Università
Ca'Foscari
Venezia

Corso di Laurea

in Economia e Finanza

ordinamento ex D.M. 270/2004

Tesi di Laurea

Peer to Peer lending: credit scoring e analisi di performance

Relatore
Ch. Prof.ssa Loriana Pelizzon
Prof. Marco Corazza

Correlatore
Prof.ssa De Cian Enrica

Laureando
Tanja Baccega
Matricola 840728

Anno Accademico
2017 / 2018

Alla mia famiglia

Sommario

L'innovazione tecnologica e la grave crisi finanziaria del 2008, hanno spinto enormemente l'evoluzione e l'incremento nel mondo dei Peer-to-Peer lending, soprattutto grazie alla loro capacità di agevolare il collegamento tra debitore e creditore a migliori condizioni di mercato. Focalizzando l'attenzione dello studio sui prestiti P2P personali, in questa analisi si vuole indagare se il rischio associato sia correttamente remunerato e quale sia l'entità della perdita in cui si potrebbe realmente incorrere.

Edit with LaTeX

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Pensando ai termini "credito" e "prestito", nell'immaginario collettivo si fa implicitamente riferimento ad istituzioni *super partes*, quali banche e istituti finanziari. L'innovazione tecnologica ha tuttavia permesso lo sviluppo di *business* finanziari altrimenti impossibili, come i *Peer-to-Peer lending*. Con il termine P2P si identificano quei business che permettono l'iterazione tra due controparti senza la presenza di un intermediario, e questo si traduce in una nuova concezione di prestito, complementare ai prestiti emessi dalle istituzioni finanziarie tradizionali. In questo contesto, il debitore non deve recarsi più presso una banca, ma è sufficiente inserire la richiesta su una piattaforma online che si occupa di incrociare gli investitori disponibili a concedere credito con le richieste di prestito, permettendo di abbattere i tempi e i costi associati all'erogazione. Questo permette la realizzazione di condizioni più favorevoli per tutte le controparti: gli investitori ottengono rendimenti maggiori, i debitori hanno accesso al credito sostenendo interessi minori, la piattaforma riceve commissioni da entrambe le parti ottenendo così un guadagno da intermediazione. Questo ha fatto si che questo modello di business trovasse, nella crisi del 2007, una spinta verso una crescita esponenziale. Ma come spesso accade, non è oro tutto quel che luccica. La facilità di accesso al credito che accompagna questa tipologia di business, ha permesso di erogare finanziamenti anche a controparti altrimenti escluse dall'erogazione di prestiti tradizionali in banca, in quanto generalmente in difficoltà nel sostenere gli interessi dovuti o con un merito di credito troppo basso. Questo si traduce in perdite potenzialmente molto elevate dato l'elevato numero di controparti in gioco, ne deriva quindi la necessità di analizzare attentamente il vantaggio economico derivante dall'investimento in un *pool* di questi prestiti in relazione al rischio complessivo. Interessante, quindi, è analizzare il comportamento di un portafoglio di P2P *lending*. Lending Club, la piattaforma più grande al mondo di prestiti peer-to-peer, posseduta dall'omonima azienda, fornisce gratuitamente i dati dei suoi prestiti, che verranno utilizzati nella tesi con lo scopo di modellizzare una distribuzione di perdita di un portafoglio composto da finanziamenti personali concessi in America dal 2007 al 2010. Dopo aver discusso, nel primo capitolo del mercato P2P, l'analisi si sofferma sul-

l'associazione di una probabilità di default al rating attribuito da Lending Club ai suoi prestiti, attraverso una regressione logistica implementata in MatLab. Per un'analisi completa del merito di credito si procederà anche alla stima dei *recovery rate*, necessari per definire il capitale soggetto effettivamente al rischio di perdita. Attraverso la ricostruzione della scala di rating e alle stime dei tassi di recupero, si potrà poi procedere al'implementazione del modello di portafoglio. In questa sede verrà utilizzato CrediRisk+, creato su logiche attuariali e distribuito dalla Credit Suisse First Boston dal 1996. La scelta è stata dettata dalla mancanza di dati necessari per l'implementazione di altri modelli. CrediRisk+ permette, invece, la sua implementazione con un numero di input minimo, permettendo comunque di considerare implicitamente anche il beneficio della diversificazione e di ottenere una *loss distribution*. Parallelamente all'analisi della distribuzione delle perdite verrà calcolato l'interesse *risk adjusted*, per verificare la remunerazione restituita dall'investimento in questa tipologia di prestiti e il rendimento restituito dal portafoglio. Lo scopo finale della tesi è quello di verificare la sostenibilità economica del mercato peer to peer, resa possibile solamente se le società che si occupano della gestione di questi prestiti applicano un tasso di interesse adeguato al rischio della controparte affidata. Il calcolo del tasso *risk adjusted* sarà necessario al fine di trarre le conclusioni su questo fondamentale aspetto. Come si vedrà nel corso della tesi non sempre l'assunto di un corretto *pricing* del rischio viene rispettato dagli operatori di mercato, in favore di un'ottica maggiormente commerciale volta ad attrarre volumi sempre maggiori di affidamenti.

Capitolo 1

P2P lending

Negli ultimi anni è emersa una sempre maggiore presenza di istituzioni finanziarie "Peer-to-Peer" (P2P) nel mercato globale. Il servizio offre soprattutto: prestiti personali (si pensi a Zopa, Prosper, Lending Club), prestiti a piccole aziende (First Circle, Kabbage), anticipo di fatture e transazioni con l'estero. Il volume di queste attività sta crescendo così rapidamente da catturare l'attenzione delle istituzioni finanziarie, spesso al fine di sviluppare una propria piattaforma. Il mercato ha percepito i P2P lending come una rivoluzione, una possibilità di sovvertire la struttura delle istituzioni finanziarie, così la percezione di "reinventare la banca" ha permesso una ingente crescita nei volumi di questo tipo di servizio. Ma questa non è l'unica conseguenza. Sopra il concetto di *peer-to-peer* e al loro successo, sono stati sviluppati moltissimi modelli di business, come per esempio Airbnb che supporta negoziazioni C2C (*consumer-to-consumer*) al posto della classica struttura B2C (*business-to-consumer*), dando vita a moltissimi nuovi modelli di business. Come verrà discusso in seguito, le piattaforme P2P offrono moltissimi vantaggi competitivi. Per esempio, consentono l'applicazione di bassi margini grazie ai bassi costi amministrativi, maggiore trasparenza e flessibilità grazie alla struttura snella. Ad ogni modo, più che una distruzione del concetto di banca, i P2P *lending* dovrebbero essere visti come un servizio complementare al modello di business bancario tradizionale (Tang 2017). Per questo motivo ci si aspetta che le banche, nel futuro, coopereranno sempre di più con questi servizi offerti da terze parti, piuttosto che competerci, nonostante il pieno sviluppo del settore sia possibile solo indirizzando anche l'interesse dei *regulators* verso il mondo dei P2P, soprattutto per la gestione del rischio ed una regolamentazione appropriata. In altre parole, l'abilità con cui le piattaforme di prestito P2P riescono a gestire adeguatamente l'attività creditizia dipende in gran parte dalla modifica e dall'adattamento dei processi esistenti, compresa l'implementazione di standard di settore, non di meno dall'applicazione

di regole al fine di supportare questa nuova forma di intermediazione, soprattutto nella gestione del rischio assunto. Secondo «The business models and economics of peer-to-peer lending» di Milne e Parboteeah, l'80% dei problemi riguarda il processo aziendale e i rischi impliciti, mentre il restante 20% si riferisce alla gestione tecnologica. Dopo una veloce analisi della tipologia di prodotto, l'oggetto del capitolo è rappresentato dall'analisi dei P2P nel mercato americano attraverso l'analisi dei dati messi a disposizione da LendingClub. I dati sono stati scaricati dal sito LC in relazione ad un periodo compreso tra il 1 Gennaio 2007 e il 31 Dicembre 2017, fornendo delle statistiche sulla composizione del portafoglio, dell'ammontare erogato, del tasso di interesse e del merito di credito delle controparti che hanno accesso a questi prestiti.

1.1 Cosa sono i peer-to-peer lending

Origine Il termine "*peer-to-peer*" descrive l'iterazione tra due parti senza la necessità di intermediazione. La crescita di internet e la possibilità di facilitare la disintermediazione tra le parti, hanno sostenuto la nascita di specifici servizi P2P, come per esempio il trasferimento di file tra utenti in BitTorrent. La storia del P2P *lending* nel settore finanziario nasce insieme alla nascita di due aziende: l'inglese Zopa nel 2005 e l'americana Prosper nel 2006. Entrambe hanno facilitato l'incontro tra debitori e creditori direttamente, creando un mercato centrale senza la necessità di chiedere intermediazione ad un istituzione finanziaria. Oltre a questo servizio se ne sono sviluppati altri: (i) *crowfounding*, in cui una somma viene raccolta tra un gruppo di creditori per un progetto specifico; (ii) piattaforme alternative di valute estere, dove le controparti si scambiano valute; (iii) anticipi di fattura senza intermediazione bancaria, con cui un'azienda può migliorare la propria liquidità chiedendo un anticipo sulle fatture; (iv) cryptovalute, come i Bitcoin, che supportano i pagamenti online con valute digitali senza intermediazione finanziaria. Tutte queste forme alternative di finanza hanno dei tratti "*peer-to-peer*".

Vantaggi La rapida espansione delle piattaforme P2P, che hanno continuato a raddoppiare il volume erogato di anno in anno nell'ultimo decennio, affiancata a costi e benefici percepiti, ha attirato l'attenzione di molti esperti. Ci sono diverse ragioni a sostegno di questa espansione. L'innovazione tecnologica è sicuramente una delle motivazioni fondamentali poiché permette alle controparti di comunicare direttamente. Ma le ragioni si estendono ad altre caratteristiche. I benefici si trovano in migliori tassi di interesse se paragonati ai tassi applicati dalle istituzioni finanziarie, sia per il creditore che per il debitore; i P2P permettono l'erogazione

di credito a categorie di debitori che non avrebbero altrimenti accesso al credito bancario. Nel mercato, inoltre, esiste la percezione che le piattaforme P2P siano maggiormente responsabili verso il valore sociale, mentre l'innovazione tecnologica permette un migliore e più veloce servizio per entrambe le parti. Secondo Milne e Parboteeah (2016), nel periodo 2010-2015, i creditori hanno ottenuto migliori guadagni da investimenti P2P rispetto a quelli che avrebbero ottenuto investendo i soldi in banca con prodotti tradizionali. Questo è reso possibile perché i costi delle piattaforme sono più bassi rispetto alle strutture bancarie. Inoltre, permettono l'incontro tra domanda e offerta senza applicare un margine di intermediazione. Quindi, sebbene il creditore di un finanziamento P2P si esponga ad un rischio maggiore (non essendoci depositi cauzionali o garanzie), viene comunque compensato da un maggiore tasso di remunerazione. Il secondo aspetto menzionato è la possibilità di affidare anche controparti molto rischiose che, altrimenti, non sarebbero finanziate dalle banche. Dalla crisi finanziaria globale sono stati applicati criteri di erogazione del credito molto più stringenti rispetto al passato, che lasciano scoperta una fetta di mercato che trova la possibilità di incontrare un investitore disponibile a concedergli un finanziamento solo nei *peer-to-peer lending*, ottenendo così un prestito ad un tasso inferiore rispetto agli strumenti tradizionali. Un altro fattore menzionato è la percezione sociale che i servizi di P2P, attraverso l'incontro diretto tra domanda e offerta, possano creare un beneficio finanziario alla società rispetto ai canali tradizionali, incolpati di non avere abbastanza riguardo per gli interessi dei propri clienti. L'ultimo vantaggio riguarda l'aspetto tecnologico. Le banche spendono ingenti quantità di denaro per mantenere in efficienza il sistema informatico in essere, piuttosto che innovarlo. Le aziende di P2P invece, possono competere con i servizi informatici degli istituti di credito perché utilizzano le nuove tecnologie, senza avere la necessità di innovare mantenendo il contatto con il sistema informativo preesistente. Questo fa si che queste società possano offrire un servizio migliore sia per il debitore, che per il creditore, oltre ad offrire nuovi schemi di investimento. Tutte le piattaforme di prestito P2P impiegano due metodi di diversificazione simili: attraverso un asta online o secondo un abbinamento automatico.

Asta online I debitori indicano quale tasso di interesse massimo sono disposti a pagare sul prestito ricevuto, mentre i creditori indicano, a loro volta, il tasso di rendimento minimo disposti ad accettare per classe di rating associata ai debitori. La piattaforma conduce una sorta di "asta inversa" abbinando le controparti con tassi simili, aumentando il tasso di interesse pagabile sino a coprire l'intera somma richiesta. Il tutto assoggettando gli abbinamenti al requisito di diversificazione

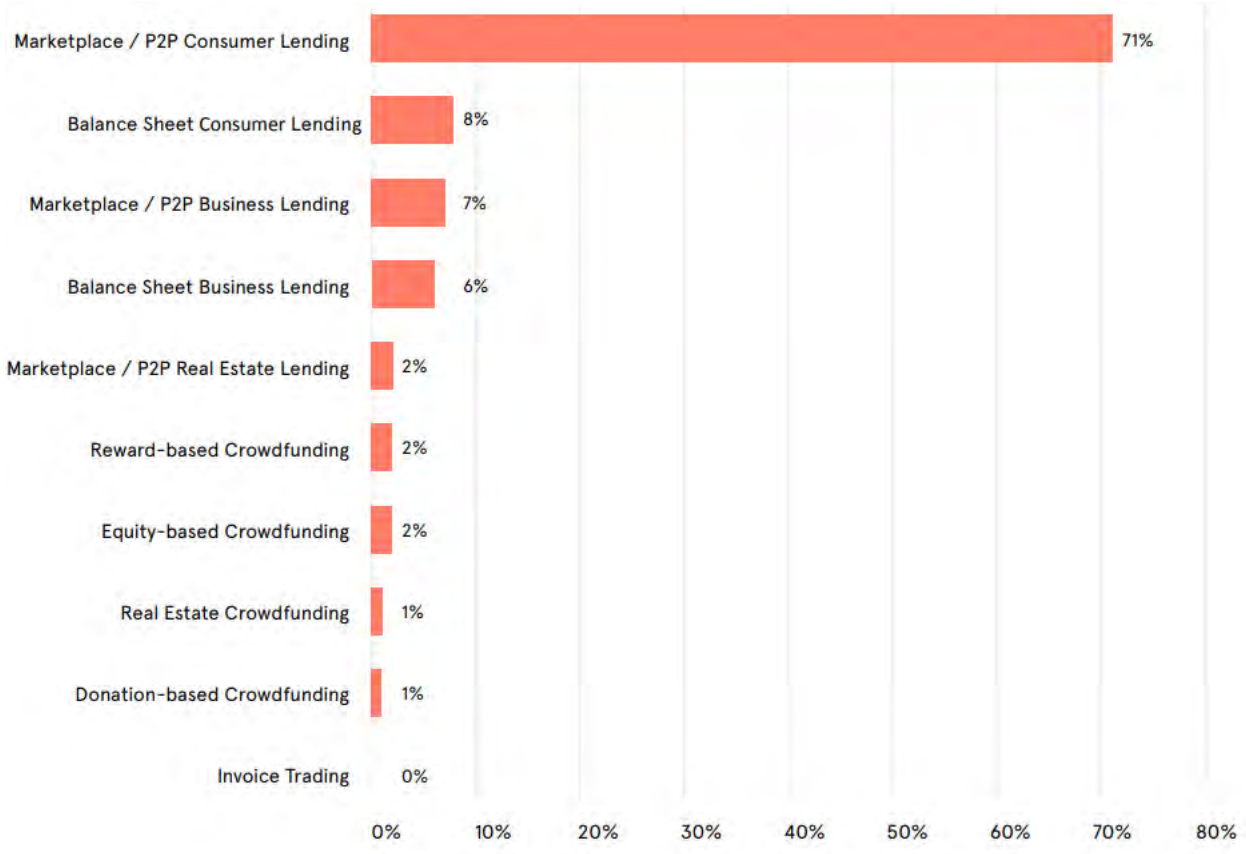
per non esporre i creditori al rischio di concentrazione. Se il tasso di interesse debitore è uguale o inferiore al tasso dichiarato dal debitore il prestito viene erogato, altrimenti viene rigettato.

Abbinamento automatico L'abbinamento avviene secondo i tassi di interesse di mercato annunciati dalla piattaforma per classe di rating. Questa metodologia, sebbene più facile da capire, può creare ritardi nel *matching* dei prestiti a seconda della numerosità delle controparti. La piattaforma può, comunque, regolare i tassi di interesse nel tempo al fine di eliminare questi *gap*.

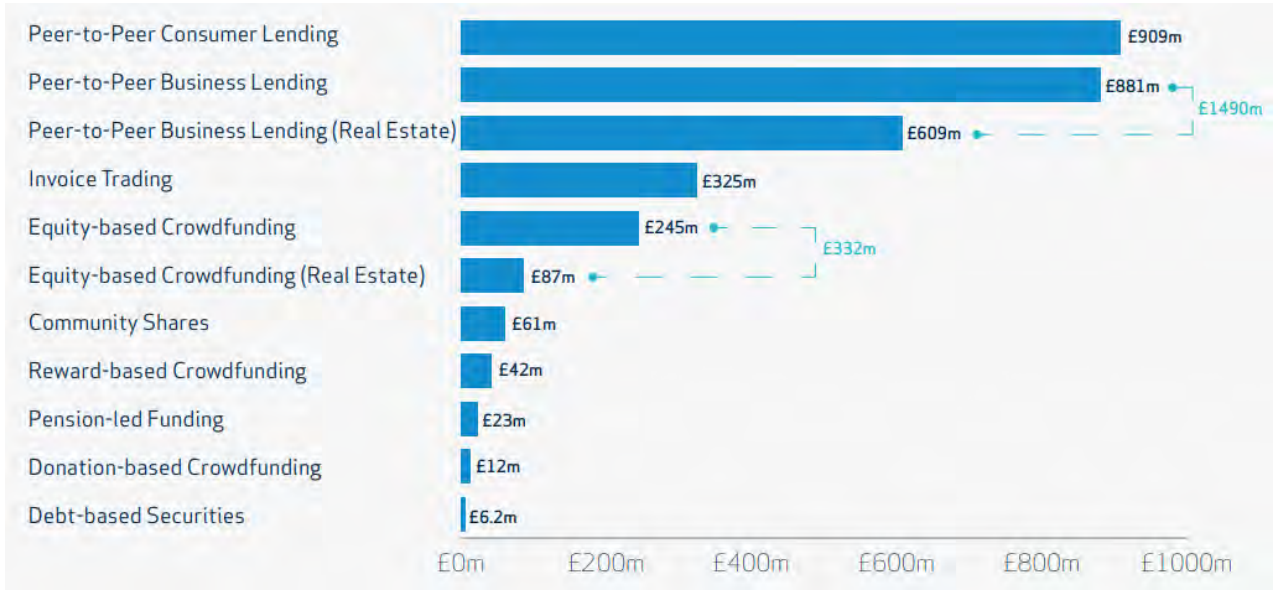
Mercati A questo punto è necessario analizzare il mercato di questi prodotti, focalizzandoci sul mercato americano. Gli Stati Uniti sono stati, insieme al Regno Unito, i pionieri dello sviluppo dei *peer-to-peer lending*. Rispetto al Regno Unito, le società americane sono più concentrate sul credito al consumo. I dati proposti da «Breaking new ground: the Americas alternative finance benchmarking report» e «Pushing boundaries: The 2015 UK alternative finance industry report», riassunti nella figura 1.1 nella pagina seguente[1], dimostrano come il volume dei prestiti al consumo americani siano dieci volte maggiori rispetto ai prestiti erogati alle piccole aziende, mentre le due categorie per i prestiti inglesi riportano capitali simili. Molte delle piattaforme americane hanno sviluppato nel tempo una *partnership* con le banche, in quanto, fornendo un nuovo modo di investire e utilizzando una tecnologia all'avanguardia, sono percepite sul mercato più come un'opportunità che come *competitors*. Un importante tratto distintivo della tecnologia impiegata nelle piattaforme P2P, è la dipendenza dall'analisi di "*big data*". Viene svolta grazie a tecniche informatiche sofisticate al fine di migliorare le misurazioni americane standard del merito di credito , come il Fair-Isaacs FICO score, e rappresentano un enorme opportunità per le banche tradizionali, oltre a rappresentare un'attività strategica per la competizione sul mercato con le altre piattaforme. La "simbiosi" nata tra piattaforme P2P e banche permette anche il rispetto di limiti normativi sui tassi di interesse di prestiti al consumo applicabili. Per superare questi controlli normativi molte società P2P erogano formalmente i prestiti attraverso banche tradizionali, dopo l'accordo delle controparti attraverso la piattaforma. Ad esempio Lending Club collabora con WebBank (sede nello Utah) prima di erogare i prestiti accordati attraverso la piattaforma [2].

Nonostante il mercato americano guardi con positività la presenza di strumenti

[1] Rispettivamente tratte da Wardrop et al. 2016, p. 31 e Zhang et al. 2016, p. 14
[2] Autorizzazioni LC



(a) Ammontare per tipologia di prestito U.S.A.



(b) Ammontare per tipologia di prestito U.K.

Figura 1.1: Confronto capitale erogato per tipologia di prestito

come i *peer-to-peer lending*, non mancano delle preoccupazioni, riservate soprattutto a soggetti finanziari più piccoli. Questo perché i debitori che accedono a questa forma di finanziamento alternativa (consumatori e piccole aziende), sono il *target* di mercato che, prima, accedeva al credito tramite le istituzioni più piccole. Alcuni istituti indipendenti, quindi, sono preoccupati della possibile migrazione della clientela a favore di queste piattaforme.

Modello di Business Ci sono diverse motivi sul perché i servizi finanziari sono sempre forniti congiuntamente da un istituto finanziario e non sono, invece, forniti attraverso aziende terze. Alla base di queste strutture c'è il problema della gestione della liquidità. I clienti proprietari di depositi bancari hanno il diritto di prelevare denaro su richiesta, inoltre un potenziale cliente per scegliere la banca apprezza anche la flessibilità nell'utilizzo dei servizi di prestito. Spesso un cliente vuole rimborsare il prestito in un momento a sua scelta al fine di evitare costi di interesse inutili. Sono disposti ad accettare tassi di remunerazione inferiori o maggiori costi di indebitamento in cambio di questi servizi, che potrebbero essere definiti "di liquidità"; ma questa libertà ha un costo che grava sull'istituto finanziario. Una banca è capace di far fronte al costo di questa struttura solo sfruttando le economie di scala, l'incertezza di questi flussi si riduce sostanzialmente solo a livello di portafoglio, abbassando il *cost of funding* e massimizzando i rendimenti ottenuti sui prestiti erogati. Per massimizzare il beneficio delle economie di scala una banca deve essere anche esperta nel valutare il merito di credito delle controparti, e quindi avere anche un vantaggio competitivo nella fornitura di prestiti attraverso una migliore gestione dei flussi di cassa e dello *screening* delle controparti. Anche la regolamentazione è un fattore a sostegno dell'erogazione congiunta di diverse servizi. Perché la disintermediazione in atto sul mercato sia veramente *peer-to-peer*, e soprattutto sia ca pace di arrivare a sostituire le banche, devono esserci benefici per entrambi le controparti di un finanziamento - debitore e creditore - sia in termini di prezzo sia in termini di miglioramento del servizio offerto. Tali benefici devono essere tali da far rinunciare alle controparti al servizio di liquidità come sopra riportato (flessibilità nell'uso del credito, prelievi e depositi). L'esperienza passata ha dimostrato come sia più difficile convincere i creditori dei benefici esistenti, piuttosto che i debitori, così le piattaforme P2P hanno dovuto ricorrere a investimenti istituzionali per far fronte alle esigenze aziendali. Concedere denaro attraverso una piattaforma P2P, per un creditore significa rinunciare alla protezione sui depositi e l'utilizzo di un prodotto non familiare o standard di cui non comprende a pieno il rischio. I rendimenti, comunque, sono molto interessanti e quindi gli investitori istituzionali continuano a finanziare le piattaforme, richie-

dendo però anche una trasparenza maggiore. Anche superando questi problemi, comunque, le piattaforme di P2P continuano ad essere un servizio complementare rispetto alle banche piuttosto che concorrenziale. Le banche stesse cercheranno di mantenere i clienti che utilizzano le piattaforme di prestiti fornendo esse stesse piattaforme simili o collaborando con esse.
Secondo Caratelli et al. (2016), sebbene ci siano elementi di elevata specificità tra piattaforme di P2P, è possibile identificare alcuni tratti caratterizzanti che permettono di suddividere questo servizio in tre modelli di business: un primo modello è denominato *client segregated account*; un secondo modello può essere identificato come *notary*; infine l'ultimo modello è chiamato *guaranteed return*. Nel modello *client segregated account*, la piattaforma all'interno del portale si occupa solo del *matching* tra creditore e debitore. I fondi da erogare, invece, vengono raccolti in un patrimonio separato, al fine di non intaccare i rapporti tra controparti utilizzanti il servizio qualora fallisse il gestore della piattaforma. In questo modello, i finanziatori offrono i fondi secondo un meccanismo ad asta e la piattaforma percepisce una commissione da entrambe le controparti per coprire i costi sostenuti dalla gestione. Una variante di questo modello è la raccolta dei finanziamenti in un fondo comune, le cui quote vengono sottoscritte appunto, dai finanziatori attraverso il portale. Nel modello *notary*, analogamente al modello appena descritto, si mantiene l'abbinamento tra le controparti attraverso la piattaforma ma i fondi vengono raccolti direttamente tra i finanziatori/creditori secondo la logica "*all o nothing*" ed erogati direttamente dalla banca depositaria utilizzata per la raccolta. La piattaforma eroga dei certificati a favore dei creditori, che attesta l'investimento e a cui viene riconosciuta la natura di valore mobiliare. L'ultimo modello, il *guaranteed return model*, affida maggiore importanza alla piattaforma, assumendola come intermediario finanziario. Questa importanza fondamentale deriva dal fatto che è la piattaforma stessa a raccogliere i fondi dai finanziatori, a cui applica un tasso di remunerazione garantito, calcolato in base al rischio che il creditore supporta. Raccolti i fondi, questi vengono erogati ai debitori. Oltre al modello di business, le piattaforme si distinguono in base alla protezione offerta in caso di perdite; ecco perché si è parlato anche di tasso di remunerazione garantito. Sotto questo profilo si trovano: (i) *unsecured platform*, nel caso in cui non siano previste tutele per il finanziatore; (ii) un gruppo ibrido, che rappresenta le piattaforme che pretendono garanzie reali o personali a copertura del debito; infine ci sono (iii) le *protected platform*, in cui viene costituito un fondo di mitigazione del rischio per rimborsare i finanziatori che non ricevono i pagamenti accordati. Parlando di *peer to peer lending*, ovviamente non si incontrano solo vantaggi, ci sono anche dei rischi potenziali da affrontare:

- Le istituzioni finanziali hanno già una serie storica di informazioni riguardanti i default, questo gli permette di creare appositi meccanismi di compensazione, mentre le piattaforme P2P hanno ancora lavoro da fare per la quantificazione delle perdite e per l'educazione degli investitori sui rischi assunti.

- Gli istituti tradizionali hanno già massimizzato le competenze per il recupero dei crediti defaultati. I tassi di recupero sono molto legati alle competenze acquisite e le piattaforme di P2P non hanno le stesse competenze delle banche nel ridurre al minimo le perdite.

- Il mercato è costretto ad affrontare tutte le perdite di un possibile fallimento di una piattaforma P2P. Cosa potrebbe succedere ai prestiti di una piattaforma dismessa?

- Gli investitori potrebbero essere costretti a subire una variazione dei tassi di interesse, derivante da decisioni delle istituzioni finanziarie che finanziano le piattaforme. Le istituzioni che investono nei P2P possono trovarsi di fronte alla modifica della propria esposizione, di conseguenza, per rimanere come investitore della piattaforma il tasso di rendimento deve essere competitivo con quello delle attività alternative. Per questo motivo potrebbe essere aggiustato.

- Infine, il cybercrimine è un problema ancora tangibile.

Una soluzione potrebbe essere la standardizzazione del settore, l'erogazione di prestiti standardizzati sotto i profili più importanti permetterebbe una migrazione dei prestiti da una piattaforma all'altra in caso di necessità, l'implementazione di misure di rischio comparabili e l'utilizzo di agenzie di raccolta per il recupero degli insoluti, migliorando la performance.

Possiamo concludere questa analisi sulle caratteristiche dei *peer-to-peer lending* e del mercato americano, riportando le caratteristiche di Lending Club.

1.2 Lending Club

LendingClub è la piattaforma *peer-to-peer lending* più grande del mondo, appartenente all'omonima società creditizia statunitense con sede a San Francisco (California). È stato il primo istituto di credito P2P a registrare, con la Securities and Exchange Commission (SEC), le proprie offerte come titoli e ad offrire prestiti su un mercato secondario. La piattaforma consente, a chi lo richiede, di ottenere un

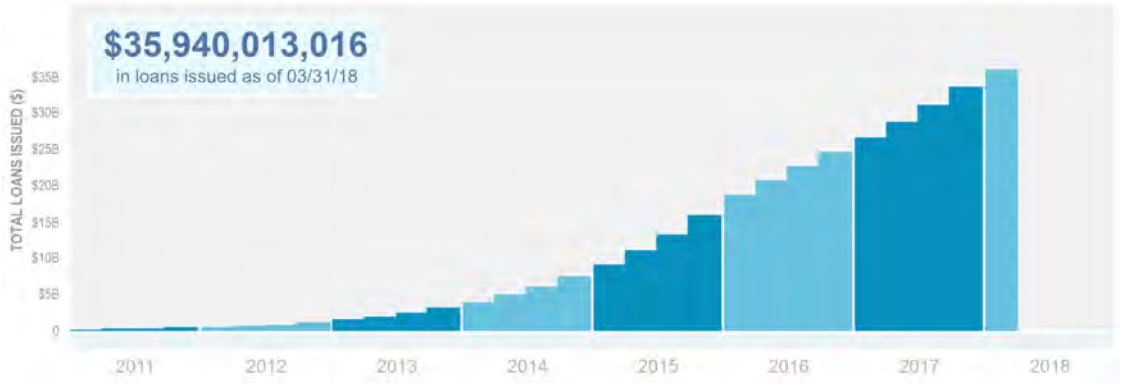


Figura 1.2: Ammontare erogato dalla piattaforma Lending Club

prestito finanziato da investitori nelle vesti di creditori, i quali acquistano "titoli" sostenuti dalle rate di rimborso dei prestiti erogati.

Lending Club consente ai richiedenti di ottenere prestiti personali non garantiti tra 1 000$ e 40 000$, la società ha erogato 35 miliardi di dollari fino al primo trimestre 2018 (figura 1.2) [3]. Il periodo di prestito standard è di tre anni, ma vi è la possibilità di scegliere una scadenza a cinque anni. Gli investitori possono selezionare i prestiti su cui desiderano investire in base alle informazioni fornite sul richiedente attraverso la piattaforma. Oltre a decidere il merito di credito delle controparti a cui erogare, possono decidere anche quanto capitale finanziare ad una singola controparte, con un investimento minimo di 25$. Gli investitori guadagnano dagli interessi ricevuti, Lending Club crea profitto trattenendo una *commissione di origine* fissa all'1% dai debitori, e una *commissione di servizio* dagli investitori, che varia dall'1.1% al 5% dell'ammontare del prestito. I debitori guadagnano sui tassi di interesse più contenuti rispetto a quelli bancari, qualcuno beneficia anche della possibilità di ottenere un finanziamento che altrimenti non gli sarebbe mai stato erogato dagli istituti di credito convenzionali. Sulla base di alcune caratteristiche della controparte, come lo FICO score, la storia creditizia, l'ammontare del prestito desiderato, nonché lo scopo per il quale è richiesto e del rapporto DTI (debt-to-income ratio)[4], Lending Club determina se il richiedente è meritevole e assegna ai prestiti approvati un rating su cui vengono determinati il tasso di interesse e le commissioni da pagare. Va sottolineato che sebbene Lending Club permette un vantaggio sui tassi, sia debitori che creditori, dichiarando un guadagno medio degli investimenti tra il 10% e il 15% (figura 1.3 nella pagina seguente [5]), la tassazione gioca a svantaggio. Poiché si tratta di prestiti personali alle persone, i guadagni sono tassabili come reddito personale invece di reddito da

[3] Statistiche LC
[4] definizione DTI
[5] Rendimenti LC

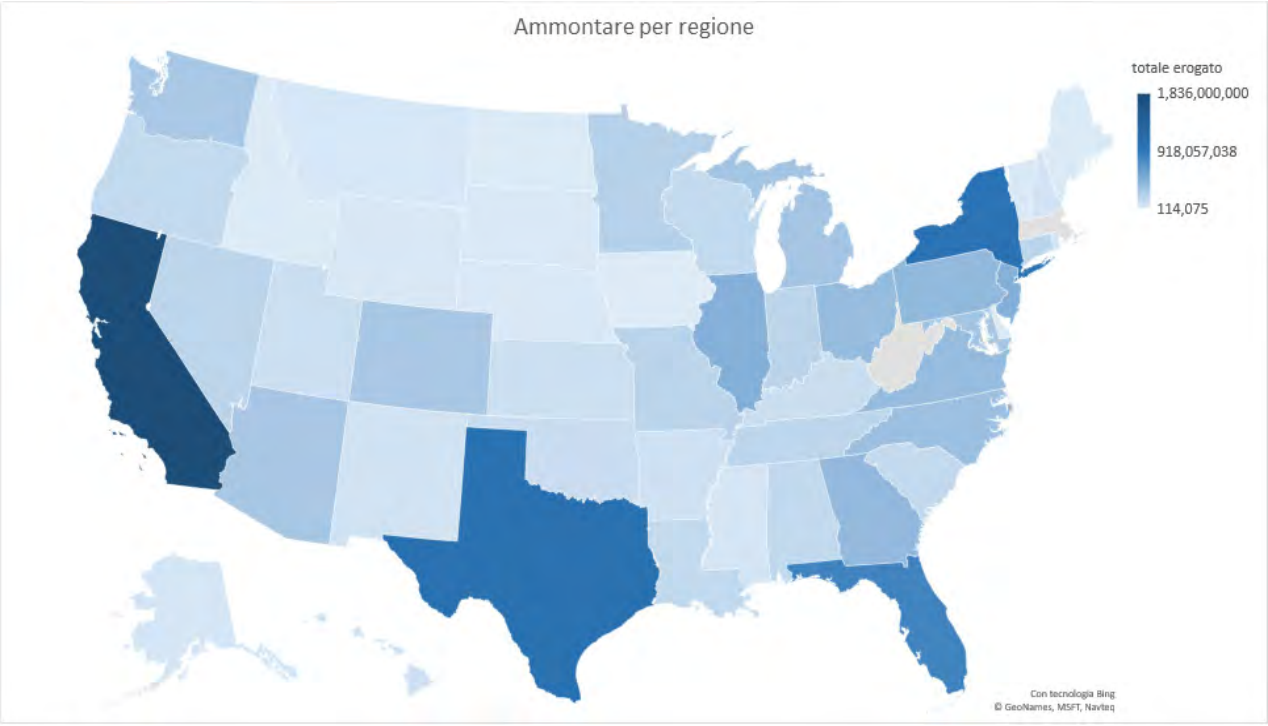
Figura 1.3: Interesse medio ponderato, per classe di rating

investimento, pertanto, i guadagni derivanti da Lending Club possono essere erosi da un tasso superiore rispetto al tasso di *capital gains*. Dalla nascita del servizio e dopo i problemi iniziali affrontati nel 2016 per attrarre capitali, Lending Club raggiunge gli investitori di 39 stati degli Stati Uniti, mentre i richiedenti possono provenire da tutti gli Stati Uniti, esclusi due stati. Nella figura 1.4 nella pagina successiva, vengono riportati i capitali erogati per stato con annessa frequenza di default. È palese che in alcuni stati i prestiti abbiano frequenze di default maggiori rispetto ad altri, da ponderare per l'ammontare erogato. Per esempio la maggior parte del capitale erogato è emesso in California, dove le insolvenze sono più basse rispetto ad altri stati, traducendosi in perdite di portafoglio proporzionali più contenute.

Per ridurre il rischio di insolvenza, Lending Club si concentra su controparti con alto merito di credito, rigettando approssimativamente il 90% delle richieste e assegnando tassi di interesse più elevati ai debitori più rischiosi. Solo i debitori con FICO score di almeno 660 punti possono ricevere il prestito.

Questi finanziamenti vengono richiesti per le più svariate motivazioni: auto, consolidamento debiti e carte di credito, istruzione, casa, acquisti importanti, salute, piccole aziende, altro, come per esempio vacanze, trasferimenti e spese mediche. Dalla figura 1.5 a pagina 14, risulta che l'80% dei prestiti viene erogato per consolidare debiti esistenti e ripagare le carte di credito, un 7% viene richiesto per la casa mentre un altro 7% indica altre motivazioni. Altri scopi, come auto, istruzione, acquisti importanti, salute o finanziamento di piccole aziende sono solo una parte minore dei prestiti erogati. A questo punto sembra necessario focalizzare l'attenzione sul consolidamento dei debiti e sul rifinanziamento delle carte di credito. Lending Club permette di richiedere un finanziamento per coprire i debiti pree-

(a) Stato dei prestiti per regione geografica



(b) Ammontare dei prestiti per regione geografica

Figura 1.4: Distribuzione ammontare e default per regione geografica

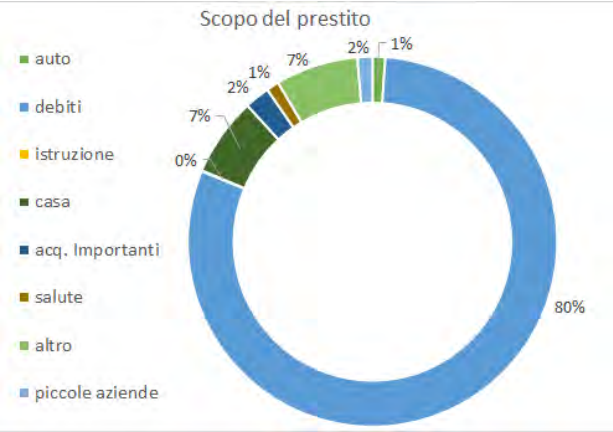
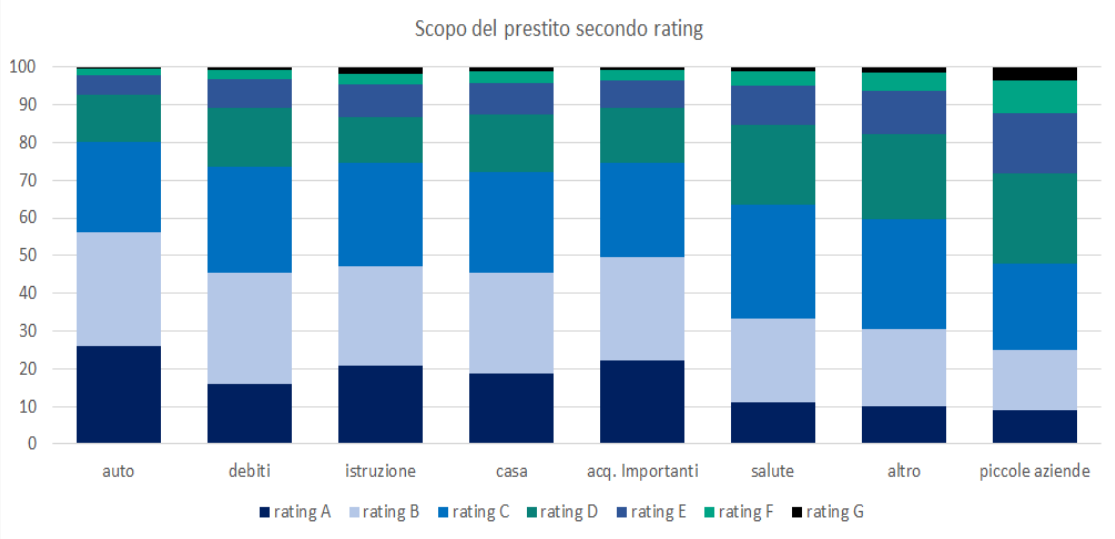


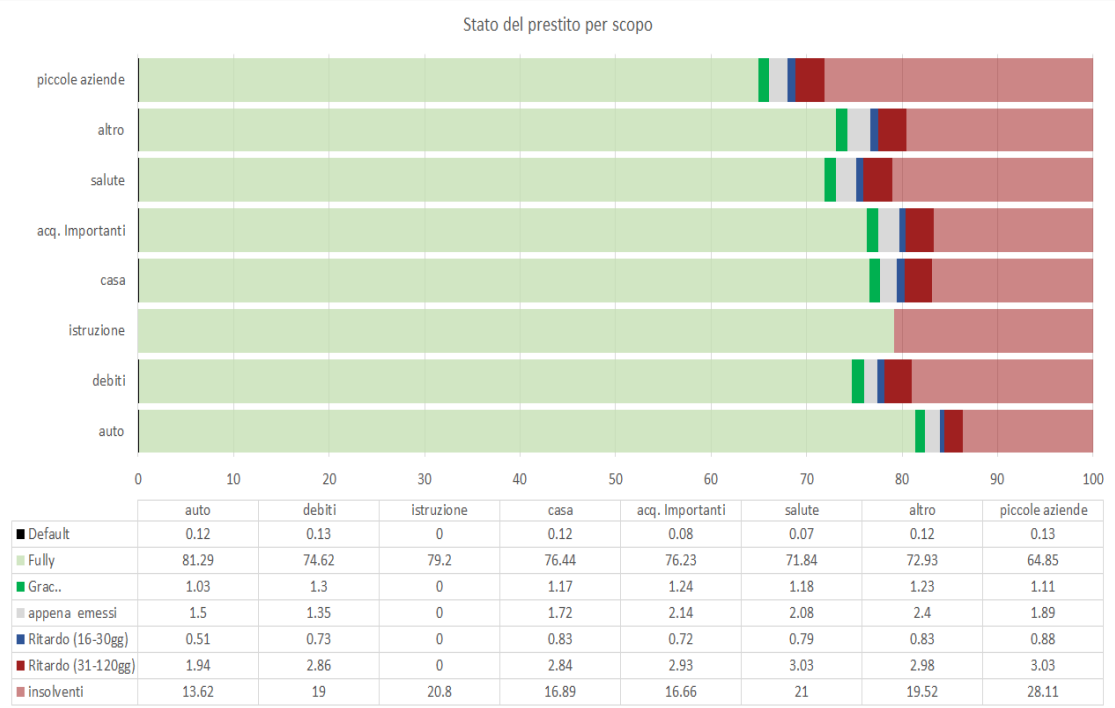
Figura 1.5: Motivi del finanziamento

sistenti. Questo permette al debitore di semplificare gli impegni finanziari in un unica rata a tasso fisso. Un altro vantaggio deriva dal tasso di interesse applicato, che come abbiamo detto, è più vantaggioso dei tassi applicati sugli stessi prestiti dalle banche. Ovviamente, questo potrebbe non dare un vantaggio economico se i tassi di mercato dovessero improvvisamente scendere, ma permette di avere una maggiore gestione e consapevolezza dei flussi di cassa in uscita e soprattutto evita qualsiasi aumento del tasso debitore qualora i tassi subissero un rialzo. Lo stesso motivo riguarda le carte di credito. La richiesta di un prestito da LC per coprire l'indebitamento attraverso una carta, spesso si traduce in un minore tasso di interesse [6]. Molto interessante è analizzare lo stato dei prestiti per motivazione. Già dalla figura 1.6a nella pagina successiva si evince che le classi più rischiose di rating sono concentrate nelle motivazioni "salute" "piccole aziende" e "altro", quindi è naturale che la maggior frequenza di default si verifichi sul gruppo di prestiti richiesti per questi scopi (figura 1.6b nella pagina seguente). Il maggior rischio si riflette anche nella situazione dei ritardi. Le stesse categorie presentano una percentuale di prestiti con pagamenti in ritardo superiore rispetto alle categorie rimanenti.

[6] Consolidamento dei debiti



(a) Merito di credito per motivazione



	auto	debiti	istruzione	casa	acq. Importanti	salute	altro	piccole aziende
Default	0.12	0.13	0	0.12	0.08	0.07	0.12	0.13
Fully	81.29	74.62	79.2	76.44	76.23	71.84	72.93	64.85
Grac..	1.03	1.3	0	1.17	1.24	1.18	1.23	1.11
appena emessi	1.5	1.35	0	1.72	2.14	2.08	2.4	1.89
Ritardo (16-30gg)	0.51	0.73	0	0.83	0.72	0.79	0.83	0.88
Ritardo (31-120gg)	1.94	2.86	0	2.84	2.93	3.03	2.98	3.03
insolventi	13.62	19	20.8	16.89	16.66	21	19.52	28.11

(b) Stato dei prestiti per scopo

Figura 1.6: Percentuale di richieste e stato dei prestiti secondo finalità

1.3 Il Dataset

L'analisi sarà svolta sul dataset di Lending Club, composto da prestiti *peer-to-peer* (P2P), che, come già detto, è una società americana con sede a San Francisco (California). È stata la prima società di prestiti P2P a registrare la sua offerta come titoli con la *"Securities and Exchange Commission"* (SEC), ed a offrire negoziazioni di prestiti in un mercato secondario. Attraverso la piattaforma online, Lending Club fornisce denaro ai consumatori o a piccole aziende attraverso un servizio online, con un incontro diretto tra domanda e offerta, trattenendo una piccola commissione da entrambe le parti. La società fornisce online moltissime informazioni sul suo operato, incluso un dataset gratuito con tutti i prestiti personali erogati tramite la piattaforma,invece non sono accessibili le informazioni sui prestiti alle aziende. Per ogni prestito vengono inserite le caratteristiche che descrivono la controparte, escludendo solo quelle protette da privacy, insieme ad un dizionario esplicativo delle variabili presenti. L'azienda permette di prendere a prestito somma da 1 000$ a 40 000$, scegliendo il rimborso tra 36 o 60 mesi. Lending Club è la piattaforma più grande e importante al mondo, quindi può essere considerata un soggetto rappresentativo del mercato americano dei *peer-to-peer*. L'azienda fornisce i dati suddivisi in due dataset:

- ***Loan Data***, raccolta dei prestiti concessi con annesse informazioni e stato dei pagamenti;
- ***Declined Loan Data***, dataset contenente le richieste rigettate, che non verrà preso in considerazione.

Nell'analisi si considerano i dati a partire dal 01 Gennaio 2007 fino al 31 Dicembre 2017 dei dataset dei prestiti erogati, in cui i dati sono aggiornati ogni quadrimestre, in questo modo si ha sempre lo stato dei prestiti attuale. Nell'appendice A a pagina 115 si riporta una spiegazione di tutte le variabili accessibili. A causa della grossa parte di prestiti ancora in corso si suddividerà il dataset in due parti: si utilizzeranno i dati dal 1 Gennaio 2007 al 31 Dicembre 2014 come campione *in sample*, mentre si utilizzeranno gli anni 2015, 2016 e 2017 come *out of sample* per la validazione dei modelli implementati. Per questo motivo si analizzeranno statisticamente solo gli anni *in sample*. Composto il dataset, quindi, si analizzano le variabili più significative.

