POLITECNICO DI TORINO

Collegio di Ingegneria Gestionale

Corso di Laurea Magistrale in Ingegneria Gestionale

Modelli di credit scoring:

Confronto delle performance tra regressioni logistiche e reti neurali sui dati di bilancio delle imprese del settore metallurgico italiano

Relatore: Prof.ssa Laura Rondi

Correlatore: Prof. Franco Varetto

Candidato:

Fabio Spina

Anno accademico 2021/2022

Indice

Introduzione

I modelli predittivi sono utilizzati in numerosi ambiti della finanza, soprattutto grazie a un costante incremento dei dati a disposizione degli analisti, tanto da far diventare imprescindibile per questo business la presenza di esperti programmatori e data scientists. Uno degli ambiti in cui tali modelli sono ampiamente utilizzati è quello del credit scoring, ossia una procedura attivata dagli intermediari finanziari per produrre un giudizio relativamente alle richieste di credito provenienti della clientela al fine di determinare la convenienza o meno di accendere delle linee di credito.

Il presente lavoro di tesi si pone l'obiettivo di valutare le performance di due tipologie di modelli di credit scoring basati sulle regressioni logistiche e sulle reti neurali nel diagnosticare eventi di default a partire dai dati di bilancio delle imprese. In particolare, le tipologie di modelli sviluppati presentano approcci e strutture molto differenti tra loro: la regressione logistica è una tecnica molto rigorosa con forti fondamenta statistiche che richiede il controllo di specifiche caratteristiche affinché il modello sviluppato possa essere preso in considerazione, quali la significatività statistica dei coefficienti e la coerenza dei segni di questi con il significato economico della variabile a cui sono associati; al contrario, le reti neurali sono una tecnica di machine learning estremamente flessibile, che viene soprattutto sviluppata con approcci empirici e intuitivi e che presenta praticamente infiniti gradi di libertà nella definizione delle architetture delle reti, permettendo così di sviluppare modelli capaci di cogliere relazioni intrinseche anche molto complesse fra le variabili.

I modelli sviluppati e analizzati sono stati programmati tramite codice Python importando nell'ambiente di sviluppo (Anaconda) le librerie necessarie per la loro costruzione, ed estraendo dal database Aida i dati di bilancio relativi alle imprese italiane appartenenti al settore metallurgico sull'orizzonte temporale 2011-2020.

La tesi è suddivisa in due macroaree dove nella prima parte, più teorica e descrittiva, sono state messe in evidenza le fondamenta teoriche su cui si sono basati gli sviluppi della seconda parte, più applicativa e sperimentale, in cui sono stati programmati e analizzati i modelli di credit scoring. Nello specifico:

- Nei primi quattro capitoli ci si è concentrati in primo luogo sulla presentazione della normativa vigente relativa alla gestione dei rischi per le banche e per gli intermediari finanziari, focalizzandosi anche sugli elementi chiave che hanno determinato le evoluzioni della regolamentazione da Basilea I fino all'attuale

Basilea III. Inoltre, sono state presentate le fondamenta teoriche alla base del rischio di credito, del credit scoring, della regressione logistica e delle tecniche di machine learning con un focus particolare sulle reti neurali.

- Nel capitolo 5 è stata condotta un'analisi macrosettoriale del settore metallurgico italiano, rielaborando anche le informazioni estratte dal database Aida e fornendo una prospettiva di andamento generale del settore, calcolando e commentando gli andamenti dei principali indici di redditività, patrimoniali e finanziari sull'orizzonte temporale 2011-2020;
- Il capitolo 6 è il cuore della parte applicativa e sperimentale della tesi e in cui sono state descritte le parti di codice Python che hanno permesso la creazione dei modelli, oltre che aver commentato le performance di classificazione ottenute;
- Nel capitolo 7 sono stati riassunti i risultati ottenuti e sono stati messi in evidenza i punti chiave che hanno caratterizzato gli sviluppi e alcune interpretazioni dei risultati più particolari che hanno fatto scaturire dei punti di riflessione durante gli sviluppi.

1. La Regolamentazione di Basilea

La regolamentazione svolge un ruolo molto importante all'interno del sistema in cui opera una banca poiché quando essa sviluppa il proprio sistema di risk management, con l'obiettivo di misurare l'assorbimento di capitale dovuto ai suoi asset rischiosi, deve necessariamente considerare anche i vincoli stabiliti dalla regolamentazione vigente. Da sempre sono presenti due correnti di pensiero contrastanti relative all'intervento della regolamentazione nelle dinamiche del sistema economico: una corrente di pensiero sostiene che *l'intervento sia superfluo e dannoso per il sistema economico* dal momento che il mercato è perfettamente in grado di autoregolarsi al meglio e più in fretta delle autorità di governo; l'altra sostiene invece che *l'intervento è necessario* poiché il sistema da solo non è in grado di autoregolarsi e produce condizioni non concorrenziali, asimmetrie di trattamento, instabilità e crisi sistemiche frequenti.

In generale si possono distinguere due impostazioni di regolamentazione bancaria:

- *Regolamentazione strutturale*: vengono imposti limiti all'operatività delle banche in modo da contenere i rischi che possono essere assunti (divieti all'entrata, all'esposizione geografica, ai prodotti offerti, alla concentrazione dei rischi, …);
- *Regolamentazione prudenziale*: l'operatività è libera ma vengono imposte forme di copertura dei rischi che si sono assunti (riserva obbligatoria, assicurazione sui depositi, standard di capitalizzazione, …).

A partire dalla Grande Depressione del 1929 negli USA il settore bancario è stato oggetto di una profonda regolamentazione al fine di scongiurare il riverificarsi di eventi simili. Tuttavia, nella metà degli anni '80 ormai da tempo aleggiava la convinzione che una regolamentazione troppo rigorosa limitasse eccessivamente le opportunità di investimento e, più in generale, la crescita dell'intero sistema economico. Iniziò quindi un processo di deregolamentazione, il quale ha sostanzialmente significato il passaggio dalla regolamentazione strutturale alla regolamentazione prudenziale appena citate, stimolando la competizione secondo l'idea che le banche che sono in grado di competere e produrre servizi finanziari a un costo più basso, oppure che sono in grado di generare prodotti nuovi tramite l'ingegneria finanziaria, sono libere di farlo a patto di rispettare specifici requisiti in termini di equity.

Uno dei principali fori di discussione e cooperazione in materia di vigilanza sull'attività bancaria internazionale è il Comitato di Basilea, un punto di riferimento a livello internazionale per l'indirizzo della regolamentazione nel settore bancario. Si tratta di un

organismo di cooperazione istituito nel 1974 e composto dai rappresentanti delle banche centrali ed autorità di vigilanza dei paesi del G10. Anche se sotto il profilo formale le decisioni del Comitato non hanno un valore giuridico, tuttavia influenzano in modo determinante le legislazioni non solo dei vari paesi del G10, ma anche di molti altri paesi che adottano volontariamente le regole da esso emanate. Il Comitato negli anni ha dedicato una particolare attenzione all'adeguatezza patrimoniale dei singoli istituti bancari, al fine di garantire la solvibilità di ognuno di essi e, più in generale, la stabilità dell'intero sistema bancario internazionale.

1.1 Basilea I

I requisiti regolamentari sul capitale basati sui capital ratio furono originariamente proposti dal Comitato di Basilea nel dicembre del 1987 per poi essere ratificati *nell'Accordo sul capitale* nel giugno del 1988. Inizialmente, tale accordo fu applicato solamente alle banche che operavano su scala internazionale, ma molte autorità nazionali, comprese quelle dell'Unione Europea, decisero di renderlo effettivo anche per tutte le banche, anche quelle con un'operatività unicamente domestica. Nello specifico, il desiderio di requisiti patrimoniali uniformi su scala internazionale rispondeva a tre esigenze:

1. Rendere più certa la solvibilità delle banche, scoraggiandole dall'assumere rischi eccessivi;
2. Garantire, attraverso l'applicazione dei requisiti su base consolidata, anche la solvibilità delle istituzioni controllate da gruppi bancari esteri, promuovendo una maggiore stabilità nei mercati finanziari internazionali;
3. Superare le distorsioni competitive legate a diverse normative nazionali, favorendo la creazione di condizioni concorrenziali uniformi per le istituzioni bancarie dei diversi Paesi.

Dalla figura 1.1 si può notare come una maggiore capitalizzazione delle banche pareva opportuna a seguito della progressiva diminuzione del livello di patrimonializzazione dei principali sistemi bancari. Infatti, il rapporto fra capitale e totale dell'attivo un tempo risultava particolarmente elevato (15-20 per cento a fine '800 per i sistemi bancari più avanzati) e si è progressivamente ridotto, toccando un punto di minimo intorno al 1970.

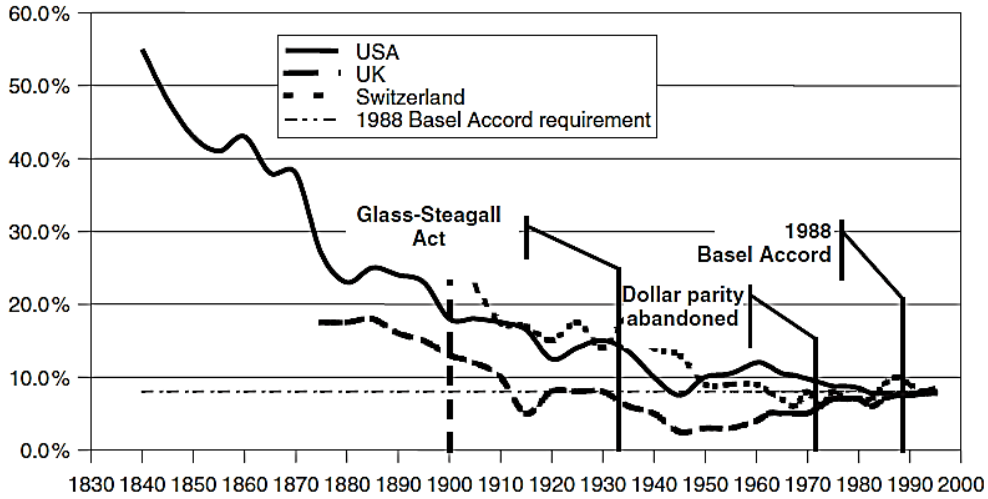


Figura 1.1: andamento del capital-to-assets ratio nei principali sistemi bancari (fonte: "Rischio e valore nelle banche" di A. Resti e A. Sironi)

Per quanto riguarda invece il terzo obiettivo relativo alle distorsioni competitive, un fine non esplicitamente dichiarato dell'Accordo fu probabilmente quello di ridurre i vantaggi di cui godevano all'epoca le banche giapponesi che, secondo le banche britanniche e statunitensi, operavano con una dotazione di capital-to-assets ratio molto inferiore rispetto ai loro principali concorrenti del G-10. Infatti, un minor capital-to-assets ratio si traduce in un minor costo medio ponderato delle fonti di finanziamento e questa situazione ha permesso alle banche giapponesi di realizzare una significativa espansione della loro quota di mercato nel settore dei prestiti bancari internazionali.

1.1.1 Il coefficiente patrimoniale

Per perseguire i tre obiettivi citati precedentemente, l'accordo del 1988 prevedeva che le banche garantissero un rapporto minimo dell'8%, detto *coefficiente patrimoniale*, tra il patrimonio, definito nell'accordo del 1988 *patrimonio di vigilanza*, e le attività ponderate per il rischio. In formula:

$$\frac{\text{Patrimonio di Vigilanza}}{\sum_i A_i \cdot w_i} \geq 8\% \qquad (1.1)$$

dove:

- A_i = attività i-esima;
- w_i = ponderazione per il rischio dell'attività i-esima.

In particolare, è richiesto che il coefficiente patrimoniale risulti pari ad all'8 per cento a livello consolidato o per le banche non appartenenti a gruppi, mentre per le singole banche appartenenti a un gruppo è richiesto che il coefficiente patrimoniale sia almeno del 7 per cento. Tale formulazione del coefficiente patrimoniale è rimasta in vigore fino al 1996, quando un emendamento del Comitato di Basilea ha esteso il requisito di capitale anche ai rischi di mercato:

$$\frac{\text{Patrimonio di Vigilanza}}{\sum_i A_i \cdot w_i + 12.5 \cdot RM} \geq 8\% \qquad (1.2)$$

dove RM indica i rischi di mercato.

1.1.1.1 Il patrimonio di vigilanza

Il patrimonio di vigilanza è suddiviso in due categorie: il *patrimonio tier 1* e il *patrimonio tier* 2. Il patrimonio tier 1 è costituito dalle poste patrimoniali più "pregiate", ovvero contraddistinte da un'elevata capacità di proteggere i terzi dagli effetti di eventuali perdite subite dalla banca. In patrimonio tier 2, invece, è costituito da strumenti maggiormente assimilabili a debito. La tabella 1.1 riassume le voci che possono andare a costituire il patrimonio di vigilanza assieme alle relative eventuali condizioni e limitazioni.

(a) Upper tier 1: - Capitale versato/azioni ordinarie - Riserve palesi (ad es. sovrapprezzo azioni o utili non distribuiti)
(b) Lower tier 1: - Strumenti innovativi di capitale (non oltre il 15 per cento del tier 1)
(c) = (a) + (b) Patrimonio tier 1 (almeno il 4 per cento delle attività ponderate per i rischi)
(d) Upper tier 2: - Riserve occulte - Riserve di rivalutazione - Accantonamenti a fondi generali per rischi su crediti - Strumenti ibridi di patrimonializzazione
(e) Lower tier 2: - Prestiti subordinati ordinari (non oltre il 50 per cento del tier 1)
(f) = (d) + (e) Patrimonio tier 2 (non oltre il patrimonio tier 1)
(g) Deduzioni: - Avviamento (dedotto dal tier 1) - investimenti in banche e simili non consolidati (dedotti dal patrimonio totale)
(c) + (f) – (g) = Patrimonio di vigilanza (almeno l'8 per cento delle attività ponderate per i rischi)
(h) Tier 3 (valido solo per i rischi di mercato):

- Prestiti subordinati a breve scadenza (non oltre il 250 per cento del tier 1 per i rischi di mercato[1])

Tabella 1.1: principali componenti del patrimonio di vigilanza (fonte: "Rischio e valore nelle banche" di A. Resti e A. Sironi)

1.1.1.2 Le ponderazioni per il rischio

I pesi assegnati alle principali poste dell'attivo dell'Accordo del 1988 erano più elevati, e quindi implicavano un requisito patrimoniale più consistente, per le attività giudicate più rischiose. La tabella 1.2 riassume i pesi assegnati ai principali assets individuabili nello Stato Patrimoniale di una banca. In particolare, gli assets sono assegnati alle varie classi di ponderazione sulla base di tre criteri:

1. *Grado di liquidità* (maggiore per la cassa, minore per i titoli e ancora minore per i prestiti e le proprietà immobiliari);
2. *Natura dei debitori* (governi centrali e banche centrali, istituzioni sovranazionali, enti pubblici, banche imprese);
3. *Paese di residenza del debitore* (Paesi dell'area OCSE e Paesi non OCSE).

$w_i = 0\%$	$w_i = 10\%$	$w_i = 20\%$	$w_i = 50\%$	$w_i = 100\%$
Cassa ed equivalenti Crediti verso governi centrali e banche centrali dei paesi OCSE	Crediti verso enti pubblici	Crediti verso banche OCSE, enti bancari internazionali e banche non OCSE, con durata residua inferiore ad 1 anno	Crediti ipotecari su immobili residenziali	Altri crediti

Tabella 1.2: valori percentuali dei pesi per diverse asset class (fonte: "Rischio e valore nelle banche" di A. Resti e A. Sironi)

1.1.2 Limiti di Basilea I e conclusioni

L'approccio raccomandato dalla regolamentazione di Basilea I presenta le seguenti limitazioni:

[1] Se si indica con C_1 il patrimonio tier 1 a fronte dei rischi di mercato e con C_3 il tier 3, la condizione espressa dalla regolamentazione è che $C_3 < 2{,}5 \cdot C_1$, da cui risulta che $C_1 > 0{,}286(C_1 + C_3)$, ovvero che i rischi di mercato devono essere coperti almeno per il 28,6 per cento dal patrimonio tier 1.

- *Focus esclusivo sul rischio di credito* – Anche se i pesi w_i tenevano in qualche modo conto della liquidità delle diverse classi di attività finanziarie, è però vero che lo schema proposto nel 1988 si concentrasse principalmente sul rischio di credito trascurando di conseguenza gli altri rischi, in particolare quello di tasso di interesse, di mercato e operativo. Infatti, proprio per questa criticità il Comitato varò nel 1966 un emendamento che estendeva i requisiti patrimoniali obbligatori anche ai rischi di mercato.
- *Scarsa differenziazione del rischio* – I pesi della tabella 1.2 consideravano come un'unica categoria di rischio tutti i crediti verso imprese private commerciali e industriali. In questo modo il risultato è che si assoggettavano al medesimo requisito patrimoniale anche imprese con rating differenti. Allo stesso modo, tutte le esposizioni verso Paesi non OCSE venivano considerate più rischiose di quelle verso i Paesi OCSE.
- *Limitato riconoscimento del legame tra scadenza e rischio di credito* – L'Accordo del 1988 ignora quasi completamente il fatto che un'esposizione creditizia presenta un grado di rischio diverso in funzione della sua vita residua. Infatti, esposizioni a lungo termine richiedono un requisito patrimoniale maggiore dal momento che hanno più probabilità, entro la fine del loro orizzonte temporale, di migrare verso una classe di rating peggiore.
- *Mancato riconoscimento della diversificazione di portafoglio* – Quando si vuole valutare il portafoglio dei prestiti di una banca è importante misurare sia i rischi relativi alle singole esposizioni che compongono il portafoglio, ma soprattutto anche le loro correlazioni. La regolamentazione proposta dall'Accordo del 1988 non permetteva la rilevazione di questo effetto, pertanto un portafoglio costituito da un pool di molti crediti ben diversificati richiedeva il medesimo requisito di capitale di un portafoglio costituito da pochi crediti accesi verso pochi clienti/nazioni/settori industriali. Il risultato di questa soluzione fu la completa assenza di un incentivo a diversificare il rischio di credito.
- *Sfruttamento di arbitraggi regolamentari* – Per via delle limitazioni descritte ai punti precedenti, il rischio misurato sulla base delle indicazioni dell'Accordo del 1988 risultava molto differente rispetto a quello stimato dai sistemi di rating interni e dai modelli VaR. Tali discrepanze permettevano alle banche di sfruttare degli arbitraggi regolamentari che le permettevano di:

- Incrementare le esposizioni creditizie caratterizzate da un requisito patrimoniale inferiore all'assorbimento di capitale misurato dai modelli interni (ad esempio i crediti nei confronti di Paesi OCSE ad alto rischio);
- Cedere a terzi, ad esempio mediante operazioni di *securitization*, le esposizioni creditizie meno rischiose, caratterizzate da un requisito patrimoniale superiore al grado di rischio effettivo.

Questi arbitraggi regolamentari hanno permesso un peggioramento della qualità dei portafogli creditizi delle banche, producendo dunque un risultato fortemente in contrasto con le finalità dalla regolamentazione che miravano a determinare invece un sistema bancario internazionale più solido.

Nonostante le limitazioni appena descritte, l'Accordo del 1988 è comunque riuscito a contrastare la tendenza verso una progressiva riduzione della capitalizzazione delle banche. Inoltre, negli anni successivo al 1988 il Comitato di Basilea non ha interrotto i propri lavori ma ha continuato a lavorare per arricchire e completare la struttura originaria della regolamentazione. Infatti, come specificato precedentemente, già nel 1996 fu approvato un emendamento relativo ai rischi di mercato e, successivamente, nel 1999 è iniziata una profonda riforma dei requisiti relativi al rischio di credito e che ha poi portato al nuovo accordo pubblicato nel 2004 di cui si discuterà nel successivo capitolo.

1.2 Basilea II

Nel 1988 il Comitato di Basilea decise di adottare un meccanismo regolamentare semplice ed omogeneo, il quale è risultato tuttavia troppo semplicistico per effettuare una concreta ed efficace valutazione dei rischi. Infatti, i limiti dello schema del 1988 e le distorsioni derivanti dalle operazioni di arbitraggio regolamentare hanno indotto le autorità di vigilanza ad avviare nel 1999 una riforma, emanata poi nella forma definitiva nel 2004, che ha cercato di conferire una maggiore rilevanza ai modelli di misurazione del rischio di credito sviluppati internamente dalle banche, testando la loro affidabilità e integrità.

La nuova riforma si fondava su tre pilastri mutuamente rinforzanti:

1) *Pillar I: requisiti minimi di capitale* – Sono state previste nuove regole di calcolo dei requisiti patrimoniali tali per cui è sempre richiesto che il capitale di vigilanza sia pari ad almeno l'8% degli assets ponderati per i rischi, con la differenza che ora è necessaria per la prima volta anche la copertura del rischio operativo oltre al rischio di credito e di mercato;

2) *Pillar II: processo di supervisione dell'autorità di vigilanza* – È stata prevista la possibilità per l'autorità di vigilanza di imporre requisiti di capitale più elevati di quelli previsti dalla regolamentazione;
3) *Pillar III: disciplina di* mercato – È stato previsto un rafforzamento della disciplina esercitata dal mercato dei capitali poiché è richiesto che le banche soddisfino particolari criteri di disclosure e trasparenza nei confronti del mercato finanziario.

È già stato specificato precedente che la regolamentazione proposta nell'accordo del 1988 si è posta come principale obiettivo l'assicurazione della solvibilità del sistema bancario internazionale e la promozione di un contesto competitivo uniforme tra le banche nei diversi paesi. Tuttavia, la nuova regolamentazione emanata nel 2004 si propone di perseguire un ulteriore obiettivo oltre a quelli appena citati: la promozione di requisiti di capitale più sensibili all'effettivo livello di rischio a cui sono esposti i portafogli delle banche, riducendo il divario tra il *capitale economico*, misurato dai modelli interni delle banche, e il *capitale regolamentare*, imposto dal Comitato di Basilea.

1.2.1 Pillar I: Requisito di capitale

Nella nuova regolamentazione proposta dal Comitato nel 2004 il vincolo con la soglia dell'8% non muta, tuttavia ora al capitale di vigilanza è richiesto di coprire anche il rischio operativo. Infatti, la banca iniziò ad essere percepita come un'attività di impresa, quindi in quanto tale esposta anch'essa a rischi come gli errori commessi nel trattamento dei dati, nella gestione delle pratiche burocratiche, nell'utilizzo dei sistemi informativi con la possibilità di innescare guasti tecnici, ecc.

Per quanto riguarda il rischio di credito, il Pillar I costituisce una rottura rispetto al passato poiché i prestiti accesi dalla banca a controparti simili richiedono requisiti di capitale differenti a seconda della rischiosità intrinseca di queste ultime, valutata da alcune agenzie di rating esterne (*approccio standard*) oppure dalla banca stessa (*approccio del rating interno*).

1.2.1.1 Approccio standard

Nell'approccio standard l'ammontare di capitale regolamentare richiesto per un prestito concesso ad una controparte dipende dal rating attribuito a tale controparte da una o più istituzioni esterne di valutazione di crediti. Tali istituzioni possono essere agenzie di rating

oppure altre istituzioni riconosciute dall'Autorità di Vigilanza e che devono soddisfare particolari requisiti minimi in termini di obiettività, indipendenza, trasparenza, pubblicità delle informazioni, risorse e credibilità.

Per quanto riguarda la ponderazione dei crediti, nella regolamentazione del 2004 a rating migliori sono associati pesi minori, come è possibile osservare alla figura 1.2, dove è possibile consultare una panoramica della struttura di attribuzione dei pesi proposta in questa nuova regolamentazione (da una parte sono presenti le categorie di controparti e dall'altra i rating secondo Standard and Poor's).

	AAA	AA A–	AA+	AA	AA–	A+	A	A–	BBB+	BBB	BBB–	BB+	BB	BB–	B+	B	B–	Below B–	Unrated	Past-due
Corporates	20%					50%			100%						150%				100%	150%
Sovereign entities	0%					20%			50 %			100%						150%	100%	
Banks	20%					50%						100%						150%	50%	
Banks, depending on the country of incorporation	20%					50%			100%									150%	100%	
Retail	75%																			
Residential real estate mortganges	35%																			100%
Non-residential real estate mortgages	From 100% to 50%, upon discretion of the national supervisory authorities																			150%

Figura 1.2: coefficienti di ponderazione dell'approccio standard (fonte: "Rischio e valore nelle banche" di A. Resti e A. Sironi)

Inoltre, l'approccio standard prevede anche uno specifico sistema di ponderazioni per le operazioni di *securitization*: tale sistema richiede consistenti requisiti patrimoniali alle banche che investono nelle tranche junior o equity, le quali sono spesso sottoscritte dalla banca originator per agevolare il collocamento dei restanti titoli dello *special purpose vehicle*, e mira ad evitare lo sviluppo incontrollato di un mercato di prestiti molto rischiosi (i cosiddetti *junk loan*).

1.2.1.2 Approccio del rating interno

L'utilizzo del presente approccio ai fini del calcolo dei requisiti patrimoniali minimi si basa sull'utilizzo dei sistemi di rating interni (SRI), dove con *sistema di rating interno* ci si riferisce all'insieme di metodi, dei processi organizzativi e di controllo che permette la raccolta e l'elaborazione delle informazioni significative per la valutazione della rischiosità delle singole esposizioni creditizie.

Le banche che applicano l'approccio del rating interno hanno l'obbligo di approvazione del proprio sistema di misurazione del rischio da parte dell'autorità di vigilanza e, inoltre,

sono direttamente responsabili per la stima del grado di rischiosità associato ad ogni singolo credito presente nel loro portafoglio. In tale contesto la regolamentazione di Basilea II identifica sei fattori di rischio, dettagliati nella tabella 1.3, che sono ritenuti in grado di determinare plausibilmente l'estensione delle perdite future relativamente alle esposizioni creditizie.

Fattore	*Significato*	*Caratteristiche*	*Note*	*Soggetto qualificato alla stima*
PD	Probabilità che la controparte risulti insolvente	Calcolata su un orizzonte temporale di 12 mesi, ma tenendo conto di possibili deterioramenti congiunturali	Si ha default quando il debitore è *"unable or unwilling to pay in full"* e comunque dopo un ritardo di oltre 90/180 giorni	La banca, se in possesso di un sistema di rating interno validato dalle autorità di vigilanza
LGD	Tasso unitario di perdita in caso di insolvenza	Calcolato tenendo conto dei costi di recupero del contenzioso e del valore finanziario del tempo	Risente della forma tecnica e della presenza di garanzie reali	Le autorità, oppure la banca, se in possesso di un sistema di rating avanzato validato dalle autorità di vigilanza
EAD	Esposizione della banca al momento dell'insolvenza	Calcolata tenendo conto dei margini disponibili su linee di credito per cassa e per firma	Costante per le forme tecniche con piano d'ammortamento prestabilito	
Vita residua o maturity (M)	Vita residua del credito	Calcolata come "duration", cioè tenendo conto dei rimborsi previsti prima della scadenza finale		
Granularità	Tendenza a erogare pochi, grandi crediti oppure molte piccole esposizioni	Non calcolata, ma fissata a priori (si suppone sia infinita)	Possibili correttivi nell'ambito del "secondo pilastro"	Le autorità di vigilanza
Correlazione	Tendenza dei diversi debitori a fallire insieme	Non calcolata, ma fissata a priori (valori diversi per diverse tipologie di clientela)		

Tabella 1.3: fattori di rischio relativi al metodo dei rating interni (fonte: "Rischio e valore nelle banche" di A. Resti e A. Sironi)

I primi quattro fattori di rischio presentati nella tabella (PD, LGD, EAD e maturity) rappresentano i parametri fondamentali che un sistema di rating deve adeguatamente misurare. In particolare, a seconda del grado di sofisticatezza del modello adottato e dai dati storici a disposizione, le banche possono essere autorizzate ad utilizzare due differenti approcci:

- *Foundation approach* – Permette alla banca di stimare con proprie metodologie interne soltanto la PD dei debitori, facendo invece riferimento a valori prefissati dalle utorità per quanto riguarda LGD, EAD e maturity;
- *Advanced approach* – Permette alla banca di misurare con metodologie proprie, la cui efficacia e robustezza deve comunque essere adeguatamente dimostrata, tutti e quattro i profili di rischio principali.

La metodologia basata sui sistemi di rating interni costituisce quindi un'alternativa all'approccio standard e, nello specifico, è stabilito che questa possa essere utilizzata solo qualora il sistema di rating interno adottato dalla banca rispetti i seguenti criteri minimi:

1) Il SRI valuta separatamente la PD e la LGD;
2) I crediti sono distribuiti tra le varie classi di rating, senza concentrazione in una specifica classe;
3) Il rating è assegnato ai debitori prima della concessione del prestito;
4) Il rating è rivisto periodicamente;
5) Il rating va utilizzato dalla banca nella gestione dei crediti e nel pricing dei prestiti;
6) La banca deve disporre di un adeguato sistema di validazione dell'accuratezza e coerenza del SRI;
7) Il SRI rispetta specifici requisiti di documentazione formale e del suo funzionamento.

1.2.2 Dal sistema di rating al capitale minimo obbligatorio

Nell'approccio standard, il patrimonio minimo associato a un'esposizione è determinato semplicemente individuando l'8 per cento degli attivi ponderati per il rischio utilizzando il sistema di pesi descritto nella precedente figura 1.2. Invece, nell'approccio basato sui rating interni il meccanismo per trasformare le caratteristiche di rischio di un prestito (PD, LGD, EAD, maturity) e del relativo portafoglio (granularità e correlazione) in un requisito patrimoniale è più complesso e fa uso di un modello VaR sul rischio di credito.

1.2.2.1 Il modello di riferimento

Il modello di riferimento per determinare il requisito di capitale è determinato partendo da un portafoglio di crediti composto da un elevato numero di piccoli prestiti, ossia un portafoglio definibile "infinitamente granulare". A questo punto si immagina, in linea col modello di Merton, che ogni imprenditore fallisca se e solo se il valore delle sue attività scende al di sotto di una certa soglia, ad esempio il valore dei debiti, al termine di un dato orizzonte temporale (tale concetto è visualizzabile graficamente alla figura 1.3). Inoltre, si immagina che il cambiamento percentuale che si verificherà nel prossimo anno nel valore degli attivi dell'i-esimo debitore possa essere espresso come:

$$Z_i = w \cdot Z - \sqrt{1 - w^2} \cdot \varepsilon_i \tag{1.3}$$

cioè come una combinazione lineare[2] di due componenti:

- Fattore Z: corrisponde al ciclo macroeconomico (quindi è uguale per tutti i debitori);
- Fattore ε_i: dipende dal rischio individuale del titolare del prestito.

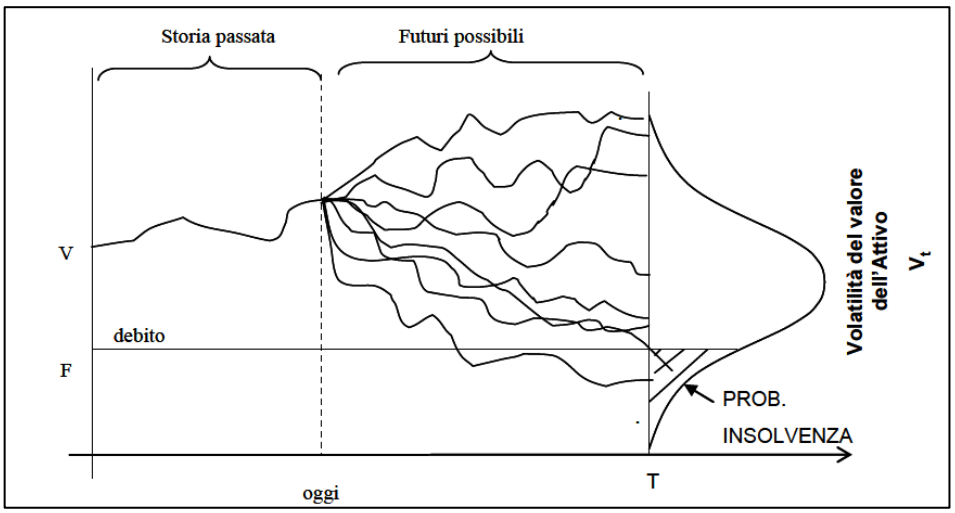


Figura 1.3: rappresentazione grafica del concetto alla base del modello di Merton per la stima della probabilità di insolvenza

[2] Dal momento che Z e ε_i hanno varianza unitaria e la varianza della somma di due termini casuali indipendenti, $var(\alpha x_1 + \beta x_2)$, è sempre pari a $\alpha^2 x_1 + \beta^2 x_2$, allora affinché Z_i segua una distribuzione normale si deve imporre che $\alpha^2 + \beta^2 = 1$. Nella fattispecie, tale condizione è stata imposta ponendo il secondo peso pari alla radice di $1 - w^2$, dove w è il primo peso.

