti di rischio costituisce il primo passo da risolvere prima di poter prendere una decisione consapevole in merito all'assunzione del rischio stesso (Bailey, 2004).

Storicamente, i primi modelli statistici in grado di fornire una soluzione al problema della valutazione del rischio, attraverso una quantificazione oggettiva della probabilità di inadempimento degli impegni di pagamento da parte dei clienti, nacquero e si svilupparono nel mondo anglosassone a partire dagli anni '60 quando, con l'arrivo e la rapida diffusione delle carte di credito, divenne evidente che una valutazione del rischio basata semplicemente sul giudizio umano non permetteva di soddisfare i requisiti di efficacia ed efficienza dettati da un mercato che andava espandendosi e da un numero di clienti che andava crescendo.

Negli anni '80, visto il successo ottenuto con le carte di credito, i sistemi automatizzati di valutazione del rischio furono adottati anche nella gestione di altri prodotti creditizi, quali prestiti e mutui[1].

Da allora, l'esigenza di sviluppare e di adottare modelli affidabili in grado di valutare preventivamente il rischio di insolvenza per coadiuvare nella decisione le persone addette a "dare credito" si è evidenziata con maggiore enfasi.

In tempi più recenti, inoltre, la crescente disponibilità di informazioni creditizie e la diffusione di soluzioni IT in grado di aggiornare e di analizzare in modo efficiente grandi quantità di dati hanno ulteriormente favorito sia l'evoluzione dei modelli di valutazione del rischio di insolvenza, sia l'industrializzazione dei processi di erogazione del credito e di gestione del portafoglio clienti, non solo nel sistema bancario, ma anche in altri tipi di imprese, quali le società di telecomunicazioni e gli istituti assicurativi, che sono chiamate a gestire le conseguenze, non solo di eventi già verificati, bensì in modo proattivo, di eventi che si possono ancora manifestare.

[1] I prodotti creditizi che tradizionalmente compongono la gamma di offerta verso la clientela privata sono rappresentati da mutui, prestiti, carte di credito e fidi di conto.

1.2 La visione cliente-centrica

Negli istituti bancari e finanziari caratterizzati da una ampia e diversificata offerta di prodotti creditizi, la struttura gestionale si presenta spesso suddivisa per tipologia di prodotto ed è quindi parcellizzata in centri di competenza indipendenti, sia nell'attività di erogazione, sia in quella di gestione della specifica tipologia di forma creditizia (Neves, 2004).

In molte di queste realtà operative, dove anche l'organizzazione della struttura informativa prevede delle banche dati "separate" e "non comunicanti" fra loro, i sistemi decisionali sono necessariamente progettati per soddisfare le esigenze valutative e per raggiungere gli obiettivi di sviluppo della singola categoria di prodotti: il tipico scenario è rappresentato da situazioni in cui, mancando l'accessibilità e l'integrazione con informazioni esterne e di carattere più generale, la soluzione al problema della valutazione del rischio resta circoscritta allo specifico prodotto in esame.

In queste circostanze, quantunque il rischio di insolvenza dei singoli prodotti sia conosciuto e gestito in modo adeguato, la mancanza di informazioni in merito alla rischiosità complessiva del cliente comporta che spesso ne venga realizzata una gestione poco coerente all'interno delle diverse parti dell'organizzazione e che, quindi, ci siano delle ricadute negative sulla relazione fra l'istituto di credito e il cliente stesso.

L'evoluzione del mercato creditizio, la sempre maggiore concorrenza fra gli operatori e l'innalzamento dei tassi di indebitamento pro capite sono alcuni fra i fattori che hanno contribuito a far sì che l'attenzione si vada spostando dal singolo prodotto verso il cliente nel suo insieme, secondo una visione che lo pone come elemento centrale da valutare e con cui interagire.

Da tempo, la maggior parte degli istituti bancari e finanziari ha riconosciuto l'esigenza di dover superare le difficoltà derivanti da una valutazione "locale" del rischio e quindi di intraprendere, nel rispetto delle normative che disciplinano il trasferimento e l'utilizzo di dati personali, una efficace condivisione delle informazioni, sia al livello dello stesso istituto

(dando origine allo sviluppo dei *"customer database"* aziendali), sia al più alto livello di sistema bancario e finanziario (contribuendo le proprie informazioni ai Sistemi di Informazioni Creditizie, noti come SIC).

Secondo questa nuova visione, perseguita anche nelle analisi presentate nel presente studio, sono andati sviluppandosi dei modelli di valutazione del rischio di insolvenza in cui le informazioni relative ai diversi prodotti di credito posseduti dal cliente non vengono analizzate separatamente e fini a se stesse, ma possono essere aggregate aumentando il proprio valore informativo nel momento in cui convergono a definire un portafoglio in base al quale è possibile desumere un quadro complessivo della tipologia di controparte in termini, soprattutto, di rischiosità creditizia.

1.3 I modelli di valutazione del rischio di insolvenza

Nel sistema bancario, il processo di quantificazione del rischio di insolvenza si realizza tramite modelli di valutazione che, sulla base di un insieme di variabili precedentemente identificate come esplicative, attribuiscono la clientela a differenti classi di rischio definite in base alle stime di probabilità d'insolvenza.

Muovendo dal principio secondo cui "il passato predice il futuro", tali modelli, applicati alla clientela di cui si ignora il rischio futuro, si prefiggono di stimarne la probabilità di insolvenza sulla base delle variabili che preventivamente, su un analogo campione di clienti di cui era noto il rischio di insolvenza, sono identificate come predittive.

Sia dato un insieme di *n* clienti su cui si osservano una variabile dipendente Y e un certo numero *k* di variabili esplicative X, formalmente si definisce il rischio noto di insolvenza per il cliente *i-esimo* con:

$$Y_i \qquad (i = 1, \dots, n),$$

dove Y può essere misurato su qualsiasi scala, più frequentemente di tipo binario (ad esempio con esito dicotomico "solvente" - "insolvente") o di tipo

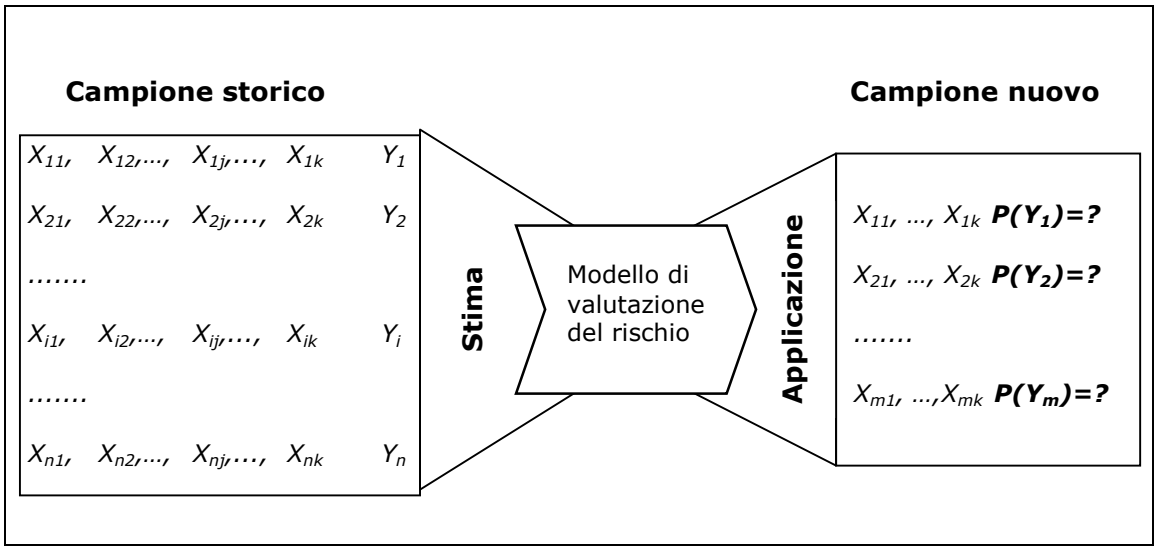
ordinale ("nessun rischio", "rischio basso", "rischio medio", "rischio alto"), e il valore della *j-esima* variabile predittiva sul cliente *i-esimo* con:

X_{ij} $(i = 1,\ldots,n;\ j = 1, \ldots k)$.

Inoltre, si supponga di aver stimato un modello di valutazione del rischio con l'esplicito scopo di riuscire a distinguere e classificare differenti gruppi di clienti a partire dal campione originale.

La regola di classificazione individuata può essere impiegata per raggruppare (nei gruppi definiti dalla variabile dipendente Y) nuove unità statistiche di cui sia ignota la variabile dipendente, ma di cui si conoscano le determinazioni delle variabili predittive, ovvero di stimarne la probabilità di insolvenza futura. Nella prassi operativa di gestione del credito, i modelli di valutazione del rischio sono costantemente utilizzati su nuovi campioni di clienti al fine di stimarne l'ignoto valore della probabilità di Y (Figura 1.1).

Figura 1.1 *Schematizzazione del processo di stima (sulla base di un campione storico di clienti) e di successiva applicazione (ad un nuovo campione di clienti) di un modello di valutazione del rischio di insolvenza*



In origine, la valutazione del rischio di insolvenza avveniva soprattutto con l'ausilio di modelli "deduttivi", ovvero tramite modelli che si basavano su informazioni la cui rilevanza nell'intercettare il rischio era

determinata "in modo esperto" dalle funzioni preposte alla gestione del credito e che, quindi, risentivano fortemente del peso dell'esperienza soggettiva pregressa (Liu, 2002).

Per quanto tali strumenti siano sempre più destinati ad estinguersi a causa della loro limitata capacità di previsione, essi trovano ancora oggi applicazione in alcune realtà di giovane costituzione in cui il patrimonio informativo disponibile non è ancora sufficientemente consolidata ai fini dello sviluppo di modelli statistici.

Nel panorama dell'attuale sistema bancario, i sistemi di supporto alle decisioni prediligono e vedono integrati al loro interno vari algoritmi di valutazione del rischio di insolvenza che si qualificano per un più elevato grado di oggettività delle stime e per una loro maggiore stabilità nel tempo.

Si tratta, infatti, di modelli di tipo "empirico" le cui variabili predittive sono state individuate e selezionate per mezzo di tecniche statistiche, in seguito ad un rigoroso processo di analisi esplorativa svolto a partire da ampie e rappresentative informazioni caratterizzate, per altro, da una consistente profondità storica.

In virtù delle metodologie e delle procedure statistiche con cui sono derivati, i modelli statistici di valutazione del rischio rappresentano un indispensabile strumento strategico in grado di:

- Migliorare la qualità degli impieghi attraverso la valutazione dell'esplicito trade off fra rischi e rendimenti,
- Assicurare, a parità di condizioni, una maggiore uniformità e oggettività delle decisioni di concessione e di gestione del credito
- Ridurre i tempi e i costi associati alle decisioni di istruzione delle pratiche in richiesta, con conseguente opportunità per gli operatori di esaminare una più ampia platea di clienti
- Razionalizzare l'impiego e le capacità delle risorse professionali dedicate all'attività di analisi e valutazione del credito.

1.4 La determinazione di gruppi caratterizzati da forte interazione

L'analisi del rischio di credito è complesso a causa della composizione dei diversi fattori caratterizzanti la clientela e la gamma di prodotti creditizi presenti nel sistema bancario.

Le soluzioni che gli istituti di credito solitamente elaborano, coadiuvati dalle società specializzate nello sviluppo di soluzioni a supporto delle decisioni, implicano articolati processi di *data mining* e fanno riferimento a modelli statistici di regressione che, mediante l'assegnazione di un punteggio, sintetizzano il posizionamento del cliente o del finanziamento all'interno di una scala ai cui due estremi stanno, rispettivamente, le situazioni meno rischiose e quelle prossime all'insolvenza.

Data l'eterogeneità e la complessità delle relazioni che intercorrono fra il rischio di insolvenza e le caratteristiche della clientela sottoposta a valutazione, generalmente non viene stimato un unico modello di valutazione del rischio da applicare a tutta la popolazione, ma si privilegia la stima di un sistema articolato composto da una pluralità di modelli.

Pertanto, la difficoltà principale da affrontare nell'analisi multidimensionale del problema della valutazione del rischio di insolvenza risiede nell'individuazione preliminare dei gruppi di clienti che sono il più possibile caratterizzati rispetto al fenomeno oggetto di studio.

Il presente studio focalizza la propria attenzione su questa delicata fase preliminare e sull'utilizzo dell'analisi di segmentazione come strumento idoneo a fornire gli elementi per individuare le interazioni fra le variabili predittive e, dunque, descrivere il più chiaramente possibile le caratteristiche dei gruppi individuati (Sonquist e Morgan, 1964; Breiman et al., 1984; Fabbris, 1997).

Il ricorso alle tecniche di segmentazione, note anche come alberi di classificazione, nella fase esplorativa può infatti facilitare l'identificazione dei gruppi più omogenei in riferimento sia al rischio di insolvenza, sia alle caratteristiche proprie della clientela privata e dei relativi portafogli di

prodotti, calibrando nel modo più adeguato i criteri di classificazione in funzione dei risultati (Frydman, Altman e Kao, 1985).

La successiva stima di distinte rette di regressione all'interno dei singoli gruppi individuati dall'analisi permette di descrivere più adeguatamente la realtà (Fabbris, 1997) e consente di creare più robuste regole di predizione del rischio di insolvenza.

CAPITOLO 2

Il campione di analisi

2.1 La base dati

Ai fini dell'analisi del rischio di insolvenza si prende in esame un campione casuale di clienti privati che alla data del 30 aprile 2006 (T_0) risultano possedere, con ruolo di richiedente principale o coobligato[2], almeno un finanziamento aperto presso un istituto di credito, e/o che sono segnalati in una parallela banca-dati dei più gravi eventi negativi associati all'insolvenza (*black-list*).

Le informazioni disponibili in merito ai clienti del campione di riferimento sono estratte dai seguenti due database:

- *Banca-dati storica dei finanziamenti erogati ai clienti*, in cui sono contemporaneamente contenute sia le informazioni di tipo "statico", come la tipologia di finanziamento o prodotto, la data di apertura del contratto, la durata, la frequenza di pagamento e altre caratteristiche del credito invarianti nel tempo, sia le informazioni di tipo "dinamico", come l'importo del saldo, l'importo degli utilizzi o degli sconfinamenti rispetto al limite di credito concesso, i giorni di ritardo nel pagamento delle rate e altre caratteristiche di gestione del credito finanziato che vengono registrate mese dopo mese;

[2] Nel presente studio non sono stati inclusi nel campione i clienti presenti con il solo ruolo di "garante" che, in una fase di analisi circoscritta a linee di credito che non sono in fase critica di recupero a legale, non sono generalmente considerati come responsabili "diretti" dell''andamento dei pagamenti e del conseguente rischio di insolvenza.

- *Banca-dati delle segnalazioni di eventi negativi gravi* (protesti, dichiarazioni fallimentari, …) relativi a persone fisiche che possono essere o meno registrate nella banca-dati storica delle linee di credito.

Le caratteristiche socio-demografiche dei clienti, come la data di nascita, l'area di residenza, la condizione lavorativa e la condizione abitativa che, in studi analoghi si sono rilevate significative, non ci sono state rese disponibili. Analogamente, non è stato fornito alcun nome o codice identificativo dei clienti al fine di garantire che tutte le informazioni estratte siano rigorosamente anonime e in nessun modo si possa risalire alla identificazione delle relative persone fisiche.

Il detto *"garbage in - garbage out"* ci rimarca l'inevitabilità di mediocri risultati di output laddove l'input non ha soddisfatto i basilari requisiti di qualità. Per questo, l'attività propedeutica alle analisi che si presentano nel seguito ha contemplato la realizzazione di un accurato processo volto a verificare e ad assicurare l'integrità del dato di input. I dati, provenienti dalle diverse fonti, sono stati "puliti" (es. eliminazione di record assolutamente identici, correzione di dati incoerenti,…) codificati e normalizzati.

Inoltre, essendo l'unità di analisi il cliente, le informazioni relative ai diversi finanziamenti sono state trasformate e aggregate in variabili riassuntive della tipologia di portafoglio di prodotti del cliente stesso.

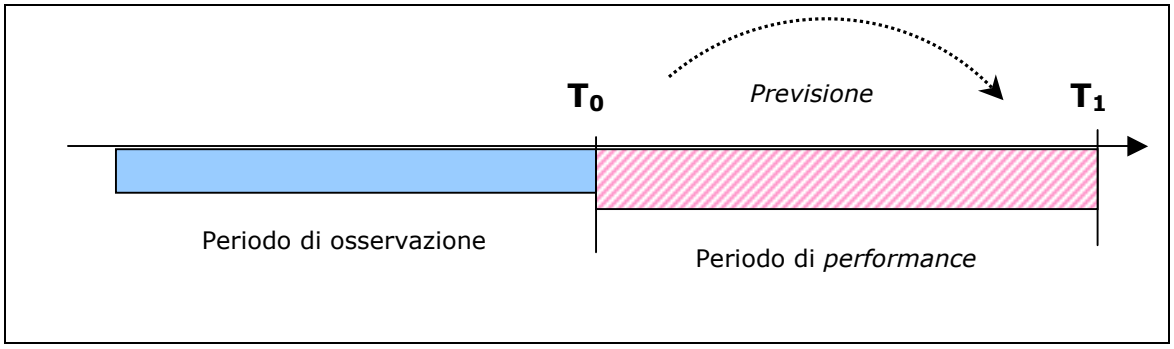
E' rientrata in questa fase anche la preparazione del campione finale di dati e quindi la predisposizione di un data-mart contenente la variabile dipendente oggetto di studio e le potenziali variabili esplicative da utilizzare nell'analisi di segmentazione.

A tal fine, le informazioni storiche disponibili sono state suddivise ed analizzate in due diversi istanti di tempo (Figura 2.1).