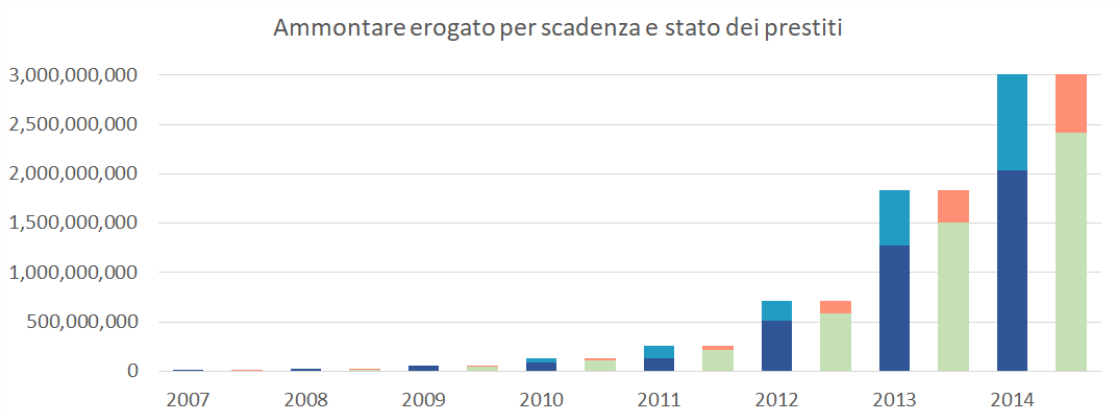
Volume e tasso di interesse Abbiamo già visto, nella figura 1.2 a pagina 11, l'enorme espansione che hanno avuto questi prestiti negli anni. Indicizzando la

variazione del capitale erogato sull'anno 2007 (base 100), si evidenza che la forte espansione del servizio è cresciuta più che proporzionalmente (figura 1.7a nella pagina successiva). Nel 2014 si è arrivati a erogare più di seicento volte il capitale erogato nel 2007. Di questi prestiti, però, è interessare notare quanto capitale si tramuta in perdita, in figura 1.7b nella pagina seguente. Nella figura 1.7c nella pagina successiva si nota la causa della variazione della frequenza dei default. La composizione di portafoglio resta, in linea generale, più o meno la stessa, ma in alcuni anni aumenta sensibilmente la presenza delle classi più rischiose a discapito delle controparti con merito di credito migliore. Questo aumenta la probabilità di default media di portafoglio. Interessante è l'analisi dei tassi di interesse e dell'ammontare dei prestiti erogati secondo scadenza. Le variabili da considerare sono:

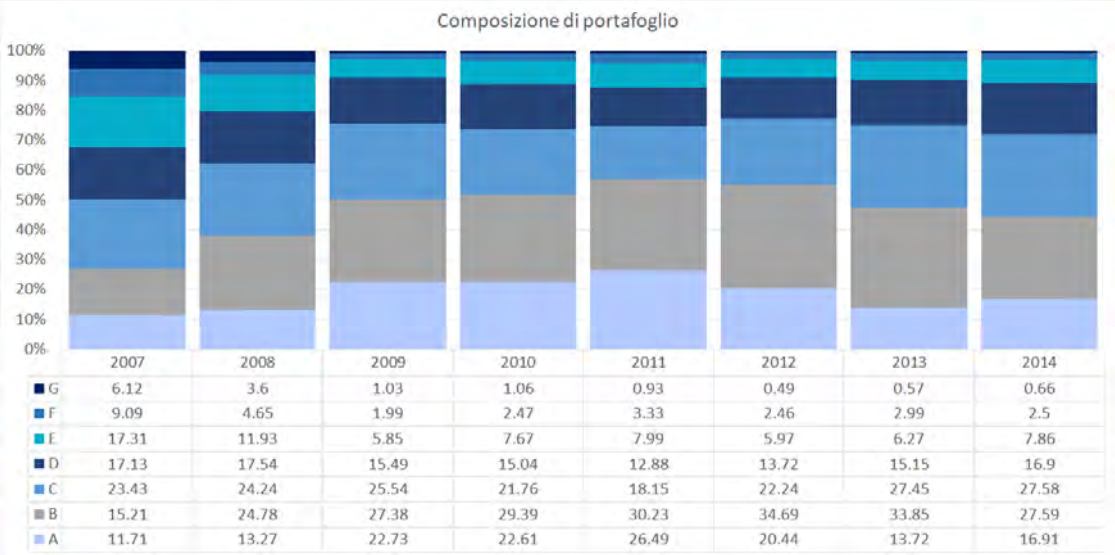
- **term**: la scadenza del prestito, si divide in due categorie, 36 o 60 mesi. Indica anche il numero di pagamenti che effettuerà la controparte.
- **loan_amnt**: lista del capitale erogato. LC eroga prestiti con un capitale da 1 000$ a 40 000$.
- **int_rate**: tasso applicato ai prestiti. È calcolato giornalmente, basato sull'anno commerciale. Lending Club assegna ad ogni potenziale debitore un rating e di conseguenza un interesse.

Si riportano, nella tabella 1.1 a pagina 19, i capitali erogati per ogni anno e suddivisi secondo scadenza del prestito.

(a) Variazione capitale erogato (indicizzazione 2007, base 100)



(b) Composizione del capitale erogato e stato dei prestiti



(c) Composizione del capitale erogato e stato dei prestiti

Figura 1.7: Composizione del capitale e del portafoglio

Tabella 1.1: Ammontare erogato e tasso applicato per classe di rating

Anno		Scadenza		Totale
		36m	60m	
2007	capitale	4819 275		4 819 275
	obs.	572		572
2008	capitale	21 100 000		21 100 000
	obs.	2 389		2 389
2009	capitale	51 700 000		51 700 000
	obs.	5 267		5 267
2010	capitale	89 400 000	42 100 000	131 500 000
	obs.	9 128	3 368	12 496
2011	capitale	133 000 000	129 000 000	262 000 000
	obs.	14 093	7 613	21 706
2012	capitale	507 000 000	210 000 000	717 000 000
	obs.	43 434	9 869	53 303
2013	capitale	1 270 000 000	562 000 000	1 832 000 000
	obs.	100 342	27 108	127 450
2014	capitale	2 040 000 000	966000000	3 006 000 000
	obs.	162 256	48 406	210 662

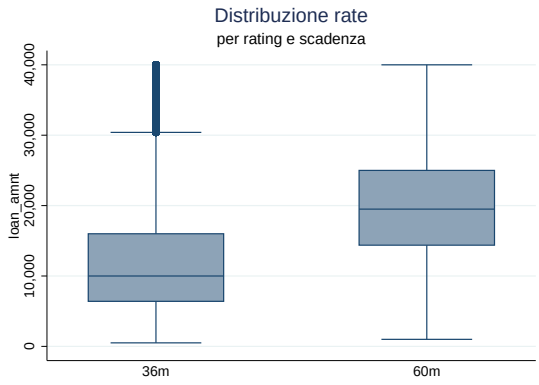


Figura 1.8: Ammontare richiesto per scadenza

Dal lancio della piattaforma online, nel 2007, al 2009 compreso, non sono stati erogati prestiti con scadenza a cinque anni, introdotti poi nel 2010 rappresentando circa il 32% del capitale erogato. Negli anni l'ammontare dei prestiti a cinque anni richiesti si è stabilizzato a circa il 50% del capitale erogato, anche se in termini di numerosità di prestiti, le erogazioni sono nettamente inferiori alla scadenza dei prestiti a tre anni. Evidentemente i prestiti a 60 mesi presentano un capitale maggiore, figura 1.8.

Tutti gli anni, comunque, presentano una composizione di portafoglio sbilanciata sulle classi di rating centrali. Per quanto riguarda i tassi di interesse, non si nota una sostanziale varianza tra gli anni, si attestano sempre in un range tra il 7% circa in classe A, ad un tasso del 25% circa nella classe di rating G. La figura 1.14 a pagina 23 riporta l'analisi della distribuzione del capitale erogato e del tasso di interesse applicato.

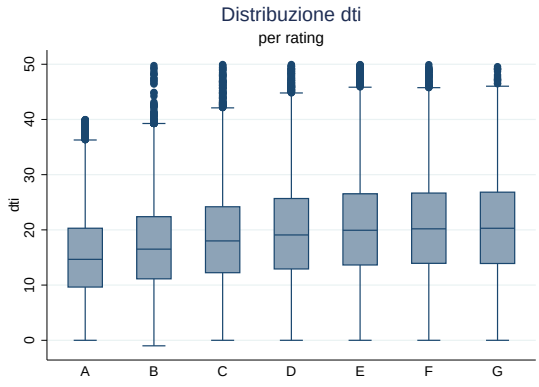


Figura 1.9: DTI secondo rating

DTI Il DTI (debt-to-income ratio) è un rapporto tra le uscite totali mensili e le entrate totale mensili. Rappresenta quanta parte del reddito è erosa dall'indebitamento e permette di quantificare il rischio di una controparte. È evidente come all'aumentare di questo indice, aumenti la rischiosità della controparte peggiorandone il merito di credito rappresentato dalla classe di rating. Una analisi più approfondita di questa variabile verrà affrontata 2.3.2 a pagina 36, in quanto tale variabile entrerà nei predittori della regressione logistica proprio grazie alla sua importanza logica.

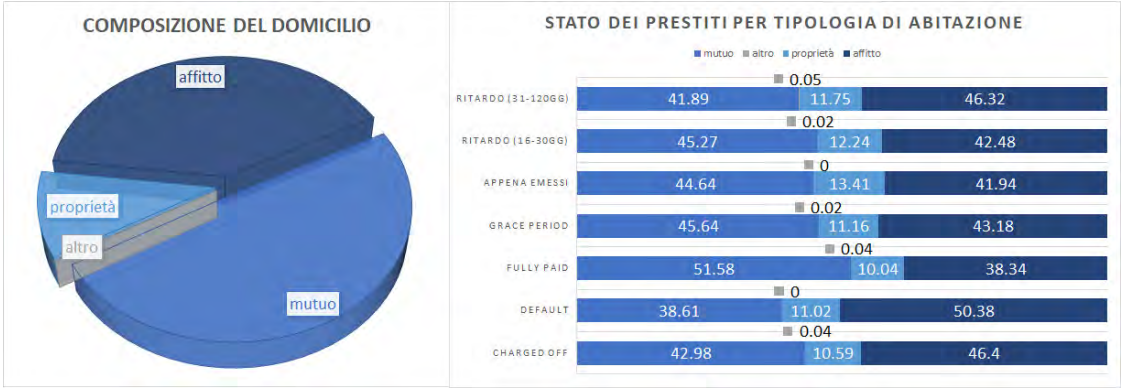


Figura 1.10: Analisi della proprietà immobiliare

Home ownership Un'altra variabile molto importante fornita da Lending Club, è lo stato della proprietà dell'immobile di residenza. La metà delle controparti ha un mutuo da ripagare, un altro 40% circa vive in affitto, mentre il restante 10% è proprietario della propria casa. Questa suddivisione rispecchia controparti con rischi diversi. Si nota nella figura 1.10 che le controparti assoggettate a mutuo sono meno rischiose delle altre, compresi i debitori proprietari del loro immobile. I prestiti completamente rimborsati, infatti presentano una percentuale di controparti con mutuo pari a più della metà delle controparti in affitto dell'immobile. La proporzione sale a più del 60% se si considerano anche le parti proprietarie rispetto agli affittuari. Come diretta conseguenza i prestiti defaultati e in fortissimo ritardo sono principalmente di controparti in affitto.

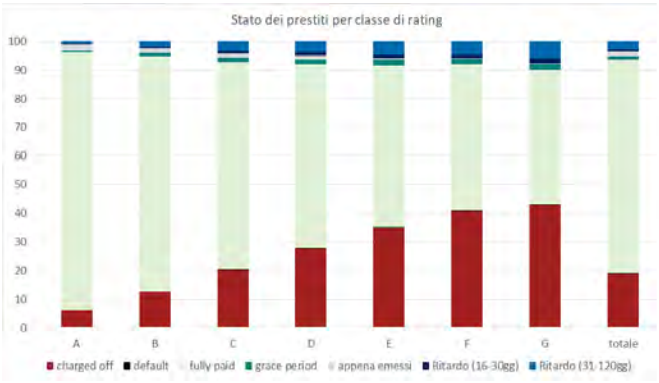


Figura 1.12: Stato dei prestiti secondo classe di rating

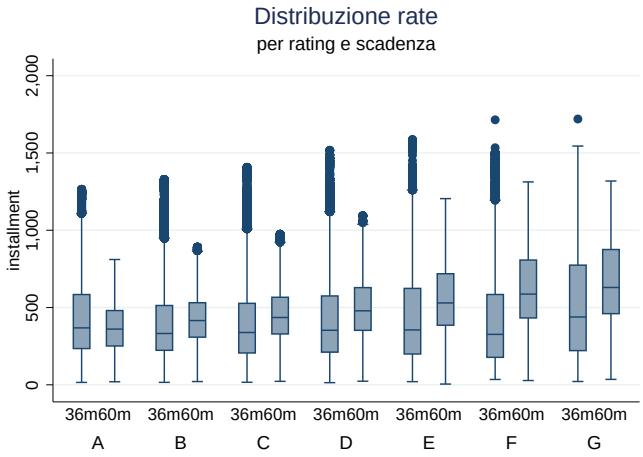


Figura 1.11: Distribuzione dell'ammontare della rata secondo classe di rating

Rata Abbiamo già parlato di ammontare del prestito e di rating della controparte. Entrambi questi concetti influenzano la rata del prestito a carico del debitore. Sebbene non sia una variabile consona a fare da predittore predittore, in quanto è una variabile risultato di una serie di fattori e quindi dipendente, è interessante notare nel grafico 1.11 come sia comunque un indicatore sommario del rischio implicito in un singolo prestito, se confrontato con le variabili che compongono il calcolo della rata. Le classi di rating migliori hanno generalmente una rata inferiore a parità di altre condizioni, questo perché il tasso di interesse applicato è inferiore. Viceversa, le classi di rating peggiori, presentano delle rate generalmente più elevate, rappresentate da un range più ampio. Questo è evidente soprattutto nel confronto degli outliers, che nelle controparti molto rischiose arrivano a rate mensili sopra i 1 500$. Effettivamente andando ad analizzare la frequenza dei default all'interno della classe di rating (figura 1.12), man mano che aumenta la classe di rating di appartenenza, maggiore è la proporzione sia dei default, sia dei prestiti con pagamenti in ritardo, come ci si aspettava.

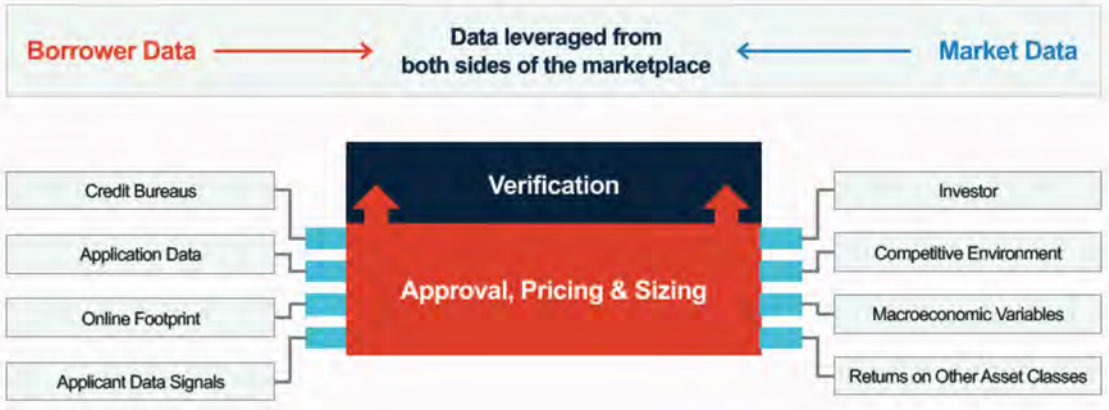


Figura 1.13: Variabili concorrenti all'assegnazione di un rating LC

Osservazioni sui dati Nella figura 1.13[7] viene riportato il modo in cui Lending Club assegna il rating alle sue controparti. È evidente che l'azienda utilizzi un sacco di dati non accessibili al pubblico. Questo potrebbe sfociare nell'implementazione di un modello non del tutto preciso. Inoltre, il dataset reso pubblico da Lending Club, viene costantemente aggiornato, non solo nei valori rappresentanti la situazione dello stato dei prestiti, ma anche nelle variabili pubblicate, rendendo il dataset disomogeneo nel tempo.

Ad ogni modo, a partire dal dataset reso pubblico, si cercherà di ricostruire uno score attraverso una regressione logistica e si calcoleranno anche i *recovery rate*. Ottenute le stime di queste componenti essenziali per l'analisi del rischio di credito si inseriranno i prestiti in un portafoglio al fine di analizzare la distribuzione di perdita, implementato secondo le logiche di CrediRisk+.

[7]Immagine tratta dall'*Investor Day Deck* (Dicembre 2017), fornito dall'azienda su richiesta.

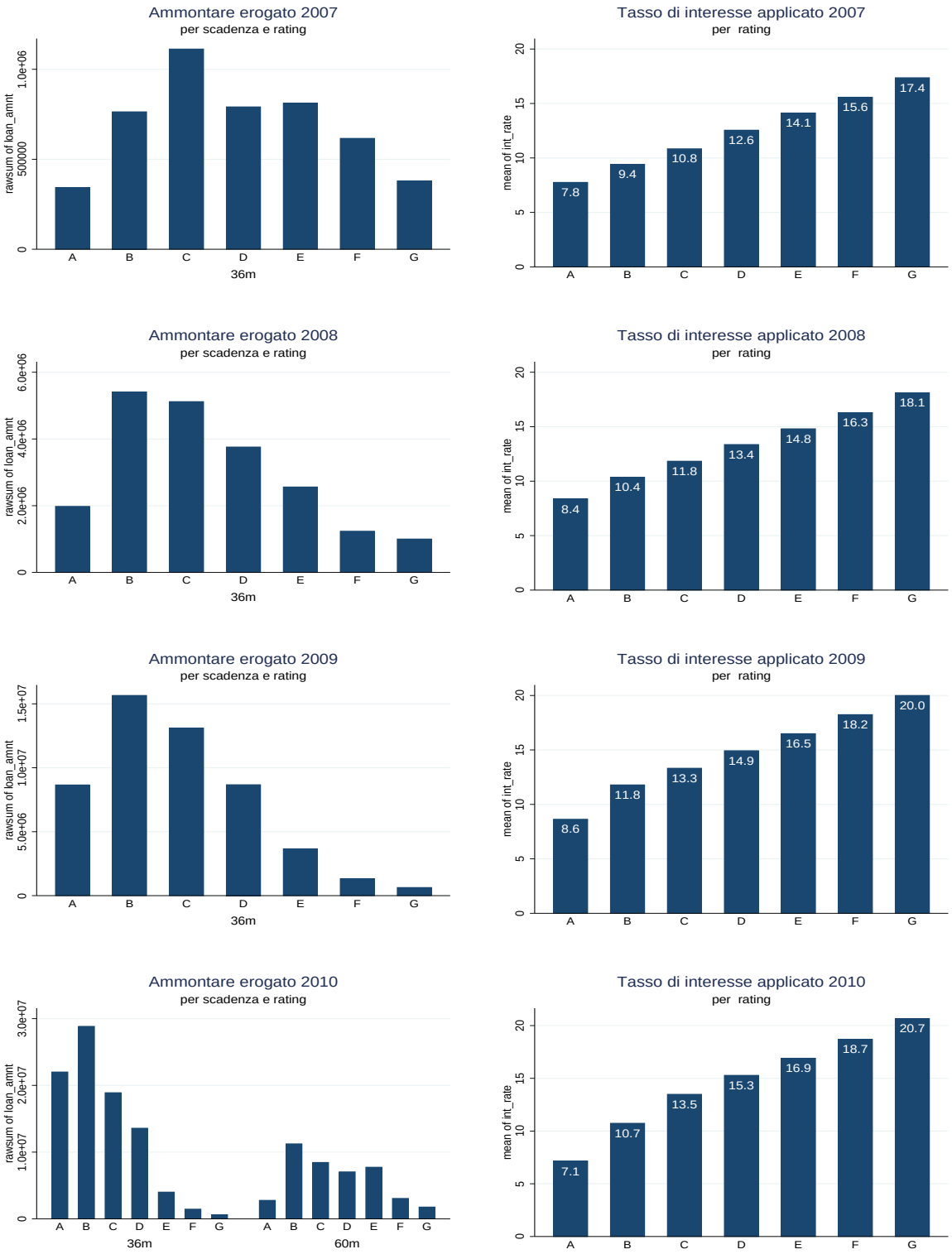


Figura 1.14: Ammontare erogato e tasso applicato per classe di rating

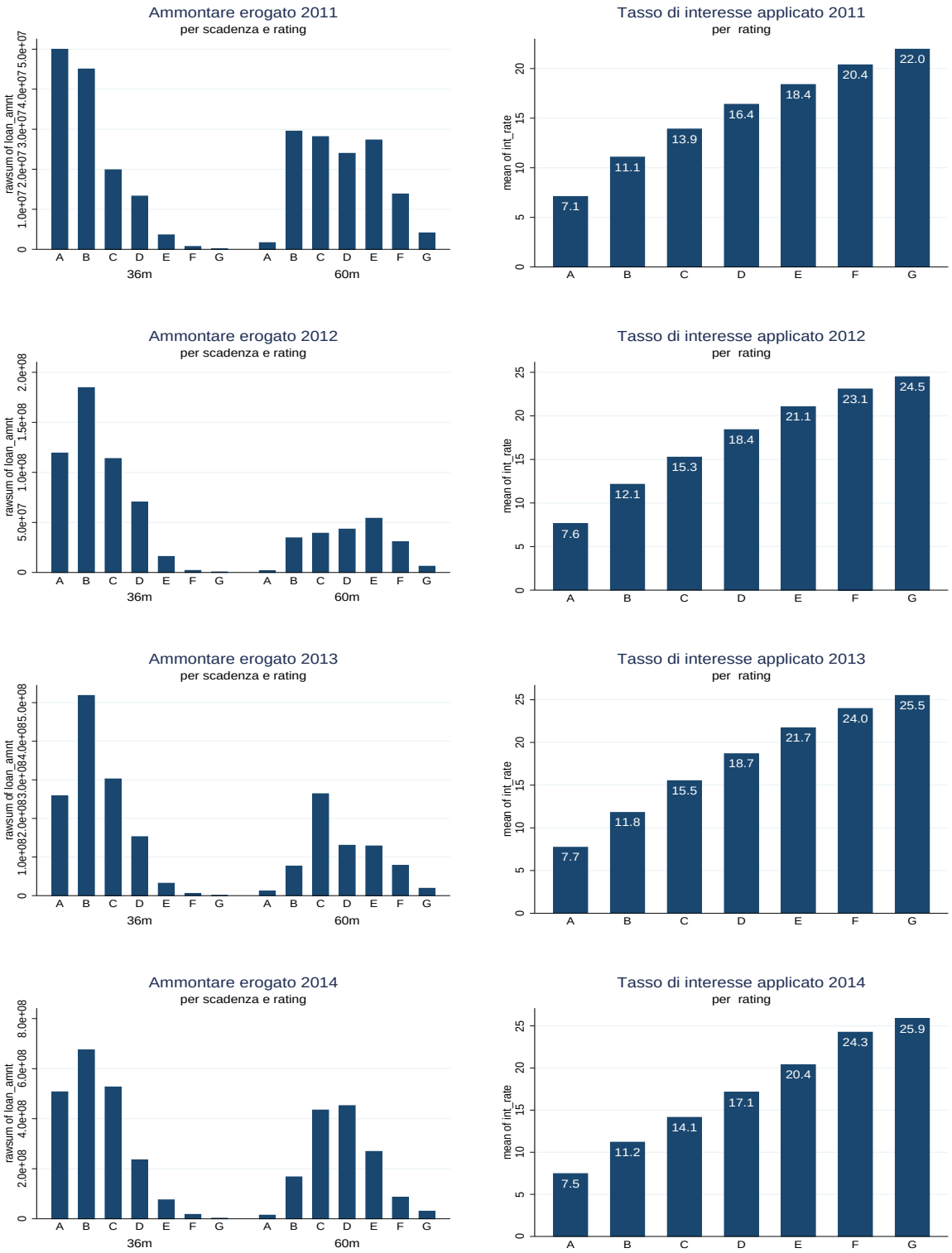


Figura 1.14: [
]Ammontare erogato e tasso applicato per classe di rating

Capitolo 2

Credit Scoring

2.1 Introduzione

Cosa significa *credit scoring*? Per rispondere a questa domanda è utile definire cosa sia il credito e a cosa serva lo scoring.

Cos'è il credito? Semplicemente un modo di definire "compra adesso, paga dopo". *Credito* deriva dalla parola latina *credo* che significa "fidarsi di". Quando si presta qualcosa a qualcuno si ha fiducia nel fatto che questa persona onori l'impegno preso; ma il processo del prestito è in realtà molto di più. I richiedenti devono dare un senso di fiducia per accedere al prestito, ma soprattutto devono ripagare il capitale ricevuto secondo i termini stabiliti, nonché un *risk premium* per la possibilità di insolvenza. Ecco perché si parla di **"merito di credito"** e annesso rischio di credito. Al giorno d'oggi si ha la possibilità di analizzare la fiducia che si merita un debitore e di misurare, al fine di mitigarlo, il rischio associato al suo prestito.
Tutti questi concetti sono particolarmente rilevanti per le banche o istituti di credito, ed effettuano grossi investimenti nel sistema informativo al fine di ottenere un vantaggio competitivo sotto questo profilo.

Cosa significa fare scoring? Aggregare tutti i dati disponibili in un valore che indica quanta qualità ha una determinata controparte.
Per creare uno score si utilizzano strumenti matematico-statistici per ordinare gli elementi della situazione in analisi, sulla base di caratteristiche oggettive e qualità percepita al fine di ottenere una decisione discriminante senza lasciare spazio alla soggettività. I modelli di score sono utilizzati allo scopo di valutare la probabilità relativa di un evento futuro basato sull'esperienza passata.

Cos'è il credit scoring? A questo punto risulta chiaro che il credit scoring è l'utilizzo di modelli statistici per trasformare dati rilevanti in misure numeriche per sostenere le decisioni di credito, *"l'industrializzazione della fiducia"*. Per la prima volta il credit scoring fu utilizzato in America negli anni '60 e da allora si è sempre più evoluto ed affermato, abbandonando progressivamente la visione dicotomica "buoni/cattivi" in cui si affidavano solo i presunti buoni e si evitavano i presunti cattivi, portando a considerare anche l'affidamento di chi può trasformarsi in un "non solvente" in quanto, l'obiettivo primario non è più discriminare chi onorerà i propri impegni da chi non lo farà, ma prezzare correttamente il rischio assunto e quindi la probabilità di incorrere in una perdita presunta, valutandone anche la gravità. Obiettivo di questo capitolo sarà, quindi, quello di analizzare il dataset di P2P lending al fine di creare un sistema di scoring adeguato ad analizzarne il rischio implicito toccando gli elementi definiti dagli Accordi di Basilea, dopo aver definito brevemente come nella storia si è arrivati ai modelli attuali e come siano talmente presenti ed importanti da essere addirittura richiesti dalle autorità di vigilanza europee ma anche statunitensi.

2.2 Breve storia del credito e metodologia del Credit Scoring

L'uso del credito è parte dell'attività umana non solo da quando l'uomo iniziò a commerciare beni e servizi, ma da molto tempo prima, tanto che un primo documento di credito molto grezzo risale agli anni 2000 a.c. della Babilonia e furono i romani i primi a regolamentare l'insolvenza,prima con *lex poetelia*, abolendo la pena di morte o la schiavitù per l'insolvenza, poi con *lex julia* introdusse i moderni concetti di separazione tra persona giuridica e proprietà, insolvenza in buona o mala fede e la equa ripartizione delle perdite tra creditori. Il credito è stato qualcosa di fondamentale in moltissime economie si pensi ai primi commerci europei del 1100 d.c., all'Italia fiorentina che a tal fine inventò la prima cambiale. Raramente,però, è stata un'attività gloriosa in passato, infatti molte religioni additarono il servizio del credito a causa degli interessi applicati e delle penalità per i mancati pagamenti. Questo fino agli anni del 1500 d.c., quando la chiesa protestante accettò di prestare denaro in cambio di un'interesse. La grossa crescita del credito si ebbe con la rivoluzione industriale, che gli diede una spinta anche nel settore dei beni di consumo,dunque non solo a livello industriale e finanziario. Inizialmente il prestito veniva concesso interamente del negoziante secondo valutazione soggettiva, ma con l'industrializzazione di beni costosi come l'auto le somme richieste aumenta-

rono notevolmente ed entrarono nel mercato anche banche e istituzioni finanziarie con scoperti e prestiti a tasso fisso. Questo fino agli anni '60,quando comparirono le prime carte di credito e revolving. Questi prodotti, caratterizzati da alti tassi di perdita, permisero l'accettazione nella pratica professionale del processo di credit scoring, sebbene fosse stato inventato già negli anni '40. Contestualmente si vide la nascita delle agenzie di rating, case editrici che pubblicavano report finanziari sugli emittenti di obbligazioni per gli investitori americani ed europei. Il primo modello di scoring fu implementato da FI nel 1958, diventandone il pioniere indiscusso. Ciò fu possibile grazie allo sviluppo dell'IT, che prima rendeva impossibile implementare i modelli in modo agevole. Inizialmente si utilizzarono solo l'analisi discriminante (DA) e la regressione lineare (LPM), ma con il tempo le cose si evolsero fino a rendere la regressione logistica la tecnica statistica più utilizzata. A fine anni '70 tutto il settore del prestito si rese conto che la modellizzazione dello score permetteva un aumento di valore non solo nel prodotto in sé ma in tutte le attività della gestione del rischio, sfociando, negli anni '80, nell'adozione di score comportamentali e nell'interesse aggiustato per il rischio che portò alla securitisation dei mutui immobiliari a metà degli anni '90.In questo contesto grazie alla continua evoluzione dell'IT e al valore apportato dal processo di scoring nel settore del credito, Moody's rilasciò RiskCalc nel 2000, primo utilizzo commerciale dell'idea di credit scoring per valutare la situazione finanziaria delle aziende. Da allora questo processo ha continuato ad evolversi e ampliarsi, tanto da trovare promozione all'interno della regolamentazione di Basilea II e seguenti. Nel futuro il credit scoring avrà sempre più seguito, soprattutto alla luce dei più recenti avvenimenti economici e alla presenza del problema dei crediti deteriorarti o, in gergo, NPLs (non performing loans).

Metodologia Le filosofie sottostanti il processo di scoring sono il pragmatismo e l'empirismo, con lo scopo di prevedere il rischio di credito e non fornirne una spiegazione. Per questo motivo il processo di scoring è basato sulle performance passate del cliente con la filosofia che il futuro replicherà il passato. Questo processo fornisce il massimo valore quando viene utilizzato per valutare i clienti nel processo di decisione del merito di credito e assume diversi nomi in base al modo in cui viene utilizzato:

Application score, usato per i nuovi business e creato a partire dalla combinazione di dati da fonti diverse, come cliente, relazioni passate con l'istituto e uffici di credito;

Behavioural score, utilizzato per la gestione degli account in essere e focalizzato sul comportamento del cliente;

Collections score, parte del processo di raccolta che incorpora dati comportamentali, dalla raccolta e dagli uffici di credito;

Customer score, che analizza il comportamento di diversi account ed è utilizzato per gestire il prestito in essere;

Bureau score, creato dagli uffici di credito su predittori di insolvenze e ritardi del dataset che gestiscono.

Appurato che lo scoring del credito è l'utilizzo di algoritmi per ordinare i clienti in base alla loro probabilità di essere solventi o non solventi secondo l'esperienza passata, è d'obbligo analizzare i diversi modi per creare un modello di scoring. Gli algoritmi possono assumere diverse forme, ma generalmente la regressione assume questa formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e \tag{2.1}$$

Le cui componenti sono:

x - variabile indipendente o variabile predittore;

b - coefficiente di regressione, parametro che pondera i predittori;

y - variabile target;

e - termine d'errore, variabile dei residui.

I coefficienti di regressione sono derivati per fornire la migliore relazione tra predittori e la variabile target. I modelli possono essere clusterizzati in un primo momento in: parametrici, che fanno ipotesi sui dati, e non parametrici, che non formano nessuna assunzione.

Tecniche parametriche Generalmente uno *scorecard* viene sviluppato con tecniche parametriche come l'analisi discriminante (DA), modello lineare (LPM) e con la regressione logistica. Il modello lineare è spesso usato come parte dell'analisi discriminante, quindi possono essere trattati come un unica identità. LPM è una tecnica molto veloce e facile da implementare, tanto che per molti anni è stata la scelta principale in questo settore; partendo dalla formula 2.1 si arriva agevolmente a identificare la funzione corrispondente alla LPM:

$$\textbf{Linear probability modeling} \quad y \cong G/(G+B) \tag{2.2}$$

dove G e B sono la frequenza di solventi (Goods) e insolventi (Bads) rispettivamente. Però questo modello presenta un limite, non è appropriato per un output binario. Al contrario la regressione logistica presenta una implementazione più lenta ma ha una maggiore adattabilità nei modelli con output duale. Questa regressione funziona attraverso una derivazione del logaritmo naturale delle probabilità:

$$\textbf{Logit} \quad y \cong e^{G/B} \tag{2.3}$$

Un altro formato usato è il Probit, che assume una distribuzione Gaussiana rispetto alla logistica.

Tecniche non parametriche A causa delle assunzioni sui dati richieste dalla tecniche parametriche, si è cercato di creare dei modelli alternativi, non parametrici. Grazie allo sviluppo della tecnologia sono nate le reti neurali (NNs), gli algoritmi genetici e il *K-nearest neighbours*. Altre tecniche non parametriche minori sono i cd. alberi decisionali e la programmazione lineare. Nonostante l'innovazione di queste tecniche e la mancanza di assunzioni, tali metodologie non hanno ancora grande successo, gli esperti del settore le accusano di mancanza di trasparenza e potenziale overfitting.

Oggigiorno la regressione logistica resta la tecnica più utilizzata per la sua adattabilità all'output binario e per la facilità di conversione dello score risultante in probabilità di default. L'abilità primaria di questo tipo di modelli è la capacità di ordinare secondo il rischio. Dall'ordinamento deriva la capacità di prezzare correttamente i prestiti secondo il rischio assunto e di determinare una perdita attesa (EL), la quale è formata da due parti. Dalla probabilità di default e dalla gravità. La gravita deriva a sua volta dall'esposizione al momento del default (EaD), dalla *loss given default* (LGD) e la maturity del prestito (M).

$$\$EL = PD\% * \$EaD * LGD\% * f(M) \tag{2.4}$$

La PD è legata alle circostanze economiche e di mercato di ogni individuo; l'EaD è il valore monetario in cui il creditore è esposto al momento del default; la LGD è la percentuale dell'esposizione che il creditore si aspetta di perdere al momento del default mentre $f(M)$ è una funzione di aggiustamento per la maturity se il prestito ha vita residua maggiore di un anno. L'aggiustamento per la maturity è inserito al fine di modellizzare il maggior rischio affrontato per maturity lunghe, però molto spesso viene eliminato perché il suo impatto spesso è trascurabile oppure perché gli altri elementi sono calcolati in un orizzonte temporale annuale, altre volte non

è possibile derivarlo. Quindi anche in questa sede non verrà considerato.
Il credit scoring è diventato base nel calcolo della perdita attesa, utilizzata per le decisioni basate sul rischio e per i modelli *value at risk* (VaR), molto importanti per definire una perdita nel peggior scenario possibile o *unexpected loss*. Maggiori sono la perdita attesa e la sua volatilità, maggiore è la perdita inattesa (UL). L'importanza di questo argomento è tale da mettere il VaR alla base dei requisiti di capitale delle banche, dopo essere stato adottato dal contesto regolamentare di Basilea II. Ad ogni modo sono argomenti che varranno trattati più dettagliatamente in seguito, mentre sembra doveroso illustrare il processo per la determinazione della PD, che si seguirà anche ai fini di questo lavoro. Il credit scoring è comunemente associato all'uso di tecniche statistiche nel processo di decisione del credito, che si articola in diverse fasi:

- Preparazione del progetto, momento in cui si definiscono gli obiettivi e si identificano le controparti;
- Preparazione dei dati, che prevede la definizione di solvenza e non solvenza, la finestra temporale dell'analisi e una prima analisi delle caratteristiche;
- Implementazione dello scorecard, ovvero la creazione del modello vero e proprio per ottenere uno score;
- Finalizzazione, in cui si valida e si calibra il modello;
- Automazione e reporting, ovvero l'utilizzo del modello come base per le decisioni aziendali in materia e stesura di tutta la documentazione necessaria.

In questa sede ci si fermerà alla finalizzazione, per ovvi motivi.

2.3 Probabilità di Default

I sistemi di rating sono la pietra angolare per il calcolo del capitale regolamentare nell'approccio di rating interno (IRB), inserito a partire da Basilea II, in quanto sono la base per la determinazione della probabilità di default delle controparti affidate. Il sistema di rating può differire in diversi modi, a seconda del tipo di debitore, il tipo di esposizione, le metodologie proprie del modello implementato (per esempio i *point-in-time* vs. *through-the-cycle*) e la disponibilità dei dati.
Con Basilea II, una banca con sistema IRB necessita di una valutazione quantitativa della probabilità di default per ogni controparte nel portafoglio. Secondo Comitato di Basilea per la vigilanza bancaria (2004, par. 452-457) il default avviene quando la banca considera la controparte incapace di pagare le sue obbligazioni

senza azioni di ricorso e/o quando il debitore è in ritardo di 90 giorni dal pagamento. Spesso il tempo di ritardo varia da banca a banca e da tipologia di prodotto in analisi, in quanto Basilea permette di definire una definizione di probabilità di default ad hoc. Ad ogni modo, i sistemi di rating si possono suddividere in due gruppi: quelli che restituiscono una probabilità di default per controparte e quelli che assegnano una probabilità di default media per gruppi di controparti. A fini di questo lavoro si è scelto di associare una probabilità di default per singola controparte, in quanto questa via permette di incorporare le informazioni correnti del credito e non richiede obbligatoriamente assunzioni sui scenari di stress in quanto questi modelli restituiscono PD che si adeguano al ciclo economico, diminuendole in situazioni di espansione ed aumentandole in recessione, ma sopratutto perché avendo un orizzonte temporale di un anno includono già lo scenario macroeconomico. Tutti i sistemi di rating, comunque, alla fine clusterizzano le probabilità di default calcolate in classi, processo che produce una *master scale*; differisce, appunto, solo il modo con cui questo avviene.

I professionisti, come riportato in *Studies on the Validation of Internal Rating System*, utilizzano termini come *"Point-In-Time (PIT)"* o *"Through-The-Cycle (TTC)"* per descrivere le caratteristiche dinamiche dei sistemi di rating. Per evitare confusione:

- PIT rappresenta sistemi di rating che rilevano i cambiamenti del ciclo economico, si focalizzano, quindi, sulle informazioni correnti della controparte. Questo tipo di rating cambia rapidamente secondo le condizioni del mercato, produce PD che tendono ad aumentare in momenti di crisi e che diminuiscono in momenti favorevoli all'economia;

- TTC sono i sistemi che producono un ordinamento delle controparti che resta stabile nonostante i cambiamenti del ciclo economico, si focalizzano sulle performance della controparte. La PD di una controparte cambia se cambiano le caratteristiche della stessa, ma la distribuzione delle controparti nelle classi di rating non cambierà con la modifica del ciclo economico.

Tra questi due poli esiste un ampia varietà di sistemi ibridi; per la tipologia di dataset che si utilizzerà in questa analisi, si costruirà un sistema di rating con caratteristiche PIT prevalenti, in quanto in ogni anno si noterà una certa volatilità nella PD media per classe di rating calcolata.

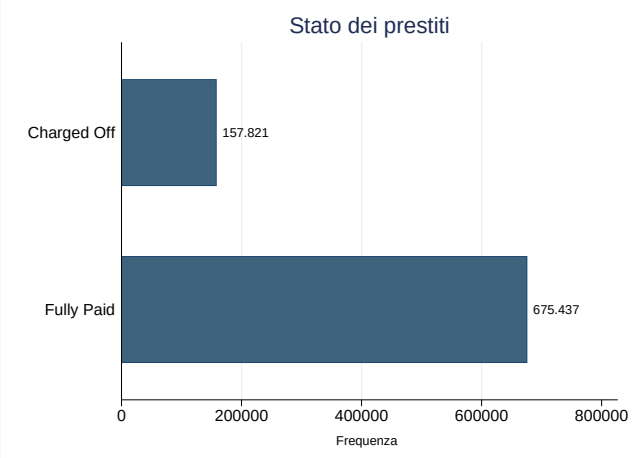


Figura 2.1: Frequenza di solventi e non solventi

2.3.1 Creazione del dataset e pulizia dei dati

La creazione del dataset e la sua pulizia è la fase più lunga e dispendiosa in termini di tempo rispetto a tutto il processo di scoring. Serve per verificare la presenza di dati sufficienti all'implementazione e alla definizione del dataset e della finestra temporale da utilizzare, infatti ogni analisi (inclusi i modelli predittivi) sono dipendenti dalla presenza o meno di dati sufficienti in termini di numerosità di casi e ammontare di informazioni disponibili per ogni caso. Nel tema in esame, secondo quanto suggerito da Raymond Anderson (2007), sono necessari almeno 1500 controparti solventi e 1500 controparti insolventi; le richieste rigettate dalla piattaforma non vengono considerate in questa sede in quanto LC fornisce già un dataset pulito dalle controparti non affidate. Non c'è nessun logica nella scelta di questi numeri, ma si continuano a tenere come target in quanto hanno sempre funzionato nella pratica passata e sono sufficientemente tanti da contrastare l'effetto della multicollinearità e dell'overfitting quando si utilizzano variabili correlate. Ai fini di questa analisi sono stati scaricati i dati in formato .csv dal sito di Lending Club [1] per costruire un dataset di 1 765 426 osservazioni costituito da 102 variabili da analizzare attraverso l'utilizzo del software Stata, in un periodo che va dal 01 Gennaio 2007 al 31 Dicembre 2017. Sono state poi eliminate tutte le variabili con oltre il 45-50% di missing values o concentrate solo in pochi anni, in quanto LC aggiorna e modifica i dati continuamente rendendo il dataset non omogeneo nel tempo. Altri regressori erano completamente privi di valori oppure concentrati su un singolo valore o con valori arbitrari, tutti diversi. Tutte queste variabili sono state eliminate subito per l'incapacità predittiva, insieme alle variabili palesemente non necessarie all'analisi. Il dataset iniziale contiene quindi, circa 35 variabili su cui

[1] lendingclub.com/download-data

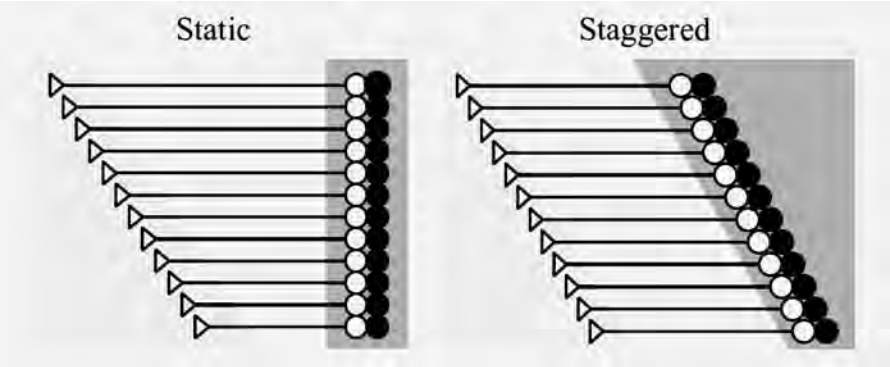


Figura 2.2: Rappresentazione grafica delle finestre di osservazione

lavorare. Per il calcolo dello score i prestiti non devono essere in corso, quindi sono state eliminate tutte le controparti con posizioni aperte, ottenendo 83 258 osservazioni finali, di cui 157 821 prestiti defaultati, rientrando nei criteri di numerosità campionaria per la non multicollinearità (figura 2.1 nella pagina precedente).

A questo punto va determinato cosa si intende per controparti solventi (in seguito anche *goods*) e controparti non solventi (o *bads*), nonché la finestra temporale dell'analisi. Per score comportamentali i dati sono analizzati *point in time*, tipicamente in un periodo di 6 o 12 mesi. Basilea richiede una finestra temporale per l'analisi di 12 mesi, che può essere statica, dove la stessa data viene utilizzata per osservare l'andamento dei prestiti, o sfalsata, che non considera una data di valutazione fissa bensì un periodo all'interno dell'osservazione costante (figura 2.2) [2]). Come la maggior parte degli sviluppi di scoring, si utilizzerà in questa sede una finestra statica di un anno, poiché permette di ottenere il massimo valore dai dati disponibili. Secondo quando riportato sul sito di LC si considereranno defaultati i crediti non pagati da oltre 150 giorni[3]. Tra gli stati LC si legge "Default" e "Charged Off"; questo sta ad indicare che i crediti defaultati hanno mancati pagamenti da 120 giorni mentre i Charged Off sono i crediti defaultati da 30 giorni per i quali non c'è più ragionevole aspettativa di essere pagati, quindi vengono iscritti a perdita[4], mentre tutto il resto sarà considerato "solvenza", in quanto a causa dei dati non troppo precisi è risultato impossibile modellizzare le perdite derivanti dai ritardi. Per quanto riguarda il dataset, si utilizzerà la finestra dal 01/01/2007 al 31/12/2014 come campione *in sample* , mentre si utilizzeranno i tre anni restanti (2015-2017) come campione *out of sample*.

[2] Raymond Anderson 2007, p. 333

[3] help.lendingclub.com, Definizioni degli stati LC

[4] Si noti che i crediti dichiarati in default prima del termine vengono convertiti subito in perdita. help.lendingclub.com, Differenza tra default e charged off

2.3.2 Implementazione

Figura 2.3: Composizione FICO score

Avendo l'obiettivo di ricreare l'ordinamento effettuato da Lending Club, associando una probabilità di default per ogni classe di rating, e sapendo che i prestiti sono di origine americana, si è cercato di simulare il FICO score [5]. Questo sistema di scoring americano viene utilizzato dal 90% delle aziende americane, che restituisce un punteggio con un range da 300 a 850 punti. Il FICO score nasce nel 1989 ed è calcolato partendo da tutti i dati creditizi disponibili nel fascicolo creditizio della controparte, i quali vengono raggruppati in cinque categorie ponderate. La ponderazione riflette l'importanza di tali categorie nel calcolo del punteggio finale:

(i) la storia creditizia della controparte pesa il 35%;

(ii) l'esposizione totale dovuta pesa il 30%;

(iii) un 15% viene rappresentato dalla lunghezza dei rapporti creditizi che il debitore abbia mai contratto;

(iv) la tipologia di crediti che la controparte detiene rappresenta un 10% dello score finale;

(v) l'ultimo 10% è rappresentato dalle caratteristiche del nuovo prestito richiesto.

Non avendo a disposizione il dataset completo delle variabili, in quanto coperti da privacy, si eviterà di pesare le variabili per ottenere il massimo valore predittivo dalla regressione.

Quindi, una volta sviluppato il dataset si è in presenza di un *panel* di variabili e caratteristiche utili per stabilire la relazione tra le caratteristiche di ogni controparte e le sue performance (intese come solvibilità/non solvibilità). Abbiamo già definito l'utilizzo della regressione logistica in questo lavoro, quindi si proseguirà nell'analizzare solo i vari step di questo modello, riassunti in figura 2.4 nella pagina successiva[6].

[5] **\href {https://www.myfico.com/credit-education/credit-report-credit-score-articles/}{www.myfic FICO score}**

[6] Siddiqi 2017, p. 175.

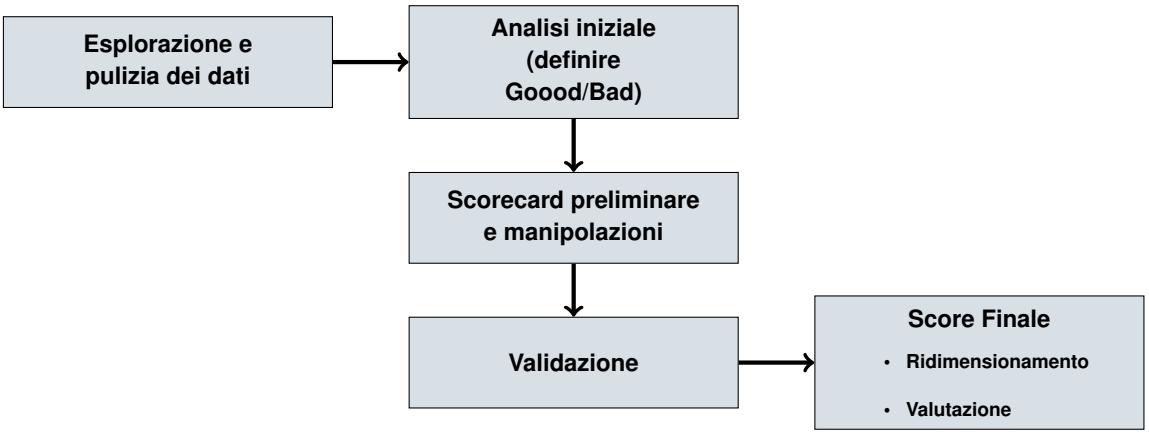


Figura 2.4: Flussi di lavoro per creare un modello di scoring

Pulizia e analisi dei dati È buona norma, nonché primo passo di sviluppo di un modello predittivo, analizzare i dati. Statistiche sommarie come media e mediana, identificazione di valori estremi (chiamati *outliers*) e i percentili dei valori assunti dalle variabili possono offrire un idea più precisa della situazione in esame, oltre a fungere come controllo per l'integrità dei dati.

Missing Values Molto rilevante, in questa analisi, è la presenza di missing values, in quanto la regressione logistica necessità di un dataset completo. I motivi per cui ci possono essere missing values sono tantissimi (variazioni nelle variabili registrate nel dataset originario, dati non disponibili, dati non inseriti dal debitore o semplicemente errori), ma ci sono pochi metodi di trattarli:

1. Escludere tutti i dati con valori mancanti, con il rischio di lavorare con troppi pochi dati;

2. Escludere le caratteristiche e le controparti con missing significativi, ad esempio se mancano più del 40% dei valori, specialmente se si prevede una persistenza futura nella proporzione di dati mancanti;

3. Trattare i valori mancanti come una categoria a parte, inserendo questo gruppo lo stesso nella regressione comportando la necessità di associarci un peso, non sempre facile da determinare;

4. Oppure si possono computare i missing values con delle tecniche statistiche.

Sebbene la via più facile sia l'esclusione di tutte le caratteristiche o delle controparti con *missing* significativi, questo non sempre può essere attuabile in quanto potrebbe ridurre a tal punto il campione osservato da rendere le statistiche poco robuste o distorte, in quanto i *missing values* possono nascondere valore predittivo. Si deve

cercare, quindi, di mantenere più valori mancanti possibili, magari trattandoli come una categoria a parte quando possibile, o di convertirli statisticamente se non si prevede una distorsione elevata nell'output finale. Sta allo sviluppatore scegliere il metodo che reputa più consono ai fini dell'analisi in atto. Le variabili inutilizzabili a causa dei troppi valori mancanti sono già state eliminate in sede di costruzione del dataset, ma va considerato anche il caso in cui un singolo debitore abbia variabili importanti vuote o che molte delle sue variabili non assumano valore. In questi ultimi casi si è deciso di mantenere i *missig* o di convertirli statisticamente.