Data tale struttura si deriva che al crescere di *w* tutti i debitori tendono ad essere sempre più correlati tra loro, mentre al diminuire di *w* le caratteristiche individuali di ciascun titolare del prestito tendono a prevalere rendendo questi ultimi sempre più indipendenti.

Se a questo punto si assume che Z e ε_i seguano una distribuzione normale standard, allora dall'equazione 1.4 si può affermare che Z_i segue anch'esso una distribuzione dello stesso tipo. Per ogni coppia di debitori *i* e *j*, la correlazione tra *asset value return* è data da:

$$\rho(Z_i, Z_j) = w^2 \tag{1.4}$$

Infatti, quanto più è elevata la dipendenza *w* delle attività di ogni azienda dal ciclo macroeconomico, tanto più alta sarà la correlazione ρ tra l'andamento delle attività delle due imprese. Dal momento che il debitore *i* diviene insolvente solo se $Z_i < \alpha$, dove α indica il suo *default point*, se si indica con $p_i = PD$ la probabilità di insolvenza di tale debitore non condizionata dal valore assunto da Z, si avrà allora che $N(\alpha) = p_i$ (figura 1.4), dove $N(\cdot)$ indica la distribuzione di probabilità cumulata normale standard.

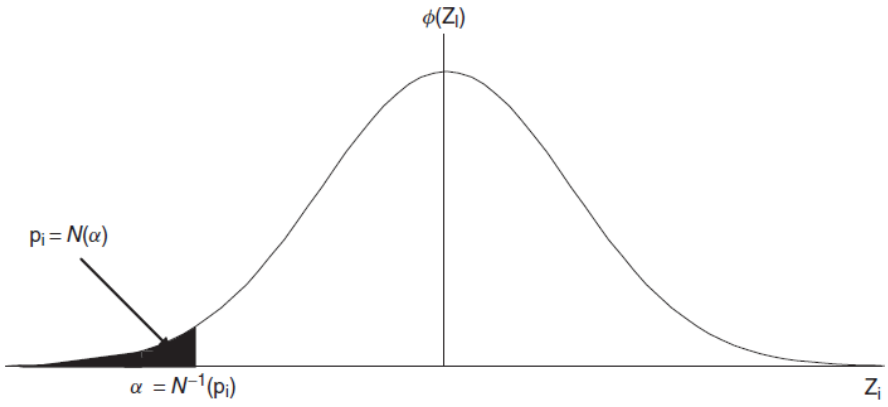


Figura 1.4: individuazione della PD dato il default point α

Se a questo punto si suppone di conoscere l'andamento della congiuntura macroeconomica nel prossimo anno (un'ipotesi irrealistica che verrà rilassata in seguito) questo equivale a supporre di conoscere il valore che verrà assunto dal fattore macroeconomico Z e che viene indicato con Z^*. Quindi:

$$Z_i = w \cdot Z^* - \sqrt{1 - w^2} \cdot \varepsilon_i \tag{1.5}$$

e l'azienda *i* diverrà insolvente solo se:

$$Z_i = w \cdot Z^* - \sqrt{1 - w^2} \cdot \varepsilon_i < \alpha \tag{1.6}$$

ovvero se:

$$\varepsilon_i < \frac{\alpha - w \cdot Z^*}{\sqrt{1-w^2}} = \frac{N^{-1}(p_i) - w \cdot Z^*}{\sqrt{1-w^2}} \tag{1.7}$$

Dal momento che ε_i segue una distribuzione normale standard, la probabilità di insolvenza per il debitore *i* condizionata a $Z = Z^*$ è:

$$p_i|_{Z=Z^*} = N\left[\frac{N^{-1}(p_i) - w \cdot Z^*}{\sqrt{1-w^2}}\right] = f(Z^*; p_i, w) \tag{1.8}$$

Pertanto, la probabilità condizionata è una funzione $f(\cdot)$ di Z, della PD non condizionata p_i e del parametro *w*. In particolare, la probabilità condizionata mostrata nella precedente equazione può essere considerata, supponendo un'esposizione di 1 euro e una LGD del 100 per cento, come *la perdita che il nostro portafoglio crediti dovrà effettivamente sopportare se il fattore macroeconomico assumerà il valore* Z^*.

Negli sviluppi riportati finora rimane tuttavia persiste una problematica non di poco conto, ovvero che non è possibile conoscere a priori il valore assunto da Z. Tuttavia, dal momento che si sa che Z segue una distribuzione normale standard dove l'*x*-esimo percentile è dato da $Z_x|N(Z_i) = x$, allora si può utilizzare l'equazione 1.8 per individuare un valore di perdita *L* che verrà superato soltanto nell'x per cento dei casi. In particolare, tale valore è dato da:

$$L = f(Z_x; p_i, w) = p_i|_{Z=Z^*} = N\left[\frac{N^{-1}(p_i) - w \cdot N^{-1}(x)}{\sqrt{1-w^2}}\right] = g(x, p_i, w) \tag{1.9}$$

Tale equazione fornisce l'ammontare di capitale e riserve necessario per fronteggiare l'$1 - x$ per cento di tutte le possibili perdite, sempre sotto l'ipotesi che la LGD sia pari al 100 per cento. Nel caso in cui si voglia rilassare quest'ultima ipotesi, la precedente equazione diventa:

$$L = LGD \cdot g(x, PD, w) = LGD \cdot N\left[\frac{N^{-1}(PD) - w \cdot N^{-1}(x)}{\sqrt{1-w^2}}\right] \tag{1.10}$$

In particolare, il Comitato di Basilea ha optato per un valore di *x* pari allo 0,1 per cento, accettando dunque che il capitale e le riserve prescritti dalla normativa del primo pilastro possano non essere sufficienti in un caso su mille. Ulteriori precisazioni sull'argomento saranno trattate nel capitolo 2.3 dopo aver discusso due differenti concetti di perdita introdotti nell'Accordo del 2004: la perdita attesa e la perdita inattesa.

1.2.3 Pillar II: un nuovo ruolo per l'Autorità di Vigilanza

Anche se si sviluppasse il più sofisticato modello tramite l'approccio del sistema di rating interno, il Pillar I fallirebbe in ogni caso a modellare con sufficiente flessibilità alcuni aspetti cruciali riguardanti il rischio di credito come la concentrazione e la correlazione delle esposizioni creditizie. Inoltre, il livello di efficacia dei modelli sviluppati internamente dalle banche dipende in modo sostanziale dalle soluzioni organizzative messe in atto dalle singole banche e dal livello di coinvolgimento del top management delle stesse. Si rivela dunque importante, al fine di risolvere tali criticità, che l'Autorità di Vigilanza effettui una revisione del processo di misurazione dei rischi attuato dalle singole banche verificandone la solidità, l'appropriatezza e richiedendo se necessario un *capital buffer* addizionale. Il Comitato di Basilea ha voluto rimarcare l'importanza della supervisione bancaria costituendo il Pillar II, il quale si fonda su quattro principi fondamentali che guidano il processo di controllo prudenziale svolto dall'Autorità di Vigilanza:

1) Le banche devono adottare un sistema di processi e di tecniche al fine di stabilire l'adeguatezza patrimoniale in funzione del loro profilo di rischio. È quindi opportuno per le banche effettuare degli *stress test* al fine di valutare l'impatto di scenari sfavorevoli che potrebbero influenzare negativamente la salute patrimoniale della banca;
2) Le autorità di vigilanza nazionali devono revisionare periodicamente i processi interni di valutazione dell'adeguatezza patrimoniale, dei rischi assunti e la qualità del capitale;
3) Le autorità di vigilanza devono aspettarsi che le banche operino con un ammontare di capitale di vigilanza superiore al minimo richiesto dalla regolamentazione. In particolare, le autorità di vigilanza possono richiedere agli istituti bancari di detenere un ammontare di capitale di vigilanza maggiore di quello richiesto dalla regolamentazione per fini prudenziali e per attenuare possibili effetti negativi prodotti dalle incertezze che colpiscono i diversi mercati e il settore bancario in generale;
4) Le autorità di vigilanza possono intervenire prontamente al fine di evitare che il capitale di vigilanza scenda sotto il minimo richiesto dalla regolamentazione tramite delle azioni correttive, quali l'intensificazione della vigilanza, delle restrizioni al pagamento dei dividendi o la definizione di un piano di ricostituzione patrimoniale.

1.2.4 Pillar III: Market Discipline

Il Pillar III stabilisce i requisiti informativi che devono essere resi noti dalle banche affinché gli operatori dei mercati finanziari siano in grado di valutare l'operatività, le esposizioni ai rischi e l'adeguatezza del patrimonio in funzione dei rischi. Il motivo per cui sono state avanzate queste richieste di disclosure nei confronti delle banche è perché queste svolgono una tipologia di business del tutto particolare:

- Mostrano un alto livello di opacità che rende complesso valutare correttamente i rischi relativi agli investimenti effettuati;
- Sono finanziate dai depositanti, i quali tipicamente non sono in grado di valutare adeguatamente i rischi collegati alla loro banca di riferimento;
- Hanno un ruolo importante nel sistema economico poiché costituiscono un canale per l'attuazione di politiche monetarie e gestiscono una larga parte del sistema dei pagamenti. Per tali ragioni alle banche può essere consentito l'utilizzo di speciali fondi da parte delle banche centrali, i quali creano un certo margine di sicurezza che scoraggia i creditori della banca a valutare la sua solidità.

Per tali ragioni i creditori delle banche difficilmente saranno sufficientemente efficienti ad effettuare le verifiche che le stesse banche effettuano nei confronti delle imprese e dei privati verso cui accendono dei prestiti e che riguardano il monitoraggio del livello di rischio assunto, le condizioni relative al tasso di interesse applicato oppure la negazione della concessione del credito quando il rischio risulta essere troppo elevato. Quello che il terzo pilastro tenta di fare è di rimuovere i fattori che rendono difficile l'applicazione di una disciplina di mercato alle banche. In particolare, il modo attraverso il quale si intende rendere possibile tale disciplina di mercato è tramite il soddisfacimento di specifici criteri di disclosure, fornendo agli investitori un preciso e dettagliato report di informazioni relative ai rischi e al capitale della banca al fine di mettere i creditori della banca in una posizione in cui è risulta più semplice effettuare una valutazione sull'attuale livello di rischio della propria banca di riferimento.

1.2.5 Limiti di Basilea II e conclusioni

L'Accordo del 2004, anche se è poi entrato effettivamente in vigore nel 2008, è considerato tra i colpevoli della severità della crisi finanziaria scoppiata nel secondo semestre del 2007. Tra le principali limitazioni e problemi si distinguono i seguenti punti:

- *Ponderazioni per il rischio nell'approccio standard* – Alcuni studi empirici hanno dimostrato come le ponderazioni per il rischio previste per le diverse fasce di rating nell'approccio standard risultano relativamente poco differenziate rispetto a quanto emerge dai dati relativi ai tassi di insolvenza storici e agli spread delle obbligazioni societarie. Ad esempio, i dati storici di Moody's relativi ai tassi di insolvenza a un anno della classe B3 sono circa 100 volte superiori a quelli relativi alla classe Baa1, tuttavia per la classe B3 è prevista una ponderazione per il rischio pari al 150 per cento mentre per la classe Baa1 una ponderazione pari al 100 per cento;
- *Efficacia della disciplina di mercato* – Se da un lato è vero che la disciplina affronta il problema della trasparenza attraverso degli obblighi di disclosure, dall'altro però non affronta i problemi legati agli incentivi per i creditori e delle banche a esercitare un'efficace azione di monitoraggio del rischio;
- *Prociclicità* – Con tale termine si fa riferimento al timore che il nuovo sistema di requisiti patrimoniali relativi al rischio di credito, fondati sui sistemi di rating interni, possa accentuare le fluttuazioni del ciclo economico e ripercuotersi negativamente sulla stabilità del sistema bancario. Infatti, dal momento che i coefficienti patrimoniali dipendono dai rating (interni o esterni) delle controparti, allora un'eventuale recessione, portando con sé tassi di insolvenza più elevati e downgrading più frequenti, condurrebbe a un aumento del capitale minimo richiesto alle banche. Dal momento che risulterebbe difficile raccogliere nuovo patrimonio durante una recessione, per mantenere le proporzioni tra capitale e attivi le banche finirebbero per concedere meno credito all'economica e ciò avrebbe come conseguenza quella di esporre le imprese a ulteriori tensioni finanziarie, accentuando così a recessione. Analogamente, in una fase di forte crescita economica e quindi di miglioramento del merito creditizio delle controparti, i coefficienti patrimoniali si allenterebbero consentendo alle banche di aumentare oltre misura l'offerta di credito all'economia.

L'Accordo del 2004, denominato Basilea II, è stato un intervento normativo ampio e capillare che ha fatto non solo da cornice normativa con alcune delle migliori metodologie in materia di rischi e capitale, ma è stato anche uno strumento di ulteriore crescita delle capacità di risk e capital management degli intermediari finanziari seppur presentando anche le criticità appena descritte.

1.3 Basilea III: l'attuale regolamentazione

Mentre le banche stavano applicando i principi previsti dalla regolamentazione di Basilea II, nel secondo semestre del 2007 è scoppiata una crisi finanziaria che è stata in grado di mettere in ginocchio i regolatori dell'epoca. Infatti, ci si è resi immediatamente conto del fatto che Basilea II, seppur rappresentando un passo in avanti verso una maggiore razionalità della regolamentazione bancaria, necessitava di una serie di provvedimenti correttivi al fine di scongiurare non solo eventuali arbitraggi regolamentari, ma soprattutto un'altra crisi come quella del 2007, in cui la sopravvivenza degli intermediari finanziari si è pesantemente basata sugli interventi a copertura delle perdite da parte dell'organismo pubblico. A seguito di tale periodo nero per l'economia mondiale il Comitato di Basilea si è nuovamente riunito per revisionare la regolamentazione vigente e proporre una nuova regolamentazione: Basilea III. Tale nuovo accordo mantiene la struttura a tre pilastri proposta da Basilea II ma ne rafforza le basi. In particolare, gli interventi di Basilea III hanno riguardato sei aspetti principali:

1) Miglioramento della qualità del capitale;
2) Introduzione di un *Capital Conservation Buffer*;
3) Introduzione di un *Countercyclical Buffer*;
4) Introduzione di una soglia massima al leverage;
5) Introduzione di due coefficienti di liquidità;
6) Nuovi requisiti per il rischio di mercato e per quello di controparte.

Nei seguenti capitoli verrà approfondito ognuno di questi aspetti.

1.3.1 Miglioramento della qualità del capitale

Il Comitato di Basilea ha previsto i seguenti interventi al fine di migliorare la qualità del capitale regolamentare[3]:

1) Innalzamento del requisito di Core Tier 1 (capitale sociale, sovrapprezzi e riserve da utili non distribuibili) al 4,5 per cento degli assets ponderati per i rischi;

[3] Nell'accordio di Basilea III il capitale regolamentare è suddiviso in Core Tier 1 (capitale sociale, riserve da utili ed altre riserve), Additional Tier 1 (strumenti finanziari senza scadenza prefissata) e Tier 2 (debito subordinato e riserve generiche per perdite su crediti).

2) Implementazione di deduzioni prudenziali dal Core Tier 1 di azioni proprie, di imposte anticipate, del patrimonio di terzi, di partecipazioni non consolidate, di avviamenti ed altri assets intangibili;
3) L'Additional Tier 1 include azioni le privilegiate irredimibili senza diritto di proprietà cumulativo;
4) La somma del Core Tier 1 e dell'Additional Tier 1 corrisponde al Tier 1 Capital;
5) Graduale esclusione dal 2013 dal patrimonio di strumenti ibridi e simili (privi di clausole bail-in);
6) Eliminazione del Tier 3.

Pertanto, il Tier 1 Capital rappresenta il patrimonio in grado di assorbire le perdite quando la banca è in condizioni di continuità aziendale, mentre il Tier 2 Capital assorbe le perdite quando la banca è in condizioni di crisi.

1.3.2 Il Capital Conservation Buffer

La regolamentazione di Basilea ha previsto dal 2016 che tale buffer debba essere pari al 2,5 per cento delle attività ponderate per i rischi al fine di garantire un cuscinetto di capitale utile all'assorbimento delle perdite che potrebbero insorgere nelle fasi di elevata turbolenza. In particolare, è previsto che nelle fasi positive del ciclo economico le banche accumulino tale Conservation Buffer fino al raggiungimento della soglia del 2,5 per cento (portando quindi il Core Tier 1 al 7% delle attività ponderate per i rischi)[4], mentre nelle fasi negative del ciclo economico le banche potranno utilizzare tale buffer per assorbire le perdite, pur garantendo un limite minimo del Core Tier 1 del 4,5 per cento delle attività ponderate per i rischi.

1.3.3 Il Countercyclical Buffer

Si tratta di un cuscinetto pari al fino al 2,5 per cento delle attività ponderate per i rischi e la cui richiesta di costituzione, la quale deve avvenire con un anno di anticipo per prevenire shock sui mercati ed agli intermediari, è a discrezione degli organismi di vigilanza nazionali a seguito di un'analisi sul grado di surriscaldamento del ciclo creditizio.

[4] Sono previsti dei limiti alla distribuzione degli utili fintanto che il cuscinetto non è stato completamente costituito.

Il Capital Conservation Buffer e il Countercyclical Buffer sono dei nuovi meccanismi introdotti per contrastare il problema della prociclicità, discusso precedentemente nelle limitazioni di Basilea II.

Un'altra richiesta avanzata da Basilea III al fine di contrastare i problemi legati alla prociclicità riguarda l'invito agli organismi contabili per introdurre criteri di calcolo delle svalutazioni di crediti più *forward-looking*: le perdite su crediti che si verificano nelle fasi negative del ciclo economico si originano nelle concessioni di prestiti effettuate nelle fasi positive del ciclo economico ed è proprio in quelle fasi che dovrebbero essere contabilizzati gli accantonamenti.

1.3.4 Nuovi coefficienti di liquidità

La crisi che ha avuto inizio nel secondo semestre del 2007 ha messo in luce importanti criticità relative alla liquidità dei mercati finanziari dal momento che ci si è resi conto che un'importantissima fonte di diffusione del contagio finanziario è stata proprio la carenza di liquidità. Per questa ragione il Comitato di Basilea ha deciso di introdurre nel nuovo accordo due nuovi indicatori per misurare il grado di liquidità.

Il primo indicatore è il *Liquidity Coverage Ratio*, un indicatore col quale viene definito che le attività liquide di alta qualità devono essere in grado di far fronte ai deflussi di cassa attesi per i successivi trenta giorni e stimati in base ad uno scenario di stress. Inoltre, è richiesto che il vincolo venga rispettato su base continuativa. In formule:

$$\text{Liq. Cov. Ratio} = \frac{\text{Attività liquide di alta qualità}}{\text{Deflussi di cassa attesi nei succ. 30 gg}} \geq 1 \qquad (1.11)$$

Le attività che rientrano nella determinazione del numeratore sono attività con un basso livello di rischio di credito e di mercato, aventi una valutazione certa e una bassa correlazione con attività rischiose.

Il secondo indicatore proposto dalla nuova regolamentazione è il *Net Stable Funding Ratio*, un indicatore col quale si prevede che debba esserci un rapporto di equilibrio (ovvero superiore a 1), tra le fonti di finanziamento stabili a medio-lungo termine e i fabbisogni di fondi a medio-lungo termine. In formule:

$$\text{Net Stable Funding Ratio} = \frac{\text{Fonti di finanziamento stabili a MLT}}{\text{Fabbisogni di fondi a MLT}} > 1 \qquad (1.12)$$

Tale requisito della regolamentazione ha l'effetto di ridurre i gradi di libertà delle banche nella trasformazione delle scadenze.

1.3.5 Nuovi requisiti per il rischio di mercato e di controparte

Infine, un ultimo aspetto introdotto dalla regolamentazione di Basilea III riguarda l'inasprimento dei requisiti per i portafogli di trading e il rafforzamento del trattamento prudenziale dei rischi connessi alla cartolarizzazione ed ai veicoli fuori bilancio. Inoltre, è stato reso più stringente il calcolo dei requisiti sui rischi di controparte nell'operatività sui derivati.

Con questi accorgimenti la regolamentazione ha voluto imporre un trattamento patrimoniale più cautelativo nei confronti di tutti quegli strumenti finanziari innovativi annoverati tra i responsabili della crisi del 2007.

1.3.6 Considerazioni su Basilea III

L'accordo di Basilea III, nato in risposta agli eventi scatenati dalla crisi finanziaria iniziata nel secondo semestre del 2007, ha influenzato in maniera significativa sulla quantità aggregata di credito disponibile. Infatti, le disposizioni regolamentari di Basilea III rispetto agli accordi delle regolamentazioni precedenti hanno comportato per le banche una detenzione di livelli di capitale più alti a fronte delle proprie attività. Tale situazione è vero che da un lato ha migliorato la stabilità del sistema bancario internazionale e la solidità dei singoli istituti di credito, ma dall'altra parte ha anche comportato un aumento del costo delle fonti di finanziamento e, di conseguenza, un aumento dei tassi di interesse applicati alla clientela, il quale ha a sua volta innescato una contrazione della domanda di credito da parte di famiglie e imprese, soprattutto in quei sistemi economici cosiddetti *bank-oriented* come l'Italia o la Germania.

Inoltre, un altro aspetto importante legato alla regolamentazione proposta da Basilea III riguarda l'attenuazione degli effetti della procicilicità tramite l'introduzione di un buffer prudenziale e di un buffer anticiclico. Infatti, durante la crisi finanziaria l'inasprimento dei requisiti di capitale ha contribuito all'instabilità del sistema finanziario, quindi è stato deciso di sfruttare i periodi positivi del ciclo economico per accumulare risorse all'interno

delle riserve al fine di coprire eventuali perdite future in periodi negativi del ciclo economico.

1.3.7 Soglia massima al leverage

Al fine di limitare la leva finanziaria e contenere gli eventuali errori di misurazione del sistema di ponderazione dei rischi è richiesto che il rapporto tra il patrimonio (Tier 1) e l'attivo contabile, comprensivo delle partite fuori bilancio (impegni, fidejussioni, accettazioni, lettere di credito, ecc.) sia fissato al 3 per cento. Tale rapporto è detto *Leverage Ratio* e in formule può essere espresso nel seguente modo:

$$\text{Leverage Ratio} = \frac{\text{Tier 1}}{\text{Attivo contabile (incluse partite fuori bilancio)}} \geq 8\% \qquad (1.13)$$

1.4 Basilea IV: le prospettive future della regolamentazione

Nel 2017 il Comitato di Basilea si è nuovamente riunito per elaborare ulteriori riforme che entreranno progressivamente in vigore tra il 2021 e il 2027. In particolare, il framework normativo di Basilea IV prevede:

- Una riduzione della complessità delle regole prudenziali per la gestione dei rischi;
- L'incremento della sensibilità delle diverse componenti di rischio;
- L'adozione di nuovi Modelli Standard per la misurazione dei rischi creditizi, di mercato e operativi maggiormente articolata e granulare;
- Il contenimento dell'applicazione dei Modelli Avanzati, imponendo dei limiti quantitativi minimi e rendendoli meno "personalizzabili", quindi più confrontabili tra le banche. In particolare, è prevista l'adozione di parametri non inferiori a valori minimi prestabiliti di PD, LGD, EAD e maturity per quantificare la corrispondente perdita attesa. Inoltre, le banche che hanno optato per l'adozione di modelli interni di risk management dovranno confrontare i valori complessivi degli asset ponderati per i rischi con quanto calcolato dal relativo Modello Standard.

Gli osservatori economici e alcune simulazioni dell'EBA (*European Banking Authority*) valutano che a regime le banche, a seguito dell'effettiva entrata in vigore del nuovo framework normativo proposto da Basilea IV, saranno sottoposte a richieste aggiuntive di capitale proprio. In particolare, tale previsione vale soprattutto per le banche di importanza

sistemica e per quelle che utilizzano sistemo di rating interni poiché saranno costrette a dover allineare le loro risultanze di rischio con quanto quantificato dall'applicazione dei nuovi Modelli Standard, rispettando comunque un *output floor* stabilito.

In conclusione, dal momento che l'avvento della regolamentazione proposta da Basilea IV renderà meno elastica l'operatività delle banche, si prevede una maggior propensione degli istituti di credito a indirizzare le loro disponibilità di capitale verso opportunità più remunerative, mentre sarà incentivato l'utilizzo di mezzi propri o di strumenti alternativi di accesso al credito.

2. Il rischio di credito

Nel capitolo precedente è stato messo in evidenza come il Comitato di Basilea negli anni abbia sempre più messo al centro l'importanza della vigilanza bancaria nei confronti delle diverse tipologie di rischio che possono mettere a repentaglio la stabilità del sistema bancario internazionale. In particolare, fra tutte le tipologie di rischio a cui è soggetta la banca, il rischio di credito rappresenta la principale causa di crisi.

In questo capitolo si dettaglieranno inizialmente gli elementi fondamentali che caratterizzano il rischio di credito e, successivamente, verrà fatto un approfondimento sulle diverse metodologie con cui è possibile effettuare delle valutazioni di credit scoring e con le quali è possibile quantificare il livello di rischio associato a diverse esposizioni creditizie.

2.1 Caratteristiche principali

Con il termine *rischio di credito* si intende la possibilità che una variazione inattesa del merito creditizio di una controparte generi una corrispondente variazione inattesa del valore corrente associato alla relativa esposizione creditizia. Tale definizione incorpora al suo interno tre concetti fondamentali:

1) *Rischio di insolvenza (default) e rischio di migrazione* – Il rischio di credito non è confinato solamente alla possibilità di default della controparte, ovvero quando si interrompono i pagamenti, ma racchiude al suo interno anche la possibilità di deterioramento del merito creditizio di quest'ultima. Infatti, un deterioramento del merito creditizio incrementa lo spread applicato al tasso risk free per una scadenza corrispondente, ossia il *risk premium* con cui si scontano i flussi di cassa futuri per determinare il valore di mercato attuale di un prestito. Pertanto, ne deriva che a livello probabilistico la misurazione e la gestione del rischio di credito non può basarsi su una distribuzione binaria "default" vs. "non default", ma deve piuttosto basarsi su una distribuzione (discreta o continua) nella quale l'evento "default" rappresenta unicamente l'evento estremo e che è preceduto da altri eventi in cui il debitore rimane solvibile ma con una probabilità di una sua insolvenza futura che è via via più elevata;
2) *Rischio inteso come un evento inatteso* – Un secondo concetto sottinteso nella definizione precedente è la variazione del merito creditizio della controparte, al fine considerata un rischio, deve essere un evento inatteso. Infatti, se una banca

avesse acceso un prestito verso una certa controparte e per cui sa già che in futuro ci sarà un deterioramento della sua qualità (profittabilità, solvibilità o liquidità), allora tale deterioramento dovrebbe già essere stato preso in considerazione nella valutazione ex ante di accensione del prestito e nel suo pricing, ossia nella determinazione del tasso di interesse debitore applicato. Pertanto, la reale componente di rischio è rappresentata dalla possibilità che le valutazioni effettuate ex ante si rivelino a posteriori errate, ossia che si verifichi un deterioramento della qualità della controparte che non era stata prevista precedentemente. Ne deriva che il concetto di rischio riguarda solo gli eventi che seppur stimabili risultano comunque inattesi;

3) *Esposizione creditizia* – Un ultimo concetto relativo alla precedente definizione risiede nel fatto che con rischio di credito non ci si limita alle "classiche" forme di credito concesse dalle banche nei confronti dei loro clienti e che vengono riportate nelle poste a bilancio, ma include anche tutte le operazioni fuori bilancio come ad esempio le garanzie prestate, gli strumenti derivati negoziati over the counter, ecc.

Un'ulteriore considerazione riguarda il fatto che la precedente definizione si riferisce al *valore di mercato* dell'esposizione creditizia, determinando di conseguenza le seguenti problematiche:

1) Molte esposizioni creditizie di una istituzione finanziaria sono iscritte a bilancio al costo storico e non al valore corrente (fair value). Una corretta valutazione del rischio di credito e dei suoi effetti richiederebbe invece valutazioni basate sul valore economico dell'esposizione, ossia sul prezzo che un compratore indipendente o un mercato secondario attribuirebbero all'esposizione se questa venisse ceduta dalla banca;
2) La maggior parte delle esposizioni creditizie delle istituzioni finanziarie consiste in assets illiquidi e per i quali non esiste un mercato secondario sviluppato. Pertanto, il valore di mercato può solamente essere stimato sulla base di un modello interno di asset pricing.

2.2 Le componenti del rischio di credito

La definizione di rischio di credito espressa nel capitolo precedente è strettamente connessa alla nozione di *perdita*, un concetto che si fonda sulla distinzione di due differenti componenti:

- *Expected Loss (EL)*;
- *Unexpected Loss (UL)*.

Precedentemente si è fatto riferimento al fatto che a costituire la vera a propria fonte di rischio sia proprio la perdita inattesa (UL), ovvero quelle variazioni nel valore di mercato dell'esposizione creditizia che non erano state previste precedentemente in fase di valutazione di concessione del prestito. La perdita attesa (EL), per contro, è già stata considerata nella fase di valutazione e, oltre che essere incorporata nel valore del tasso di interesse attivo per la banca, è anche eventualmente transitata tramite il conto economico in una specifica riserva per coprire eventuali perdite future previste.

Alla seguente figura 2.1 è possibile visualizzare graficamente i due concetti di perdita appena descritti e che saranno approfonditi nei due capitoli che seguono. In particolare, nella figura a destra è possibile osservare una densità di probabilità in funzione delle perdite potenziali sui crediti in cui si nota la tipica forma asimmetrica, proprio perché la forma della distribuzione delle perdite sui crediti ha la caratteristica di essere asimmetrica (figura a sinistra).

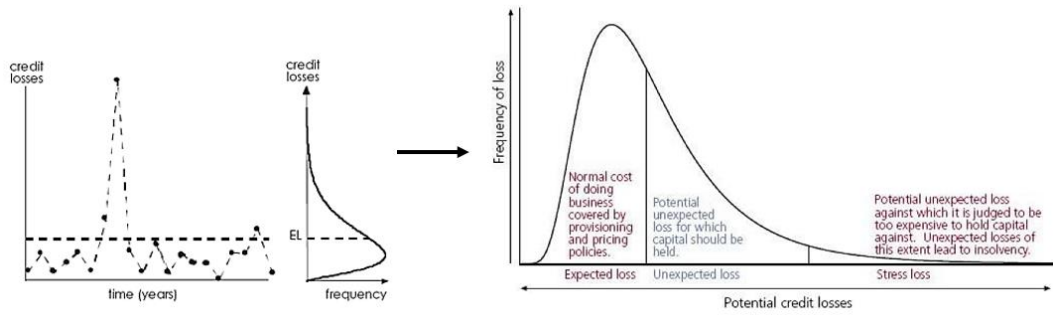


Figura 2.1: distribuzione di probabilità in funzione delle perdite potenziali (a destra) e distribuzione di probabilità caratteristica delle perdite sui crediti (a sinistra)

2.2.1 Expected Loss (EL)

La perdita attesa è rappresentata dal valore medio della distribuzione delle perdite. Come detto precedentemente la perdita attesa non rappresenta in senso stretto un rischio dal momento che essa viene già stimata ex-ante dall'istituto finanziario e che viene coperta applicando un opportuno spread al tasso applicato al prestito e costituendo una riserva in bilancio. In conclusione, se la perdita attesa dovesse effettivamente manifestarsi nella misura prevista l'istituto finanziario conseguirebbe esattamente il rendimento netto che si era originariamente prefissato.

Per effettuare una stima della perdita attesa di un'esposizione creditizia è necessaria a sua volta la stima di tre parametri:

1) Il valore atteso dell'esposizione in caso di insolvenza, anche detta *Exposure At Default* (EAD), ossia una variabile casuale data dall'esposizione corrente più la possibile variazione nella dimensione del prestito che interverrà nel lasso di tempo che va da oggi al possibile evento di default;
2) La probabilità di insolvenza della controparte detta *Probability of Default* (PD);
3) Il tasso di perdita atteso in caso di insolvenza, detto *Loss Given Default* (LGD), ossia la percentuale dell'esposizione che la banca prevede di non riuscire a recuperare. Corrisponde anche a $1 - RR$, dove RR sta per *Recovery Rate*, ossia tasso di recupero atteso sull'esposizione.