Al tempo T_0 (chiamato "data di osservazione") vengono generalmente rilevate ed elaborate tutte le informazioni che si ritengono potenzialmente esplicative nel comportamento creditizio della controparte, mentre al tempo T_1 ("data di *performance*") si analizza come la controparte si è

effettivamente comportata nell'assolvere i pagamenti.

Figura 2.1 *Rappresentazione della suddivisione temporale delle informazioni*



Il campione utilizzato per le analisi si compone di 81.686 clienti, dei quali alla data di osservazione il 30,06% è presente solo nella banca-dati negativa, il 65,33% solo nel data base storico delle informazioni creditizie e il restante 4,61% in entrambe le fonti.

2.2 La variabile dipendente: il rischio di insolvenza

Nel presente studio, si intende definire ed analizzare il rischio di insolvenza secondo una visione "cliente-centrica" (cfr. Paragrafo 1.3).

Pertanto, al fine di determinare il rischio di insolvenza del cliente, dapprima vengono analizzate le singole "storie dei pagamenti" di tutte le linee di credito presenti nel portafoglio del cliente stesso, e solo in un secondo momento ne viene sintetizzata la situazione a livello di cliente.

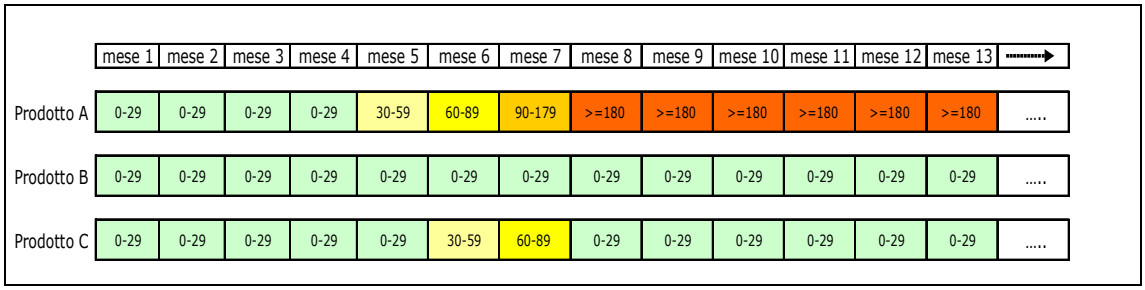
La storia dei pagamenti di un prodotto di credito viene definita da una sequenza temporale di "fotografie" che sintetizzano mese per mese il numero di giorni di ritardo.

Nella Figura 2.2 sono riportate graficamente tre storie di pagamento esemplificative: la storie dei pagamenti si possono presentare "sempre regolari" o "con alcune lievi irregolarità rapidamente sanate", come nel caso dei prodotti B e C, rispettivamente, oppure possono presentarsi "con irregolarità crescenti", come nel caso del prodotto A.

L'analisi della storia dei pagamenti dei prodotti posseduti da un cliente rappresenta la fonte informativa principale da cui muove l'identificazione e la definizione del rischio di insolvenza del cliente stesso, oggetto ultimo di valutazione.

Spesso, oltre alle informazioni sulla storia dei pagamenti, è possibile che in merito alla rischiosità creditizia delle controparti siano reperite fonti informative supplementari, come la banca-dati negativa resa disponibile per il presente studio. In tali circostanze, le informazioni in merito a gravi eventi negativi in capo al cliente concorrono ad affinarne la definizione di rischio di insolvenza.

Figura 2.2 *Rappresentazione grafica delle storie dei pagamenti di tre prodotti*



In generale, il percorso di analisi che conduce all'identificazione della variabile indicatrice della rischiosità del cliente, e quindi della variabile dipendente, si articola nei 5 passaggi chiave:

- *Definizione del periodo di performance (cfr. Paragrafo 2.2.1)*
- *Analisi della storia dei pagamenti dei singoli prodotti (cfr. Paragrafo 2.2.2)*
- *Analisi del portafoglio di prodotti di ciascun cliente (cfr. Paragrafo 2.2.3)*
- *Analisi ed integrazione di altre fonti informative pertinenti (cfr. Paragrafo 2.2.4)*
- *Definizione della variabile di rischiosità del cliente (cfr. Paragrafo 2.2.5)*

2.2.1 Definizione del periodo di performance

Il periodo di *performance* è definito dall'insieme di mesi che intercorrono fra la data di osservazione o di valutazione (T_0), identificata al 30 aprile 2006, e una successiva data di *performance* (T_1), generalmente collocata a 12 mesi di distanza ma limitata al 31 marzo 2007 per mancanza di dati. Il periodo di *performance* ha, quindi, un'ampiezza di 11 mesi.

2.2.2 Analisi della storia dei pagamenti dei singoli prodotti

Per ciascun prodotto di credito presente nella banca-dati alla data di osservazione (T_0) è stata identificata la storia dei pagamenti relativa al periodo di *performance*, ne è stata misurata l'ampiezza (i prodotti che si chiudono fra T_0 e T_1 hanno storie di pagamento di ampiezza inferiore a quella dell'intero periodo di *performance* e possono non concorrere alla definizione complessiva di insolvenza del cliente) e ne è stata colta la più grave manifestazione di irregolarità di pagamento, ovvero il più elevato numero di giorni di ritardo.

Da una analisi preliminare si evince che il 60% dei prodotti dei clienti analizzati presenta delle storie di pagamento completamente regolari mentre il 16,4% presenta delle irregolarità di pagamento. Fra i contratti con un ritardo di pagamento superiore ai 30 giorni, il 50% circa presenta comunque dei ritardi di pagamento che non superano i 60 giorni (Tabella 2.1.).

Si rileva, inoltre, un restante il 23,6% di prodotti che viene definito "senza *performance*", o perché senza un proseguimento della storia dei pagamenti durante il periodo di *performance*, o perché con una storia dei pagamenti regolare ma troppo breve per poterne dedurre con ragionevole sicurezza la bontà di comportamento[3].

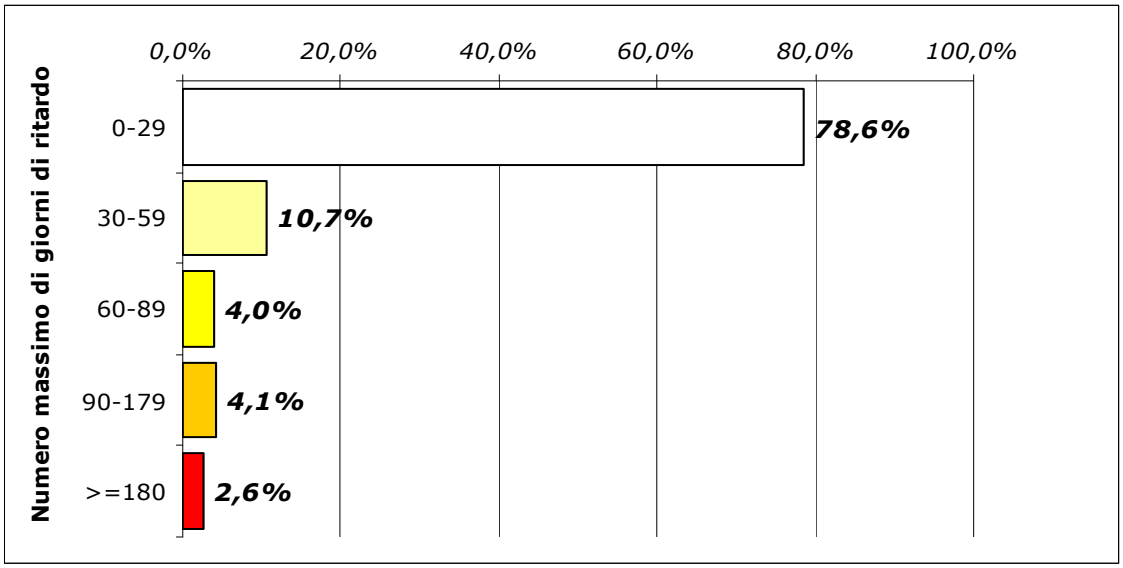
[3] In molti studi condotti sui rischi di insolvenza viene sovente adottato un criterio conservativo secondo cui "è *sufficiente un mancato pagamento per poter essere definiti con performance "non good", ma per essere definito come "good" ci vogliono almeno sei consecutivi pagamenti regolari*".

Tabella 2.1 *Distribuzione di frequenza dei prodotti secondo la più grave irregolarità di pagamento rilevata durante il periodo di performance*

Più grave irregolarità di pagamento del singolo prodotto (espressa in giorni di ritardo)	*Prodotti (#)*	Prodotti (%)
0-29	115.646	*60,04*
30-59	15.688	*8,14*
60-89	5.879	*3,05*
90-179	6.099	*3,17*
>=180	3.875	*2,01*
Senza *performance* (NR)	45.440	*23,59*
Totale	192.627	*100,00*

La distribuzione di frequenza dei soli prodotti con *performance* secondo la più grave irregolarità di pagamento (Figura 2.3), evidenzia che solo un quarto dei casi presenta ritardi di pagamento uguali o superiori ai 30 giorni.

Figura 2.3 *Rappresentazione grafica della distribuzione di frequenza percentuale dei soli prodotti con performance, secondo la più grave irregolarità di pagamento (numero massimo di giorni di ritardo)*



2.2.3 Analisi del portafoglio di prodotti di ciascun cliente

Per ciascun cliente del campione di riferimento sono state analizzate le manifestazioni di irregolarità del portafoglio di prodotti durante il periodo di *performance*.

Anche in questo caso, secondo una prassi cautelativa che impone di considerare come indicatore di sintesi la peggior situazione manifestata durante il periodo di *performance*, è stata individuata la più grave manifestazione di irregolarità di pagamento, ovvero il più elevato numero di giorni di ritardo, su tutto il portafoglio di prodotti.

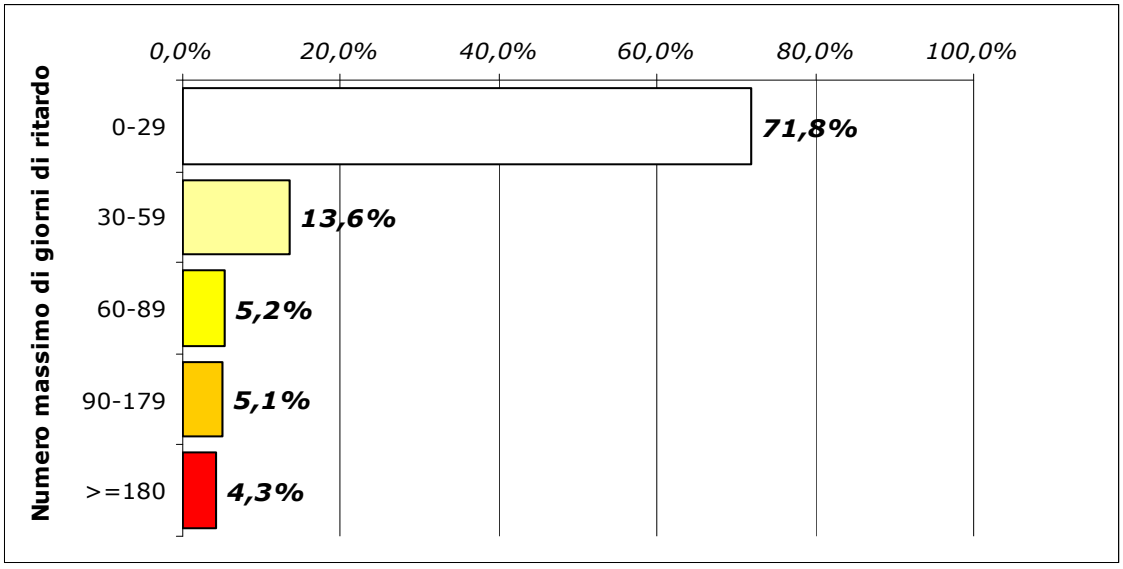
L'analisi della più grave irregolarità di pagamento manifestatasi nel portafoglio prodotti del cliente evidenzia che nel 50.4% dei casi i clienti hanno dei pagamenti completamente regolari su tutti i prodotti (Tabella 2.2). Lasciando da parte il gruppo di clienti per i quali mancano informazioni circa il comportamento di pagamento durante il periodo di *performance* (29,8% del campione), la distribuzione di frequenza della peggiore irregolarità di pagamento mette in luce che la percentuale di clienti regolari è pari al 72% dei casi. I clienti con ritardi compresi fra 30 e 59 giorni rappresentano il 13,5% dei casi, mentre quelli con 60-89 e 90-179 giorni di ritardo rappresentano ciascuno il 5% circa dei casi. Infine, il 4,3% dei clienti presenta ritardi uguali o superiori ai 180 giorno (Figura 2.4).

Tabella 2.2 *Distribuzione di frequenza dei clienti secondo la più grave irregolarità di pagamento sull'intero portafoglio di prodotti posseduti, rilevata durante il periodo di performance*

Più grave irregolarità di pagamento del cliente (espressa in giorni di ritardo)	*Clienti (#)*	Clienti (%)
0-29	41.188	*50,42*
30-59	7.782	*9,53*
60-89	2.986	*3,66*
90-179	2.935	*3,59*
>=180	2.447	*3,00*
Senza *performance* (NR)	24.348	*29,81*
Totale	81.686	*100,00*

La distribuzione delle frequenze è caratterizzata da una forte concentrazione sulla modalità che denota la nullità del fenomeno (clienti con ritardi di pagamento inferiori ai 30 giorni) e da un decadimento molto rapido in corrispondenza delle modalità su cui si collocano le massime intensità di gravità.

Figura 2.4 *Rappresentazione grafica della distribuzione di frequenza percentuale dei soli clienti con performance, secondo la più grave irregolarità di pagamento (numero massimo di giorni di ritardo) sul portafoglio di prodotti posseduti*



2.2.4 Analisi ed integrazione di altre fonti informative pertinenti

Per ciascun cliente del campione di riferimento, oltre o in sostituzione della storia dei pagamenti, è stata verificata anche la presenza del cliente all'interno di una *black list*, condizione che rivela un elevato grado di rischiosità creditizia del cliente. Nel campione considerato, il 35,5% dei clienti risulta segnalato nella banca-dati negativa (Tabella 2.3.).

L'integrazione delle informazioni desunte dalla storia dei pagamenti con quelle relative alla presenza/assenza di segnalazioni nella banca-dati negativa hanno rappresentato il successivo passo verso la definizione della variabile indicatore della rischiosità del cliente.

Tabella 2.3 *Distribuzione di frequenza dei clienti secondo la presenza o meno di segnalazioni nella banca-dati negativa, rilevata durante il periodo di performance*

Presenza dei clienti nella banca-dati negativa	*Clienti (#)*	Clienti (%)
Non segnalati	52.665	*64,47*
Segnalati	29.021	*35,53*
Totale	81.686	*100,00*

Definita la *frequenza di segnalazione* nella banca-dati degli eventi negativi come:

$$\textit{Frequenza di segnalazione} = \frac{\textit{Totale clienti segnalati}}{\textit{Totale clienti}} \textit{ x 100}$$

l'analisi della distribuzione congiunta delle informazioni relative alla rischiosità del cliente (Tabella 2.4. e Figura 2.5) evidenzia come la *frequenza di segnalazione* nella banca-dati degli eventi negativi cresca con l'incrementarsi del livello di irregolarità nei pagamenti.

Tabella 2.4 *Distribuzione di frequenza dei clienti secondo la più grave irregolarità di pagamento rilevata sull'intero portafoglio di prodotti posseduti e la presenza di segnalazioni nella banca-dati negativa, durante il periodo di performance*

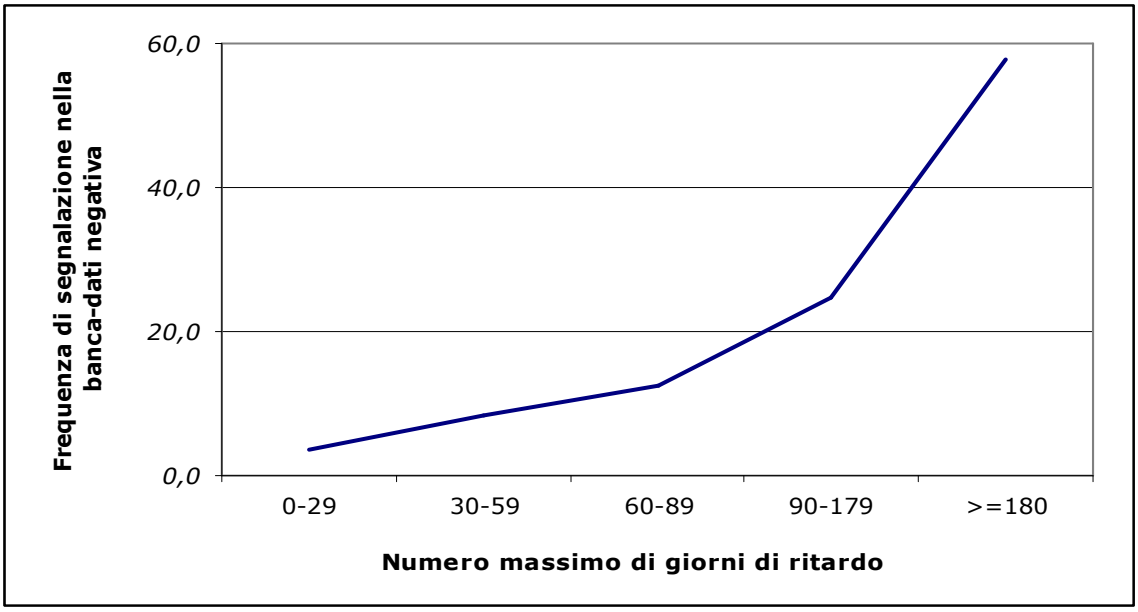
Più grave irregolarità di pagamento del cliente (espressa in giorni di ritardo)	*Clienti non segnalati (#)*	*Clienti segnalati (#)*	*Totale Clienti (#)*	Frequenza di segnalazione
0-29	39.666	1.522	41.188	*3,70*
30-59	7.143	639	7.782	*8,21*
60-89	2.609	377	2.986	*12,63*
90-179	2.213	722	2.935	*24,60*
>=180	1.034	1.413	2.447	*57,74*
Senza *performance*	0	24.348	24.348	*100,00*
Totale	52.665	29.021	81.686	*35.53*

Mente solo il 3,7% dei clienti con pagamenti regolari risulta contemporaneamente segnalato nella banca-dati negativa, per i clienti con

30-59 giorni di ritardo la percentuale è più che doppia (8,2%). In corrispondenza dei livelli più gravi di irregolarità nei pagamenti, risultano registrati nella banca-dati negativa il 24,6% e il 57,7%, rispettivamente, dei clienti con 90-179 e almeno 180 giorni di ritardo.