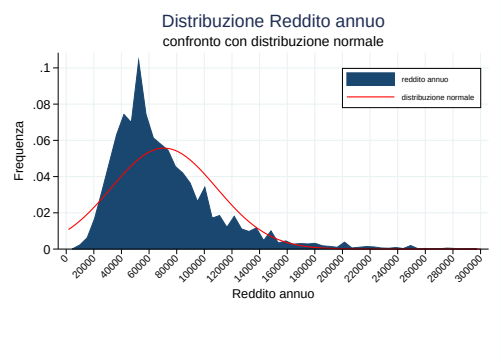
Outliers Un altro aspetto da analizzare con molta attenzione è quello dei valori estremi e l'analisi dei percentili per ogni variabile aiuta a capire quali sono i valori "fuori scala". Questi estremi, sebbene a volte possano essere veritieri, possono portare ad una distorsione dei risultati della regressione e quindi devono essere generalmente esclusi. In alcuni casi, quando i risultati finali non vengono distorti e ci sono poche osservazioni fuori scala, gli *outliers* possono essere sostituiti con delle misure statistiche come la media, per esempio. Ad ogni modo andrebbero sempre analizzati per ogni singola variabile per valutare come meglio procedere nella situazione in esame.

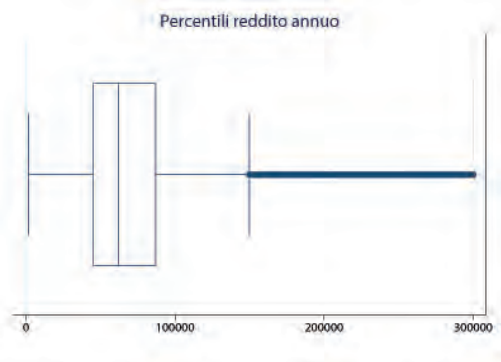
Si è quindi deciso, in questa sede, di eliminare tutte le controparti con reddito annuo assente o fuori scala, mantenendo le osservazioni all'interno del 99-esimo percentile, poiché il reddito annuo è considerata una variabile molto importante che incide notevolmente sulla capacità di ripagare i debiti contratti. Questo si traduce nell'eliminazione di tutte le controparti con reddito superiore ai 500mila dollari. A questo punto si nota dai grafici in figura 2.5 nella pagina seguente che la distribuzione del reddito è asimmetrica a destra, con una mediana di circa 60mila dollari. Il grosso delle osservazioni si concentra tra i 40mila dollari e i 90mila, che corrispondono circa al 25esimo e 75esimo percentile, come sottolineato dal boxplot. Un altra variabile molto importante è l'indice *DTI*, che per natura deve essere compreso esclusivamente tra 0 e 100%. Il *DTI* o *Debt-to-income ratio* [7] rappresenta un semplice calcolo per analizzare velocemente la capacità di ripagare il debito, calcolato come:

$$dti = \frac{Totale\ dei\ debiti\ da\ ripagare\ mensilmente}{Totale\ entrate\ mensili\ lorde} * 100 \tag{2.5}$$

Sono esclusi dal totale mensile dei debiti i pagamenti per *utilities*, il finanziamento per la macchina o prestiti studenteschi, così come le bollette telefoniche. Il *dti* è

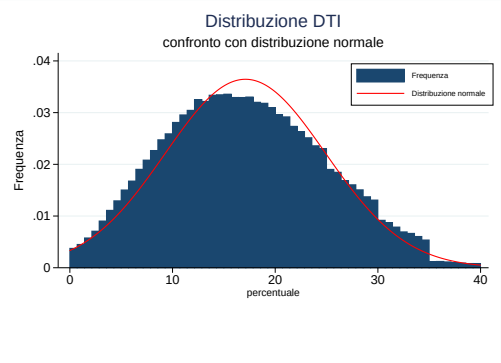
[7] www.lendingclub.com, Calcolo del dti

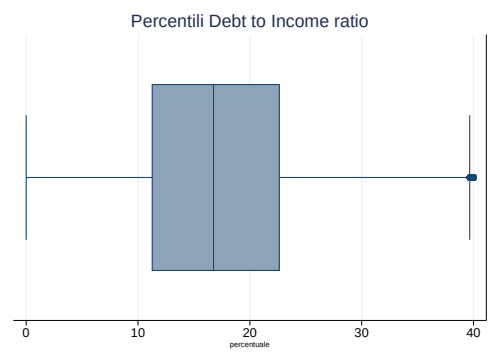




Statistiche Reddito annuo			
Min	1896	25esimo	45000
Max	500000	50esimo	62000
Media	72028.22	75esimo	88000
SD	42796.39	90esimo	120000
Asim.	2.635	95esimo	150000
Curtosi	16.133	99esimo	235000

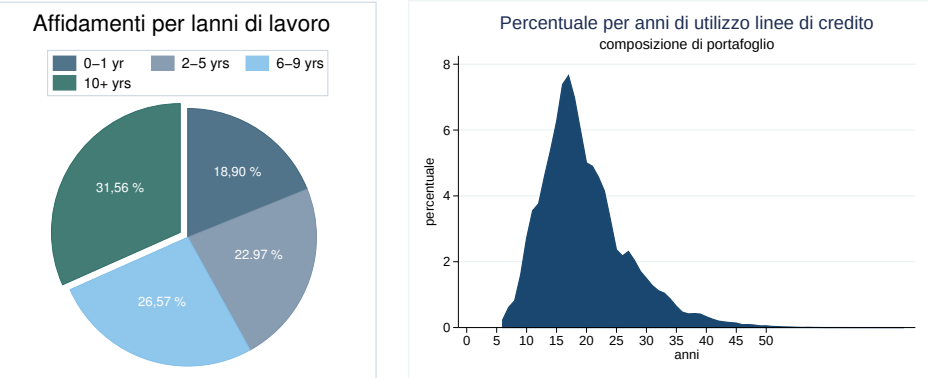
Figura 2.5: Analisi della variabile Reddito annuo





Statistiche DTI			
Min	0 %	25esimo	11.22%
Max	40%	50esimo	16.71%
Media	17.067%	75esimo	22.6%
SD	7.824	90esimo	27.82%
Asimmetria	0.197	95esimo	30.58%
Curtosi	2.462	99esimo	34.57%

Figura 2.6: Analisi della variabile DTI



(a) *Anni di lavoro* **(b)** *Anni di accesso al credito*

Figura 2.7: Distribuzioni temporali

molto importante perché se risulta molto alto la controparte potrebbe non essere in grado di ripagare il prestito e quindi il finanziatore perderebbe i suoi soldi, generalmente si considera buono un DTI inferiore al 40%. L'analisi delle statistiche sommarie conferma questa regola in quanto nemmeno al 99esimo percentile questo indice arriva al 40%, ma si attesta quasi sei punti percentuali sotto con una media del 17%. La maggior parte delle osservazioni indica un DTI compreso tra l'11% e il 23%.

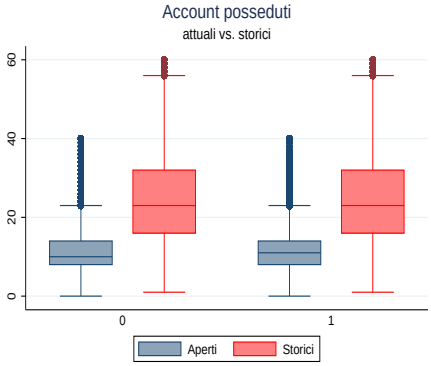
Delle 35 variabili mantenute nel dataset, non tutte saranno utilizzate al fine della regressione ma serviranno successivamente per l'analisi delle altre componenti del rischio di credito. Oltre a queste due, si utilizzeranno inizialmente altre 14 variabili, che verranno poi analizzate nella loro capacità predittiva. Queste variabili coprono tutti i settori definiti a inizio sezione mentre si parlava del FICO score. Per quanto riguarda la lunghezza della storia di accesso al credito sono presenti l'anno di accesso al primo prestito (yr_earliest_cr_line) e la lunghezza della vita lavorativa (emp_length). È evidente che l'azienda preferisca affidare controparti con una storia lavorativa lunga perché spesso è sinonimo di maggiore stabilità finanziaria, ma la composizione del portafoglio (figura 2.7a) non è così sbilanciata, anzi, anche controparti molto giovani nel mercato del lavoro sono affidate in buona misura. La stessa logica si nota nella variabile che conta il numero di anni di "esperienza" nel settore creditizio. Come si può notare in figura 2.7b distribuzione ha corpo tra i cinque e i vent'anni di esperienza, forse dovuto al fatto che istituzioni creditizie sono la prima scelta per persone che si fanno affidare per la prima volta e/o perché controparti con maggior storia creditizia garantiscono una migliore capacità di prezzare il loro rischio. Molto importante al fine dell'affidamento è l'analisi del numero di linee di credito aperte dalla controparte richiedente (open_acc), in quanto maggiore sono le linee aperte, maggiore è la probabilità

di risultare insolventi (acc_now_delinq) come dimostrato dalla linea di regressione in figura 2.8b a pagina 40. Importante è anche la percentuale di utilizzo di tali linee aperte in cui non si è verificato lo stato di insolvenza (revol_util). La variabile che rappresenta il numero di linee in cui la controparte è insolvente è rappresentato da una variabile categoriale che assume valore 0 quando il debitore è assolutamente solvente, 1 quando è insolvente in una linea di credito, oppure valore 2 se l'insolvenza ha raggiunto almeno due affidamenti. Accanto alle linee attualmente attive si sono analizzate le linee aperte durante tutta la storia creditizia (total_acc) ma come si può notare in figura 2.8a nella pagina seguente la variabile non sembra essere significativa per determinare la differenza tra una controparte solvente e una no, quindi si procede ad eliminarla e non inserirla nella regressione. Per quanto riguarda le percentuali di utilizzo, si osserva una relazione a prima vista inversa rispetto al numero di account aperti. Verrebbe quasi spontaneo pensare che all'aumentare di linee aperte la percentuale di utilizzo delle stesse aumenti, in quanto una controparte potrebbe aprire un nuovo account perché ha esaurito la disponibilità in quelli già esistenti, invece l'apertura di nuovi account non sembra corrispondere a questa teoria in quanto all'aumentare del numero di account posseduti l'utilizzo diminuisce. Questo è logico in quanto la percentuale è influenzata dalla misura di scala dell'ammontare affidato e, all'aumentare di linee aperte, si guadagna "un'economia di scala". Ad ogni modo la media utilizzata delle linee di credito possedute corrisponde circa al 50% dell'ammontare complessivo.

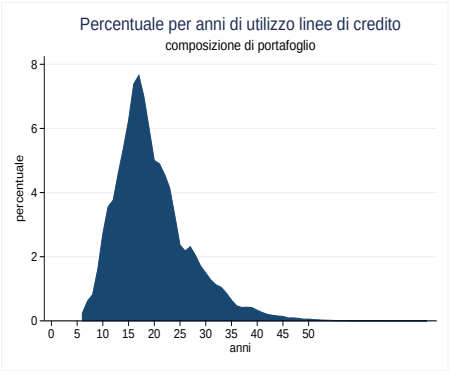
Parlando del FICO score nella parte 2.3.2 a pagina 33, si è sottolineata la presenza di variabili che rappresentano la storia creditizia. A questo proposito si sono inserite quattro variabili:

- Indagini nell'ultimo anno (inq_last_12m)
- I ritardi di oltre trenta giorni registrati del file creditizio della controparte negli ultimi due anni (delinq_2yrs), per analizzare la storia creditizia della controparte in un più lungo periodo
- Numero di deroghe di dominio pubblico effettuate (pub_rec)
- Numero di fallimenti di dominio pubblico (pub_rec_bankruptcies)

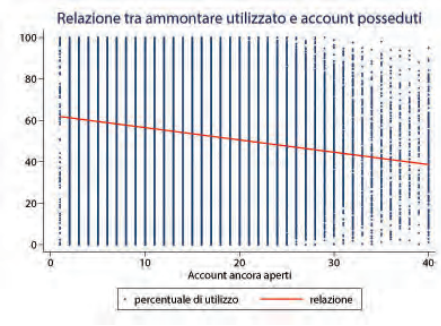
Osservando la figura 2.9 a pagina 41 si nota che queste variabili sono state "categorizzate", nel senso che sono stati mantenuti i valori, discretizzati, entro l'intervallo del 99% mentre tutti i valori oltre questo limite sono stati sostituiti dal valore massimo consentito. Si può notare come queste variabili presentino dei *missing values* che si è deciso di mantenere in quanto non sono in misura tale da inficiare



(a) Nr linee per stato del prestito



(b) Relazione tra nr. di account e insolvenza

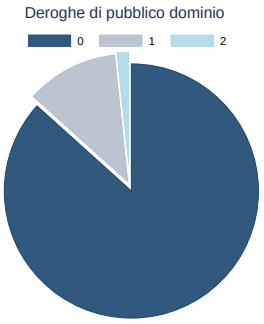


(c) Relazione tra nr.account e utilizzo

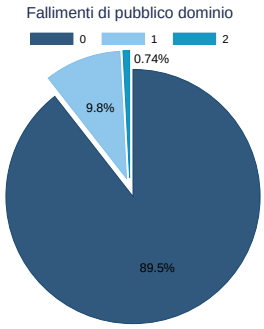
Statistiche revol_util			
Min	0 %	25esimo	38.7%
Max	40%	50esimo	57.1%
Media	55.77%	75esimo	74.4%
SD	7.824	90esimo	86.9%
Asimmetria	-0.2531403	95esimo	92.3%
Curtosi	2.311596	99esimo	97.9%

(d) Statistiche sull'utilizzo

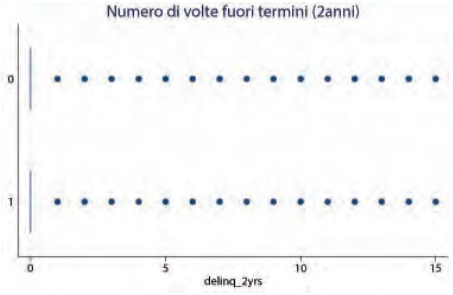
Figura 2.8: Analisi della quantità di linee di credito



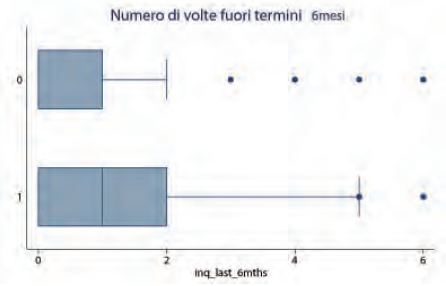
(a) Composizione della variabile pub_rec



(b) Composizione della variabile pub_rec_bankruptcies



(c) Volte fuori termine per stato del prestito



(d) Numero di inchieste per stato del prestito

Figura 2.9: Analisi variabili per la storia creditizia

l'ordinamento della regressione. Le variabili che riguardano i dati di pubblico dominio sono state suddivise in tre categorie, "0" identifica nessuna deroga ai termini del prestito che sia di dominio pubblico, "1" identifica una deroga mentre "2" ne identifica due o più. Per quanto riguarda le azioni legali si sono considerati due boxplot per vedere la relazione con lo stato del prestito, identificando con zero lo stato di solvenza e con uno lo stato di default. Nel grafico delle inchieste negli ultimi sei mesi si nota come le controparti insolventi abbiano un più alto numero di azioni legali a carico rispetto alle controparti solventi avendo una media di un'azione (figura 2.9d), netta è la distinzione se si prende in considerazione il valore dal 25esimo al 75esimo percentile, dove gli insolventi hanno un valore compreso tra zero e due, mentre per i solventi si limita solo fino alla prima classe. Più difficile è l'analisi delle azioni legali negli ultimi due anni, in quanto le controparti portano ad una concentrazione dei valori su zero. Questo però al momento non significa che la variabile non sia predittiva in quanto sono comunque presenti molte osservazioni fuori dallo zero, andrà testato in sede di regressione il valore informativo del regressore. Insieme alle variabili appena analizzate si utilizzeranno le variabili loan_amnt addr_state, home_ownership, purpose e term analizzare nel capitolo 1.2.

Analisi iniziale Completata la pulizia dei dati si dovrebbe analizzare la significatività predittiva di ogni singola variabile e la sua relazione con la situazione di default/non default. Questa analisi, definita selezione univariata, serve ad evidenziare relazioni illogiche e a selezionare le variabili con predittività maggiore, i cui valori verranno poi raggruppati *"processo di binning"*. Questo modo di procedere offre dei vantaggi:

- Nel caso vengano mantenuti outliers, i loro effetti vengono ridotti attraverso il raggruppamento;
- Rendono più facile comprendere la relazione esistente con l'output, aiutando l'analista a capire in quali valori dei dati con cui cambia il comportamento della controparte;
- Permette di modellizzare in maniera lineare anche relazioni non lineari;
- Permette di modellare lo score finale modellando il binning, molto utile quando si lavora con dataset piccoli o parziali;
- In questo modo diventa evidente il comportamento della variabile predittiva anche a fini strategici.

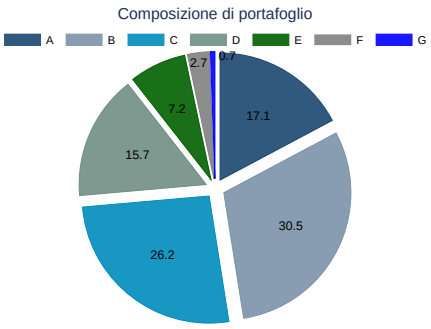


Figura 2.10: Composizione

Essendo richiesta la monotonicità della curva finale dello score, si utilizzerà per il binning un *"algoritmo monotono per gruppi adiacenti"* o MAPA, descritto in Raymond Anderson (2007, p. 371) come *classificatore grossolano monotono di massima verosimiglianza - Maximun Likelihood Coarse Classifier* in quanto permette di individuare un WOE monotono. Una seconda argomentazione in favore di questa scelta è la presenza di un portafoglio iniziale non omogeneo nella rischiosità delle controparti come mostrato nella figura 2.10, dove i gradi B e C coprono circa il 60% degli affidati.

Il punto iniziale è il calcolo della cumulata dei default per ogni livello di score:

$$CumulataBads_{k,v} = \sum_{i=V_{k-1}+1}^{v} B_i \Big/ \sum_{i=V_{k-1}+1}^{v} (B_i + G_i) \tag{2.6}$$

dove Cum è la cumulata dei default; G e B sono la frequenza dei solventi e non solventi rispettivamente; V è il vettore che contiene i punti di cut-off determinati;

v è il livello di score dell'ultimo cut-off determinato; i e k sono l'indicizzazione degli score e dello score al cut-off. I punti di taglio vengono identificati come livello di score con massimo livello di default:

$$\textbf{MAPA} \quad V_k = max\{v|C_{k,v} = max\{C_{k,v}\}\}, \qquad \forall v > V_{k-1} \tag{2.7}$$

Questo processo è ripetuto finché il massimo della cumulata dei default è associata allo score più alto presente. Un esempio permetterà di comprendere il funzionamento dell'algoritmo
[8]:

Esempio 1. Applicazione algoritmo MAPA [9]

Bin	Good	Bad	Cut-Off1	Cut-Off2	Cut-Off3	Cut-Off4
$[-Inf, 33000)$	127	107	0.543			
$[33000, 38000)$	194	90	0.620	0.683		
$[38000, 42000)$	135	78	0.624	0.662		
$[42000, 47000)$	164	66	0.645	0.678	0.713	
$[47000, Inf]$	183	56	0.669	0.700	0.740	0.766

L'algoritmo, dopo aver diviso in gruppi omogenei le osservazioni, determina i punti di taglio sulla proporzione di *Good* osservata. Nella colonna Cut-Off1 il valore 0.543 è la quantità di buoni diviso il totale delle controparti osservate(127/(127+107)), e così via. Il primo taglio viene effettuato dove c'è il minimo della cumulata, ovvero qui 0.620. L'algoritmo procede con queste azioni finchè non esaurisce le osservazioni.

Quanto discusso sulla predittività, invece, si implementa attraverso l'utilizzo di due misure specifiche: l'Information Value - (IV) ed il Weight of Evidence (WOE) restituiti dal binning in MatLab. Ad ogni modo, misurare la forza statistica in termini di WOE e IV non è l'unico fattore per scegliere le caratteristiche utili al modello. Per analizzare la forza predittiva delle variabili in questo tipo di analisi va considerato anche l'ordinamento logico.

Alla nascita dell'idea di credit scoring, Fair Isaac adottò una misura che chiamò Information Value per misurare la forza predittiva di una caratteristica, basata sulla misura di divergenza di Kullback utilizzata per misurare la distanza tra due distribuzioni. Secondo Siddiqi (2017, p.185) si identifica una "regola del pollice"

[8] it.mathworks.com, Autobinning in MatLab
[9] Thomas 2002

per identificare la forza di ogni caratteristica attraverso questo indice: un valore minore di 0.02 è generalmente non predittivo, ma soprattutto sotto lo 0.01 sarebbe meglio eliminarlo; da 0.02 a 0.1 è debole; con un valore compreso tra 0.1 e 0.3 si ha una capacità predittiva media; in presenza di un IV maggiore di 0.3 e minore di 0.5 si hanno variabili con predittività forte; nei casi in cui si superi lo 0.5 si dovrebbe indagare per *overpredicting* e sarebbe meglio togliere queste caratteristiche dal modello.

La formula dell'information value è:

$$IV = \sum_{i=1}^{n} \big(DistrGood_i - DistrBad_i\big) * ln\big(DistrGood_i / DistrBad_i\big) \tag{2.8}$$

$$= \sum_{i=1}^{n} \left[\left(\frac{N_i}{\sum N} - \frac{P_i}{\sum P} \right) * WOE_i \right] \tag{2.9}$$

dove N sono i solventi (negativi allo stato di default), P sono i non solventi (positivi allo stato di default), WOE è il weight of evidence indicizzato alla i-esima caratteristica e n è il numero totale delle variabili. Va sottolineato che l'IV può assumere solo valori positivi e in questa formula le distribuzioni devono essere inserite in formato decimale.

Nella tabella 2.1 calcolata sul dataset si osserva come l'IV è in buona parte delle volte molto basso, ma come sottolineato nei libri già citati, essendo in presenza di una relazione logica molto forte con la variabile risultato, è meglio mantenere le variabili più deboli lo stesso nel modello. Viceversa, ci sono delle variabili che in alcuni anni presentano IV praticamente pari a zero o quasi deterministici che si è proceduto ad eliminare dalla regressione, secondo le indicazioni sopra riportate. Delle variabili rimanenti verrà analizzato il p-value, ove non significativo verrà eliminato dalla regressione, come riportato in tabella 2.2 a pagina 52.

Nel 1950, Irving John (Jack) Good pubblicò un libro che conteneva il modo in cui le persone prendono decisioni sulla base degli eventi che succedono. Le persone decidono di attraversare una strada in base a quanto traffico vedono al momento, oppure decidono se portarsi l'ombrello guardando le previsioni del meteo o il tempo fuori. Per ogni decisione si raccolgono e valutano le circostanze e si determina un peso dell'evidenza - weight of evidence - che converte, semplicemente, il rischio associato ad una scelta da fare in una relazione lineare facilmente comprensibile al cervello. Il WOE è una misura di separazione tra *goods/bads* applicabile ad ogni caratteristica, per l'esattezza misura la differenza tra proporzione di solventi

Tabella 2.1: Information Value per variabile secondo anno

	2007	2008	2009	2010	2011	2012	2013	2014
loan_amnt	0.014	0.069	0.017	0.002	0.030	0.042	0.034	0.036
yr_earliest_cr_line	0.065	0.013	0.011	0.007	0.012	0.005	0.010	0.011
acc_now_delinq	1	0.79	1	1	1	0	0	0
pub_rec_bankruptcies	0.046	0.050	0.003	0.032	0.013	0.002	0	0
pub_rec	0.031	0.033	0.012	0.038	0.015	0.003	0	0
addr_state	0.779	0.062	0.016	0.003	0	0.003	0	0.001
annual_inc	0.035	0.025	0.028	0.055	0.054	0.033	0.044	0.038
revol_util	0.004	0.072	0.056	0.074	0.088	0.042	0.031	0.025
delinq_2yrs	0.003	0	0	0.003	0.007	0.001	0	0.001
dti	0.024	0.050	0.019	0.013	0.021	0.034	0.063	0.067
emp_length	0.036	0.024	0	0.026	0.016	0.002	0.005	0.007
home_ownership	0.027	0.044	0.013	0.007	0.009	0.007	0.014	0.012
term	0.469	0.786	1	0.161	0.218	0.140	0.248	0.274
purpose	0.104	0.053	0.061	0.053	0.057	0.019	0.010	0.005
inq_last_6mths	0.291	0.209	0.095	0.142	0.039	0.036	0.024	0.020
open_acc	0.025	0.013	0.003	0.012	0.005	0.004	0.003	0.006
perc_int_rate	0.465	0.228	0.219	0.370	0.378	0.298	0.437	0.485
obs.	572	2389	5267	7635	21706	53256	127450	210662

e insolventi per ogni variabile. Si calcola come $(DistrGood/DistrBad)$ o in modo da rendere il risultato di più immediata comprensione:

$$\textbf{Weight of evidence} \quad W_i = \Big[ln\big(DistrGood/DistrBad\big)\Big] * 100 \tag{2.10}$$

$$= \ln\left(\left(\frac{N_i}{\Sigma N}\right)/\left(\frac{P_i}{\Sigma P}\right)\right) \tag{2.11}$$

L'esempio in figura 2.11 nella pagina successiva[10] può aiutare a capire il funzionamento dei questo indicatore. Risultati negativi sono possibili e implicano che quel particolare gruppo di quella caratteristica rappresenta una maggiore presenza di *bads* che *goods*. Per quanto riguarda le variabili selezionate per la regressione è bene osservare la loro relazione logica con lo stato di default o meno, attraverso la figura 2.13 a pagina 48. Analizzando la variabile "ammontare del prestito - loan_amnt" il WOE presenta un trend decrescente, così come per "utilizzo linee - revol_util", "numero di account da sempre aperti - open_acc" e "tasso di interesse - perc_int_rate". Questo significa, come riportato anche nella figura 2.11 nella pagina successiva, che ad un aumento dell'utilizzo o dell'ammontare richiesto aumenta il rischio di default. Inutile spiegare che variabili come "account attualmente in default - acc_now_delinq", "fallimenti di dominio pubblico

[10] Raymond Anderson 2007, p. 193.

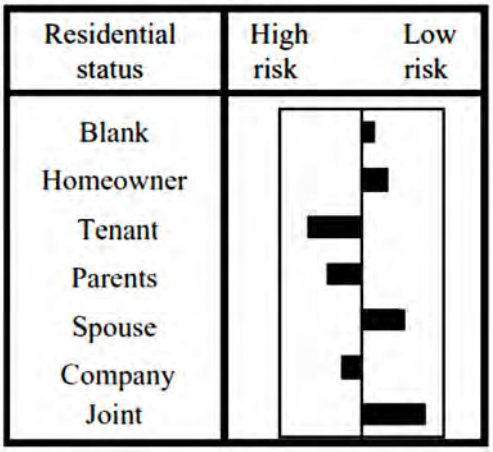


Figura 2.11: Applicazione esempio del WOE

- pub_rec_bankruptcies", "inchieste di dominio pubblico - pub_rec", "insolvenze e ritardi negli ultimi due anni - delinq_2yrs" e "inchieste negli ultimi sei mesi - inq_last_6mths" presentano lo stesso andamento. Questo trend è logico perchè maggiore è la somma richiesta più possono aumentare le difficoltà nel ripagarla e, allo stesso tempo, le variabili che indicano la storia della bontà dei pagamenti dovuti del debitore rafforza il maggior rischio in base al comportamento della controparte. Un altra variabile che presenta un trend logicamente decrescente del WOE è rappresentata dalla "scadenza - term". Maggiore la *maturity* o comunque maggiore è il tempo di recupero del capitale affidato, aumenta l'incertezza (rischio), quindi i prestiti a cinque anni sono per costruzione più incerti e rischiosi dei prestiti a tre anni. Le variabili che rappresentano le entrate presentano un trend inverso. A maggiori entrate infatti corrisponde un WOE crescente. Questo è visibile nella variabile "reddito annuo - annual_inc" e "anni di lavoro - emp_length" perchè a maggior carriera lavorativa, considerata anche una minima propensione al risparmio, dovrebbe corrispondere un maggior patrimonio". Le variabili categoriali sono più difficili da analizzare invece. Per quanto riguarda lo "scopo-purpose" si nota come i prestiti per piccole aziende di persone, consolidamento dei debiti e la categoria "altro", dove sono compresi prestiti per viaggi, vacanze e matrimonio, sono le più rischiose, mentre i prestiti per la casa e la macchina si sono rivelati i più sicuri.

Particolare attenzione va posta sulla variabile "tipologia di domicilio - home_ownership". Sebbene sia logico che chi paga l'affitto sia un debitore più rischioso di chi ha la casa di proprietà, non pare logico il fatto che chi abbia ancora a carico il mutuo sia però meno rischioso di chi abbia già finito di pagarlo o non lo abbia affatto.

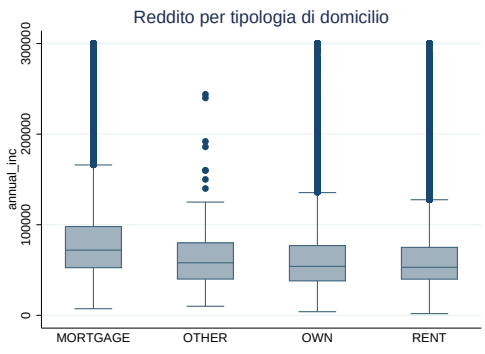


Figura 2.12: Relazione tra reddito e tipologia di abitazione

La spiegazione potrebbe risiedere sul fatto che chi ha ricevuto un mutuo abbia condizioni finanziarie migliori altrimenti non avrebbe avuto accesso al prestito, mentre chi ha pagato la casa potrebbe aver eroso il proprio patrimonio (figura 2.12). Verrebbe da pensare che una casa con mutuo aperto sia più giovane o comunque meglio mantenuta, rispetto ad una casa i cui proprietari non investono in lavori di manutenzione rappresentando una garanzia implicita maggiore . Non si può, invece, fare alcuna considerazione sullo "stato di residenza - addr_state" in quanto il WOE è qui influenzato dalle variabili macroeconomiche nazionali, che si riflettono sul merito di credito del debitore.

Scorecard L'analisi iniziale ha permesso di identificare un set di variabili che vanno considerate nel modello finale, eliminando se necessario quelle che permetteranno di ottenere maggiore valore predittivo senza la loro presenza. Generalmente il modello finale viene creato con una regressione logistica quando il risultato è binario, che conterrà un numero di variabili variabile tra 8 e 15, selezionate considerando la correlazione tra le variabili e la forza statistica del modello [11]. Come già citato, questo tipo di regressione non ammette *missing values*, quindi quelli mantenuti devono essere trattati in qualche modo. Si è deciso di considerarli come zeri, in modo da assegnare un WOE nullo a quella determinata variabile mancante per un precisa controparte.

Regressione logistica Come riportato in Raymond Anderson (2007, p. 171), questo tipo di regressione ha radici nello studio della crescita della popolazione. Nel 1798 Malthus affermò che la popolazione umana cresceva in progressione geometrica, tesi valida, secondo l'epoca, vista la crescita massiva della popolazione di molti stati europei. Ma quarant' anni dopo un astronomo belga, Alphonse Quetelet, realizzò che tale crescita non poteva continuare indefinitamente e incaricò il suo pupillo Pierre F.Verhulst di lavorare al tema, il quale definì una curva a forma di S inserendo un limite massimo alla crescita geometrica, fornendo la formula

[11] Per l'implementazione del modello e la declinazione di tutte le sue caratteristiche sono state utilizzate le funzioni *built-in* fornite da MatLab e le loro spiegazioni. it.mathworks.com/help/finance, Creditscorecard analysis

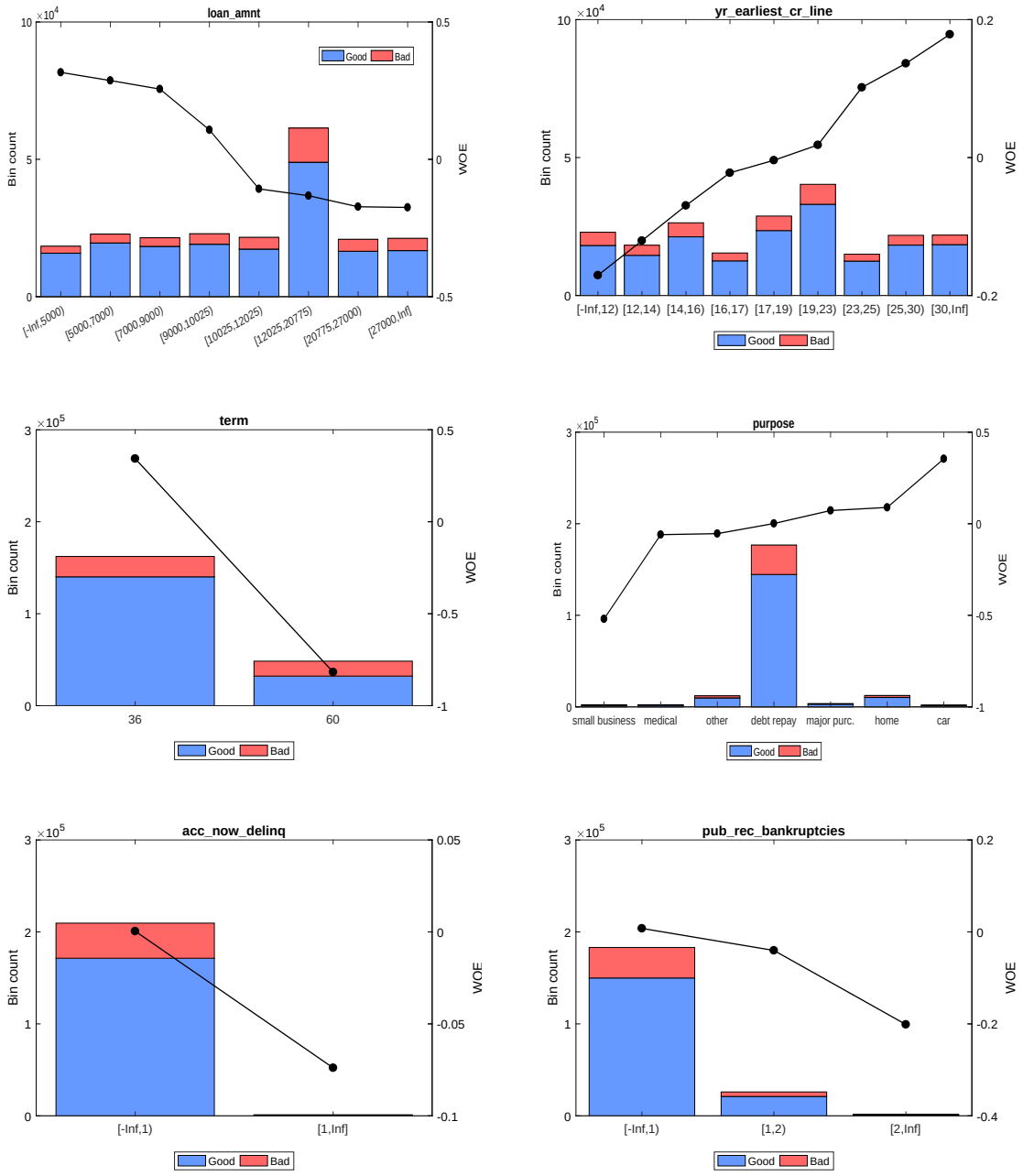


Figura 2.13: Analisi andamento del WOE

$P(Z) = exp(Z)/(1 + exp(Z))$ chiamandola curva logistica.

La regressione logistica usa, come tutti i metodi per creare modelli statistici, un set di predittori per modellizzare la probabilità di un evento (target), secondo l'equazione:

$$Logit\ (p_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e \tag{2.12}$$

dove:

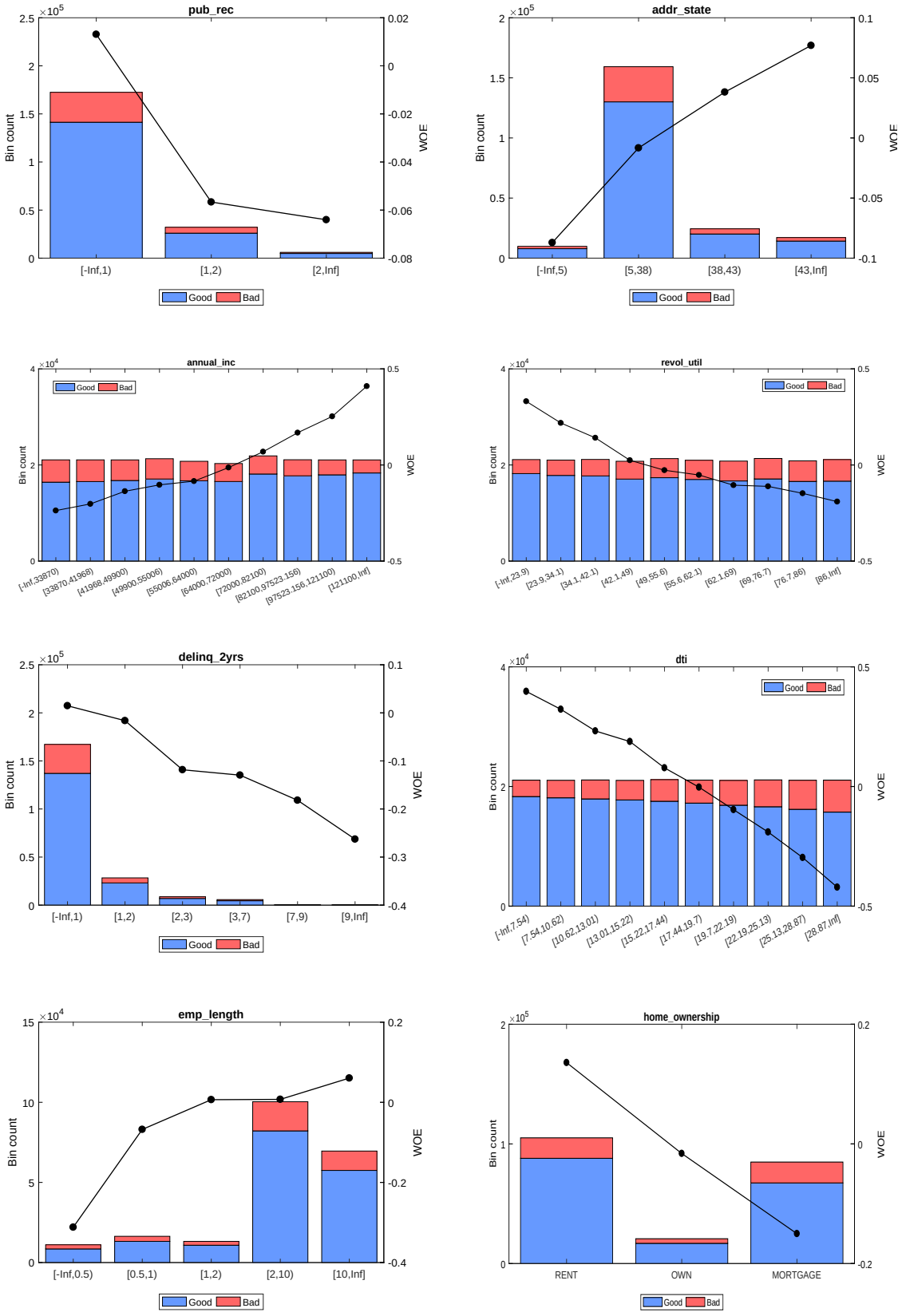


Figura 2.13: Analisi andamento del WOE

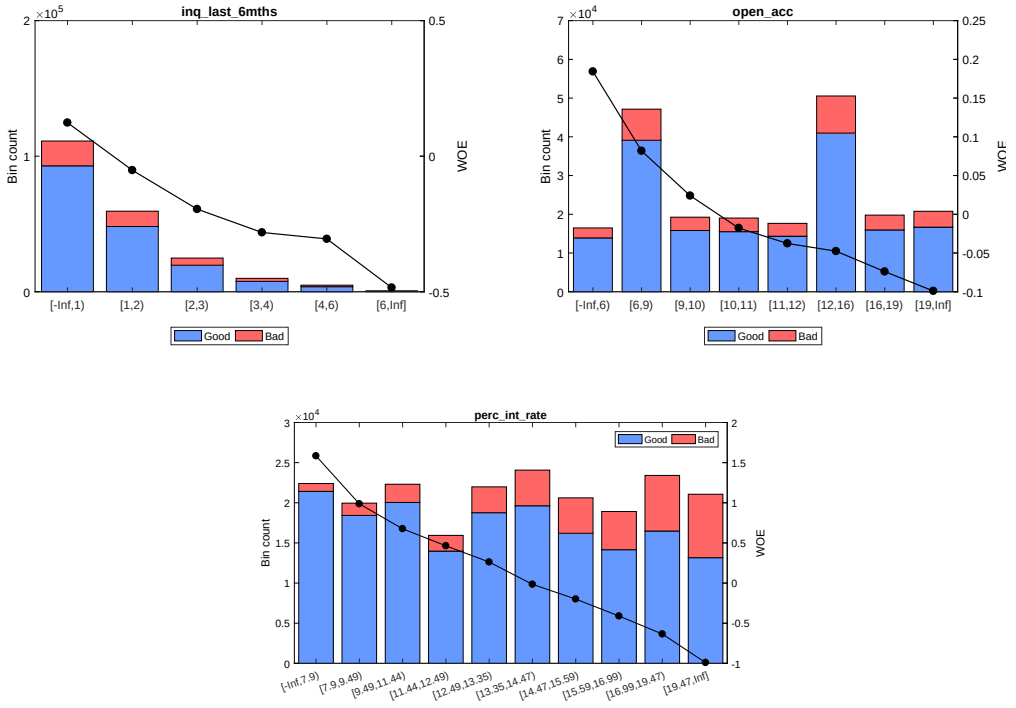


Figura 2.13: Analisi andamento del WOE

p=probabilità dell'evento target,
x=variabili input
β_0=intercetta della linea di regressione
β_k=parametri
e=termine d'errore

Poiché $\frac{p_i}{1-p_i}$ assume valori tra 0 e ∞, $log\left(\frac{p_i}{1-p_i}\right)$ assume valori tra $-\infty$ e $+\infty$. Va ricordato che la regressione logistica richiede delle assunzioni: richiede una variabile target categoriale;in secondo luogo deve esserci una relazione lineare con la funzione logaritmica delle probabilità, tutto ciò che non lo è dovrebbe essere linearizzato o trasformato in variabile dummy ove possibile; i termini d'errore devono essere indipendenti e i predittori non correlati, altrimenti ci sarebbe un problema di multicollinearità. Infine si devono utilizzare variabili rilevanti. Il risultato *Logit* (p_i) rappresenta la trasformazione logaritmica dell'output, ovvero $log(p[evento]/p[nonevento])$, ed è utilizzata per linearizzare la probabilità in modo da avere un risultato dicotomico, 0 o 1. I parametri β_1 e β_k vengono stimati con il metodo di massima verosomiglianza e misurano il tasso di variazione nel modello per una variazione unitaria del valore della variabile input considerata, considerato anche un aggiustamento per tutte le altre variabili presenti nella regressione. I parametri sono influenzati dall'unità di input, si pensi a caratteristiche espresse in

percentuale piuttosto che in valore assoluto, quindi devono essere standardizzate oppure si può sostituire il valore della caratteristica di input con il WOE per ogni raggruppamento di essa creato. Nella nostra analisi si utilizzerà questa seconda soluzione:

$$Logit\ (p_i) = \beta_0 + \beta_1 WOE_1(i) + \cdots + \beta_k WOE_k(i) + e \tag{2.13}$$

in quanto l'utilizzo del WOE al posto dei valori grezzi delle variabili, non solo permette di bypassare l'influenza della misura di scala ma anche di tenere in considerazione l'esatto trend e la scala della relazione tra un gruppo e l'altro dei valori del predittore. In altre parole, considerare le variabili in termini di WOE associato ai raggruppamenti dei loro valori, permette al modello di considerare le relazioni non lineari dei predittori e di assicurare che l'assegnazione del punteggio ad ogni gruppo della variabile mantenga una scala di score logica e reale.

Esempio 2.
Per chiarire, la differenza tra un *loan-to-value* del 10% e uno del 20% non sarà la stessa della differenza tra un *loan-to-value* del 70% e uno di 80%.

Lo score creato sul *WOE-binning* restituisce sempre uno score determinato dalle curve WOE delle variabili e permette di inserire anche tutte le variabili selezionate (da cui deriva l'appellativo *"all possible regression"*). Ad ogni modo lo sviluppatore dovrà sempre affidarsi a misure statistiche per verificare la qualità del modello finale, come il p-value o il coefficiente di Gini di cui si parlerà in seguito. Quello che si vuole fare è assegnare punti ad ogni caratteristiche di ogni variabile esplicativa, la cui somma restituirà lo score finale che potrà essere interpretato come probabilità di default. In questa situazione la regressione non sarà omogenea in tutti gli anni perché il dataset è modificato da Lending Club in modo molto frequente, così come la numerosità campionaria ed i casi presenti. Come risultato si hanno variabili che in alcuni anni sono concentrate su un valore solo e quindi inutili al fine della regressione, oppure variabili non significative posto un $\alpha = 5\%$. Ovviamente saranno eliminate dal modello. La tabella 2.2 nella pagina successiva riporta le variabili utilizzate con i coefficienti stimati e i relativi p-value.

Tabella 2.2: P-value per variabile secondo anno

	2007		2008		2009		2010		2011		2012		2013		2014	
	Est.	pVal	Est.	pVal	Est.	pVal	Est.	pVal	Est.	pVal	Est.	pVal	Est.	pVal	Est.	pVal
(Int.)	1.015	0.000	1.344	0.000	1.842	0.000	1.816	0.000	1.726	0.000	1.645	0.000	1.650	0.000	1.501	0.000
loan_amnt			0.612	0.003	1.200	0.001					0.269	0.001	0.694	0.000	0.513	0.000
yr_earliest_cr_line											0.418	0.021	0.427	0.000	0.729	0.000
acc_now_delinq																
pub_rec_bankruptcies													1.252	0.016		
pub_rec			0.718	0.009	0.722	0.038	0.549	0.000	0.503	0.001	0.680	0.002			0.626	0.005
addr_state			1.099	0.000	1.316	0.000	1.219	0.027			1.109	0.000	1.044	0.000	0.996	0.000
annual_inc					1.552	0.000	1.220	0.000	1.458	0.000	1.489	0.000	1.194	0.000	0.977	0.000
revol_util							0.513	0.000	0.448	0.000					0.220	0.000
delinq_2yrs									0.471	0.043					1.089	0.000
dti					0.940	0.002					0.281	0.000	0.426	0.000	0.436	0.000
emp_length	1.284	0.019					1.002	0.000	1.159	0.000	1.221	0.000	1.093	0.000	1.144	0.000
home_ownership			0.668	0.009	0.715	0.045					0.597	0.000	0.547	0.000		
term							0.726	0.000	0.612	0.000	0.517	0.000	0.576	0.000	0.571	0.000
purpose	1.233	0.000	0.839	0.000	0.774	0.000	1.004	0.000	0.858	0.000	0.890	0.000	0.173	0.029	0.503	0.000
inq_last_6mths	0.690	0.000	0.859	0.000	0.819	0.000	0.762	0.000	0.572	0.000	0.578	0.000	0.300	0.000	0.352	0.000
open_acc	1.596	0.014									0.666	0.002	0.983	0.000	0.863	0.000
perc_int_rate	1.019	0.000	0.797	0.000	0.729	0.000	0.635	0.000	0.661	0.000	0.771	0.000	0.702	0.000	0.710	0.000
obs.	572		2389		5267		7635		21706		53256		127450		210662	

In sintesi le regressioni annue si riducono a [12]:

$$\begin{aligned}
def2007 =& Int. + emp_length + purpose + inq_last_6mths + open_acc \\
&+ perc_int_rate
\end{aligned} \tag{2.14}$$

$$\begin{aligned}
def2008 =& Int. + loan_amnt + pub_rec + addr_state + purpose \\
&+ home_ownership + inq_last_6mths + perc_int_rate
\end{aligned} \tag{2.15}$$

$$\begin{aligned}
def2009 =& Int. + loan_amnt + pub_rec + addr_state + annual_inc \\
&+ dti + home_ownership + purpose \\
&+ inq_last_6mths + perc_int_rate
\end{aligned} \tag{2.16}$$

$$\begin{aligned}
def2010 =& Int + pub_rec + addr_state + annual_inc + revol_util \\
&+ delinq_2yrs + emp_length + term + purpose \\
&+ inq_last_6mths + perc_int_rate
\end{aligned} \tag{2.17}$$

$$\begin{aligned}
def2011 =& Int + pub_rec + annual_inc + revol_util + emp_length \\
&+ purpose + delinq_2yrs + inq_last_6mths \\
&+ term + perc_int_rate
\end{aligned} \tag{2.18}$$

$$\begin{aligned}
def2012 =& Int + loan_amnt + yr_earliest_cr_line + pub_rec + dti \\
&+ addr_state + annual_inc + emp_length + term \\
&+ home_ownership + purpose + open_acc \\
&+ inq_last_6mths + perc_int_rate
\end{aligned} \tag{2.19}$$

$$\begin{aligned}
def2013 =& Int + loan_amnt + yr_earliest_cr_line + addr_state \\
&+ pub_rec_bankruptcies + annual_inc + emp_length \\
&+ home_ownership + term + dti + purpose + open_acc \\
&+ perc_int_rate
\end{aligned} \tag{2.20}$$

$$\begin{aligned}
def2014 =& Int + loan_amnt + yr_earliest_cr_line + annual_inc \\
&+ pub_rec + revol_util + delinq_2yrs + dti + emp_length \\
&+ term + addr_state + purpose + inq_last_6mths \\
&+ open_acc + perc_int_rate
\end{aligned} \tag{2.21}$$

ed è evidente come gli anni 2012-2014 siano segnati da un enorme ampliamento del business e dal miglioramento del dataset, infatti presentano regressioni pressoché identiche.