Analiticamente si può riassumere quanto appena specificato nel seguente modo[5]:

$$EL = \overline{EAD} \cdot PD \cdot \overline{LGD} \tag{2.1}$$

2.2.1.1 Exposure At Default (EAD)

L'EAD rappresenta una variabile casuale la cui aleatorietà dipende dalla tipologia di finanziamento concessa al richiedente. Infatti, nel caso di un'apertura di credito in conto corrente la banca mette a disposizione una certa quantità di fondi al cliente che sceglie quale porzione del fido utilizzare e questo fa sì che la dimensione effettiva del finanziamento può variare nel tempo per effetto di decisioni esterne alla banca. Tuttavia, è anche vero che in molti casi l'EAD è deterministica e facilmente quantificabile, come ad esempio nel caso di un titolo obbligazionario dove i flussi di cassa sono interamente definiti al momento dell'emissione. In conclusione, quindi, un'esposizione creditizia può essere *a valore certo* (ad esempio titolo obbligazionario, mutuo, ecc.) o a *valore incerto* (ad esempio un fido bancario). In particolare, la stima dell'EAD a valore incerto richiede di conoscere:

- La *Drawn Portion* (DP), ossia la quota di fido utilizzata;
- La *Undrawn Portion* (UP), ossia la quota di fido inutilizzata;

[5] Per semplicità si ipotizza che i tre fattori di rischio che costituiscono la EL siano indipendenti. Infatti, se così non fosse la stima della PD e dei valori attesi di EAD e LGD non sarebbe sufficiente per ricavare la perdita attesa, poiché occorrerebbe conoscere e tenere in considerazione anche le covarianze tra i diversi fattori di rischio.

- Il *Credit Converion Factor* (CCF), ossia la percentuale di fido inutilizzato che ci si attende venga utilizzata al momento del verificarsi dell'evento di insolvenza.

Analiticamente si può determinare l'EAD nel seguente modo:

$$\overline{EAD} = DP + UP \cdot CCF \tag{2.2}$$

2.2.1.2 Probability of Default (PD)

La probabilità di default (PD) può essere definita come la possibilità che la controparte diventi inadempiente alle obbligazioni sottostanti alla propria esposizione creditizia. A seconda delle informazioni a disposizione e, più nello specifico, di quali si desidera tenere conto per il calcolo della PD, si possono distinguere tre diverse metodologie di stima di tale fattore:

- La prima metodologia prevede l'impiego dei dati derivanti dai sistemi di controllo di gestione dell'impresa;
- La seconda metodologia prevede l'impiego di modelli matematici in grado di valutare aspetti sia qualitativi che quantitativi;
- La terza metodologia si avvale dei rating creditizi rilasciati da agenzie di rating specializzate.

2.2.1.3 Loss Given Default (LGD)

La LGD è definita come la perdita subita dall'istituto di credito in caso di insolvenza della controparte e può essere definita come il complemento a uno del *Recovery Rate* (RR), ossia il tasso che l'istituto di credito si aspetta di recuperare in seguito al default del debitore. Come già espresso precedentemente in questa seconda versione la LGD può essere scritta come:

$$LGD = 1 - RR \tag{2.3}$$

Per quanto riguarda il tasso di recupero, questo è influenzato da diversi fattori quali:

- *Caratteristiche del finanziamento* – Esistenza di garanzie reali o finanziarie, liquidità ed efficacia delle garanzie, grado di subordinazione rispetto ad altri creditori, tipi di contenzioso previsto per il recupero;
- *Caratteristiche dell'impresa finanziata* – Settore produttivo e specificità dei beni dell'impresa, paese o regione geografica dell'impresa;

- *Caratteristiche della banca* – Politiche interne di recupero crediti, efficienza dei servizi legali interni;
- *Fattori esterni* – Livello dei tassi di interesse, stato del ciclo economico.

Dal punto di vista analitico il RR è esprimibile nel seguente modo:

$$RR = \frac{\sum_{t=1}^{n} \frac{ER_t - AC_t}{(1+i)^t}}{EAD} \tag{2.4}$$

dove:

- ER_t = *Expected Recovery*, ossia l'importo recuperato nel periodo t;
- AC_t = *Administrative Costs*, ossia i costi amministrativi sostenuti nel periodo t;
- EAD = *Exposure At Default*;
- $i =$ tasso di attualizzazione dei flussi di cassa;
- $T =$ periodo di tempo stimato per realizzare il recupero.

In figura 2.2 sono raffigurate le distribuzioni tipiche del RR e della LGD, ovvero delle distribuzioni bimodali: sono presenti moltissimi casi in cui la banca non recupera praticamente nulla dall'esposizione creditizia, ma allo stesso tempo sono presenti altrettanti casi in cui il recupero invece è totale.

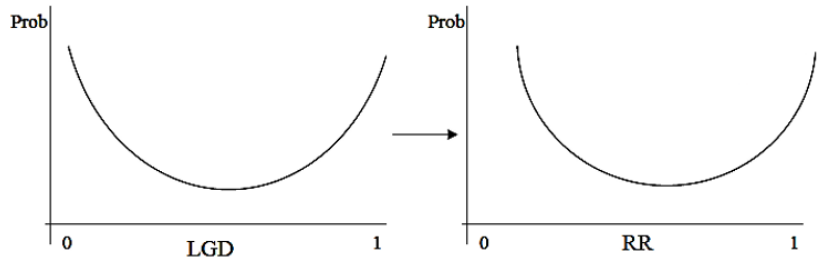


Figura 2.2: rappresentazioni grafiche delle distribuzioni di probabilità della LGD e del RR

2.2.2 Unexpected Loss (UL)

Il vero rischio di credito, ossia il rischio che la perdita a posteriori si dimostri superiore a quella inizialmente stimata, è legato alla perdita inattesa. Questa in generale può essere definita come la variabilità della perdita attorno al suo valore medio, un concetto osservabile in figura 2.3.

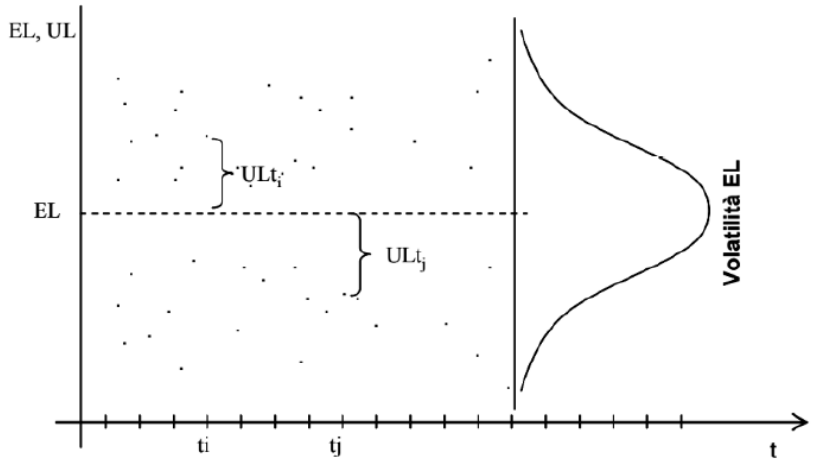


Figura 2.3: definizione della perdita inattesa come volatilità delle perdite attorno alla perdita attesa

La distinzione tra perdita attesa e inattesa risulta particolarmente rilevante quando si considera un portafoglio di impieghi: la perdita attesa di un portafoglio è semplicemente pari alla somma delle perdite attese degli impieghi che lo compongono, mentre la variabilità della perdita totale sul portafoglio è in generale minore della somma delle variabilità delle perdite sui singoli impieghi e, in particolare, è tanto minore quanto più basso è il grado di correlazione tra i vari impieghi. Pertanto, a differenza della perdita attesa che non può essere ridotta diversificando il portafoglio, la perdita inattesa può essere ridotta mediante un'adeguata strategia di ripartizione del rischio e, quindi, a parità di rendimento atteso è possibile ridurre l'ammontare complessivo di rischio di credito.

2.2.2.1 Modello di riferimento

Il modello di riferimento base per la valutazione della perdita inattesa si basa sul modello binomiale costituito da due soli eventi: "default" e "non default". Alla seguente tabella 2.1 è possibile osservare una schematizzazione degli eventi.

	EVENTI	
	Default	Non default
Probabilità	PD	1-PD
Perdita	LGD	0

Tabella 2.1: schematizzazione degli eventi nel modello binomiale

$$MEDIA = EL = PD \cdot LGD + (1 - PD) \cdot 0 = PD \cdot LGD \tag{2.5}$$

$$VOLATILITA' = UL = LGD \cdot \sqrt{PD \cdot (1 - PD)} \tag{2.6}$$

In questo caso l'espressione della volatilità è coerente solo se vengono rispettate le seguenti ipotesi:

- Il tasso di perdita (LGD) è deterministico;
- Vi è indipendenza tra la PD e la LGD.

Nel caso in cui la LGD avesse natura stocastica, ma fosse comunque mantenuta l'indipendenza tra la PD e la LGD, l'espressione della volatilità si modifica nel seguente modo:

$$UL = \sqrt{PD \cdot (1 - PD) \cdot LGD^2 + PD \cdot \sigma_{LGD}^2} \tag{2.7}$$

Infine, se vengono rilassate tutte le ipotesi l'espressione della volatilità assume la seguente forma funzionale:

$$\begin{aligned} UL = COV(PD^2, LGD^2) + [PD(1 - PD) + PD^2] \cdot [\sigma_{LGD}^2 + LGD^2] \\ - [COV(PD, LGD) + PD \cdot LGD]^2 \end{aligned} \tag{2.8}$$

2.3 Quantificazione del requisito di capitale per il sistema basato sui rating interni di Basilea II

Il capitale quantificato tramite l'equazione 1.10 copre ogni possibile perdita futura fino a un livello di confidenza del 99,9 percento. Tale concetto è riportato graficamente alla figura 2.4 in cui è possibile osservare che la copertura patrimoniale suggerita nell'equazione 1.10 fronteggia tutte le possibili perdite tranne lo 0,1 per cento di situazioni peggiori (area scura a destra). Ciò significa che tale copertura include sia le perdite attese (EL) che una certa misura di perdite inattese (UL) destinate a verificarsi solo in determinati scenari estremi:

$$L = EL + UL \tag{2.9}$$

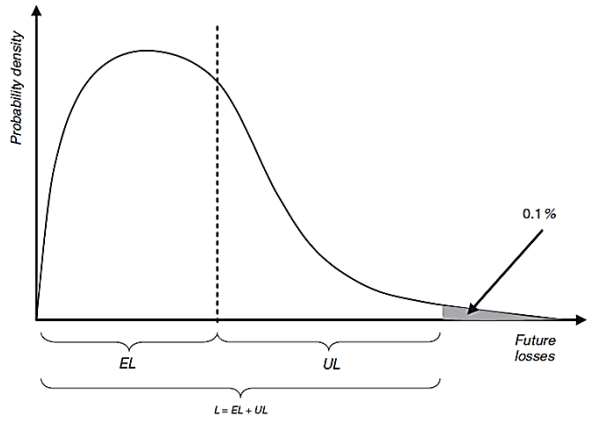


Figura 2.4: perdita attesa e perdita inattesa

È stato visto precedentemente che la EL dovrebbe essere considerata come un costo di produzione piuttosto che come un rischio e che dovrebbe essere fatta transitare in conto economico per essere accantonata come riserva. Per contro, la UL deve essere coperta con il capitale degli azionisti al fine di evitare il fallimento se le perdite effettive dovessero superare il valore atteso. È quindi opportuno distinguere le due componenti presenti nell'equazione 2.9: la perdita attesa, da coprire con riserve, e la perdita inattesa, da fronteggiare con capitale. Si può scrivere che[6]:

$$EL = PD \cdot LGD \tag{2.10}$$

$$UL = LGD \cdot N\left[\frac{N^{-1}(PD) - w \cdot N^{-1}(0{,}1\%)}{\sqrt{1-w^2}}\right] - PD \cdot LGD \tag{2.11}$$

L'Accordo del 2004 prevede alle banche che adottano il metodo dei rating interni di quantificare separatamente le perdite attese e inattese, stabilendo in particolare che:

- La perdita attesa *può* essere coperta attraverso gli accantonamenti a riserva effettuati dalla banca. Qualora tali accantonamenti non siano sufficienti, la quota di EL scoperta può essere fronteggiata con capitale;
- La perdita inattesa *deve* essere coperta con capitale.

L'equazione 2.11 non rappresenta tuttavia la versione finale della formula approvata nell'Accordo del 2004 per il calcolo delle perdite inattese dal momento che resta da considerare l'impatto esercitato sul rischio della vita residua del prestito. Infatti, l'equazione 1.10 e l'equazione 2.11 derivano da un modello binomiale in cui si verificano

[6] È necessario imporre l'ipotesi che PD e/o LGD siano note con certezza, o che PD e LGD siano due variabili casuali indipendenti di cui stiamo utilizzando la media.

perdite su un credito soltanto se questo risulta insolvente. Tuttavia, si è visto che i prestiti a lungo termine possono subire una diminuzione nel loro valore teorico anche in assenza di un vero e proprio default, semplicemente perché il rating del debitore è stato declassato. Pertanto, per tenere conto di questo rischio di downgrading è necessario modificare l'equazione 2.11[7] in funzione della vita residua del prestito (la maturity M) moltiplicandola per il seguente *maturity adjustment factor*:

$$b = \frac{1 + (M - 2{,}5)(\alpha - \beta \ln PD)^2}{1 - 1{,}5 \cdot (\alpha - \beta \ln PD)^2} \tag{2.12}$$

dove α e β sono due parametri pari, rispettivamente, all'11,852 per cento e al 5,478 per cento. Pertanto, la formula che definisce il requisito patrimoniale ($UL^* = b \cdot UL$) diviene:

$$UL^* = b \cdot \left\{ LGD \cdot N\left[\frac{N^{-1}(PD) - w \cdot N^{-1}(0{,}1\%)}{\sqrt{1 - w^2}}\right] - PD \cdot LGD \right\} \tag{2.13}$$

Infine, per evitare che il passaggio al metodo dei rating interni possa ridurre eccessivamente il capitale delle banche il Comitato di Basilea si riserva di inserire nel calcolo del requisito di capitale finale, che verrà chiamato UL^{**}, un fattore di scala σ concordato a livello internazionale. Pertanto, la formula completa per il calcolo della UL^{**} rettificata per il maturity adjustment factor e per il fattore di scala diviene:

$$UL^{**} = \sigma \cdot b \cdot \left\{ LGD \cdot N\left[\frac{N^{-1}(PD) - w \cdot N^{-1}(0{,}1\%)}{\sqrt{1 - w^2}}\right] - PD \cdot LGD \right\} \tag{2.14}$$

Moltiplicando l'equazione 2.14 e l'equazione 2.10 per la EAD è dunque possibile calcolare, rispettivamente, la quantità di capitale e di accantonamenti a riserva richiesti a fronte di ogni prestito.

2.4 Le principali tipologie di rischio di credito

Il rischio di credito comprende le seguenti principali tipologie:

- *Rischio di insolvenza* – È il rischio connesso all'insolvenza della controparte, che dichiara fallimento o comunque smette di onorare regolarmente i pagamenti

[7] Le variazioni del rating hanno per definizione natura inattesa, pertanto solo il valore della UL deve essere corretto per il rischio di downgrading mentre la formula per la perdita attesa resta invariata.

previsti sul prestito. Tale rischio si traduce in una perdita pari al prodotto fra l'esposizione al default (EAD) e il tasso di perdita in caso di insolvenza (LGD);

- *Rischio di migrazione* – Si tratta del rischio connesso a un deterioramento del merito creditizio della controparte;
- *Rischio di spread* – È il rischio connesso a un eventuale rialzo degli spread richiesti dal mercato ai debitori, il quale può essere dovuto anche a eventi esogeni al rapporto tra l'istituto creditizio e la controparte debitrice come, ad esempio, l'ascesa di un improvviso clima di sfiducia e avversione al rischio fra gli investitori;
- *Rischio di recupero* – Indica il rischio che il valore economico dell'ammontare effettivamente recuperato da una controparte divenuta insolvente risulti inferiore a quanto originariamente stimato;
- *Rischio di esposizione* – Indica il rischio che l'ammontare dell'esposizione subisca un incremento in prossimità del default;
- *Rischio di pre-regolamento* o di *sostituzione* – Indica il rischio che la controparte di una transazione in derivati negoziati in mercato over the counter (OTC) divenga insolvente prima della scadenza dello stesso e rende dunque necessario per la banca "sostituire" la posizione sul mercato a condizioni contrattuali differenti;
- *Rischio Paese* – Indica il rischio che una controparte non residente non sia in grado di adempiere alle proprie obbligazioni a causa di eventi di natura politica o legislativa.

3. Il credit scoring

I modelli più diffusi per la previsione dell'insolvenza di un'impresa sono modelli di natura statistica e generalmente noti come modelli di *scoring*. In particolare, si tratta di modelli multivariati[8], ovvero che studiano la variazione simultanea di due o più variabili casuali, che, utilizzando come input i principali indicatori economico-finanziari di un'impresa e attribuendo ad ognuno di essi una ponderazione che riflette la sua rilevanza relativa nel prevedere l'insolvenza, giungono a una valutazione del merito creditizio sintetizzata in un valore numerico rappresentativo della probabilità di insolvenza, lo *score*.

In questo capitolo si analizzeranno tre categorie di modelli di scoring: l'analisi discriminante lineare, i modelli di regressione logistica e le recenti tecniche di machine learning. In particolare, è interessante specificare che mentre le prime due categorie sono fondate su un approccio deduttivo volto a spiegare le cause economiche dell'insolvenza, la terza segue invece un approccio puramente empirico e di tipo induttivo.

3.1 L'analisi discriminante lineare

L'analisi discriminante lineare trova le fondamenta negli studi iniziati da Fisher già nel 1936 e basa il suo funzionamento sull'identificazione delle variabili, tipicamente indicatori economico-finanziari di natura contabile, che consentono di "discriminare" fra imprese sane e imprese anomale, ossia insolventi (fallite o assoggettate a ristrutturazione finanziaria) o classificate come "in sofferenza" dal sistema bancario. Nello specifico, l'analisi discriminante usa informazioni estrapolate dai dati di un campione di imprese per delineare un confine tra imprese sane e insolventi. In figura 3.1 è possibile osservare graficamente l'idea su cu si fonda il modello di Fisher, dove in questo esempio le imprese affidabili (gruppo A) e le imprese insolventi (gruppo B) sono descritte da due sole variabili economico-finanziarie (x_1 e x_2).

[8] Si tratta di un'evoluzione che cerca di superare i limiti dei modelli univariati sviluppati in prima istanza de Beaver e che si fondavano sull'uso individuale delle variabili economico-finanziarie senza combinarle tra di esse.

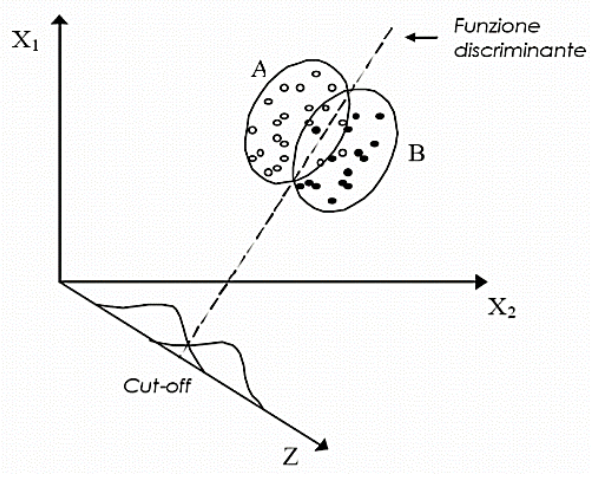


Figura 3.1: funzione discriminante

Sull'asse z è rappresentato uno score generato combinando linearmente tra loro le due variabili originarie e tale score prende il nome di *funzione discriminante*. Più in generale, date n variabili indipendenti per la generica impresa i-esima si ha:

$$z_i = \sum_{j=1}^{n} \gamma_j x_{i,j} \tag{3.1}$$

I coefficienti γ_j di tale combinazione lineare sono scelti tra tutte le infinite soluzioni possibili in modo tale da ottenere uno score z che discrimini in modo quanto più possibile netto le imprese anomale da quelle sane. Detto in altri termini, gli z_i ottenuti devono essere tali da massimizzare la distanza tra le medie (*centroidi*) Z_A e Z_B dei due gruppi di imprese sane e anomale. In pratica si vuole che le z_i delle imprese sane siano il più possibile simili tra loro e il più possibile diverse dalle imprese anomale. È possibile dimostrare tale condizione è soddisfatta se il vettore dei coefficienti $\boldsymbol{\gamma}$ è calcolato come segue:

$$\boldsymbol{\gamma} = \boldsymbol{\Sigma^{-1}} \cdot (\mathbf{x_A} - \mathbf{x_B}) \tag{3.2}$$

dove con $\boldsymbol{\Sigma^{-1}}$ si indica l'inversa della matrice di varianza/covarianza tra le n variabili indipendenti e con $\mathbf{x_A}$ e $\mathbf{x_B}$ si indicano rispettivamente i vettori contenenti i valori medi delle n variabili indipendenti per il gruppo delle imprese sane e per quello delle imprese anomale.

Sulla base di tale modello ogni impresa riceve sulla base dei propri valori delle variabili indipendenti un punteggio discriminante. In figura 3.2 si può osservare il risultato di un esempio tratto dal libro di testo "Rischio e valore nelle banche" di A. Resti e A. Sironi in

cui è stato analizzato un pool di 38 imprese di cui 24 sane (gruppo A) e 14 anomale studiando i valori assunti da due variabili indipendenti.

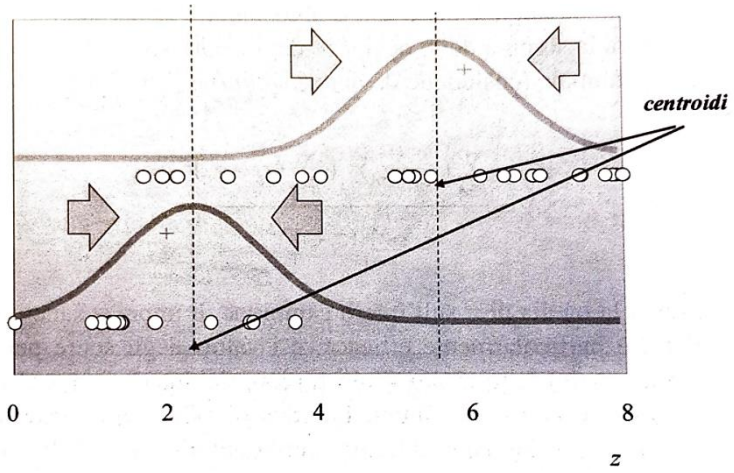


Figura 3.2: rappresentazione grafica dell'esempio tratto dal libro "Rischio e valore nelle banche" di A. Resti e A. Sironi

La linea più chiara in alto rappresenta la distribuzione dei valori assunti dalle z_i nel gruppo di imprese anomale, mentre la linea scura in basso rappresenta la distribuzione dei valori assunti dalle z nel gruppo di imprese sane. Quello che si può osservare è che i due gruppi risultano addensati attorno ai rispettivi centroidi z_A e z_B, pari rispettivamente a $-2{,}32$ e $-5{,}61$. A questo punto dell'analisi quello che è necessario fare è fissare una soglia, detta *cut-off point*, al di sotto della quale un'impresa viene scartata, ossia la banca le rifiuta il credito poiché giudicata troppo rischiosa. Ad esempio, considerando l'esempio in questione si potrebbe pensare di utilizzare come soglia di cut-off la metà della distanza tra i due centroidi, quindi:

$$\alpha = \frac{z_A + z_B}{2} = -3{,}97 \tag{3.3}$$

Tuttavia, dal confronto tra il valore di α e i valori assunti delle z_i delle imprese nell'esempio proposto sul libro, quello che emerge è che tale scelta di cut-off point condurrebbe a rifiutare il credito a ben sei imprese sane e a concedere il credito a un'impresa anomala. Tale risultato mette in risalto le problematiche che affliggono il modello discriminante nella capacità di discriminazione, la quale risulta piuttosto limitata rispetto ad altre soluzioni.

3.1.1 Lo Z-score di Altman

Il più noto score discriminante applicato nei modelli di rischio di credito è quello sviluppato da E. Altman nel 1968 per le imprese quotate statunitensi. Esso è funzione di cinque variabili indipendenti ed è formulato nel seguente modo:

$$z_i = 1{,}2 \cdot x_{i,1} + 1{,}4 \cdot x_{i,2} + 3{,}3 \cdot x_{i,3} + 0{,}6 \cdot x_{i,4} + 1{,}0 \cdot x_{i,5} \qquad (3.4)$$

dove:

- $x_1 =$ capitale circolante/totale attivo;
- $x_2 =$ utili non distribuiti/totale attivo;
- $x_3 =$ utile ante interessi e imposte/totale attivo;
- $x_4 =$ valore di mercato del patrimonio/valore contabile delle passività verso terzi;
- $x_5 =$ fatturato/totale attivo.

Maggiore è il valore di z di un'impresa, migliore è la sua qualità, ossia minore è la sua probabilità di insolvenza. Altman fissa una soglia di cut-off tra imprese sane e anomale in corrispondenza del valore 1,81 (valore ottenuto in modo del tutto uguale a quello impiegato nell'equazione 3.3).

3.2 Modelli di regressione

3.2.1 La regressione semplice e multipla

I primi modelli previsionali sviluppati sono stati di tipo statistico. Il caso più semplice è rappresentato dalla regressione lineare con un singolo regressore, con la quale è possibile studiare la relazione tra due variabili individuando l'equazione di una retta generica per l'intera popolazione:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (3.5)$$

dove:

- Y è la *variabile dipendente*;
- X è la *variabile indipendente*, anche detta *regressore*;
- β_0 è l'*intercetta* della retta di regressione della popolazione;
- β_1 è la *pendenza* della retta di regressione della popolazione;
- ε è l'*errore* o *disturbo*.

In particolare, la pendenza β_1 rappresenta la variazione di Y associata a una variazione unitaria di X, mentre l'intercetta rappresenta il valore della retta di regressione quando $X = 0$. Tali parametri devono essere stimati dal momento che inizialmente sono ignoti e il metodo di gran lunga più utilizzato è quello di scegliere la retta che corrisponde alla stima dei "minimi quadrati" per il dataset utilizzato, ovvero di utilizzare lo stimatore dei minimi quadrati ordinari (OLS). Tale stimatore ha la caratteristica di scegliere i parametri della regressione in modo che la retta di regressione risultante sia il più possibile vicina ai dati osservati, minimizzando quindi la somma dei quadrati degli errori $u_i = Y_i - (\beta_0 + \beta_1 X_i)$ che si commettono, dove con i si indica la i-esima osservazione nel dataset.

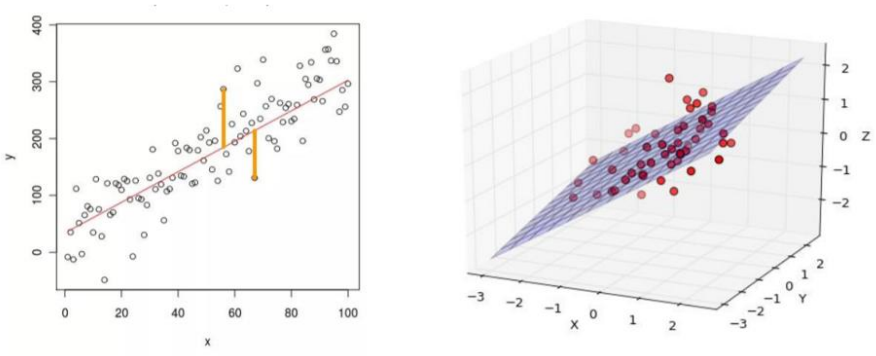


Figura 3.3: retta di regressione (a sinistra) e iperpiano di regressione (a destra)

La logica alla base della regressione lineare risulta pertanto piuttosto semplice e può essere complicata a piacere. Infatti, alle volte risulta necessario inserire più di un regressore per analizzare in modo più completo il comportamento della variabile dipendente Y e in questo caso si parla di *regressione multipla* e, se indichiamo con k il numero dei regressori esplicitati, la retta di regressione multipla può essere generalizzata nella seguente maniera:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (3.6)$$

Nel caso in cui ad esempio $k = 2$, il risultato di una regressione lineare multipla consiste in un iperpiano (un esempio grafico è riportato alla figura 3.3). L'equazione 3.6 presenta la medesima struttura dell'equazione y con la differenza che in questo caso i parametri β_i ($i = 1, \dots, k$) esprimono l'entità della variazione della variabile dipendente Y relativamente ad una variazione unitaria del regressore a cui sono associate.

3.2.2 La regressione logistica (modello logit)

Capita spesso nelle analisi di regressione che ci si trovi di fronte a situazioni in cui la variabile dipendente Y di cui si vuole studiare il comportamento sia dicotomica e quindi, dal punto di vista matematico, sia codificata come 0 o 1. Un esempio potrebbe essere la scelta da parte di una banca se concedere o no un mutuo a seconda di determinate caratteristiche appartenenti al richiedente. Più in generale sono riconducibili a questo ambito tutti i problemi di classificazione e di appartenenza a gruppi.

In questi casi la funzione di regressione è interpretata come una probabilità predetta e, sebbene sia ugualmente possibile applicare il modello della regressione semplice o multipla visti precedentemente, da un punto di vista matematico un modello non lineare risulta più appropriato. Infatti, nella formulazione $P(Y = 1|X) = \beta_0 + \beta_1 X$ il modello lineare implica che i valori della variabile dipendente Y siano compresi nell'intervallo $(-\infty\,, +\infty)$, tuttavia tale soluzione non risulta appropriata alla variabile dipendente dicotomica che, come detto, può assumere esclusivamente i valori 0 oppure 1 oltre che essere interpretata come una probabilità. Per affrontare questo problema è necessario adottare modelli non lineari, come la regressione logistica.

In figura 3.4 è possibile osservare un confronto grafico tra un modello di regressione lineare e un modello di regressione logistica, dove si vede che quest'ultima si adatta particolarmente bene alle casistiche in cui la variabile sia dipendente dicotomica dal momento che il suo andamento tende asintoticamente a 0 e a 1 per valori estremi della X.

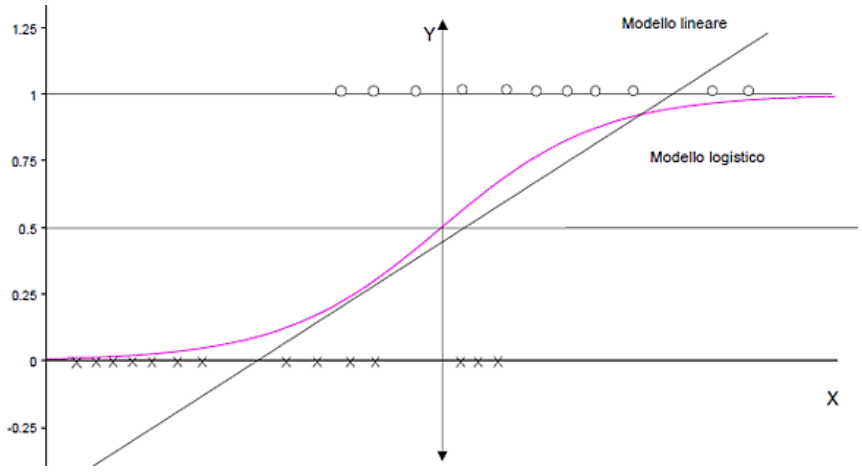


Figura 3.4: confronto delle rette di regressione di un modello lineare e di un modello logistico

La non linearità del modello logistico non consente di applicare il metodo OLS a meno che non si proceda con opportune trasformazioni che linearizzino la relazione, come ad esempio la *trasformazione logaritmica*. In particolare, per esprimere la relazione tra la

variabile indipendente e la dipendente in termini lineari si può partire dalla formulazione della regressione lineare vista precedentemente, dove il valore atteso condizionato della variabile dipendente Y data la variabile indipendente X può essere espresso come un'equazione lineare in X:

$$P(Y = 1|X) = \beta_0 + \beta_1 X \tag{3.7}$$

Come già precisato precedentemente, tale modello non è adeguato poiché i valori della probabilità sono compresi tra 0 e 1, mentre il termine $\beta_0 + \beta_1 X$ può assumere valori che vanno da $-\infty$ a $+\infty$. Per risolvere parzialmente il problema si applica in prima istanza la trasformazione esponenziale:

$$P(Y = 1|X) = e^{\beta_0 + \beta_1 X} \tag{3.8}$$

Tuttavia, anche questa formulazione non risolve il problema dal momento che, seppur è vero che ora l'intervallo è limitato inferiormente dal valore 0, non esiste comunque un limite superiore. A tal scopo si può applicare la *trasformazione logistica*, la quale consente di controllare i valori e di restringerli nell'intervallo $(0,1)$:

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{3.9}$$

A questo punto è possibile esprimere la probabilità mediante l'odds, ossia un modo di esprimere la probabilità mediante un rapporto tra le frequenze osservate in un livello ($Y = 1$) e quelle osservate nell'altro ($Y = 0$). Matematicamente:

$$odds_{Y=1} = \frac{P(Y = 1)}{1 - P(Y = 1)} \tag{3.10}$$

dove $P(Y = 0) = 1 - P(Y = 1)$. Sostituendo i relativi termini nella precedente espressione si ottiene:

$$odds_{Y=1} = \frac{\frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}{\frac{1}{1 + e^{\beta_0 + \beta_1 X}}} = e^{\beta_0 + \beta_1 X} \tag{3.11}$$

Infine, una volta calcolato l'*odds* è possibile calcolare il suo logaritmo naturale, ossia il *logit*:

$$\ln[odds_{Y=1}] = \beta_0 + \beta_1 X \tag{3.11}$$

Nel caso di una regressione logistica multipla con k variabili indipendenti, è possibile ripercorrere lo stesso ragionamento e ottenere la seguente funzione:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \qquad (3.12)$$

Ad esempio, nel caso in cui $k = 2$, il risultato di una regressione logistica consiste in una superficie tridimensionale (un esempio grafico è riportato alla figura 3.5).