Inoltre, si mette in evidenza che circa il 30% di clienti per i quali non è stato possibile definire una *performance* di pagamento è invero segnalato nella banca-dati negativa. In particolare, questi risultano essere i clienti che, in seguito ad una segnalazione nella banca-dati negativa, non risultano possedere alcuna forma di finanziamento durante il periodo di *performance*.

Figura 2.4 *Percentuale di segnalazione dei clienti nella banca dati negativa, secondo la più grave irregolarità di pagamento (numero massimo di giorni di ritardo) sul portafoglio di prodotti posseduti*



2.2.5 Definizione della variabile di rischiosità del cliente

La variabile indicatrice della rischiosità del cliente è definita alla luce dei risultati emersi dall'integrazione delle diverse informazioni disponibili in merito alla rischiosità creditizia.

Il criterio adottato per la definizione della variabile di rischiosità

prevede che sia definito come estremamente rischioso il cliente con almeno una segnalazione nella banca-dati "negativa" o con un ritardo di pagamento di almeno 180 giorni. I livelli di rischiosità inferiore sono, di conseguenza, definiti in base alla sola irregolarità di pagamento.

La variabile ordinale distingue i seguenti 5 livelli di rischiosità:

- Assenza di rischio: corrispondente ai clienti che, durante il periodo di *performance*, hanno un ritardo di pagamento massimo di 29 giorni e che, nel contempo, non hanno nessuna segnalazione della banca-dati negativa,
- Rischio basso: per i clienti con 30-59 giorni di ritardo nell'adempimento degli impegni di pagamento e nessuna segnalazione della banca-dati negativa,
- Rischio medio: per i clienti con 60-89 giorni di ritardo e nessuna segnalazione della banca-dati negativa,
- Rischio alto: per i clienti con 90-179 giorni di ritardo e nessuna segnalazione della banca-dati negativa,
- Rischio molto alto: per i clienti con almeno 180 giorni di ritardo e/o almeno una segnalazione della banca-dati negativa.

Oltre la metà dei clienti presenta un livello di rischiosità contenuto (Tabella 2.5.): il 48,6% è sostanzialmente rispettoso delle scadenze di pagamento e un ulteriore 8,7% raggiunge al massimo 59 giorni di ritardo, pari al mancato pagamento di una sola rata mensile. Le situazioni di rischiosità intermedia (da 60 a 89 giorni di ritardo) interessano il 3,2% dei clienti, quelle di rischiosità alta (ritardi compresi fra 90 e 179 giorni) il 2,7%.

I casi più gravi, corrispondenti ai clienti che mancano il pagamento per almeno 6 mesi consecutivi, quindi almeno 180 giorni, o che sono stati segnalati alla banca-dati negativa, anche se con ritardi inferiori ma per altri gravi incidenti di morosità, coinvolgono il 36,8% dei clienti del campione.

Tabella 2.5 *Distribuzione della variabile ordinale di rischiosità del cliente, definita sulla base delle informazioni di credito e delle segnalazioni nella banca-dati negativa*

Livello di rischiosità del cliente	*Clienti (#)*	Clienti (%)
Assenza di rischio (0-29 giorni di ritardo e nessuna segnalazione negativa)	39.666	*48,56*
Rischio basso (30-59 giorni di ritardo e nessuna segnalazione negativa)	7.143	*8,74*
Rischio intermedio (60-89 giorni di ritardo e nessuna segnalazione negativa)	2.609	*3,19*
Rischio alto (90-179 giorni di ritardo e nessuna segnalazione negativa)	2.213	*2,71*
Rischio molto alto (almeno 180 giorni di ritardo e/o almeno una segnalazione negativa)	30.055	*36,79*
Totale	81.686	*100,00*

Dalla combinazione o dalla selezione delle modalità della variabile di rischiosità sono definite le diverse variabili dipendenti (binarie e ordinali) utilizzate nelle analisi di segmentazione binaria del campione di clienti (cfr. Paragrafo 3.3).

Una volta individuate le categorie che definiscono i diversi livelli di rischiosità, la verifica della coerenza interna della variabile indicatrice della rischiosità è svolta analizzando la matrice di transizione (Tabella 2.6 e Figura 2.5) definita a partire dalla manifestazione della rischiosità del cliente rilevata in due orizzonti temporali successivi, distinti e della medesima ampiezza.

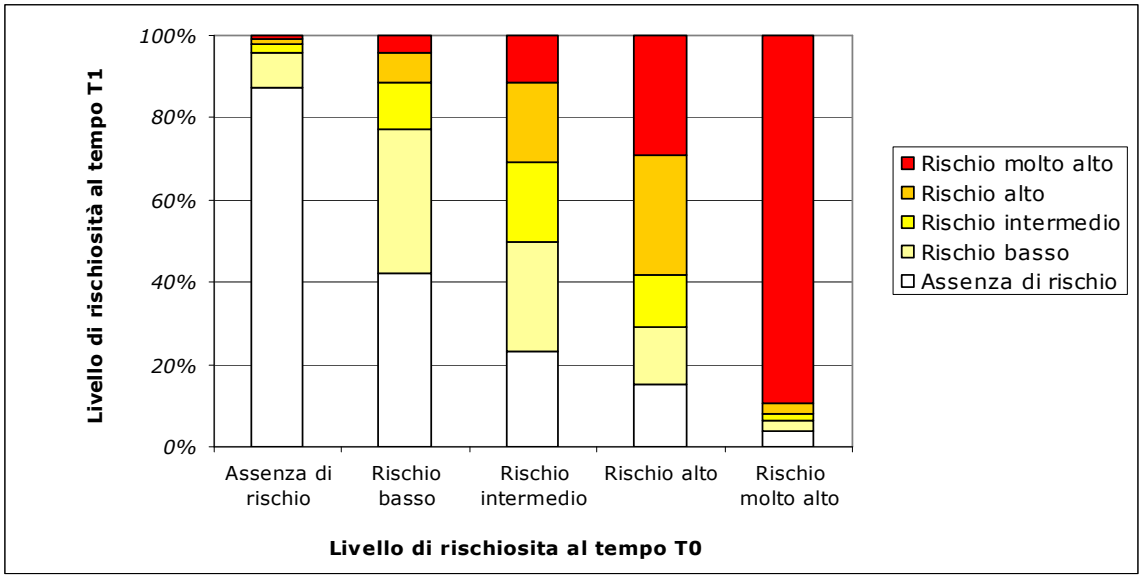
Nella matrice, le righe corrispondono ai possibili livelli di rischiosità osservata al tempo T_0 e, viceversa, le colonne ai possibili livelli osservati, sulla medesima scala di misura, al successivo tempo T_1.

Per ciascuno dei livelli di rischiosità osservato in un primo intervallo temporale, l'analisi della distribuzione percentuale dei casi sulla riga corrispondente fornisce indicazioni in merito alla probabilità di "scivolamento" verso un diverso livello di rischiosità nel susseguente periodo.

Tabella 2.6 *Matrice di transizione fra livelli di rischiosità registrati in due periodi distinti e successivi (T_0 e T_1)*

Livello di rischiosità al tempo T_0	Livello di rischiosità al tempo T_1					
	Assenza di rischio	*Rischio basso*	*Rischio intermedio*	*Rischio alto*	*Rischio molto alto*	*Totale*
Assenza di rischio	***87,4***	*8,1*	*2,2*	*1,3*	*0,9*	*100,0*
Rischio basso	*42,2*	***34,9***	*11,5*	*7,1*	*4,3*	*100,0*
Rischio intermedio	*23,3*	*26,6*	***19,4***	*19,3*	*11,4*	*100,0*
Rischio alto	*15,2*	*13,8*	*12,8*	***29,2***	*29,0*	*100,0*
Rischio molto alto	*3,9*	*2,3*	*1,6*	*2,8*	***89,4***	*100,0*

Figura 2.5 *Rappresentazione grafica della matrice di transizione fra livelli di rischiosità registrati in due periodi distinti e successivi (T_0 e T_1)*



I livelli di "assenza di rischio" e di "rischio molto alto" presentano le probabilità di "scivolamento" più contenute, ovvero tendono a mantenersi "inalterati" nel tempo per quasi il 90% dei casi (87,4% e 89,4%, rispettivamente).

Viceversa, a differenza delle situazioni estreme, gli altri livelli di rischiosità sono molto più "incerti" e appaiono come stati "transitori": le percentuali riportate lungo la diagonale principale, indicatrici della proporzione di clienti che permangono nel medesimo stato di rischio, sono pari al 34,9% per i clienti a "rischio basso", al 19,4% per i clienti a "rischio intermedio" e al 29,2% per i clienti a "rischio alto".

Tali constatazioni portano a riconoscere la coerenza interna delle definizione della variabile indicatrice della rischiosità del cliente: da un lato, è in grado di identificare con certezza le situazioni estreme del fenomeno ("assenza di rischio" e "rischio molto alto") e, dall'altro, permette di cogliere anche le diverse gradazioni che, non risultando nettamente sbilanciate verso uno dei due poli estremi di rischiosità, il che indicherebbe uno stato tendente verso l'assenza di rischio o verso il rischio estremo, si presentano effettivamente come "intermedie".

2.3 Le variabili predittive

L'obiettivo principale dell'analisi di segmentazione è di pervenire ad una partizione dei clienti in esame in gruppi omogenei, sulla base delle variabili esplicative scelte come potenziali predittori.

Per effettuare una corretta analisi di segmentazione, è opportuno definire accuratamente l'insieme di variabili predittive. Pertanto, nella fase preliminare di controllo della qualità del dato sono state condotte una serie di analisi finalizzate a verificare la stabilità temporale delle distribuzioni delle variabili disponibili, a gestire e codificare le informazioni mancanti e a trasformare il livello di misura di alcune variabili (categorizzazione).

Ai fini della costruzione delle variabili da utilizzare nell'analisi, si è innanzitutto dovuto tener conto della diversa natura e ricchezza informativa delle fonti di riferimento: essendo la banca-dati degli eventi negativi una mera *black list* di clienti, l'unica informazione desumibile è stata la presenza del cliente fra i soggetti segnalati come estremamente rischiosi. La maggior parte delle variabili utilizzate nell'analisi è stata pertanto derivata a partire

dalla ben più ricca banca-dati storica delle linee di credito.

Le variabili predittive sono derivate dalle informazioni disponibili fino alla data di osservazione (T_0), e in sostanza si riferiscono alle seguenti due caratteristiche del cliente:

- *Tipologia di portafoglio posseduto*: rientrano in questo gruppo le variabili quali la combinazione di prodotti creditizi posseduti (prestiti, carte di credito, fidi di conto e mutui), la tipologia di prodotti posseduti (rateali e/o non rateali), il numero di prodotti posseduti, il numero di differenti tipologie di prodotti posseduti, le anzianità del primo e dell'ultimo prodotto aperto, il numero di contratti aperti negli ultimi 3 o 6 mesi, il numero di istituti di credito presso cui sono aperte le linee di credito, …
- *Comportamento di pagamento passato*: rientrano in questo gruppo le variabili relative alla peggior insolvenza raggiunta nel corso dell'ultimo anno, al numero di contratti con irregolarità di pagamento, al numero di mesi intercorsi dall'ultima irregolarità di pagamento, alla massima percentuale di utilizzo rispetto al limite di credito concesso, e simili.

L'insieme di 20 variabili, alcune di tipo nominale e altre di tipo numerico, è stato sottoposto ad una preliminare analisi volta a definire le eventuali aggregazioni di modalità o di valori. Per poter agevolmente svolgere l'analisi si segmentazione binaria, una volta identificate le opportune aggregazioni, a partire da ciascuna variabile sono state generate tante variabili dicotomiche quante sono le modalità della variabile meno una.

L'insieme delle variabili utilizzate nell'analisi di segmentazione con le rispettive modalità sono elencate nella Tabella 2.7, con l'indicazione dell'acronimo della variabile riportato nell'illustrazione successiva dei risultati in forma grafica.

Tabella 2.7 *Lista delle potenziali variabili predittive utilizzate nelle analisi di segmentazione*

Variabile predittiva	*Acronimo della variabile*	Modalità
Tipologia di finanziamenti posseduti (rateali e/o non rateali)	*Flag_F*	prodotti con limite di credito altri prodotti, rateali
Presenza di finanziamenti attivi	*Status_t0_F*	senza linee di credito in essere con linee di credito in essere
Presenza di almeno una segnalazione negativa	*In_db_negativo_F*	Sì No
Presenza di ritardi di almeno ai 180 giorni	*Gia_con_almeno_180dpd_F*	Sì No
Presenza di ritardi fra 90 e 179 giorni	*Gia_con_90_179dpd_F*	Sì No
Presenza di ritardi inferiori ai 90 giorni	*Con_meno_di_90dpd_F*	Sì No
Numero di finanziamenti posseduti (1)	*Num_cont_1F*	<=1 >=2
Numero di finanziamenti posseduti (2)	*num_cont_2F*	<=2 >=3
Numero di finanziamenti posseduti (3)	*num_cont_3F*	<=3 >=4
Numero di finanziamenti posseduti (4)	*num_cont_4F*	<=4 >=5
Numero di finanziamenti aperti nel corso dell'ultimo anno (0)	*cont_aperti_ultimo_anno_0F*	0 >=1
Numero di finanziamenti aperti nel corso dell'ultimo anno (1)	*cont_aperti_ultimo_anno_1F*	<=1 >=2
Numero di finanziamenti aperti nel corso dell'ultimo anno (2)	*cont_aperti_ultimo_anno_2F*	<=2 >=3

Numero di finanziamenti aperti nel corso dell'ultimo anno (3)	*cont_aperti_ultimo_anno_3F*	<=3 >=4
Numero di finanziamenti aperti nel corso dell'ultimo anno (4)	*cont_aperti_ultimo_anno_4F*	<=4 >=5
Peggior ritardo di pagamento (0)	*peggior_ritardo_0f*	) dpd >0 dpd
Peggior ritardo di pagamento (1)	*peggior_ritardo_1f*	<30 dpd >=30 dpd
Peggior ritardo di pagamento (2)	*peggior_ritardo_2f*	<60 dpd >=60 dpd
Peggior ritardo di pagamento (3)	*peggior_ritardo_3f*	<90 dpd >=90 dpd
Peggior ritardo di pagamento (4)	*peggior_ritardo_4f*	<120 dpd >=120 dpd
Peggior ritardo di pagamento (5)	*peggior_ritardo_5f*	<150 dpd >=150 dpd
% Massima di utilizzo del limite di credito (1)	*max_utilizzo_limite_cred_1F*	)% >0%
% Massima di utilizzo del limite di credito (2)	*max_utilizzo_limite_cred_2F*	<=80% >80%
% Massima di utilizzo del limite di credito (3)	*max_utilizzo_limite_cred_3F*	<=100% >100%
% Massima di utilizzo del limite di credito (4)	*max_utilizzo_limite_cred_4F*	<=130% >130%
Numero di istituti di credito presso cui sono aperti i finanziamenti (1)	*num_istituti_diversi_1F*	<=1 >=2
Numero di istituti di credito presso cui sono aperti i finanziamenti (2)	*num_istituti_diversi_2F*	<=2 >=3
Numero di istituti di credito presso cui sono aperti i finanziamenti (3)	*num_istituti_diversi_3F*	<=3 >=4
Numero di prodotti diversi posseduti (1)	*num_prodotti_diversi_1F*	<=1 >=2

Numero di prodotti diversi posseduti (2)	*num_prodotti_diversi_2F*	<=2 >=3
Numero di prodotti diversi posseduti (3)	*num_prodotti_diversi_3F*	<=3 >=4
Numero di mesi dall'apertura del primo finanziamento (1)	*eta_primo_prodotto_1F*	<=6 >6
Numero di mesi dall'apertura del primo finanziamento (2)	*eta_primo_prodotto_2F*	<=12 >12
Numero di mesi dall'apertura del primo finanziamento (3)	*eta_primo_prodotto_3F*	<=18 >18
Numero di mesi dall'apertura del primo finanziamento (4)	*eta_primo_prodotto_4F*	<=24 >24
Numero di mesi dall'apertura dell'ultimo finanziamento (1)	*eta_ultimo_prodotto_1F*	<=6 >6
Numero di mesi dall'apertura dell'ultimo finanziamento (2)	*eta_ultimo_prodotto_2F*	<=12 >12
Numero di mesi dall'apertura dell'ultimo finanziamento (3)	*eta_ultimo_prodotto_3F*	<=18 >18
Numero di mesi dall'apertura dell'ultimo finanziamento (4)	*eta_ultimo_prodotto_4F*	<=24 >24
Numero di mesi intercorsi dall'ultima irregolarità di pagamento 30-89dpd[4] (1)	*eta_ultimo_ritardo_30_89dpd_1F*	<=3 >3
Numero di mesi intercorsi dall'ultima irregolarità di pagamento 30-89dpd (2)	*eta_ultimo_ritardo_30_89dpd_2F*	<=6 >6
Numero di mesi intercorsi dall'ultima irregolarità di pagamento 90dpd (1)	*eta_ultimo_ritardo_oltre90dpd_1F*	<=3 >3
Numero di mesi intercorsi dall'ultima irregolarità di pagamento 90dpd (2)	*eta_ultimo_ritardo_oltre90dpd_2F*	<=6 >6
Presenza di almeno un mutuo	*loan_Hous_T0_F*	Sì No

[4] Con l'abbreviazione anglosassone "dpd" si intende indicare i "*days past due,* ovvero i giorni di ritardo.