Ottenuto il risultato finale (in termini di intercetta, stima dei parametri e per-

[12] Le regressioni ottenute sono simili a quanto riscontrato nel paper *Modeling default for peer-to-peer loans*

formance statistiche), si può procedere a ridimensionare il range e il formato dello score risultante. In generale, la relazioni tra le probabilità e lo score possono essere rappresentate come una relazione lineare[13]:

$$Score = Factor * ln(odds) + offset \tag{2.22}$$

$$= \sum_{j,i=1}^{k,n} \left(-\left(WOE_j * \beta_i + \frac{a}{n} \right) * factor + \frac{offset}{n} \right) \tag{2.23}$$

dove:

WOE=weight of evidence per ogni gruppo di attributo, non moltiplicato per 100;
a=intercetta della regressione;
β=coefficiente di regressione per ogni caratteristica;
n=numero di caratteristiche nel modello;
k=numero di gruppi per caratteristiche.

Questa operazione non intacca la capacità predittiva del modello, ma è semplicemente una decisione presa su considerazioni come la possibilità di implementazione nel processo di *application*, ovvero la domanda per l'accesso al credito, o per facilitare la comprensione. Lending Club non affida nessuna controparte sotto un FICO score di 660 punti e sapendo che questo score arriva ad un massimo di 800, si è deciso di ridimensionare la regressione considerando un *WorstAndBestScore* con un range di 660-800 punti [14], ovviamente. A questo punto si è ottenuto uno score finale per ogni debitore nel portafoglio con annessa probabilità di default. La figura 2.14 dimostra come la curva delle probabilità di default di tre anni *in sample* presi a campione corrisponda ai criteri di monotonicità crescente e figura a "S".

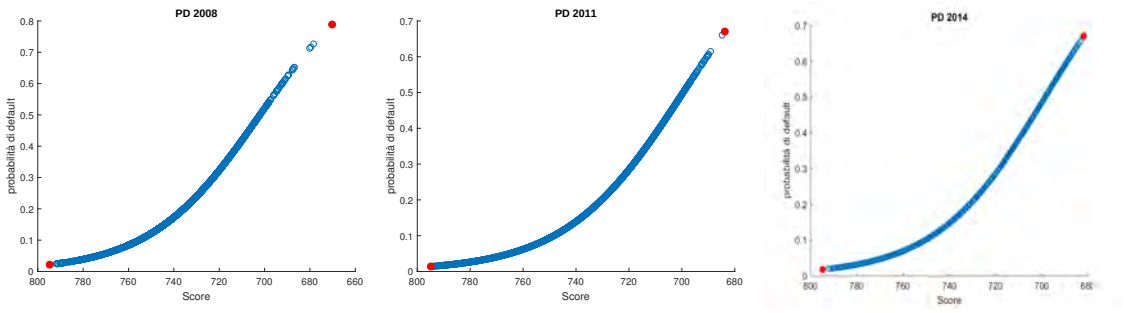


Figura 2.14: Curve di probabilità di default per gli anni campione

[13] per lo sviluppo vedere Siddiqi 2017, p. 242
[14] www.valuepenguin.com/personal-loans/lending-club, Eligibility LC

Prima di procedere alla creazione della *master scale* per definire la probabilità di default da associare al rating, si deve analizzare la capacità predittiva del modello implementato.

Misure di separazione e divergenza

Il modello di scoring necessita un'analisi per verificare la divergenza della distribuzione sottostante il modello di scoring creato da quella del modello reale. Esiste, per questo scopo, una lunga lista di misure per valutare quanto bene il modello descrive i dati o prevede lo stato di insolvenza. Si tratta di statistiche bivariate che hanno origine da diverse discipline, che hanno lo scopo di misurare:la**capacità di ordinamento** intesa come la capacità di identificare la dipendenza fra caratteristiche, score e risultato binomiale. Questo è il principale punto di interesse dell'analisi in quanto maggiore capacità di ordinamento aggiunge valore al modello. La **divergenza** ovvero variazione tra i risultati attesi e ottenuti, cercando la minore differenza possibile.

A questo fine si utilizzeranno la curva KS, la CAP curve e la curva ROC-Receiver Operating Characteristic, ognuna delle quali ha differenti punti di forza e di debolezza a seconda della situazione, quindi non dovrebbero mai essere usate singolarmente.

Kolmogorov-Smirnov (KS) Una delle statistiche più usate nel credit scoring, così come in molte altre discipline, è la "curva KS", conosciuta anche come "grafico a occhio di pesce". Questa statistica è stata sviluppata da due matematici sovietici, A.N.Kolmogorov e N.V.Smirnov per l'appunto, e proposta in Italia per la prima volta nel 1933.
Come il coefficiente di Gini, è una delle diverse statistiche costruite sull'analisi della distribuzione cumulata empirica, una per *bads* e una per i *goods*, ed identifica la massima deviazione tra le distribuzioni modellate rispetto alle distribuzioni generatrici dei dati. Secondo Raymond Anderson (2007, p. 196) il valore KS dovrebbe essere compreso tra il 20% e il 70%; sotto la soglia inferiore ci si dovrebbe interrogare come migliorare il modello, mentre sopra tale range è probabilmente "troppo bello per essere vero". La distribuzione cumulata dei solventi e quella dei non solventi vengono rappresentate contro lo score come in figura 2.15[15], identificando la percentuale di "buoni" e di "cattivi" che si osserva sotto una soglia di

[15]Raymond Anderson 2007, p. 196.

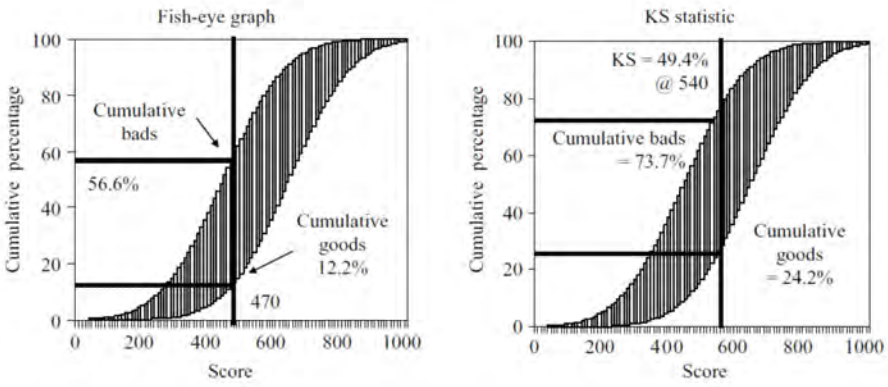


Figura 2.15: Esempi di curve KS

score, identificata come il punto di maggiore divergenza.

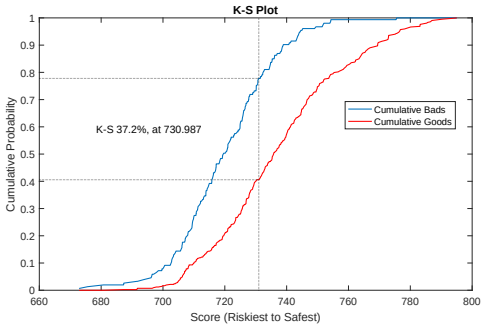
$$D_{KS} = max\{abs(cpY - cpX)\} \tag{2.24}$$

Si riportano i grafici del campione *in sample* per verificare il rispetto del range per la validità (figura 2.16 nella pagina successiva). Tutti gli anni sono quindi sopra la soglia minima, anche se non hanno un indice KS molto alto. La performance peggiore si riscontra nell'anno Sebbene questa misura sia molto facile da interpretare, spesso potrebbe rivelarsi troppo semplicistica quindi solitamente viene utilizzata insieme ad altre misure.

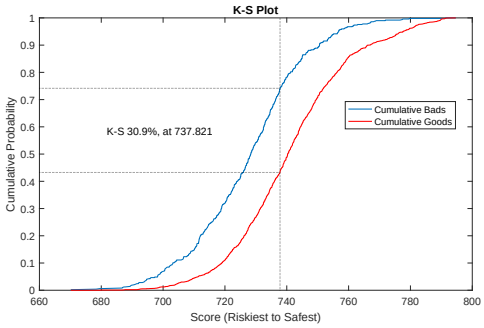
Curva di Lorenz e coefficiente di Gini Tra il 1800 e il 1900 si è vista una crescente attenzione per la distribuzione del reddito nei vari stati. Un matematico americano sviluppò, nel 1905, la curva di Lorenz, per rappresentare graficamente la distribuzione del reddito nella società (fig. 2.17 a pagina 58)[16].
Le entrate sono ordinate in ordine decrescente e viene calcolata la cumulata dei flussi in entrata e della popolazione come $cpV_i = \frac{\Sigma_{j=1}^{i} V_i}{\Sigma V}$ il cui risultato viene rappresentato in un grafico XY come in figura 2.17 a pagina 58 dove viene indicato che il 70% delle entrate viene detenuto dal 20% della popolazione totale.

La perfetta uguaglianza giace esattamente sulla diagonale, mentre la perfetta disuguaglianza dovrebbe coprire tutto lo spazio sopra di essa; quindi l'effettiva disuguaglianza è rappresentata nel grafico dallo spazio compreso tra la diagonale e la curva.
Nei modelli di credit scoring viene utilizzata come analisi della capacità del modello di separare *goods* e *bads accounts*, rappresentando la cumulata dei defaultati in

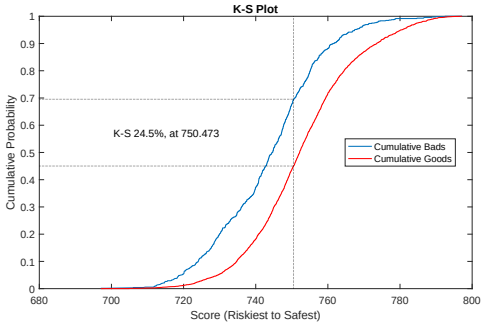
[16]Raymond Anderson 2007, p. 203.



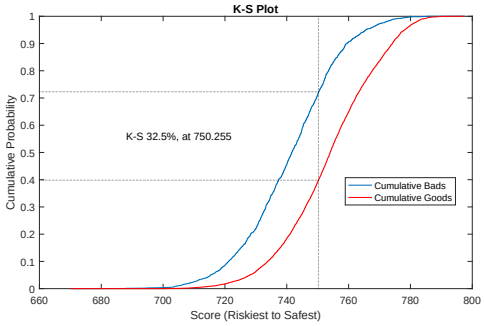
(a) Anno 2007



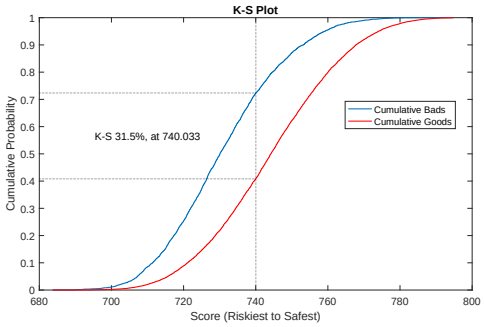
(b) Anno 2008



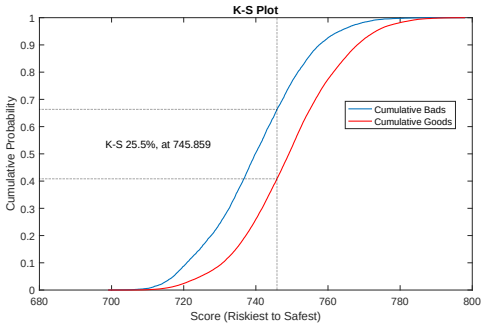
(c) Anno 2009



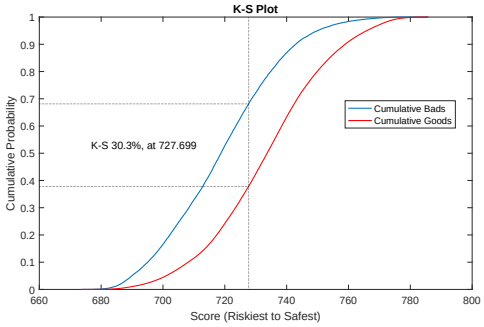
(d) Anno 2010



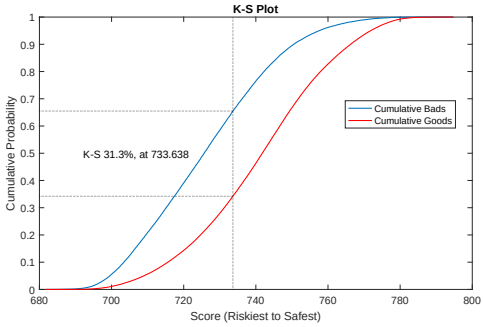
(e) Anno 2011



(f) Anno 2012



(g) Anno 2013



(h) Anno 2014

Figura 2.16: Curve KS per gli anni in sample

un asse e la cumulata dei solventi nell'altro. Un modello senza capacità predittiva implica perfetta uguaglianza, mentre la perfetta disuguaglianza è propria di un modello perfettamente capace di separare solventi da non solventi.

Coefficiente di Gini Nel 1910 un altro contributo arrivò da Corrado Gini che confrontò la disuguaglianza tra stati con un metodo riconosciuto oggi come coefficiente di Gini.
Attraverso la formula:

$$D = 1 - \sum_{i=1}^{n} \big((cpY_i - cpY_{i-1})(cpX_i + cpX_{i-1}) \big) \tag{2.25}$$

in cui cpY è la cumulata percentuale delle entrate e cpX è la cumulata percentuale della popolazione, si identifica l'area tra la curva e la diagonale, come percentuale dell'area sopra la diagonale.
Il risultato è un coefficiente di correlazione di rango utilizzato non tanto per i test di ipotesi, ma come misura di separazione. Il coefficiente di Gini misura quanto bene un modello di scoring è in grado di distinguere tra *goods* e *bads*. Questa misura soffre però di alcuni limiti. Questa misura può essere esagerata aumentando il range degli indeterminati come per esempio i correnti, ed è sensibile alle definizioni categoriali, va quindi prestata attenzione alla sua interpretazione e se ne raccomanda l'utilizzo parallelo ad altre misure.
Quando il coefficiente di Gini ha un valore accettabile? Sebbene non esistano regole diverse da *"la regola del pollice"*, un valore del coefficiente di Gini maggiore o uguale del 50% può essere considerato soddisfacente, mentre sotto il 35% è considerato sospetto. Un valore di tale statistica sotto il 30% è inaccettabile, se possibile.

ROC curve-Receiver Operating Characteristic curve Sebbene non si utilizzeranno in questo lavoro, sembrava opportuno ricordare la curva di Lorenz e il coefficiente di Gini per comprendere meglio la curva ROC.
Negli anni '40 venne sviluppata la curva ROC allo scopo di misurare la capacità dei radar di distinguere tra un segnale vero e un rumore. La stessa statistica venne adottata, nel periodo 1950-1960, del campo della psicologia per lo studio di modelli comportamentali che non avrebbero potuto essere spiegati con le teorie esistenti.
Al giorno d'oggi la ROC curve viene usata in moltissimi settori, incluso il credit scoring, in quanto riesce ad analizzare: la **sensitività** o meglio la capacità di individuare i *veri positivi* e la **specificità** ,ovvero l' abilità di identificare *veri negativi*.

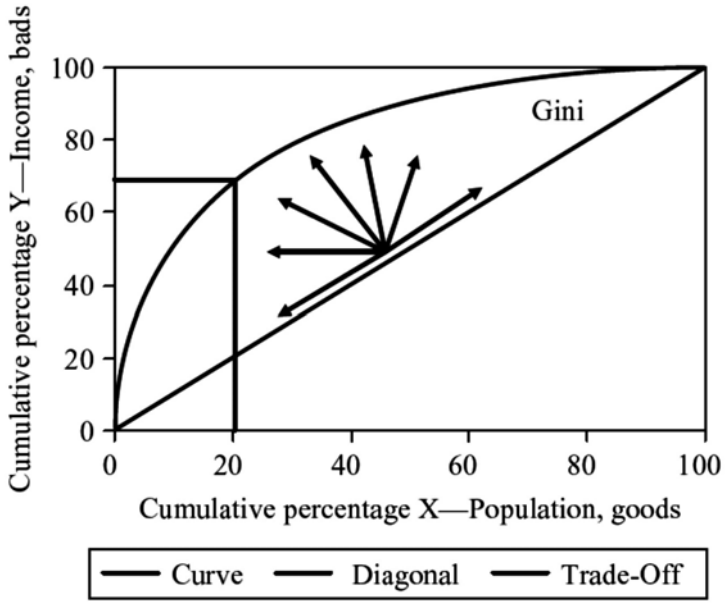


Figura 2.17: Esempio di curva di Lorenz

Secondo la sua costruzione ne deriva un grafico[17] di facile comprensione dove le ascisse rappresentano l'indice di sensitività e le ordinate, invece, il complemento a uno della specificità. Si disegna quindi un curva come in figura 2.18, dove:

$$X = Pr[S_{FP} \leq S_{Cut-off}] \tag{2.26}$$

$$Y = 1 - \big(Pr[S_{TP} \leq S_{Cut-off}]\big) \tag{2.27}$$

La concavità nella curva è equivalente alla probabilità di default condizionale come funzione decrescente dello score, mentre la non concavità indica un utilizzo sub-ottimale delle informazioni utilizzate per la specificazione dello score. Anche in questo caso è presente una statistica sommaria che rappresenta l'area sotto la curva come percentuale dell'area totale sopra la diagonale. Si tratta del coefficiente AUROC, conosciuto anche come c-statistic, la cui formula è:

$$\textbf{AUROC} \quad c_{P,N} = Pr[S_{TP} < S_{TN}] + 0.5Pr[S_{TP} = S_{TN}] \tag{2.28}$$

ed è possibile mettere in relazione con il coefficiente di Gini attraverso il rapporto $c \approx (D+1)/2$.
In altri termini l'area sotto la curva rappresenta la probabilità che il rating di un defaultato sia minore di quello di una controparte solvente, più il 50% di probabilità che i due rating siano uguali. Ovviamente un AUROC del 50% rappresenta

[17] tratto da Raymond Anderson 2007, p. 207

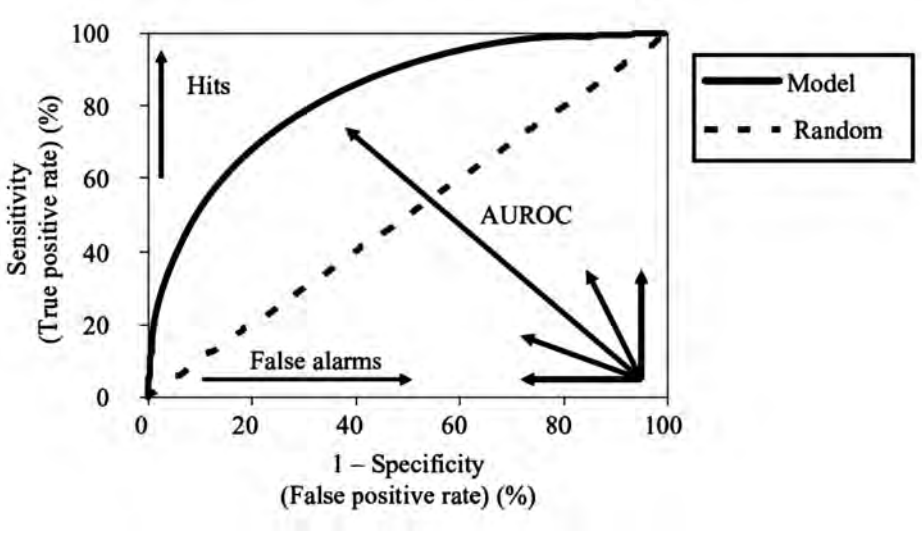
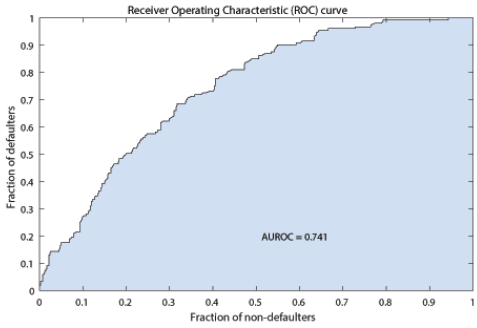


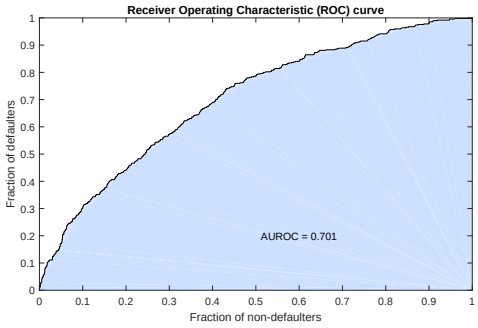
Figura 2.18: Esempio di ROC curve

un modello random mentre più si avvicina al 100%, più il modello sarà capace di predire i default.
Risulta evidente, a questo punto, che la curva di Lorenz e la ROC sono quasi esattamente la stessa cosa, il coefficiente di Gini e l'AUROC sono estremamente correlati.
Nel tempo si è arrivati, quindi, ad utilizzare la ROC curve con la relativa statistica AUROC, al posto della curva di Lorenz e del coefficiente di Gini. Si nota come, all'aumentare della numerosità campionaria e al migliorare dei dati raccolti, migliori il coefficiente AUROC e la ROC copra sempre di più l'area sopra la diagonale (figura 2.19 nella pagina successiva). Come per la statistica KS, l'AUROC presenta valori sopra la soglia minima sebbene non eccellenti, infatti tutti gli anni del campione *in sample* si attestano intorno ad un indice di circa 0.70.

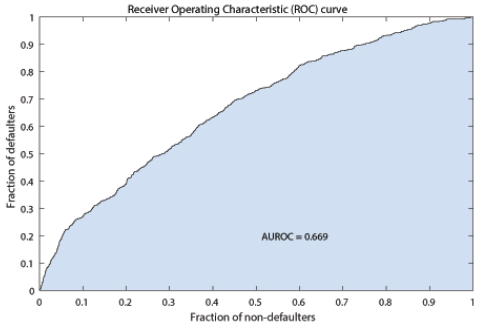
CAP curve-Cumulative Accuracy Profile Sulla base della curva di Lorenz e il coefficiente di Gini è stata sviluppata la curva CAP, le cui caratteristiche statistiche sono le stesse della curva ROC. La concavità della curva disegnata con questa statistica indica che le probabilità di default dato uno score sottostante, creano una distribuzione di score decrescente, viceversa, la non concavità indica un utilizzo sub-ottimale delle informazioni utilizzate per creare il modello. La forma di questa curva dipende dalla proporzione tra solventi e non solventi, quindi una comparazione tra portafogli diversi può essere fuorviante. Secondo quanto riportato da *Studies on the Validation of Internal Rating System*, p. 38, l'Accuracy Ratio dovrebbe assumere un valore tra il 50% e l'80%, ad ogni modo dovrebbe essere interpretato con attenzione in quanto fortemente dipendenti dal numero di



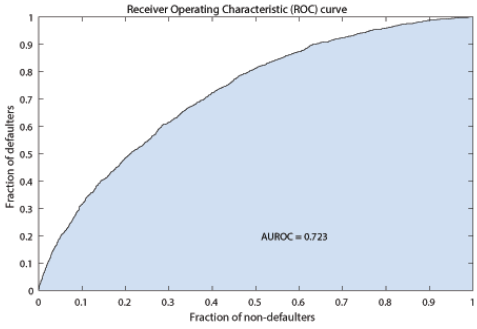
(a) Anno 2007



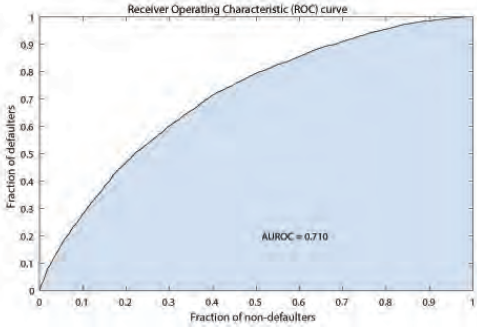
(b) Anno 2008



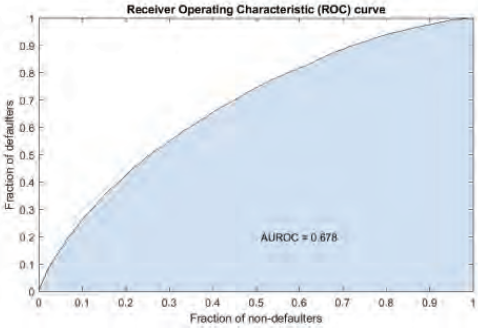
(c) Anno 2009



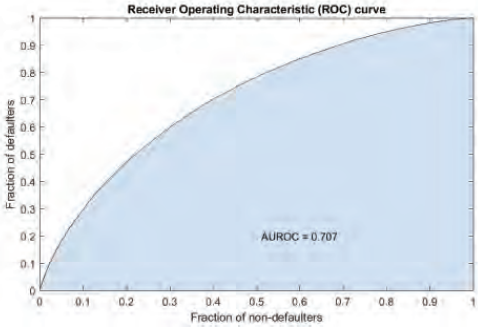
(d) Anno 2010



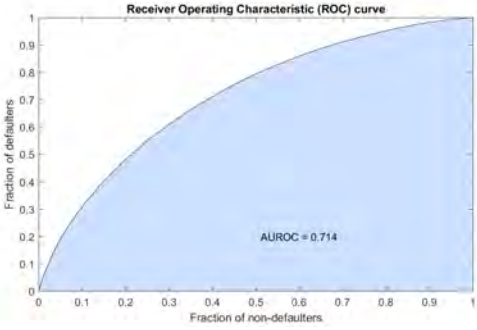
(e) Anno 2011



(f) Anno 2012



(g) Anno 2013



(h) Anno 2014

Figura 2.19: Curve ROC per gli anni in sample

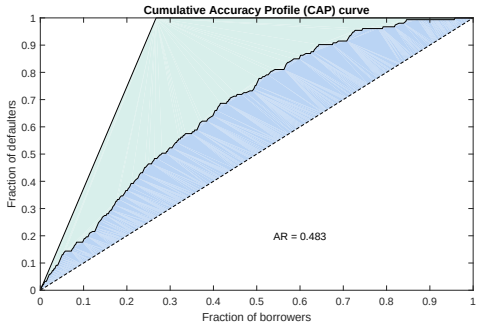
default presenti nel portafoglio. Per completezza si riporta anche la CAP curve dei tre anni presi a campione. La figura **??** a pagina ?? dimostra come l'AR si attesti sempre intorno al 40%, inferiore alla soglia minima consigliata ma, essendo le altre statistiche corrette ed essendo il migliore valore ottenibile dal dataset analizzato, si mantengono le regressioni come riportate nell'equazione 2.22 a pagina 54.

Si riportano tutte le statistiche per tutti gli anni analizzati nella tabella 2.3. Si puo notare una certa persistenza dei valori nell'intorno di 0.45 per l'indice AR, di 0,7 per l'AUROC e di 0,3 della statistica KS

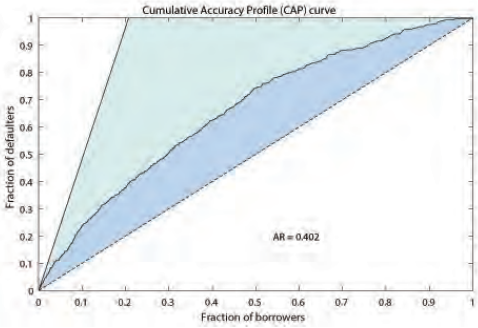
Tabella 2.3: Statistiche di separazione e divergenza

Misura	2007	2008	2009	2010	2011	2012	2013	2014
AR	0.483	0.402	0.338	0.447	0.420	0.356	0.415	0.429
AUROC	0.741	0.701	0.669	0.723	0.710	0.678	0.707	0.714
KS stat.	0.372	0.309	0.245	0.325	0.315	0.255	0.303	0.313
obs.	572	2389	5267	7635	21706	53256	127450	210662

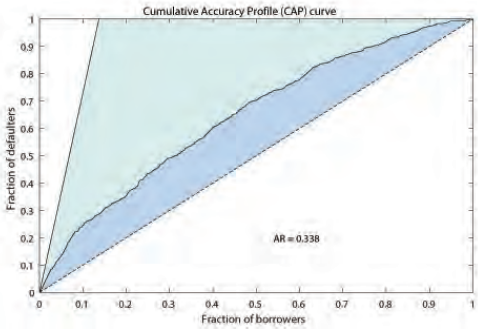
Soddisfatti della forza predittiva delle regressioni effettuate, riassunta nella tabella 2.3, si procede a creare la master scale per associare una probabilità di default al rating. Per arrivare ad una scala di rating come siamo abituati a vederla, nel mondo degli istituti di credito e dei professionisti, si utilizza un cut-off per ampliamento delle classi. In questo lavoro, invece si taglieranno le classi seguendo la composizione di Lending Club attraverso la composizione di portafoglio, al fine di catturare la frequenza dei default per ogni classe correttamente. Attraverso questo procedimento si ottiene una *master scale* annuale come in figura 2.21 a pagina 64, su cui verrà calcolata la media del *panel* e la deviazione standard necessari per il modello di portafoglio. Per ogni anno si riporta anche la distribuzione delle classi, utile per avere una prima idea della probabilità di default complessiva del gruppo di controparti. Per scendere nei dettagli si riporta anche una tabella che presenta numericamente l'output finale. La struttura del rating creata presenta alcune probabilità sicuramente diverse da quelle stimate da LC, ma questo è dovuto dalla mancanza di alcuni dati protetti da privacy come, per esempio, il FICO score iniziale della controparte. Mediamente comunque, la classe minore si attesta intorno al 4% di probabilità di default, mentre la classe peggiore intorno al 50%.



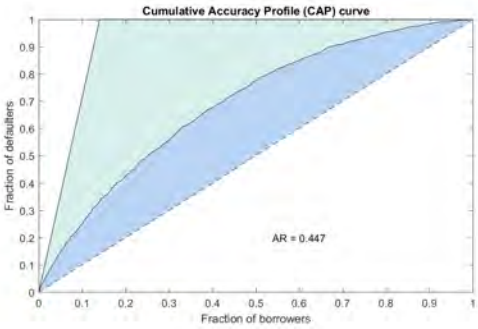
(a) Anno 2007



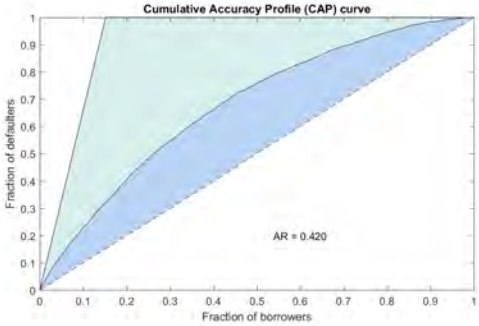
(b) Anno 2008



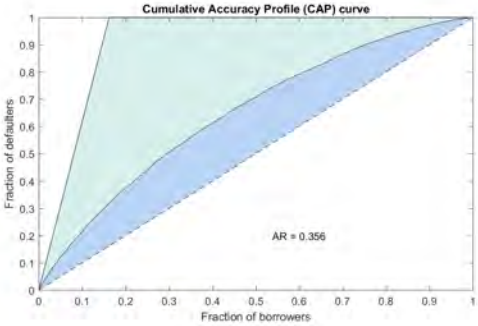
(c) Anno 2009



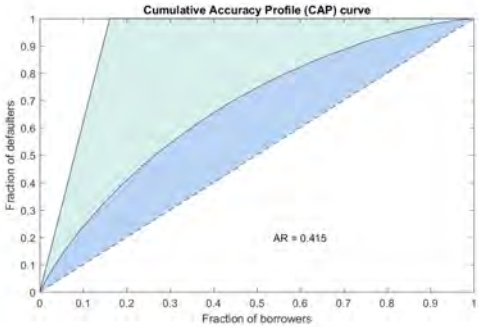
(d) Anno 2010



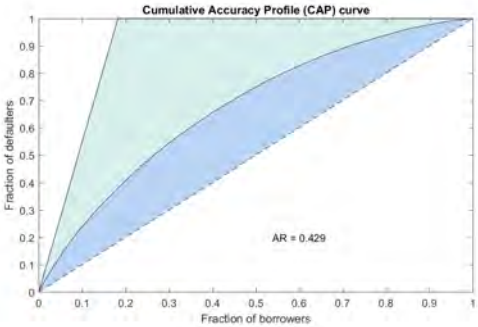
(e) Anno 2011



(f) Anno 2012



(g) Anno 2013



(h) Anno 2014

Figura 2.20: Curve CAP per gli anni in sample

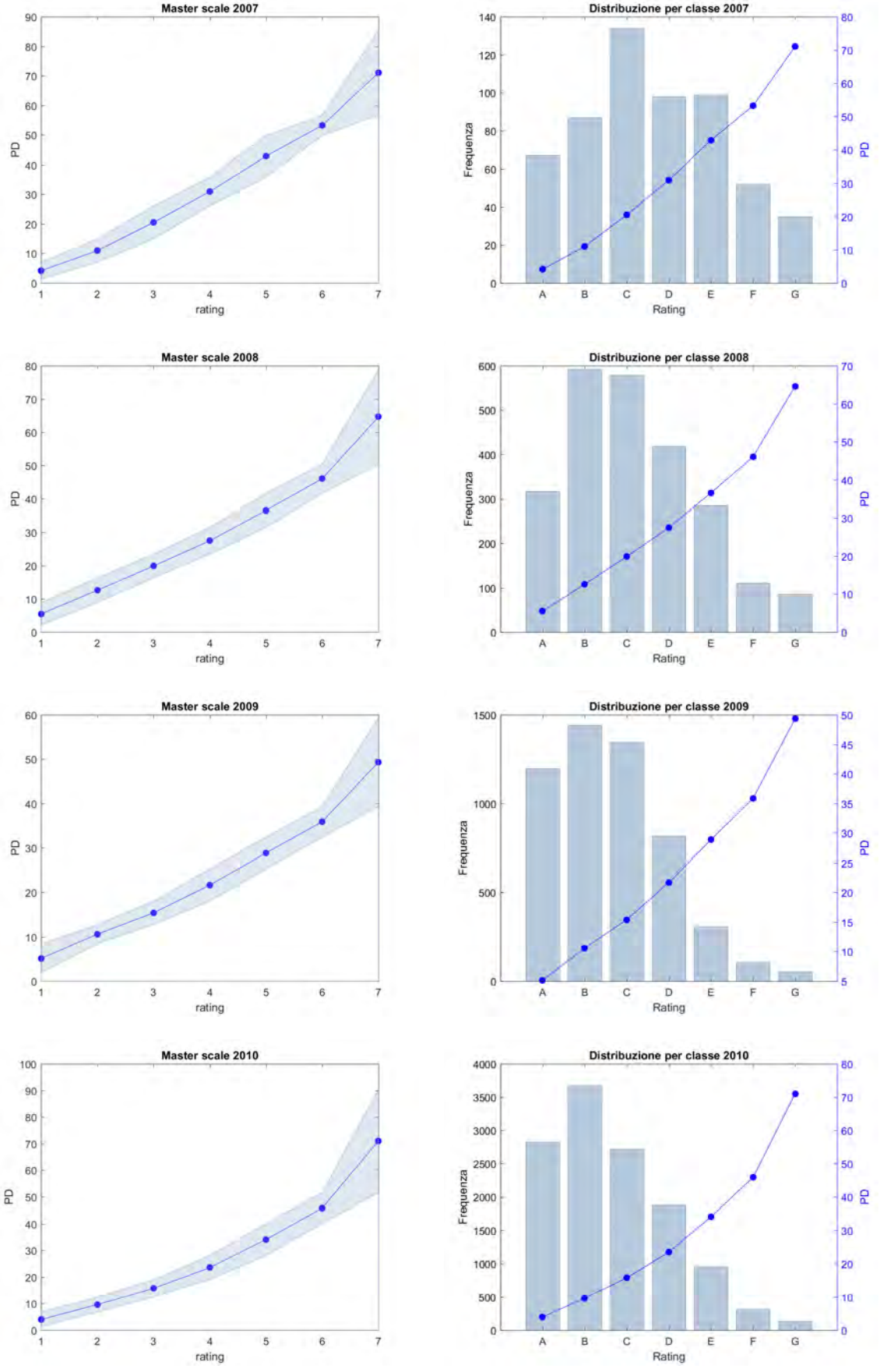


Figura 2.21: Creazione delle classi di rating per anno

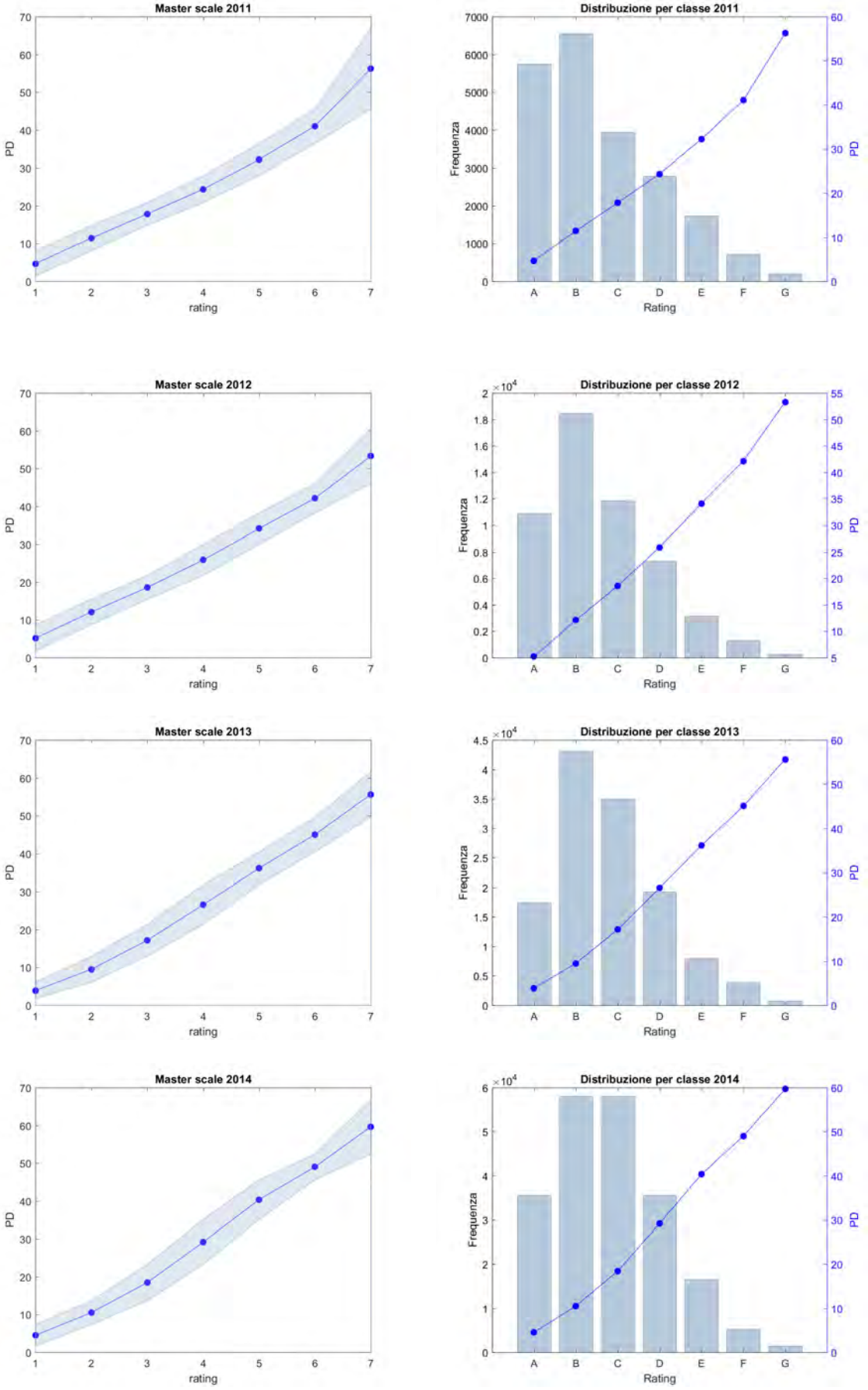


Figura 2.21: Creazione delle classi di rating per anno

Tabella 2.4: Tabella delle masterscale annuali

	Anno 2007			Anno 2008			Anno 2009			Anno 2010		
	Range(%)		PD(%)	Range(%)		PD(%)	Range(%)		PD(%)	Range(%)		PD(%)
A	0.00	- 7.00	4.14	0.00	- 8.86	5.49	0.00	- 8.40	5.13	0.00	- 6.82	4.00
B	7.01	- 14.95	10.98	8.87	- 16.30	12.58	8.41	- 12.77	10.59	6.83	- 12.46	9.64
C	14.96	- 26.10	20.53	16.31	- 23.46	19.88	12.78	- 18.03	15.40	12.47	- 18.92	15.69
D	26.11	- 35.86	30.98	23.47	- 31.50	27.48	18.04	- 25.30	21.67	18.93	- 28.08	23.50
E	35.87	- 49.78	42.82	31.51	- 41.59	36.55	25.31	- 32.40	28.85	28.09	- 40.06	34.07
F	49.79	- 56.65	53.22	41.60	- 50.46	46.03	32.41	- 39.34	35.87	40.07	- 51.58	45.82
G	56.66	- 85.47	71.06	50.47	- 78.82	64.64	39.35	- 59.37	49.35	51.59	- 90.46	71.02

	Anno 2011			Anno 2012			Anno 2013			Anno 2014		
	Range(%)		PD(%)	Range(%)		PD(%)	Range(%)		PD(%)	Range(%)		PD(%)
A	0.00	- 8.03	4.69	0.00	- 8.72	5.19	0.00	- 6.07	3.88	0.00	- 7.34	4.49
B	8.04	- 14.84	11.44	8.73	- 15.45	12.09	6.08	- 12.94	9.50	7.35	- 13.66	10.50
C	14.85	- 20.83	17.84	15.46	- 21.75	18.60	12.95	- 21.37	17.15	13.67	- 23.26	18.50
D	20.84	- 27.96	24.40	21.76	- 29.97	25.86	21.38	- 31.78	26.57	23.27	- 35.08	29.17
E	27.97	- 36.43	32.20	29.98	- 38.24	34.10	31.79	- 40.44	36.11	35.09	- 45.56	40.32
F	36.44	- 45.54	40.99	38.25	- 46.00	42.12	40.45	- 49.61	45.03	45.57	- 52.43	50.00
G	45.55	- 67.10	56.27	46.01	- 60.66	53.32	49.62	- 61.65	55.63	52.44	- 66.93	59.67

2.3.3 Validazione e Backtesting

Una volta definito il modello di score in tutti i suoi dettagli, si devono validare i risultati ottenuti con una validazione *"standard"*, ovvero creata su un *out-of-sample*, per confermare l'applicabilità del modello. Generalmente si utilizzano il 70-80% delle osservazioni come campione per l'implementazione, mantenendo il restante 20-30% come *out of sample*. In alternativa, come suggerito da Siddiqi (2017, p. 258), in presenza di pochi dati, si può sviluppare il modello su tutto il *dataset* disponibile, estraendo poi casualmente tra 50% e 80% delle osservazioni per procedere alla validazione. In questa analisi si utilizzeranno gli ultimi 3 anni come *out of sample*, ossia il periodo compreso tra 1 Gennaio 2015 e 31 Dicembre 2017; procedendo ad analizzare la divergenza tra le distribuzione di *goods* e *bads in sample* con quelle *out of sample*, nonché le statistiche di forza. Ovviamente il modello verrà validato se la deviazione non sarà significativa. Solitamente un esame visivo è sufficiente a tal fine ma un ulteriore aiuto è fornito dal calcolo della divergenza:

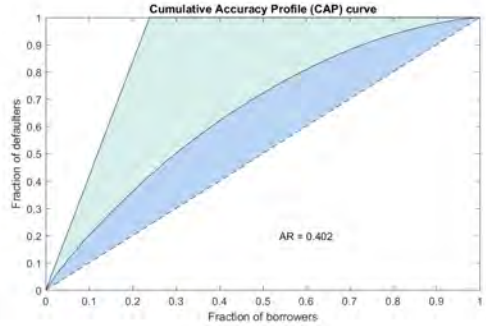
$$Divergenza = \Big((mean_G - mean_B)^2 / \big[0.5(var_G + var_B)\big]\Big) \tag{2.29}$$

dove $mean_G$, $mean_B$, $varG$ e var_B sono la media e la varianza rispettivamente di *goods* e *bads*.

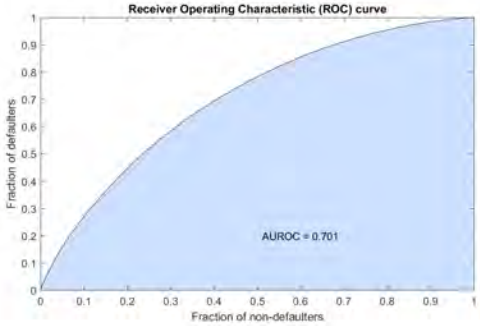
Si osservano prima i risultati della curva CAP e della curva ROC (figura 2.22 a pagina 67). Prendendo come riferimento il 2014, ultima finestra temporale di un anno considerata, si notano delle divergenze in tali indici nei primi due anni *out of sample*, mentre il 2017 presenta una netta diminuzione in termini percentuali. Per la curva ROC si traduce rispettivamente in -1.79% nel 2015, del -3.3

Tabella 2.5: Statistiche di separazione e divergenza out of sample

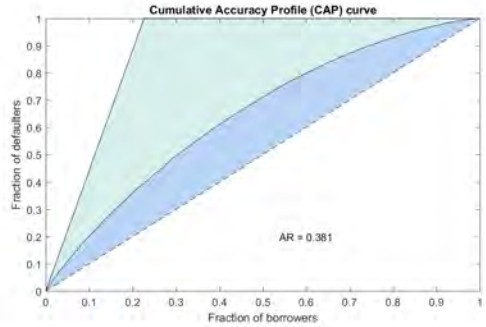
Misura	2014	2015	2016	2017
AR	0.429	0.402	0.381	0.386
Δ AR		−6.220%	−11.400%	−9.960%
AUROC	0.714	0.701	0.690	0.693
Δ AUROC		−1.790%	−3.300%	−2.900%
KS stat.	0.313	0.293	0.274	0.283
Δ KS		−6.400%	−12.300%	−9.700%
obs.	210662	189438	155350	5267



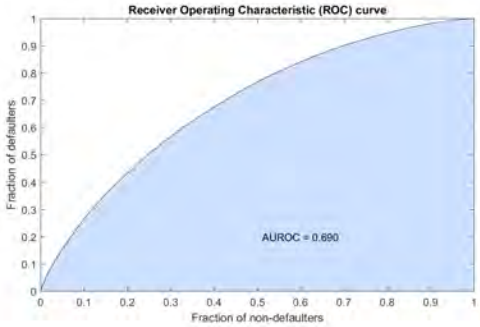
(a) CAP curve anno 2015



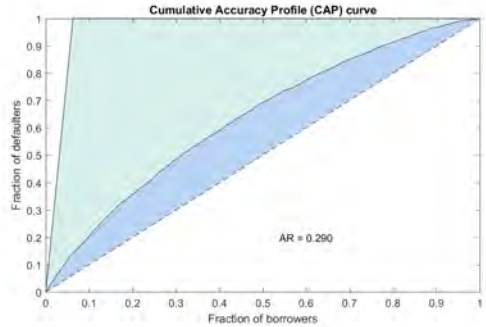
(b) ROC curve anno 2015



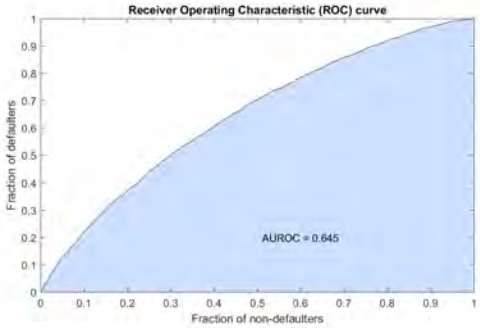
(c) CAP curve anno 2016



(d) ROC curve anno 2016



(e) CAP curve anno 2017



(f) ROC curve anno 2017

Figura 2.22: Analisi delle statistiche per il campione out of sample

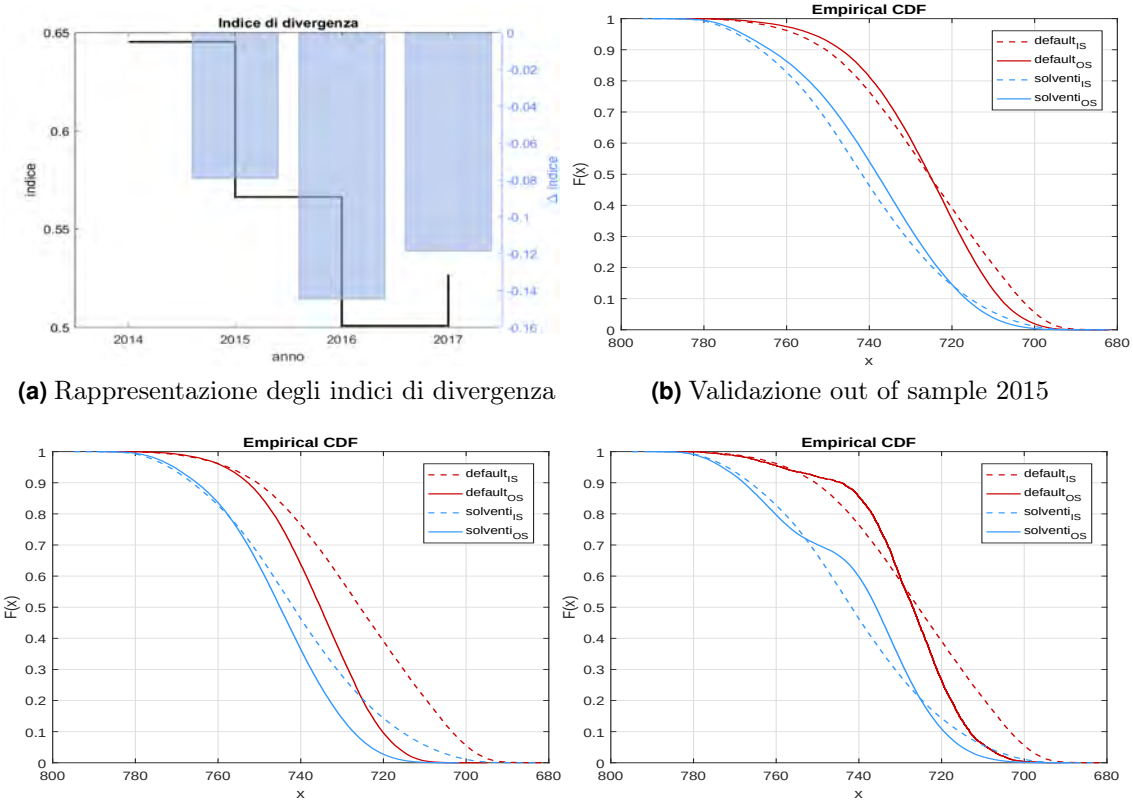
La tabella 2.5 a pagina 67 riporta anche la statistica KS. I valori assunti dalle statistiche restano in zona di validazione del modello e sono simili alle misure calcolate per il campione. Solo l'anno 2016 presenta un sensibile peggioramento. A questo punto è utile analizzare la divergenza tra le distribuzioni, riportata in tabella 2.6 a pagina 68 e rappresentate nella figura 2.23 a pagina 68. Le distribuzioni in esame divergono fino al 15% circa nell'anno 2016 rispetto alla distribuzione originaria del 2014. Ma comunque risulta coerente con quanto osservato nelle altre misure di divergenza e comunque il modello si può considerare valido. Questa divergenza nelle distribuzioni nel modello applicato al dataset *out of sample* trova parziale spiegazione nel campionamento delle controparti, in quando gli anni utilizzati per la validazione sono anni con moltissimi prestiti ancora in corso e questo sicuramente distorce le statistiche. Non a caso l'anno 2015 si avvicina molto di più rispetto ai risultati del campione 2014, infatti presenta il minore numero di prestiti ancora in vita. Ad ogni modo si è cercato di ottenere il miglior modello possibile con i dati a disposizione e per fini didattici si può considerare validato.