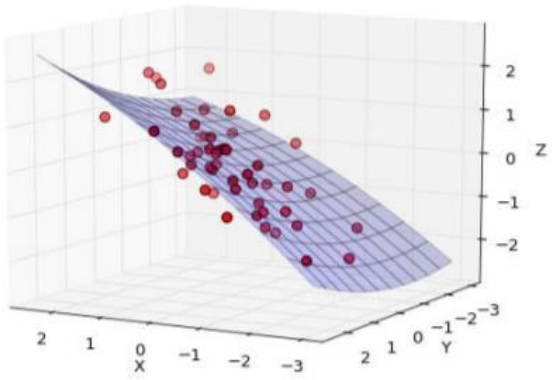


Figura 3.5: superficie tridimensionale relativa ad una regressione logistica con k = 2 regressori

È importante evidenziare che, come si evince dagli sviluppi, la probabilità, l'*odds* e il *logit* sono tre differenti modi di esprimere il medesimo concetto e la trasformazione in *logit* ha l'unico scopo di garantire la correttezza matematica dell'analisi.

3.2.2.1 Stima dei parametri del modello

Per quanto riguarda invece l'interpretazione della relazione tra variabili indipendenti e variabile indipendente, come nella regressione lineare anche nella regressione logistica tale procedura avviene mediante la valutazione dei parametri del modello. Come già specificato precedentemente, nella stima dei parametri della regressione non può essere impiegato il metodo OLS, quindi i parametri vengono stimati differentemente mediante un algoritmo ad hoc. Uno degli algoritmi più comuni per assolvere questo scopo è l'*algoritmo di massima verosimiglianza*, il quale attribuisce dei valori ai parametri incogniti del modello al fine di massimizzare la *funzione di verosimiglianza*: l'algoritmo viene avviato assegnando dei valori arbitrari di prova ai parametri del modello e che vengono successivamente aggiornati per verificare se la funzione può essere migliorata. Il procedimento viene iterato per diversi stage fintanto che la capacità di miglioramento della funzione di massima verosimiglianza è infinitesimale.

Riprendendo la nomenclatura utilizzata precedentemente, si definisce $\pi(X) = P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, il quale fornisce la probabilità condizionata che $Y = 1$ data X e $1 - \pi(X) = P(Y = 0)$, ossia la probabilità condizionata che $Y = 0$ data X. Per una osservazione (y_i, x_i) in cui $y_i = 1$ il contributo alla funzione di verosimiglianza è $\pi(x_i)$, mentre dove $y_i = 0$ il contributo risulta essere $1 - \pi(x_i)$. Pertanto, per l'osservazione (y_i, x_i) questo può essere espresso come:

$$\pi(x_i)^{y_i} \cdot [1 - \pi(x_i)]^{1-y_i} \tag{3.13}$$

L'assunzione che viene fatta per questo algoritmo è che le osservazioni siano indipendenti, quindi la funzione di verosimiglianza può essere ottenuta come prodotto dei termini dati dalla precedente espressione per ogni osservazione nel dataset:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \cdot [1 - \pi(x_i)]^{1-y_i} \tag{3.14}$$

dove $\boldsymbol{\beta}$ è il vettore dei parametri incogniti del modello logistico che devono essere stimati affinché $l(\boldsymbol{\beta})$ sia massimizzata. Da un punto di vista matematico risulta più semplice operare col logaritmo naturale di questa equazione:

$$\ln\big(l(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{3.15}$$

3.3 Modelli di machine learning

I modelli di scoring esaminati nei capitoli precedenti sono tutti caratterizzati dal tentativo di identificare le relazioni fondamentali che spiegano l'equilibrio economico-finanziario al fine di poter successivamente effettuare delle previsioni di insolvenza. Pertanto, si tratta di modelli fondati sulle caratteristiche strutturali che descrivono le condizioni di salute di un'impresa e dove la scelta delle variabili rilevanti, anche se calibrata con tecniche statistiche, riflette sempre una scelta a priori fondata sul ragionamento economico e condotta da un analista. Al contrario, le reti neurali adottano un procedimento induttivo: si parte da un campione di dati e se si riscontra una certa regolarità, allora tale regolarità viene utilizzata per cercare di prevedere il default di altre imprese. Pertanto, in questo caso al posto di affidarsi a regole determinate in via deduttiva, ci si affida ad un approccio meramente empirico.

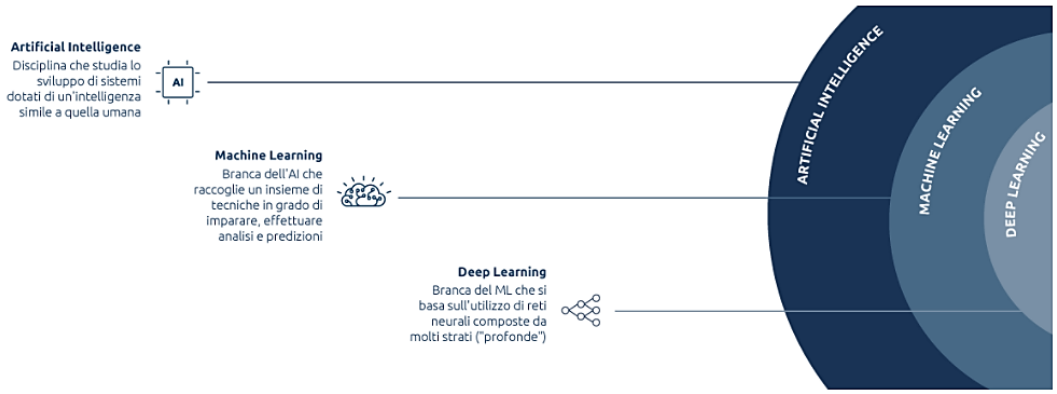


Figura 3.6: schematizzazione insiemistica dei concetti di IA, ML e deep learning

Le reti neurali appartengono alla famiglia delle tecniche di machine learning (ML), una branca dell'intelligenza artificiale (IA) che esplora la costruzione di modelli e algoritmi in grado di apprendere delle informazioni dai dati forniti in input ed effettuare delle previsioni. Le operazioni eseguite sono fondate su modelli statistici che permettono ad un computer di operare in modo estremamente efficiente ed in autonomia, ed è infatti per questa ragione che il ML si avvicina molto all'ambito della statistica computazionale ma alla quale aggiunge lo studio dell'ottimizzazione matematica che porta a metodi, teorie ed applicazioni che sono proprie del Machine Learning.

Tipicamente le tecniche di ML sono classificate in tre categorie:

- *Apprendimento supervisionato* – In questa prima tipologia di apprendimento i dati in input (*training data*) forniti al modello di ML sono associazioni di input e output desiderati già noti a priori e da cui si vuole dedurre la funzione che lega gli input agli output con l'obiettivo di mappare nuovi dati in input. Le tipologie di algoritmi utilizzati in questo caso sono le tecniche di *regressione*, quando l'obiettivo è identificare una relazione tra un insieme di variabili predittive e una variabile di output che può assumere valori nel continuo, e di *classificazione*, quando i dati sono di tipo categorico e si vuole determinare la classe più opportuna da assegnare a un determinato elemento;

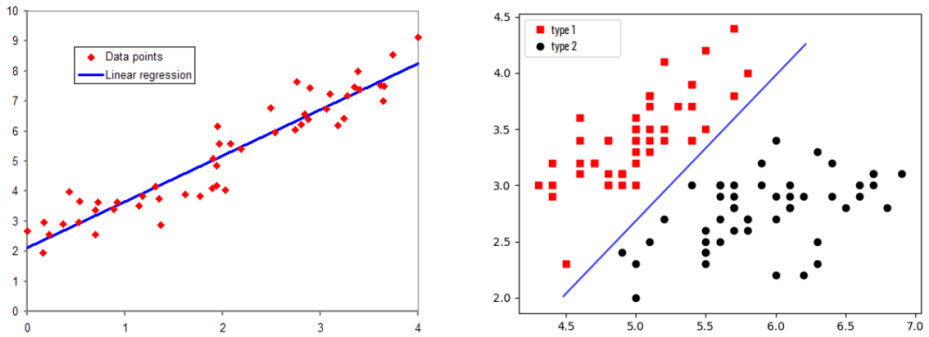


Figura 3.7: modello di regressione (a sinistra) e modello di classificazione (a destra)

- *Apprendimento non supervisionato* – In questa seconda tipologia di apprendimento i dati in input forniti al modello sono sprovvisti dei corrispondenti output, pertanto si conoscono solamente gli input della funzione che si vorrebbe individuare per mappare nuovi input. I modelli che si basano sull'apprendimento non supervisionato sono tipicamente modelli di clustering: acquisiscono i dati in input e li raggruppano in classi sulla base delle caratteristiche che rendono simili certe tipologie di dati. In particolare, le classi non sono note a priori ma vengono generate dal modello estrapolando informazioni dalla natura dei dati stessi;

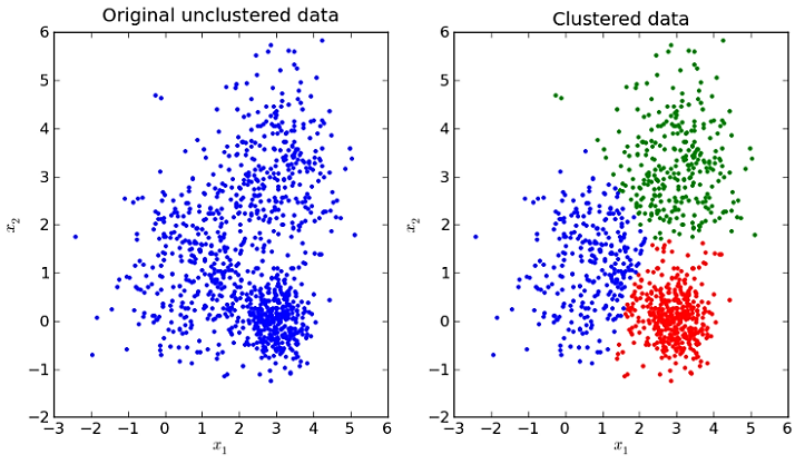


Figura 3.8: mappatura dei dati in cluster

- *Apprendimento con rinforzo* – I modelli che si basano su questo tipo di apprendimento sono caratterizzati dall'interazione con l'ambiente circostante. In particolare, dopo le prime fasi in cui il modello opera in modo casuale grazie a dei segnali, detti appunto "rinforzi", il modello è in grado di appurare la qualità delle azioni intraprese fino a quel momento e di correggersi al fine di migliorare la propria performance.

4. Le reti neurali

Le reti neurali sono degli strumenti di calcolo che prendono ispirazione dal funzionamento del cervello umano. La loro architettura è costituita da elementi di elaborazione operanti in parallelo i quali, presi singolarmente, sono in grado di effettuare semplici operazioni: integrano l'informazione proveniente da altri elementi elaborandola e comunicando il risultato di tale elaborazione ad altri elementi a cui sono collegati. Tali elementi di elaborazioni si ispirano per l'appunto ai neuroni biologici del cervello umano e dei quali rappresentano una semplificazione.

4.1 Il neurone biologico e il neurone artificiale

All'interno del cervello umano è possibile identificare un vastissimo numero di centri di elaborazione chiamati neuroni. Essi sono composti da tre elementi:

- *Il corpo cellulare (soma)*: contiene il nucleo del neurone ed è rivestito da una membrana contenente dei canali che permettono la comunicazione tra l'interno e l'esterno del soma;
- *I dendriti*: rappresentano i canali di input del neurone e ricevono i segnali provenienti dai neuroni a cui sono connessi;
- *L'assone*: rappresenta il canale di output e costituisce il percorso attraverso il quale il segnale emesso dal neurone si propaga verso altre parti del sistema nervoso, anche molto remote.

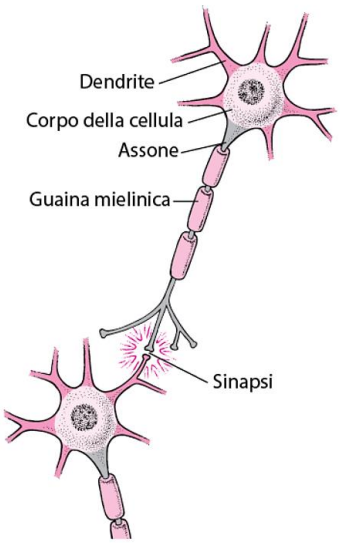


Figura 4.1: raffigurazione di neurone biologico

Il trasferimento dell'informazione da assone a dendrite avviene in zone di contatto dette *sinapsi* e per mezzo di processi elettrochimici: l'ingresso di ioni attraverso le sinapsi dei dendriti provoca una differenza di potenziale tra il soma e l'esterno e, non appena la differenza di potenziale oltrepassa una certa soglia, il neurone emette un impulso elettrico che si propaga nell'assone rilasciando a sua volta degli ioni. I segnali ricevuti in ingresso tramite i dendriti sono sommati facendo sì che l'informazione trasmessa possa essere di due tipologie:

- Di natura eccitatoria: la somma degli stimoli ricevuti oltrepassa la soglia di eccitazione, quindi il neurone viene stimolato e genera a sua volta un impulso che si propaga nell'assone;
- Di natura inibitoria: la somma degli stimoli in input on oltre passa la soglia di eccitazione, quindi non viene trasmesso nessun segnale elettrico in uscita.

In questo meccanismo risulta particolarmente importante una proprietà che caratterizza ogni legame sinaptico e che è conosciuta come *forza del legame sinaptico*, la quale può variare da sinapsi a sinapsi e determina l'efficacia di trasmissione del segnale elettrico. Pertanto, il segnale complessivo ricevuto in input da un neurone dipende sia dagli impulsi trasmessi dagli altri neuroni che sono stati eccitati che dalla forza del legame sinaptico attraverso cui vengono forniti in input i vari impulsi in ingresso.

Le reti neurali artificiali si ispirano al comportamento del cervello e sono composte da unità elementari di elaborazione (i neuroni) e da connessioni pesate ed orientate tra esse (le sinapsi). Ad ogni neurone è associato un valore numerico che rappresenta il valore che verrà trasferito dal neurone stesso e che dipende:

- Dai segnali di ingresso trasmessi dalle sinapsi;
- Da una funzione di attivazione (calcola il valore di input del neurone partendo dai valori pesati di output dei neuroni precedenti);
- Da una funzione di output che utilizza il valore della funzione di input per determinare il valore che verrà trasmesso al neurone successivo.

Anche alle sinapsi è associato un valore che determina l'efficacia della trasmissione: è positivo nel caso di sinapsi eccitatorie, mentre è negativo nel caso di sinapsi inibitorie.

Nel seguente capitolo verranno descritti i principali modelli teorici sviluppati per la modellizzazione artificiale dei neuroni.

4.2 Principali modelli teorici

4.2.1 Il modello di McCulloch e Pitts

Il primo modello di neurone artificiale risale al 1943 su proposta di McCulloch e Pitts ed è caratterizzato da input e output binari. La struttura è schematizzata alla figura 4.2.

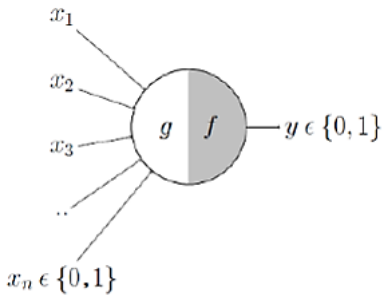


Figura 4.2: neurone artificiale secondo McCulloch e Pitts

Il suo funzionamento inizia con la ricezione degli input x_i, i quali sono pesati per i rispettivi valori w_i per tenere conto dell'efficacia della connessione sinaptica e, successivamente, sono sommati tra di loro. In particolare, se tale somma risulta superiore a uno specifico valore soglia *T*, allora il neurone si attiva e l'output *y* corrisponderà a 1, altrimenti risulta pari a 0. Il ragionamento può essere formalizzato matematicamente nel seguente modo:

$$y = \theta(h - T) \tag{4.1}$$

dove:

$$h = \sum_{i=1}^{n} w_i x_i \tag{4.2}$$

$$\theta(h - t) = \begin{cases} 1, & (h - t) > 0 \\ 0, & altrimenti \end{cases} \tag{4.3}$$

4.2.2 Il neurone moderno

Negli anni il campo dell'intelligenza artificiale ha attirato sempre più studiosi e ricercatori, i quali hanno permesso di apportare significative innovazioni ai modelli di rappresentazione dei neuroni. In figura 4.3 è schematizzato il funzionamento di un neurone nella sua versione più moderna.

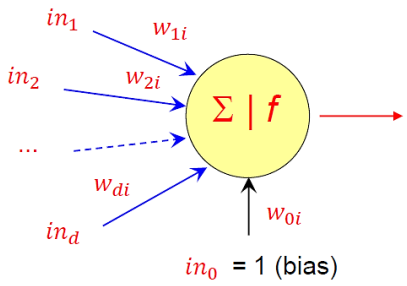


Figura 4.3: neurone moderno

dove:

- $in_1, \dots, in_d = d$ input che il neurone i riceve dai neuroni precedenti oppure dal problema in fase di inizializzazione;
- $w_{1i}, \dots, w_{di} =$ pesi che determinano l'efficacia delle connessioni sinaptiche dei dendriti relativi al neurone i;
- $w_{0i} =$ ulteriore peso, detto *bias*, che è collegato ad un input fittizio pari a 1 al fine di "tarare" il punto di lavoro ottimale del neurone;
- $net_i =$ livello di eccitazione globale del neurone i;
- $f(\cdot) =$ funzione di attivazione che determina l'output $out_i = f(net_i)$ del neurone i in funzione del suo livello di eccitazione net_i.

Dal punto di vista matematico il livello di eccitazione può essere espresso nel seguente modo:

$$net_i = \sum_{j=1}^{d} w_{ji} \cdot in_j + w_{0i} \tag{4.4}$$

Tipicamente le reti neurali più comunemente utilizzate operano con livelli continui in cui $f(\cdot)$ è una funzione non lineare, continua e differenziabile. La non linearità è una caratteristica fondamentale che deve possedere la funzione di attivazione se si vuole essere in grado di eseguire un mapping complesso dell'informazione di input, mentre la continuità e la differenziabilità sono caratteristiche imprescindibili per generare la retro-propagazione dell'errore (si veda in seguito l'algoritmo di *error back-propagation* al capitolo 4.6).

4.3 Le funzioni di attivazione

Si è visto, parlando dei principali modelli teorici che modellizzano matematicamente il comportamento dei neuroni biologici, che la funzione di attivazione determina il valore del

segnale in uscita da un neurone in funzione della somma pesata dei segnali in ingresso del neurone stesso. Le funzioni di attivazione che possono essere prese in considerazione sono molteplici, tuttavia tipicamente le più utilizzati sono:

- La funzione a soglia;
- La funzione di Sigmund;
- La funzione rettificatrice;
- La funzione tangente iperbolica.

Nei seguenti sotto-capitoli sono passate in rassegna le funzioni di attivazione appena elencate in funzione della somma pesata degli input in ingresso nel neurone (ΣwX).

4.3.1 Funzione a soglia

Tale tipologia di funzione costituisce il modello più semplice: la funzione restituisce il valore 1 nel caso in cui la somma pesata dei segnali in input è maggiore o uguale a un certo valore soglia t e zero negli altri casi. Il grafico di questa funzione è quello riportato alla figura seguente.

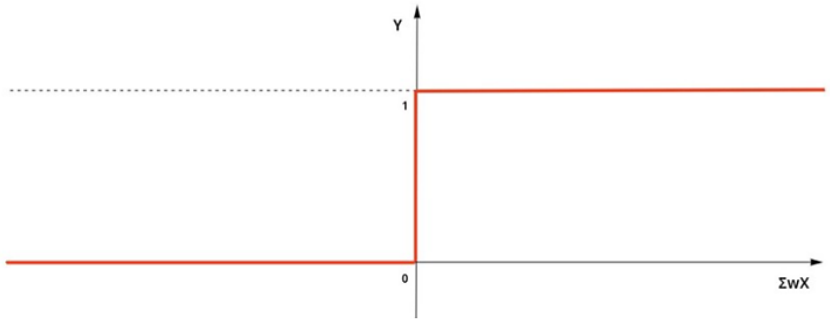


Figura 4.4: funzione a soglia

In formule il valore di output del neurone è rappresentabile nel seguente modo:

$$O = \theta(\Sigma wX) \tag{4.5}$$

dove con θ si indica la *funzione di Heaviside* $\theta: \mathbb{R} \rightarrow \{0,1\}$:

$$\theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{4.6}$$

L'utilizzo di questa funzione risulta molto utile qualora si abbia bisogno di un segnale in output binario del tipo Y/N.

4.3.2 Funzione di Sigmund

Tale funzione è del tutto uguale a quella impiegata per effettuare la regressione logistica. Pertanto, anche in questo caso il codominio della funzione, ossia i valori che essa può assumere in output, è continuo e compreso dall'intervallo $(0,1)$. Il grafico della funzione di Sigmund è riportato alla seguente figura.

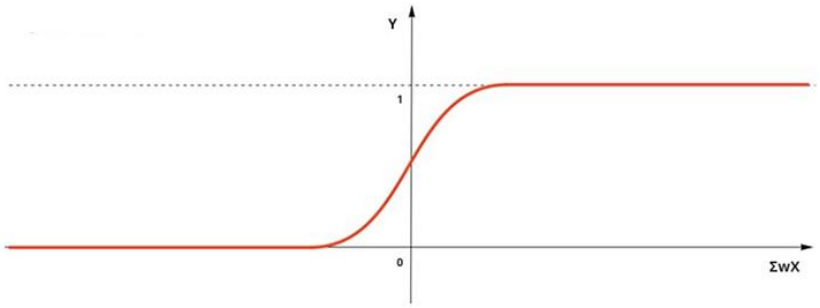


Figura 4.5: funzione di Sigmund

In formule:

$$Y = \frac{1}{1 + e^{-\Sigma wX}} \tag{4.7}$$

Tale funzione può essere utilizzata al posto della funzione a soglia considerando il valore in uscita non come Y ma come una probabilità che Y sia uguale ad 1.

4.3.3 Funzione rettificatrice

La funzione rettificatrice rappresenta la funzione di attivazione più utilizzata. Tale funzione ha la caratteristica di restituire 0 qualora la somma pesata degli input di un neurone sia minore o uguale a 0 e ΣwX in tutti gli altri casi. Ne deriva che il codominio di questa funzione è continuo e compreso nell'intervallo $[0, +\infty)$. Il grafico è riportato alla seguente figura.

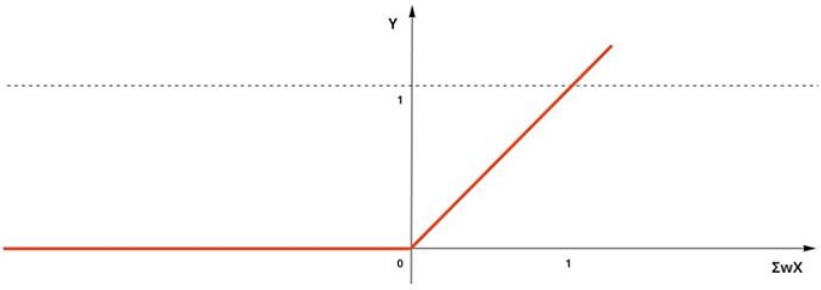


Figura 4.6: funzione rettificatrice

In formule:

$$Y = \max\{0, \Sigma wX\} \tag{4.8}$$

4.3.4 Funzione tangente iperbolica

La funzione tangente iperbolica è molto simile alla funzione di Sigmund e differisce da quest'ultima per il semplice fatto che il suo codominio è compreso nell'intervallo $(-1, +1)$. Il grafico è riportato alla seguente figura.

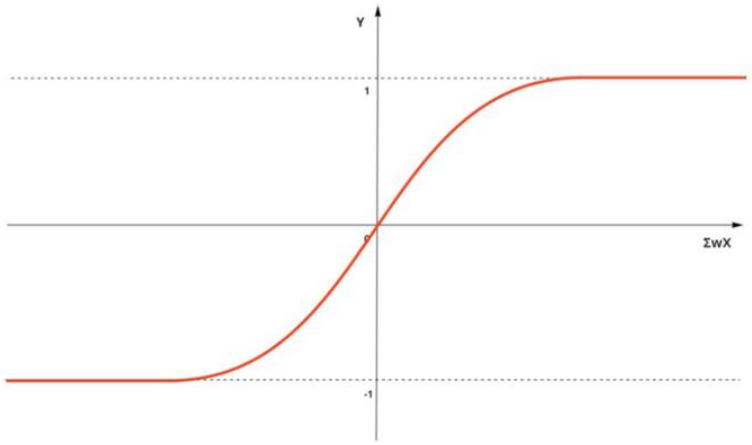


Figura 4.7: funzione tangente iperbolica

In formule:

$$Y = \frac{1 - e^{-2\Sigma wX}}{1 + e^{-2\Sigma wX}} \tag{4.9}$$

L'impiego di tale funzione si rivela particolarmente utile nel caso in cui si necessiti di mappare l'output di un neurone come una variazione percentuale.

4.4 Architettura delle reti

I neuroni artificiali con le relative funzioni di attivazione possono essere organizzati in differenti architetture per formare una rete neurale artificiale. In generale, i neuroni che costituiscono una rete neurale possono essere classificati in tre categorie a seconda della loro funzione:

- *Neuroni di input*: sono i neuroni le cui attivazioni rappresentano i valori di input della rete;

- *Neuroni di output*: sono i neuroni le cui attivazioni rappresentano i valori di output della rete;
- *Neuroni nascosti*: si tratta dei neuroni che non sono collocati ai confini della rete neurale e che non sono visibili dall'esterno.

4.4.1 Reti feed-forward

Nell'architettura feed-forward i neuroni sono raggruppati in diversi strati concepiti come sottoinsiemi disgiunti e ordinati a seconda della loro funzione: sono presenti uno strato di input, uno o più strati nascosti e uno strato di output e ad ogni strato sono connessi i neuroni degli strati adiacenti. Il modello più diffuso ed utilizzato prevede che ogni neurone sia connesso a tutti i neuroni degli strati adiacenti e che non ci siano connessioni tra neuroni dello stesso strato. I neuroni dello strato di input non hanno connessioni in ingresso e la loro attivazione consiste nel vettore *pattern* (input del problema) tramite il quale una funzione trasferisce il valore di attivazione senza eseguire calcoli ai neuroni dello strato nascosto, i quali calcolano la loro attivazione e la trasferiscono o ai neuroni di un altro strato nascosto oppure a quelli di output (l'attivazione di questi rappresenta l'output della rete).

In tale architettura il flusso informativo è unidirezionale: i neuroni ricevono input solo dallo strato precedente e lo trasmettono solo a quello successivo ed è per questo motivo che tale architettura è detta "feed-forward". Infatti, non sono consentite connessioni all'indietro o connessioni verso lo stesso livello.

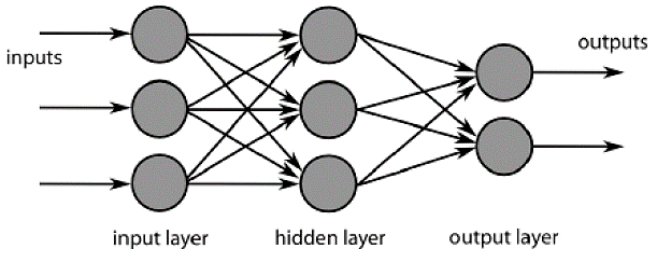


Figura 4.8: rete feed-forward

4.4.2 Reti ricorrenti

Nelle reti ricorrenti sono previste connessioni di feedback che in genere sono indirizzate verso i neuroni dello stesso livello, ma possono anche essere indirizzate ai neuroni del livello precedente. Questa caratteristica rende l'addestramento di tali reti un po' più

complesso dal momento che viene richiesto di considerare il comportamento in più instanti temporali.

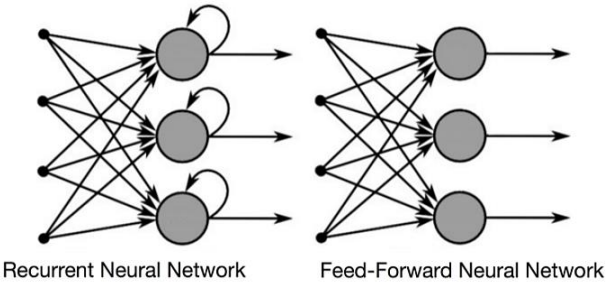


Figura 4.9: confronto grafico tra una rete ricorrente (a sinistra) e una rete feed-forward (a destra)

Tali reti sono indicate per la gestione di sequenze (ad esempio audio, video, frasi in linguaggio naturale, ecc.) perché dotate di un effetto memoria che al tempo t rende disponibile l'informazione processata al tempo $t-1, t-2$, ecc.

4.4.3 Reti convoluzionali

Le reti convoluzionali sono un'architettura di rete neurale in cui la struttura di connettività tra i neuroni è ispirata all'organizzazione della corteccia visiva animale, dove i singoli neuroni sono disposti in maniera tale da rispondere alle regioni di sovrapposizione che tassellano il campo visivo. Il principale campo applicativo di tale architettura riguarda l'impiego in problemi di identificazione e di classificazione di immagini e video.

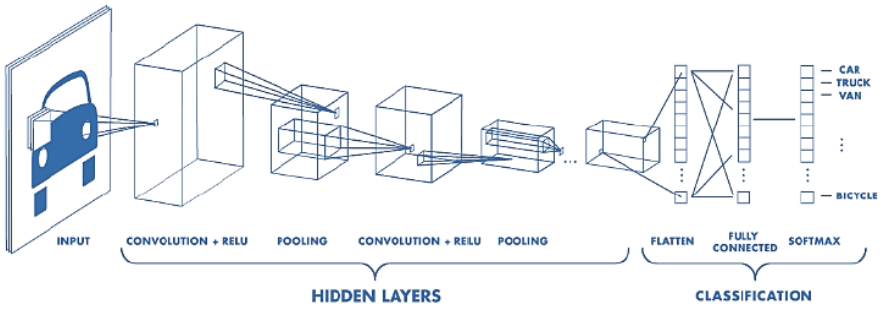


Figura 4.10: rete convoluzionale

La struttura è costituita da più livelli:

- Il primo livello ha la funzione di recepire gli input esterni, solitamente rappresentati da un insieme di pixel convertiti in cifre;
- Successivamente, i dati passano ai livelli convoluzionali i quali hanno lo scopo di individuare nelle immagini la presenza di elementi come angoli e curve e passare

i loro output ai *ReLu*[9], i quali tramite la loro funzione di attivazione caratteristica annullano i valori negativi degli strati precedenti;

- L'immagine a questo punto è semplificata dal livello *pool* e questo processo si ripete finché la rete non elabora la propria previsione nel livello *fully connected* e producendo come output un vettore di dimensione pari al numero delle classi tra le quali l'algoritmo deve scegliere, dove ad ogni classe è associata la probabilità che l'immagine in input appartenga ad ogni specifica classe.

4.5 Il Perceptron e il Multilayer Perceptron

Il *perceptron* (in italiano percettrone) è un modello di neurone sviluppato da Rosenblatt nel 1956 che utilizza una funzione di attivazione lineare a soglia, dove la soglia è rappresentata dal valore θ. Alla seguente figura è riportato una schematizzazione del suo funzionamento.

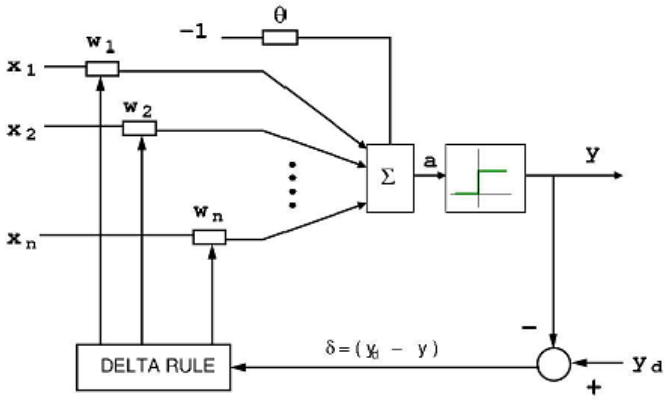


Figura 4.11: il percettrone

Un singolo percettrone o una rete di percettroni a due soli livelli (input e output) può essere addestrato con una regola, detta *delta rule*, che prevede che se l'output desiderato d e quello effettivo y prodotto dalla rete neurale non coincidono, i pesi delle connessioni sinaptiche vengono modificate come segue:

$$w_i(t+1) = w_i(t) + \eta\delta x_i \qquad (4.10)$$

dove:

- $w_i(t+1)$ = peso della sinapsi i dopo la modifica;

[9] ReLu sta per Rectified Linear Units, ossia la funzione di attivazione rettificatrice.