Presenza di almeno un prestito personale	*loan_PerCon_T0_F*	Sì No
Presenza di almeno un fido di conto	*loan_RevOve_T0_F*	Sì No
Presenza di almeno una carta di credito	*cards_T0_F*	Sì No

CAPITOLO 3

L'analisi di segmentazione di campioni

Nel presente capitolo si presentano gli aspetti metodologici distintivi (Paragrafo 3.1) e gli obiettivi (Paragrafo 3.2) dell'analisi di segmentazione, la procedura di analisi utilizzata (Paragrafo 3.3), la descrizione dell'impostazione degli approcci di analisi che sono stati condotti sul campione in base alla duplice definizione della variabile dipendente, su scala binaria e su scala ordinale (Paragrafo 3.4.) e una riflessione finale sugli indicatori idonei a misurare l'efficacia delle segmentazioni ottenute (Paragrafo 3.5.).

3.1 I metodi per la ricerca dei gruppi a rischio

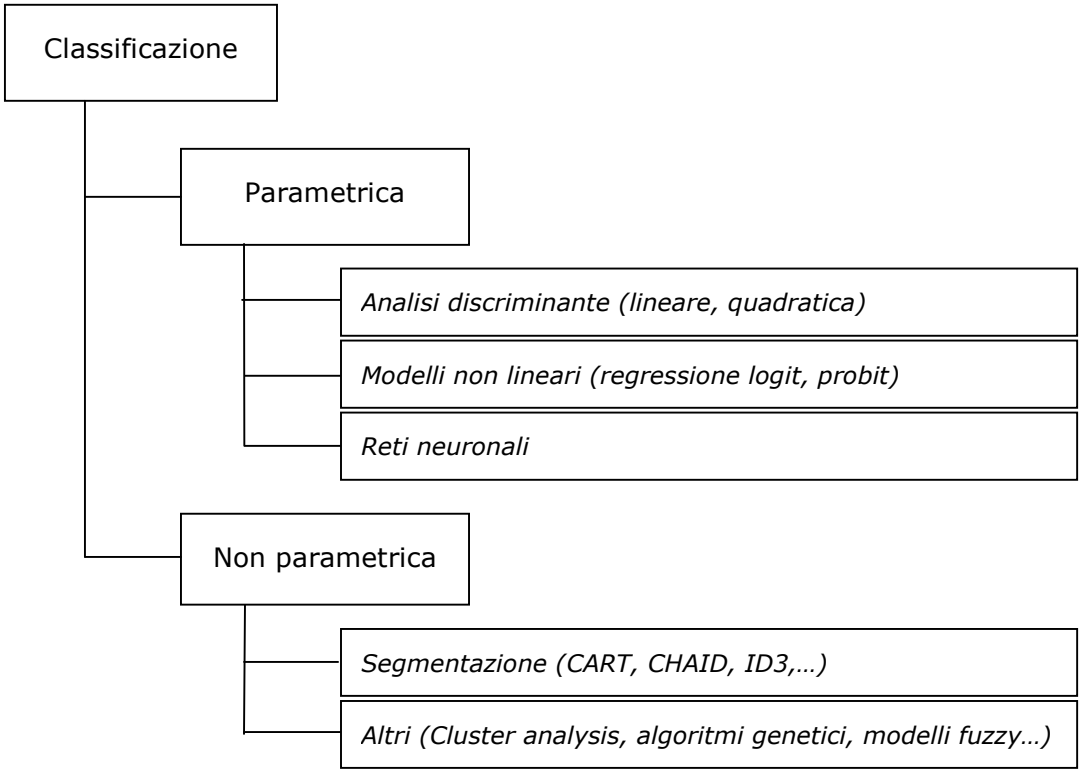
Dato un insieme di *n* unità indipendenti su cui si osserva il rischio di insolvenza (Y) e un certo numero *k* di variabili esplicative X, da un punto di vista formale, il problema della stima di un modello di valutazione del rischio di insolvenza consiste nell'identificare e nel ponderare le relazioni che le *k* variabili esplicative X hanno con la variabile Y.

A tale scopo, il panorama della letteratura statistica (Breiman et al., 1984; Fabbris, 1997; Liu, 2002) mette a disposizione diversi metodi di analisi che si distinguono principalmente per il tipo di assunzione che viene fatta in merito alla distribuzione dei parametri delle variabili predittive: mentre l'applicazione dei metodi *parametrici* richiede che siano verificate condizioni di normalità e di indipendenza distributiva dei parametri, l'applicazione dei metodi *non parametrici* è svincolata da tali assunzioni

(Figura 3.1).

Fra i metodi parametrici, l'*analisi discriminate* e i *modelli di regressione non lineare* sono stati fra i primi ad essere utilizzati nel sistema bancario per la stima del rischio di insolvenza (Altman, 1968; Martin, 1977). Nel tempo, l'espansione delle basi dati del sistema creditizio e l'evoluzione tecnologica hanno favorito lo sviluppo e l'applicazione di nuove tecniche quali le *reti neuronali* (Enache, 1998) e la *segmentazione* di campioni (Sonquist, 1970; Bouroche e Tennenhaus, 1971; Kass, 1980; Breiman et al., 1984). Le altre tecniche non parametriche quali la *cluster analysis*, gli *algoritmi genetici* e i *modelli fuzzy* non trovano altrettanto diffusa applicazione nella stima dei modelli di valutazione del rischio, ma sono per lo più utilizzati per confermare le soluzioni individuate tramite altri metodi e per incrementarne l'accuratezza (Galindo e Tamayo, 2000).

Figura 3.1 *Schematizzazione delle diverse tipologie di classificazione e dei relativi algoritmi di analisi*



Tuttavia, anche se nella realtà i diversi metodi sono tutti euristici, nel senso che stimano il fenomeno partendo prevalentemente dai dati disponibili, e non esiste il metodo assolutamente ”migliore”, dato che ogni ricerca differisce in termini di struttura, disponibilità e qualità dei dati, il metodo della segmentazione di campioni, che rientra nei metodi *non parametrici*, presenta notevoli vantaggi di flessibilità e di robustezza laddove non si possano assumere specifiche distribuzioni delle variabili predittive (Kaltofen, Paul & Stein, 2006).

Inoltre, rispetto alla tecniche di *cluster analysis*, finalizzate anche queste all'individuazione di gruppi di osservazioni omogenee fra loro e diverse da quelle degli altri gruppi, la segmentazione di campioni si distingue per alcune sostanziali caratteristiche:

- Mentre l'analisi di segmentazione è un metodo di analisi esplorativo asimmetrico e richiede la conoscenza a priori della classe di appartenenza delle unità, il metodo della *cluster analysis* costruisce i gruppi di unità statistiche secondo un'analisi simmetrica.
- La segmentazione è operata con riferimento ad una sola variabile predittiva alla volta (selezionata fra tutte quelle rese disponibili per l'analisi), mentre la formazione dei gruppi nella *cluster analysis* viene fatta usando contemporaneamente tutte le variabili a disposizione senza alcun ordine gerarchia.
- La regola di classificazione individuata attraverso gli algoritmi di segmentazione viene usata per prevedere la collocazione di unità statistiche di cui non si conosce la classe di appartenenza.

3.2 Obiettivi dell'analisi di segmentazione

Il motivo che ha portato alla scelta della tecnica di segmentazione risiede nella duplice possibilità di individuare gruppi di clienti con profilo il più possibile omogeneo e di ricercare le determinanti del rischio di insolvenza, tenendo simultaneamente conto delle variabili esplicative rilevanti.

Più in generale, gli obiettivi che l'analisi di segmentazione consente di raggiungere sono (Fabbris, 1997):

- *Individuazione delle determinanti della variabile dipendente*: l'analisi individua le variabili esplicative che meglio spiegano la variabile dipendente. La segmentazione permette, cioè, di individuare le variabili esplicative rilevanti all'interno di sottoinsiemi di unità progressivamente definiti in modo più chiaro.
- "*Eliminazione" dell'informazione ridondante*. In subordine, l'analisi di segmentazione è in grado di eliminare anche le variabili che ripetono l'informazione contenuta in altre e le variabili che non sono rilevanti per spiegare la variabile dipendente.
- *Ricerca di interazioni fra variabili predittive*. L'interazione è l'effetto che una combinazione di modalità ha sulla variabile dipendente e si può scoprire solo analizzando le distribuzioni condizionate (gerarchia delle variabili)
- *Produzione di regole di previsione o di classificazione*. Sulla base del campione osservato si crea una regola di previsione del valore della variabile dipendente di un nuovo soggetto per il quale si siano osservate le variabili esplicative.
- *Ricerca di relazioni non lineari o non monotone*. Poiché alla base degli alberi di classificazione non c'è un modello che lega la variabile dipendente alle variabili esplicative, tale relazione può essere di forma qualsiasi.

3.3 L'analisi di segmentazione

Dato un insieme di *n* unità indipendenti su cui si osservano una variabile dipendente Y e un certo numero *k* di variabili esplicative X, l'analisi di segmentazione permette di esplorare le relazioni fra variabili mediante la suddivisione progressiva del campione iniziale in gruppi via via più omogenei al loro interno rispetto alla variabile dipendente, detta anche "criterio" dell'analisi (Fabbris, 1997).

Nel presente studio, l'analisi che si è stabilito di compiere è orientata alla segmentazione di un campione di clienti, partendo dalle relazioni che le variabili esplicative disponibili al momento dell'osservazione (T_0) fanno rilevare con la variabile dipendente di risposta rilevata nel successivo periodo di *performance* (T_1).

Inizialmente, il campione sottoposto a valutazione è al suo interno fortemente disomogeneo rispetto alla variabile obiettivo perché racchiude tutti i clienti del sistema bancario, senza alcuna distinzione o classificazione derivante dalle le variabili esplicative.

Grazie all'analisi di segmentazione, si possono esaminare le relazioni fra queste variabili attraverso la suddivisione progressiva del collettivo in gruppi via via più omogenei al loro interno rispetto alla variabile dipendente, ovvero al rischio di insolvenza, e se ne possono individuare le determinanti. La distinzione fra clienti rischiosi e non rischiosi ottenuta a livello finale è più netta rispetto alla classificazione di primo livello.

Il tipo di segmentazione che si realizza è definito in base al numero di sottoinsiemi che si possono formare ad ogni passo. Nel presente studio si esamina il caso specifico della *segmentazione binaria*, con partizione a due vie: la procedura di segmentazione inizia con la suddivisione del campione di *n* unità in due sottoinsiemi, definiti dalle modalità di una fra le variabili esplicative, e, nei passi successivi, i sottoinsiemi ottenuti al passo precedente vengono ulteriormente suddivisi in due, fino all'arresto del processo.

La segmentazione gerarchica è quindi una procedura di tipo *stepwise* dove, ad ogni passo, la migliore segmentazione viene selezionata sulla base di un criterio di omogeneità interna (funzione criterio), detta anche *"purezza"* (Breiman et al., 1984) dei gruppi che si vengono a creare dalle suddivisioni.

Ogni gruppo formato ad uno stadio del processo può essere poi ulteriormente suddiviso negli stadi seguenti, fino a quando tale processo viene portato a termine con riferimento ad una prefissata regola di arresto.

I risultati delle tecniche di segmentazione sono visualizzati attraverso

strutture grafiche gerarchiche dette *alberi di classificazione*. Ogni albero è costituito da un insieme finito di elementi:

- i *nodi*, che rappresentano ciascuno un gruppo di unità a diversi stadi del processo di classificazione. Un nodo viene chiamato "*genitore*" rispetto ai nodi che esso genera, e "*figlio*" rispetto al nodo da cui discende. I valori soglia di una variabile che dividono le unità di un determinato nodo sono chiamati cut-off o split.

- i *rami*. che sono le condizioni che hanno determinato la suddivisione

- le *foglie*, che sono i nodi terminali per i quali non si ritiene utile una ulteriore suddivisione

3.3.1 La procedura di segmentazione utilizzata

L'analisi di segmentazione implica una successione di scelte e di impostazioni, derivanti sia dalla valutazione di esperienze di analisi analoghe, sia dagli strumenti tecnici a disposizione per l'analisi stessa[5].

Nel presente studio, la procedura di segmentazione binaria del campione ha visto realizzarsi le seguenti scelte:

- *Tipo di variabile dipendente*
- *Tipo di predittori e dicotomizzazione delle variabili esplicative* (cfr. Paragrafo 2.3). Per la scelta delle variabili da utilizzare nel modello si è cercato di sfruttare soprattutto le considerazioni e l'esperienza maturata in studi analoghi volti all'analisi del rischio di insolvenza.
- *Individuazione della funzione criterio*. Per la verifica della significatività delle relazioni fra variabili è stato adottato il coefficiente del χ^2. Il valore di significatività che deve raggiungere il test che valuta la bontà della partizione prima che questa possa essere applicata è stato posto pari a 0,05 ed è stata adottato

[5] L'analisi dei dati è stata condotta con gli algoritmi di segmentazione implementati nel software KnowledgeSEEKER IV Versione 4.5.2 (ANGOSS Software Company).

aggiustamento di Bonferroni. Sotto tale soglia un gruppo genitore è considerato compatto e omogeneo e, quindi, non ulteriormente divisibile.

- *Definizione delle regole di arresto*. Per prevenire la formazione di gruppi di numerosità ridotta si impone una soglia sotto la quale la numerosità dei gruppi può non essere sufficiente per garantire l'attendibilità delle stime: nel nodo genitore, il numero minimo di casi è fissato pari al 6% dei casi totali del campione analizzato; nel nodo foglio, il numero minimo è fissato pari al 3%. Questo fa sì che l'analisi non possa proseguire per i gruppi la cui ulteriore suddivisione creerebbe dei nodi figli di numerosità insufficiente a garantire stime attendibili e criteri di classificazione generalizzabili.

Inoltre, nel presente studio, per l'applicazione della tecnica di segmentazione, il campione iniziale di clienti (*n* grande) è suddiviso in modo casuale in due sottocampioni distinti e denominati, rispettivamente, come:

- *learning sample,* pari al 70% del campione iniziale
- *testing sample,* o *hold out sample,* pari al restante 30% del campione.

Il primo costituisce il campione di apprendimento su cui si determinano i gruppi caratterizzati da interazione, mentre il secondo viene utilizzato nella fase successiva per verificare l'affidabilità e la stabilità della classificazione individuata sul precedente campione di clienti.

3.4 Gli approcci con variabile dipendente binaria e ordinale

Al fine di individuare i fattori di rischio che determinano l'insolvenza creditizia della clientela privata del sistema bancario, sono stati analizzati e messi a confronto gli esiti ottenuti sul campione di dati in seguito all'applicazione delle tecniche di segmentazione realizzata secondo due approcci distinti e complementari.

La scelta di procedere nell'analisi di segmentazione rifacendosi a tale duplice approccio è ispirata dalla possibilità di identificare e di verificare

possibili differenze fra le componenti del rischio "estremo" e il rischio "intermedio ed estremo" di insolvenza, distinguendo le cause più importanti.

3.4.1 L'approccio con variabile binaria

L'analisi di segmentazione del rischio di insolvenza con variabile dipendente binaria è finalizzata a costruire classificare i clienti in gruppi omogenei in funzione di significative variabili esplicative del "rischio estremo" di insolvenza.

Dal momento che il concetto di "rischio estremo" può essere suscettibile di interpretazioni diverse, si propongono due varianti per la definizione della variabile dipendente. Definito lo stato di "assenza di rischio" come quello corrispondente ai clienti che, durante il periodo di *performance*, hanno avuto ritardi inferiori ai 30 giorni e nessuna segnalazione nella banca-dati negativa, per la definizione dello stato di "rischio estremo" si ipotizzano le seguenti definizioni:

- la prima definizione, derivando dalla comune prassi bancaria, intravede il "rischio estremo" di insolvenza già al manifestarsi di ritardi di pagamento superiori ai 90 giorni,
- la seconda, invece, definisce il "rischio estremo" solo per le situazioni estremamente gravi, corrispondenti a ritardi superiori ai 180 giorni.

A prescindere dal numero di giorni di ritardo con cui il cliente ha adempiuto ai pagamenti durante il periodo di *performance*, la presenza di almeno una segnalazione nella banca-dati negativa concorre a definirne comunque uno stato di "rischio estremo".

Le Tabelle 3.1 e 3.2 riportano la distribuzione della variabile dipendente dicotomica, rispettivamente secondo le due definizioni, sul campione iniziale di clienti.

Mentre il campione su cui si osserva la variabile dicotomica che definisce il "rischio estremo" per ritardi di almeno 90 giorni è composto da 71.934 clienti e presenta un tasso di rischiosità pari al 44,7%, il secondo

campione, quello su cui si osserva la variabile dicotomica che definisce il "rischio estremo" per ritardi di almeno 180 giorni, è composto da 69.721 clienti e presenta un tasso di rischiosità pari al 43,1%,

Tabella 3.1 *Distribuzione della variabile dicotomica con rischio estremo definito per almeno 90 giorni di ritardo*

Livello di rischiosità del cliente	*Clienti (#)*	Clienti (%)
Assenza di rischio (0-29 giorni di ritardo e nessuna segnalazione negativa)	39.666	*55,14*
Rischio estremo (almeno 90 giorni di ritardo e/o almeno una segnalazione negativa)	32.268	*44,86*
Totale	71.934	*100,00*

Tabella 3.2 *Distribuzione della variabile dicotomica con rischio estremo definito per almeno 180 giorni di ritardo*

Livello di rischiosità del cliente	*Clienti (#)*	Clienti (%)
Assenza di rischio (0-29 giorni di ritardo e nessuna segnalazione negativa)	39.666	*56,89*
Rischio estremo (almeno 180 giorni di ritardo e/o almeno una segnalazione negativa)	30.055	*43,11*
Totale	69.721	*100,00*

3.4.2 L'approccio con variabile ordinale

In questo secondo approccio dell'analisi di segmentazione, si adotta una definizione della variabile dipendente che ricalca esattamente la definizione della variabile indicatrice della rischiosità del cliente analizzata nel Capitolo precedente (cfr. Paragrafo 2.2.5).