Tabella 2.6: Indice di divergenza tra in sample e out of sample

Misura	2014	2015	2016	2017
divergenza	0.645	0.566	0.501	0.527
scarto OS-IS		−7.900%	−14.450%	−11.860%
obs.	210662	189438	155350	54625

Per completezza si riportano anche i coefficienti stimati e il relativo p-value del campione *out of sample* . Come per gli anni *in sample*, è evidente come la regressione del 2014 sia molto simile alle variabili significative degli anni *out of sample*.

Per quanto riguarda il rating si riportano le relative scale con annesse distribuzioni di portafoglio (figura 2.24 nella pagina seguente), oltre alla tabella 2.9 a pagina 71 esplicativa.



(a) Rappresentazione degli indici di divergenza

(b) Validazione out of sample 2015

(c) Validazione out of sample 2016

(d) Validazione out of sample 2016

Figura 2.23: Cumulative distribution function e indice di divergenza

Tabella 2.7: P-value per variabile secondo anno per il campione out of sample

	2014		2015		2016		2017	
	Est.	pVal	Est.	pVal	Est.	pVal	Est.	pVal
(Int.)	1.50	0.00	1.16	0.00	1.23	0.00	2.71	0.00
loan_amnt	0.51	0.00	0.75	0.00	1.05	0.00	1.39	0.00
yr_earliest_cr_line	0.73	0.00	0.27	0.00	0.23	0.00	−1.52	0.02
acc_now_delinq					0.98	0.02		
pub_rec_bankruptcies			−1.92	0.00	−1.24	0.00		
pub_rec	0.63	0.01	1.65	0.00	1.37	0.00		
addr_state	1.00	0.00	1.01	0.00	1.04	0.00	1.09	0.02
annual_inc	0.98	0.00	0.54	0.00	0.75	0.00	0.41	0.01
revol_util	0.22	0.00	0.49	0.00	0.63	0.00		
delinq_2yrs	1.09	0.00	1.21	0.00	1.32	0.00	0.77	0.01
dti	0.44	0.00	0.50	0.00	0.49	0.00	0.62	0.00
emp_length	1.14	0.00	0.94	0.00	0.90	0.00	0.87	0.00
home_ownership			0.86	0.00	0.85	0.00	0.66	0.00
term	0.57	0.00	0.28	0.00	0.17	0.00	−1.98	0.00
purpose	0.50	0.00			0.45	0.00	0.78	0.00
inq_last_6mths	0.35	0.00	0.43	0.00	0.60	0.00	0.48	0.00
open_acc	0.86	0.00	0.92	0.00	0.82	0.00	4.25	0.00
perc_int_rate	0.71	0.00	0.78	0.00	0.75	0.00	0.96	0.00
obs.	210662		189438		155350		210662	

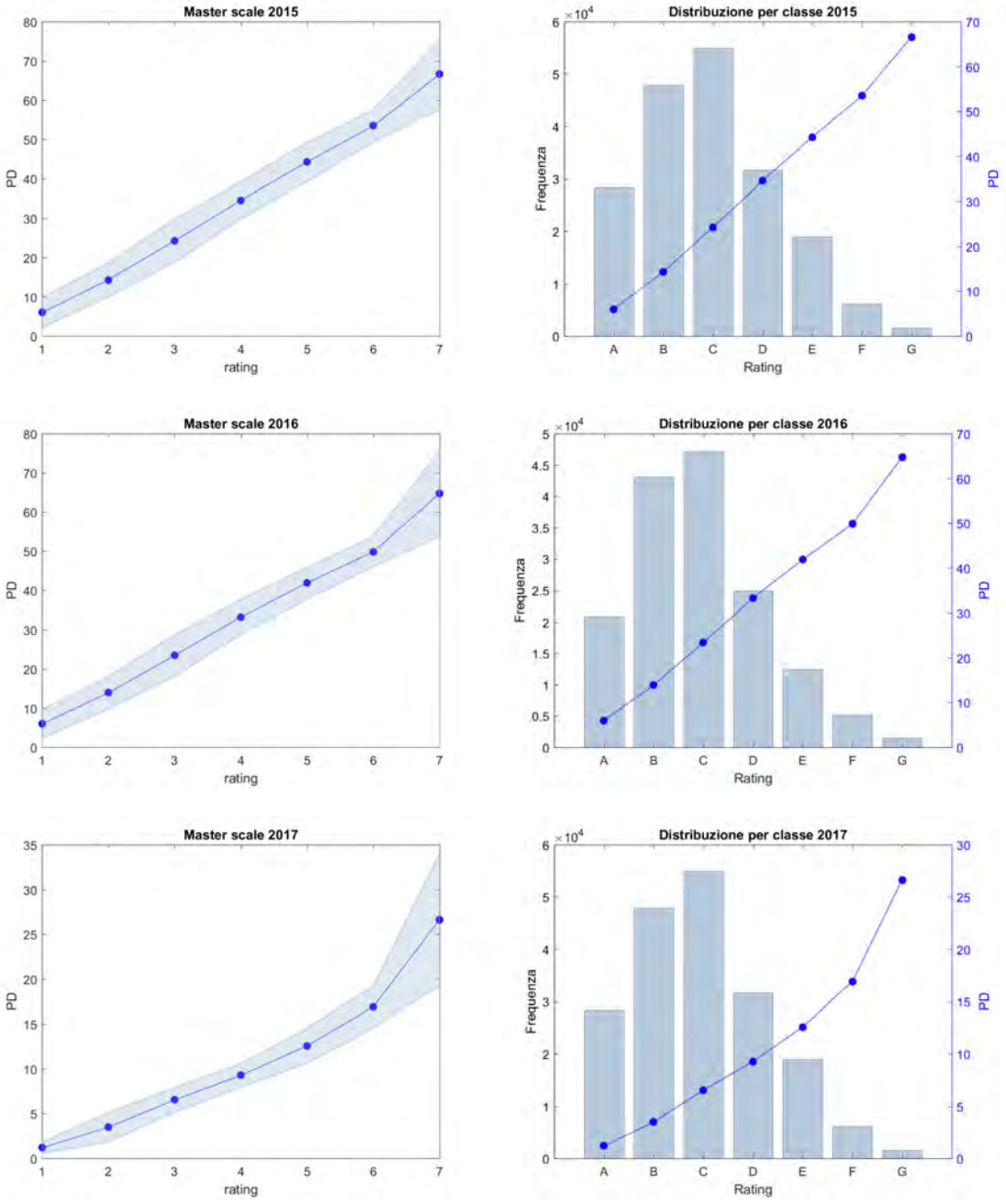


Figura 2.24: Sistema di rating out of sample

Tabella 2.8: Masterscale per il campione out of sample

	Anno 2015			Anno 2016			Anno 2017		
	Range (%)		PD (%)	Range (%)		PD (%)	Range (%)		PD (%)
A	0.00	- 9.94	5.99	0.00	- 9.66	5.98	0.00	- 1.81	1.17
B	9.95	- 18.67	14.30	9.67	- 18.10	13.89	1.82	- 5.14	3.48
C	18.68	- 29.84	24.26	18.11	- 28.77	23.44	5.15	- 7.92	6.53
D	29.85	- 39.23	34.53	28.78	- 37.75	33.26	7.93	- 10.62	9.27
E	39.24	- 49.25	44.24	37.76	- 45.88	41.82	10.63	- 14.54	12.58
F	49.26	- 57.65	53.45	45.89	- 53.75	49.82	14.55	- 19.26	16.90
G	57.66	- 75.55	66.60	53.76	- 75.74	64.75	19.27	- 34.04	26.64

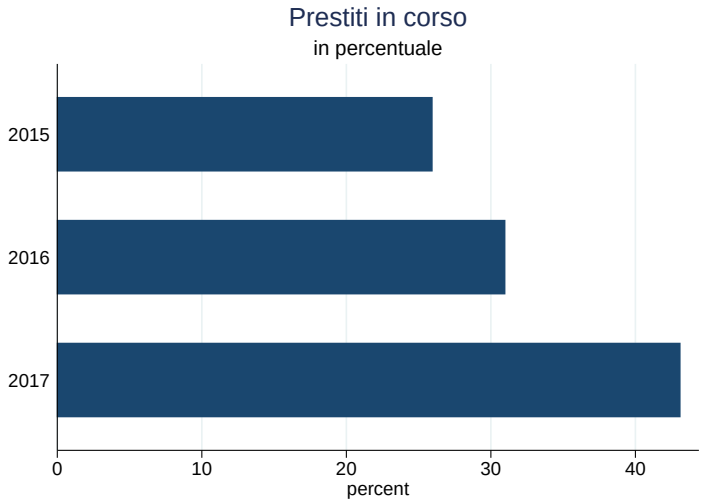


Figura 2.25: Percentuale dei prestiti non ancora terminati

Nonostante le regressioni dei tre anni utilizzati come *out of sample* siano molto simili alla regressione dell'anno 2014, sembra comunque esserci una piccola divergenza, diventa quindi molto importante il confronto delle stime ottenute con la frequenza relativa di default *actual*. Questa è calcolata come:

$$freq.\ relativa_{classe\ x} = \frac{nr\ default_{classe\ x}}{nr\ controparti_{classe\ x}} \tag{2.30}$$

i cui risultati per l'anno 2014 sono riportati in tabella 2.26b nella pagina seguente. Una rappresentazione grafica aiuta a comprendere come la stima delle probabilità di default non sia del tutto esatta per le tre classi di rating peggiori. Nella figura 2.26a nella pagina successiva si riporta la *masterscale* costruita per l'anno 2014, con le linee grigie discontinue si rappresentano i range per classe di rating, mentre la linea continua blu rappresenta la media della classe. Nello stesso grafico si riportano le frequenze relative per classe, rappresentate dai trattini rossi. Per esempio,

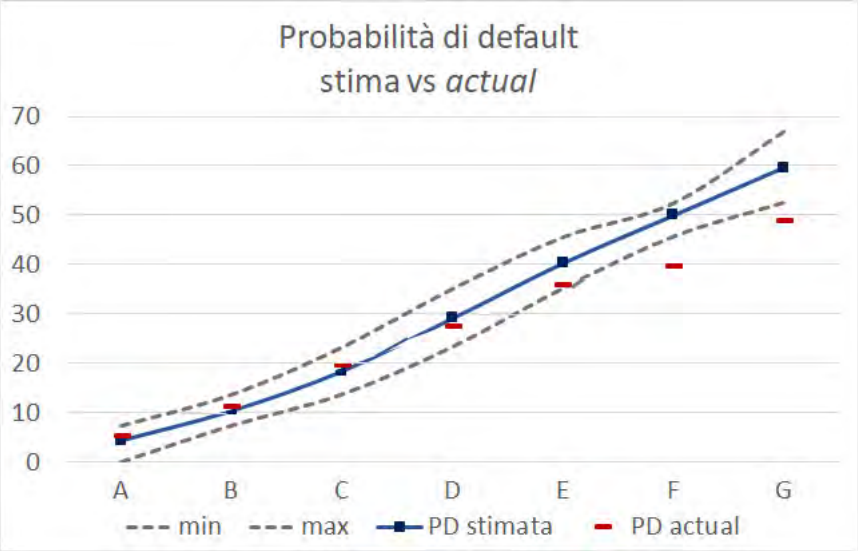
Tabella 2.9: Masterscale per il campione out of sample

	Anno 2015				Anno 2016				Anno 2017			
	Range (%)			PD (%)	Range (%)			PD (%)	Range (%)			PD (%)
A	0.00	-	9.94	5.99	0.00	-	9.66	5.98	0.00	-	1.81	1.17
B	9.95	-	18.67	14.30	9.67	-	18.10	13.89	1.82	-	5.14	3.48
C	18.68	-	29.84	24.26	18.11	-	28.77	23.44	5.15	-	7.92	6.53
D	29.85	-	39.23	34.53	28.78	-	37.75	33.26	7.93	-	10.62	9.27
E	39.24	-	49.25	44.24	37.76	-	45.88	41.82	10.63	-	14.54	12.58
F	49.26	-	57.65	53.45	45.89	-	53.75	49.82	14.55	-	19.26	16.90
G	57.66	-	75.55	66.60	53.76	-	75.74	64.75	19.27	-	34.04	26.64

nel 2014 le tre classi di rating peggiori, ovvero classe E, F e G, presentano una probabilità di default media maggiore rispetto alla frequenza relativa . Tutte le altre classi di rating presentano dei valori molto vicini alla media stimata. Questo porta a ritenere che il modello abbia capacità predittiva.

Nel complesso, il modello sembra avere la capacità di catturare le dinamiche dei default anche alla luce della mancanza dei dati più sensibili, sebbene ci sia una lieve divergenza in alcuni casi.

Calcolato un *panel* di probabilità di default per ogni classe di rating si passa a calcolare le altre componenti del merito di credito, ossia la *loss given default* e l'*exposure at default*. Calcolati tutti questi elementi si procederà a calcolare la media e la deviazione standard delle probabilità di default come richiesto dal modello di portafoglio che sarà utilizzato nel capitolo 4 a pagina 84, in cui subentreranno anche le altre due componenti succitate.



(a) Rappresentazione della masterscale e della frequenza dei default anno 2014

	Anno 2014				
	Range (%)			PD(%) stimate	PD(%) actual
A	0	-	7.34	4.49	5.39
B	7.35	-	13.66	10.5	11.26
C	13.67	-	23.26	18.5	19.56
D	23.27	-	35.08	29.17	27.63
E	35.09	-	45.56	40.32	36
F	45.57	-	52.43	50	39.83
G	52.44	-	66.93	59.67	48.89

(b) Confronto PD stimate con frequenze relative

Figura 2.26: Confronto probabilità di default stimate con frequenza relativa

Capitolo 3

Le altre componenti del rischio di credito

Loss Given Default e Exposure at default

Le misure di rischio moderne dipendono da tre parametri: probabilità di default (PD), esposizione al momento del default (EaD) e perdita avvenuto il default (LGD). La probabilità di default descrive la probabilità che il *lender* debba affrontare l'insolvenza della controparte, l'esposizione al momento del default stima l'ammontare esposto al verificarsi del default e indica, quindi, la massima perdita per l'affidamento. La LGD calcola la perdita che la banca deve sostenere al verificarsi del default di una controparte. A fine di determinare il rischio complessivo per ogni controparte e perseguire nell'analisi di portafoglio si dovrà dunque procedere alla stima dei due parametri residui. Prima di tutto si analizzeranno quali fattori incidono sui tassi di recupero e quali criteri dovrebbero essere seguiti per la loro stima, successivamente si procederà alla loro stima e al confronto con il campione *out of sample*. Stimati i RR, si proseguirà con delle considerazioni sull'*exposure at default* in quanto, sebbene sia un parametro molto importante è anche molto difficile da poter modellizzare e, avendo scelto un modello di portafoglio con base attuariale a causa dell'assenza di dati, non si necessita della stima di questa componente, che rimarrà analizzata solo a livello teorico.

3.1 Loss Given Default e Recovery rate

Il tasso LGD è la percentuale di perdita per un creditore se il debitore diventa insolvente. È dato dal complemento a uno del *recovery rate* - *RR* ed è compreso

tra 0 e 1, ebbene possano esserci delle eccezioni. Formalmente:

$$LGD = 1 - RR \tag{3.1}$$

La *loss given default* non è mai conosciuta in modo preciso finchè non avviene il default, soprattutto se non esiste un mercato secondario liquido. Al contrario, se esiste, si possono derivare le percentuali di *recovery rate* e *loss given default* sulla base dei prezzi di mercato post default. Ad ogni modo i dati sono certi e corretti solo al termine del processo di *recovery*.

3.1.1 Fattori che influenzano il *recovery rate*

Secondo Resti A. (2007) nel suo **book:sironi** la LGD viene da influenzata da quattro gruppi di fattori: le caratteristiche dell'esposizione, quelle della controparte, le caratteristiche della banca che gestisce il processo di recupero ed infine, da fattori esterni. Le caratteristiche dell'esposizione comprendono: la presenza di *collateral*, ovvero *asset* finanziari o reali a garanzia; il loro livello di liquidità e la priorità dell'esposizione, che può essere *senior* o subordinata alle altre esposizioni. Infine eventuali garanzie di terze parti. Le caratteristiche del debitore includono invece : l'industria o il settore in cui opera/lavora, il luogo geografico in cui opera e ha residenza, che influenza la procedura dell'eventuale fallimento, e la situazione finanziaria. L'istituto che gestisce il processo di recupero influenza i *recovery rates* dipendentemente dal grado di efficienza con cui segue quest'attività. Infine i fattori esterni, come lo stato dell'economia e il livello dei tassi di interesse influenzano il valore attuale delle *recoveries*, così come la probabilità di default della controparte. La tabella 3.1 riassume le variabili che hanno un impatto sul tasso di recupero e la componente della LGD che influenza maggiormente.

3.1.2 Stima dei *recovery rate* - RR

I tassi di recupero per esposizioni in default possono essere calcolati in quattro modi alternativi, come riportato anche in Luisa Izzi (2012): *workout* LGD, LGD di mercato, LGD implicita e LGD storica. Gli ultimi due modi sono considerati impliciti, in quanto non si basano sulle LGD già stimate su posizioni insolventi. Per di più l'approccio utilizzante LGD storiche è permesso solo per esposizioni retail. L'approccio che prevede l'utilizzo di LGD di mercato o il *workout* LGD, invece, sono tecniche che possono essere usate solo in circostanze limitate, ossia quando il mercato è profondo e liquido. La *market* LGD, infatti, utilizza i prezzi di mercato delle esposizioni defaultate come stima dei RR; nella *workout* LGD si utilizzano

Tabella 3.1: Fattori che influenzano i Recovery Rate

Gruppo	Fattori	Componenti influenzati
Caratteristiche dell'esposizione	*Collateral* *Seniority* Garanzie di soggetti terzi	Ammontare recuperato
Caratteristiche della controparte	Settore Nazione Situazione finanziaria	Capacità di trovare una controparte e prezzo di vendita Durata del processo Ammontare recuperato
Fattori interni della banca	Velocità ed efficienza procedura di recupero Vendite di NPL e uso di insediamenti extragiudiziali	Ammontare recuperato e durata del processo
Fattori macroeconomici, esterni	Stato dell'economia Tassi di interesse	Ammontare recuperato Valore attuale delle *recoveries*

dati e informazioni di mercato in quanto oggettivi e sempre aggiornati, ma sono disponibili solo per controparti *corporate*.

Normalmente la stima della *loss given default* è considerata una funzione delle garanzie e del merito di credito della controparte, si applica osservando una struttura in sezioni che suddivide le parti garantite da quelle non protette del prestito come in figura 3.1 nella pagina successiva, che dimostra come la *loss given default* sia definita da una parte coperta da garanzie (defintia *"secured"*) e una parte scoperta dei prestiti (*"unsecured"*), dove la $LGD_{secured}$ è calcolata come percentuale dell'esposizione che ci si aspetta di perdere dalla parte collateralizzata, la $LGD_{unsecured}$ è la percentuale di perdita attesa sulla parte del portafoglio non coperta da garanzie. Da queste considerazioni appare evidente come la stima della *loss given default* abbia tre principali componenti:

- I *recovery rate*, tassi di recupero della parte del prestito dopo azioni legali e recupero dalle garanzie a copertura del prestito;
- La percentuale del prestito coperta dalle garanzie;
- La percentuale del prestito non coperta che ci si aspetta di poter perdere.

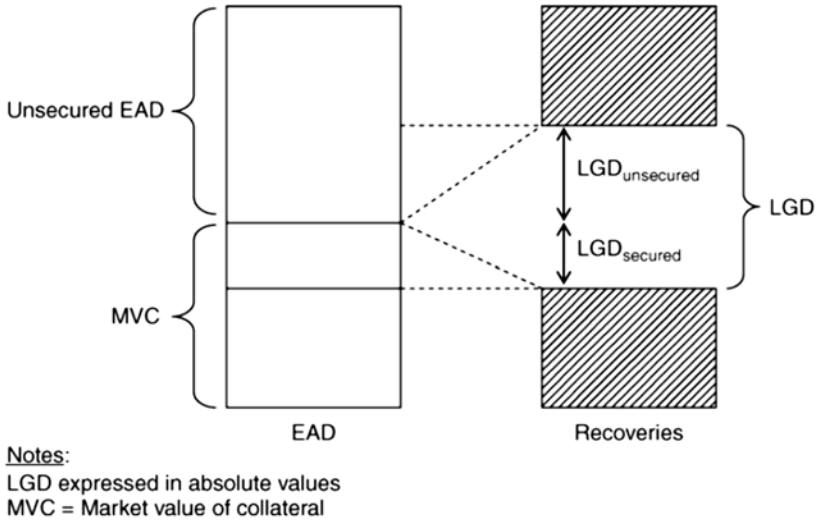


Figura 3.1: Schema della struttura per la stima della LGD

Da qui deriva che la stima della LGD storica di un dataset di posizioni defaultate, quindi basata sulle *recoveries*, parte dall'*exposure at default* recuperata dal processo di recupero e dalle eventuali garanzie, dal *cost of carry*, costo associato all'intervallo temporale che intercorre tra il momento del default e il momento del recupero, nonché dai costi amministrativi, interni ed esterni sostenuti per intraprendere il processo di recupero.

Essendo già noto come i prestiti P2P in esame non siano protetti da garanzie, si concentrerà l'analisi dei *recovery rate* e della LGD per parti *unsecured*.

LGD di prestiti *unsecured* Per i prestiti senza annesse garanzie ci si aspetta che qualche forma di recupero venga messa in atto al fine di recuperare quanto più possibile dell'esposizione e di contenere le perdite. Il tasso di recupero è quindi mera funzione del merito di credito della controparte. Per calcolare la LGD di prestiti *unsecured* si deve conoscere: l'ammontare del prestito non coperto, l'ammontare recuperato del capitale dal pagamento delle rate (*total principal*), il tempo impiegato per il recupero e i costi sostenuti per il processo. Il primo passo da fare è calcolare il valore attuale dell'ammontare di capitale recuperato (NPV - net present value):

$$NPV(R_{unsecured}) = \sum_{i=1}^{n} \frac{recovery_i \; asset \; non \; coperti}{(1 + tasso)^t} \tag{3.2}$$

dove t è il numero di anni trascorsi dal default alla fine del processo di recupero. Il tasso di recupero realizzato, a questo punto, è funzione del valore attuale

dell'ammontare recuperato e dei costi in relazione dell'ammontare dell'esposizione:

$$RR = \frac{NPV(R_{unsecured}) - NPV(costi)}{EaD} \tag{3.3}$$

Stimati tutti i RR per ogni controparte, si deve determinare il tasso di recupero medio per ogni classe di rating, perché, come già detto, la LGD per prestiti non garantiti dipende solo dal merito di credito della controparte:

$$RR_{unsecured\ classe\ m} = \frac{1}{n} * \sum_{i=1}^{n} RR_i \tag{3.4}$$

$$LGD_{unsecured\ classe\ m} = 1 - RR_{unsecured\ classe\ m} \tag{3.5}$$

dove:

m=lettera assegnata ad ogni classe di rating

n=osservazioni all'interno di ogni classe di rating

Nel settembre del 2004, il Comitato di Basilea e il suo gruppo di implementazione, impostarono un lavoro per condividere le idee e chiarire l'approccio alla stima della LGD, da cui emerse, tra le altre cose, che la disponibilità dei dati limitati rappresentava un'importante sfida da affrontare. Questo problema permane tutt'ora, infatti anche nella situazione in esame, la mancanza dei tempi impiegati per il recupero, l'assenza dei costi sostenuti e l'estrema difficoltà nel selezionare un tasso di sconto adeguato, rende di fatto impossibile calcolare i RR e la LGD secondo quanto sopra illustrato. Ci si accontenterà, a fini didattici, di utilizzare i dati messi a disposizione da Lending Club per effettuare una stima più o meno approssimativa. Per stimare i *recovery rate* si utilizzeranno due variabili presenti nel dataset:

- total_rec_prncp (*principal received to date*), che rappresenta il capitale recuperato dai pagamenti della controparte
- recoveries, ammontare recuperato dopo il default, qualunque sia il mezzo con cui è stato riscosso.

Purtroppo l'assenza di tempistiche e costi sostenuti non permette di calcolare i tassi di recupero nel modo sopra illustrato. Vista la forte influenza della variabile tempo sulla stima dei *recovery rate*, si è deciso di procedere suddividendo la stima secondo scadenza del prestito. Si avrà quindi una stima in sample per i *recovery*

rate di prestiti a 36 mesi con orizzonte gennaio 2007 - dicembre 2014, mentre il calcolo per i prestiti a 60 mesi utilizzerà una finestra temporale che si estende da gennaio 2010 a dicembre 2012, anni in cui sono emessi e terminati tutti i prestiti a cinque anni. Abbiamo già detto che i tassi di recupero per prestiti non coperti da garanzia sono fortemente dipendenti dal merito di credito della controparte, quindi si stimeranno i *recovery rate* come media di tutti i tassi di recupero delle controparti appartenenti a quella classe di rating, secondo la formula:

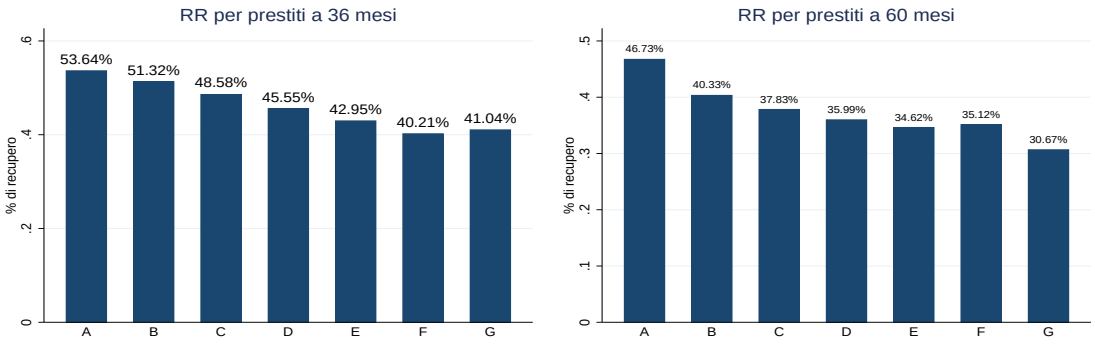
$$recoveryrate_i = \frac{total\_rec\_prncp_i + recoveries_i}{loan\_amnt_i} \tag{3.6}$$

$$RR_{grado\ x} = \frac{\sum_{i=1}^{n} recovery\ rate_{i,\ grado\ x}}{nr\ controparti_{grado\ x}} \tag{3.7}$$

dove: total_rec_prncp è il capitale restituito dal debitore, recoveries è l'ammontare recuperato dopo il default e loan_amnt è la somma richiesta. I risultati sono riportati in tabella 3.2, da cui si evince come prestiti più brevi abbiano tassi di recupero superiori di dieci punti percentuali rispetto a prestiti con pagamenti più dilazionati nel tempo (figura 3.2). Ad ogni modo, per i prestiti P2P di Lending Club si nota un tasso di recupero molto alto sebbene siano prestiti *unsecured*, compreso tra il 53% e il 41% circa per prestiti a "breve", mentre si stima un intervallo dal 43% al 30% circa per i prestiti più a lungo termine. Alcune società di consulenza specializzate riportano un $RR_{unsecured}$ tra il 6% e l'8%,molto più basso dei prodotti in esame. questa caratteristica permette di mitigare il rischio delle probabilità di default elevate .Documentazione consultata al Convegno *"Securitisation, management and valuation of NPLs"*, EY, Aprile 2018 Ca' Foscari.

Tabella 3.2: Stima dei Recovery Rate

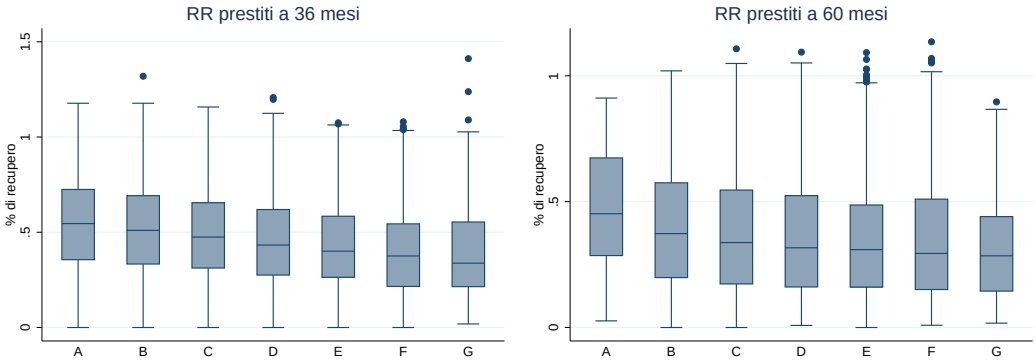
	36 mesi (%)	60 mesi (%)
A	53.64	43.76
B	51.32	40.33
C	48.58	37.83
D	45.55	35.99
E	42.95	34.62
F	40.21	35.12
G	41.04	30.67



(a) Differenza dei tassi di recupero per termine **(b)** Recovery rate medio per classe di rating

Figura 3.2: Rappresentazione dei Recovery Rate

Per completare l'analisi dei *recovery rate* si riporta, in figura 3.3, la rappresentazione grafica dei percentili della distribuzione dei tassi di recupero al fine di verificare che sia rispettata la relazione inversa con la probabilità di default. Come da aspettativa, all'aumentare della classe di rating, conseguentemente anche della PD, si osserva un netto peggioramento dei tassi di recupero.



(a) Prestiti a 36 mesi **(b)** Prestiti a 60 mesi

Figura 3.3: Distribuzione per classe di rating dei RR

Validazione Come per le altre componenti del rischio di credito è richiesta la validazione. Perseguendo la logica seguita per identificare i dataset *in sample*, si confrontano le stime per i tassi di recupero a 36 mesi con le stime dei tassi di recupero dell'anno 2015, mentre per i *recovery rate* dei prestiti a cinque anni si confronteranno le stime ottenute con i tassi dell'anno 2013, in modo da confrontare stime calcolate su portafogli di prestiti prevalentemente già terminati.

Come si evince dalla figura 3.4, i tassi di recupero del campione *out of sample* rispetto alle stime *in sample* sono coerenti. Per i prestiti a cinque anni le stime

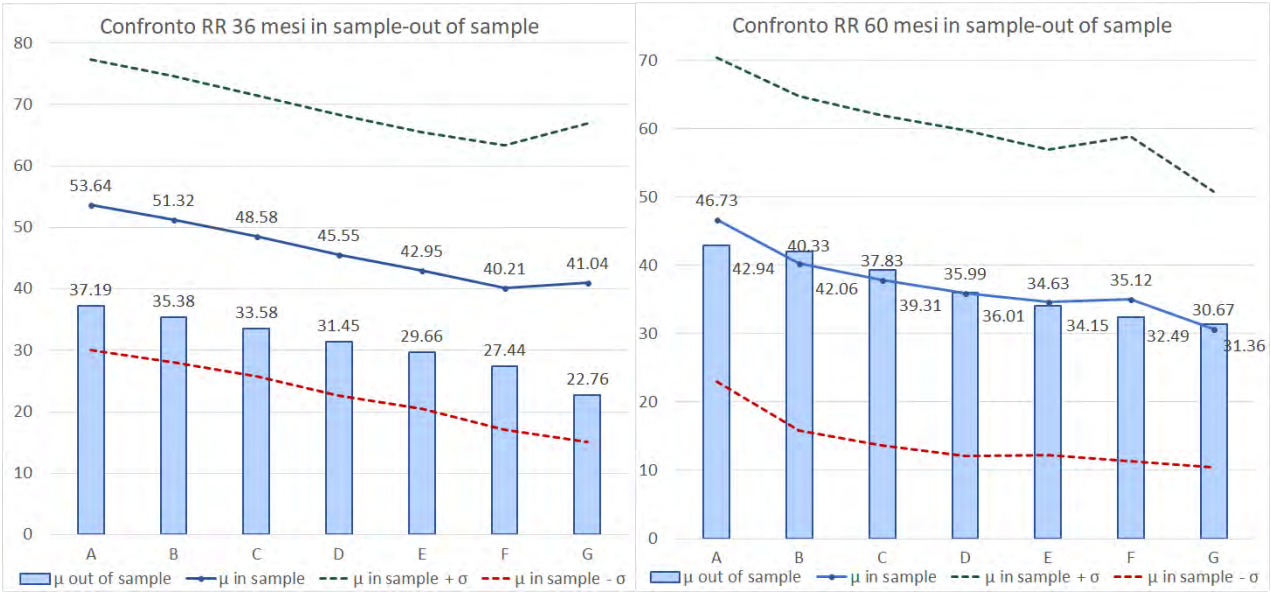


Figura 3.4: Confronto tra RR in sample e RR out of sample

out of sample risultano molto vicine alle medie stimate nel dataset *in sample*; i prestiti a tre anni, invece, presentano tassi di recupero inferiori alle stime effettuate con il campione *in sample* ma comunque sono compresi nel range della deviazione standard.

3.2 Exposure at default

Un altro parametro da calcolare è rappresentato dall'*exposure at default - EaD*, ovvero l'ammontare atteso del prestito al momento del default. Studi empirici[1] hanno dimostrato che, quando un debitore entra in difficoltà finanziaria , utilizzerà in modo più massiccio i suoi affidamenti; quindi per non sottostimare le perdite è importante basare i calcoli sull'utilizzo atteso dell'affidamento al momento del default piuttosto che sull'utilizzo medio atteso. La metodologia per il calcolo dell'EaD si basa sulla suddivisione in due categorie. La prima categoria comprende esposizioni **certe**, come ad esempio mutui e prestiti, basati su un piano di ammortamento; la seconda riguarda le esposizioni **incerte**, di cui non si ha un utilizzo prestabilito, ma dipende solo dal comportamento della controparte.

Per le esposizioni certe si conosce lo schema dei pagamenti in ogni momento (mentre non si conosce il momento del default); in un approccio *"snapshot"*, illustrato in Luisa Izzi (2012), la stima dell'esposizione al momento del default equivale

[1] Luisa Izzi 2012

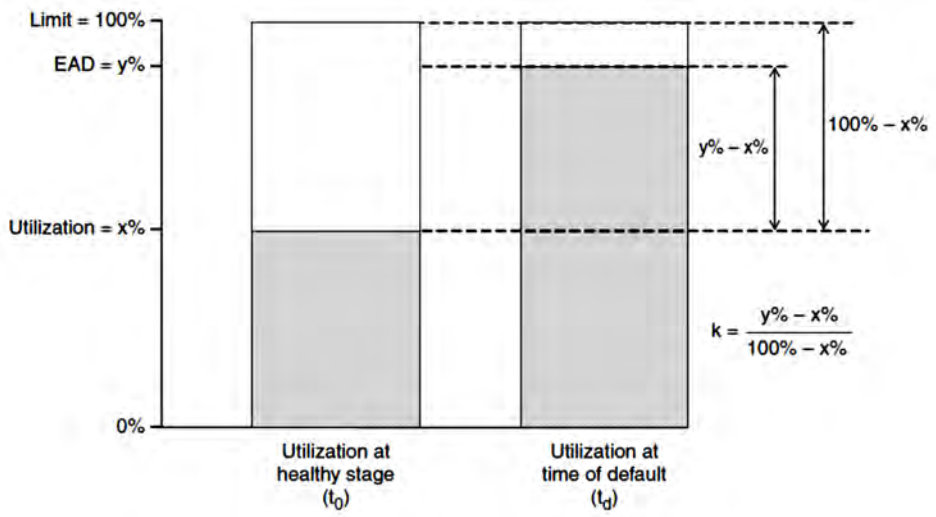


Figura 3.5: Rappresentazione del caso generale del fattore k

alla somma erogata:

$$EaD_{certain} = exposure_0 \tag{3.8}$$

Le difficoltà iniziano nel determinare le esposizioni incerte, in quanto i futuri prelievi non conosciuti incidono sulla stima dell'EaD in modo non facilmente prevedibile. Per questo tipo di esposizioni, l'*exposure at default* è dato dalla somma dell'utilizzo corrente più un ulteriore utilizzo atteso, richiesto dalla controparte in difficoltà finanziaria:

$$EaD_{uncertain} = U + k * (L - U) \tag{3.9}$$

dove:

- U è l'utilizzo corrente della linea di credito;
- L è il limite massimo concesso;
- k è una misura di ulteriore utilizzo dell'affidamento, calcolata come proporzione della parte non ancora utilizzata, definita anche fattore di conversione del credito - CCF:

$$k = \frac{EaD - U}{L - U} \tag{3.10}$$

Per definizione, k, può essere applicabile solo quando U e L possono cambiare nel tempo, e si considera compreso tra 0 e 1, sebbene esistano delle eccezioni, una rappresentazione grafica di tale componente è rappresentato in 3.5[2]. Poiché k è l'unico fattore sconosciuto dell'EaD, stimare l'esposizione al momento del default significa, sostanzialmente, stimare il CCF.

[2]Luisa Izzi 2012

Stima dell'EaD Per calcolare l'esposizione al momento del default è, quindi, necessario conoscere l'utilizzo corrente dell'affidamento, il limite massimo del fido e il fattore di conversione k, quando applicabile. Le prime due variabili sono facilmente ottenibili, in quanto dati oggettivi, mentre il CCF deve essere stimato attraverso l'esperienza passata, di cui si utilizzerà l'utilizzo dell'esposizione al default storico (U_i), l'utilizzo ($\overline{U_i}$) e il limite prima del default ($\overline{L_i}$) da cui, partendo dall'equazione 3.9, si arriva a:

$$k_i = \frac{U_i\% - \overline{U_i}\%}{1 - \overline{U_i}\%} \tag{3.11}$$

Alla base della modello CCF c'è la stima del calcolo $\overline{U_i}\%$ che deve essere basato su l'utilizzo medio degli ultimi anni degli affidamenti di controparti in difficoltà e il limite medio di utilizzo oltre il quale gli affidamenti si considerano sofferenti. Questo significa che $\overline{U_i}\%$ riflette l'esperienza passata sulle sofferenze. Come conseguenza, $U_i\%$ può essere calcolato dividendo l'utilizzo al momento del default sul limite medio di utilizzo calcolato oltre il quale si considera sofferenza ($\overline{L_i}\%$). La media del fattore k per un gruppo di n controparti è data da:

$$\overline{k} = E(k) = \frac{1}{n} * \sum_{i=1}^{n} k_i \tag{3.12}$$

In altre parole il fattore k medio ($\overline{k}$) è la media dei fattori k di ogni controparte nel gruppo. Risulta evidente che k è influenzato da diversi fattori delle caratteristiche della controparte. Idealmente, queste componenti possono essere calcolate usando dati delle controparti di diversi mesi e in diversi momenti del mese. Una volta che $\overline{k}$ è calcolato dovrebbe rimanere fisso nel tempo finchè non verrà aggiornato, quindi per una generica esposizione, se conosciamo il fattore medio $\overline{k}$ per ogni gruppo di controparti, il limite massimo dell'affidamento L_i e l'utilizzo U_i, è possibile determinare la corrispondente *exposure at default* come:

$$EaD_i = U_i + \overline{k} * (L_i - U_i) \tag{3.13}$$

Ad ogni modo, a causa della mancanza dei dati necessari al calcolo, si è deciso di utilizzare un modello di portafoglio con base attuariale che non necessita del calcolo di questa componente, ma deriverà le perdite dalla manipolazione delle LGD delle singole controparti, per cui l'esposizione al momento del default verrà considerata come l'ammontare ricevuto dal debitore meno l'ammontare recuperabile nel caso di insolvenza.

Capitolo 4

Creditrisk +

Negli ultimi quindici anni, la misurazione e la gestione del rischio di credito sono diventate parte rilevante nella gestione del rischio nelle le istituzioni finanziarie. Si sono verificate grandissime innovazioni nell'ambito della modellizzazione del merito di credito. Ad oggi, sono stati sviluppati tre approcci principali: il modello "*Merton style*", un'approccio al problema con una logica puramente econometrica oppure la modellizzazione attraverso un approccio attuariale. Sebbene ognuno di questi approcci abbia prodotto moltissimi modelli, tutti hanno l'obiettivo di determinare la distribuzione delle probabilità delle perdite di un portafoglio di strumenti di credito. La necessità di stimare la distribuzione delle perdite è un profilo essenziale nel *risk management* nel campo del *credit risk*, in quanto permette di calcolare il *Value at risk* - *VaR* - e il capitale economico necessario a fronteggiare tali perdite. In questo capitolo si focalizzerà l'attenzione su uno dei modelli più conosciuti: CreditRisk+. Dagli anni '90 la Credit Suisse First Boston (CSFB) sviluppa e distribuisce metodi di gestione del rischio. Nel 1993 ha lanciato, in parallelo, un progetto per modernizzare il suo *risk management* e, grazie ai suoi esperti, lancia CreditRisk+ nel 1996, modello per l'analisi del rischio di credito basato su logiche attuariali, che è diventato in poco tempo uno dei modelli più importanti. La popolarità è dovuta grazie a diverse motivazioni, tra cui i pochi dati necessari alla sua implementazione, coincidenti con i dati richiesti anche dai modelli interni di Basilea II; la capacità di determinare una distribuzione di perdita in modo analitico ed infine, la considerazione dei più importanti rischi del credito come la concentrazione. In questo lavoro si è riprodotto il modello originale attraverso l'utilizzo di MatLab, considerando due tipologie di applicazione, una previsionale sui prestiti ancora in corso, una a consuntivo sui prestiti conclusi per verificare la capacità predittiva del modello.

4.1 Il modello CreditRisk+ e le sue componenti

Un approccio moderno alla gestione del rischio di credito dovrebbe affrontare tutti gli aspetti sopra citati, dalla modellizzazione quantitativa allo sviluppo di pratiche per la gestione (figura 4.1 nella pagina successiva)[1]. CreditRisk+ riflette i requisiti di una gestione moderna e presenta tre principali caratteristiche. Il principale tratto distintivo è l'utilizzo di un approccio di portafoglio con tecniche analitiche largamente applicate nel mondo delle assicurazioni; in secondo luogo, questo modello è anche una metodologia per calcolare il capitale economico adeguato a coprire il rischio di credito; infine permette di fare previsione attraverso un mezzo per misurare la concentrazione e la diversificazione al fine di supportare la gestione del portafoglio creditizio. Questo modello si fonda su un approccio di portafoglio che considera informazioni legate alla dimensione e alla *duration* di un'esposizione, al merito creditizio e al rischio sistematico di una controparte. CreditRisk+ utilizza tecniche statistiche per modellizzare il rischio di default senza fare assunzioni sulle cause dell'insolvenza. Va sottolineato che questo modello non utilizza la correlazione tra i default come dato di input, in quanto si assumono le probabilità di default come stocastiche, con media pari alla PD della classe di rating . La PD viene considerata come variabile casuale continua che incorpora la volatilità dei tassi di default, al fine di catturarne l'incertezza. Spesso fattori macroeconomici, come il ciclo economico, possono incidere sulla correlazione tra i default sebbene non ci sia nessuna relazione causale tra le controparti. Questi effetti sono incorporati nel modello attraverso la considerazione della volatilità delle probabilità di default e l'analisi settoriale, piuttosto che considerare la correlazione un input esplicito. Questa impostazione permette a CreditRisk+ di catturare le caratteristiche della frequenza dei default e di calcolare una *loss distribution* per un portafoglio di controparti creditizie. Il risultato ottenuto può essere utilizzato per determinare il livello di capitale economico necessario a coprire le perdite inattese del rischio di credito insito nel portafoglio in esame.

4.2 Il modello

Nella rappresentazione della realtà attraverso modelli statistici ci sono tre tipi di rischio operativo da considerare:

1. **Rischio di processo**, in quanto le osservazioni reali sono soggette a variazioni anche quando il modello utilizzato per le stime è appropriato. Gene-

[1] Tabella riportata da *CreditRisk+: A credit risk management framework*

Figura 4.1: Tabella con componenti del modello CR+

CreditRisk+			
Misure per il Rischio di Credito		Capitale Economico	Applicazioni
Esposizione	Tasso di Default	Default	Approvvigionamento
Tassi di recupero	Volatilità PD	Distribuzione di perdita	Limiti
CreditRisk+ Model		Analisi di scenario	Gestione di portafoglio

ralmente questa incertezza nel risultato viene affrontata esprimendo la stima effettuata in termini di un appropriato livello di confidenza.

2. **Incertezza dei parametri**. Questo problema nasce dalla difficoltà di stimare i parametri utilizzati nel modello. L'unica informazione che può essere ottenuta sul processo sottostante al modello è estrapolata osservando i risultati passati. È possibile poi, valutare l'incertezza dei parametri attraverso analisi di sensitività sui parametri stessi.

3. **Termine d'errore**, questo parametro nasce perché il modello proposto potrebbe non riflettere accuratamente il processo reale. La divergenza viene catturata da una variabile "errore (ε)".

Questi tre rischi vengono affrontati in CreditRisk+ in diversi modi. In primo luogo non sono effettuate assunzioni sulle cause del default. Questo comportamento è presente principalmente nella gestione del rischio di mercato, dove non vengono fatte ipotesi sulla causa del movimento dei prezzi, è stato così ripreso nel modello in esame. Questo permette non solo di ridurre il termine di errore, ma conduce all'implementazione di un modello sviluppabile solo analiticamente, togliendone la soggettività. In secondo luogo i dati di input richiesti da Creditrisk+ sono stati mantenuti al minimo possibile, permettendo di minimizzare l'errore derivante dalla stima degli stessi. In questo settore i dati sono difficili da ottenere e, comunque, possono subire ampi scostamenti da un anno all'altro. Infine l'incertezza residua sui parametri può essere affrontata con un'analisi di scenario, grazie alla quale si possono "stressare" i dati e quantificare gli effetti sulle perdite.

Il rischio di credito ha origine dalla variazione del merito di credito di una controparte e si identifica in due tipologie: (i) Rischio di spread, in portafogli in cui lo spread creditizio è negoziato sul mercato e la cui variazione incide nel valore del portafoglio; (ii) Rischio di default o che la controparte risulti insolvente, tipico di tutte le esposizioni creditizie. Se si verifica lo stato di default si incorre in una

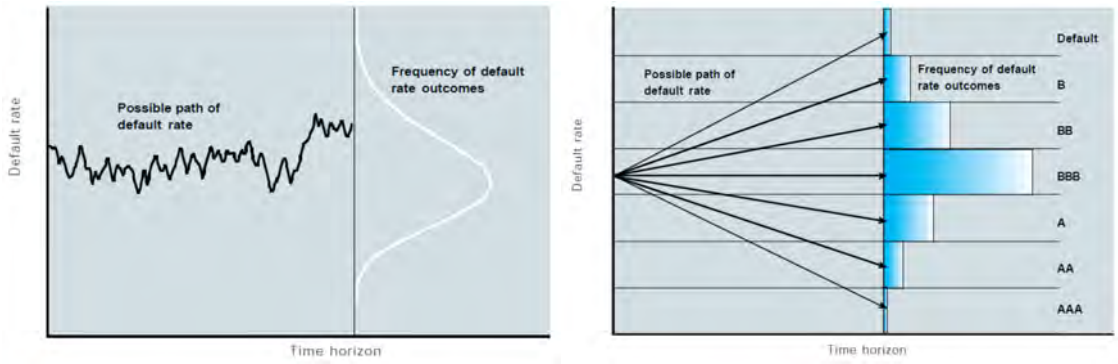
perdita immediata legata all'esposizione, che può essere totale o parziale a seconda della presenza o meno di garanzie a copertura o dei tassi di recupero (*recovery rate*). Diversa è la situazione di una variazione del merito creditizio in uno stato di solvenza, che si può tradurre sia in una perdita immediata diminuendo il valore dell'esposizione, oppure in una perdita futura al momento del default. Nella situazione in esame non si analizza con prestiti soggetti a spread, ma prestiti mantenuti in portafoglio fino a *maturity*. Quindi nel seguito del lavoro ci si concentrerà solo sul rischio di default.