- w_i = peso della sinapsi *i* prima della modifica;
- η = coefficiente di apprendimento (*learning rate*): valori elevati portano a grandi modifiche delle sinapsi ad ogni passo comportando una possibile instabilità dell'apprendimento, mentre valori bassi portano a modifiche piccole;
- x_i = valore trasmesso dalla sinapsi *i*;
- $\delta = (d - y)$.

Tale modello rappresenta l'esempio più semplice di rete neurale, tuttavia il suo limite è che non risulta adatto a risolvere problemi caratterizzati da ingressi non linearmente separabili. In particolare, quando si parla di *problemi linearmente separabili* ci si riferisce a problemi per i quali, una volta che gli ingressi sono disposti in uno spazio, non è possibile identificare un iperpiano che divida nettamente gli elementi appartenenti alle diverse classi del problema. Tale problema può essere superato utilizzando il *Multilayer Perceptron* (MLP), ossia un'architettura feed-forward con almeno tre livelli di cui almeno uno nascosto e con funzioni di attivazione non lineari. Alla figura 4.12 è schematizzata la struttura generica di un MLP costituita da tre livelli $(d: n_H: s)$:

- Livello di input: *d* neuroni;
- Livello nascosto: n_H neuroni;
- Livello di output: *s* neuroni.

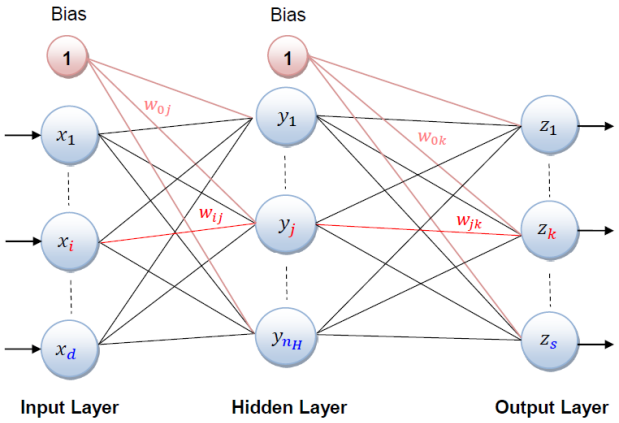


Figura 4.12: Multilayer Perceptron

Sulla base della nomenclatura riportata in figura 4.12, il *k*-esimo valore di output può essere calcolato nel seguente modo:

$$z_k = f\left(\sum_{j=1\dots n_H} w_{jk} \cdot y_j + w_{0k}\right)$$
$$= f\left(\sum_{j=1\dots n_H} w_{jk} \cdot f\left(\sum_{j=1\dots d} w_{ij} \cdot x_i + w_{0j}\right) + w_{0k}\right) \quad (4.11)$$

A questo punto, fissati il numero di livelli e di neuroni, l'addestramento di una rete neurale consiste nel calibrare il valore dei pesi *w* che determinano il mapping desiderato tra input e output. Tuttavia, sebbene i primi neuroni artificiali risalgono agli anni '40, fino alla metà degli anni '80 non erano disponibili algoritmi di training efficaci e si è dovuto aspettare fino al 1986 quando Rumelhart, Hinton e Williams svilupparono l'*algoritmo di error back-propagation*. Il funzionamento di tale algoritmo avviene in due fasi:

- Nella prima fase l'attivazione delle unità di input viene propagata in avanti tramite le funzioni di attivazione (*forward phase*);
- Nella seconda fase vengono modificati i pesi delle connessioni sinaptiche tramite la tecnica del *gradient descent* con la quale l'errore delle unità di output viene propagato all'indietro (*backward phase*).

4.6 L'algoritmo di error back-propagation

Si è detto precedentemente che lo scopo dell'addestramento è quello di calibrare i pesi delle connessioni sinaptiche cosicché sia possibile implementare il mapping desiderato tra input e output. Tipicamente, quando si sviluppa un modello di machine learning che necessita di un addestramento per calibrare i parametri caratteristici, il dataset a disposizione viene diviso in tre porzioni:

- Il *training set* costituisce la porzione di dataset che viene impiegata esclusivamente per l'addestramento del modello, ossia l'insieme di dati tramite i quali il modello imparerà le relazioni tra il pattern $\boldsymbol{x}$ in input e l'output desiderato $\boldsymbol{y}$;
- Il *validation set* è una porzione di dataset che viene impiegato per scongiurare il fenomeno dell'overfitting, ossia la creazione di un modello che è perfettamente in grado di predire i dati utilizzati in fase di addestramento ma che non è adeguato a generalizzare dati nuovi mai visti prima. In particolare, tramite il validation set vengono forniti al modello sviluppato nella fase di training dei dati che non ha mai visto prima per elaborare una predizione. Nel caso in cui le performance fossero

scarse, si dovrà rieseguire il train finché il risultato del validation set non risulterà soddisfacente. Il validation set, pertanto, gioca un ruolo cruciale nell'addestramento della rete neurale perché il punto di forza dei modelli di machine learning è proprio la generalizzazione, ossia la capacità di un modello di effettuare predizioni corrette su dati che non ha mai analizzato in precedenza;

- Il *test set* viene utilizzato per fornire al modello elaborato col vaglio del validation set ulteriori nuovi dati mai processati dal modello. Tale set a volte viene omesso e viene tipicamente utilizzato per visualizzare le prestazioni e il funzionamento del modello.

Tuttavia, prima di procedere all'aggiornamento dei pesi per calibrare opportunamente il la rete neurale è necessario costruire una metrica che permetta di valutare la validità del modello, ovvero una *loss function*.

Widrow e Hoff sono due studiosi che nel secolo scorso ebbero il merito di introdurre per la prima volta il concetto di errore determinando che la variazione dei pesi è proporzionale al gradiente dell'errore. In particolare, è definito *errore quadratico assoluto* relativo ad un pattern in ingresso la sommatoria dei quadrati delle differenze tra l'output generato e quello desiderato per ogni nodo di output della rete neurale. Se si considera con $\boldsymbol{z} = [z_1, \dots, z_s]$ l'output prodotto dalla rete in corrispondenza del pattern $x = [x_1, \dots, x_d]$ fornito in ingresso e con $\boldsymbol{t} = [t_1, \dots, t_s]$ l'output desiderato, la loss function appena descritta può essere espressa in formule nel seguente modo:

$$J(\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{2} \cdot \sum_{c=1}^{s} (t_c - z_c)^2 \tag{4.12}$$

dove *s* corrisponde al numero di neuroni di output della rete. La funzione $J(\boldsymbol{x}, \boldsymbol{w})$ può essere minimizzata modificando i pesi $\boldsymbol{w}$ in direzione opposta al gradiente di J: il gradiente in analisi matematica indica la direzione di maggior crescita di una funzione di più variabili, quindi muovendosi in direzione opposta si identifica la direzione di maggior decrescita della loss function. In particolare, quando la minimizzazione della funzione di errore avviene attraverso passi in direzione opposta al gradiente, l'algoritmo di Back-Propagation è denominato anche *gradient descent*.

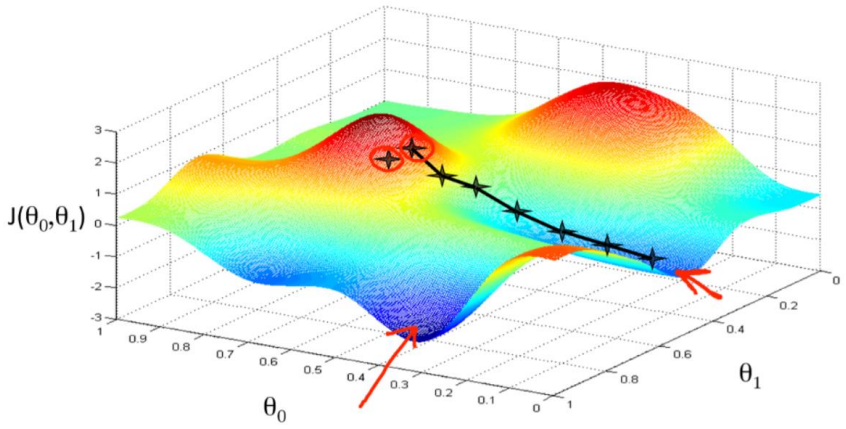


Figura 4.13: rappresentazione grafica dell'algoritmo di gradient descent su una generica funzione

Tipicamente per gestire un problema di classificazione con s classi e pattern con dimensione pari a d si è soliti utilizzare una rete neurale con $d: n_H: s$ neuroni nei tre livelli. Di seguito viene descritta dal punto di vista matematico:

- La modifica dei pesi w_{jk} corrispondente all'interfaccia *hidden-output*;
- La modifica dei pesi w_{ij} corrispondente all'interfaccia input-*hidden*.

Modifica dei pesi hidden-output:

$$\frac{\partial J}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}}\left(\frac{1}{2}\sum_{c=1}^{s}(t_c - z_c)^2\right) = (t_k - z_k)\cdot\frac{\partial(-z_k)}{\partial w_{jk}}$$

Dal momento che solo z_k è influenzato da w_{jk}, si può procedere con lo sviluppo come segue:

$$= (t_k - z_k)\cdot\frac{\partial\left(-f(net_k)\right)}{\partial w_{jk}}$$

$$= -(t_k - z_k)\cdot\frac{f(net_k)}{\partial net_k}\cdot\frac{\partial net_k}{\partial w_{jk}}$$

$$= -(t_k - z_k)\cdot f'(net_k)\cdot\frac{\sum_{s=1}^{n_H} w_{sk}\cdot y_s}{\partial w_{jk}}$$

$$= -(t_k - z_k)\cdot f'(net_k)\cdot y_j$$

Se si definisce $\delta_k = (t_k - z_k)\cdot f'(net_k)$, allora:

$$\frac{\partial J}{\partial w_{jk}} = -\delta_k \cdot y_j$$

Pertanto, il peso w_{jk} può essere aggiornato come:

$$w_{jk} = w_{jk} + \eta \cdot \delta_k \cdot y_j$$

dove con η si indica il coefficiente di apprendimento (*learning rate*).

Modifica dei pesi input-hidden:

$$\frac{\partial J}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}}\left(\frac{1}{2}\sum_{c=1}^{s}(t_c - z_c)^2\right)$$

$$= -\sum_{c=1}^{s}(t_c - z_c) \cdot \frac{\partial z_c}{\partial w_{ij}}$$

$$= -\sum_{c=1}^{s}(t_c - z_c) \cdot \frac{\partial z_c}{\partial net_c} \cdot \frac{\partial net_c}{\partial w_{ij}}$$

$$= -\sum_{c=1}^{s}(t_c - z_c) \cdot f'(net_c) \cdot \frac{\partial net_c}{\partial w_{ij}}$$

Sviluppando l'ultimo termine della precedente sommatoria si ottiene:

$$\frac{\partial net_c}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}}\sum_{r=1}^{n_H} w_{rc} \cdot y_r$$

$$= \frac{\partial}{\partial w_{ij}}\sum_{r=1}^{n_H} w_{rc} \cdot f(net_r)$$

Dal momento che solo net_j è influenzato da w_{ij} si può proseguire lo sviluppo come segue:

$$= \frac{\partial}{\partial w_{ij}}\left(w_{jc} \cdot f\left(net_j\right)\right)$$

$$= w_{jc} \cdot \frac{\partial f\left(net_j\right)}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}}$$

$$= w_{jc} \cdot f'\left(net_j\right) \cdot \frac{\partial}{\partial w_{ij}}\sum_{q=1}^{d} w_{qj} \cdot x_q$$

$$= w_{jc} \cdot f'(net_j) \cdot x_i$$

Pertanto, tenendo in considerazione che $\delta_k = (t_k - z_k) \cdot f'(net_k)$ si può riscrivere la loss function come:

$$\frac{\partial J}{\partial w_{ij}} = -\sum_{c=1}^{s} \delta_c \cdot w_{jc} \cdot f'(net_j) \cdot x_i$$

$$= -x_i \cdot f'(net_j) \sum_{c=1}^{s} \delta_c \cdot w_{jc}$$

Se a questo punto si definisce $\delta_j = f'(net_j) \cdot \sum_{c=1}^{s} w_{jc} \cdot \delta_c$, allora si ottiene che:

$$\frac{\partial J}{\partial w_{ij}} = -\delta_j \cdot x_i$$

Pertanto, il peso w_{ij} può essere aggiornato come:

$$w_{ij} = w_{ij} + \eta \cdot \delta_j \cdot x_i$$

dove con η si indica il coefficiente di apprendimento (*learning rate*).

5. Analisi dei dati del settore metallurgico italiano

La scelta del settore merceologico sui cui sviluppare il modello di regressione logistica e la rete neurale è ricaduta sul settore metallurgico, individuato dal codice Ateco 24. Nello specifico, la metallurgia è quella scienza che studia il complesso delle tecniche e dei processi che permettono di produrre i diversi metalli a partire dai minerali che li contengono. Secondo la classificazione Ateco, le attività svolte dalle aziende appartenenti a tale settore sono:

- Fabbricazione di ferro, acciaio e ferroleghe;
- Fabbricazione di tubi, condotti, profilati cavi e relativi accessori in acciaio;
- Stiratura a freddo di barre, laminazione a freddo di nastri, profilatura mediante formatura o piegatura a freddo, trafilatura a freddo;
- Produzione di metalli di base preziosi e altri metalli non ferrosi, trattamento dei combustibili nucleari;
- Fonderie.

In questo capitolo si è effettuerà in primo luogo un'analisi del settore di riferimento relativamente al periodo 2011 – 2020, per poi successivamente descrivere le operazioni che hanno portato alla determinazione del campione di analisi e le operazioni di pulizia e correzione che sono state messe in atto al fine di rendere i dati raccolti fruibili come input per i modelli.

5.1 Analisi macrosettoriale

L'industria metallurgica riveste un'importanza strategica fondamentale a livello nazionale riconducibile all'uso estensivo che si fa dei metalli e in particolare dell'acciaio. L'Italia è infatti comunemente conosciuta come un paese fortemente manifatturiero e i settori che fanno ampiamente uso dei metalli sono molti e anche molto diversi e trasversali tra loro, come le costruzioni, la meccanica, l'automotive, l'alimentare, il medicale, ecc.

Il settore metallurgico italiano ha recentemente archiviato dei risultati connotati da evidenti difficoltà. Infatti, già nel 2019 il comparto ha dovuto affrontare un rallentamento dell'economia globale causata dal protrarsi delle tensioni commerciali internazionali tra Cina e Stati Uniti, dalla presenza di sovracapacità produttiva, da un incremento dei prezzi

delle materie prime e dalla "questione Ilva"; tuttavia, oltre a questi eventi che hanno influito negativamente sulla redditività del settore, l'anno 2020 nei suoi primi mesi ha portato con sé l'emergenza sanitaria causata dal Covid-19, un evento straordinario e non prevedibile che ha causato uno stop forzato per quasi tutti i settori produttivi e cali della produzione con picchi fino al 50%. In generale l'economia nazionale, già in fase di stagnazione, è stata tra le più colpite dalla crisi economico-sanitaria e la frenata di essa ha interessato tutte le componenti del PIL, anche se sono stati registrati dei cali più marcati dalle esportazioni (-13,8%) e dalle importazioni (-12,6%).

Alla seguente figura 5.1 è possibile osservare i dati elaborati da Federacciai per l'Italia e l'Europa relativi alla produzione dell'acciao, la cui produzione e lavorazione risulta essere uno dei principali componenti del settore metallurgico:

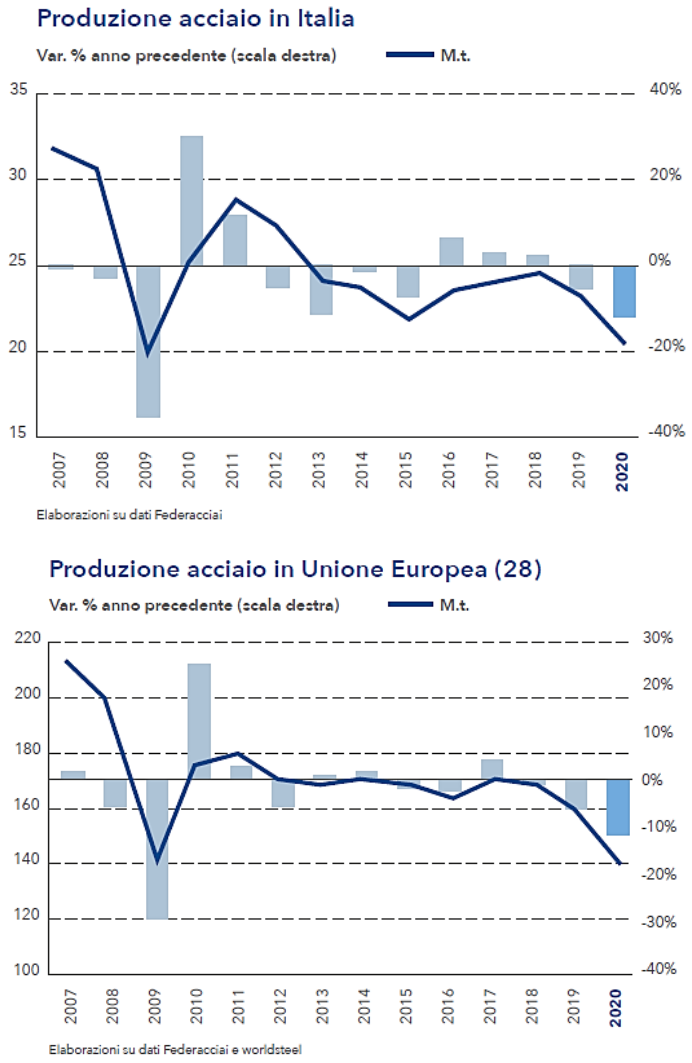


Figura 5.1: variazioni percentuali e assolute di produzione dell'acciaio in Italia (figura in alto) e in Europa (figura in basso)

Dai precedenti grafici si può osservare come il calo produttivo, espresso in M.t., registrato per l'anno 2020 ha interessato in generale tutti i Paesi dell'Unione, seppur con intensità differenti (con riferimento all'anno 2019: Germania -10,0%, Italia: -12,1%, Francia -19,8%, Spagna -19,1%). L'Italia nell'anno 2020 si è confermata tra i principali produttori europei posizionandosi al secondo posto con una quota di mercato pari al 14,6% e alle spalle della Germania con una quota di mercato pari al 25,6%. In generale, se si guarda l'andamento produttivo nazionale del settore si osserva che non sono mai stati eguagliati i livelli di produzione precedenti alla crisi finanziaria iniziata nella seconda metà del 2007, un evento di grandissima portata e i cui effetti hanno avuto pesanti ripercussioni per l'anno 2009 ma anche per quelli successivi sino ad oggi. Infatti, a seguito di un profondo crollo iniziato nel 2008 e che ha raggiunto il suo apice nell'anno 2009, al netto di un fisiologico rimbalzo relativo all'anno 2010, i tassi di variazione dei livelli produttivi sono stati o negativi o molto prossimi allo 0. È quindi interessante notare come gli effetti di eventi scaturiti sui mercati finanziari possano avere non solo pesanti ripercussioni sulle economie nazionali nel breve periodo, ma anche su un orizzonte temporale medio-lungo.

Nei seguenti capitoli sono presentati diverse analisi al fine di approfondire le dinamiche in termini di redditività, situazione patrimoniale, finanziaria e produttiva del campione di imprese estratto ed analizzato dal database Aida. In particolare, i dati di partenza che sono poi stati elaborati per effettuare le analisi citate sono stati ottenuti aggregando per anni i bilanci delle imprese appartenenti del campione, ottenendo così un unico bilancio per ogni anno dell'orizzonte temporale in analisi, dal 2011 al 2020.

5.1.1 Analisi della redditività

L'analisi della redditività ha l'obiettivo di verificare l'attitudine a produrre un reddito sufficiente a coprire i costi e generare un utile. Per valutare la redditività del settore metallurgico si è scelto di approfondire i valori dei seguenti indici di bilancio:

- Return on Equity (ROE);
- Return on Sales (ROS);
- ROI.

Il R*eturn on Equity* (ROE) esprime la redditività del capitale proprio ed è pari al rapporto tra l'utile di esercizio e l'equity:

$$\text{ROE} = \frac{\text{Utile di esercizio}}{\text{Patrimonio netto}} \tag{5.1}$$

Nello specifico, considerando che il Patrimonio netto contiene al suo interno sia la ricchezza apportata dai soci che quella prodotta nel corso del tempo per effetto degli utili accantonati e non distribuiti, l'indice in questione misura il rendimento dell'investimento effettuato dai soci delle società appartenenti al campione. Nel seguente grafico è possibile osservare l'andamento del ROE riferito al campione in analisi: la redditività del capitale versato dai soci nei primi anni del campione ha presentato un trend decrescente col quale ha toccato un minimo nel 2013 riportando un valore pari al -7,40 per cento, dal quale si è poi seguito un trend in crescita fino al 2018 in cui si è toccato un massimo di +6,67 per cento e a seguito del quale c'è stato un assestamento intorno a valori di poco inferiori al +2 per cento fino alla fine dell'orizzonte temporale.

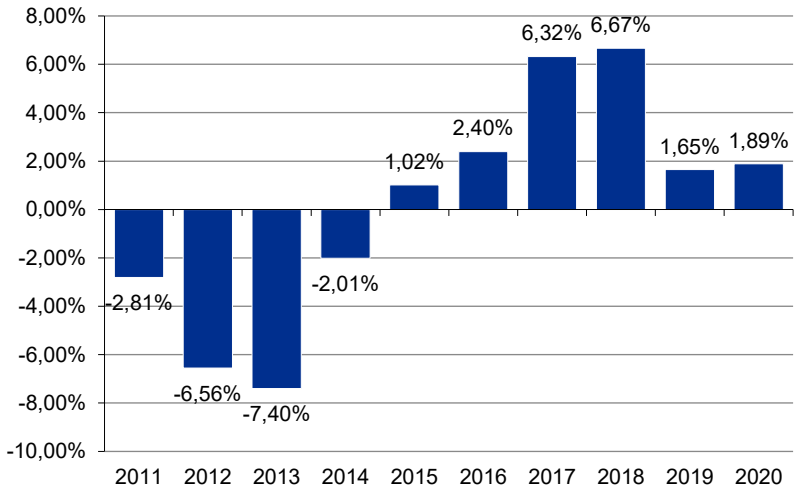


Figura 5.2: valori del ROE

La remunerazione del capitale investito dagli azionisti nel settore nel periodo di riferimento risulta piuttosto volatile con variazioni da un anno all'altro anche piuttosto importanti.

Il secondo indicatore scelto per valutare la redditività del settore in analisi è il *Return on Sales* (ROS), un indice che esprime la redditività delle vendite e, nello specifico, quanta parte dei ricavi è assorbita dalla gestione operativa. In particolare, tale indicatore risulta molto interessante poiché è diretta conseguenza delle condizioni di efficienza interna e delle situazioni esterne di mercato: le prime incidono sulla capacità di contenere i costi, di mantenere un equilibrio economico e di realizzare determinati volumi produttivi; le seconde influiscono invece sulle dinamiche dei prezzi di vendita, dei costi di acquisto e sulle variabili commerciali. L'indice in questione è calcolato nel seguente modo:

$$\text{ROS} = \frac{\text{EBIT}}{\text{Fatturato}} \tag{5.2}$$

Al seguente grafico è possibile osservare l'andamento del ROS nel periodo di riferimento del campione in analisi:

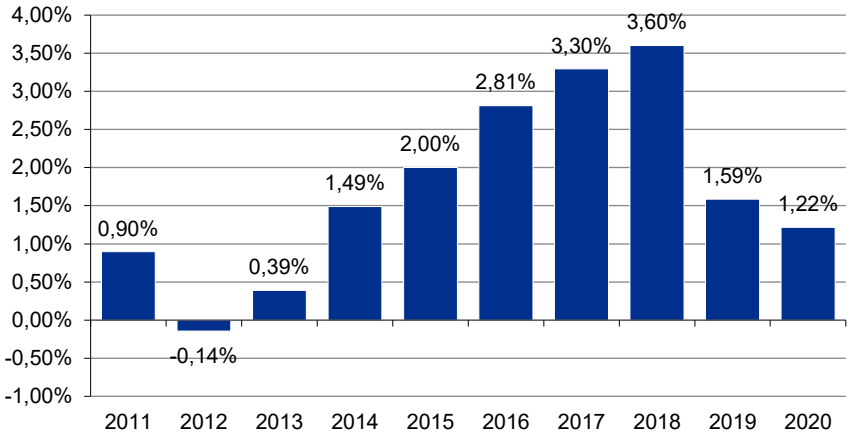


Figura 5.3: valori del ROS

Si può osservare che, similmente all'andamento del ROE, è presente nei primi periodi dell'orizzonte temporale di riferimento una flessione al ribasso per poi recuperare fino a raggiungere un massimo nel 2018 con, infine, i valori per gli anni 2019 e 2020 che si assestano su dimensioni simili fra loro e nettamente inferiori rispetto a quella del 2018. In generale, la redditività delle vendite risulta positiva seppur con valori inferiori al 4 per cento anche nell'anno 2018, un risultato che mette luce il grande peso che hanno i costi della produzione nel settore metallurgico. Tuttavia, per indagare ulteriormente i risultati ottenuti col ROS e valutare l'effettiva fonte dei costi riportati a conto economico che abbattono così fortemente il fatturato, si è deciso di studiare l'andamento del seguente rapporto di bilancio:

$$\frac{\text{EBITDA}}{\text{Fatturato}} \tag{5.3}$$

Il seguente grafico riporta un confronto tra il grafico del ROS e quello del rapporto appena citato:

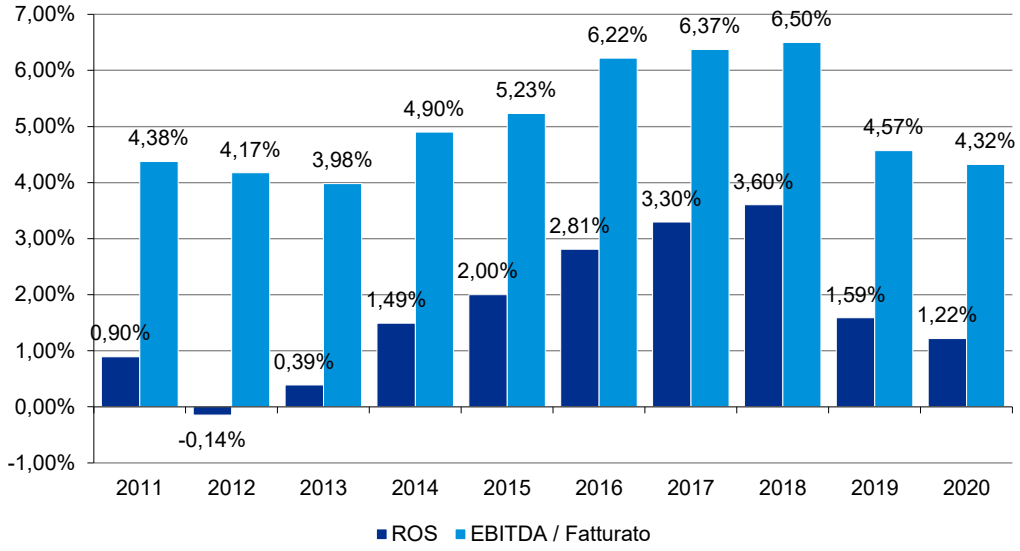


Figura 5.4: confronto tra i valori del ROS e dell'EBITDA/Fatturato

Dai valori riportati si può osservare come una parte significativa dei costi che incidono sul fatturato e producono il reddito operativo provenga dalle voci a conto economico di due componenti: gli ammortamenti, a causa dell'indispensabile necessità per le imprese del settore di macchinari ed attrezzature molto costose per realizzare le lavorazioni e con vita limitata nel tempo; le svalutazioni relative a crediti e immobilizzazioni.

Infine, il terzo indicatore che si è deciso di approfondire è il *Return on Investment* (ROI), un indice di bilancio che, a differenza del ROE, misura la redditività dell'impresa invece che degli azionisti, ossia la congruità del rendimento economico rispetto al capitale investito. Infatti, ogni società dispone di un livello di attività proporzionato alle esigenze operative e che può variare ad esempio a seconda della dimensione, del business, della struttura finanziaria. Il ROI serve dunque ad accertare se il reddito operativo (EBIT) generato dall'impresa sia in grado di remunerare adeguatamente gli investimenti effettuati. Tale indicatore è calcolato nel seguente modo:

$$\text{ROI} = \frac{\text{EBIT}}{\text{Totale Attivo}} \tag{5.4}$$

Nel grafico riportato di seguito è possibile osservare l'andamento del ROI nel periodo di riferimento e in particolare si possono effettuare le seguenti osservazioni: in primo luogo il trend seguito dal ROI è molto simile a quello presentato dal ROS, dove infatti si osserva una prima fase decrescente senza presentare tuttavia valori negativi eccetto che per l'anno 2012, seguita poi da una fase crescente con un massimo nel 2018 con poi valori di molto inferiori per gli anni 2019 e 2020; in secondo luogo, dal momento che i valori sono

generalmente positivi, si può affermare che il core business delle imprese appartenenti al campione ha prodotto per l'orizzonte temporale in analisi una remunerazione positiva relativa per investimenti effettuati.

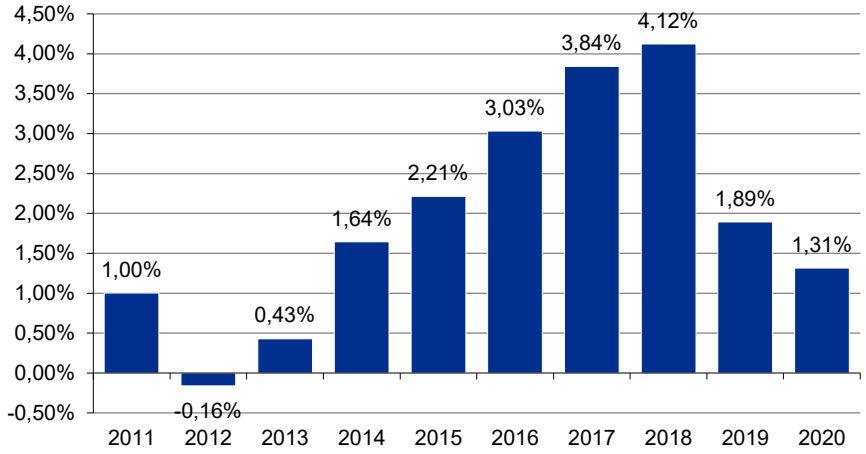


Figura 5.5: valori del ROI

Di seguito una tabella riepilogativa dei valori degli indici trattati nell'analisi di redditività:

Anno	ROE	ROS	EBITDA/Fatturato	ROI
2011	-2,81%	0,90%	4,38%	1,42%
2012	-6,56%	-0,14%	4,17%	-0,23%
2013	-7,40%	0,39%	3,98%	0,60%
2014	-2,01%	1,49%	4,90%	2,29%
2015	1,02%	2,00%	5,23%	2,98%
2016	2,40%	2,81%	6,22%	3,92%
2017	6,32%	3,30%	6,37%	4,93%
2018	6,67%	3,60%	6,50%	5,17%
2019	1,65%	1,59%	4,57%	2,33%
2020	1,89%	1,22%	4,32%	1,58%

Tabella 5.1: riepilogo dei valori assunti dagli indici di redditività analizzati

5.1.2 Analisi patrimoniale

La seconda tipologia di analisi che si è deciso di condurre è l'analisi patrimoniale, ossia un'analisi che permette di indagare la struttura degli investimenti e dei finanziamenti per valutare la capacità delle imprese del settore di mantenere nel tempo una situazione di equilibrio strutturale. I rapporti di bilancio che si è deciso di approfondire sono i seguenti:

- Indipendenza del I tipo;
- Indipendenza del II tipo;
- Rigidità.