Pertanto, per la misura del "rischio intermedio ed estremo", si utilizza una scala ordinale a 5 modalità, la quale consente di evidenziare, con buon dettaglio, le diverse gradazioni di rischiosità della clientela.

La variabile dipendente misurata su scala ordinale distingue i seguenti livelli di rischio: "assenza di rischio", "rischio basso", "rischio intermedio", "rischio alto" e "rischio molto alto".

3.5 La misura dell'efficacia di segmentazioni binarie

L'analisi di segmentazione consente di determinare, a partire da un campione di apprendimento per il quale siano note le determinazioni della variabile dipendente e quelle delle variabili esplicative, una regola di classificazione dei clienti in gruppi il più possibile omogenei (Breiman et al., 1984). Una volta determinata la regola di classificazione si pone il problema di valutarne le seguenti componenti:

- La *capacità discriminante*, che denota l'abilità nel discernere *ex ante* il livello di rischiosità dei clienti
- La *stabilità*, che denota l'abilità a mantenere la medesima capacità discriminante anche quando applicata a campioni diversi e in successivi periodi di tempo.

Queste valutazioni devono essere effettuate, oltre che sul campione di apprendimento, anche e soprattutto su campioni che non hanno concorso alla stima del modello.

Pertanto, quando si dispone di un secondo campione di dati, non utilizzato ai fini della determinazione della regola di classificazione o perché tenuto preventivamente da parte rispetto al campione di apprendimento ("*out of sample*") o perché facente riferimento ad un diverso periodo temporale ("*out of time*"), è possibile approfondire sia l'affidabilità della procedura di assegnazione dei clienti ai gruppi individuati dall'analisi di segmentazione, applicando i medesimi criteri di classificazione al nuovo campione e quindi misurandone la capacità discriminante, sia la possibilità di generalizzare la soluzione individuata, confrontando le distribuzioni che campioni analoghi ma indipendenti di clienti presentano sulle classi individuate dall'analisi.

Per verificare la capacità discriminante e la stabilità della soluzione ottenuta è possibile anche alterare casualmente un certo numero di dati, ripetere l'analisi e verificare se e in quale misura il nuovo albero di segmentazione differisce da quello ottenuto nella prima fase (*Cross-*

Validation)[6]. Nel caso di campioni di ridotta numerosità, il principale vantaggio di tale metodologia risiede nel fatto che consente di utilizzare tutte le osservazioni in maniera più efficiente.

Nella prassi nazionale e internazionale, laddove vengono sviluppati modelli di valutazione del rischio di insolvenza con rispetto ad una variabile dipendente misurata su scala dicotomica ("*good*", "*bad*"), gli indicatori statistici più frequentemente utilizzati per valutare la capacità discriminante della soluzione ottenuta sono l'indicatore *KS di Kolmogorov-Smirnov* e l' *Indice di Gini*.

Data una regola di classificazione che identifica un certo numero *p* di classi*,* ciascuna con associato un tasso di rischiosità R_i *(i= 1,..,p),* entrambi gli indicatori vengono calcolati una volta che il campione è stato ordinato in senso decrescente rispetto al tasso di rischiosità R_i osservato su ciascuna delle *p* classi.

L'indicatore *KS di Kolmogorov-Smirnov* esprime la massima distanza fra le distribuzioni percentuali cumulate dei clienti riconosciuti come estremamente rischiosi *("bad")* e poco rischiosi *("good"),* rispetto alla classificazione individuata. Note tali distribuzioni sulle *p* classi, e calcolata per ciascuna classe *i-esima* la differenza assoluta fra la distribuzione dei clienti estremamente rischiosi $(\%Cum.B_i)$ e quella dei clienti poco rischiosi $(\%Cum.G_i)$, allora l'indicatore KS della classificazione corrisponde al valore massimo, su tutte le *p* classi, di tali differenze.

$$KS = Max_i \left| (\%Cum.B_i - \%Cum.G_i) \right|$$

per $i= 1,..,p.$

L'indicatore *KS* varia fra 0 e 100: valori superiori al 60 sono indicativi di una buona capacità discriminante.

[6] Nella realizzazione di queste analisi si è solitamente vincolati alle opzioni disponibili all'interno degli algoritmi di segmentazione utilizzati

L'Indice di Gini è un indicatore di concentrazione che, misurando il livello di concentrazione delle frequenza di clienti rischiosi nelle singole classi, fornisce una misura di quanto l'adozione di una regola di classificazione permette di migliorare la valutazione del rischio, rispetto alla totale assenza di indicazioni predittive. Anche questo indicatore prende in considerazione le distribuzioni percentuali cumulate di clienti *good (G)* e *bad (B)* sulle p classi individuate, e formalmente viene espresso dalla seguente formula:

$$\textit{Indice di Gini} = \left| 1 - \Sigma_i (\%Cum.G_i + \%Cum.G_{i-1})(\%Cum.B_i - \%Cum.B_{i-1}) \right|$$

per $i = 1,..,p.$

L'*Indice di Gini* assume valori fra 0 e 100, dove i valori estremi corrispondono rispettivamente al minimo e al massimo potere discriminante della classificazione. L'indice è nullo nel caso in cui la soluzione individuata non apporta alcun beneficio rispetto ad una valutazione casuale della rischiosità, ossia quando i clienti rischiosi sono egualmente distribuiti sulle *p* classi senza alcuna concentrazione in alcune limitate classi, mentre è massimo quando la regola di classificazione determina, idealmente, una perfetta differenziazione dei clienti rischiosi da quelli che non lo sono, ossia quando i clienti *"bad"* sono concentrati in un poche classi iniziali.

Un ulteriore modo, più immediato, per valutare la capacità discriminante di un modello sta nell'apprezzamento del tasso di corretta classificazione: valori crescenti superiori al 50% (corrispondenti al risultato di una pura scelta casuale), indicano una buona capacità discriminante del modello.

Per la valutazione della stabilità di classificazioni, ottenute su campioni diversi a fronte dell'applicazione della medesima regola, si fa ricorso all'*Indice di stabilità* che, in linea di principio, si basa sul concetto dell'entropia o incertezza (Shannon e Weaver, 1949). Tale indice viene utilizzato sia per misurare la stabilità delle distribuzioni di frequenze dei

clienti sulle *p* classi, sia per misurare la stabilità dei tassi di rischiosità osservati sulle medesime classi.

Nel primo caso, definita con $\%Tot.L_i$ la percentuale di clienti che, nel campione di apprendimento, ricadono nella classe *i-esima* e con $\%Tot.T_i$ la percentuale di clienti che, nel campione di convalida, ricadono nella medesima classe, l'indice assume la seguente formulazione:

$$Indice\ di\ stabilità_{(Freq)} = (\%Tot.L_i - \%Tot.T_i)\ \ln(\%Tot.L_i / \%Tot.T_i)$$

per $i = 1,..,p.$

Nel secondo caso, definiti con $\%Risk.L_i$ e $\%Risk.T_i$ i tassi di rischiosità rilevati sulla medesima classe *i-esima,* rispettivamente nel campione di apprendimento e in quello di convalida, l'indice assume la seguente formulazione:

$$Indice\ di\ stabilità_{(Risk)} = (\%Risk.L_i - \%Risk.T_i)\ \ln(\%Risk.L_i / \%Risk.T_i)$$

per $i = 1,..,p.$

Valori dell'indice inferiori allo 0,05% danno conferma dell'equivalenza in distribuzione fra i campioni analizzati e identificano soluzioni stabili.

I modelli di valutazione che si ottengono tramite l'analisi di segmentazione sono valutati anche in relazione ad altre dimensioni di carattere più "qualitativo" e maggiormente legate alle possibilità di utilizzo del modello stesso nella quotidiana prassi operativa di valutazione del rischio.

A parità di "bontà" dei modelli, vengono prediletti quelli che soddisfano anche i criteri di *parsimoniosità*, ossia che si basano su un insieme contenuto di variabili esplicative, e di *semplicità interpretativa*, ossia che contemplano variabili esplicative ottenute evitando complesse

trasformazioni dei dati elementari.

CAPITOLO 4

La stima del rischio estremo di insolvenza

4.1 La determinazione dei gruppi con rischio estremo

La determinazione dei gruppi caratterizzati da un "rischio estremo" di insolvenza è stata condotta applicando l'analisi di segmentazione al campione di clienti di cui si conosce l'esito della variabile dipendente binaria, definita secondo le due varianti descritte nel capitolo precedente (cfr. Paragrafo 3.2.1).

4.1.1 Il rischio estremo pari ad almeno 90 giorni di ritardo

L'analisi di segmentazione, di tipo binario, è stata eseguita prima su un campione di apprendimento di 50.358 clienti (*learning sample*) e poi applicata ad un secondo campione di 21.576 clienti (*testing sample*) al fine di verificarne la capacità discriminante e la stabilità.

Entrambi i campioni sono stati selezionati casualmente dal campione di 71.934 clienti per i quali è stato possibile definire il "rischio estremo" di insolvenza in corrispondenza a ritardi superiori ai 90 giorni (Tabella 4.1).

Tabella 4.1 *Numerosità e tassi di "rischio estremo" (almeno 90 giorni di ritardo) del campione iniziale di clienti e dei due campioni selezionati casualmente per l'analisi di segmentazione*

Caratteristiche dei campioni	*Campione iniziale*	*Learning sample*	*Testing sample*
Numerosità campionaria	71.934	50.358	21.576
Tasso di "rischio estremo" (almeno 90 giorni di ritardo)	44,86	45.10	44,30

L'applicazione dell'analisi di segmentazione binaria sul campione di clienti definito come *learning sample* ha prodotto l'albero di classificazione riportato nella Figura 4.1. L'applicazione sul *testing sample* è invece riportata nella Figura 4.2. Entrambi gli alberi sono quindi formati da 14 nodi, di cui 8 foglie terminali (ciascuna identificate dal relativo numero).

Figura 4.1 *Albero di segmentazione ottenuto sul learning sample, con variabile dipendente binaria "rischio estremo" (almeno 90 giorni di ritardo)*

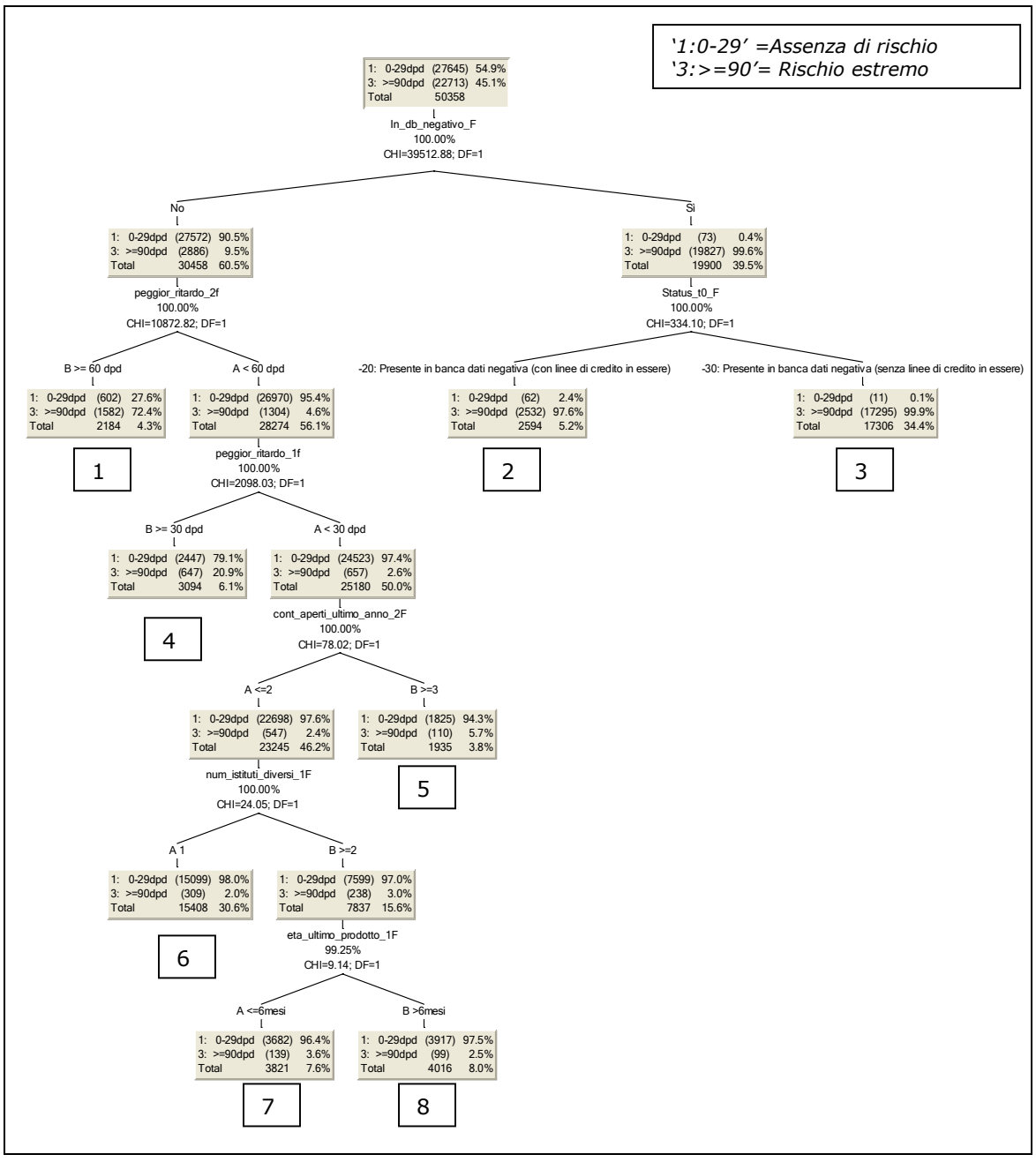
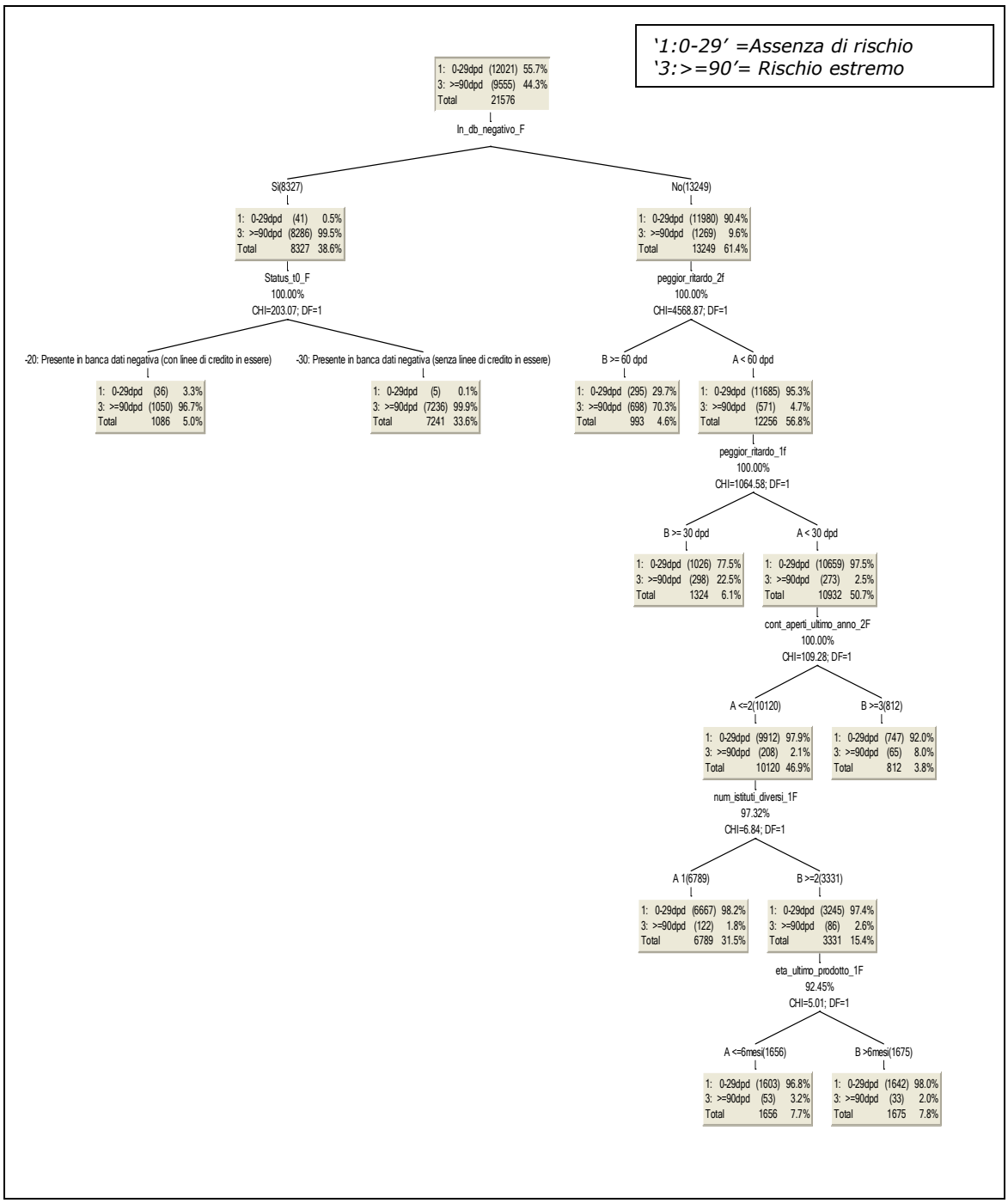


***Figura 4.2**. Albero di segmentazione ottenuto sul testing sample, con variabile dipendente binaria "rischio estremo" (almeno 90 giorni di ritardo)*



Delle variabili predittive disponibili, 6 sono entrate nel processo di segmentazione e una ("Peggior ritardo di pagamento") è entrata due volte determinando una tri-partizione delle modalità.