Rischio di default Il rischio di default rappresenta il rischio che una controparte sia incapace di adempiere ai suoi obblighi finanziari. Nell'insolvenza di una controparte, un creditore incorre in una perdita equivalente all'ammontare posseduto dal debitore al momento del default meno un ammontare recuperato da azioni legali, come la liquidazione o la ristrutturazione del debito. La principale difficoltà nel passare dalla stima del rischio di credito a livello di singola esposizione, al rischio a livello aggregato, è insita nella necessità di stimare la probabilità di default congiuntamente per tutte le controparti e le relazioni intercorrenti tra esse.

Tassi di default Il processo del default può essere rappresentato in due modi: come variabile continua o come variabile discreta. Quando i default sono trattati come variabile continua si descrive la frequenza dei default in un lasso temporale predefinito come distribuzione, la quale può essere specificata a partire dalla probabilità di default e dalla sua volatilità. I dati richiesti in questo caso sono analoghi a quelli richiesti per prezzare le opzioni, situazione rappresentata in figura 4.2a nella pagina seguente[2] . Mentre l'assunzione dei tassi di default come variabile discreta è una semplificazione della situazione appena descritta e necessita di ulteriori informazioni per ottenere un modello predittivo; deriva per esempio dalla mappatura dei tassi di default sul rating. Nella fattispecie si necessita di una matrice di transizione, la quale fornisce la probabilità di mantenere lo stesso livello di rating più la probabilità di muoversi verso altre classi di rating, come rappresentato in figura 4.2b nella pagina successiva[3].

I due approcci , quello discreto con la matrice di transizione o viceversa, l'approccio in continuo, sono rappresentazioni differenti del comportamento dei tassi di default (tabella 4.2c nella pagina seguente) che restituiscono entrambi una *default distribution*. CreditRisk+ considera i tassi di default come variabile random

[2] *CreditRisk+: A credit risk management framework*
[3] Le figure 4.2 sono tratte da *CreditRisk+: A credit risk management framework*



(a) Variabile continua

(b) Variabile discreta

Trattamento variabili	Dati richiesti
Variabile Continua	• Probabilità di default • Volatilità delle PD
Variabile Discreta	• Rating • Matrice di transizione

(c) Riassunto dei differenti processi secondo probabilità di default

Figura 4.2: Rappresentazione del default

continua e incorpora la volatilità delle probabilità di default al fine di catturare l'incertezza della frequenza.

Misurazione La gestione del rischio di credito viene effettuata con l'ausilio di diverse misure, tra cui:

- **Distribuzione delle perdite (*loss distribution*)**: a partire dal portafoglio in esame si cerca di ottenere una distribuzione di perdita per ottenere il capitale perso con un certo intervallo di confidenza.
- **Calcolo dei risultati estremi**. Il *risk manager* è interessato a identificare risultati estremi o catastrofici, difficilmente calcolabili con modelli statistici, ma simulabili attraverso un'analisi di scenario.

In prima istanza, il rischio di credito può essere gestito attraverso la diversificazione in quanto il numero delle esposizioni all'interno del portafoglio è solitamente elevato. Nella pratica questo viene tradotto nel controllo del limite di capitale erogato, della *maturity* e diversificando geograficamente o a livello settoriale i prestiti concessi. In questo modo si possono creare portafogli ben diversificati controllando i principali quattro fattori che influenzano il rischio di credito di un portafoglio, ossia l' ammontare dell'esposizione, la durata del prestito, la probabilità di default della controparte e il rischio sistemico e di concentrazione.

Orizzonte temporale Una decisione cruciale per l'implementazione del modello è la scelta dell'orizzonte temporale. CreditRisk+ non impone un orizzonte temporale predefinito, ma suggerisce due modalità di analisi. Si può scegliere un orizzonte temporale costante, come un anno, che permette di considerare tutte le esposizioni alla stessa data futura. Spesso si utilizza l'anno come riferimento perché le azioni di copertura devono essere eseguite annualmente, così come il controllo del capitale economico, oppure si può mantenere un orizzonte*hold-to-maturity*. In questo modo si riesce a riconoscere la struttura dei tassi di default dell'intera vita delle esposizioni. Questa metodologia permette di comparare esposizioni di diverse durate e diversi meriti di credito.

4.2.1 Dati di input

Qualsiasi modello di gestione del rischio di credito dipende completamente dalla qualità dei dati che si utilizzano. Una scarsa quantità può inficiare l'accuratezza delle misure effettuate e di conseguenza, sviare le decisioni da prendere sulla base

stime ottenute.
I dati necessari per l'implementazione di CreditRisk+ sono:

- L'esposizione del creditore,
- La probabilità di default della controparte affidata,
- La deviazione standard della PD,
- I *Recovery Rates*.

L'esposizione Tutte le esposizioni di una controparte dovrebbero essere aggregate. Il modello è in grado di trattare qualsiasi tipo di esposizione creditizia, come prestiti, mutui, lettere di credito ed esposizioni in derivati. Per molte di queste, però, è necessario fare un'assunzione sul livello dell'esposizione al momento del default.

Esempio 3. Esempio di esposizione creditizia
Per esempio, una lettera di credito verrà utilizzata prima del default, quindi l'*exposure at default* dovrebbe essere considerata pari al valore nominale.

In un approccio pluri-annuale si dovrà inoltre, modificare accuratamente la variazione dei livelli di esposizione. Nel caso in esame si costruiranno tre distribuzioni: una sui prestiti correnti e due per la validazione del modello suddividendo i prestiti a tre anni da quelli a cinque anni. A causa di limiti computazionali, in quanto le iterazioni nel modello per il calcolo della distribuzione sono moltissime, si considerano sotto-campioni composti da 30 000 controparti estratte casualmente mantenendo la stessa composizione del portafoglio originale in termini di *rating mix*, al fine di poter ragionare per approssimazione secondo termini percentuali. Verranno suddivisi i prestiti secondo scadenza perché, per effettuare una validazione confrontando la perdita stimata dal modello con la perdita effettivamente sostenuta, servono prestiti terminati, in modo da sapere per ogni controparte se è risultata solvente o no. Questo comporta un taglio temporale all'anno 2012 per i prestiti a cinque anni e un taglio al 2014 per i prestiti a tre anni.

Probabilità di default La probabilità di default rappresenta la probabilità che un debitore risulti insolvente. Il rating è l'opinione sulle capacità finanziarie di un debitore ad assolvere le sue obbligazioni, quindi ogni controparte avrà una PD associata al suo merito di credito. Una valutazione della natura dell'obbligazione, inclusa la *seniority*, dovrebbe essere effettuata durante l'analisi dei *recovery rate*.

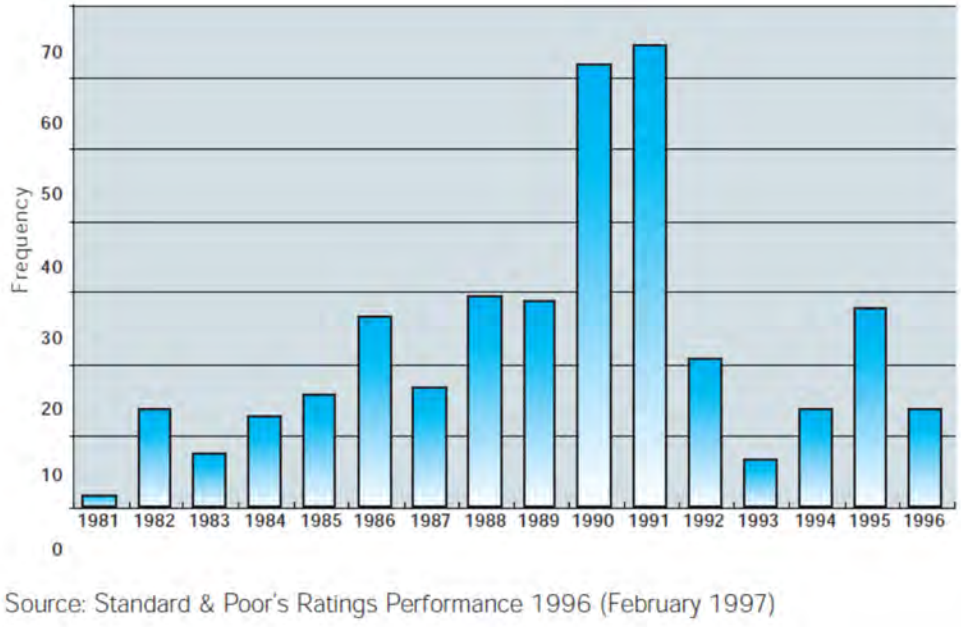


Figura 4.3: Livelli di default

Va sottolineato che le probabilità di default annuali possono subire significative variazioni da un anno all'altro, come dimostrato dalla figura 4.3 tratta da Boston, p. 12. Durante i periodi di recessione i numeri di default possono essere di molto superiori a quelli osservati in altre fasi del ciclo economico, questo spiega perché nel modello venga inserita anche la volatilità delle PD come dato di input. Nel portafoglio di prestiti current si utilizzeranno le probabilità di default come stimate nel secondo capitolo, per la validazione del modello, invece, sono state calcolate, sempre con la regressione logistica, le probabilità di default medie e la loro deviazione standard al taglio dell'orizzonte temporale.

Volatilità dei tassi di default Come già citato, le probabilità di default variano negli anni dalla loro media. La divergenza può essere descritta attraverso la volatilità, rappresentata dalla deviazione standard - σ.Si inseriscono le probabilità di default e la loro deviazione standard utilizzate per i portafogli in tabella 4.1 nella pagina seguente.

Recovery Rate Nel default della controparte, un creditore incorre in una perdita pari all'ammontare ancora dovuto dal debitore meno un ammontare recuperato attraverso azioni legali, come la liquidazione. I tassi di recupero stimati dovrebbero tenere in considerazione il livello di *seniority* del prestito ed eventuali garanzie a copertura. I RR verranno sottratti al capitale erogato, al fine di ottenere un portafoglio di solo capitale a rischio.

Tabella 4.1: PD e SD di input, stimate con la regressione del Capitolo 2

	Cinque anni - 2012		Tre anni - 2014		Current	
Rating	Media(%)	SD(%)	Media (%)	SD(%)	Media(%)	SD(%)
A	12.47	1.25	4.36	0.76	4.63	0.60
B	33.06	1.10	10.32	1.27	10.91	1.09
C	19.99	0.60	16.84	2.00	17.94	1.82
D	58.51	6.91	24.36	2.66	26.20	3.04
E	25.80	1.40	32.85	4.28	35.63	4.43
F	42.18	2.27	40.93	6.53	44.76	5.25
G	5.38	1.52	52.18	11.42	60.12	8.07

4.2.2 Correlazione e fattori macroeconomici

La correlazione dei default impatta sulla variabilità delle perdite di portafoglio. CreditRisk+ incorpora gli effetti di questa correlazione utilizzando la volatilità delle probabilità di default e dell'appartenenza settoriale. I default avvengono come conseguenza di situazioni che rendono impossibile definire l'esatto momento del default o il numero esatto di debitori insolventi, infatti spesso ci sono fattori macroeconomici che incidono sulla correlazione dei default senza che vi sia un nesso causale tra di essi, come per esempio il tasso di crescita dell'economia e il tasso di interesse applicato. È ormai risaputo che il ciclo economico di uno stato e/o il settore di appartenenza di un debitore hanno impatto diretto sul merito di credito della controparte, ma la forza dell'influenza dipende dalla sensitività del debitore ai vari fattori. I modelli costruiti cercando di catturare esplicitamente l'effetto dei cambiamenti economici sulla frequenza dei default, presentano diversi limiti, tra cui una tangibile difficoltà di verifica dell'accuratezza del modello che deriva probabilità di default per nazione o settore, in quanto ci sono dati pubblici limitati. Inoltre, anche se ci fosse una relazione causale tra le situazioni di default ed una variabile economica, non è detto che questa relazione si mantenga stabile nel tempo. Per cui sono stati vagliati diversi modelli alternativi che cercano di catturare la variabilità della frequenza dei default senza calcolare esplicitamente la variazione dello scenario economico. È possibile incorporare l'incidenza dei fattori economici utilizzando una distribuzione delle frequenze di default incorporando la loro deviazione standard nel modello. CreditRisk+ si inserisce in questo secondo filone, per diverse ragioni:

- **Instabilità delle correlazioni dei default**: generalmente, le correlazioni calcolate a partire dai dati finanziari presentano una forte instabilità, oltre ad

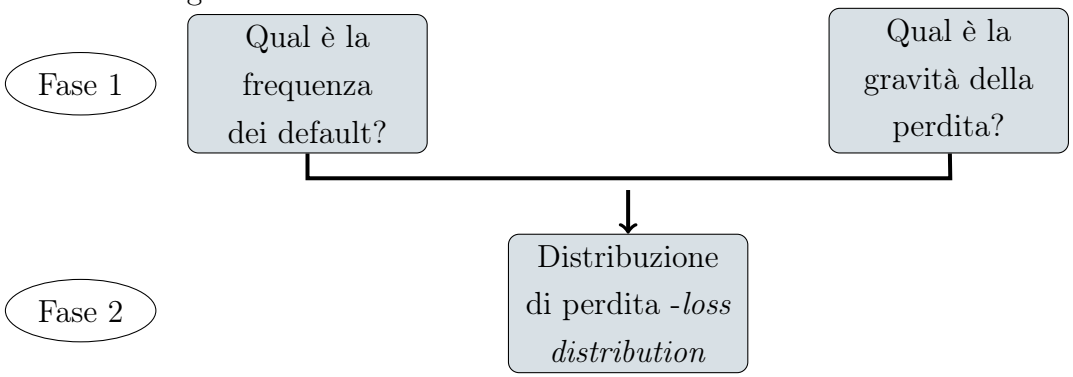
essere dipendenti dal periodo considerato per il calcolo. Lo stesso problema nasce nell'analisi del merito di credito.

- **Mancanza di dati empirici**: l'analisi del merito di credito presenta pochi dati, a causa della privacy o comunque perchè i default non sono così frequenti da permettere il calcolo delle correlazioni in modo esplicito.

Ad ogni modo, entrambi gli approcci comportano un'ispessimento delle code della distribuzione e di conseguenza un aumento delle perdite stimate. È risaputo che lo stato dell'economia di differenti nazioni può cambiare nel tempo e i diversi settori industriali interni possono esserne influenzati con diversa intensità. Un portafoglio di esposizioni potrebbe avere una concentrazione di controparti in un determinato settore o area geografica, perciò è importante che il modello selezionato catturi l'effetto del rischio di concentrazione, in CreditRisk+ si utilizzerà a tal fine l'analisi settoriale, suddividendo una particolare esposizione tra i settori di appartenenza. In questa sede si suddivideranno le controparti per area geografica. Gli stati americani sono suddivisi in "South - S", "West - W","MidWest - MW" e "NorthEast - NE", a cui si aggiunge un settore più generale del mercato americano globale. Non si utilizza l'appartenenza ai singoli stati in quanto aumentando il numero di fattori nel modello, aumenta l'ordine dell'approssimazione della matrice di correlazione.

4.3 Implementazione

La modelizzazione del rischio di credito prevede un processo a due fasi, come illustrato nel diagramma sottostante:



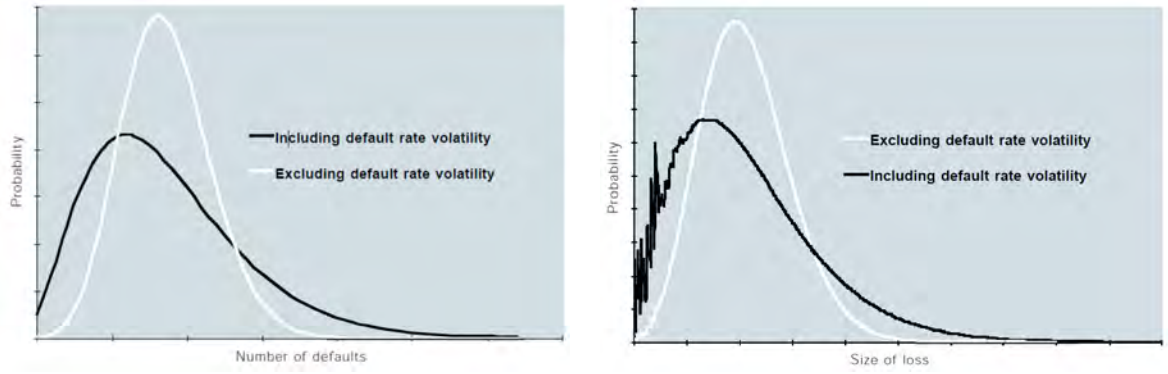
Calcolando la distribuzione dei default è possibile valutare se la qualità del portafoglio creditizio stia migliorando o peggiorando; la distribuzione delle perdite, invece, permette di valutare l'impatto delle perdite potenziali, oltre a misurare l'ammontare della diversificazione e della concentrazione di portafoglio.

Distribuzione Il default avviene come sequenza di eventi che rendono impossibile prevedere l'esatto momento della perdita e la frequenza dei default totali. Esiste anche la possibilità di perdita derivante dall'intero portafoglio anche se la probabilità di default associata ad ogni singola controparte è bassa. Questa situazione si può rappresentare con una distribuzione di Poisson. Se si considera una *default distribution* di un portafoglio composto da un certo *rating mix* senza considerarne la volatilità si può rappresentare approssimativamente come distribuzione di Poisson. Questa è la base di partenza del modello, a cui verrà aggiunta la deviazione standard, per incorporare la variazione dei rating nel tempo.

CrediRisk+ modellizza quindi il tasso di default sottostante attraverso una probabilità di default media e la sua volatilità, per tenere in considerazione la possibile variazione temporale in modo pragmatico, senza introdurre ulteriore termine d'errore nel modello con ulteriori input da stimare. L'effetto della considerazione della volatilità può essere riassunto nella figura 4.4a nella pagina successiva tratta da *CreditRisk+: A credit risk management framework*, p. 18, la quale mostra una distribuzione di default generata con CreditRisk+, dapprima senza considerare la volatilità, inserendo poi nella stessa distribuzione la deviazione standard. In questa seconda ipotesi la distribuzione presenta una forte asimmetria a destra con conseguente aumento del rischio di eventi estremi, le code risultano infatti più spesse. Calcolato il numero di default, si può procedere a creare la *loss distribution* dell'intero portafoglio, inserendo anche l'ammontare di capitale delle perdite, che dipende direttamente dall'esposizione dei debitori. A differenza della distribuzione dei default, la volatilità nell'ammontare dei prestiti risulta in una distribuzione delle perdite che non può essere approssimata in una *Poisson distribution*. Tuttavia, è possibile descrivere la *loss distribution* complessiva perché la sua funzione generatrice delle probabilità ha una forma chiusa, calcolabile in modo analitico. Abbiamo già parlato dei *recovery rate*. Per evitare il numero di dati da processare, si utilizzano i RR per scontare le esposizioni al fine di calcolare le *loss given default*, poi le esposizioni nette verranno suddivise in gruppi, creati come multipli di una unità scelta arbitrariamente tenendo conto della grandezza di scala delle esposizioni. Attraverso CrediRisk+ si può calcolare la probabilità che avvenga una perdita ragionando in termini di multipli ed ottenendo una funzione seghettata come quella in figura 4.4b[4]. Vanno notate alcune caratteristiche importanti:

- Entrambe le *loss distribution* hanno lo stesso livello di perdita attesa.
- La differenza sostanziale, tra l'inclusione e non della volatilità nella distribuzione delle perdite, è l'ispessimento delle code della *loss distribution* con

[4]tratta da Boston 1997



(a) Distribuzione dei default senza volatilità **(b)** Distribuzione dei default con volatilità

Figura 4.4: Aspetto delle distribuzioni del modello

l'inserimento della deviazione standard. Per esempio, al 99-esimo percentile la perdita è significativamente più alta.

Poiché le code della distribuzione si sono ispessite, mentre la perdita attesa è rimasta invariata, si può concludere che la varianza della *loss distribution* è aumentata. L'aumento della volatilità è dovuto alla correlazione tra controparti, incorporata nel modello attraverso l'analisi settoriale, oltre che dalla deviazione standard, per misurare i benefici della concentrazione e della diversificazione. Spesso la diversificazione avviene in modo naturale quando il numero delle controparti nel portafoglio è alta, ma questo potrebbe comportare un rischio di concentrazione, che nasce quando dei debitori entrano in sofferenza perchè influenzati da un fattore comune. Per analizzare il rischio di concentrazione, occorre avere ben chiara la differenza tra rischio sistemico e rischio idiosincratico. Il rischio sistemico è causato da fattori sistemici: sono quelle variabili che intaccano il merito di credito di un gruppo di controparti con un'influenza in comune, come per esempio il domicilio. Il rischio idiosincratico trova causa nei fattori specifici: caratteristiche proprie della controparte che intaccano il merito di credito proprio del debitore, pesano prevalentemente sulle perdite estreme di un portafoglio di controparti creditizie. Il rischio di concentrazione del portafoglio dipende principalmente da fattori sistemici, ed è misurabile attraverso l'analisi settoriale, inserendo nel modello una matrice con l'appartenenza settoriale delle controparti, da cui derivano diverse situazioni:

(i) **Allocazione di tutte le controparti in un singolo settore**
La versione più diretta e veloce del modello prevede l'appartenenza di tutti i debitori ad un singolo settore. In questo modo un singolo fattore sistemico influenza la volatilità delle probabilità di default delle controparti. In questo modo, inoltre, si cattura nel modello tutto il rischio di concentrazione del

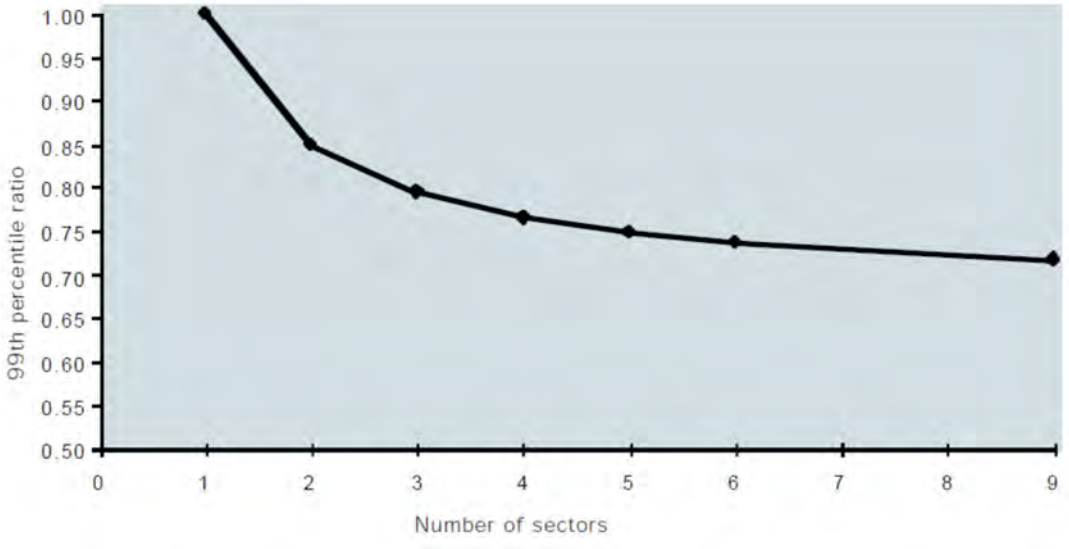


Figura 4.5: Variazione dell'impatto del rischio di concentrazione

portafoglio e si esclude il beneficio della diversificazione. Questa è la via più prudente per la stima delle componenti estreme di perdita.

(ii) **Allocazione delle controparti in diversi settori**
Al fine di riconoscere il beneficio della diversificazione, si può assumere che ogni controparte venga influenzata da un fattore sistemico, responsabile di tutta la volatilità della probabilità di default delle controparti influenzate da esso. Per esempio, i debitori possono essere divisi per regione di appartenenza. In questo modo si *clusterizza* la volatilità delle PD in gruppi i cui individui sono influenzati dallo stesso fattore.

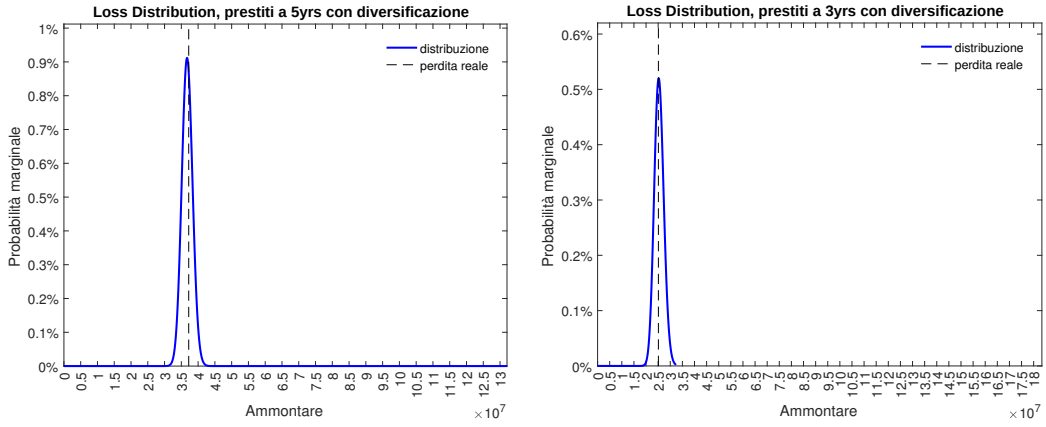
(iii) **Suddivisione di ogni controparte in più settori**
Il modello più generale permette di assumere l'influenza di diversi fattori per ogni singolo debitore. CreditRisk+ permette di attribuire la varianza delle PD a diversi fattori di rischio, ponderandola in base all'entità dell'influenza. Questa logica rappresenta la declinazione del modello meno prudente.

La figura 4.5[5] mostra come, all'aumentare del numero di settori, l'impatto del rischio di concentrazione sia minore. Il grafico mostra la variazione dell'entità della perdita al 99esimo percentile in base al numero di settori considerati.

Backtesting Come tutti i modelli, anche per il portafoglio è prevista una validazione. In questa sede è consigliabile effettuarla prima di procedere all'implementazione del modello con prestiti *current*, in quanto è necessario confrontare

[5]tratta da *CreditRisk+: A credit risk management framework*

la perdita attesa stimata con le perdite realmente sostenute per verificare quanto il modello sia capace di prevedere il rischio da affrontare. La validazione viene effettuata sul portafoglio di prestiti a cinque anni dell'anno 2012 e successivamente sui prestiti a tre anni erogati nel 2014.



(a) Confronto del portafoglio a cinque anni **(b)** Confronto del portafoglio a tre anni

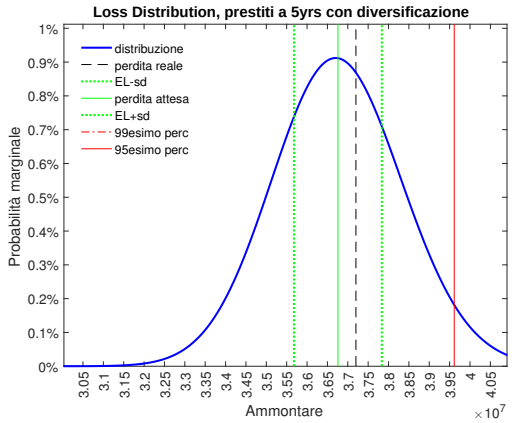
Figura 4.6: Confronto perdita attesa stimata con la perdita reale

Dalla figura 4.6 risulta evidente come in entrambi i casi la *loss distribution* stimata contenga al suo interno la perdita reale sopportata, rappresentata dalla linea nera tratteggiata, anche se nel portafoglio di prestiti a cinque anni, la stima dell'*expected loss* risultante dal modello è inferiore alle perdita reale, mentre nel portafoglio di prestiti a tre anni risulta leggermente superiore. La perdita attesa che restituisce il modello è quindi molto vicina alla perdita reale, soprattutto per i prestiti a tre anni. Infatti nella figura 4.7a, che riporta il dettaglio della distribuzione delle perdite del portafoglio di prestiti a cinque anni, la linea nera tratteggiata, che rappresenta l'ammontare di perdita realmente perso nell'anno, è spostata a destra nella distribuzione rispetto alla linea verde continua, la quale indica il capitale perso stimato attraverso CreditRisk+; la perdita reale è comunque compresa dentro l'intervallo creato attraverso la deviazione standard $[\text{EL}-\sigma;\ \text{EL}+\sigma]$. La linea rossa inserita nella distribuzione, rappresenta il capitale perso con un intervallo di confidenza del 95%, mentre risulta assente una linea rossa tratteggiata che indica la perdita al 99esimo percentile. Questo è molto interessante poiché fa pensare che grazie alla diversificazione geografica il portafoglio abbia una perdita con un intervallo di confidenza del 99% nulla, che per l'esattezza corrisponde ad una *unexpected loss* di soli 24 dollari, arrotondata dal modello a zero. Il portafoglio di prestiti a tre anni viene rappresentato in figura 4.7b nella pagina seguente. La notazione utilizzata per rappresentare le componenti inserite nel grafico rimane sempre la stessa. In questo caso specifico la linea nera tratteggiata della perdita

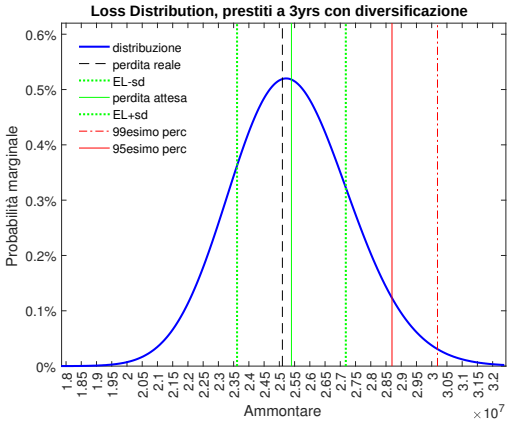
reale è leggermente spostata verso sinistra rispetto alla linea verde continua delle perdite attese stimate, ma comunque inserita nel tunnel verde tratteggiato creato con la considerazione della volatilità. Diversamente dal grafico analizzato prima, qui si nota anche la linea rossa tratteggiata inserita per rappresentare il 99esimo percentile, si può notare come questo tagli la distribuzione sulla parte della coda più estrema, rappresentando perdite inattese di tre punti percentuali in più rispetto alla perdita attesa, pari al 13.89%, ossia un 16.5% del capitale a rischio. Il modello quindi può essere validato rispetto alla stima delle perdite future. Nella tabella 4.7c si riportano le perdite reali e le perdite stimate, calcolate anche in termini percentuali rispetto al capitale a rischio erogato in quanto, sebbene sia un campione ridotto e non si possa parlare in termini assoluti di capitale, sembra ragionevole poterlo utilizzare come *proxy*, presentando un *rating mix* uguale al campione totale.

Le distribuzioni hanno, comunque, una caratteristica evidente: rispetto alla distribuzione di esempio della figura 4.8 ottenuta con il codice di prova fornito da CSFB, non presentano "seghettatura". Questa differenza nella forma della distribuzione è dovuta alla bassa deviazione standard delle probabilità di default (tabella 4.1 a pagina 92), rispetto alla volatilità fornita nell'esempio di CSFB.

Per completezza si è provato ad aggiungere il rischio di concentrazione ai portafogli, ipotizzando l'appartenenza di tutte le controparti allo stesso settore, per verificare come variano la capacità predittiva e la distribuzione. Si può notare (tabella 4.9e) che la concentrazione settoriale non ha ovviamente intaccato la perdita attesa, essendo semplicemente la sommatoria della probabilità di default della controparte per il suo capitale erogato. Si è, invece, sensibilmente ispessita la coda destra della distribuzione. Nel portafoglio di prestiti a tre anni si traduce in perdite maggiori di un punto percentuale; mentre nell'aggregato di prestiti a cinque anni si presenta una perdita al 99esimo percentile, che prima veniva mitigata, corrispondente al 32% del capitale a rischio. Come appena notato, il maggior rischio di un portafoglio intacca solo le code della *loss distribution*, quindi è utile analizzare un'altra misura di rischio: il *Credit Value at Risk* o *Credit Var*. Secondo quanto riportato in Hull (2006), il "valore a rischio" rappresenta il tentativo di riassumere in un solo numero il rischio complessivo di un portafoglio. Indica l'ammontare di perdita che, ad un certo livello di confidenza α, non verrà oltrepassato entro l'orizzonte temporale considerato. Il VaR creditizio, quindi, può essere definito alla stesso modo, ed è considerato il patrimonio da accantonare a fronte dei rischi di credito. Nei portafogli creditizi, il C.VaR viene calcolato come differenza tra le perdite ad un determinato livello di confidenza e la perdita attesa stimata. In questa analisi il C.VaR calcolato con le perdite inattese corrisponde al capitale



(a) Portafoglio a cinque anni



(b) Portafoglio a tre anni

	5 anni		3 anni	
	Capitale	(%)	Capitale	(%)
Capitale	131 577 772.35	100.00	182 892 827.60	100.00
Perdita	37 200 000.00	28.27	25 100 000.00	13.72
SD	1 078 009.74	0.82	1 784 476.65	0.98
EL $-\sigma$	35 680 120.52	27.12	23 612 277.94	12.91
EL	36 758 129.93	27.94	25 396 754.59	13.89
EL actual	37 200 000.00	28.00	25 100 100.00	13.72
EL $+\sigma$	37 836 139.34	28.76	27 181 231.24	14.86
95esimo perc.	39 612 858.00	30.11	28 696 552.00	15.69
UL - 99esimo perc.	0.00	0.00	30 189 792.00	16.51
CVaR 95esimo	2 854 728.07	2.17	3 299 797.41	1.80
CVaR 99esimo	0.00	0.00	4 793 037.41	2.62

(c) Ammontare delle perdite nei portafogli con geo-diversificazione

Figura 4.7: Perdite stimate con il modello di portafoglio

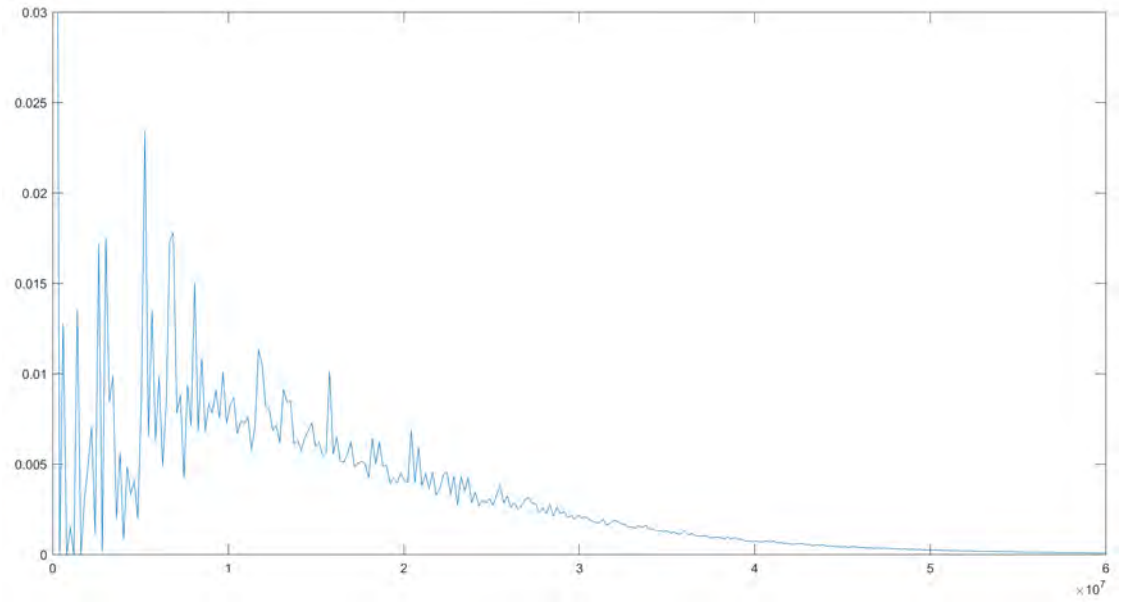


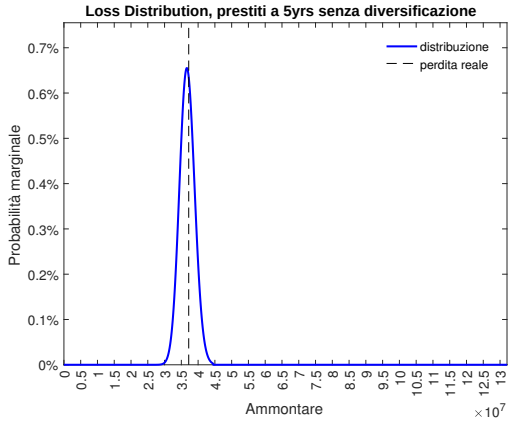
Figura 4.8: Loss distribution con dati di prova forniti da CSFB

economico.

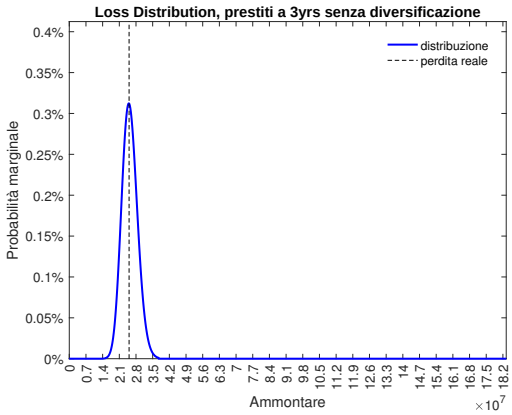
In ogni caso il modello viene considerato predittivo, si procede quindi alla sua implementazione con prestiti *current*. Di tutti i prestiti ancora in vita di Lending Club ci si aspetta di perdere il 14% circa del capitale non coperto da garanzie o da tassi di recupero, con una perdita inattesa che arriva al 16.30% al 99esimo percentile. L'inserimento di dati annuali, fa si che questa perdita sia da considerarsi annuale. Come per gli altri portafogli si è proceduto al confronto con il portafoglio senza il beneficio della geo-diversificazione, apprezzando un ispessimento della coda di destra che arriva a identificare una perdita del 18% circa rispetto al capitale non protetto, nonché un aumento della volatilità della perdita attesa di circa un punto percentuale.

Nella tabella 4.2 a pagina 103 si riportano le perdite percentuali rispetto al capitale erogato (al lordo dei recovery rate), rappresentate poi in figura 4.11 a pagina 104. Si nota come, sebbene le probabilità di default siano elevate, il capitale perso è comunque abbastanza contenuto, soprattutto nel portafoglio con i prestiti ancora in corso, in cui ci sono sia prestiti a tre anni, sia prestiti a cinque anni. Questo è permesso da tassi di recupero alti e da forze mitigatrici come la contribuzione al rischio negativa e all'effetto di diversificazione geografica.

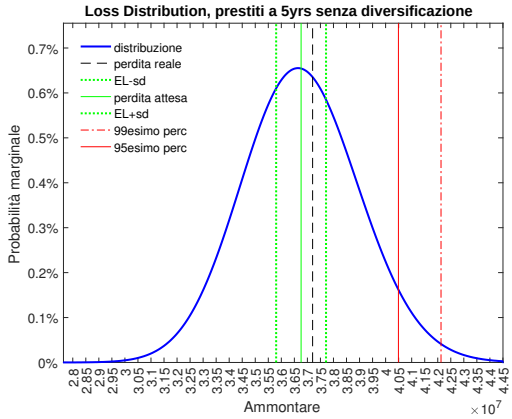
In appendice B a pagina 129 viene riportato il modello CreditRisk+ in forma estesa con tutte le formule per i calcoli necessari e la loro derivazione.



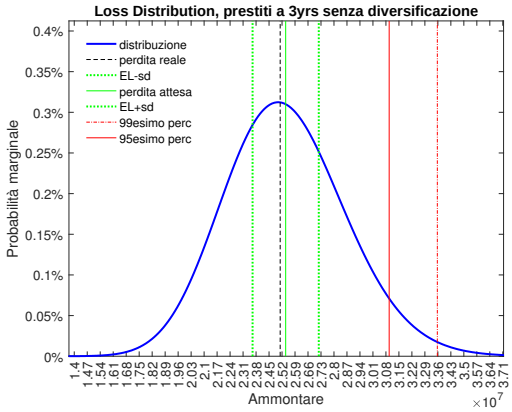
(a) Portafoglio a cinque anni



(b) Portafoglio a tre anni



(c) Percentili portafoglio a cinque anni

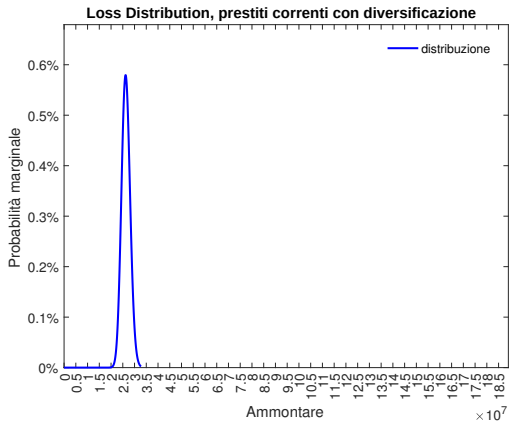


(d) Percentili portafoglio a tre anni

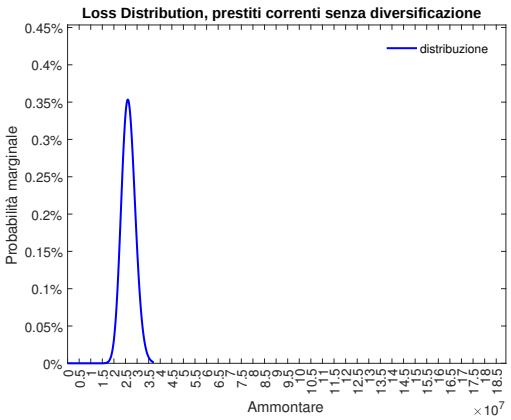
	5 anni		3 anni	
	Capitale	(%)	Capitale	(%)
Capitale	131 577 772.35	100.00	182 892 827.60	100.00
Perdita	37 200 000.00	28.27	25 100 000.00	13.72
SD	956 773.74	0.73	1 784 526.65	0.98
EL $-\sigma$	35 801 356.19	27.21	23 612 277.94	12.91
EL	36 758 129.93	27.94	25 396 754.59	13.89
EL actual	37 200 000.00	28.00	25 100 100.00	13.72
EL $+\sigma$	37 714 903.67	28.66	27 181 231.24	14.86
95esimo perc.	40 491 430.00	30.77	30 975 282.00	16.94
UL - 99esimo perc.	42 125 444.00	32.02	33 571 394.00	18.36
C.VaR 95esimo	3 733 300.07	2.84	5 578 527.41	3.05
C.VaR 99esimo	5 367 314.07	4.08	8 174 639.41	4.47

(e) Ammontare delle perdite dei portafogli senza diversificazione geografica

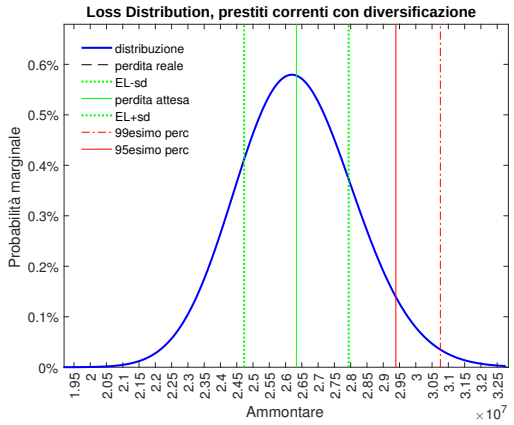
Figura 4.9: Risultati del modello senza diversificazione



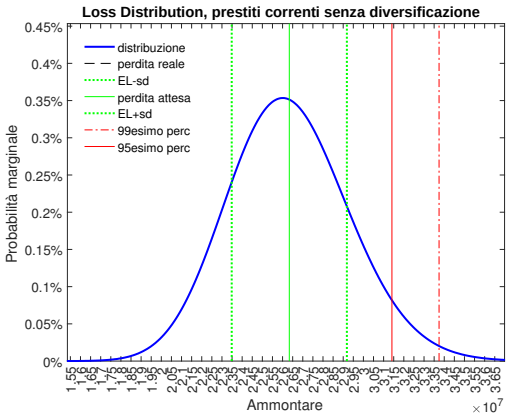
(a) Portafoglio current con diversificazione



(b) Portafoglio current senza diversificazione



(c) Percentili portafoglio current con diversificazione



(d) Percentili portafoglio current senza diversificazione

	con diversificazione		senza diversificazione	
	Capitale	(%)	Capitale	(%)
Capitale	188 667 917.72	100.00	188 667 917.70	100.00
SD	1 605 158.72	0.85	2 849 515.49	1.51
EL $-\sigma$	24 727 730.54	13.11	23 483 373.77	12.45
EL	26 332 889.26	13.96	26 332 889.26	13.96
EL $+\sigma$	27 938 047.97	14.81	29 182 404.74	15.47
95esimo perc.	29 387 323.00	15.58	31 411 226.00	16.65
UL - 99esimo perc.	30 754 708.00	16.30	33 743 702.00	17.89
C.VaR 95esimo	3 054 433.74	1.62	5 078 336.74	2.69
C.VaR 99esimo	4 421 818.74	2.34	7 410 813.00	3.93

(e) Ammontare delle perdite attese per i prestiti Current

Figura 4.10: Risultati del modello sui prestiti correnti

Tabella 4.2: Perdite percentuali rispetto al capitale erogato

	con diversificazione			senza diversificazione		
	60mesi	36mesi	current	60mesi	36mesi	current
Capitale erogato	210 000 000.00	378 000 000.00	378 000 000.00	210 000 000.00	378 000 000.00	378 000 000.00
Capitale	188 667 917.72	182 892 827.60	131 577 772.35	188 667 917.72	182 892 827.60	131 577 772.35
EL (%)	17.50	6.72	6.97	17.50	6.72	6.97
95esimo (%)	18.86	7.59	7.77	19.28	8.19	8.31
99esimo (%)	0.00	7.99	8.14	20.06	8.88	8.93
C.VaR 95esimo	1.36	0.87	0.81	1.78	1.48	1.34
C.VaR 99esimo	0.00	1.27	1.17	2.56	2.16	1.96

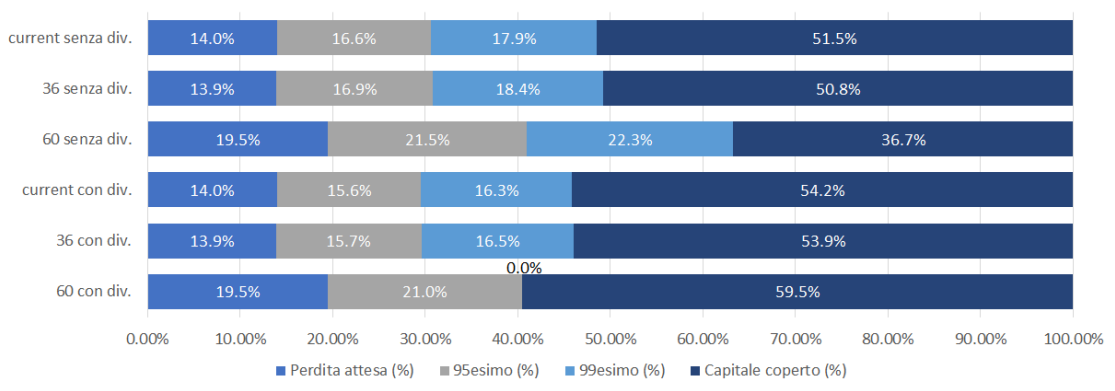


Figura 4.11: Rappresentazione della composizione delle perdite

Limiti

CreditRisk+ presenta indubbiamente dei grandi vantaggi, derivanti soprattutto dalla formulazione chiusa per la derivazione delle grandezze di interesse (perdita attesa, perdita inattesa e capitale economico), permettendo l'agevole computo di controparti ulteriori a portafogli già in essere o della rimozione di parte di essi dal portafoglio. Questo permette agli istituti finanziari di identificare le politiche di gestione del rischio e di copertura più adeguate. Non ultimo, la limitata necessità di dati di input. Come ogni modello però, CreditRisk+ presenta anche delle limitazioni[6]. L'algoritmo originale presentato da CSFB per l'implementazione del modello, si basa su una formulazione ricorsiva che, nel contesto del rischio di credito presenta due problemi dovuti alla limitata precisione che i computer utilizzati per l'implementazione hanno. In primo luogo la formulazione ricorsiva non riesce a sostenere portafogli arbitrariamente grandi. Man mano che il numero dei default aumenta, aumenta l'imprecisione. In secondo luogo, l'algoritmo è numericamente instabile, l'errore accumulato durante il calcolo è significativo. A questo fine sono state proposte diverse soluzioni alternative, come l'utilizzo dell'algoritmo di Panjer o l'utilizzo della *Fast Fourier Transformation*. Per ulteriori approfondimenti in merito si consigliano le letture *Creditrisk+ by fast fourier transform*, Melchiori (2004) e *Review and Implementation of Credit Risk Models*, AVESANI et al. (2014). Un'ulteriore difetto è rappresentato dall'impossibilità di considerare in modo implicito variazioni della qualità del credito. Non vi è modo di inserire, quindi, una matrice di transizione nella computazione del modello, se non utilizzando analisi di scenario successivi.