L'*indipendenza finanziaria del I tipo* è un indicatore che permette di valutare la dimensione del patrimonio netto rispetto al totale delle fonti di finanziamento. Tale indicatore permette quindi di comprendere come il settore finanzia i propri impieghi. Matematicamente la formula è la seguente:

$$\text{Indip. fin. I tipo} = \frac{\text{Patrimonio netto}}{\text{Totale passivo}} \tag{5.5}$$

Nel grafico seguente è possibile osservare i valori di tale indicatore relativamente all'orizzonte temporale del campione di analisi:

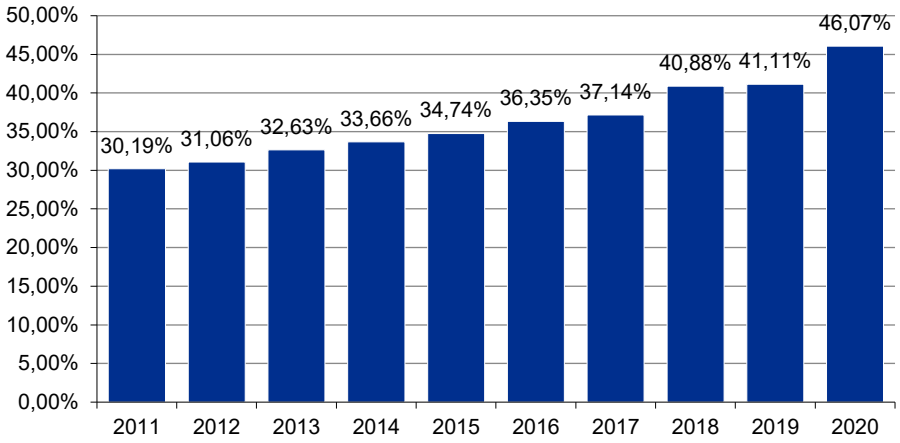


Figura 5.6: valori dell'indipendenza finanziaria del I tipo

Dal grafico si può osservare che, seppur le fonti di capitale di terzi rimangono la principale fonte di finanziamento degli impieghi delle imprese del settore in analisi, la tendenza nell'orizzonte temporale di riferimento è quella di una sempre maggior capitalizzazione e, quindi, di finanziamento degli impieghi con capitale degli azionisti.

Il secondo indice relativamente all'analisi patrimoniale è l'*indipendenza finanziaria di II tipo*, un indicatore molto simile al a quello appena trattato ma che permette di approfondire la proporzione tra le fonti di finanziamento mettendo in relazione il patrimonio netto con i debiti finanziari:

$$\text{Indip. fin. II tipo} = \frac{\text{Patrimonio netto}}{\text{Patrimonio netto} + \text{Debiti finanziari}} \tag{5.6}$$

Il grafico relativo a questo indicatore è riportato di seguito e da quest'ultimo si possono trarre le seguenti osservazioni: in primo luogo la tendenza osservata nel grafico precedente è confermata e si nota una tendenza ad una maggior capitalizzazione nel tempo.

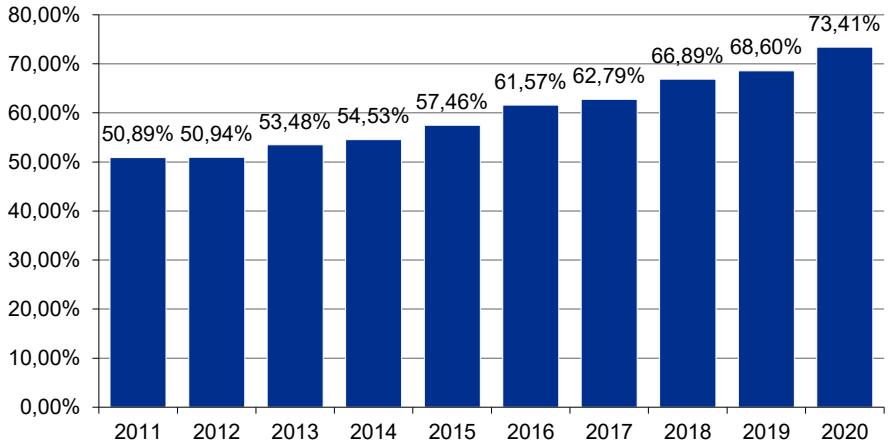


Figura 5.7: valori dell'indipendenza finanziaria del II tipo

Il terzo indice che si è deciso di approfondire è la *rigidità*, al fine di indagare la composizione degli impieghi delle imprese appartenenti al campione in analisi. La formula è la seguente:

$$\text{Rigidità} = \frac{\text{Totale immobilizzazioni}}{\text{Totale attivo}} \tag{5.7}$$

Al seguente grafico è riportato l'andamento di tale rapporto per l'orizzonte temporale in analisi e dal quale si può dedurre che la maggior parte degli impieghi è costituita da attività correnti.

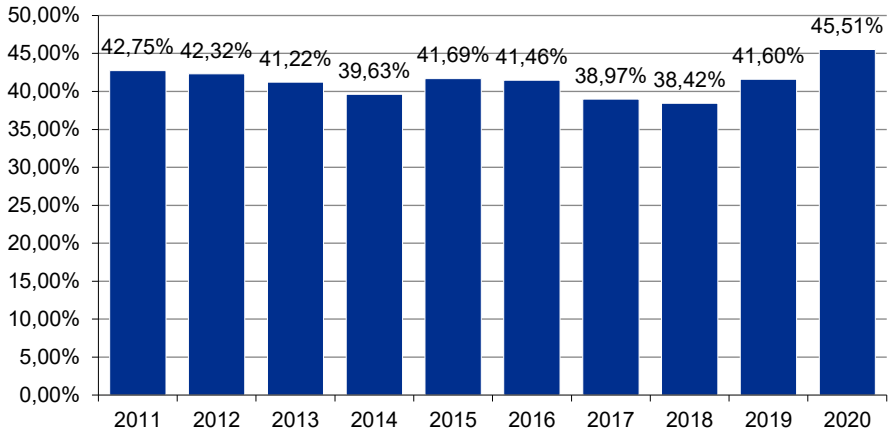


Figura 5.8: valori della rigidità

Visto il risultato appena ottenuto, si è deciso di indagare la composizione delle attività correnti e da un'analisi più approfondita si è rilevato come queste siano costituite per la maggior parte da rimanenze e crediti commerciali esigibili entro l'esercizio come osservabile al seguente grafico:

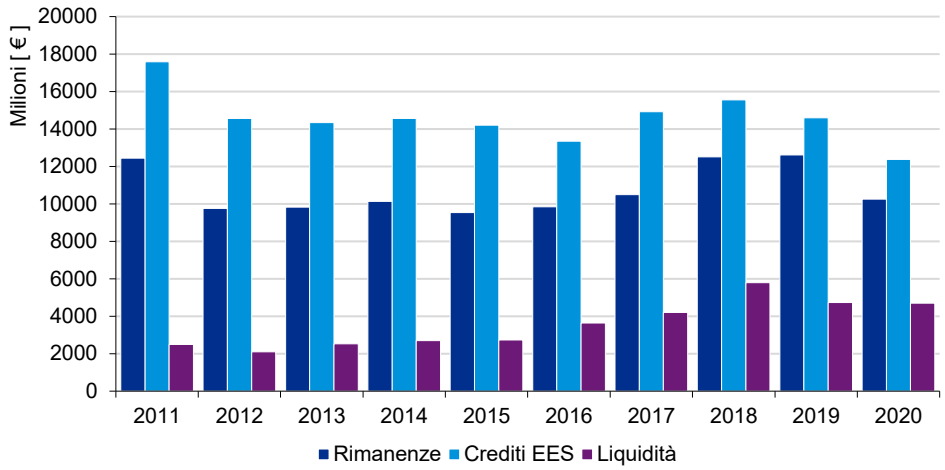


Figura 5.9: confronto fra i valori di rimanenze, crediti EES e liquidità

Si può pertanto affermare che le imprese appartenenti al campione in analisi sono caratterizzate da un alto livello di crediti commerciali a breve termine, un alto livello di rimanenze di magazzino e da un discreto livello di liquidità. Queste caratteristiche dovrebbero porre tali imprese nella condizione di poter rispondere efficacemente ad eventi imprevisti, a condizione che le rimanenze e i crediti riescano effettivamente e tramutarsi in liquidità. Di seguito una tabella riepilogativa dei valori degli indici trattati:

Anno	Indipendenza finanziaria del I tipo	Indipendenza finanziaria del II tipo	Rigidità
2011	30,19%	50,89%	42,75%
2012	31,06%	50,94%	42,32%
2013	32,63%	53,48%	41,22%
2014	33,66%	54,53%	39,63%
2015	34,74%	57,46%	41,69%
2016	36,35%	61,57%	41,46%
2017	37,14%	62,79%	38,97%
2018	40,88%	66,89%	38,42%
2019	41,11%	68,60%	41,60%
2020	46,07%	73,41%	45,51%

Tabella 5.2: riepilogo dei valori assunti dagli indici patrimoniali analizzati

5.1.3 Analisi finanziaria

L'analisi finanziaria permette di verificare se le imprese appartenenti al campione sono equilibrate o meno da un punto di vista finanziario. Gli indicatori che si è deciso di approfondire sono i seguenti:

- Indice di disponibilità;
- Indice di liquidità secondaria;
- Margine di struttura.

L'*indice di disponibilità* è un indicatore che permette di valutare quante volte le attività correnti liquidabili sono superiori o inferiori alle corrispondenti passività correnti. In linea teorica tale indicatore è da considerarsi positivo quando è maggiore di 1. La formula è la seguente:

$$\text{Indice di disponibilità} = \frac{\text{Attività correnti}}{\text{Passività correnti}} \tag{5.8}$$

Nel grafico che segue è possibile osservare i valori assunti dell'indice di disponibilità nell'orizzonte temporale di analisi.

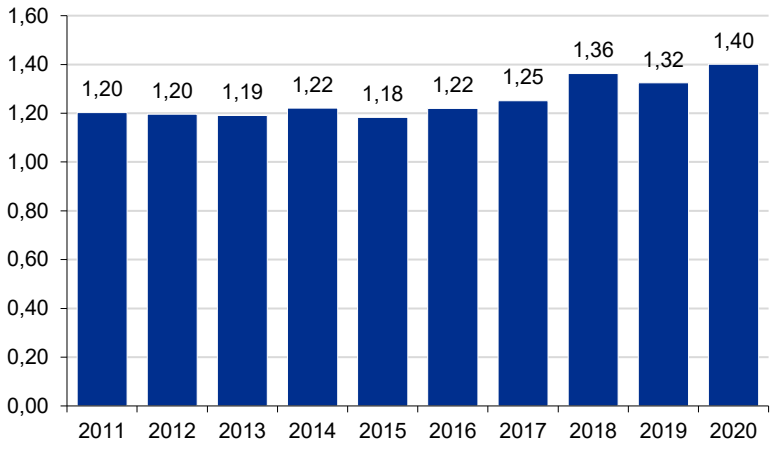


Figura 5.10: valori dell'indice di disponibilità

Dal grafico si possono trarre le seguenti conclusioni: in primo luogo l'indice presenta sempre un valore superiore ad 1, pertanto le imprese del settore sono mediamente in grado di far fronte alle loro passività di breve periodo; in secondo luogo, dal 2016 è evidente un trend in crescita dell'indice in questione.

Il secondo indice che si è deciso di indagare per valutare la situazione finanziaria del settore è l'*indice di liquidità secondaria*, ossia un indicatore che esprime la capacità di far fronte

alle passività a breve termine esclusivamente con le risorse liquide disponibili. Tale indicatore è calcolato nel seguente modo:

$$\text{Indice di liquidità secondaria} = \frac{\text{Liquidità}}{\text{Passività correnti}} \qquad (5.9)$$

Al seguente grafico è possibile osservare come tale indice riferito alle imprese nel campione in analisi per l'orizzonte temporale di riferimento presenti un trend in crescita e che porta nel 2020 la liquidità ad essere quasi un quarto delle passività correnti, un risultato molto positivo.

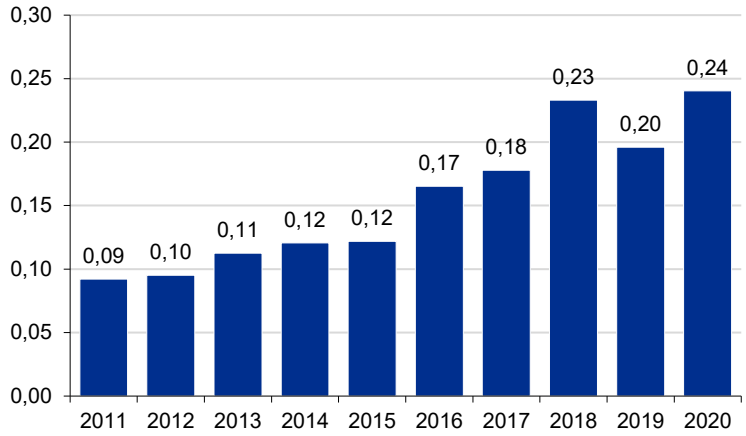


Figura 5.11: valori dell'indice di liquidità secondaria

Un ulteriore indicatore che si è deciso di approfondire è il *margine di struttura*, un indicatore che permette di valutare quale quota delle immobilizzazioni è finanziata con capitale di rischio. La formula del margine di struttura è la seguente:

$$\text{Margine di struttura} = \text{Patrimonio netto} - \text{Totale immobilizzazioni} \qquad (5.10)$$

Dal seguente grafico è possibile osservare che per quasi tutti i periodi dell'intervallo di riferimento l'indice in questione ha sempre presentato valori negativi; pertanto, il patrimonio netto non si è quasi mai rivelato sufficiente a coprire la totalità delle immobilizzazioni che, di conseguenza, devono essere state necessariamente finanziate con altre fonti di finanziamento. Tuttavia, è importante sottolineare come sia presente un trend fortemente in crescita e in linea con quanto commentato precedentemente relativamente ad una tendenza di maggior capitalizzazione nelle fonti di finanziamento delle imprese del settore.

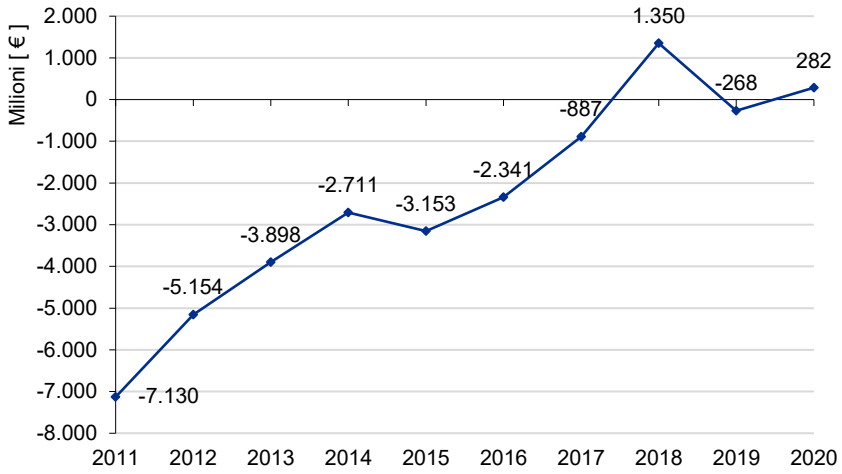


Figura 5.12: andamento del margine di struttura

Di seguito una tabella riepilogativa dei valori degli indici trattati nell'analisi patrimoniale:

Anno	Indice di disponibilità	Indice di liquidità secondaria	Patrimonio Netto - Immobilizzazioni
2011	1,20	0,09	-7.130.025.054
2012	1,20	0,10	-5.154.354.293
2013	1,19	0,11	-3.898.352.256
2014	1,22	0,12	-2.710.689.808
2015	1,18	0,12	-3.153.196.826
2016	1,22	0,17	-2.341.170.292
2017	1,25	0,18	-886.720.594
2018	1,36	0,23	1.350.136.152
2019	1,32	0,20	-268.373.399
2020	1,40	0,24	282.153.392

Tabella 5.3: riepilogo dei valori assunti dagli indici finanziari analizzati

5.2 Estrazione dei dati dal database Aida – Bureau Van Dijk

La ricerca e l'elaborazione dei dati di bilancio delle imprese appartenenti al settore metallurgico italiano per lo sviluppo dei modelli di regressione logistica e di machine learning è avvenuta tramite la banca dati AIDA grazie ad una licenza offerta dal Politecnico di Torino. Si tratta di una banca dati della società Bureau Van Dijk, un'impresa acquisita nel 2017 da Moody's e incorporata in Moody's Analytics, che offre una vasta gamma di

soluzioni per la ricerca, la consultazione, l'analisi e l'elaborazione delle informazioni economico-finanziarie, anagrafiche e commerciali di tutte le società di capitali operanti in Italia. Nello specifico, per ciascuna società è disponibile il bilancio dettagliato secondo lo schema completo della IV direttiva CEE, la serie storica fino a 10 anni, la scheda anagrafica completa di descrizione dell'attività svolta ed il bilancio ottico. Pertanto, la banca dati AIDA può risultare una soluzione ideale per svolgere le seguenti attività:

- Valutare l'affidabilità finanziaria di aziende italiane e monitorare il rischio di credito e di insolvenza del proprio portfolio per un credit e risk management strategico ed efficace;
- Effettuare analisi di benchmarking, grazie alla valutazione immediata di aziende e settori;
- Svolgere attività di due diligence e di business intelligence;
- Individuare istantaneamente le società più solide e potenziali aree di sviluppo per attività di business development;
- Ottimizzare la gestione dell'ufficio approvvigionamenti, attraverso la rapida determinazione delle parti correlate, l'attività di scouting e la realizzazione di vendor list, la creazione e il monitoraggio del proprio albo fornitori;
- Mappare i gruppi societari e le relazioni tra società, identificare immediatamente il titolare effettivo e l'azionista di riferimento, per rispondere efficacemente alle disposizioni in materia di compliance e antiriciclaggio;
- Monitorare costantemente le news della stampa economica su prospect, clienti e concorrenti;
- Integrare automaticamente il proprio sistema CRM e ERP grazie ai sistemi di data integration, aggiornando le informazioni esistenti grazie alle informazioni presenti nei database BvD;
- Svolgere ricerche, studi di settore e confronti personalizzati, grazie alle analisi statistiche avanzate.

5.2.1 Download dei dati di bilancio

Grazie alla licenza messa a disposizione dal Politecnico di Torino è stato possibile scaricare i dati di bilancio relativi alle aziende appartenenti al settore metallurgico per gli anni che vanno dal 2011 al 2020, ovvero la massima estensione temporale disponibile per questo settore. In particolare, la scelta di estrarre il maggior numero di dati di bilancio disponibili

è stata fatta al fine di poter rilevare un numero adeguato di eventi anomali per poter poi costruire modelli soddisfacenti.

Per poter scaricare i dati di bilancio relativi al settore metallurgico per l'orizzonte temporale individuato, dalla home page di AIDA è stata effettuata una ricerca per "Classificazione merceologica" cliccando sull'apposita scelta nel menù a comparsa.

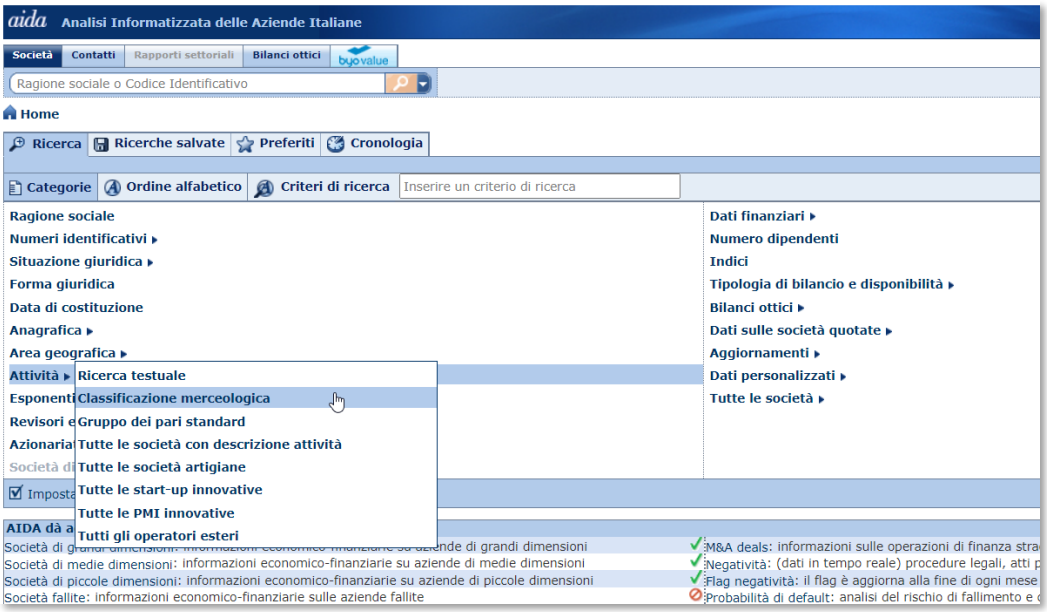


Figura 5.13: home page della banca dati Aida. Selezione dalla ricerca per classificazione merceologica

Al click si viene reindirizzati alla pagina dedicata a tale ricerca dove sono state selezionate tutte le aziende corrispondenti al codice Ateco 24, il quale individua le attività appartenenti al settore metallurgico e rilevando un totale di 3.200 imprese nel database.

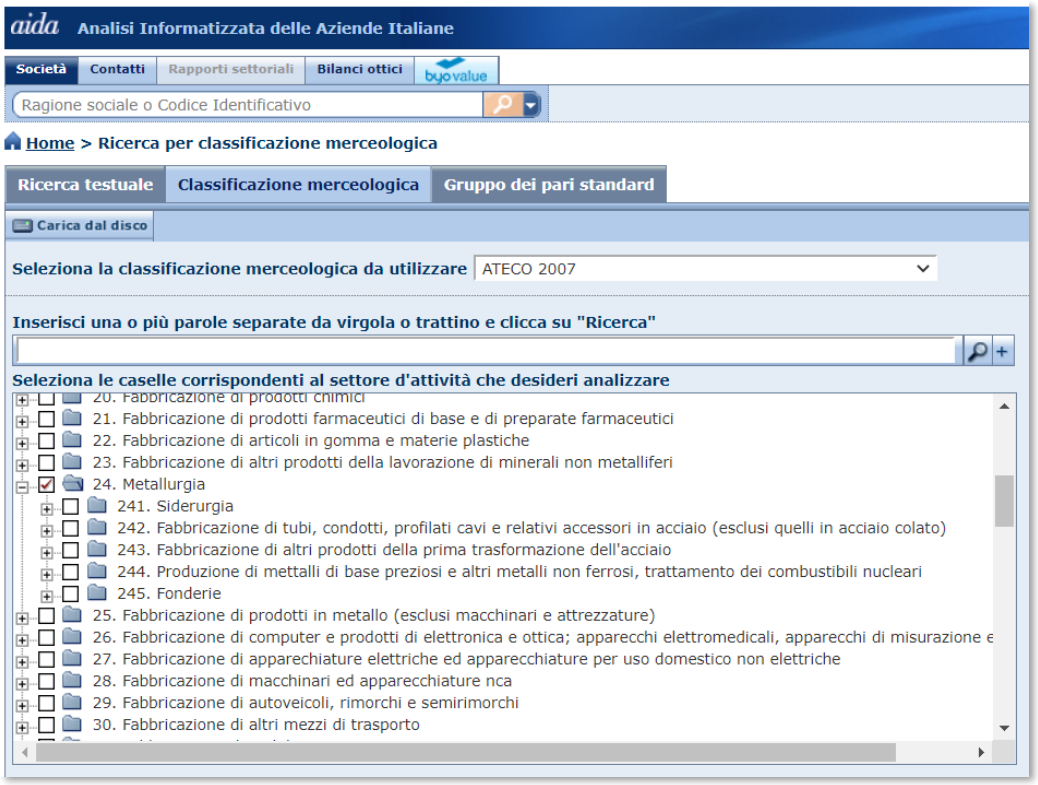


Figura 5.14: selezione per codice ATECO 24 (Metallurgia)

Una volta identificate le imprese sono state selezionate le informazioni di bilancio di interesse, ossia le informazioni anagrafiche, tutte le voci di Stato Patrimoniale e Conto Economico e i principali indici di bilancio e si è provveduto al download di queste per ognuno degli anni dell'orizzonte temporale specificato popolando un file Excel di analisi su cui sono state effettuate diverse operazioni descritte nel seguito.

Fra le informazioni scaricate, di particolare importanza al fine del raggiungimento dell'obiettivo della tesi è l'informazione anagrafica relativa allo stato giuridico ("ditta attiva", "ditta in liquidazione", "ditta in fallimento", "ditta sospesa", "ditta inattiva", "ditta cessata", "ditta cessata per trasferimento") e l'eventuale procedura subita. In particolare, sulla base delle informazioni contenute in quest'ultimo campo relativo alla procedura subita, sul file Excel di analisi sono stati assegnati cinque flag di status corrispondenti a specifiche procedure associate alle società del campione:

- **Flag = 0**: società sana senza particolari segnalazioni;
- **Flag = 1**: società classificata come anomala poiché presenta una delle seguenti procedure:
  - Concordato preventivo
  - Fallimento

- Amministrazione giudiziaria
- Accordo di ristrutturazione dei debiti
- Chiusura del fallimento
- Altre cause
- Liquidazione giudiziaria
- Motivo non precisato
- Stato di insolvenza
- Sequestro giudiziario
- Concordato fallimentare
- Amministrazione controllata
- Cancellazione per comunicazione piano di riparto
- Amministrazione straordinaria
- Chiusura per fallimento o liquidazione
- Decreto cancellazione tribunale
- Liquidazione coatta amministrativa
- Scioglimento per atto dell'autorità
- Sequestro conservativo di quote
- Bancarotta

- **Flag = 2**: società classificata come sana ma in condizioni particolari poiché presenta una delle seguenti procedure:
  - Liquidazione volontaria
  - Scioglimento e liquidazione
  - Scioglimento
  - Chiusura della liquidazione
  - Chiusura dell'unità locale
  - Cessazione di ogni attività
  - Cancellata d'ufficio ai sensi art. 2490 c.c. (bilancio di liquidazione)
  - Liquidazione
  - Scioglimento e messa in liquidazione
  - Chiusura per liquidazione
  - Scioglimento senza messa in liquidazione
  - Cessazione delle attività nella provincia
  - Cessazione d'ufficio
- **Flag = 3**: società classificata come sana ma in condizioni particolari poiché presenta una delle seguenti procedure:

- Fusione mediante incorporazione in altra società
- Scissione
- Trasferimento sede all' estero
- Fusione mediante costituzione di nuova società
- Cessione azienda
- Mancata ricostituzione pluralità dei soci
- Trasformazione in sede legale
- Trasformazione natura giuridica

- **Flag = 4**: società classificata come sana ma in condizioni particolari poiché presenta una delle seguenti procedure:
  - Cessata
  - Cancellata dal registro impresa
  - Trasferimento in altra provincia
  - Cancellata d'ufficio a seguito istituzione cciaa di fermo, di monza,..
  - Cessata d'ufficio perché già iscritta nel registro ditte e non transitata nel registro imprese
  - Provvedimento di cancellazione dal registro imprese

Sulla base della precedente classificazione per flag, sono state definite quattro colonne flag:

- FLAG DI STATUS S/A - SOCIETA': identifica una società come anomala con un "1" se questa presenta una procedura che rientra nella casistica "Flag = 1", altrimenti la società è identificata con uno "0" se rientra in tutte le altre casistiche;
- FLAG DI STATUS S/A - ANNO: identifica l'anno in cui è avvenuta la procedura che rientra nella casistica "Flag = 1" con un "1", altrimenti viene segnato uno "0";
- FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2): identifica una società come anomala con un "1" se questa presenta una procedura che rientra nella casistica "Flag = 1" o nella casistica "Flag = 2", altrimenti la società è identificata con uno "0" se rientra in tutte le altre casistiche;
- FLAG soc sana/anom+ in liquidazione - ANNO (flag=1+2): identifica l'anno in cui è avvenuta la procedura che rientra nella casistica "Flag = 1" o nella casistica "Flag = 2" con un "1", altrimenti viene segnato uno "0" se rientra in tutte le altre casistiche.

Inoltre, nel medesimo file Excel di analisi sono state aggregate le voci di Stato Patrimoniale e Conto Economico al fine di ottenere la versione abbreviata del bilancio con cui sono stati poi calcolati 46 indicatori di bilancio suddivisi in tre categorie:

- Indicatori di redditività;
- Indicatori di produttività e struttura operativa;
- Indicatori di liquidità e struttura finanziaria.

Sui dati grezzi scaricati dal database Aida sono poi state effettuate delle operazioni di pulizia e correzione, descritte nel capitolo successivo.

5.2.2 Pulizia e correzione dei dati

Dopo aver scaricato i dati di bilancio, assegnato i flag e calcolato il bilancio in forma abbreviata e i principali indicatori, si è proceduto a eseguire due controlli fondamentali:

1. È stata effettuata un'operazione di controllo e pulizia al fine di rendere "*machine readable*" il contenuto di alcune celle. In particolare, tutte le celle contenenti valori "n.d.", "n.s." sono state sostituite con degli "0", mentre tutte le celle contenenti degli errori del tipo "#DIV/0!" o "#NUM!" sono state sostituite con valori compatibili per la successiva alimentazione dei modelli statistici e di machine learning. Inoltre, sono stati eliminati tutti i record relativi a bilanci con un Totale attivo nullo ottenendo un numero totale di records per il campione in analisi pari a 19.250;
2. Tramite tre colonne di controllo sono state svolte delle verifiche di correttezza delle informazioni esportate da Aida:
    - Colonna di controllo dell'Attivo: verifica che la somma effettiva delle voci di bilancio dell'Attivo dello Stato Patrimoniale esportate da Aida corrisponda effettivamente al Totale attivo esportato;
    - Colonna di controllo del Passivo: verifica che la somma effettiva delle voci di bilancio del Passivo dello Stato Patrimoniale esportate da Aida corrisponda effettivamente al Totale passivo esportato;
    - Colonna di controllo del Conto economico: verifica che la somma effettiva delle voci di ricavo e costo accese a conto economico ed esportate da Aida corrisponda all'utile esportato.

    Nello specifico, per ogni anno si è controllato che tutte e tre le colonne riportassero valori pari a zero e laddove ciò non accadesse si è provveduto a individuare le cause delle incongruenze a livello di bilancio al fine di correggerle. In particolare, è interessante notare come la totalità delle incongruenze individuate era frutto di errori nelle somme delle singole voci di bilancio con cui si perveniva alle varie

categorie dell'Attivo, del Passivo e ai saldi intermedi nel Conto economico. Tale operazione, seppur molto dispendiosa a livello di tempo poiché richiede un'analisi minuziosa delle voci presenti a bilancio, si è rivelata tuttavia necessaria per poter avere una solida base con cui poi individuare successivamente le variabili che hanno fornito gli input ai modelli. In particolare, nel prossimo capitolo verrà descritto il procedimento con cui sono state individuate le variabili che hanno alimentato i modelli statistici e di machine learning sviluppati.

5.2.3 Individuazione delle variabili di input

Una volta conclusa la pulizia dei dati, è stato ottenuto un database dove ogni record possedeva informazioni, per uno specifico anno, sulla situazione contabile e di "salute" per una specifica impresa. Infatti, l'organizzazione e la corretta strutturazione dei dati costituiscono un passaggio fondamentale quando si opera con database di grandi dimensioni, sia per una questione di ordine ma anche perché tale impostazione consente di ottenere la massima resa dall'analisi statistica.

Le variabili con cui alimentare i modelli sono state scelte fra i 46 indicatori di bilancio calcolati nel foglio di analisi Excel e ottenuti a partire dalla versione abbreviata dei bilanci esportati da Aida, come specificato precedentemente al termine del capitolo 5.2.1. In particolare, per pervenire alla scelta delle variabili di input sono state effettuate le seguenti operazioni preliminari:

1. Gestione degli outliers;
2. Calcolo della correlazione tra ognuno dei 46 indicatori di bilancio e le colonne "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)" e "FLAG soc sana/anom+ in liquidazione - ANNO (flag=1+2)";
3. Calcolo della correlazione tra i 46 indicatori di bilancio.

Per quanto riguarda la gestione degli outliers, ossia quelle osservazioni che assumono valori estremi rispetto agli altri dati compresi nel dataset, per ogni indicatore è stato calcolato il 5° e il 95° percentile al fine di limitare i valori presentati da tutti gli indicatori all'interno di questo intervallo: ogni valore inferiore al 5° percentile è stato sostituito con il 5° percentile, mentre ogni valore superiore al 95° percentile è stato sostituito con il 95° percentile. Successivamente, per ogni indicatore è stata calcolata la sua correlazione con le colonne flag "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)" e "FLAG soc sana/anom+ in liquidazione - ANNO (flag=1+2)" controllando sempre che il segno della

correlazione fosse coerente con il significato economico dell'indicatore. Nello specifico, la motivazione per cui si è scelto di prendere in considerazione solo le colonne "flag = 1+2" è che le colonne per SOC e ANNO con "flag = 1" presentavano tassi di default pari rispettivamente a 2,72% e 2,67% (524 e 514 eventi di default), valori troppo esigui per sviluppare modelli di credit scoring significativi. Invece, i tassi di default per le colonne con "flag = 1+2" presentavano valori pari a 12,31% e 12,28% (2370 e 2363 eventi di default), rispettivamente per SOC e ANNO. Il calcolo della correlazione tra gli indicatori e le colonne con i flag si è rivelato di fondamentale importanza al fine di individuare le variabili maggiormente esplicative, anche se ciò non basta per identificare le variabili di input perché se da una parte sono state scelte le variabili con maggiore capacità diagnostica, dall'altra sono state scartate quelle variabili che, seppur fossero anche molto esplicative, presentavano una correlazione molto elevata con un'altra variabile più esplicativa. Infatti, se in un modello si procede ad inserire indicatori con buona capacità diagnostica ma fra loro molto correlati, quello che succede è che il più debole fra questi perde il segno, comportando quindi una perdita di significatività del modello oltre che ad uno spreco di tempo e di capacità di calcolo.