Le variabili che definiscono l'albero di segmentazione sono pertanto le

seguenti:

- Presenza di almeno una segnalazione negativa (*In_db_negativo_F)*
- Presenza di finanziamenti attivi (*Status_t0_F)*
- Peggior ritardo di pagamento (1) (*peggior_ritardo_1f)*
- Peggior ritardo di pagamento (2) (*peggior_ritardo_2f)*
- Numero di finanziamenti aperti nel corso dell'ultimo anno (2) (*cont_aperti_ultimo_anno_2F)*
- Numero di istituti di credito presso cui sono aperti i finanziamenti (1) (*num_istituti_diversi_1F)*
- Numero di mesi dall'apertura dell'ultimo finanziamento (1) (*eta_ultimo_prodotto_1F)*

La Tabella 4.2 riporta, per ciascuna delle 8 foglie dell'albero di segmentazione, una descrizione sintetica della regola di classificazione che ne determinato l'individuazione, la numerosità (assoluta e percentuale) dei clienti del campione e il relativo tasso di "rischio estremo".

La distribuzione delle foglie, e delle rispettive caratteristiche, è riportata ordinandole in senso decrescente rispetto al tasso di "rischio estremo": le foglie 3 e 2, corrispondenti entrambe ai clienti con almeno una segnalazione nella banca-dati negativa già al momento dell'osservazione, ma distinte per la presenza o meno di un finanziamento attivo, racchiudono il 39,5% dei clienti e presentano i tassi di rischio più elevato (99,9% e 97,6%, rispettivamente).

I clienti che non hanno alcuna segnalazione nella banca-dati negativa, ma che nei mesi di osservazione hanno già evidenziato ritardi superiori ai 60 nell'espletamento degli impegni di pagamento rappresentano il terzo gruppo più rischioso (tasso di "rischio estremo" pari al 72,4%).

Una volta isolati questi 3 gruppi, i clienti rimanenti rappresentano circa il 57% del campione iniziale e presentano un tasso di rischiosità estremamente contenuto (4,6%), pari a un decimo di quello osservato sul

collettivo di partenza.

Tabella 4.2 *Caratteristiche delle foglie dell'albero di segmentazione ottenuto con l'analisi si segmentazione con variabile dicotomica "rischio estremo" (almeno 90 giorni di ritardo)*

Foglia	Descrizione della regola di classificazione	*Clienti (#)*	Clienti (%)	Tasso di rischiosità
3	Presenza di almeno una segnalazione negativa e nessun finanziamento attivo	17.306	*34,37*	*99,94*
2	Presenza di almeno una segnalazione negativa e almeno un finanziamento attivo	2.594	*5,15*	*97,61*
1	Nessuna segnalazione negativa, ritardi di pagamento >= 60gg	2.184	*4,34*	*72,44*
4	Nessuna segnalazione negativa, ritardi di pagamento di 30-59gg	3.094	*6,14*	*20,91*
5	Nessuna segnalazione negativa, ritardi di pagamento di <30gg e almeno 3 finanziamenti aperti nell'ultimo anno	1.935	*3,84*	*5,68*
7	Nessuna segnalazione negativa, ritardi di pagamento di <30gg, meno di 3 finanziamenti aperti nell'ultimo anno, di cui uno negli ultimi 6 mesi ed esposizione con almeno 2 istituti di credito diversi	3.821	*7,59*	*3,64*
8	Nessuna segnalazione negativa, ritardi di pagamento di <30gg, meno di 3 finanziamenti aperti nell'ultimo anno, di cui uno da più di 6 mesi ed esposizione con almeno 2 istituti di credito diversi e	4.016	*7,97*	*2,47*
6	Nessuna segnalazione negativa, ritardi di pagamento di <30gg, meno di 3 finanziamenti aperti nell'ultimo anno ed esposizione presso un solo istituto di credito	15.408	*30,60*	*2,01*
Totale		50.358	*100,00*	*45,10*

A partire da questo, non essendo raggiunti i criteri di arresto (cfr. Paragrafo 3.1.2), la procedura si segmentazione individua ulteriormente altri significativi gruppi di clienti, molto omogenei al loro interno.

Quindi, al passo successivo, la variabile relativa al "massimo numero di giorni di ritardo già raggiunti in passato" viene selezionata nuovamente come la migliore, e permette di isolare un'ulteriore gruppo di clienti (foglia 4) molto rischiosi. Un'ulteriore apprezzabile risultato della segmentazione è dato dal fatto che venga identificato (foglia 6) il gruppo di clienti decisamente meno rischioso (tasso di rischiosità pari al 2%).

La valutazione della capacità discriminante ottenuta verificando i valori assunti dall'indicatore *KS di Kolmogorov-Smirnov* e l' *Indice di Gini* ai risultati ottenuti sul *learning sample* e sul *testing sample* conferma l'efficacia della segmentazione (Tabella 4.3).

Tabella 4.3 *Valori assunti dagli indicatori dell'efficacia della classificazione ottenuta tramite l'analisi di segmentazione con variabile dicotomica "rischio estremo" (almeno 90 giorni di ritardo)*

Indicatori di efficacia	*Learning sample*	*Testing sample*
KS di Kolmogorov-Smirnov	*91,8*	*91,2*
Indice di Gini	*96,6*	*96,2*

La valutazione della stabilità della soluzione ottenuta per mezzo dell'*Indice di stabilità* (cfr. Paragrafo 3.3) fornisce elementi utili per ipotizzare aggregazioni più coerenti e robuste delle foglie ottenute, o per procederne allo sfoltimento. L'esigua numerosità di casi che nel *testing sample* ricadono nelle due foglie più rischiose ha suggerito di aggregarle nuovamente, ovvero di considerare come foglia il gruppo genitore da cui sono state derivate. In seguito a tale operazione di "*pruning*", ossia di sfoltimento dell'albero (Breiman et al., 1984), la nuova classificazione che ne deriva distingue 7 gruppi.

A questo punto, il confronto fra le distribuzione di frequenza dei clienti e dei tassi di rischiosità dei due campioni, rispetto alla nuova

classificazione, eseguito per mezzo dell'*Indice di stabilità,* conferma la stabilità nella capacità di ordinamento del rischio generata dalle nuove classi e quindi la possibilità di generalizzare la regola (Tabella 4.4).

Tabella 4.4 *Indice di stabilità della soluzione ottimale ottenuta con la segmentazione con variabile dicotomica "rischio estremo" (almeno 90 giorni di ritardo)*

	Tasso di rischiosità			Clienti (%)		
Foglia	*Learning sample*	*Testing sample*	*Indice di stabilità (risk)*	*Learning sample*	*Testing sample*	*Indice di stabilità (freq)*
3+2	*99,94*	*99,5*	*0,000002*	*39,5*	*38,6*	*0,000218*
1	*72,44*	*70,3*	*0,000644*	*4,3*	*4,6*	*0,000158*
4	*20,91*	*22,5*	*0,001174*	*6,1*	*6,1*	*0,000000*
5	*5,68*	*8,0*	*0,007941*	*3,8*	*3,8*	*0,000016*
7	*3,64*	*3,2*	*0,000560*	*7,6*	*7,7*	*0,000010*
8	*2,47*	*2,0*	*0,001109*	*8,0*	*7,8*	*0,000057*
6	*2,01*	*1,8*	*0,000229*	*30,6*	*31,5*	*0,000243*

4.1.2 Il rischio estremo pari ad almeno 180 giorni di ritardo

L'analisi di segmentazione binaria svolta con variabile dipendente binaria, qui definita con "rischio estremo" per ritardi uguali o superiori ai 180 giorni, è stata anche in questo caso svolta su un campione di apprendimento e poi convalidata su un secondo campione di clienti (Tabella 4.5).

Tabella 4.5 *Numerosità e tassi di "rischio estremo" (almeno 180 giorni di ritardo) del campione iniziale di clienti e dei due campioni selezionati casualmente per l'analisi di segmentazione*

Caratteristiche dei campioni	*Campione iniziale*	*Learning sample*	*Testing sample*
Numerosità campionaria	68.721	48.804	20.917
Tasso di "rischio estremo" (almeno 180 giorni di ritardo)	43,11	43.30	42,50

L'applicazione dell'analisi di segmentazione binaria ai due campione di *learning* e di *testing* ha portato ad identificare i due alberi di classificazione riportati nelle Figure 4.3 e 4.4, rispettivamente.

Figura 4.3 *Albero di segmentazione ottenuto sul learning sample, con variabile dipendente binaria "rischio estremo" (almeno 180 giorni di ritardo)*

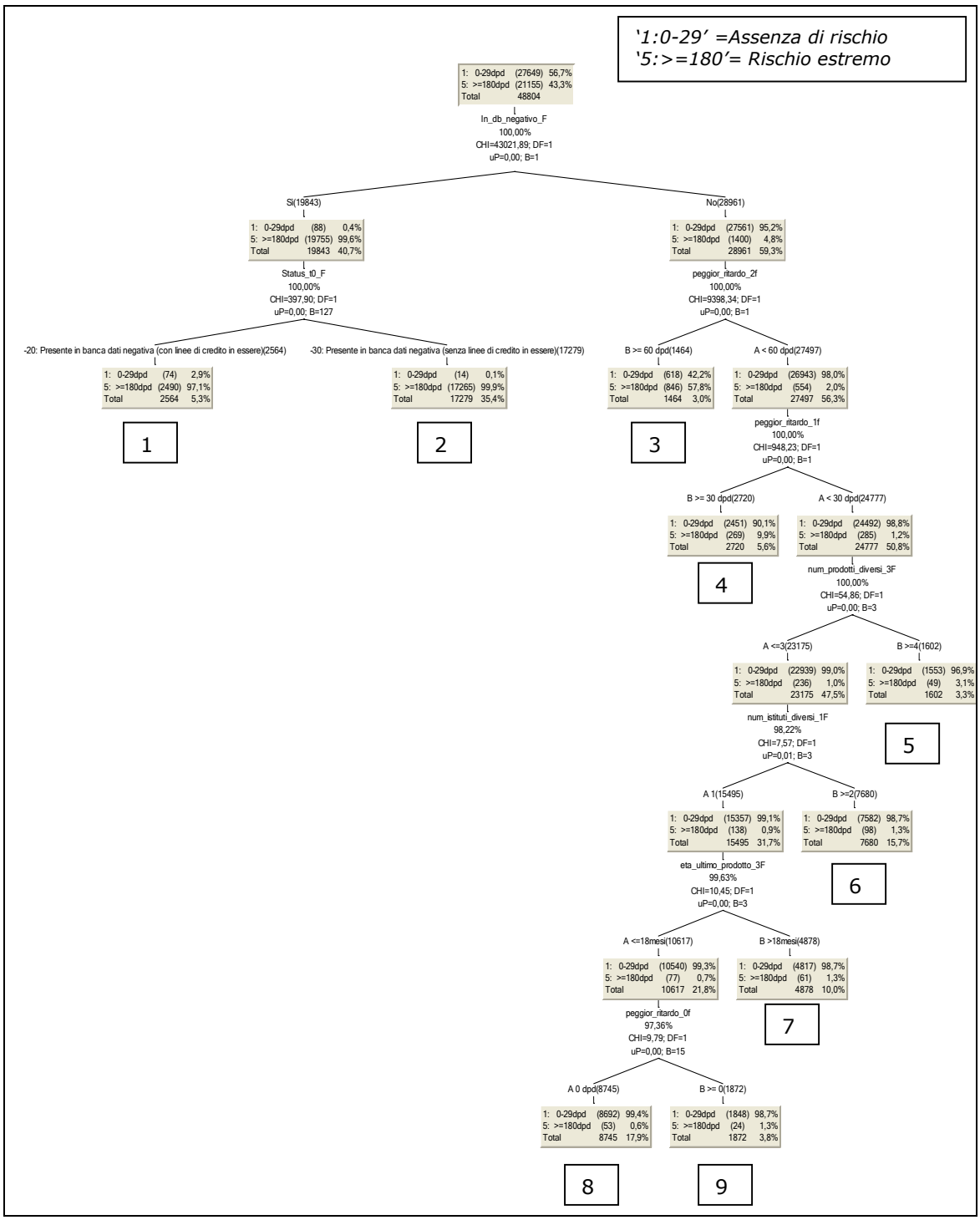
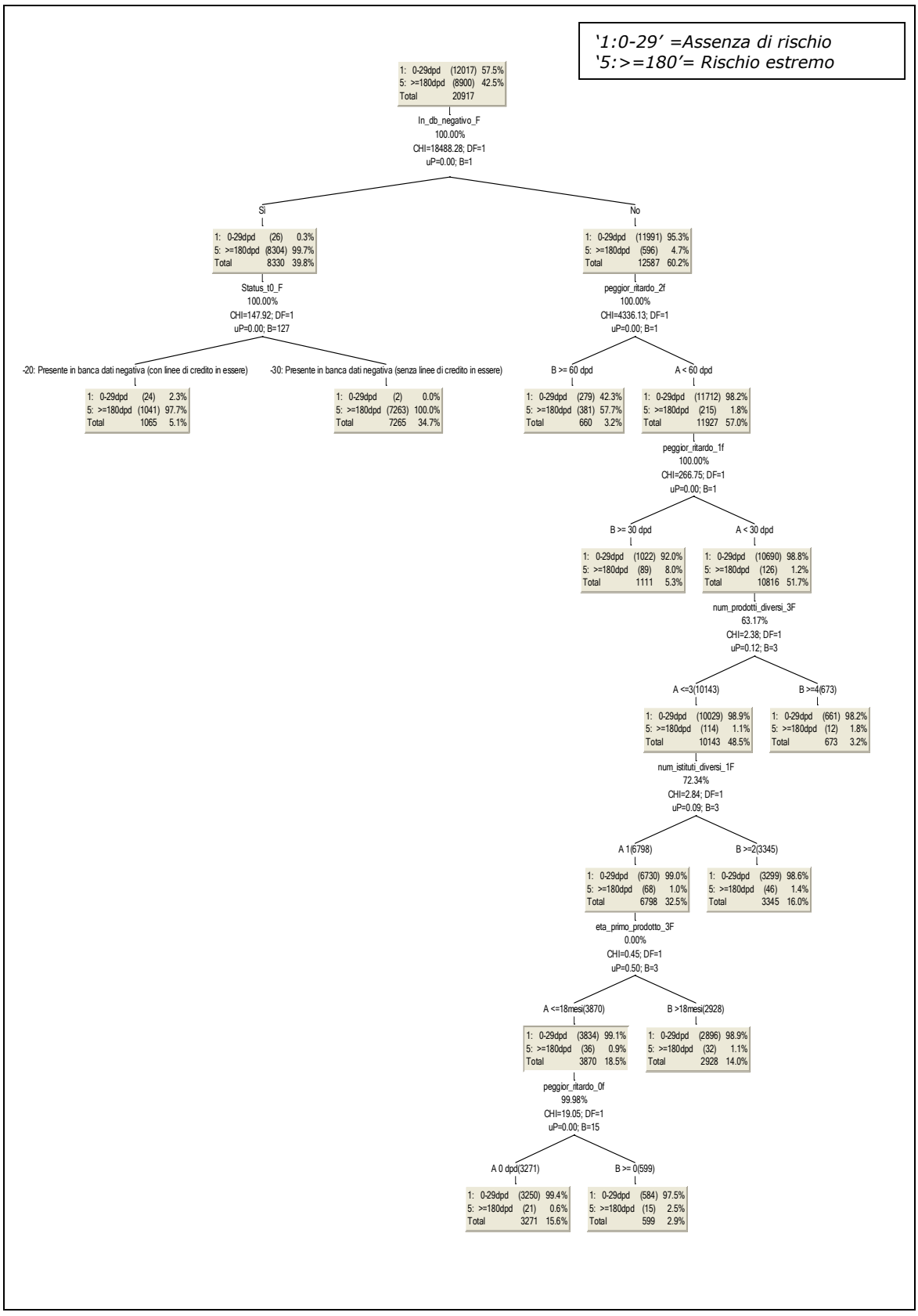


Figura 4.4 *Albero di segmentazione ottenuto sul testing sample, con variabile dipendente binaria "rischio estremo" (almeno 180 giorni di ritardo)*



Anche in questo caso, 6 delle variabili predittive disponibili entrano nel processo di segmentazione. La variabile relativa al "Peggior ritardo di pagamento" entra addirittura tre volte, generando una partizione a quattro vie delle modalità. Le variabili che definiscono l'albero di segmentazione nel caso di "rischio estremo" (almeno 180 giorni di ritardo), sono le seguenti:

- Presenza di almeno una segnalazione negativa (*In_db_negativo_F)*
- Presenza di finanziamenti attivi (*Status_t0_F)*
- Peggior ritardo di pagamento (0) (*peggior_ritardo_0f)*
- Peggior ritardo di pagamento (1) (*peggior_ritardo_1f)*
- Peggior ritardo di pagamento (2) (*peggior_ritardo_2f)*
- Numero di prodotti diversi posseduti (3) (*num_prodotti_diversi_3F)*
- Numero di istituti di credito presso cui sono aperti i finanziamenti (1) (*num_istituti_diversi_1F)*
- Numero di mesi dall'apertura dell'ultimo finanziamento (1) (*eta_ultimo_prodotto_1F)*

L'albero ottenuto si compone di 16 nodi, di cui 9 foglie finali. La distribuzione di queste, con le relative caratteristiche, è riportata nella Tabella 4.6.