[6]Si può trovare una sintesi in *Review and Implementation of Credit Risk Models*

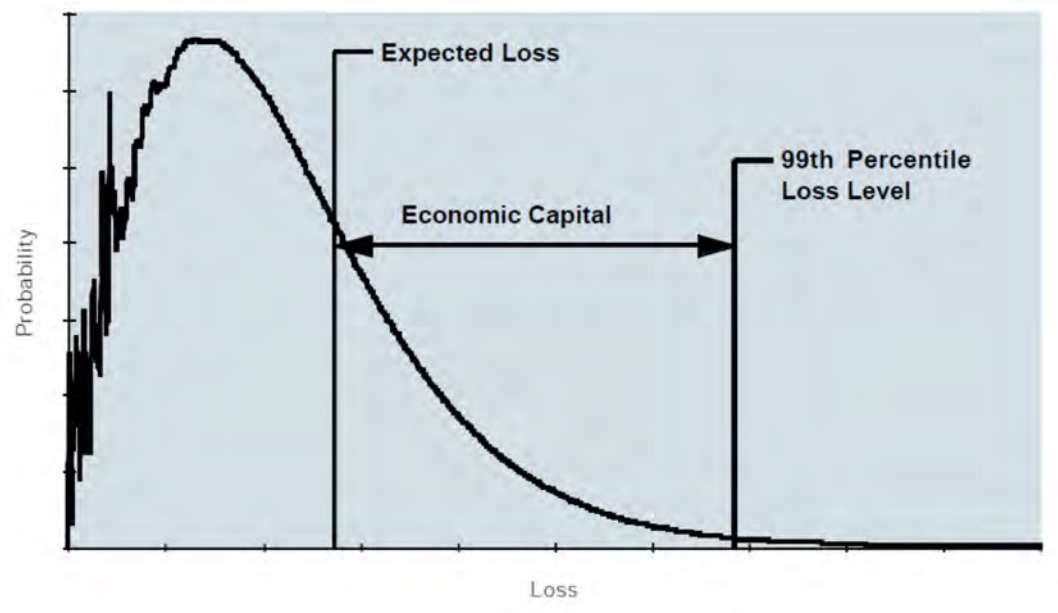


Figura 4.12: Rappresentazione del capitale economico

Il capitale economico e tassi di interesse *risk adjusted*

L'analisi del rischio è la base del *risk management*, quindi, l'analisi delle perdite inattese in un portafoglio creditizio è fondamentale per la gestione del rischio di credito. Per affrontare tali perdite è necessario generare profitti adeguati attraverso l'applicazione di tassi aggiustati per il rischio. Il capitale economico, o loss absorbing capacity (LAC), è un *buffer* di liquidità pensato per attutire le perdite inattese, perchè il livello di perdita subito in un determinato momento potrebbe essere maggiore di quanto ci si aspetti. La conoscenza della distribuzione delle perdite (*loss distribution*), restituisce le informazioni sull'ammontare di capitale che si potrebbe perdere; determinato un percentile si può arrivare a determinare il livello di capitale economico necessario per coprire le perdite a quel determinato livello di confidenza. Per catturare una significante proporzione delle code della *loss distribution* in un orizzonte temporale di un anno, è suggerito il 99esimo percentile (figura 4.12 nella pagina successiva). Si utilizza il concetto di capitale economico in quanto è la misura più appropriata specificata dalla normativa, creata su una logica di portafoglio tiene quindi in considerazione il beneficio della diversificazione. Questa misura riesce a differenziare portafogli diversi poiché tiene in considerazione la qualità del credito e l'ammontare delle esposizioni. Al suo interno considera anche i cambiamenti del rischio aggregato, è quindi utilizzata per l'ottimizzazione di portafoglio. Si riporta in tabella 4.3 il capitale economico per i casi in esame. Si può notare come, ovviamente, nel caso di assenza di diversificazione, il capitale economico richiesto sia maggiore dei casi in cui il portafoglio

Tabella 4.3: Capitale economico per i portafogli analizzati

		Ammontare	% sul capitale erogato
con diversificazione	60 mesi	-36 758 129.93	-17.50
	36 mesi	4 793 037.41	1.27
	current	4 421 818.74	1.17
senza diversificazione	60 mesi	5 367 314.07	2.56
	36 mesi	8 174 639.41	2.16
	current	7 410 812.745	1.96

Tabella 4.4: Percentili della loss distribution del portafoglio a cinque anni con diversificazione

Percentile	Ammontare
media	36,758,130
50	36,748,309
75	37,852,207
95	39,612,858
97.5	40,408,465
99	0
99.5	0
99.75	0
99.9	0
capitale economico	3,650,335
	1.74%

consideri la geo-diversificazione. Il caso particolare del portafoglio di prestiti a cinque anni con diversificazione richiede un'analisi ulteriore, in quanto non subendo perdite al 99esimo percentile, restituisce un capitale economico negativo. Questo non significa che il portafoglio non possa subire perdite estreme, ma semplicemente che le perdite si concentrano nei percentili precedenti. In tabella 4.4 nella pagina successiva si riportano tutte le stime effettuate per livello di confidenza selezionato, rendendo ovvio che il calcolo del capitale economico verrà effettuato ad un livello di 97.5. A questo livello di confidenza si richiede che l'1.74% del capitale venga accantonato al fine di coprire le perdite inattese.

Un altro aspetto interessante da valutare è il calcolo dell'interesse aggiustato per il rischio. Nella gestione del rischio di credito, l'interesse *risk adjusted*, rappresenta una parte molto importante, poiché permette di capire se si sta ottenendo una corretta remunerazione dagli investimenti effettuati rispetto al rischio che ci si sta accollando. La formulazione, supponendo un investitore neutrale al rischio non

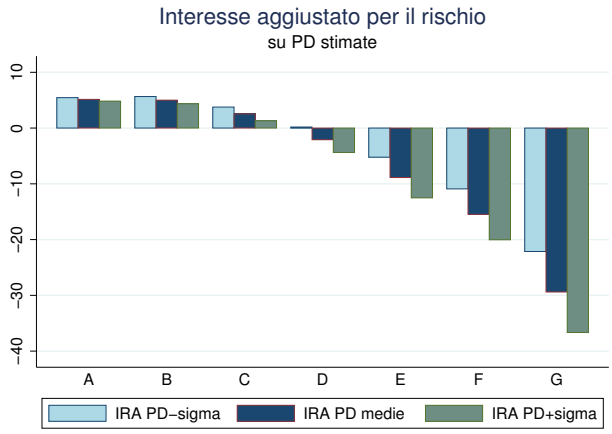
Tabella 4.5: Confronto dell'interesse Risk Adjusted

	StimaPD$-\sigma$ (%)	Frequenze (%)	StimaPD (%)	StimaPD$+\sigma$ (%)
A	4.83	4.64	5.14	5.45
B	4.36	4.72	5.01	5.65
C	1.31	1.66	2.54	3.77
D	-4.37	-1.83	-2.09	0.19
E	-12.52	-7.10	-8.87	-5.22
F	-20.04	-9.86	-15.47	-10.91
G	-36.67	-14.69	-29.40	-22.14

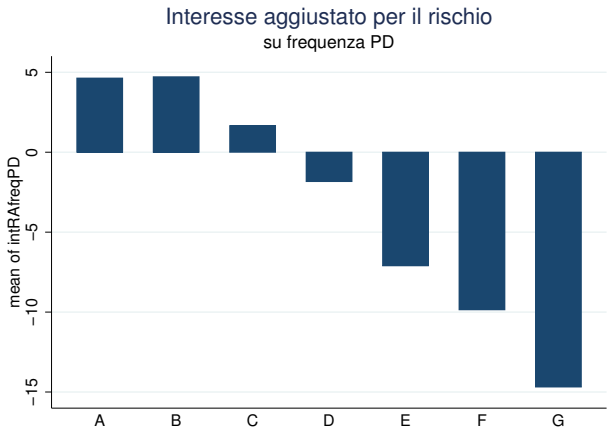
potendo fare assunzioni sul rendimento atteso soggettivo, è:

$$(1 + r) = (1 - PD) * (1 + i) + (PD * RR) \tag{4.1}$$

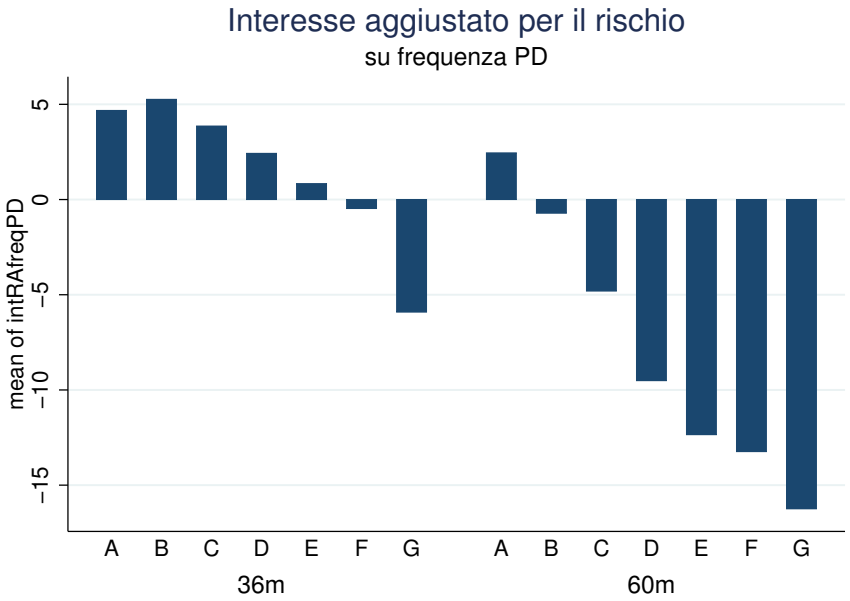
dove r è il tasso di interesse aggiustato per il rischio, i è il tasso applicato al debitore, PD e RR sono , rispettivamente, le probabilità di default e i *recovery rates* stimati in precedenza. Avendo tutte le componenti rientranti nella formulazione si procede a stimare l'r medio ponderato per ammontare suddiviso secondo classe di rating (tabella 4.5). L'interesse aggiustato per il rischio calcolato con le probabilità di default stimate, riporta una range dal 5.14% (rating A) al -29.4% (rating B). A causa della volatilità implicita nelle probabilità di default, è stato stimato anche l'interesse *risk adjusted* considerando il limite inferiore, calcolato come PD medie meno la deviazione standard, e il limite superiore, rappresentato dalle PD medie più la deviazione standard. Il trend decrescente è molto marcato, quindi la maggior parte delle classi di rating risulta sottoprezzata rispetto al rischio intrinseco. Questo si può notare nella figura 4.13a nella pagina successiva. Allo stesso modo si confronta l'interesse *risk adjusted* sulle frequenze di default storiche, in quanto abbiamo visto in figura 2.26a a pagina 72 che le probabilità di default stimate divergono leggermente dalle frequenze relative. La rappresentazione dei risultati ottenuti è contenuta nella figura 4.13b nella pagina successiva e nella tabella 4.5 nella pagina precedente. Tutte le stime dell'interesse *risk adjusted* stimato sulle frequenze reali sono comprese nel tunnel creato con la stima utilizzante le probabilità di default stimate. La caratteristica di sotto prezzare i prestiti erogati, comporta la possibilità, per l'investitore, di incorrere in una perdita, non necessariamente in termini monetari, ma quanto meno in termini di opportunità. Risulta utile a tal fine analizzare l'interesse aggiustato per il rischio raggruppando le classi di rating secondo scadenza dei prestiti (figura **??** a pagina ??). Il risultato è interessante, Lending Club non riesce a catturare nel prezzo tutto il rischio prevalentemente



(a) IRA con PD stimate



(b) IRA con frequenze di default



(c) Analisi dell'Interest Risk Adjusted per classe di rating e scadenza

Figura 4.13: Calcolo dell'Interest Risk Adjusted per classe di rating

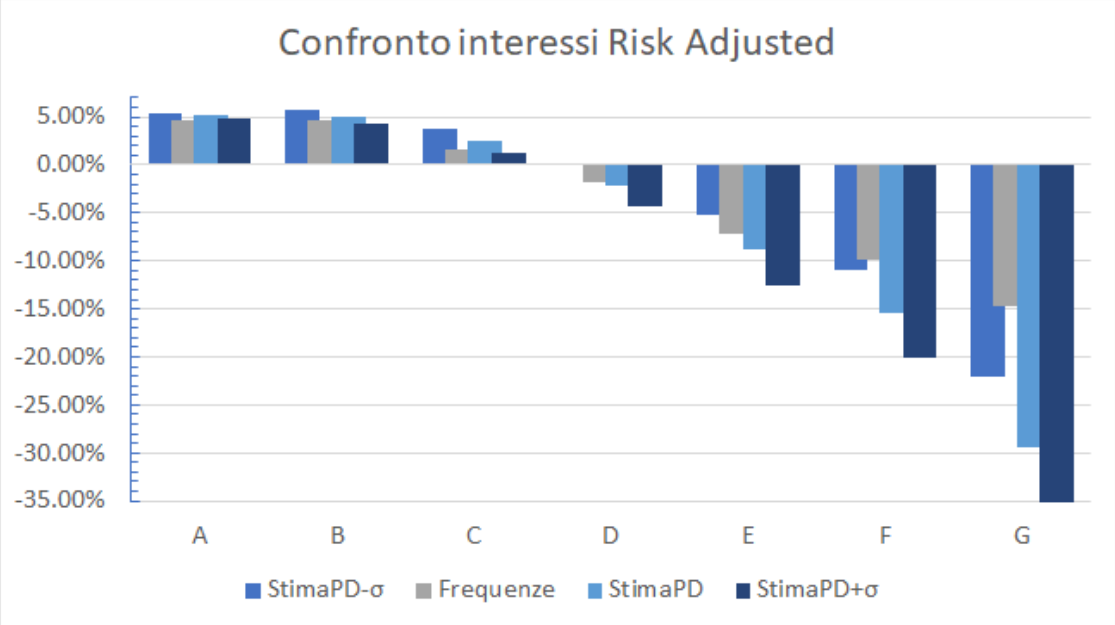


Figura 4.14: Confronto dei diversi livelli di Interest Risk Adjusted

nei prestiti a cinque anni. Questo comporta una notevole influenza nella media dell'interesse *risk adjusted* anche nei prestiti a tre anni, categoria in cui la maggior parte delle classe di rating sia prezzata correttamente. Ad ogni modo come si vede dalla figura 4.14 a pagina 109 solo le classi di rating peggiori, ossia E, F e G, presentano una divergenza sensibile tra l'interesse corretto per il rischio calcolato sulle frequenze relative e l'interesse calcolato con le probabilità di default stimate. Questo perchè la regressione ha determinato, per queste classi di rating, una probabilità di default maggiore rispetto alla frequenza reale dei default passati. Tutte le altre classi sono in linea con l'evidenza empirica. Sebbene il risultato non sia una notizia positiva per il mercato degli investimenti, non dovrebbe stupire. Qualsiasi ricerca sull'analisi dei tassi *risk adjusted* di prestiti personali restituisce una sistematica tendenza a sotto prezzare questa categoria di prodotti, in virtù della concorrenza spietata del mercato. Se per prestiti ingenti o controparti corporate la normativa è molto più stringente, permettendo di poter apprezzare prestiti correttamente prezzati e quindi di ricevere una remunerazione dall'investimento effettuato, questo non è altrettanto vero per i prestiti personali. A parità di servizi offerti, un debitore chiederà un prestito all'istituto che applicherà condizioni più favorevoli, traducendosi in un minore tasso di interesse applicato. Per mantenere alti i volumi e i prodotti concorrenziali, le istituzioni ragionano in ottica prettamente commerciale, applicando un tasso di interesse più basso rispetto al dovuto per rendere il prodotto più attraente agli occhi dei potenziali clienti/debitori. in questo specifico contesto, questo è ancora più vero in virtù del fatto che non vi è

Tabella 4.6: Rendimento ottenuto dai prestiti a scadenza

	Portafoglio		36mesi		60mesi	
	μ (%)	σ (%)	μ (%)	σ (%)	μ (%)	σ(%)
2007	-2.210	-0.847	0.517	-0.847		
2008	-1.338	-0.684	-0.031	-0.684		
2009	7.850	8.207	8.563	8.207		
2010	7.511	7.759	8.007	7.440	8.437	
2011	10.322	10.515	10.708	8.034	13.073	
2012	11.623	11.742	11.860	9.531	17.079	
2013	10.923	10.995	11.066	11.140	10.666	
2014	6.177	6.234	6.290	8.909	0.583	

nemmeno l'obbligo di accantonare capitale a fronte del rischio sostenuto, perchè le perdite sono interamente a carico degli investitori.

4.4 Rendimenti

Visto quanto fino a qui emerso, sembra doveroso porre attenzione anche al rendimento ottenuto da questi portafogli. Lending Club obbliga le controparti a delle commissioni ogni qualvolta non rispettino la scadenza per il pagamento delle rate. Questa componente crea profitto fin tanto che il debitore non defaulta, quindi a fine prestito i flussi in entrata sul capitale erogato sono:

$$\pi = \frac{total\_rec\_prncp + total\_rec\_int + total\_rec\_late\_fee + recoveries}{loan\_amnt} - 1 \tag{4.2}$$

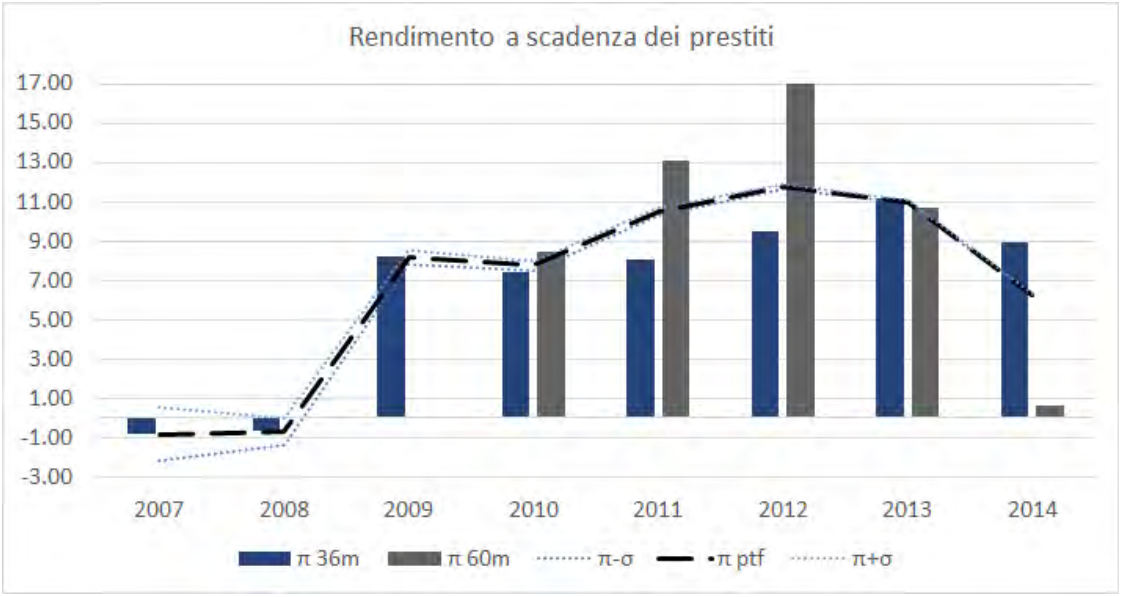
dove total_rec_prncp è il capitale restituito, total_rec_int sono la quota capitale pagata dal debitore, total_rec_late_fee sono tutte le commissioni ricevute, infine recoveries sono i capitali recuperati in caso di insolvenza. Nella figura 4.15a si riportano i rendimenti ponderati per il capitale ottenuti dal portafoglio alla scadenza di ogni prestito, contenuti nella tabella 4.6.

A livello aggregato sembra che alla scadenza dei prestiti erogati, il portafoglio restituisca un rendimento. Derivando da un orizzonte temporale di tre o cinque anni, va quindi diviso per il tempo al fine di ottenere una *proxy* del rendimento annuo. La media dei rendimenti annui viene riportata in tabella 4.7 e rappresentata in figura 4.15b. Sicuramente questo è dovuto principalmente dal fatto che il portafoglio è composto per oltre il 75% da prestiti in classi di rating A, B e C, ovvero prezzati correttamente, come riportato nelle figure 2.10 e 4.14, ma anche da altre forze difficilmente analizzabili come la contribuzione al rischio, probabil-

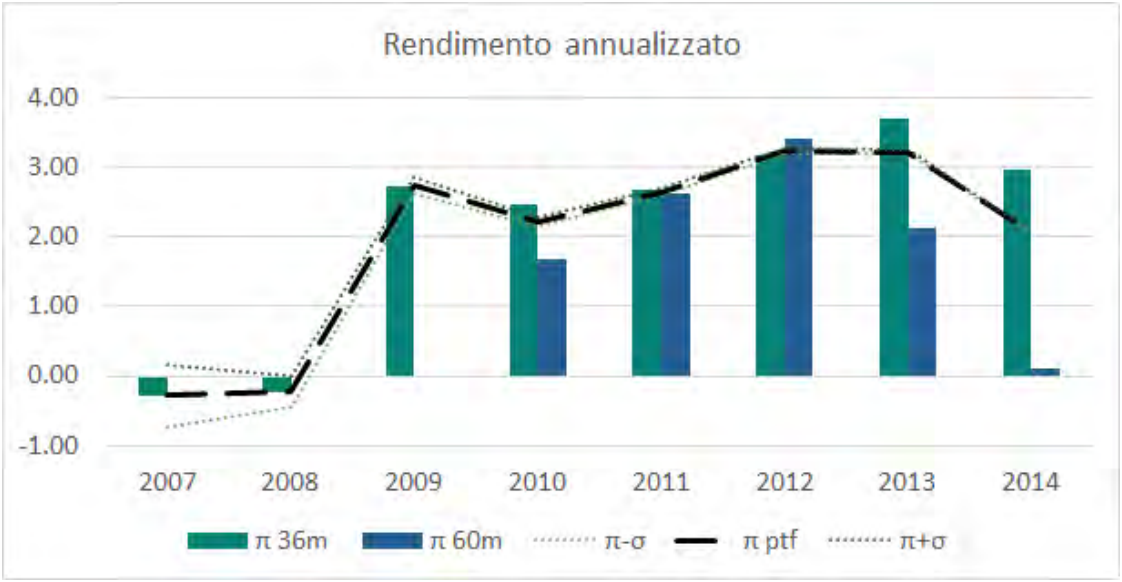
Tabella 4.7: Rendimento annualizzato dei prestiti

	Portafoglio		36mesi		60mesi	
	μ (%)	σ(%)	μ(%)	σ (%)	μ (%)	σ (%)
2007	-0.737	-0.282	0.172	-0.282		
2008	-0.446	-0.228	-0.010	-0.228		
2009	2.617	2.736	2.854	2.736		
2010	2.158	2.226	2.295	2.480	1.687	0.128
2011	2.598	2.647	2.696	2.678	2.615	0.088
2012	3.213	3.247	3.281	3.177	3.416	0.081
2013	3.208	3.229	3.249	3.713	2.133	0.045
2014	2.037	2.053	2.069	2.970	0.117	0.033

mente negativa per buona parte dei prestiti, e dalle correlazioni. Appare evidente come, al termine dell'analisi, sembra che questa tipologia di investimento non sia molto conveniente per un singolo investitore, che in media incorre in una perdita investendo capitale in questi asset. Gli istituti finanziari, invece, potendo investire grosse quantità di capitale per finanziare portafogli con un numero di controparti molto ampio, sembra possano trovare l'investimento interessante, in quanto grazie a dinamiche di contribuzione al rischio e correlazioni che mitigano il rischio assunto, in media si riesce ad ottenere un rendimento che anche se contenuto, presenta scarsa volatilità.



(a) Rendimento a scadenza dei prestiti



(b) Rendimento annualizzato dei prestiti

Figura 4.15: Analisi del rendimento a livello di portafoglio

Conclusioni

L'analisi posta in essere era indirizzata principalmente a calcolare il vantaggio economico di un portafoglio di prestiti *peer-to-peer lending*. Essendo Lending Club la piattaforma più grande a livello globale, si sono considerati i dati forniti attraverso il loro sito come approssimazione dell'intero mercato. Una volta analizzato il modello di business e il funzionamento di questo prodotto, si è proceduto ad analizzare il dataset in esame, nonché alla sua pulizia al fine di stimare i dati di input necessari all'implementazione del modello di portafoglio CreditRisk+. Sono state calcolate le probabilità di default attraverso una regressione logistica, Capitolo 2, le quali si attestano in un range dal 4% al circa 50%, evidenziando in prima battuta la rischiosità elevata delle controparti. Per validare il modello sono state, altresì confrontate con le frequenze relative dei default, evidenziando come il modello abbia la capacità di prevedere in modo corretto i default. Nel terzo capitolo sono stati stimati i *recovery rate*, parte essenziale per l'analisi completa del rischio di credito e per l'implementazione di CreditRisk+. È interessante notare che i tassi di recupero sono piuttosto elevati, ciò permette di mitigare il rischio insito nello scarso merito di credito. Grazie alla stima di queste variabili, nel capitolo 4 si è potuto analizzare la *loss distribution* di portafoglio, nonché il rendimento aggiustato per il rischio, obiettivo finale dell'elaborato. È emerso come i prestiti a cinque anni tendano ad essere generalmente sottoprezzati, mentre i prestiti a tre anni presentino un adeguato rendimento in quasi tutte le classi di rating, escluse le due peggiori. Non è agevole scoprire il motivo dell'applicazione di tassi di interesse tendenzialmente troppo bassi rispetto al rischio intrinseco, ma una prima idea potrebbe essere che il modello di *pricing* a cinque anni debba ancora essere calibrato, essendo prodotti nati qualche anno dopo rispetto agli iniziali 36 mesi. Un'altra ipotesi probabile è che sia un problema di commercializzazione, l'obiettivo principale dell'azienda è quello di attrarre volumi. Erogando prestiti a basso costo, Lending Club si mette in una posizione di vantaggio rispetto ai concorrenti, perchè in un mercato di servizi standard si applica una strategia di prezzo ed è risaputo come i prestiti, anche in banca, tendano ad essere sottoprezzati per rendere il prodotto più commerciale. Le perdite, in fondo, vengono subite dagli

investitori. Ad ogni modo, il portafoglio, nel complesso, riporta un rendimento positivo, che si attesta intorno al 2% annuo. Sicuramente questo è giustificato dal fatto che buona parte del portafoglio complessivo è composta da prestiti prezzati correttamente, infatti in media le prime tre classi, miste di prestiti a cinque anni e di prestiti a tre anni, presentano un interesse risk adjusted positivo dal 3 al 5%. Analizzando più attentamente i prestiti differenziandoli per scadenza si arriva alla conclusione che i prestiti a tre anni, i quali compongono da soli più della metà del capitale erogato, sono quasi tutti prezzati correttamente. Inoltre a livello aggregato subentrano forze, come contribuzione al rischio negativa e diversificazione, che permettono di concentrare e contenere le perdite, come si è visto nel modello di portafoglio, permettendo, su mezzo milione di prestiti e oltre 35 miliardi erogati, di avere un piccolo rendimento. Quindi, sebbene un privato non possa sostenere un investimento simile e assodato che in media sperpera capitale investendo in questi prestiti, una istituzione finanziaria, in grado di investire enormi capitali in portafogli molto ampi, potrebbe trovare l'investimento interessante.

Appendice

Appendice A

Variabili incluse nel dataset

Variabile	**Descrizione**
acc_now_delinq	Numero di account in cui il debitore è insolvente.
acc_open_past_24mths	Numero di compravendite aperte negli ultimi 24 mesi. La variabile presenta dei valori concentrati solo negli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
addr_state	Domicilio del debitore.
all_util	Saldo dal limite di credito su tutte le transazioni. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
annual_inc	Reddito annuo dichiarato.
annual_inc_joint	Reddito annuo congiunto dichiarato. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
application type	Se la richiesta è a carico di un debitore o di più debitori. La variabile non presenta valori significativi ai fini del lavoro, è concentrata prevalentemente su un unico valore.
avg_cur_bal	Saldo medio attuale di tutti i conti. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
bc_open_to_buy	Ammontare totale disponibile delle carte bancarie revolving. La variabile presenta dei valori solo negli ultimi anni, periodo troppo breve rispetto al campione per utilizzarla nell'analisi.

bc_util	Rapporto tra il saldo corrente totale e il limite massimo di credito concesso in tutti i conti bancari. La variabile si concentra solo sugli ultimi anni, mentre non è presente nella maggioranza degli anni passati, non verrà utilizzata.
chargeoff_within_12_mths	Numero di default negli ultimi 12 mesi. La variabile presenta dei valori concentrati solo negli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
collection_recovery_fee	Commissioni raccolte post default. La variabile è concentrata su un periodo temporale troppo ristretto rispetto al campione totale, quindi non verrà utilizzata.
collections_12_mths_ex_med	Numero di insoluti in 12 mesi escluse le insolvenze mediche. La variabile presenta dei valori concentrati solo negli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
delinq_2yrs	Il numero di insolvenze di oltre 30 giorni della storia creditizia del debitore negli ultimi 2 anni.
delinq_amnt	Ammontare dovuto nei conti in cui il debitore è insolvente. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
desc	Descrizione della richiesta del prestito. La variabile è soggettiva, quindi presenta troppi valori per poter essere utilizzata.
dti	Indice calcolato come divisione della somma di tutti i pagamenti dovuti sul totale delle entrate mensili.

dti_joint	DTI calcolate su entrate e debiti mensili congiunti. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
earliest_cr_line	Anno di apertura della prima linea di credito. Nel lavoro verrà convertita in uno scalare che rappresenta un numero di anni.
emp_length	Esperienza lavorativa in anni.
emp_title	Professione del debitore. La variabile presenta descrizioni soggettive fornite dai debitori, non può essere utilizzata a fini statistici.
funded_amnt	L'importo totale erogato per il prestito. Non sarà necessario utilizzare questa variabile nel lavoro, in quanto si ragionerà sull'ammontare del prestito erogato (loan_amnt).
funded_amnt_inv	L'importo totale impegnato per quel prestito. Non sarà necessario utilizzare questa variabile nel lavoro, in quanto si ragionerà sull'ammontare del prestito erogato (loan_amnt).
grade	Rating.
home_ownership	Lo stato di proprietà della casa.
id	Un codice identificativo LC univoco.
il_util	Indice del totale del saldo corrente del credito maggiore sul limite concesso su tutti i conti aperti. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.

initial_list_status	Lo stato di quotazione iniziale del prestito. I valori possibili sono - W, F - la variabile non ha significato statistico e non verrà utilizzata.
inq_fi	Numero di inchieste finanziarie. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
inq_last_12m	Numero di inchieste creditizie negli ultimi 12 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
inq_last_6mths	Il numero di inchieste negli ultimi 6 mesi (escluse le inchieste di auto e mutui ipotecari).
installment	Rata mensile dovuta.
int_rate	Tasso di interesse applicato.
issue_d	Data di emissione. Ai fini del lavoro verrà convertita in anni.
last_credit_pull_d	Il mese più recente in cui LC ha ottenuto investimenti per il prestito. La variabile presenta una concentrazione temporale troppo ristretta rispetto al campione utilizzato, quindi non verrà utilizzata.
last_pymnt_amnt	Ultimo importo ricevuto. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
last_pymnt_d	Data dell'ultima rata ricevuta. La variabile non è necessaria ai fini del lavoro.
loan_status	Stato attuale del prestito.
loan_amnt	Ammontare del prestito.

max_bal_bc	Saldo corrente dovuto in tutti i conti aperti. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
member_id	Un ID univoco assegnato da LC per il membro.
mo_sin_old_il_acct	Mesi trascorsi dall'apertura del più vecchio conto bancario. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mo_sin_old_rev_tl_op	Mesi trascorsi dall'apertura del più vecchio conto revolving. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mo_sin_rcnt_rev_tl_op	Mesi trascorsi dall'apertura del più recente conto revolving. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mo_sin_rcnt_tl	Mesi trascorsi dall'apertura del più recente conto bancario. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mort_acc	Numero di mutui. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
mths_since_last_delinq	Il numero di mesi dall'ultima insolvenza del debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mths_since_last_major_derog	Mesi trascorsi dall'ultimo peggioramento del rating. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mths_since_last_record	Mesi trascorsi dall'ultima inchiesta pubblica. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.

mths_since_rcnt_il	Mesi trascorsi dall'ultima apertura di linee di credito a rate. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mths_since_recent_bc	Mesi trascorsi dall'apertura dell'ultimo conto bancario. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
mths_since_recent_bc_dlq	Mesi trascorsi dalla più recente insolvenza su carta di credito. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
mths_since_recent_inq	Mesi trascorsi dalla più recente inchiesta. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
mths_since_recent_revol_delinq	Mesi dalla più recente insolvenza sui conti revolving. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
next_pymnt_d	Data della prossima rata. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_accts_ever_120_pd	Numero di conti insolventi da almeno 120 giorni o più. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_actv_bc_tl	Numero di conti bancari attualmente attivi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_actv_rev_tl	Numero di conti bancari solventi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.

num_bc_sats	Numero di conti revolving attualmente attivi. La variabile è concentrata solo in un periodo temporale troppo ristretto rispetto al campione considerato, quindi non verrà utilizzata.
num_bc_tl	Numero di conti bancari. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_il_tl	Numero di account rateali. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_op_rev_tl	Numeri di conti revolving aperti. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_rev_accts	Numeri di conti revolving. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_rev_tl_bal_gt_0	Numero di posizioni revolving con saldo positivo. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_sats	Numero di conti solventi. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
num_tl_120dpd_2m	Numero di conti attualmente insolventi da 120 giorni (aggiornato negli ultimi 2 mesi). I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_tl_30dpd	Numero di conti attualmente insolventi da 30 giorni (aggiornato negli ultimi 2 mesi). I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
num_tl_90g_dpd_24m	Numero di conti attualmente insolventi da almeno 90 giorni (aggiornato negli ultimi 24 mesi). I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.

num_tl_op_past_12m	Numero di conti aperti negli ultimi 12 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
open_acc	Numero di linee di credito aperte nella storia creditizia del debitore.
open_acc_6m	Numero di transazioni aperte negli ultimi 6 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
open_il_12m	Numero di rateizzazioni aperte negli ultimi 12 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
open_il_24m	Numero di rateizzazioni aperte negli ultimi 24 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
open_il_6m	Numero di transazioni in rate attive. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
open_rv_12m	Numero di transazioni revolving aperte negli ultimi 12 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
open_rv_24m	Numero di transazioni revolving aperte negli ultimi 24 mesi. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
out_prncp	Capitale ancora da rimborsare. Non è necessaria ai fini del lavoro, verrà utilizzato l'ammontare di capitale già rimborsato (total_rec_prncp).
out_prncp_inv	Capitale ancora da rimborsare agli investitori. Non è necessaria ai fini del lavoro, verrà utilizzato l'ammontare di capitale già rimborsato (total_rec_prncp).

pct_tl_nvr_dlq	Percentuale di transazioni mai state in sofferenza. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
percent_bc_gt_75	Percentuale carte di credito utilizzate per più del 75 % del limite. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
policy_code	Indica se la contropartes utilizza nuovi prodotti non disponibili pubblicamente. La variabile presenta un unico valore quindi, non essendo significativa, non verrà utilizzata.
pub_rec	Numero entrate pubbliche negative.
pub_rec_bankruptcies	Numero di fallimenti.
purpose	Motivazione della richiesta.
pymnt_plan	Indica la presenza o meno di un piano di pagamento per il prestito. La variabile presenta un solo valore, quindi non essendo significativa, verrà eliminata.
recoveries	Ammontare recuperato dopo il default.
revol_bal	Saldo totale del credito revolving. La variabile è concentrata solo sugli ultimi anni e non permette inferenza sugli anni passati.
revol_util	Tasso di utilizzo della linea revolving o importo del credito revolving che il debitore utilizza in relazione a tutti i crediti revolving disponibili. La variabile presenta dei valori solo negli ultimi anni, periodo troppo breve rispetto al campione per poterla utilizzare nell'analisi.

sub_grade	Sottorating. Ai fini dell'analisi si utilizzerà solo la classe di rating.
tax_liens	Numero di gravami fiscali. La variabile presenta dei valori concentrati solo degli ultimi anni, periodo troppo breve rispetto al campione, quindi non verrà utilizzata.
term	Scadenza del prestito.
title	Il titolo del prestito fornito dal richiedente. Non presenta valore statistico essendo una variabile soggettiva con troppi valori al suo interno.
tot_coll_amt	Ammontare degli insoluti dovuti nella storia creditizia del richiedente. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
tot_cur_bal	Saldo attuale di tutti i conti disponibili. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
tot_hi_cred_lim	Totale del credito maggiore sul limite di credito. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
total_acc	Il numero totale di linee di credito attualmente nel file di credito del debitore.
total_bal_ex_mort	Saldo totale dei crediti escluso il mutuo. La variabile si concentra solo sugli ultimi anni, mentre non è presente nella maggioranza degli anni passati, non verrà utilizzata.
total_bal_il	Saldo corrente totale di tutte le linee di credito rateizzate. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
total_bc_limit	Credito maggiore concesso attraverso una carta, sul limite di credito. La variabile si concentra solo sugli ultimi anni, mentre non è presente nella maggioranza degli anni passati, non verrà utilizzata.

total_cu_tl	Numero di operazioni finanziarie. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
total_il_high_credit_limit	Maggiore credito rateizzato sul limite di credito. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
total_pymnt	Pagamenti ricevuti fino a questo momento per l'importo totale finanziato. Non è necessaria ai fini del lavoro, verrà utilizzato l'ammontare di capitale già rimborsato (total_rec_prncp).
total_pymnt_inv	Pagamenti ricevuti dagli investitori fino a questo momento per l'importo totale finanziato. Non è necessaria ai fini del lavoro, verrà utilizzato l'ammontare di capitale già rimborsato (total_rec_prncp).
total_rec_int	Ammontare di interesse ricevuti fino a questo momento.
total_rec_late_fee	Commissioni per i ritardi ricevute fino a questo momento.
total_rec_prncp	Capitale rimborsato fino a questo momento.
total_rev_hi_lim	Massimo credito revolving sul limite di credito. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
url	URL per la pagina LC con i dati dell'elenco. Non è necessaria.
verification_status	Indica se il reddito annuo dichiarato è stato verificato oppure no. La variabile presenta un valore solo quindi, non essendo significativa, non verrà utilizzata.

verified_status_joint	Indica se il reddito congiunto annuo dichiarato è stato verificato oppure no. Non risulta necessaria.
zip_code	Le prime tre cifre del codice postale. La variabile non presenta valori utilizzabili, in quanto parzialmente omessi al fine di garantire la privacy.
revol_bal_joint	Somma del saldo dei crediti revolving dei co-mutuatari, al netto dei saldi duplicati. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_earliest_cr_line	Anno della richiesta della prima linea di credito del co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_inq_last_6mths	Crediti insoluti negli ultimi 6 mesi per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_mort_acc	Numero di mutui a carico del co-debitore al momento della richiesta. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_open_acc	Numero di transazioni aperte al momento della richiesta per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_revol_util	Indice del saldo corrente del credito maggiore sul limite di credito di tutti i conti revolving per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_open_il_6m	Numero di transazioni rateizzate attive al momento della richiesta per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_num_rev_accts	Numero di conti revolving al momento della richiesta per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.

sec_app_chargeoff_within_12_mths	Numero di default negli ultimi 12 mesi per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_collections_12_mths_ex_med	Numero di insoluti in 12 mesi escluse le insolvenze mediche per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.
sec_app_mths_since_last_major_derog	Mesi trascorsi dall'ultimo peggioramento del rating per il co-debitore. I missing values sono maggiori del 40% quindi la variabile non sarà utilizzata.

Appendice B

Appendice creditrisk+

Il default avviene dopo una serie di eventi che rendono impossibile prevedere l'esatto momento del default e il numero di controparte defaultate. In questa sezione si riporta la derivazione teorica alla base del rischio di credito utilizzata per la costruzione del modello, ripresa da *CreditRisk+: A credit risk management framework* dalla Boston.

B.1 Default

Si consideri un portafoglio composto da N controparti. In linea con le assunzioni del capitolo 4 a pagina 84, ogni controparte ha una probabilità di default con un orizzonte temporale di un anno, che è data come conosciuta. Quindi:

$$p_A = PD\ annuale\ della\ controparte\ A \tag{B.1}$$

Per analizzare la distribuzione delle perdite dell'intero portafoglio va introdotta la funzione generatrice dei dati in termini variabile ausiliaria z come:

$$F(z) = \sum_{n=0}^{\infty} p(n\ defaults) z^n \tag{B.2}$$

nel senso che un debitore può essere insolvente o solvente. La funzione generatrice dei dati per una singola controparte può essere agilmente esplicitata:

$$F_A(z) = 1 - p_A + p_A z = 1 + p_A(z-1) \tag{B.3}$$

Come conseguenza dell'indipendenza tra gli eventi del default, la funzione generatrice dei dati dell'intero portafoglio è una produttoria di tutte le funzioni

individuali:

$$F(z) = \prod_A F_A(z) = \prod_A \big(1 + p_A(z-1)\big) \tag{B.4}$$

$$\equiv logF(z) = \sum_A log\big(1 + p_A(z-1)\big) \tag{B.5}$$

Adesso supponiamo, che le probabilità di default delle controparti siano uniformemente basse. Questa caratteristica è spesso presente nei portafogli creditizi, quindi il logaritmo può essere sostituito con:

$$log\big(1 + p_A(z-1)\big) = p_A(z-1) \tag{B.6}$$

Quindi il limite dell'equazione (B.5) diventa:

$$\begin{aligned} F(z) =& e^{\sum_A p_A(z-1)} = e^{\mu(z-1)} \\ \text{dove } \mu =& \sum_A p_A \end{aligned} \tag{B.7}$$

e rappresenta il numero atteso di default all'interno del portafoglio in un anno. Per capire la distribuzione sottostante a questa funzione generatrice dei dati possiamo espandere $F(z)$ nella sua serie di Taylor:

$$\begin{aligned} F(z) &= e^{\mu(z-1)} \\ &= e^{-\mu} e^{\mu z} \\ &= \sum_{n=0}^{\infty} \frac{e^{-\mu}\mu^n}{n!} z^n \end{aligned} \tag{B.8}$$

Quindi se le probabilità di default individuali sono basse, sebbene non necessariamente uguali, si deduce, dall'equazione (B.8), che la probabilità di generare n default nel portafoglio è:

$$P(n\ defaults) = \frac{e^{-\mu}\mu^n}{n!} \tag{B.9}$$

Ovvero la distribuzione di Poisson utilizzata per le assunzioni iniziali. Va notato che:

- La distribuzione ha un solo parametro, la frequenza attesa dei default. La distribuzione non dipende dal numero delle esposizioni in portafoglio o dalle probabilità di default individuali.
- Non è necessario che tutte le controparti abbiano PD uguali, ogni debitore

Tabella B.1: Notazione per la suddivisione delle esposizioni

Variabile	Simbolo
Controparte	A
Esposizione	L_A
Probabilità di default	P_A
Perdita attesa	λ_A

può mantenere la sua se vi sono informazioni sufficienti.

La distribuzione di Poisson teorica ha media μ e deviazione standard pari a $\sqrt{\mu}$. Però i dati presentano una deviazione standard sicuramente maggiore, quindi i tassi di default non sono fissi, come prima assunto.

B.2 Distribuzione dei default

Abbiamo ottenuto la distribuzione della frequenza dei default annuali nel portafoglio, ora si può procedere a derivare la probabilità di subire determinati livelli di perdita. Le due distribuzioni sono diverse in quanto lo stesso livello di perdita potrebbe derivare, in ugual misura, dall'inadempienza di una singola esposizione molto ingente o dal default di un gruppo di esposizioni più piccole. Ne deriva che i diversi importi delle esposizioni determinano una *loss distribution* che non può essere generalizzata in una distribuzione di Poisson, inoltre, l'informazione sulla distribuzione delle diverse esposizioni è essenziale per la distribuzione complessiva di portafoglio. Si può descrivere la distribuzione generale perché la sua funzione di generazione di probabilità ha una forma chiusa, semplice da calcolare.

Il primo passo per arrivare alla *loss distribution* di portafoglio, è suddividere le esposizioni in gruppi secondo multipli o sottomultipli di un'unità determinata in modo arbitrario tenuto conto della grandezza di scala delle esposizioni. In questo modo si riduce significativamente l'ammontare dei dati necessari al calcolo, sebbene si introduca un elemento di approssimazione. Ad ogni modo è un approssimazione talmente piccola che può essere trascurata, in quanto il numero di esposizioni è molto cospicuo rispetto all'ampiezza dei raggruppamenti che si effettuano e l'ammontare preciso dell'esposizione nel portafoglio non è un elemento critico per il rischio totale. Nella tabella (B.1) si riporta la notazione che si utilizzerà nello sviluppo teorico del modello. Al fine di suddividere le esposizioni in gruppi, viene determinata una base unitaria L secondo quanto sopra indicato. Per ogni controparte possono adesso essere calcolate l'esposizione (v_A) e la perdita attesa (ε_A)

Tabella B.2: Notazione per l'aggregazione del portafoglio

Variabile	Simbolo
Esposizione comune nella classe j in unità di L	v_j
Perdita attesa nella classe j in unità di L	ε_j
Frequenza di default attesa nella classe j	μ_j

come multipli dell'unità. Per ogni controparte A:

$$L_A = L * v_A \tag{B.10}$$

$$\lambda_A = L * \varepsilon_A \tag{B.11}$$

L'azione fondamentale è l'arrotondamento di ogni esposizione v_A all'intero maggiore più vicino, portando ogni ammontare L_A al più vicino multiplo di L. Se l'unità L è stata scelta in modo appropriato, si avranno un numero relativamente piccolo di classi rispetto al numero alto delle esposizioni iniziali. Ci saranno quindi pochi valori per v_A, che saranno condivisi da più contropartì. Il portafoglio può essere adesso diviso in m classi, indicizzate in j, da cui deriva la notazione in tabella (B.2). La relazione per esprimere la perdita attesa in termini di probabilità di default è:

$$\varepsilon_j = v_j * \mu_j \tag{B.12}$$

$$\text{quindi } \mu_j = \frac{\varepsilon_j}{v_j} \tag{B.13}$$

$$= \sum_{A:v_A=v_j} \frac{\varepsilon_A}{v_A} \tag{B.14}$$

$$\text{dove } \mu = \sum_{j=1}^{m} \mu_j \tag{B.15}$$

$$= \sum_{j=1}^{m} \frac{\varepsilon_j}{v_j} \tag{B.16}$$

Va notato che, avendo arrotondato v_j all'intero, la perdita attesa ε_A potrebbe essere inficiata, a meno che non venga considerato un aggiustamento per l'arrotondamento, applicato alla frequenza attesa dei default μ_j. Secondo la notazione già definita, si possono calcolare la perdita attesa di portafoglio (ε) e la deviazione standard

(σ), espresse in unità di L come:

$$\varepsilon = \sum_{j=1}^{m} \varepsilon_j \tag{B.17}$$

$$\sigma^2 = \sum_{j=1}^{m} v_j * \varepsilon_j \tag{B.18}$$

L'errore introdotto con l'utilizzo dell'arrotondamento in classi può essere espresso:

$$\begin{aligned} \widetilde{v_j} &= v_j + \tau_j \\ \text{dove} 0 \leq &\tau_j \leq 1 \end{aligned} \tag{B.19}$$

Assunto che le esposizioni sono arrotondate per eccesso,τ_j è un valore positivo. Dalla notazione sopra riportata, si può notare che la perdita attesa non è inficiata dall'arrotondamento effettuato, in quanto la sua derivazione è indipendente. Per la deviazione standard abbiamo:

$$\sigma^2 \leq \widetilde{\sigma^2} = \sum_{j=1}^{m} \widetilde{v_j} * \varepsilon_j = \sigma^2 + \sum_{j=1}^{m} \tau_j * \varepsilon_j \leq \sigma^2 + \sum_{j=1}^{m} \varepsilon_j = \sigma^2 + \varepsilon \tag{B.20}$$

dove ε è la perdita attesa del portafoglio. Prendendo le radici quadrate e trascurando i termini di ordine superiori al secondo nella serie di Taylor, otteniamo:

$$\sigma \leq \widetilde{\sigma} \leq \sigma\Big(1 + \frac{\varepsilon}{2\sigma^2}\Big) = \sigma + \frac{\varepsilon}{2\sigma} \tag{B.21}$$

Per un portafoglio reale, la perdita attesa ε e la quantità 2σ sono dello stesso ordine. Si può concludere, quindi, che la perdita attesa calcolata dal modello non è inficiata dalla creazione delle classi e la deviazione standard è sovrastimata di una quantità pari all'unità scelta.

B.3 Volatilità dei tassi di default

Va ricordato che, a causa della volatilità, c'è la possibilità che i tassi di default stimati siano maggiori delle aspettative, questo comporta una maggiore probabilità di incorrere in perdite estreme. Ne deriva che:

i Le probabilità di default stimate possono variare nel tempo anche se le controparti hanno un merito di credito analogo.

ii La variabilità delle probabilità di default può essere ricondotta all'incertezza dei fattori sottostanti, come il ciclo economico, che influenzano la capacità di ripagare dei debitori affidati.

iii Il cambiamento dei fattori macroeconomici non comportano con certezza l'insolvenza.