Al termine di queste operazioni è stato possibile individuare le seguenti variabili di input in ordine di capacità diagnostica e, quindi, in ordine di logica di inserimento nel corso dello sviluppo dei modelli:

Indicatore	FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)
ROE	0,4127
Utile corrente/ricavi	0,3736
Debiti totali/EBITDA	0,3621
Riserve+utile/AN	0,3577
Ln(RIC)	0,3323
servizi esterni/costi operativi	0,3180
Deb finanziari (stimati)/Ric	0,2983
Auto Lordo-comp straord/AN	0,2904
Patr netto tan/Debiti tot+PN	0,2675
Ricavi/AN	0,2296
Cap Circ/AN	0,1878
Ln(AN)	0,1831
Deb finanziari (stimati)/VA	0,1562
Deb totali/VA	0,1019
gg magazz	0,0391
OFN/RIC	0,0218
Patr netto/debiti totali	0,0144
OFN/AN	0,0124

Indicatore	FLAG soc sana/anom+ in liquidazione - ANNO (flag=1+2)
ROE	0,4136
Utile corrente/ricavi	0,3746
Debiti totali/EBITDA	0,3630
Riserve+utile/AN	0,3581
Ln(RIC)	0,3338
servizi esterni/costi operativi	0,3185
Deb finanziari (stimati)/Ric	0,2984
Auto Lordo-comp straord/AN	0,2909
Patr netto tan/Debiti tot+PN	0,2678
Ricavi/AN	0,2286
Cap Circ/AN	0,1880
Ln(AN)	0,1853
Deb finanziari (stimati)/VA	0,1540
Deb totali/VA	0,0995
gg magazz	0,0388
OFN/RIC	0,0243
Patr netto/debiti totali	0,0140
OFN/AN	0,0137

Tabella 5.4.1 (sx) e 5.4.2 (dx): report delle correlazioni in valore assoluto fra gli indicatori di bilancio selezionati e la colonna flag=1+2 per SOC e ANNO

Come si può osservare le correlazioni tra le due colonne sono molto simili, proprio perché la percentuale di società anomale nelle sue classificazioni sono molto simili. Tuttavia, le analisi e gli sviluppi dei modelli logistici e di machine learning sono stati condotti per entrambe le casistiche.

Infine, una volta individuati gli indicatori con cui alimentare i modelli, il dataset estratto da Aida e corretto è stato diviso in due set di dati al fine di poter non solo allenare i modelli ma anche testare le loro performance:

- **Training set** pari a 2/3 delle osservazioni totali (12.833 records) con un tasso di default per SOC e ANNO per la colonna "flag = 1+2" pari rispettivamente a 12,38% e 12,35% (1.589 e 1.585 eventi di default);
- **Test set** pari a 1/3 delle osservazioni totali (6.417 records) con un tasso di default per SOC e ANNO per la colonna "flag = 1+2" pari rispettivamente a 12,17% e 12,12% (781 e 778 eventi di default).

Su entrambi i datasets così definiti è stata poi fatta un'operazione di *normalizzazione*. Infatti, dal momento che il range dei valori può variare anche molto tra un indicatore e un altro, è fortemente consigliato un preprocessamento dei dati in input con una normalizzazione prima di procedere con l'alimentazione e il training dei modelli per una serie di ragioni:

- Gli algoritmi di ottimizzazione basati sull'algoritmo *gradient descent* tendono a convergere più rapidamente;
- Si tende ad evitare la cosiddetta *trappola del NaN*, ossia una condizione in cui un numero durante la fase di training del modello assume un valore NaN, ad esempio quando durante il training un numero si trovi ad assumere un valore che eccede il limite di precisione floating-point, il quale causa a sua volta, per via di operazioni matematiche, che ogni altro valore nel modello si trovi ad assumere anch'esso il valore NaN;
- Supporta il modello a determinare dei pesi adeguati ad ogni input dal momento che senza una normalizzazione quello che succederebbe sarebbe che il modello presterebbe un'eccessiva attenzione agli input con una range più ampio e tralascerebbe gli altri con un range più ristretto.

Una tecnica di normalizzazione che viene tipicamente utilizzata in queste casistiche è la *Z-Score*, la quale consiste nel processare ogni valore *x* degli indicatori selezionati nel seguente modo:

$$x' = \frac{x - \mu}{\sigma} \tag{5.11}$$

Ossia per ogni valore di uno specifico indicatore, il valore aggiornato è determinato sottraendo la media di quell'indicatore e dividendo il tutto per la deviazione standard dell'indicatore stesso. In questo modo, i valori aggiornati x' per ogni indicatore rappresentano quanto il valore originario x si discosta dalla media dell'indicatore a cui appartiene in termini di deviazioni standard.

A questo punto, terminato il preprocessamento dei dati, i dataset di *training* e di *test* sono stati convertiti in due file in formato .csv per favorire l'importazione col codice Python e si è successivamente passati allo sviluppo dei modelli statistici e di machine learning tramite codice e i cui dettagli saranno descritti nei capitoli seguenti.

6. Sviluppo dei modelli di credit scoring

La scelta della tecnologia con cui sviluppare i modelli statistici e di machine learning basati sulle reti neurali è ricaduta su Python, uno dei linguaggi di programmazione più famosi e utilizzati al mondo grazie alla sua sintassi concisa e semplice che gli permette di essere utilizzato negli ambiti più disparati, come ad esempio lo sviluppo di applicazioni Web o desktop, la realizzazione di GUI, l'esecuzione di task di amministrazione di sistema, il calcolo scientifico e numerico, lo sviluppo di videogames, la modellazione 3D, ecc.

Una delle caratteristiche e dei punti di forza di Python è che è molto ricco di librerie. Infatti, oltre al fatto che ogni installazione di Python vanta di default una collezione di oltre 200 moduli che permette agli utenti di svolgere numerosi compiti, è possibile scaricare ed aggiungere ulteriori moduli nel caso si rendesse necessario l'esecuzione di operazioni specifiche, come nel caso in questione per lo sviluppo di modelli di credit scoring basati sulle regressioni logistiche o sulle reti neurali. Nello specifico i moduli che sono stati selezionati e importati nell'ambiente di sviluppo sono i seguenti:

- **statsmodels**: libreria open source che fornisce classi e funzioni per diverse finalità, come ad esempio lo sviluppo di diversi modelli statistici, l'esecuzione di test statistici e l'esplorazione statistica dei dati;
- **TensorFlow**: libreria open source che fornisce moduli ottimizzati per lo sviluppo di algoritmi di machine learning e che è stata sviluppata e rilasciata per la prima volta nel 2015 da Google con una licenza open source. Negli anni tale piattaforma è diventata molto famosa tanto da essere utilizzata in molti contesti operativi aziendali, come ad esempio all'interno di Google per delle funzionalità su Gmail e Google Translate e da grandi imprese come Airbus, Coca Cola, Airbnb, Qualcomm, Twitter, PayPal.

Parallelamente alle librerie appena citate sono stati importati degli ulteriori moduli da una famosa libreria open source, la libreria **Scikit-learn**, che offre diverse funzioni di data science e machine learning. La necessità di utilizzare tale libreria è nata dal bisogno di generare delle metriche di performance comuni ad entrambe le classi di modelli al fine di poter poi effettuare un confronto al termine degli sviluppi. Nello specifico, sono stati importati i seguenti metodi dal modulo *metrics*:

- **roc_auc_score**: metodo che permette il calcolo della metrica AUC (*Area Under the ROC Curve*). In particolare, con ROC si intende il *Receiver Operating*

Characteristic, ossia uno schema di valutazione delle performance utilizzato per la valutazione di modelli di classificazione binaria. La seguente figura mostra un generico schema ROC:

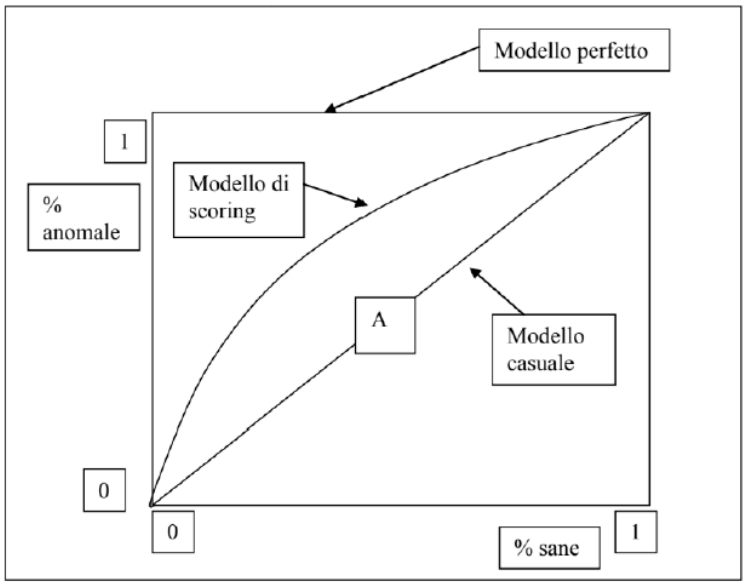


Figura 6.1: rappresentazione grafica dello schema ROC in cui è possibile visualizzare un modello perfetto, un modello generico e un modello casuale

Dalla figura si può osservare che sulle ascisse è collocata la totalità delle imprese sane mentre sulle ordinate la totalità delle imprese anomale. Pertanto, ne deriva che un eventuale modello perfetto presenterebbe una curva che percorrerebbe esattamente l'asse delle ordinate fino al valore $y = 1$ e percorrerebbe successivamente la retta orizzontale superiore fino alla retta verticale in corrispondenza di $x = 1$. Al contrario, un modello totalmente casuale, e quindi pessimo da un punto di vista della classificazione, percorrerebbe esattamente la retta diagonale fra i punti $(0,0)$ e $(1,1)$. L'AUC in questo contesto risulta particolarmente utile poiché, calcolando l'area sottesa al ROC, fornisce una metrica globale di validità del modello, dove un valore pari a 1 rappresenta un modello perfetto, un valore pari a 0,5 un modello casuale e un valore pari a 0 un modello totalmente errato che non è in grado di effettuare nessuna predizione corretta. Un altro aspetto molto interessante riguardo alla AUC per le finalità della tesi è che tale metrica viene calcolata utilizzando diverse soglie di classificazione, pertanto permette un confronto diretto e svincolato dalla scelta di una specifica soglia di classificazione per i vari modelli sviluppati, sia che siano basati su tecniche statistiche che su tecniche machine learning;

- **confusion_matrix** e **ConfusionMatrixDisplay**: metodi che permettono rispettivamente il calcolo e la visualizzazione delle matrici di confusione, ossia uno strumento di valutazione delle performance per modelli di classificazione binaria che riassume in uno schema suddiviso in quattro quadranti le performance ottenute dal modello. Lo schema è costruito confrontando i valori veri con quelli predetti e, quindi, calcolando i valori *true positive*, *true negative*, *false positive*, *false negative*. Alla seguente figura è riportato un esempio di tale schema:

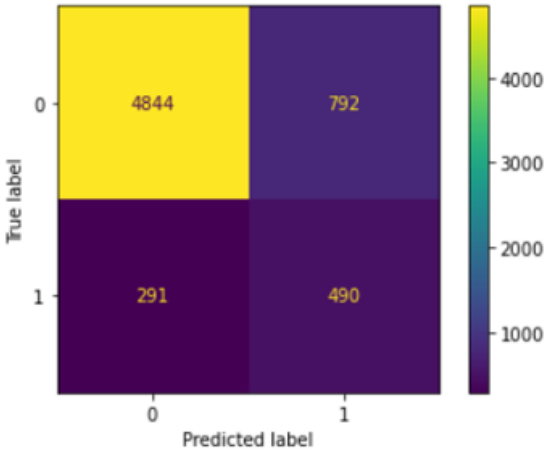


Figura 6.2: esempio di una confusion matrix

Tale schema riporta sulle righe le classi vere, mentre riporta sulle colonne le classi predette. All'interno di ogni quadrante è riportato il numero totale di osservazioni che rientrano in quella specifica casistica, quindi per esempio nello schema riportato nella precedente figura le osservazioni *false negative* ammontano a 291 (quadrante in basso a sinistra).

Una caratteristica piuttosto importante che differenzia questo schema dall'AUC è che le performance del modello di classificazione binaria rilevate con questo strumento dipendono fortemente dalla soglia di classificazione scelta, dal momento che i valori predetti sono generati a partire da tale soglia. In particolare, alla seguente figura è riportato un esempio di come la soglia di classificazione sia in grado di influenzare i valori assunti dalle quattro classi schematizzate nella confusion matrix:

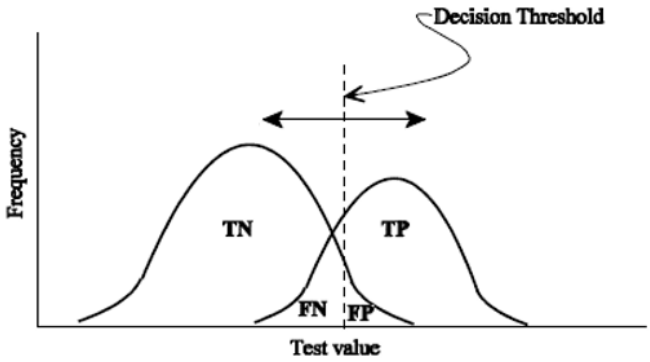


Figura 6.3: effetto dello spostamento della soglia di classificazione

La presenza di uno strumento di rilevazione delle performance che tenga in considerazione la soglia di classificazione scelta è necessaria al fine di poter valutare il modello sviluppato anche per le sue capacità predittive in uno specifico contesto operativo.

In conclusione, l'utilizzo di questi due schemi permette di avere un punto di vista generale sulle performance dei modelli sviluppati e con la quale risulta agibile un confronto fra essi al fine di poter determinare quale effettivamente risulta migliore per diagnosticare situazioni potenziali di default a partire dai dati di bilancio delle imprese sottoposte ai modelli.

6.1 Sviluppo dei modelli basati sulla regressione logistica

La prima tipologia di modelli che è stata sviluppata è quella di tipo statistica e basata sulla regressione logistica. Gli sviluppi sono stati realizzati richiamando specifiche funzioni della libreria Python open source *statsmodels*.

Dopo alcune prove per prendere confidenza con la libreria scelta, la strategia che è stata seguita per sviluppare i diversi modelli è stata quella di iniziare ad alimentare il primo modello con l'indicatore di bilancio individuato come più discriminante (indicatore di bilancio in cima alle tabelle 5.4.1 e 5.4.2), per poi aggiungere di volta in volta sempre più indicatori scorrendo le tabelle in ordine dall'alto ed effettuando i seguenti controlli:

1. Verifica della significatività statistica dei coefficienti dei regressori inseriti nel modello;
2. Verifica della coerenza del segno assunto dai coefficienti con il significato economico dei regressori a cui sono associati.

Il processo appena descritto è proseguito fino a quando una delle verifiche appena elencate ha prodotto un esito negativo. Infatti, uno degli elementi che caratterizza i modelli basati sulle regressioni logistiche è il forte rigore matematico e statistico che accompagna il processo di costruzione e per il quale se anche un solo coefficiente non rispetta una delle due condizioni appena elencate, allora l'intero modello deve necessariamente essere rigettato.

Alla seguente figura 6.4 è possibile visualizzare un report delle prove effettuate, dove è anche indicato nella colonna "Note" il modello ritenuto migliore (modello 7 per entrambe le colonne flag) da un punto di vista sia di consistenza statistica che di performance per quanto riguarda l'individuazione di società anomale. In particolare, il modello in questione è stato utilizzato come benchmark di confronto per valutare la performance ottenuta con la miglior rete neurale.

Un'osservazione importante da fare, e che è osservabile alla seguente figura 6.4, è che i risultati ottenuti dai vari modelli sviluppati in riferimento alla colonna flag=1+2 SOC e alla colonna flag=1+2 ANNO sono pressoché i medesimi. Il motivo di ciò è che la differenza in termini di società anomale tra una colonna e l'altra è esigua (la colonna SOC presenta 7 eventi di default in più rispetto alla colonna ANNO a fronte di 19.250 record del database). Tuttavia, per completezza nella produzione del lavoro e nelle analisi condotte, è stato deciso di sviluppare i modelli su entrambe le classi di problemi.

Regressione logistica su SOC					Train				Test				
Num. Modello	Nome file	Check segno coefficienti	Check significatività coefficienti	Corr. logit vs flag s/a (train)	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	Note
1	1_RL_SOC_1 reg	OK	OK	0.4193	0.7008	0.8582	0.4356	0.4915	0.6893	0.8563	0.4192	0.4686	
2	2_RL_SOC_2 reg	OK	OK	0.4627	0.7392	0.8545	0.4350	0.5859	0.7384	0.8563	0.4329	0.5826	
3	3_RL_SOC_3 reg	OK	OK	0.4615	0.7408	0.8290	0.3829	0.6237	0.7434	0.8312	0.3822	0.6274	Correlazione logit-flag inferiore rispetto al modello precedente
4	4_RL_SOC_4 reg	OK	OK	0.4685	0.7395	0.8314	0.3867	0.6174	0.7423	0.8351	0.3888	0.6197	
5	5_RL_SOC_5 reg	OK	OK	0.4742	0.7433	0.8348	0.3941	0.6218	0.7454	0.8386	0.3961	0.6223	p-value utile corrente/ricavi = 0.043
6	6_RL_SOC_6 reg	OK	KO	0.4801									p-value utile corrente/ricavi = 0.464
7	7_RL_SOC_5 reg	OK	OK	0.4697	0.7436	0.8329	0.3908	0.6249	0.7468	0.8362	0.3922	0.6287	Modello di riferimento

Regressione logistica su ANNO					Train				Test				
Num.	Nome file	Check segno coefficienti	Check significatività coefficienti	Corr. logit vs flag s/a (train)	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	Note
1	1_RL_ANNO_1 reg	OK	OK	0.4201	0.7014	0.8586	0.4358	0.4927	0.6903	0.8568	0.4192	0.4704	
2	2_RL_ANNO_2 reg	OK	OK	0.4637	0.7400	0.8549	0.8549	0.5874	0.7396	0.8568	0.4329	0.5848	
3	3_RL_ANNO_3 reg	OK	OK	0.4624	0.7413	0.8292	0.3827	0.6246	0.7447	0.8317	0.3822	0.6298	Correlazione logit-flag inferiore rispetto al modello precedente
4	4_RL_ANNO_4 reg	OK	OK	0.4694	0.7403	0.8317	0.3867	0.6189	0.7429	0.8354	0.3883	0.6208	
5	5_RL_ANNO_5 reg	OK	OK	0.4742	0.7433	0.8353	0.3950	0.6211	0.7453	0.8384	0.3958	0.6223	p-value utile corrente/ricavi = 0.048
6	6_RL_ANNO_6 reg	OK	KO										p-value utile corrente/ricavi = 0.485
7	7_RL_ANNO_5 reg	OK	OK	0.4707	0.7446	0.8335	0.3913	0.6265	0.7483	0.8372	0.3931	0.6311	Modello di riferimento

Figura 6.4: report dei modelli più significativi sviluppati con regressioni logistiche relativamente alla colonna flag=1+2 SOC (in alto) e ANNO (in basso)

6.1.1 Descrizione del codice dei modelli basati sulle regressioni logistiche

Nel presente capitolo verrà presentato il codice prodotto per sviluppare il modello statistico basato sulla regressione logistica che ha ottenuto le migliori performance per entrambe le colonne flag, ossia il modello numero 7. In particolare, nonostante il codice che verrà descritto sarà quello che fa riferimento alla colonna flag SOC, la descrizione risulterà ugualmente valida anche per la tipologia di modello basata sulla colonna flag ANNO dal momento che la differenza è minima e circoscritta al testo contenuto in una stringa in cui è sufficiente sostituire il termine "SOC" con "ANNO" per richiamare una colonna anziché l'altra.

La prima parte di codice che è stata prodotta riguarda l'importazione delle librerie fondamentali per poter gestire i dati degli indicatori scelti come input del modello:

Codice 6.1: importazione delle librerie necessarie

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import statsmodels.api as sm
from scipy import stats
from statsmodels.formula.api import logit
from sklearn.metrics import roc_auc_score
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
```

Successivamente sono stati importati i datasets di *training* e *test* sottoforma di file in formato .csv per agevolare la lettura con Python:

Codice 6.2: importazione dei dataset di training e di test

```
df_train = pd.read_csv("train_dataset_norm.csv", delimiter = ";")

df_test = pd.read_csv("test_dataset_norm.csv", delimiter = ";")
```

Il passo successivo è stato quello di definire le variabili di input e la colonna flag desiderata con cui sviluppare il modello statistico richiamando la libreria *statsmodels:*

Codice 6.3: definizione dei regressori e della variabile dipendente

```
# Definizione delle variabili di input
n_train = len(df_train["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"])

x = df_train["ROE"]
x = sm.add_constant(x)

x["Utile corrente/ricavi"] = df_train["Utile corrente/ricavi"]
x["Debiti totali/EBITDA"] = df_train["Debiti totali/EBITDA"]
x["Riserve+utile/AN"] = df_train["Riserve+utile/AN"]
x["Ln(AN)"] = df_train["Ln(AN)"]

# Definizione della colonna flag di riferimento
y_var = df_train["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"]
```

Codice 6.4: esecuzione della funzione di calcolo del modello logistico

```
1 Logit_x = sm.Logit(y_var, x)
2
3 result_x = Logit_x.fit(cov_type="hc1")
4
5 print(result_x.summary())
```

In particolare, il parametro *cov_type="hc1"* settato alla riga 3 nel precedente blocco di testo permette di costruire il modello con una regressione logistica che calcoli dei coefficienti che abbiano degli errori standard robusti all'eteroschedasticità. Si tratta di una condizione molto importante poiché se non si specificasse, e quindi si assumesse che gli errori standard fossero omoschedastici, si rischierebbe di considerare dei coefficienti statisticamente significativi quando in realtà questi non lo sono. L'output generato al lancio del blocco di codice in cui si richiama la funzione della libreria *statsmodels* è il seguente:

```
1 Logit_x = sm.Logit(y_var, x)
2
3 result_x = Logit_x.fit(cov_type="hc1")
4
5 print(result_x.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.292246
         Iterations 7
                                    Logit Regression Results
==============================================================================================================
Dep. Variable:     FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)   No. Observations:                12833
Model:                                                            Logit   Df Residuals:                    12827
Method:                                                             MLE   Df Model:                            5
Date:                                                  Sat, 22 Jan 2022   Pseudo R-squ.:                0.2196
Time:                                                          15:33:28   Log-Likelihood:              -3750.4
converged:                                                         True   LL-Null:                     -4805.6
Covariance Type:                                                    hc1   LLR p-value:                     0.000
==============================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                    -2.3534      0.034    -68.679      0.000      -2.421      -2.286
ROE                      -0.3596      0.036     -9.938      0.000      -0.431      -0.289
Utile corrente/ricavi    -0.2339      0.032     -7.250      0.000      -0.297      -0.171
Debiti totali/EBITDA      0.2677      0.038      7.015      0.000       0.193       0.342
Riserve+utile/AN         -0.3367      0.044     -7.600      0.000      -0.424      -0.250
Ln(AN)                   -0.1440      0.034     -4.178      0.000      -0.211      -0.076
==============================================================================
```

Figura 6.5: output della libreria statsmodels per la regressione logistica

Come si può osservare dal report riportato alla figura 6.5, la regressione logistica si è basata sul modello *logit* e la metodologia di calcolo si è fondata sulla *Maimum Likelihood Estimation* (MLE). Tutti i regressori utilizzati presentano dei coefficienti che risultano statisticamente significativi oltre che avere dei segni coerenti con il corrispondente significato economico.

Una volta ottenuti i valori dei coefficienti, questi sono stati estratti tramite il metodo *result_x.params.values* che crea un vettore contenete il valore della costante e di tutti i coefficienti calcolati nel modello (in posizione 0 è presente il valore della costante mentre in ultima posizione il valore dell'ultimo coefficiente, ossia *Ln(AN)* in questo caso).

A questo punto, una volta raccolta l'informazione relativa al valore dei coefficienti, si è proceduto a effettuare le seguenti operazioni, sempre sul training set:

1. Calcolo dei vettori *logit* e PD per ogni record compreso nel training set;
2. Calcolo della matrice di correlazione tra la colonna flag SOC e il vettore dei logit: per testare la qualità del modello sviluppato: si tratta di un test alternativo allo pseudo R^2 e che permette di valutare la qualità del modello sviluppato;
3. Calcolo dell'AUC e della matrice di confusione (per un riferimento si rimanda al codice riportato per il test set al blocco di codice 6.7 in cui l'unica differenza rispetto a quello prodotto per i training set è il dataframe *df_test* da cui sono richiamati i dati, dove nel caso del training set ha la dicitura *df_train*).

Di seguito è riportato il codice prodotto per effettuare le operazioni appena elencate. Per completezza si riporta di seguito anche la formula presente alla riga 11 che per motivi di spazio è stata tagliata e con cui è stata calcolata la generica posizione *i* all'interno del vettore contenente i valori dei *logit*:

$$logit\_train[i] = const + \sum_{j=\{regressori\}} coeff_j \cdot var_{ij} \qquad (6.1)$$

Dove l'indice *i* rappresenta la riga del training dataset a cui si sta fendo riferimento, mentre var_{ij} indica il valore assunto alla posizione *i* dal regressore *j* inserito nel modello.

Codice 15: calcolo del vettore logit, del vettore delle PD, dei valori predetti e della matrice di correlazione tra logit e variabile dipendente

```
# Calcolo vettore logit e PD per classificazione
n_anomale_train = 0
for i in range(n_train):
    if df_train["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"][i] == 1:
        n_anomale_train += 1
threshold = n_anomale_train / n_train

logit_train, probDef_train, y_pred_train = [0] * n_train, [0] * n_train, [0] * n_train

for i in range(n_train):
    logit_train[i] = const + ROE_coeff * df_train["ROE"][i] + UtileCorrente_Ricavi_coeff * df_train["Utile corrente/ricavi"]
    probDef_train[i] = 1 / (1 + np.exp(- logit_train[i]))
    if probDef_train[i] >= threshold:
        y_pred_train[i] = 1
    else:
        y_pred_train[i] = 0

# Inserimento nel dataframe colonna logit e predizioni default
df_train["Logit"] = logit_train
df_train["flag_pred s/a"] = y_pred_train

# Calcolo matrice di correlazione tra logit e flag s/a
df_corr = pd.DataFrame({
    "flag s/a": df_train["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"],
    "Logit": df_train["Logit"]})

print("Matrice di correlazione:")
df_corr.corr()
```

```
Matrice di correlazione:
```

	flag s/a	Logit
flag s/a	1.0000	0.4697
Logit	0.4697	1.0000

Per quanto riguarda la soglia di classificazione, dopo aver effettuato numerose prove partendo dal canonico settaggio pari a 0,5 e provando diversi valori sia al di sopra che al di sotto di tale parametro, le soglie di classificazione che hanno prodotto le migliori performance sin dai primi modelli sono state quelle con dei valori prossimi al tasso di default campionario del training set. Pertanto, alla riga 6 nella precedente figura, si può osservare come la soglia di classificazione scelta sia stata proprio la percentuale di default nel campione di training. In alternativa, quello che si sarebbe potuto fare per ottenere un risultato molto puntuale è, tramite un ciclo programmato su Python, individuare il punto di lavoro in grado di garantire le migliori performance di classificazione; tuttavia, in questo modo la soglia sarebbe stata stabilita prescindendo da qualsiasi ragionamento economico e affidandosi completamente ad un algoritmo. Per tale ragione, alla fine quello che è stato deciso di fare è stabilire una soglia di cut-off pari al tasso di default campionario.

Successivamente, fra la riga 10 e la riga 16 tramite un ciclo for e una condizione if-else è stato calcolato dapprima il vettore dei *logit* secondo la formula 6.1 e con il quale è stato poi calcolato il vettore delle PD secondo la seguente formula:

$$probDefault\_train[i] = \frac{1}{1 + e^{-logit\_train[i]}} \tag{6.2}$$

Successivamente, tramite una condizione if-else, ogni valore contenuto nel vettore contenente le PD è stato filtrato utilizzando la soglia di classificazione scelta ed è stato quindi prodotto un vettore di valori predetti 0-1 chiamato y_pred_train.

Infine, per calcolare la matrice di correlazione tra il vettore contenente i logit e la colonna flag è stato utilizzato il metodo *.corr()* e il cui output è visibile nella al precedente blocco di codice 6.5.

Un controllo che è stato scelto di fare per valutare la consistenza del modello statistico deriva da un'osservazione della letteratura che stabilisce che la somma di tutte le probabilità di default corrisponde alla somma degli eventi di default osservati, ossia alla somma degli "1" presenti nella colonna flag. Tale concetto può essere espresso matematicamente secondo la seguente formula:

$$n_a = \sum_{i=1}^{n} PD_i \tag{6.3}$$

Dove con n_a si indica il numero di osservazioni anomale nel campione di training e con n il numero di osservazioni totali nel medesimo campione. Pertanto, per ogni modello è stata

effettuata tale verifica e che ha sempre dato esiti positivi. Alla seguente figura è riportato il codice sviluppato per effettuare tale controllo:

Codice 6.6: codice per calcolo di consistenza modello statistico tale per cui PD = somma eventi anomali

```
# VERIFICA CORRETTEZZA SOMMA PD = SOMMA ANOMALE
sum_pd = 0
for i in range(n_train):
    sum_pd += probDef_train[i]

print("Somma PD:\n", sum_pd)
print("\nN. anomale nel train set:\n", n_anomale_train)
```

```
Somma PD:
 1588.999999999994

N. anomale nel train set:
 1589
```

Una volta terminate tutte le operazioni sul training set, si è proceduto a testare il modello sviluppato sul test set per valutarne le performance. Anche in questo caso sono state ripetute le operazioni svolte sul training set per determinare il vettore dei logit, il vettore delle PD e il vettore delle predizioni sulla base della soglia di classificazione impostata.

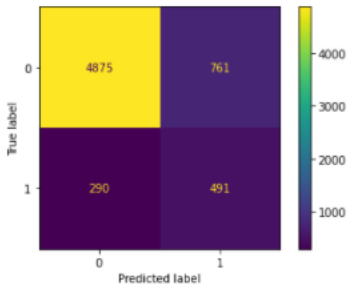
Infine, con l'ultimo blocco di codice riportato di seguito, si è proceduto a determinare anche per il test set il valore dell'AUC e a plottare la confusion matrix:

Codice 6.7: calcolo metriche di performance sul test set

```
1  # Calcolo AUC
2  auc_score_test = roc_auc_score(df_test["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"],
3                                 df_test["flag_pred s/a"], average=None)
4  print("Metriche di performance:\n- AUC: ", float("{:.4f}".format(auc_score_test)))
5
6  # Metriche performance
7  tp, tn, fp, fn = 0, 0, 0, 0
8
9  for i in range(n_test):
10     if df_test["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"][i] == 1 and df_test["flag_pred s/a"][i] == 1:
11         tp += 1
12     elif df_test["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"][i] == 0 and df_test["flag_pred s/a"][i] == 0:
13         tn += 1
14     elif df_test["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"][i] == 1 and df_test["flag_pred s/a"][i] == 0:
15         fn += 1
16     elif df_test["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"][i] == 0 and df_test["flag_pred s/a"][i] == 1:
17         fp += 1
18
19 accuracy = (tp + tn) / n_test #fraction of predictions our model got right
20 precision = tp / (tp + fp) #What proportion of positive identifications was actually correct?
21 recall = tp / (tp + fn) #What proportion of actual positives was identified correctly?
22
23 print("- Accuracy: ", float("{:.4f}".format(accuracy)),
24       "\n- Precision: ", float("{:.4f}".format(precision)),
25       "\n- Recall: ", float("{:.4f}".format(recall)))
26
27 # Plot Confusion Matrix
28 cm = confusion_matrix(df_test["FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)"], df_test["flag_pred s/a"])
29 cm_disp = ConfusionMatrixDisplay(confusion_matrix=cm)
30 cm_disp.plot()
```