Dall'analisi della significatività del test χ^2 sul campione di *testing* e dalla stessa distribuzione dei tassi di rischiosità sulle foglie individuate si evince che gli ultimi passi della segmentazione, nell'isolamento di gruppi di clienti molto diversi fra loro rispetto alla variabile dipendente, hanno solo parzialmente apportato dei benefici significativi.

Il contributo che tale segmentazione dà alla valutazione del rischio di insolvenza risiede principalmente nell'aver messo in evidenza l'importanza della variabile relativa al "Peggior ritardo di pagamento", dopo quella relativa alla "Presenza di almeno una segnalazione negativa".

Tabella 4.6 *Caratteristiche delle foglie dell'albero di segmentazione ottenuto con l'analisi si segmentazione con variabile dicotomica "rischio estremo" (almeno 180 giorni di ritardo)*

Foglia	Descrizione della regola di classificazione	*Clienti (#)*	Clienti (%)	Tasso di rischiosità
2	Presenza di almeno una segnalazione negativa e nessun finanziamento attivo	17.279	*35,40*	*99,92*
1	Presenza di almeno una segnalazione negativa e almeno un finanziamento attivo	2.564	*5,25*	*97,11*
3	Nessuna segnalazione negativa, ritardi di pagamento >= 60gg	1.464	*3,00*	*57,79*
4	Nessuna segnalazione negativa, ritardi di pagamento di 30-59gg	2.720	*5,57*	*9,89*
5	Nessuna segnalazione negativa, ritardi di pagamento di <30gg e almeno 4 prodotti diversi in portafoglio	1.602	*3,28*	*3,06*
9	Nessuna segnalazione negativa, ritardi di pagamento fra 1 e 29gg, meno di 4 prodotti diversi in portafoglio con lo stesso istituto e con il prodotto più recente aperto da meno di un anno e mezzo	1.872	*3,84*	*1,28*
6	Nessuna segnalazione negativa, ritardi di pagamento di <30gg, meno di 4 prodotti diversi in portafoglio aperti con almeno 2 istituti diversi	7.680	*15,74*	*1,28*
7	Nessuna segnalazione negativa, ritardi di pagamento di <30gg, meno di 4 prodotti diversi in portafoglio con lo stesso istituto e con il prodotto più recente aperto da più di un anno e mezzo	4.878	*10,00*	*1,25*
8	Nessuna segnalazione negativa e assolutamente nessun ritardo, meno di 4 prodotti diversi in portafoglio con lo stesso istituto e con il prodotto più recente aperto da meno di un anno e mezzo	8.745	*17,92*	*0,61*
Totale		48.804	*100,00*	*43,35*

I valori assunti dall'indicatore *KS* e dall' *Indice di Gini,* sia sul *learning* che sul *testing sample,* confermano l'efficacia della soluzione di segmentazione individuata (Tabella 4.7).

Tabella 4.7 *Valori assunti dagli indicatori dell'efficacia della classificazione ottenuta tramite l'analisi di segmentazione con variabile dicotomica "rischio estremo" (almeno 180 giorni di ritardo)*

Indicatori di efficacia	*Learning sample*	*Testing sample*
KS di Kolmogorov-Smirnov	*94,8*	*95,0*
Indice di Gini	*98,4*	*98,4*

La valutazione della stabilità della soluzione ottenuta per mezzo dell'*Indice di stabilità* suggerisce di sfoltire e di aggregare alcune delle foglie individuate, al fine di ottenere una regola di classificazione più stabile e generalizzabile.

La nuova classificazione distingue 6 classi ottimali (Tabella 4.8), che si descrivono nel paragrafo 4.2.

Tabella 4.8 *Indice di stabilità della soluzione ottimale ottenuta con la segmentazione con variabile dicotomica "rischio estremo" (almeno 180 giorni di ritardo)*

	Tasso di rischiosità			Clienti (%)		
Foglia	*Learning sample*	*Testing sample*	*Indice di stabilità (risk)*	*Learning sample*	*Testing sample*	*Indice di stabilità (freq)*
2+1	*99,56*	*99,69*	*0,000002*	*40,7*	*39,8*	*0,000106*
3	*57,79*	*57,73*	*0,000001*	*3,0*	*3,2*	*0,000079*
4	*9,89*	*8,01*	*0,003959*	*5,6*	*5,3*	*0,000126*
5+9+6	*1,53*	*1,58*	*0,000015*	*22,8*	*22,1*	*0,000272*
7	*1,25*	*1,09*	*0,000212*	*10,0*	*14,0*	*0,013484*
8	*0,61*	*0,64*	*0,000021*	*17,9*	*15,6*	*0,003105*

4.2 Le determinanti del rischio estremo di insolvenza

Nelle analisi di segmentazione, ottenute con variabile dipendente definita secondo le due varianti del "rischio estremo" (almeno 90 e almeno 180 giorni di ritardo), entrano nella soluzione sostanzialmente le stesse variabili. L'unica differenza è rappresentata dal tipo di variabile che intende quantificare la "dimensione" del portafoglio di finanziamenti del cliente nel momento in cui lo si sta valutando: mentre nella prima soluzione entra la variabile "Numero di finanziamenti aperti nel corso dell'ultimo anno (2)", nella seconda soluzione entra la variabile "Numero di prodotti diversi posseduti (3)".

Quasi tutta la variabilità della variabile dipendente risulta spiegata dalla "Presenza di almeno una segnalazione negativa" nella relativa banca-dati. Lungo la dimensione dell'albero definita dall'assenza di informazioni negative, in entrambe le varianti la variabile "Peggior ritardo di pagamento" entrata al secondo e terzo livello, con diversa combinazione delle modalità: prima per isolare i clienti con ritardi di pagamento superiori ai 60 giorni e successivamente per isolare, fra i clienti con meno di 60 giorni di ritardo, quelli che ne presentano più di 30.

Definito l'*Odds* come il rapporto fra la probabilità *P* che la variabile dipendente *Y* assuma valore 1 (rischio estremo) e quella che assuma valore 0 (assenza di rischio), condizionatamente al valore assunto dalla variabile predittiva (*X=x)*:

$$Odds(x) = \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \frac{P(Y=1 \mid X=x)}{1-P(Y=1 \mid X=x)} = \frac{\pi(x)}{1-\pi(x)}$$

L'*Odds Ratio (OR)*, nel caso di una variabile esplicativa che distingue, attraverso combinazione delle proprie modalità, due gruppi x_1 e x_2, assume la seguente forma:

$$Odds\ Ratio\ (OR) = \frac{\pi(x_1)/(1-\pi(x_1))}{\pi(x_2)/(1-\pi(x_2))}$$

L'analisi delle variabili esplicative tramite gli *Odds Ratio*, o *proportional Odds* (McCullagh e Nelder, 1989), aiuta a confermare che le irregolarità evidenziate nel comportamento di pagamento passato rappresentano le determinanti principali del futuro rischio di insolvenza.

La Tabella 4.9 riporta la sintesi degli *Odds Ratio* osservati nelle due varianti dell'analisi di segmentazione.

Tabella 4.9 *Odds Ratio delle variabili esplicative entrate nella definizione degli alberi decisionali ottenuti dall'analisi si segmentazione con variabile dicotomica "rischio estremo"*

	Odds Ratio (OR)	
Variabile esplicativa	*Rischio estremo (almeno 90 giorni di ritardo)*	*Rischio estremo (almeno 180 giorni di ritardo)*
Presenza di almeno una segnalazione negativa	*2372,05*	*4938,50*
Peggior ritardo >= 60gg	*54,40*	*67,11*
Peggior ritardo = 30-59gg	*9,90*	*9,05*
Contratti aperti nell'ultimo anno >=3	*2,46*	--
Prodotti diversi posseduti >=4	--	*3,17*
Istituti di credito diversi >=1	*0,66*	*1,45*
Mesi dall'apertura ultimo prodotto > 6	*1,46*	--
Mesi dall'apertura ultimo prodotto > 18	--	*1,87*
Peggior ritardo = 1-29gg	--	*2,18*

Se si verifica la "Presenza di almeno una segnalazione negativa" il "rischio estremo" è esplosivo. Questa conseguenza dipende dal fatto che la registrazione nella banca-dati negativa si qualifica come uno stato "assorbente" dal quale è difficile uscire: se alla data della valutazione, a fronte di un evento negativo grave, si è già verifica la segnalazione del

cliente alla banca-dati negativa, è molto probabile che, se non sono "scaduti" i termini di conservazione delle informazioni nel database, la segnalazione sia ancora presente e valida durante il periodo di *performance* su cui viene misurato il rischio.

La seconda determinante del rischio, per ordine di importanza, è data da ritardi di pagamento superiori ai 60 giorni: il verificarsi di questo evento tende a far crescere di oltre 50 volte il "rischio estremo" (*OR* = 54,4), e la gravità dell'evento è tale che il rischio relativo estremo è molto maggiore (*OR* = 67,1) quando si adotta una variante di definizione della variabile dipendente che estremizza le situazioni più gravi (almeno 180 giorni di ritardo).

Quando l'ampiezza del portafoglio aumenta in modo considerevole (nel campione la media di contratti *pro capite* risulta pari a 2,7), ovvero quando risulta che il cliente abbia aperto più di 3 contratti di credito nel corso dell'ultimo (*OR* = 2,5) o che possegga già più di 4 finanziamenti attivi di tipo diverso (*OR* = 3,2), allora i rischi relativi aumentano nella stessa direzione, confermando che un incremento dell'esposizione del cliente, oltre i valori medi del sistema, porta con sé notevoli probabilità di rischio.

CAPITOLO 5

La stima del rischio estremo e intermedio di insolvenza

Nel capitolo 4 sono stati evidenziati, mediante segmentazione binaria, i gruppi della clientela privata che sono soggetti a livelli estremi del rischio di insolvenza. Nel presente capitolo si analizza, sempre applicando la tecnica della segmentazione binaria, i gruppi della clientela privata che sono soggetti a diversi livelli di rischio, ovvero al "rischio estremo e intermedio" di insolvenza.

5.1 La determinazione dei gruppi con vari livelli di rischio

L'analisi di segmentazione finalizzata alla determinazione dei gruppi con vari livelli di rischio di insolvenza è stata condotta considerando come variabile dipendente la stessa variabile indicatrice della rischiosità del cliente, definita al paragrafo 1.4.

La variabile dipendente è pertanto misurata su una scala ordinale a 5 livelli (cfr. Paragrafo 3.2.2).

Anche in questo caso, l'analisi di segmentazione è stata dapprima eseguita su un campione di apprendimento di 57.180 clienti (*learning sample*) e poi applicata ad un secondo campione di 24.506 clienti (*testing sample*).

Entrambi i campioni sono stati selezionati casualmente dal campione iniziale composto da 81.686 clienti. Nella tabella 5.1 sono riportati le numerosità campionarie e la distribuzione dei tassi di rischio per i diversi livelli della variabile dipendente.

Tabella 5.1 *Numerosità e tassi di rischiosità del campione iniziale di clienti e dei due campioni selezionati casualmente per l'analisi di segmentazione*

Caratteristiche dei campioni	*Campione iniziale*	*Learning sample*	*Testing sample*
Numerosità campionaria	81.686	57.180	24.506
Assenza di rischio	48,56	48,50	48,80
Rischio basso	8,74	8,80	8,60
Rischio intermedio	3,19	3,20	3,30
Rischio alto	2,71	2,71	2,70
Rischio molto alto	36,79	36,80	36,70

L'applicazione dell'analisi di segmentazione binaria al campione di *learning* e ha portato ad identificare un 'albero di classificazione che, per questioni di "visibilità grafica", è stato riportato nelle Figure 5.1.a e 5.1.b

Figura 5.1.a *Albero di segmentazione ottenuto sul learning sample, con variabile dipendente ordinale*

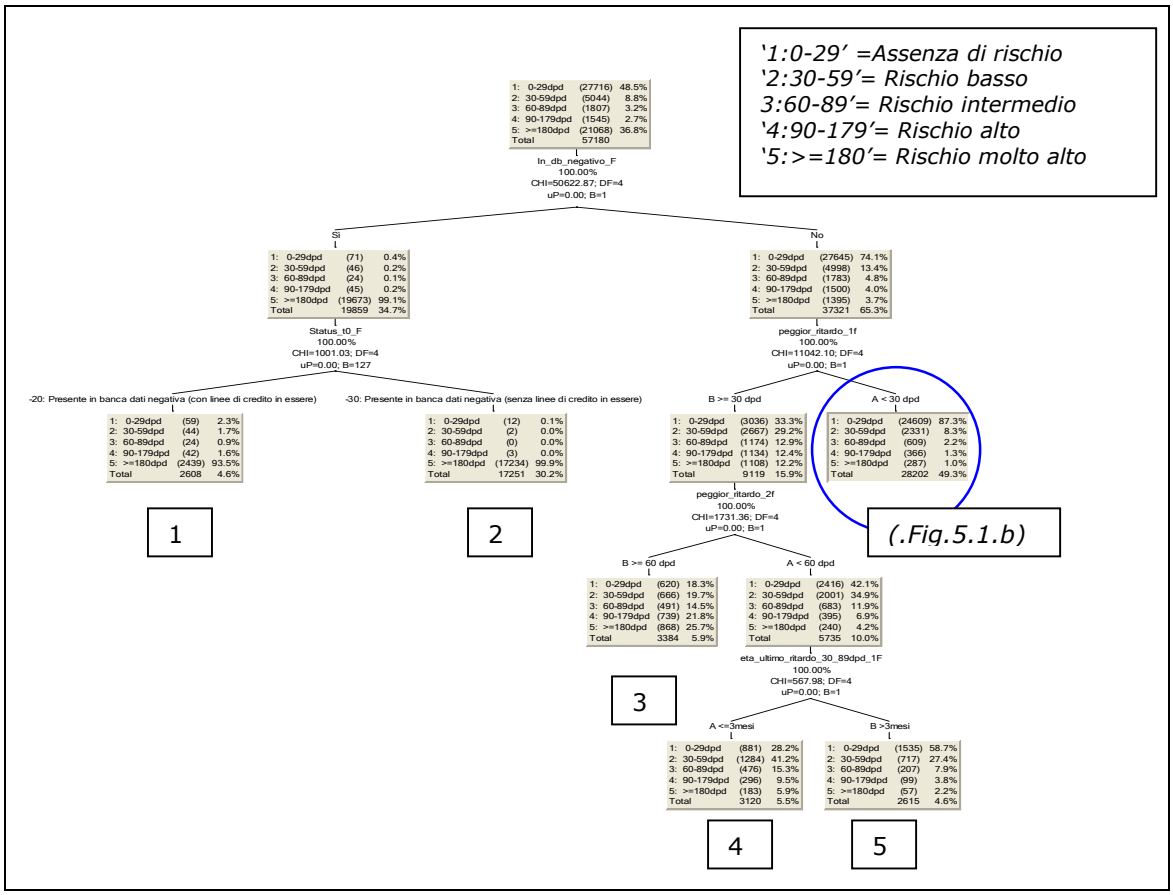
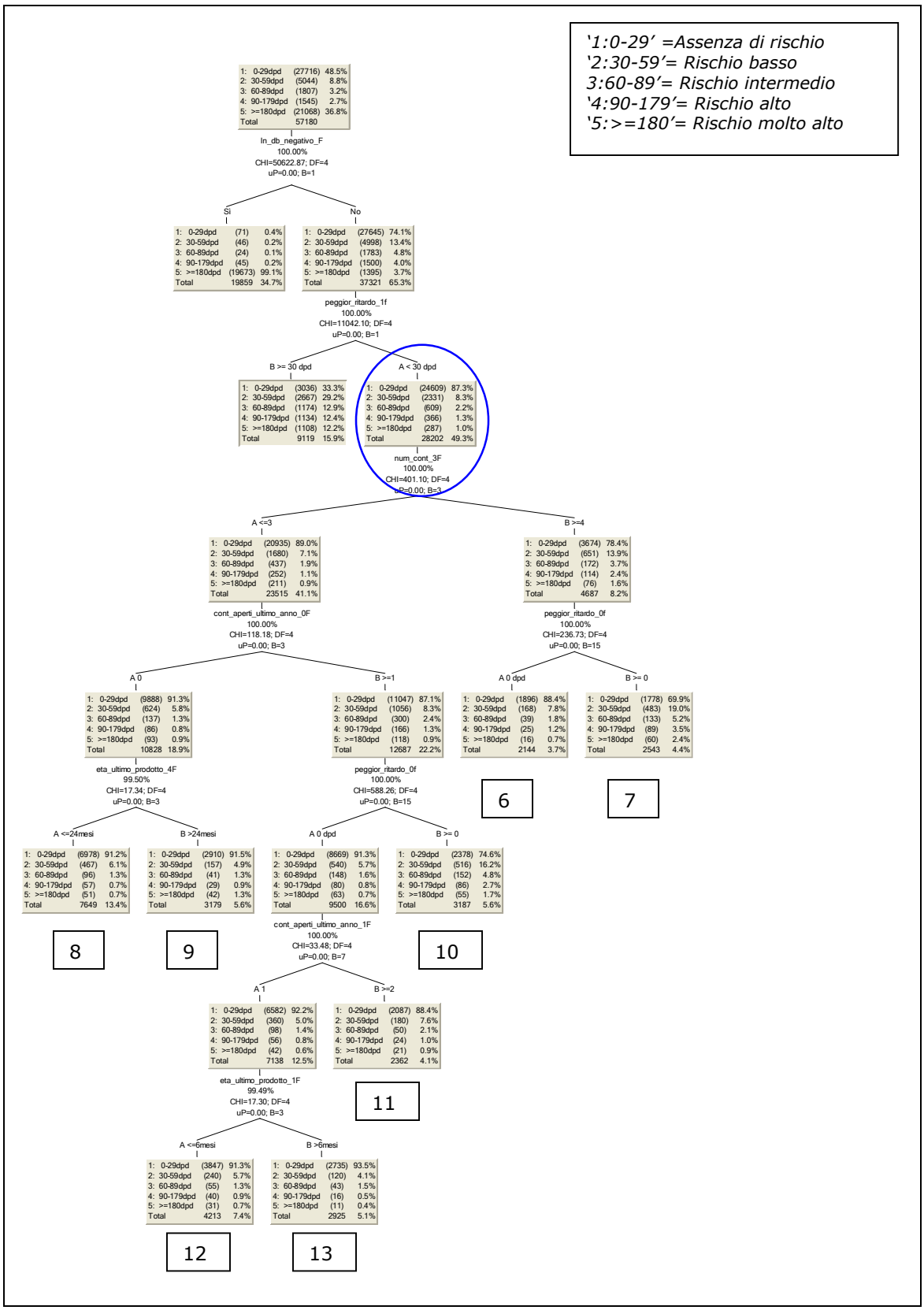