Per questo motivo l'analisi settoriale risulta molto importante, in quanto l'incertezza derivante da questi fattori potrebbe influenzare un ampio numero di controparti.

Analisi di settore L'analisi di settore permette di quantificare l'influenza dei fattori esterni sulle frequenze di default. Per esempio, l'economia della posizione geografica del domicilio può influenzare la parte di controparti che vive in quello stato ma può non intaccare il merito di credito del resto dei debitori. CrediRisk+ permette di misurare l'influenza dei fattori esterni, suddividendo le controparti secondo l'appartenenza ponderata ai settori in esame. Per suddividere il portafoglio in settori e considerare la volatilità intrinseca di ogni fattore, deve essere introdotta ulteriore notazione: $S_k : 1 \leq k \leq n$, per i settori, ognuno dei quali deve essere visto come un sottoinsieme di controparti. CreditRisk+ assume che un unico fattore influenzi ogni settore, così che questo *underlying* spieghi tutta la variabilità delle frequenze di default misurate nel settore stesso, modellizzandolo come una variabile random con media μ_k e deviazione standard σ_k, definita per ogni settore. La deviazione standard rifletterà il grado in cui le probabilità di default nel portafoglio sono soggette a divergenza dai loro livelli medi. Ad esempio, in un settore costituito da un numero elevato di debitori con bassa qualità creditizia, il tasso medio predefinito potrebbe essere del 5 % annuo e la deviazione standard del tasso di default effettivo potrebbe essere una quantità simile. Quindi ci sarà una probabilità di incremento sostanziale della PD che potrebbe divenire, ad esempio, 10 % anziché 5 %. La tabella (B.3) riporta la notazione necessaria per l'implementazione del modello considerando l'analisi di settore. In particolare, viene introdotta una variabile casuale x_k, che rappresenta la frequenza di default del settore. La media di x_k sarà μ_k, mentre la deviazione standard σ_k. Per ogni settore, quindi, vengono richiesti i dati riportati in tabella (B.4) La media μ_k è legata alla perdita attesa secondo la relazione:

$$\mu_k = \sum_{j=1}^{m(k)} \frac{\varepsilon_j^{(k)}}{v_j^{(k)}} \tag{B.22}$$

Per ogni settore, oltre alla frequenza di default attesa definita nell'equazione (B.22),va specificata una deviazione standard della media μ_k, la cui derivazione risulta age-

Tabella B.3: Notazione per la suddivisione del portafoglio con analisi settoriale

Settori	$S_k : 1 \leq k \leq n$
Variabile random che rappresenta la media della frequenza dei default	x_k
Media di x_k, frequenza di lungo periodo	μ_k
Deviazione standard di x_k	σ_k

Tabella B.4: Dati richiesti per il modello di portafoglio con analisi settoriale

Dati dell'esposizione per settore	Notazione precedente	Nuova notazione
Unità base dell'esposizione	L	L
Dimensione dell'esposizione in unità	$L_j = Lv_j$	$L_j^{(k)} = Lv_j^{(k)}$
	$1 \leq j \leq m$	$1 \leq k \leq n; 1 \leq j \leq m(k)$
Perdita attesa in unità per classe	$\lambda_j = L\varepsilon_j$	$\lambda_j^{(k)} = L\varepsilon_j^{(k)}$
	$1 \leq j \leq m$	$1 \leq k \leq n; 1 \leq j \leq m(k)$

vole espandendo l'equazione per il calcolo della media come sommatoria:

$$\mu_k = \sum_A \frac{\varepsilon_A}{v_A} \tag{B.23}$$

Dove la sommatoria si estende a tutte le controparti A nel settore k e la relazione $\varepsilon_A/v_A = p_A$ esprime la probabilità di default media della controparte nel periodo considerato. Per ottenere anche σ_k, si assume che sia nota la deviazione standard per ogni controparte (σ_A), facilmente derivabile se si considera dipendente solo dal merito di credito della controparte, da cui si deriva agilmente σ_k attraverso un processo di calcolo della media. A questo punto si può modellizzare la distribuzione random x_k, che sarà una proporzione della frequenza di default della classe di controparti A e la cui media sarà dipendente dalla probabilità di default delle controparti. Si può, quindi, esprimere questa assunzione identificando con x_A la probabilità di default random di una controparte A:

$$x_A = \frac{\varepsilon_A}{v_A}\frac{x_k}{\mu_k} \tag{B.24}$$

Si noti che la media di x_A è correttamente specificata come p_A, in particolare:

$$\sum_A \sigma_A = \sum_A \frac{\varepsilon_A}{v_A}\frac{\sigma_k}{\mu_k} = \sigma_k \frac{1}{\mu_k}\sum_A \frac{\varepsilon_A}{v_A} = \sigma_k \tag{B.25}$$

usata nell'equazione (B.23). La sommatoria comprende tutte le controparti appartenenti al settore. Abbiamo stimato la deviazione standard del settore in modo da

rispettare questa condizione. Perciò la deviazione standard della media dei tassi di default del settore è stimata come somma delle deviazioni standard per ogni controparte appartenente al settore. Un modo più intuitivo di derivare σ_k è considerarlo come media della deviazioni standard di ogni controparte pesata per la sua contribuzione al tasso di default:

$$\begin{aligned} \frac{\sigma_k}{\mu_k} &= \frac{\sum_A \sigma_A}{\sum_A p_A} \\ &= \frac{\sum_A p_A \left(\frac{\sigma_A}{p_A}\right)}{\sum_A p_A} \end{aligned} \tag{B.26}$$

Secondo l'esperienza passata, l'indice σ_A/p_A è tipicamente nell'ordine di uno, così la deviazione standard della frequenza dei default osservata in un anno, nella stessa classe di rating, è tipicamente della stessa grandezza d'ordine della media annuale dei default. L'equazione (B.26) dimostra che, come ci si aspetta, questo sia vero per ogni settore. In assenza di dati dettagliati, la stima dell'indice (B.3) per una controparte specifica, può essere sostituito da un indice fisso, denominato ω_k, che riduce l'equazione (B.26) in:

$$\sigma_k = \omega_k * \mu_k \tag{B.27}$$

Se la natura del settore rende più appropriato stimare la deviazione standard σ_k direttamente, dovrebbe essere equivalente a stimare l'indice fisso ω_k direttamente.

Tasso di default con probabilità variabili Ottenuta la distribuzione dei default attraverso la stima della funzione generatrice dei dati con probabilità di default fisse (equazione (B.7)), dobbiamo rigenerarla inserendo probabilità di default variabili. Riprendendo la funzione generatrice dei dati come nell'equazione (B.2):

$$F(z) = \sum_{n=0}^{\infty} p(n\ defaults) z^n \tag{B.28}$$

possiamo riscriverla come produttoria dei settori, essendo gli stessi indipendenti:

$$F(z) = \prod_{k=1}^{n} F_k(z) \tag{B.29}$$

Per ogni settore si deve, quindi, determinare $F(z)$. Nella tabella B.3, abbiamo definito la media dei default in un settore k come una variabile random (x_k) con media μ_k e deviazione standard σ_k. Condizionatamente al valore x_k, possiamo

scrivere la funzione generatrice dei dati per la distribuzione dei default come:

$$F_k(z)\Big|[x_k = x] = e^x(z-1) \tag{B.30}$$

derivata dall'equazione (B.7). Supponendo che x_k ha una funzione di densità di probabilità pari a $f_k(x)$, possiamo scrivere:

$$P(x \leq x_k \leq x + dx) = f_k(x)dx \tag{B.31}$$

Quindi, la funzione generatrice dei dati per gli eventi dei default in un settore è la media della funzione generatrice dei dati condizionale, dell'equazione (B.30), calcolata su tutti i possibili valori medi di tassi di default, computata come segue:

$$\begin{aligned} F_k(z) =& \sum_{n=0}^{\infty} P(n\ defaults)z^n = \\ =& \sum_{n=0}^{\infty} z^n \int_{x=0}^{\infty} P(n\ defaults\big|x)f(x)dx = \\ =& \int_{x=0}^{\infty} e^{x(z-1)} f(x)dx \end{aligned} \tag{B.32}$$

Al fine di ottenere una formula esplicita per la funzione generatrice dei dati, si deve scegliere una distribuzione per X_k. Si deve formulare un' assunzione fondamentale: che x_k abbia una distribuzione Gamma con media μ_k e deviazione standard σ_k. La *Gamma distribution* è scelta grazie alla possibilità di trattarla analiticamente attraverso due parametri $\big(\Gamma(\alpha, \beta)\big)$. Riprendendo l'equazione (B.31), si può esplicitare la funzione di densità di probabilità, distribuzione random di X, come:

$$\begin{aligned} P(x \leq x_k \leq x + dx) =& f_k(x)dx = \\ =& \frac{1}{\beta_\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1} dx \\ \text{dove } \Gamma(\alpha) =& e^{-x}x^{\alpha-1}dx \text{ è la funzione Gamma} \end{aligned} \tag{B.33}$$

La distribuzione Gamma è descritta interamente da due parametri, media e deviazione standard definite come:

$$\mu = \alpha\beta \tag{B.34}$$

$$\sigma^2 = \alpha\beta^2 \tag{B.35}$$

Quindi, per il settore k, i parametri della distribuzione Gamma sono:

$$\alpha_k = \mu_k^2/\sigma_k^2 \tag{B.36}$$
$$\beta_k = \sigma_k^2/\mu_k \tag{B.37}$$

Con la scelta della distribuzione Gamma per la funzione $F(x)$, l'espressione per la funzione generatrice delle probabilità:

$$F_k(z) = \int_{x=0}^{\infty} e^{x(z-1)} f(x) dx \tag{B.38}$$

si può calcolare direttamente per sostituzione:

$$\begin{aligned} F_k(z) =& \int_{x=0}^{\infty} e^{x(z-1)} \frac{e^{-x/\beta} x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} dx = \\ =& \frac{1}{\beta_\alpha \Gamma(\alpha)} \int_{y=0}^{\infty} \Big(\frac{y}{\beta^{-1} + 1 - z} \Big)^{\alpha-1} e^{-y} \frac{dy}{\beta^{-1} + 1 - z} = \\ =& \frac{\Gamma(\alpha)}{\beta^\alpha \Gamma(\alpha)(1 + \beta^{-1} - z)^\alpha} = \\ =& \frac{1}{\beta^\alpha (1 + \beta^{-1} - z)^\alpha} \end{aligned} \tag{B.39}$$

Riarrangiando per un singolo settore, si ottiene:

$$\begin{aligned} F_k(z) =& \left(\frac{1 - p_k}{1 - p_k z} \right)^{\alpha_k} \\ \text{dove } \; p_k =& \frac{\beta_k}{1 + \beta_k} \end{aligned} \tag{B.40}$$

Questa è la funzione generatrice dei dati della distribuzione dei default per il settore k, da cui è possibile derivare la distribuzione della frequenza dei default sottostante la funzione generatrice dei dati. Espandendo $F_k(z)$ nella serie di Taylor:

$$\begin{aligned} F_k(z) =& (1 - p_k)^{\alpha_k} \sum_{n=1}^{\infty} \left(\begin{bmatrix} n + \alpha_k - 1 \\ n \end{bmatrix} \right) p_k^n z^n \\ \text{da cui } \; P(n\ defaults) =& (1 - p_k)^{\alpha_k} \left(\begin{bmatrix} n + \alpha_k - 1 \\ n \end{bmatrix} \right) p_k^n z^n \end{aligned} \tag{B.41}$$

Questa formula può essere identificata come la densità della distribuzione binomiale negativa. Riassumendo, quindi, la funzione generatrice dei dati a livello di

portafoglio è data da:

$$F(z) = \prod_{k=1}^{n} F_k(z) = \prod_{k=1}^{n} \left(\frac{1 - p_k}{1 - p_{kz}} \right)^{\alpha_k} \tag{B.42}$$

dove i parametri α_k ,β_k e p_k sono calcolati come:

$$\alpha_k = \mu_k^2 / \sigma_k^2 \tag{B.43}$$

$$\beta_k = \sigma_k^2 / \mu_k \tag{B.44}$$

$$\text{e } p_k = \frac{\beta_k}{1 + \beta_k} \tag{B.45}$$

Quindi, sebbene la distribuzione dei default a livello di portafoglio non sia riconducibile ad una distribuzione binomiale negativa, si può comunque approssimare ad una sommatoria di distribuzioni binomiali negative indipendenti.

Perdita con tassi di default variabili

L'equazione (B.42) restituisce le informazioni necessarie per i tassi di default del portafoglio, ma per passare alle perdite causate dalle insolvenze si devono computare nel calcolo anche le informazioni sulle esposizioni, introducendo la funzione generatrice dei dati per le perdite del portafoglio:

$$G(z) = \sum_{n=0}^{\infty} p(perdite\ di\ portafoglio = n * L) z^n \tag{B.46}$$

Come per la distribuzione dei default, l'indipendenza settoriale permette una produttoria della funzione generatrice dei dati.

$$G(z) = \prod_{k=1}^{n} G_k(z) \tag{B.47}$$

dove $G_k(z)$ è la funzione generatrice delle probabilità per il settore k, $1 \leq k \leq n$. Possiamo definire un polinomio $P_k(z)$, $1 \leq k \leq n$:

$$\begin{aligned} P_k(z) =& \frac{\sum_{j=1}^{m(k)} \left(\frac{\varepsilon_j^{(k)}}{v_j^{(k)}} \right) z^{v_j^{(k)}}}{\sum_{j=1}^{m(k)} \left(\frac{\varepsilon_j^{(k)}}{v_j^{(k)}} \right)} \\ =& \frac{1}{\mu_k} \sum_{j=1}^{m(k)} \left(\frac{\varepsilon_j^{(k)}}{v_j^{(k)}} \right) z^{v_j^{(k)}} \end{aligned} \tag{B.48}$$

in cui viene utilizzata l'equazione (B.22). La componente $P_k(z)$ rappresenta la connessione tra l'evento del default e le perdite, in quanto esiste la relazione per ogni settore:

$$G_k(z) = F_k\big(P_k(z)\big) \tag{B.49}$$

infatti espandendo la formula (B.49) come somma di individui appartenenti al settore k, si ottiene:

$$\begin{aligned} P_k(z) =& \frac{\sum_A \frac{\varepsilon_A}{v_A} z^{v_A}}{\sum_A \frac{\varepsilon_A}{v_A}} = \\ =& \frac{1}{\mu_k} \sum_A \frac{\varepsilon_A}{v_A} z^{v_A} \end{aligned} \tag{B.50}$$

Riprendendo l'equazione (B.24), abbiamo:

$$e^{-\sum_A x_A + \sum_A x_A z^{v_A}} = e^{-\sum_A x_A (z^{v_A}-1)} = e^{x_k/\mu_k \sum_A \varepsilon_A/v_A (z_{v_A}-1)} = e^{x_k(P_z-1)} \tag{B.51}$$

dove la parte sinistra dell'equazione (B.51) è la funzione generatrice delle probabilità, in cui ogni controparte (A) ha probabilità di default x_A. Proprio come nell'equazione (B.38), che esprime $F_k(z)$ come integrale della funzione di generazione di probabilità di Poisson, un argomento di probabilità condizionale mostra che $G_k(z)$ è l'integrale del lato sinistro dell'equazione (B.51) nello spazio dei valori x_k. Quindi:

$$G_k(z) = \sum_{n=0}^{\infty} z^n \int_{x_k=0}^{\infty} P(loss\ of\ nL \big| x_k) f_k(x_k) dx_k = \tag{B.52}$$

$$= \int_{x_k=0}^{\infty} e^{\sum_A x_A (z^{v_A}-1)} f_k(x_k) dx_k = \tag{B.53}$$

$$= \int_{x_k=0}^{\infty} e^{x_k(P_k(z)-1)} f_k(x_k) dx_k \tag{B.54}$$

Sostituendo nell'equazione (B.40) e producendo alla produttoria di ogni settore, si ottiene:

$$G(z) = \prod_{k=1}^{n} G_k(z) = \tag{B.55}$$

$$= \prod_{k=1}^{n} \left(\frac{1-p_k}{1 - p_k/\mu_k \sum_{j=1}^{m(k)} \varepsilon_j^{(k)}/v_j^{(k)} z^{v_j^{(K)}}} \right)^{\alpha_k} \tag{B.56}$$

L'equazione (B.56) è la forma della funzione generatrice delle probabilità più vicina alla reale. Per computare la distribuzione delle perdite da questa formulazione è necessario derivare una relazione ricorrente. Supponiamo a livello generale una serie:

$$G(z) = \sum_{n=0}^{\infty} A_n z^n \tag{B.57}$$

si definisca $G(z)$ come una funzione che soddisfa l'equazione differenziale:

$$\frac{d}{dz}\big(logG(z)\big) = \frac{1}{G(z)}\frac{dG(z)}{dz} = \frac{A(z)}{B(z)} \tag{B.58}$$

in cui A e B sono polinomi dati rispettivamente da:

$$A(z) = a_0 + \cdots + a_r z^r \tag{B.59}$$

$$B(z) = b_0 + \cdots + b_r z^r \tag{B.60}$$

In altre parole, si richiede che la derivata logaritmica di $G(z)$ sia una funzione razionale. Quindi, i termini della serie (B.57) soddisfano la relazione ricorrente seguente:

$$A_{n+1} = \frac{1}{b_0(n+1)}\left(\sum_{i=0}^{min(r,n)} a_i A_{n-1} - \sum_{j=0}^{min(s-1,n-1)} b_{j+1}(n-1)A_{n-j}\right) \tag{B.61}$$

Per dimostrare questo, si può riarrangiare l'equazione (B.58) come:

$$B(z)\frac{dG}{dz} = A(z)G \tag{B.62}$$

Differenziando G secondo ogni suo termine, si arriva a:

$$\left(\sum_{j=0}^{s} b_j z^j\right)\left(\sum_{n=0}^{\infty}(n+1)A_{n+i}z^n\right) = \left(\sum_{i=0}^{r} a_i z^i\right)\left(\sum_{n=0}^{\infty} A_n z^n\right) \tag{B.63}$$

Per $n \geq 0$ il termine in z^n a sinistra e destra rappresentano rispettivamente:

$$\sum_{j=0}^{min(s,n)} b_j(n+1-j)A_{n+1-j} \tag{B.64}$$

$$\sum_{i=0}^{min(r,n)} a_i A_{n-1} \tag{B.65}$$

che riarrangiando diventa:

$$b_0(n+1)A_{n+1} = \sum_{i=0}^{min(r,n)} a_i A_{n-i} - \sum_{j=1}^{min(s,n)} b_j(n+1-j)A_{n+1-j} \tag{B.66}$$

o equivalentemente:

$$b_0(n+1)A_{n+1} = \sum_{i=0}^{min(r,n)} a_i A_{n-i} - \sum_{j=0}^{min(s-1,n-1)} b_{j+1}(n-j)A_{n-j} \tag{B.67}$$

come richiesto. Riprendendo l'equazione (B.56), in cui abbiamo derivato la funzione generatrice dei dati delle perdite, possiamo derivarla rispetto a z, trovando:

$$\frac{G'(z)}{G(z)} = \sum_{k=1}^{n} \frac{G'_k(z)}{G_k(z)} = \sum_{k=1}^{n} \frac{\frac{p_k \alpha_k}{\mu_k} \sum_{j=1}^{m(k)} \varepsilon_j^{(k)} z^{v_j^{(k)}-1}}{1 - \frac{p_k}{\mu_k} \sum_{j=1}^{m(k)} \frac{\varepsilon_j^{(k)}}{v_j^{(k)} z^{v_j^{(k)}}}} \tag{B.68}$$

In questo modo si riesce ad esprimere $\frac{G'(z)}{G(z)}$ come una funzione razionale. Secondo questa formulazione, dopo il calcolo polinomiale di $A(z)$ e $B(z)$, si può riscrivere:

$$\frac{A(z)}{B(z)} = \sum_{k=1}^{n} \frac{\frac{p_k \alpha_k}{\mu_k} \sum_{j=1}^{m(k)} \varepsilon_j^{(k)} z^{v_j^{(k)}-1}}{\frac{p_k \alpha_k}{\mu_k} \sum_{j=1}^{m(k)} \varepsilon_j^{(k)} z^{v_j^{(k)}-1}} \tag{B.69}$$

Questa formula permette un agevole calcolo della distribuzione dell'ammontare delle perdite, attraverso una relazione ricorrente. Si noti che l'equazione (B.69) permette di fare ciò attraverso una semplice sommatoria delle *loss given default* attraverso i settori. Nella derivazione di CrediRisk+ la funzione generatrice delle probabilità per la distribuzione delle perdite, è suddivisa per settore, ognuno dei quali è un gruppo di controparti. Questa situazione corrisponde alla suddivisione delle controparti in classi, ognuna delle quali è influenzata da un fattore ed è indipendente dalle altre. A questo punto si può considerare una situazione più generale in cui, un numero di fattori spiega la volatilità e la frequenza dei default nel portafoglio, senza che il default di una controparte dipenda necessariamente da un fattore. In questa situazione più generale, CreditRisk+ incorpora la situazione allo stesso modo, sostituendo il concetto di settore con rischio sistematico. Riprendendo la funzione generatrice delle probabilità espressa come produttoria

dei settori, (equazione (B.56)):

$$G(z) = \prod_{k=1}^{n} G_k(z) = \tag{B.70}$$

$$= \prod_{k=1}^{n} \left(\frac{1 - p_k}{1 - {}^{p_k}/_{\mu_k} \sum_{j=1}^{m(k)} {}^{\varepsilon_j^{(k)}}/_{v_j^{(k)}} z^{v_j^{(K)}}} \right)^{\alpha_k} \tag{B.71}$$

si può riscrivere come integrale multiplo:

$$G(z) = \int_{x_1=0} \cdots \int_{x_n=0} e^{\sum_{k=1}^{n} x_k (P_k(z)-1)} f_k(x_k) dx_k \tag{B.72}$$

L'equazione (B.72) rappresenta la funzione di densità di una distribuzione di Poisson per ogni gruppo di valori con media x_k, $1 \leq l \leq n$. La funzione di densità è quindi integrata sullo spazio di tutti i possibili stati, rappresentati dai valori di x_k pesati per la propria funzione di densità. Riprendendo l'equazione (B.51), possiamo esaminare l'esponente dell'integrale nella sua forma equivalente:

$$\sum_{k=1}^{n} = x_k(P_k(z) - 1) = \sum_{k=1}^{n} \sum_{A} \frac{x_k}{\mu_k} \frac{\varepsilon_A}{v_A} (z^{v_A} - 1) = \sum_{A,k} \delta_{A,k} \frac{x_k}{\mu_k} \frac{\varepsilon_A}{v_A} (z^{v_A} - 1) \tag{B.73}$$

in cui viene utilizzata la notazione:

$$\delta_{A,k} = \begin{cases} 1 & A \notin k \\ \\ 0 & A \in k \end{cases} \tag{B.74}$$

Per generalizzare il concetto di un singolo settore nella situazione in cui una singola controparte è influenzata da più di un fattore x_k, si deve sostituire la funzione *delta* con l'allocazione delle controparti tra i settori, rispettando la condizione:

$$\theta_{Ak} : \sum_{k=1}^{n} \theta_{Ak} = 1 \tag{B.75}$$

L'allocazione θ_{Ak} rappresenta la situazione in cui la probabilità di default di una controparte A è influenzata da diversi fattori. Generalizzando l'equazione (B.73) si ottiene:

$$\sum_{k=1}^{n} x_k \Big(P_k(z) - 1 \Big) = \sum_{A,k} \theta_{Ak} \frac{x_k}{\mu_k} \frac{\varepsilon_A}{v_A} (z^{v_A} - 1) \tag{B.76}$$

in cui ogni controparte contribuisce con un termine:

$$\begin{aligned} & x_A(z^{v_A} - 1) \\ & \text{dove :} \\ & x_A = \frac{\varepsilon_A}{v_A} \sum_{k=1}^{n} \theta_{Ak} \frac{x_k}{\mu_k} \end{aligned} \tag{B.77}$$

L'equazione (B.50) è sostituita da:

$$\begin{aligned} & P_k(z) = \frac{1}{\mu_k} \sum_A \theta_{Ak} \frac{\varepsilon_A}{v_A} z^{v_A} \\ & \text{dove :} \\ & \mu_k = \sum_A \theta_{Ak} \frac{\varepsilon_A}{v_A} \end{aligned} \tag{B.78}$$

Il numero θ_{Ak} rappresenta la proporzione dei settori che influenzano il merito di credito della controparte A. La media di ogni settore è la somma della contribuzione di ogni controparte, ora pesata per l'allocazione θ_{Ak}. Poiché:

$$\mu_k = \sum_A \theta_{Ak} \mu_A \tag{B.79}$$

Ora, analogamente all'equazione (B.26), esprimiamo l'indice $\frac{\sigma_k}{\mu_k}$ come media ponderata delle contribuzione al rischio di ogni controparte:

$$\begin{aligned} \frac{\sigma_k}{\mu_k} = & \frac{\sum_A \theta_{Ak} \mu_A \frac{\sigma_A}{\mu_A}}{\sum_A \theta_{Ak} \mu_A} \\ \text{poichè } \sigma_k = & \sum_A \theta_{Ak} \sigma_A \end{aligned} \tag{B.80}$$

L'equazione (B.80) permette di stimare la deviazione standard per ogni fattore.

Finora abbiamo ipotizzato che tutta la variabilità dei tassi di default nel portafoglio fosse sistematica, per questo serve un settore aggiuntivo per modellizzare anche il rischio idiosincratico di ogni controparte. Abbiamo osservato in precedenza che assegnare una variazione zero a un settore equivale a supporre che il settore sia a sua volta un portafoglio composto da un gran numero di sottosettori. Quindi, per un portafoglio contenente un gran numero di controparti, solo un settore è necessario per incorporare fattori specifici (per ipotesi settore 1). Quindi, per ciascun debitore A, la proporzione della varianza della frequenza di default prevista per una controparte determinata è spiegata dal rischio specifico θ_{A1}. Al settore 1 verrà

assegnata una deviazione standard totale data da equazione (B.80). Tuttavia, solo per il settore del rischio sistemico, la deviazione standard può essere impostata su zero. Il settore per il rischio di mercato si comporta quindi come il limite di un gran numero di settori, uno per ciascuna controparte nel portafoglio, con variabilità indipendente nella loro frequenza di default. La deviazione standard delle perdite, rappresentata da σ_1, è la misura del beneficio della presenza di fattori specifici nel portafoglio.

Contribuzione al rischio e correlazione

Contribuzione al rischio La correlazione e la contribuzione al rischio, sono due delle misure collegate alla distribuzione della frequenza dei default.

- La contribuzione al rischio è, semplicemente, il contributo di ciascun debitore all'*unexpected loss* del portafoglio, calcolata determinando un determinato percentile o deviazione standard.
- La correlazione dell'evento del default, invece, è una misura che cerca di misurare il rischio di concentrazione all'interno del portafoglio.

Per quanto riguarda una controparte "A", possiamo definire la sua contribuzione al rischio l'effetto marginale della sua esposizione "E_A" alla deviazione standard della *loss distribution* di portafoglio. Alternativamente, la contribuzione al rischio può essere definita come l'effetto marginale della presenza dell'esposizione "E_A" sulle altre misure di rischio a livello aggregato, come la perdita per percentile. Nel primo caso si può determinare la *risk contribution* con una formula analitica:

$$RC_A = E_A \frac{\partial \sigma}{\partial E_A} \equiv \frac{E_A \partial \sigma^2}{2\sigma \partial E_A} \tag{B.81}$$

La formula (B.81) dipende dalla deviazione standard in quanto:

$$\sigma^2 = \sum_{A,B} \rho_{AB} E_A E_B \sigma_A \sigma_B \tag{B.82}$$

dove σ_A e σ_B sono le deviazioni standard delle probabilità di default per ogni controparte. La varianza è qui espressa come un polinomio quadratico omogeneo delle esposizioni, quindi, attraverso una generale proprietà dei polinomi omogenei si arriva a:

$$\sum_A RC_A = \frac{1}{2\sigma} \sum_A E_A \frac{\partial \sigma^2}{\partial E_A} = \frac{2\sigma^2}{2\sigma} = \sigma \tag{B.83}$$

Se la contribuzione al rischio è scelta come effetto marginale ad un determinato percentile, non è possibile arrivare ad un risultato in modo analitico. Si deve

Tabella B.5: Notazione per lo sviluppo settoriale

Variabile	Simbolo	Media	Varianza
Polinomio dell'entità della perdita	$P(z)$		
Funzione generatrice dei dati per probabilità di default condizionali	$E(z,x)$	μ_E	σ_E^2
Funzione di densità per la media x	$f(x)$	μ_f	σ_f^2
Funzione generatrice della probabilità di default	$F(z)$	$\mu_f = \mu_k$	σ_F^2
Funzione generatrice dei dati di CreditRisk+	$G(z)$	$\mu_G = \varepsilon_k$	σ_G^2

ricorrere ad una approssimazione. Posto ε, σ e l'*expected loss* (X), la deviazione standard delle perdite e l'ammontare di perdita ad un determinato percentile della distribuzione. Definito un multiplo per un determinato percentile come ξ, dove:

$$\widetilde{RC_A} = \varepsilon_A + \xi RC_A \tag{B.84}$$

per cui l'equazione (B.83) diventa:

$$\sum_A \widetilde{RC_A} = \sum_A (\varepsilon_A + \xi RC_A) = \varepsilon + \xi\sigma = X \tag{B.85}$$

Fino ad ora ci si è concentrati sulla determinazione della contribuzione alla deviazione standard. Per valutare correttamente anche la parte destra dell'equazione (B.81), bisogna derivare una formula analitica per calcolare la media e la varianza della distribuzione dei default e della *loss distribution*. Poiché la media e la varianza della distribuzione delle perdite in CrediRisk+, sono additive tra i settori, per sviluppare il concetto si può concentrare l'attenzione ad un settore solo. Per snellire la notazione, verrà soppressa l'indicizzazione k (tabella B.5). Qui μ_k e ε_k sono rispettivamente la media del numero dei default e la perdita attesa nel settore k. Abbiamo:

$$G(z) = F\big(P(z)\big) \tag{B.86}$$

che rappresenta una semplice manipolazione dell'equazione (B.49). Lo stesso, dall'equazione (B.38):

$$F(z) = \int_x E(z,x) f(x) dx \tag{B.87}$$

Per la funzione generatrice dei dati E, F e G, secondo la proprietà generale della funzione generatrice dei dati:

$$\mu_E = \frac{dE}{dz}(1), \qquad \sigma_E^2 + \mu_E^2 = \frac{d^2E}{dz^2}(1) + \frac{dE}{dz}(1); \tag{B.88}$$

$$\mu_F = \frac{dF}{dz}(1), \qquad \sigma_F^2 + \mu_F^2 = \frac{d^2F}{dz^2}(1) + \frac{dF}{dz}(1); \tag{B.89}$$

$$\text{e } \mu_G = \frac{dG}{dz}(1), \qquad \sigma_G^2 + \mu_G = \frac{d^2G}{dz^2}(1) + \frac{dG}{dz}(1) \tag{B.90}$$

Dalla definizione di x otteniamo:

$$\mu_E(x) = x \tag{B.91}$$

Poiché $E(z, x)$ è la funzione generatrice delle probabilità secondo una distribuzione di Poisson, abbiamo:

$$\sigma_E^2 = \mu_E \tag{B.92}$$

Differenziano le equazioni (B.87) e (B.90) secondo la variabile ausiliaria z otteniamo:

$$\begin{aligned} \mu_F &= \int_x \mu_E(x) f(x) dx = \\ &= \int_x x f(x) dx = \\ &= \mu_f \end{aligned} \tag{B.93}$$

In modo analogo, usando l'equazione (B.87) e (B.90), si arriva:

$$\sigma_F^2 + \mu_F^2 = \int_x (\sigma_E^2 + \mu_E^2) f(x) dx = \tag{B.94}$$

$$= \int_x (\mu_E + \mu_E^2) f(x) dx = \tag{B.95}$$

$$= \mu_f + \sigma_f^2 + \mu_f^2 \tag{B.96}$$

Quindi:

$$\sigma_F^2 = \mu_F + \sigma_f^2 \tag{B.97}$$

Le equazioni (B.93) e (B.97) rappresentano la media e la varianza della frequenza dei default. Per rappresentare la connessione con i momenti della *loss distribution*

possiamo utilizzare l'equazione (B.86), che produce:

$$\frac{dG}{dz}(z) = \frac{dF}{dz}\big(P(z)\big)\frac{dP}{dz} \tag{B.98}$$

$$\frac{d^2G}{dz^2}(z) = \frac{d^2F}{dz^2}\big(P(z)\big)(\frac{dP}{dz})^2 + \frac{dF}{dz}\big(P(z)\big)\frac{d^2P(z)}{dz^2} \tag{B.99}$$

poiché:

$$\sigma_G^2 = \frac{d^2F}{dz^2}\big(P(1)\big)\frac{dP}{dz}(1)^2 + \frac{dF}{dz}\big(P(1)\big)\frac{d^2P}{dz^2}(1) + \frac{dF}{dz}\big(P(1)\big)\frac{dP}{dz}(1) - \left(\frac{dF}{dz}\big(P(1)\big)\frac{dP}{dz}\right)^2 \tag{B.100}$$

Successive differenziazioni dell'equazione (B.78), conducono:

$$P(1) = 1 \tag{B.101}$$

$$\frac{dP}{dz}(1) = \frac{1}{\mu_k}\sum_A \theta_{Ak}\varepsilon_A = \frac{\varepsilon_k}{\mu_k} \tag{B.102}$$

$$\frac{d^2P}{dz^2}(1) = \frac{1}{\mu_k}\sum_A \theta_{Ak}\varepsilon_A(v_A - 1) \tag{B.103}$$

Sostituendo le equazioni (B.97) e (B.103), nell'equazione (B.100), otteniamo:

$$\sigma_G^2 = (\sigma_F^2 + \mu_k^2 - \mu_k)\left(\frac{1}{\mu_k}\sum_A \theta_{Ak}\varepsilon_A\right)^2 + \sum_A \theta_{Ak}\varepsilon_A v_A - \left(\sum_A \theta_{Ak}\varepsilon_A\right)^2 \tag{B.104}$$

Sostituendo per ε_k, si ottiene:

$$\sigma_G^2 = (\sigma_k^2 + \mu_k^2)\left(\frac{\varepsilon_k}{\mu_k}\right)^2 + \sum_A \theta_{Ak}\varepsilon_A v_A - \varepsilon_A v_A - \left(\sum_A \theta_k^2\right) = \tag{B.105}$$

$$= \sigma_k^2\left(\frac{\varepsilon_k}{\mu_k}\right)^2 + \sum_A \theta_{Ak}\varepsilon_A v_A \tag{B.106}$$

Infine, sommando attraverso i settori, si ottiene la deviazione standard della *loss distribution* di portafoglio restituita da CreditRisk+:

$$\sigma^2 = \sum_{k=1}^{n} \varepsilon_k^2\left(\frac{\sigma_k}{\mu_k}\right)^2 + \sum_A \varepsilon_A v_A \tag{B.107}$$

che rappresenta la deviazione standard della distribuzione delle perdite. Come definito prima, σ_k rappresenta la deviazione standard dei fattore che conducono la frequenza di default in ogni settore. Adesso si può derivare la contribuzione al rischio direttamente dall'equazione (B.107). Per cui, riprendendo l'equazione

(B.81), si arriva a:

$$RC_A = \frac{E_A}{2\sigma}\frac{\partial}{\partial E_A}\left(\sum_B \varepsilon_B v_B + \sum_k \left(\frac{\sigma_k}{\mu_k}\right)^2 \varepsilon_k^2\right) = \tag{B.108}$$

$$= \frac{E_A}{2\sigma}\left(2E_A\mu_A + \sum_k \left(\frac{\sigma_k}{\mu_k}\right)^2 \varepsilon_k \theta_{Ak}\mu_A\right) \tag{B.109}$$

in cui sono stati scambiati E_A e v_A per agevolare la notazione. Quindi:

$$RC_A = \frac{E_A\mu_A}{\sigma}\left(E_A + \sum_k \left(\frac{\sigma_k}{\mu_k}\right)\varepsilon_k\theta_{Ak}\right) \tag{B.110}$$

Questa rappresenta la formula per il calcolo della contribuzione al rischio della deviazione standard. Come già accennato precedentemente, l'equazione (B.110) si può scrivere in modo esplicito:

$$\sum_A RC_A = \sum_A \frac{E_A\mu_A}{\sigma}\left(E_A + \sum_k \left(\frac{\sigma_k}{\mu_k}\right)^2 \theta_{Ak}\varepsilon_k\right) = \tag{B.111}$$

$$= \sum_A \frac{E_A^2\mu_A}{\sigma} + \sum_A\sum_k \left(\frac{\sigma_k}{\mu_k}\right)^2 \frac{E_A\mu_A}{\sigma}\theta_{Ak}\varepsilon_k \tag{B.112}$$

Quindi, utilizzando l'equazione (B.107), si arriva come richiesto a:

$$\sum_A RC_A = \sum_A \frac{\varepsilon_A v_A}{\sigma} + \sum_k \left(\frac{\sigma_k}{\mu_k}\right)^2 \varepsilon_k \sum_A \theta_{Ak}\frac{E_A\mu_A}{\sigma} = \tag{B.113}$$

$$= \frac{1}{\sigma}\left(\sum_A \varepsilon_A v_A + \sum_k \left(\frac{\sigma_k}{\mu_k}\right)^2 \varepsilon_k^2\right) \tag{B.114}$$

$$= \frac{\sigma^2}{\sigma} = \tag{B.115}$$

$$= \sigma \tag{B.116}$$

Correlazione Per quanto riguarda la correlazione in un periodo di tempo Δt, ad ogni controparte va associata la propria funzione di indicazione (I_A), variabile random con valori

$$I_A = \begin{cases} 1 & \text{se la controparte defaulta} \\ 0 & \text{altrimenti} \end{cases} \tag{B.117}$$

Tabella B.6: Notazione per il calcolo della correlazione

Variabile	Controparte A	Controparte B
Periodo temporale	Δt	Δt
Probabilità di default	p_A	p_B
Frequenza default attesi	$\mu_A = 1 - e^{-p_A \Delta t} = p_A \Delta t$	$\mu_B = 1 - e^{-p_B \Delta t} = p_B \Delta t$
Suddivisione del settore	$\theta_{Ak}; 1 \leq k \leq n$	$\theta_{Bk}; 1 \leq k \leq n$

A questo punto, la correlazione del default (ρ) tra due controparti, A e B, nel periodo Δt, è definita come:

$$\rho_{AB} = \rho(I_A, I_B) \tag{B.118}$$

Che rappresenta la correlazione statistica tra le controparti considerate. Se i valori attesi di I_A e I_B e del prodotto $I_{A,B}$, sono rispettivamente μ_A, μ_B e μ_{AB}, sono anche il numero di default attesi per le controparti considerate (A,B,AB). Avendo la funzione di identificazione un valore compreso tra 0 e 1, l'espressione standard della correlazione si riduce alla forma seguente:

$$\rho_{AB} = \frac{\mu_{AB} - \mu_A \mu_B}{(\mu_A - \mu_A^2)^{1/2} (\mu_B - \mu_B^2)^{1/2}} \tag{B.119}$$

Si deve, ora, cercare un' espressione per calcolare l'espressione (B.119) nel contesto CreditRisk+, in quanto il termine μ_{AB} è sconosciuto. Consideriamo sempre due controparti, A e B, a cui associamo la notazione nella tabella. B.6 Essendo μ_{AB} la probabilità di default congiunta attesa, e poiché A e B sono distinti per ogni realizzazione della media di settore $x_k, 1 \leq k \leq n$, allora gli eventi del default sono indipendenti e x_A, x_B possono essere riscritti come nell'equazione (B.77):

$$\mu_{AB} = \int_{x_1} \cdots \int_{x_n} x_A x_B \prod_{k=1}^{n} f_k(x_k) dx_k \tag{B.120}$$

dove, come indicato in tabella B.6, si approssimano gli integrali ignorando il potere della probabilità di default attesa secondo:

$$(1 - e^{-x_A \Delta t})(1 - e^{-x_B \Delta t}) \approx x_A x_B \tag{B.121}$$

A partire dall'equazione (B.77), secondo un' ottica di decomposizione settoriale, otteniamo:

$$x_A = \sum_{k=1} \frac{x_k}{\mu_k} \theta_{Ak} \mu_A \tag{B.122}$$

$$x_B = \sum_{k=1} \frac{x_k}{\mu_k} \theta_{Bk} \mu_B \tag{B.123}$$

Definendo per convenienza, un coefficiente $\omega_{kk'}$:

$$\omega_{kk'} = \frac{\theta_{Ak}\theta_{Bk'}}{\mu_k \mu_{k'} \mu_A \mu_B} \tag{B.124}$$

Allora:

$$\mu_{AB} = \int_{x_1} \cdots \int_{x_n} \sum_{k,k'} \omega_{kk'} x_k x_{k'} \prod_{k=1}^{n} f_k(x_k) dx_k \tag{B.125}$$

e si può dedurre che:

$$\mu_{AB} = \sum_{k \neq k'} \omega_{kk'} \partial_{x_k} \partial_{x_{k'}} x_k x_{k'} f_k(x_k) f_k(x_{k'}) dx_k dx_{k'} \partial_{x_j : j \neq k,k'} \prod_{j=1; i \neq k,k'}^{n} f_j(x_j) dx_j \tag{B.126}$$

$$+ \sum_{k=1}^{n} \omega_{kk} \partial_{x_k} x_k^2 f_k(x_k) dx_k \partial_{x_j : j \neq k} \prod_{j=1; i \neq k,k'}^{n} f_j(x_j) dx_j \tag{B.127}$$

ovvero:

$$\mu_{AB} = \sum_{k,k=1, k \neq k'} \omega_{kk'} \mu_k \mu_k + \sum_{k=1} \omega_{kk} (\mu_k^2) \sigma_k^2 \tag{B.128}$$

Comunque:

$$\sum_{k,k'=1} \omega_{kk'} \mu_k \mu_{k'} = \left(\sum_{k=1}^{n} \frac{\theta_{Ak} \mu_A}{\mu_k} \mu_k \right) \left(\sum_{k=1}^{n} \frac{\theta_{Bk} \mu_B}{\mu_k} \mu_k \right) = \tag{B.129}$$

$$= \mu_A \mu_B \tag{B.130}$$

Poiché $\mu_{AB} = \mu_A\mu_B + \sum_{k=1}^{n} \omega_{kk}\sigma_k^2$, sostituendo per ω_{kk}, otteniamo:

$$\rho_{AB} = \frac{\mu_{AB} - \mu_A\mu_B}{(\mu_A - \mu_A^2)^{1/2}(\mu_B - \mu_B^2)^{1/2}} = \tag{B.131}$$

$$= (\mu_A\mu_B)^{1/2} \sum_{k=1}^{n} \theta_{Ak}\theta_{Bk}(\frac{\sigma_k}{\mu_k})^2 \tag{B.132}$$

$$\text{che si semplifica in} \tag{B.133}$$

$$\rho_{AB} = (\mu_A\mu_B)^{1/2} \sum_{k=1}^{n} \theta_{Ak}\theta_{Bk}\left(\frac{\sigma_k}{\mu_k}\right)^2 \tag{B.134}$$

L'equazione (B.134) rappresenta la formula della correlazione dei default per due controparti, A e B.

Bibliografia

Altman, Edward I (1968). *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.*

AVESANI, RENZO G et al. (2014). *Review and Implementation of Credit Risk Models.*

Barro, Diana (2004). «Un'introduzione ai modelli di rischio di credito per portafogli finanziari». In: *Università Ca'Foscari di Venezia* 124.

Berkovich, Efraim (2011). *Search and herding effects in peer-to-peer lending: evidence from prosper. com.*

Boston, Credit Suisse First (1997). *CreditRisk+: A credit risk management framework.*

Bürgisser, Peter et al. (1999). *Integrating correlations.*

Caratelli, Massimo et al. (2016). *Il mercato del peer-to-peer lending nel mondo e le prospettive per l'Italia.*

Carmichael, Don (2014). *Modeling default for peer-to-peer loans.*

Comitato di Basilea per la vigilanza bancaria (2004). *International convergence of capital measurement and capital standards: a revised framework.* Bank for International Settlements.

CreditRisk+ in the Banking Industry (2004). 1ª ed. Springer Finance. Springer-Verlag Berlin Heidelberg.

Crouhy, Michel, Dan Galai e Robert Mark (2000). *A comparative analysis of current credit risk models.*

David W. Hosmer, Stanley Lemeshow (2000). *Applied logistic regression (Wiley Series in probability and statistics).* 2ª ed. Wiley Series in probability and statistics. Wiley-Interscience Publication.

Faia, Ester e Monica Paiella (2017). *P2P Lending: Information Externalities, Social Networks and Loans' Substitution.*

Falkenstein, Eric (2002). *Credit scoring for corporate debt.*

— (2008). *DefProb: A Corporate Probability of Default Model.*

Falkenstein, Eric, Andrew Boral e Lea Carty (2000). *RiskCalc for private companies: Moody's default model.*

Fernandes, João (2005). *Corporate credit risk modeling: Quantitative rating system and probability of default estimation.*

Gordy, Michael B (2000). *A comparative anatomy of credit risk models.*

— (2002). *Saddlepoint approximation of CreditRisk+.*

Gregoriou, Greg N. (2009). *Operational Risk Toward Basel III: Best Practices and Issues in Modeling, Management, and Regulation (Wiley Finance).* 1ª ed.

Gunter Loeffler, Peter N. Posch (2007). *Credit Risk Modeling using Excel and VBA.* Har/DVD. Wiley finance series. Wiley.

Hilbe, Joseph (2007). *Negative binomial regression.* 1ª ed. Cambridge University Press.

Hilbe, Joseph M. (2015). *Practical Guide to Logistic Regression.* Chapman e Hall/CRC.

Hull, John C (2006). *Opzioni, futures e altri derivati. Manuale delle soluzioni.* Pearson Italia Spa.

Ju Yang (auth.), Dash Wu (eds.) (2011). *Quantitative Financial Risk Management.* 1ª ed. Computational Risk Management 1. Springer-Verlag Berlin Heidelberg.

Klafft, Michael (2008). *Online peer-to-peer lending: a lenders' perspective.*

Koyluoglu, H Ugur e Andrew Hickman (1998). *A generalized framework for credit risk portfolio models.*

Luisa Izzi Gianluca Oricchio, Laura Vitale (auth.) (2012). *Basel III Credit Rating Systems: An Applied Guide to Quantitative and Qualitative Models.* Palgrave Macmillan Finance and Capital Markets Series. Palgrave Macmillan UK.

Mach, Traci, Courtney Carter e Cailin Slattery (2014). *Peer-to-peer lending to small businesses.*

Magee, Jack R (2011). *Peer-to-peer lending in the United States: surviving after Dodd-Frank.*

Marek Capiński, Tomasz Zastawniak (2017). *Credit Risk.* 1ª ed. Mastering Mathematical Finance. Cambridge University Press.

Melchiori, Mario R (2004). *Creditrisk+ by fast fourier transform.*

Milne, Alistair e Paul Parboteeah (2016). «The business models and economics of peer-to-peer lending». In:

Mingo, John e Eric Falkenstein (2000). *Internal Credit Risk Rating Systems Must Evolve to Stay Relevant.*

Morgan, JP (1997). *Creditmetrics-technical document.*

On Banking Supervision, Basel Committee (2005). *Studies on the Validation of Internal Rating System.*

Papini, Matteo (2010). *L'impatto della microfinanza, delle asimmetrie informative e del peer-to-peer bancario nei mercati in via di sviluppo.*

Raymond Anderson (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, USA.

Refaat, Mamdouh (2006). *Data Preparation for Data Mining Using SAS (The Morgan Kaufmann Series in Data Management Systems)*.

Resti A., Sironi A. (2007). *Risk Management and Shareholders' Value in Banking: From Risk Measurement Models to Capital Allocation Policies*.

Siddiqi, Naeem (2017). *Intelligent credit scoring : building and implementing better creditrisk scorecards*. 2nd edition. Wiley.

Simitas, Fotios (2015). *Peer To Peer Lending*.

Svetlozar T. Rachev Stoyan V. Stoyanov, Frank J. Fabozzi CFA (2008). *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimization: The Ideal Risk, Uncertainty, and Performance Measures (Frank J. Fabozzi Series)*. Frank J. Fabozzi Series. Wiley.

Tang, Huan (2017). *P2P lenders versus Banks, substitutes or complements*.

Thomas David Edelman, Jonathan Crook (2002). *Credit scoring and its applications*. 1st. SIAM monographs on mathematical modeling and computation. Society for Industrial e Applied Mathematics.

Vandendorpe, Antoine et al. (2008). «On the parameterization of the CreditRisk+ model for estimating credit portfolio risk». In: *Insurance: Mathematics and Economics* 42.2, pp. 736–745.

Wang, Shaun (1998). «Aggregation of correlated risk portfolios: models and algorithms». In: *Proceedings of the Casualty Actuarial society*. Vol. 85. 163, p. 92.

Wardrop, Robert et al. (2016). «Breaking new ground: the Americas alternative finance benchmarking report». In: *Cambrige Centre for Alternative Finance https://www. jbs. cam. ac. uk/fileadmin/user_ upload/research/centres/alternative-finance/downloads/2016-americas-alternative-finance-benchmarking-report. pdf (accessed Fegruary 19, 2017)*.

Wilson, Thomas C (1997). *Portfolio credit risk*.

Zanetti, Silvia (2017). «Peer to Peer Lending: intermediazione finanziaria online». B.S. thesis. Università Ca'Foscari Venezia.

Zhang, Bryan et al. (2016). «Pushing boundaries: The 2015 UK alternative finance industry report». In: *Cambridge Centre for Alternative Finance*.