Figura 5.1.b *Albero di segmentazione ottenuto sul learning sample, con variabile dipendente ordinale*



Le variabili che concorrono a definire l'albero di segmentazione sono 7, alcune delle quali entrano più volte nella soluzione, aggregando in modo diverso le modalità. Pertanto, i rami dell'albero sono definiti dalle seguenti combinazioni di modalità delle variabili::

- Presenza di almeno una segnalazione negativa (*In_db_negativo_F)*
- Presenza di finanziamenti attivi (*Status_t0_F)*
- Peggior ritardo di pagamento (0) (*peggior_ritardo_0f)*
- Peggior ritardo di pagamento (1) (*peggior_ritardo_1f)*
- Peggior ritardo di pagamento (2) (*peggior_ritardo_2f)*
- Numero di finanziamenti posseduti (3) (*num_cont_3F)*
- Numero di mesi intercorsi dall'ultima irregolarità di pagamento 30-89dpd[7] (1) (*eta_ultimo_ritardo_30_89dpd_1F)*
- Numero di finanziamenti aperti nel corso dell'ultimo anno (0) (*cont_aperti_ultimo_anno_0F)*
- Numero di finanziamenti aperti nel corso dell'ultimo anno (1) (*cont_aperti_ultimo_anno_1F)*
- Numero di mesi dall'apertura dell'ultimo finanziamento (1) (*eta_ultimo_prodotto_1*
- Numero di mesi dall'apertura dell'ultimo finanziamento (4) (*eta_ultimo_prodotto_4F)*

L'albero è formato da 24 nodi, di cui 13 foglie terminali sintetizzante nelle loro caratteristiche nella Tabella 5.2. Le osservazioni sono riportate per livelli decrescenti del rischio.

I gruppi di clienti più rischiosi sono anche in questo caso identificati dalla "Presenza di almeno una segnalazione negativa" e dai ritardi di pagamento superiori ai 60 giorni.

[7] "dpd", ovvero "*days past due,* indica i giorni di ritardo.

Tabella 5.2 *Caratteristiche delle foglie dell'albero di segmentazione ottenuto con l'analisi si segmentazione con variabile ordinale*

				Livello di rischio				
Foglia	Descrizione della regola di classificazione	Clienti (#)	*Clienti (%)*	*Assenza*	*Basso*	*Medio*	*Alto*	*Molto alto*
2	Presenza di almeno una segnalazione negativa e nessun finanziamento attivo	17.251	*30,17*	*0,1*	*0,0*	*0,0*	*0,0*	*99,9*
1	Presenza di almeno una segnalazione negativa e almeno un finanziamento attivo	2.608	*4,56*	*2,3*	*1,7*	*0,9*	*1,6*	*93,5*
3	Nessuna segnalazione negativa, ritardi di pagamento di >= 60gg	3.384	*5,92*	*18,3*	*19,7*	*14,5*	*21,8*	*25,7*
4	Nessuna segnalazione negativa, ritardi di pagamento di <60gg, d cui l'ultimo avvenuto negli ultimi 3 mesi	3.120	*5,46*	*28,2*	*41,2*	*15,3*	*9,5*	*5,9*
7	Nessuna segnalazione negativa, ritardi di pagamento di 1-28gg e più di 4 prodotti attivi	2.543	*4,45*	*69,9*	*19,0*	*5,2*	*3,5*	*2,4*

5	Nessuna segnalazione negativa, ritardi di pagamento di <60gg, d cui l'ultimo accaduto prima degli ultimi 3 mesi	2.615	*4,57*	*58,7*	*27,4*	*7,9*	*3,8*	*2,2*
10	Nessuna segnalazione negativa, ritardi di pagamento di 1-29gg e più di un contratto aperto nell'ultimo anno	3.187	*5,57*	*74,6*	*16,2*	*4,8*	*2,7*	*1,7*
9	Nessuna segnalazione negativa, ritardi di pagamento <30gg e contratto più recente aperto da meno di 2 anni	3.179	*5,56*	*91,5*	*4,9*	*1,3*	*0,9*	*1,3*
11	Nessuna segnalazione negativa, mai nessun ritardo e più di un contratto aperto nell'ultimo anno	2.362	*4,13*	*88,4*	*7,6*	*2,1*	*1,0*	*0,9*
6	Nessuna segnalazione negativa, mai nessun ritardo e più di 4 prodotti attivi	2.144	*3,75*	*88,4*	*7,8*	*1,8*	*1,2*	*0,7*
8	Nessuna segnalazione negativa, ritardi di pagamento <30gg e contratto più recente aperto da più di 2 anni	7.649	*13,38*	*91,2*	*6,1*	*1,3*	*0,7*	*0,7*

12	Nessuna segnalazione negativa, mai nessun ritardo e un solo contratto aperto da meno di 6 mesi	4.213	*7,37*	*91,3*	*5,7*	*1,3*	*0,9*	*0,7*
13	Nessuna segnalazione negativa, mai nessun ritardo e un solo contratto aperto da al massimo 6-12 mesi	2.925	*5,12*	*93,5*	*4,1*	*1,5*	*0,5*	*0,4*
Totale		57.180	*100,00*	*48,50*	*8,80*	*3,20*	*2,71*	*36,80*

5.2 Le determinanti del rischio estremo e intermedio di insolvenza

Anche in questo caso, quasi tutta la variabilità della variabile dipendente risulta spiegata dalla "Presenza di almeno una segnalazione negativa" nella relativa banca-dati.

Per quanto concerne l'informazione relativa al "Peggior ritardo di pagamento" registrato in passato, si osserva che la variabile entra ripetutamente nella soluzione al fine di identificare i gruppi caratterizzati dal rischio intermedio ed estremo.

Rispetto a quanto ottenuto nell'analisi di segmentazione con variabile dicotomica, in questo caso l'informazione relativa ai giorni di ritardo viene ritenuta fra i predittori più significativi anche quando distingue gruppi caratterizzati da ritardi di pagamento inferiori ai 30 giorni (Tabella 5.3).

Nell'analisi di segmentazione finalizzata a determinare diversi livelli di rischio, e non solo il "rischio estremo", i clienti che hanno avuto anche solo qualche giorni di ritardo presentano un rischio relativo decisamente superiore (*OR* = 36,2).rispetto ai clienti che sono sempre stati regolari in tutti gli adempimenti di pagamento (Peggior ritardo = 0 giorni, cioè nessun giorno di ritardo).

Tabella 5.3 *Odds Ratio della variabile "Peggior ritardo" secondo le aggregazioni di modalità con cui entra nella soluzione dell'analisi di segmentazione con variabile ordinale*

Variabile esplicativa	Odds Ratio (OR)
Peggior ritardo >= 60gg vs Peggior ritardo < 60gg	*37,97*
Peggior ritardo di >= 30gg vs Peggior ritardo < 30gg	*36,15*
Peggior ritardo > 0gg vs Peggior ritardo = 0gg	*36,22*

Quando, invece, si sono registrati dei ritardi superiori ai 30 giorni, ma comunque ancora al di sotto dei 60 giorni, diventa discriminante la "distanza temporale" dell'evento rispetto al momento in cui si valuta il cliente: il confronto fra le foglie 4 e 5 mette in luce come il verificarsi di ritardi in tempi molto recenti (meno di 3 mesi) sia quasi dieci volte più rischioso (*OR* = 10,6).

Anche questa soluzione identifica nel numero di 3 la modalità discriminante della variabile indicatrice della dimensione del portafoglio contratti. Nell'albero di classificazione vengono separati i clienti con un numero di contratti al più pari alla media dell'interno campione e i clienti con un'esposizione superiore alla media: i gruppi identificati da questa variabile sono relativi ai clienti con al massimo 3 prodotti e ai clienti con 4 e più prodotti

I clienti con un'esposizione, in termine di linee di credito, superiore alla media sono altamente rischiosi nel 3,8% dei casi, mediamente o poco rischiosi nel 17,6% dei casi e assolutamente non rischiosi nel 78,4% dei casi. I clienti del segmento complementare, ossia quelli con un numero di linee di credito inferiore a 3, sono altamente rischiosi solo nel 2% dei casi, mediamente o poco rischiosi nel 10,7% e assolutamente non rischiosi nell'88,4% dei casi.

L'*Odds Ratio* (OR) = 8,77 mostra come il rischio relativo "intermedio ed estremo" sia quasi nove volte superiore per i clienti con più di 3 finanziamenti attivi.

CAPITOLO 6

Conclusioni

L'evoluzione del mercato del credito, in termini di varietà nella gamma dei prodotti offerti e di allargamento della clientela di base e l'intensificarsi del confronto concorrenziale incentivano gli istituti di credito a ricercare strumenti di valutazione idonei ad identificare anticipatamente il rischio di insolvenza delle controparti ed a ottimizzare, una volta integrati all'interno dei propri sistemi decisionali, la prassi quotidiana di erogazione e di gestione del credito.

La conoscenza delle tipologie di clienti più a rischio risulta funzionale ad impostare lo sviluppo di sistemi di valutazione del rischio di insolvenza che siano più efficaci e precisi nell'intercettare preventivamente quelle situazioni che possono degenerare con una probabilità maggiore verso l'insolvenza.

Data la complessità del fenomeno del rischio di credito, è largamente diffusa l'opinione secondo la quale, per definire efficaci politiche di erogazione e di gestione del credito, non sia possibile prescindere da analisi e strategie di segmentazione della clientela.

L'analisi di segmentazione, da un lato, risponde all'immediato bisogno di dipanare il problema della valutazione del rischio di insolvenza e dell'identificazione delle sue determinanti, e dall'altro offre un primo contributo per orientare l'approfondimento delle successive analisi del rischio di insolvenza.

L'utilizzo della tecnica della segmentazione binaria, adottata nel presente studio, si presenta come un metodo efficace a individuare le determinanti del rischio di insolvenza in situazioni caratterizzate da un

elevato grado di eterogeneità. L'analisi di segmentazione che si è condotta ha permesso di approfondire l'interazione fra le variabili, ossia l'effetto che le modalità combinate di variabili esplicative hanno sulla variabile dipendente, e di individuare i fattori determinanti del rischio, isolandoli da una moltitudine di variabili osservate.

Per la misura del rischio di insolvenza, nell'analisi di segmentazione si è presa in esame, in un primo momento, la variabile dicotomica "ritardo di pagamento di almeno 180 giorni" verso "ritardo inferiore" e, in un secondo momento, una scala di misura più dettagliata definita tramite cinque modalità ordinali. Questo approccio si qualifica con la duplice valenza di aiutare a comprendere se i clienti caratterizzati da livelli di rischio intermedio sono una sottocategoria dei clienti molto rischiosi o solo una categoria "sporca" dei clienti meno rischiosi, e di obbligare quindi a chiarire anche il significato della variabile dipendente.

Le determinanti del rischio messe in rilievo dai diversi approcci sono sostanzialmente le medesime, e le poche differenze sono legate al fatto che alcune variabili risultano significative a livelli diversi.

Innanzitutto, sono stati isolati in un unico gruppo, omogeneo ed estremamente rischioso, i clienti che al momento della valutazione risultano già registrati nella banca-dati degli eventi negativi.

La variabile relativa al peggior ritardo di pagamento ha contribuito ad individuare ulteriori gruppi molto rischiosi: il fattore legato al "massimo numero di giorni di ritardo raggiunti nel recente passato" si evidenzia come significativo in tutte le analisi di segmentazioni e permette l'identificazione dei clienti che hanno già manifestato vari livelli di difficoltà nel rispetto degli adempimenti di pagamento assunti rispetto ai finanziamenti in essere, e che presentano i rischi relativi maggiori.

Tra i clienti con una storia dei pagamenti pregressa abbastanza regolare, quindi con ritardi di pagamento inferiori ai 30 giorni, i livelli di rischiosità creditizia futura e l'eterogeneità del gruppo sono notevolmente inferiori. Nelle diverse applicazioni, si è tuttavia osservato che l'informazione relativa all'ampiezza del portafoglio di prodotti posseduti dal cliente può

ancora apportare un contributo significativo per affinare la classificazione ottenuta. Infatti, i clienti che, come richiedenti principali, possiedono o stipulano più di 3 contratti di finanziamento entro un breve arco temporale, tendono ad aumentare la propria esposizione diretta verso il sistema creditizio e più facilmente possono incorrere in problemi di solvibilità.

E' opportuno ribadire che la base informativa che ci è stata resa disponibile per l'analisi di segmentazione manca di vere e proprie variabili di tipo "strutturale", e questo ha parzialmente limitato le possibilità di inferenza causale e l'identificazione di altre interazioni. Di conseguenza, ci si sente di rivolgere l'invito a voler favorire la documentazione e l'analisi di altri aspetti descrittivi del cliente, in modo particolare delle informazioni sulle sue condizioni economiche e socio-demografiche, che possono contribuire a condizionare la regolare solvibilità dei pagamenti e quindi il rischio del cliente stesso.

Una considerazione aggiuntiva, riguarda le possibilità di approfondire l'analisi della segmentazione. Benché gli algoritmi di segmentazione disponibili sul mercato si equivalgono nella sostanza, alcuni software sono più "avanzati" rispetto ad altri in quanto permettono:

- di gestire i dati di input in modo più efficiente (trattamento dei dati mancanti, definizione delle modalità delle variabili predittive, trasformazione della variabile dipendente,...),
- di impostare opzioni di analisi metodologicamente anche molto specifiche (Schievano, 2002),
- di valutare e confrontare le soluzioni ottenute tramite procedure statistiche (es. *Cross-Validation*) e opportuni indicatori statistici e con il supporto di rappresentazioni grafiche *users friendly*.

In questo modo si amplia lo spettro delle possibilità di indagine e di approfondimento conoscitivo dei fenomeni oggetto di studio.

BIBLIOGRAFIA

Altman E.I. (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy". In: *Journal of Finance,* Vol.23, pp.: 589-609.

Bailey M. (2004), "An introduction to the Principles". In: Bailey M.(a cura di), *Credit Scoring: The Principles and Practicalities*, Windsor, PIC Solution, London, pp.: 1-6.

Bouroche J.M., Tennenhaus M. (1971), *Some segmentation methods*, Metra, 7, pp.: 407-418.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), *Classification and Regression Trees*, Wadsworth Inc., Belmont California.

Du Toit S.H.C., Steyn A.G.W., Stumpf R.H. (1986), *Graphical Exploratory Data Analysis,* Springer-Verlag, New York Inc.

Enache D. (1998), "Künstliche neuronale Netze zur Kreditwürdigkeits-überprüfung von Konsumentenkrediten". In: Bomsdorf E., Kösters W., Matthes W. (a cura di), *Quantitative Ökonomie,* Vol.86, Lohmar, Cologne.

Fabbris L. (1997), *Statistica multivariata, analisi esplorativa dei dati*, McGraw-Hill Libri Italia srl, Milano.

Frydman H., Altman E.I., Kao D.L. (1985), "Introducing Recursive Partitioning for Financial Classification: The case of Financial Distress". In: *Journal of Finance*, Vol.11, pp.: 269-291.

Galindo J., Tamayo P. (2000), "Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modelling Applications". In: *Computational Economics*, Vol.15, pp.: 107-143.

Kass G.V. (1980), "An exploratory technique for investigating large quantities of categorical data". In: *Journal of Applied Statistics*, Vol.29, pp.:119-127.

Kaltofen D., Paul S., Stein S. (2006), *Retail Loans & Basel II, using portfolio segmentation to reduce capital requirements,* European Credit Research Institute, Research Report No.8, Brussels.

Liu Y. (2002), *A framework of data mining application process for credit scoring*, Institut für Wirtschaftsinformatik, Georg-August Universität Göttingen, Nr.01/2002, Göttingen.

Martin D. (1977), "Early Warning of Bank failure – a logit regression approach". In: *Journal of Banking and Finance,* Vol.1, pp.:249-276.

McCullagh P., Nelder J.A. (1989), *Generalised Linear Models,* Cambridge University Press, Cambridge.

Neves E. (2004), "Customer Scoring". In: Bailey M. (a cura di), *Credit Scoring: The Principles and Practicalities*, Windsor, PIC Solution, London, pp.: 122-132.

Schievano C. (2002), "LAID-OUT.1: un programma per l'analisi di segmentazionr binaria con riferimento ad una variabile dicotomica trasformata in logit". In: Puggioni G. (a cura di) *Modelli e metodi per l'analisi dei rischi sociali e sanitari,* Cleup, Padova, pp.: 21-36.

Shannon C.E., Weaver W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois.

Sonquist J.A. (1970), *Multivariate Model Building. The Validation of a Search Strategy*, Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.

Sonquist J.A., Morgan J.N. (1964), *The Detection of Interaction Effects*, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.