```
Metriche di performance:
- AUC:  0.7468
- Accuracy:  0.8362
- Precision:  0.3922
- Recall:  0.6287
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1a8211aeb80>
```



Come si può notare, tra la riga 9 e la riga 17, per arricchire il codice sono state calcolate tre ulteriori metriche di performance che forniscono una prospettiva più di dettaglio sul modello, anche se tali punti di vista sono già implicitamente integrati nel calcolo dell'AUC. Nello specifico, i significati delle metriche calcolate sono i seguenti:

- **Accuracy** $= \frac{TP+TN}{TP+TN+FP+FN}$: metrica che fornisce la percentuale di osservazioni classificate correttamente sulla base del totale delle osservazioni nel dataset;
- **Precision** $= \frac{TP}{TP+FP}$: metrica che stabilisce quale porzione delle diagnosi anomale totali, ossia pari a 1, è effettivamente corretta;
- **Recall** $= \frac{TP}{TP+FN}$: metrica che stabilisce quale porzione delle osservazioni anomale è stata identificata correttamente.

L'output ottenuto eseguendo il precedente blocco di codice consiste in un breve report delle metriche calcolate e il plot della confusion matrix sulla base della soglia di classificazione impostata precedentemente.

6.1.2 Commenti sui modelli logistici sviluppati

Il codice descritto nel capitolo precedente è stato prodotto facendo sì che con delle piccole modifiche potesse essere adattato a diversi regressori in ingresso e ad entrambe le colonne flag. In totale sono state effettuate sette prove per la colonna flag per SOC e altrettante per la colonna flag ANNO, per un totale quindi di quattordici modelli.

La seguente tabella 6.1 riporta le strutture dei modelli logistici che sono stati sviluppati prendendo come riferimento la colonna flag SOC:

Regressori	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
ROE	-0,8438*** (0,021)	-0,6485*** (0,025)	-0,5654*** (0,029)	-0,3599*** (0,037)	-0,3851*** (0,036)	-0,3725*** (0,036)	-0,3596*** (0,036)
Utile corrente/ricavi		-0,4487*** (0,024)	-0,3341*** (0,029)	-0,2681*** (0,031)	**-0,0861** (0,043)**	**-0,0326 (0,044)**	-0,2339*** (0,032)
Debiti totali/EBITDA			0,2468*** (0,038)	0,2676*** (0,038)	0,2407*** (0,038)	0,2507*** (0,039)	0,2677*** (0,038)
Riserve+utile/AN				-0,3570*** (0,045)	-0,3378*** (0,044)	-0,3304*** (0,042)	-0,3367*** (0,044)
Ln(RIC)					-0,2623*** (0,042)	**-0,1097** (0,050)**	
Servizi esterni/costi operativi						0,2870*** (0,041)	
Ln(AN)							-0,1440*** (0,034)
Intercetta	-2,2435*** (0,032)	-2,3000*** (0,033)	-2,3287*** (0,034)	-2,3464*** (0,034)	-2,3554*** (0,034)	-2,3688*** (0,034)	-2,3534*** (0,034)

*Tabella 6.1: report regressioni logistiche su colonna flag=1+2 SOC. Gradi di significatività dei coefficienti: * se p < 0,10; ** se p < 0,05; *** se p < 0,01*

Come già indicato in precedenza, la strategia seguita per costruire i diversi modelli è stata quella di iniziare ad alimentare il primo modello con l'indicatore individuato come più discriminante per poi aggiungere di volta in volta sempre più indicatori avendo premura di verificare la significatività statistica dei coefficienti, la consistenza del segno assunto dai coefficienti con il significato economico della rispettiva variabile e la correlazione tra la colonna flag e il vettore dei logit. In generale, a livello di performance sono sempre stati riscontrati dei miglioramenti aggiungendo nuovi regressori, tuttavia la stessa cosa non può dirsi per la significatività statistica dei coefficienti. Infatti, il primo modello che ha iniziato a dare segni di perdita di significatività statistica è stato il modello 5, dove il regressore *Utile corrente/ricavi* è passato per la prima volta da un p-value pari a 0 a un p-value pari a 0.043, come osservabile dalla seguente figura:

```
Optimization terminated successfully.
         Current function value: 0.291161
         Iterations 7
                                  Logit Regression Results
==============================================================================================================
Dep. Variable:     FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)   No. Observations:                12833
Model:                                                            Logit   Df Residuals:                    12827
Method:                                                             MLE   Df Model:                            5
Date:                                                  Wed, 19 Jan 2022   Pseudo R-squ.:                 0.2225
Time:                                                          22:52:17   Log-Likelihood:                -3736.5
converged:                                                         True   LL-Null:                       -4805.6
Covariance Type:                                                    hc1   LLR p-value:                    0.000
==============================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                     -2.3554      0.034    -68.674      0.000      -2.423      -2.288
ROE                       -0.3851      0.036    -10.577      0.000      -0.457      -0.314
Utile corrente/ricavi     -0.0861      0.043     -2.021      0.043      -0.170      -0.003
Debiti totali/EBITDA       0.2407      0.038      6.305      0.000       0.166       0.316
Riserve+utile/AN          -0.3378      0.044     -7.740      0.000      -0.423      -0.252
Ln(RIC)                   -0.2623      0.042     -6.318      0.000      -0.344      -0.181
==============================================================================
```

Figura 6.6: output regressione logistica con iniziale perdita di significatività per il coefficiente del regressore "Utile corrente/ricavi" a seguito dell'inserimento della variabile Ln(RIC)

La perdita di significatività del coefficiente in questione è stata prodotta a seguito dell'inserimento nel modello del regressore *Ln(RIC)*, il quale presentava una correlazione molto forte con il denominatore del regressore in esame. In particolare, è interessante notare come nel modello 6, riportato alla seguente figura, la perdita di significatività statistica è stata accentuata inserendo un ulteriore regressore che ha avuto l'effetto di portare il p-value della variabile *Utile corrente/ricavi* ad un valore pari a 0,464:

```
Optimization terminated successfully.
         Current function value: 0.288768
         Iterations 7
                                  Logit Regression Results
==============================================================================================================
Dep. Variable:     FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)   No. Observations:                12833
Model:                                                            Logit   Df Residuals:                    12826
Method:                                                             MLE   Df Model:                            6
Date:                                                  Wed, 19 Jan 2022   Pseudo R-squ.:                 0.2289
Time:                                                          22:58:04   Log-Likelihood:                -3705.8
converged:                                                         True   LL-Null:                       -4805.6
Covariance Type:                                                    hc1   LLR p-value:                    0.000
==============================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                         -2.3688      0.034    -68.716      0.000      -2.436      -2.301
ROE                           -0.3725      0.036    -10.274      0.000      -0.444      -0.301
Utile corrente/ricavi         -0.0326      0.044     -0.733      0.464      -0.120       0.055
Debiti totali/EBITDA           0.2507      0.039      6.480      0.000       0.175       0.327
Riserve+utile/AN              -0.3304      0.042     -7.808      0.000      -0.413      -0.247
Ln(RIC)                       -0.1097      0.050     -2.208      0.027      -0.207      -0.012
servizi esterni/costi operativi  0.2870    0.041      7.065      0.000       0.207       0.367
==============================================================================
```

Figura 6.7: output regressione logistica con perdita di significatività per il coefficiente del regressore "Utile corrente/ricavi" e iniziale perdita di significatività per il coefficiente del regressore "Ln(RIC)" a seguito dell'inserimento della variabile "servizi esterni/costi operativi" al modello della figura 6.6

Inoltre, anche il coefficiente del regressore *Ln(RIC)* si può osservare come inizi a perdere di significatività statistica. La giustificazione alla quale si è giunti relativamente a tale perdita di significatività statistica dei coefficienti è che i valori *Ln(Ric)* e *Utile corrente / ricavi* sono rispettivamente un flusso e un rapporto tra flussi, ossia dei valori di bilancio generati a conto economico e che in queste tipologie di analisi tendono spesso ad assumere

dei comportamenti più instabili rispetto ad altri parametri, come ad esempio i valori di stock presenti Stato Patrimoniale. Infatti, per risolvere tale problematica sono stati fatti diversi ulteriori tentativi e la soluzione che è stata trovata è stata quella di partire nuovamente dal modello 4 e inserire come ulteriore regressore il *Ln(AN)*, ossia un dato di stock:

```
Optimization terminated successfully.
         Current function value: 0.292246
         Iterations 7
                                    Logit Regression Results
==============================================================================================================
Dep. Variable:     FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)   No. Observations:                12833
Model:                                                           Logit   Df Residuals:                    12827
Method:                                                            MLE   Df Model:                            5
Date:                                                 Thu, 20 Jan 2022   Pseudo R-squ.:                 0.2196
Time:                                                         18:15:18   Log-Likelihood:               -3750.4
converged:                                                        True   LL-Null:                      -4805.6
Covariance Type:                                                   hc1   LLR p-value:                   0.000
==============================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                   -2.3534      0.034    -68.679      0.000      -2.421      -2.286
ROE                     -0.3596      0.036     -9.938      0.000      -0.431      -0.289
Utile corrente/ricavi   -0.2339      0.032     -7.250      0.000      -0.297      -0.171
Debiti totali/EBITDA     0.2677      0.038      7.015      0.000       0.193       0.342
Riserve+utile/AN        -0.3367      0.044     -7.600      0.000      -0.424      -0.250
Ln(AN)                  -0.1440      0.034     -4.178      0.000      -0.211      -0.076
==============================================================================
```

Figura 6.8: output regressione logistica di riferimento per colonna flag=1+2 SOC

Tale strategia si è rivelata vincente dal momento che non solo è stato possibile inserire un ulteriore regressore al modello 4 senza incorrere in perdite di significatività dei coefficienti del modello, ma si sono anche ottenuti i valori di AUC più alti rispetto a tutti i modelli sviluppati precedentemente per entrambi i dataset.

Alla seguente figura 6.9 è possibile osservare le confusion matrix sviluppate sui test set per i quattro modelli logistici individuati come più significativi. Una prima osservazione che si può fare è che il modello 1 è quello che è meglio in grado a diagnosticare le imprese sane, ma allo stesso tempo anche il peggiore a diagnosticare le imprese effettivamente anomale, presentando inoltre il più alto numero di classificazioni come *false negative*. Dal momento che per un intermediario finanziario risulta peggiore lo scenario in cui si accolga la richiesta di finanziamento di un soggetto non meritevole di credito piuttosto che lo scenario in cui si rigetti la richiesta di finanziamento da un soggetto meritevole, si è scelto di scartare il modello 1 e di concentrarsi sui rimanenti modelli. In particolare, la motivazione economica per cui il primo scenario descritto risulta peggiore è che in quella casistica l'istituto di credito rischierebbe di perdere sia il capitale prestato che gli interessi maturati su questo, mentre nel secondo scenario la perdita sarebbe costituita solo dai mancati interessi attivi che si sarebbero percepiti sul capitale concesso a credito.

Scartato il modello 1, dal momento che un confronto qualitativo risulta proibitivo essendo i quadranti dei modelli rimanenti molto simili fra loro, al fine di determinare quale sia il

modello migliore sulla base della soglia di classificazione scelta in precedenza si è deciso di attribuire i seguenti pesi *w* agli errori di classificazione commessi per determinate delle funzioni di costo con cui individuare il modello migliore:

- $\mathrm{w} = 1$ per le osservazioni classificate come *false positive*;
- $\mathrm{w} = 20$ per le osservazioni classificate come *false negative*.

L'espressione della funzione di costo (*C*) è la seguente:

$$C = w_{FP} \cdot n_{FP} + w_{FN} \cdot n_{FN} \tag{6.4}$$

Pertanto, le funzioni di costo prodotte per i modelli in esame sono le seguenti:

- Modello 1: $507 \cdot 1 + 415 \cdot 20 = 8.807$
- Modello 4: $761 \cdot 1 + 297 \cdot 20 = 6.701$
- Modello 5: $741 \cdot 1 + 295 \cdot 20 = 6.641$
- Modello 7: $761 \cdot 1 + 290 \cdot 20 = 6.561$

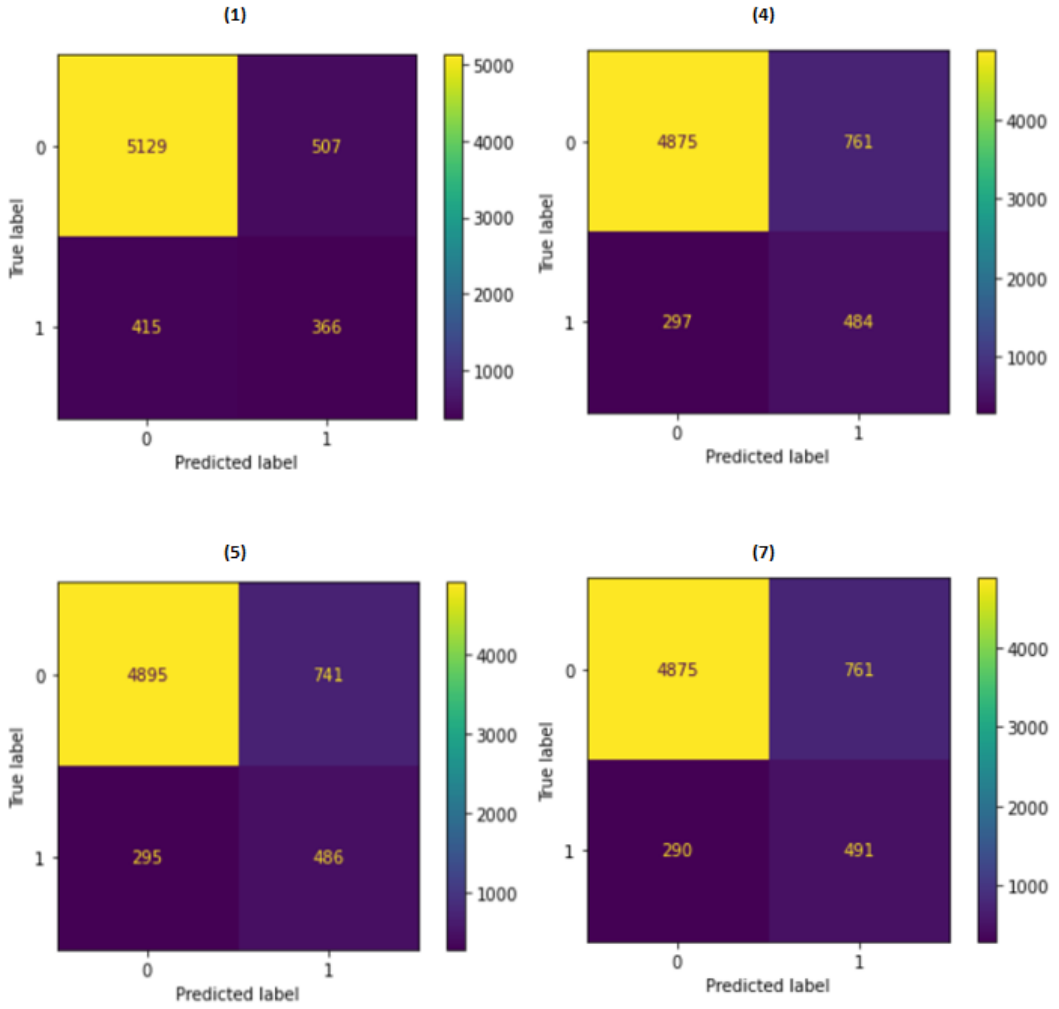


Figura 6.9: confusion matrix dei quattro modelli più significativi relativi alla colonna flag=1+2 SOC

In conclusione, a seguito delle analisi condotte considerando lo score dell'AUC e delle funzioni di costo per fornire una lettura alle confusion matrix, il modello migliore è risultato essere il modello numero 7.

Di seguito è riportata una tabella in cui sono messe in evidenza le strutture dei modelli logistici costruiti prendendo come riferimento la colonna flag ANNO e le matrici di confusione più significative per tali modelli. Come si nota, i risultati sono molto simili a quelli ottenuti per i modelli sviluppati sulla base della colonna flag SOC e il motivo alla base di ciò è che la differenza di numerosità di osservazioni anomale tra un flag e l'altro è davvero minima e relativa a poche unità. Tuttavia, come già espresso in precedenza, per completezza si è deciso di effettuare nuovamente le medesime analisi già condotte sulla colonna flag SOC e le conclusioni alle quali si è pervenuti sono le medesime, ossia prevedono che il modello migliore sia il numero 7.

Regressori	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ROE	-0,8456*** (0,021)	-0,6498*** (0,025)	-0,5662*** (0,029)	-0,3613*** (0,037)	-0,3870*** (0,036)	-0,3744*** (0,036)	-0,3610*** (0,036)
Utile corrente/ricavi		-0,4501*** (0,024)	-0,3349*** (0,029)	-0,2690*** (0,031)	**-0,0843** (0,043)**	**-0,0311 (0,044)**	-0,2332*** (0,032)
Debiti totali/EBITDA			0,2482*** (0,038)	0,2690*** (0,038)	0,2416*** (0,038)	0,2517*** (0,039)	0,2691*** (0,038)
Riserve+utile/AN				-0,3561*** (0,045)	-0,3366*** (0,044)	-0,3292*** (0,042)	-0,3348*** (0,044)
Ln(RIC)					-0,2662*** (0,042)	**-0,1151** (0,050)**	
Servizi esterni/costi operativi						0,2846*** (0,041)	
Ln(AN)							-0,1513*** (0,034)
Intercetta	-2,2481*** (0,032)	-2,3051*** (0,033)	-2,3342*** (0,034)	-2,3518*** (0,035)	-2,3610*** (0,034)	-2,3743*** (0,035)	-2,3595*** (0,034)

*Tabella 6.2: report regressioni logistiche su colonna flag=1+2 ANNO. Gradi di significatività dei coefficienti: * se p < 0,10; ** se p < 0,05; *** se p < 0,01*

Le confusion matrix più significative sono state quelle relative ai medesimi modelli sviluppati per la colonna flag SOC:

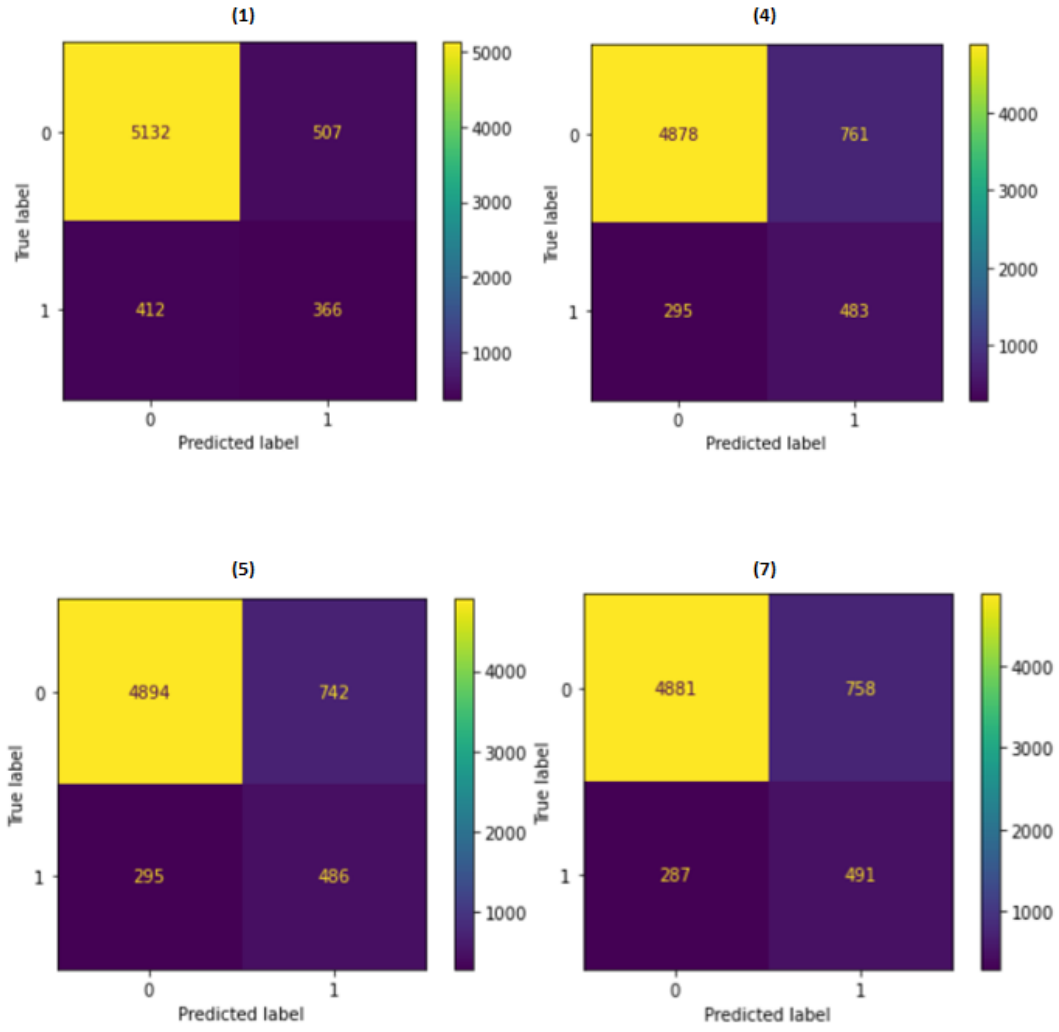


Figura 6.10: confusion matrix dei quattro modelli più significativi relativi alla colonna flag=1+2 ANNO

Anche in questo caso sono state calcolate tre funzioni di costo per determinare il modello migliore e il risultato ottenuto è stato il medesimo del gruppo d modelli precedente, identificando il modello 7 come migliore sia per quanto riguarda lo score dell'AUC che per le funzioni di costo calcolate:

- Modello 1: $507 \cdot 1 + 412 \cdot 20 = 8.747$
- Modello 4: $761 \cdot 1 + 295 \cdot 20 = 6.661$
- Modello 5: $742 \cdot 1 + 295 \cdot 20 = 6.642$
- Modello 7: $758 \cdot 1 + 287 \cdot 20 = 6.498$

In conclusione, si può dire che entrambi i modelli sviluppati generino delle performance discrete considerando la metrica AUC, la quale riporta dei valori pari a 0,7468 e 0,7483 per il modello 7 e relativamente alla colonna "flag 1+2" SOC e ANNO. Infatti, tenendo in considerazione che un modello che genera predizioni totalmente casuali possiede un'AUC pari a 0,5, si può dire che i modelli ottenuti siano sostanzialmente a metà strada tra un modello perfetto e uno totalmente casuale.

Nel prossimo capitolo verrà descritto il processo tramite il quale sono state determinate le reti neurali, i relativi commenti sulle performance ottenute e le criticità riscontrate.

6.2 Modelli basati sulle reti neurali

La seconda tipologia di modelli che è stata sviluppata si basa sulle reti neurali. Dopo un primo numero di prove per prendere confidenza con la libreria di TensorFlow, la strategia che è stata seguita per la costruzione dei modelli è stata quella provare diverse architetture, sia in termini di neuroni in input nella rete che di numerosità degli strati nascosti e dei relativi neuroni contenuti al loro interno. Infatti, la letteratura e gli studi condotti sino a questo momento non forniscono delle regole universalmente valide e ottimali per quanto riguarda la costruzione delle architetture e il settaggio degli iperparametri dei modelli, ossia quei parametri non addestrabili e che vengono forniti dagli sviluppatori della rete per regolare le dinamiche di apprendimento dell'algoritmo. Nello specifico, nel modello sviluppato gli iperparametri settati sono stati il *learning rate*, il numero di *epochs* e la dimensione del *batch size*. In particolare, con tali termini si fa riferimento ai seguenti concetti:

- **Learning rate**: coefficiente che determina con quale estensione avvengono le modifiche dei pesi della rete a seguito della determinazione della loss function dopo una specifica iterazione;
- **Epochs**: numero di cicli che l'algoritmo di apprendimento svolge sul training set;
- **Batch size**: numero di campioni processati per ogni epoch (generalmente si cerca di impostare valori che siano dei multipli di 2 per via del fatto che l'architettura dei computer basa il suo funzionamento sul calcolo binario).

In particolare, dopo numerose prove si è appurato che gli iperparametri che determinavano performance migliori nei modelli erano i seguenti: epochs = 200; batch size = 64; learning rate = 0,01 (a volte è stato necessario ridurre il learning rate di un ordine di grandezza portandolo a 0,001 per risolvere problemi di overfitting). Per quanto riguarda il valore della soglia di attivazione del neurone è stato scelto di impostare un valore pari a quello utilizzato con la regressione logistica (0,123821) al fine di avere un confronto diretto con i modelli sviluppati precedentemente. La funzione di attivazione scelta per i neuroni della rete è la funzione di Sigmoid, per la sua ottima adattabilità ai problemi di classificazione binaria.

Alla seguente figura 6.11 è possibile osservare un report dei modelli più significativi sviluppati in riferimento alla colonna flag "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)". Come si può osservare per ogni riga sono stati specificati i neuroni di input, l'architettura degli strati nascosti e le metriche di performance per ogni dataset. Non è stata invece specificata la numerosità dei neuroni nello strato di output delle reti poiché quest'ultimo è sempre stato composto da un solo neurone essendo l'obiettivo della rete una classificazione binaria anomala/sana.

Un'altra osservazione che può essere fatta sulla figura 6.11 è relativa alla costituzione di un ulteriore set rispetto al training set e al test set utilizzati per i modelli basati sulla regressione logistica: il *validation set*. Come già dettagliato nel capitolo 4.6 in cui si è trattato l'algoritmo di error back-propagation, i dati appartenenti a tale set vengono forniti di volta in volta al modello durante la fase di training per elaborare una predizione e valutare l'eventuale esistenza di overfitting. Infatti, il motivo per cui si è deciso di non definire un validation set nei modelli basati sulla regressione logistica è che in questi ultimi non è previsto un training perché le predizioni sono fatte a partire dai valori dei coefficienti, i quali sono stati ottenuti a seguito del lancio dell'algoritmo di *Maximum Likelihood Estimation* (MLE). Nello specifico, è stato deciso di effettuare la seguente suddivisione:

- Training set pari all'80% del training set definito per il calcolo dei modelli basati sulle regressioni logistiche (80% di 2/3 delle osservazioni);
- Validation set pari al 20% del training set definito per il calcolo dei modelli basati sulle regressioni logistiche (20% di 2/3 delle osservazioni);
- Test set pari a 1/3 delle osservazioni.

Pertanto, quello che è stato deciso di fare è mantenere il medesimo test set utilizzato nei modelli basati sulla regressione logistica al fine di avere un confronto diretto delle performance delle due tipologie di modelli sulle medesime osservazioni, mentre è stato sacrificato il 20% dei dati del training set per avere la possibilità di valutare al termine di ogni epoch in fase di addestramento l'esistenza di un eventuale overfitting.

Nel successivo capitolo saranno descritte le porzioni di codice prodotte per sviluppare le varie architetture presentate alla figura 6.11. In particolare, verrà descritto il codice relativo al modello 1 e relativo alla colonna flag "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)". Come nel caso dei modelli basati sulla regressione logistica, dal momento che le differenze sia in termini di codice Python che di performance ottenute è minima, si presenterà il codice solo per la colonna "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)".

Reti neurali su colonna flag = 1+2 SOC

								Train		Validation			Test		
Num. Modello	Nome file	n. input	hidden layers	Learning rate	Epochs	Batch size	Threshold	AUC	Loss	AUC	Loss	Fitting	AUC	Loss	Note
1	M3_7_ANN_SOC_7IN	7	5 | 10 | 4	0,001	200	64	0,123821	0,8400	0.2777	0,8103	0.2730	good	0,8379	0,2763	
2	M4_7_ANN_SOC_7IN	7	6 | 10 | 5	0,01	200	64	0,123821	0,8427	0,2699	0,8192	0,2748	overfitting	0,8414	0,2734	Overfitting risolto nel modello successivo abbassando il learning rate
3	M5_7_ANN_SOC_7IN	7	6 | 10 | 5	0,001	200	64	0,123821	0,8304	0,2775	0,8094	0,2749	good	0,8333	0,2761	
4	M6_7_ANN_SOC_7IN	7	6 | 7 | 7 | 4	0,01	200	64	0,123821	0,8295	0,2778	0,8067	0,2748	good	0,8327	0,2769	
5	M7_7_ANN_SOC_7IN	7	6 | 8 | 8 | 4	0,001	200	64	0,123821	0,8255	0,2783	0,8029	0,2789	good	0,8240	0,2810	
6	M8_7_ANN_SOC_7IN	7	6 | 9 | 9 | 4	0,001	200	64	0,123821	0,8326	0,2769	0,8131	0,2729	good	0,8370	0,2759	
7	M9_7_ANN_SOC_7IN	7	6 | 10 | 10 | 4	0,001	200	64	0,123821	0,8269	0,2779	0,8071	0,2750	good	0,8283	0,2778	
8	M21_7_ANN_SOC_7IN	7	5 | 10 | 3	0,001	200	64	0,123821	0,8283	0,2787	0,8090	0,2780	good	0,8297	0,2796	
9	M22_7_ANN_SOC_7IN	7	6 | 10 | 3	0,001	200	64	0,123821	0,8266	0,2792	0,8131	0,2717	good	0,8321	0,2777	
10	M10_8_ANN_SOC_8IN	8	5 | 10 | 4	0,001	200	64	0,123821	0,8288	0,2757	0,8108	0,2719	good	0,8342	0,2763	
11	M11_8_ANN_SOC_8IN	8	6 | 10 | 5	0,001	200	64	0,123821	0,8268	0,2781	0,8066	0,2729	good	0,8307	0,2762	
12	M12_8_ANN_SOC_8IN	8	6 | 7 | 7 | 4	0,001	200	64	0,123821	0,8289	0,2760	0,8165	0,2701	good	0,8327	0,2759	
13	M13_8_ANN_SOC_8IN	8	6 | 8 | 8 | 4	0,001	200	64	0,123821	0,8261	0,2778	0,8028	0,2768	good	0,8268	0,2789	
14	M14_8_ANN_SOC_8IN	8	6 | 9 | 9 | 4	0,001	200	64	0,123821	0,8298	0,2771	0,8070	0,2760	good	0,8352	0,2771	
15	M15_8_ANN_SOC_8IN	8	6 | 10 | 10 | 4	0,001	200	64	0,123821	0,8295	0,2769	0,8096	0,2760	good	0,8300	0,2803	
16	M16_ANN_SOC_5IN	5	3 | 8 | 2	0,001	200	64	0,123821	0,8175	0,2858	0,7945	0,2804	good	0,8223	0,2823	
17	M17_ANN_SOC_5IN	5	3 | 10 | 2	0,001	200	64	0,123821	0,8177	0,2888	0,8056	0,2778	good	0,8243	0,2827	
18	M18_ANN_SOC_5IN	5	3 | 8 | 8 | 3	0,001	200	64	0,123821	0,8224	0,2816	0,8104	0,2713	good	0,8282	0,2789	
19	M19_ANN_SOC_5IN	5	3 | 8 | 8 | 2	0,001	200	64	0,123821	0,8211	0,2826	0,8074	0,2734	good	0,8289	0,2787	
20	M20_ANN_SOC_5IN	5	3 | 10 | 10 | 2	0,001	200	64	0,123821	0,8162	0,2871	0,7964	0,2805	good	0,8229	0,2828	
21	M23_7_ANN_SOC_7IN	7	3 | 6 | 2	0,001	200	64	0,123821	0,8223	0,2811	0,8080	0,2719	good	0,8295	0,2779	

Reti neurali su colonna flag = 1+2 ANNO

								Train		Validation			Test		
Num. Modello	Nome file	n. input	Architettura hidden layers	Learning rate	Epochs	Batch size	Threshold	AUC	Loss	AUC	Loss	Fitting	AUC	Loss	Note
1	M3_7_ANN_ANNO_7IN	7	5 | 10 | 4	0,001	200	64	0,123821	0.8242	0.2799	0.8057	0.2771	good	0.8298	0.2783	
6	M8_7_ANN_ANNO_7IN	7	6 | 9 | 9 | 4	0,001	200	64	0,123821	0.8298	0.2764	0.8169	0.2704	good	0.8346	0.2747	
14	M14_8_ANN_ANNO_8IN	8	6 | 9 | 9 | 4	0,001	200	64	0,123821	0.8301	0.2759	0.8158	0.2709	good	0.8350	0.2754	
16	M16_ANN_ANNO_5IN	5	3 | 8 | 2	0,001	200	64	0,123821	0.8186	0.2863	0.8034	0.2763	good	0.8247	0.2810	
18	M18_ANN_ANNO_5IN	5	3 | 8 | 8 | 3	0,001	200	64	0,123821	0.8169	0.2860	0.7978	0.2802	good	0.8254	0.2812	
21	M23_7_ANN_ANNO_7IN	7	3 | 6 | 2	0,001	200	64	0,123821	0.8219	0.2820	0.8080	0.2720	good	0.8296	0.2780	

Figura 6.11: report dei modelli più significativi sviluppati con reti neurali relativamente alla colonna flag=1+2 SOC (in alto) e ANNO (in basso)

6.2.1 Descrizione del codice dei modelli basati sulle reti neurali

La prima operazione che è stata fatta è stata quella di importare le librerie fondamentali per poter gestire e analizzare i dati relativi agli indicatori scelti come input della rete:

Codice 6.8: importazione delle librerie necessarie

```
import pandas as pd
import tensorflow as tf
from matplotlib import pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
```

Successivamente sono stati importati il training set e il test set, due file salvati in formato .csv per agevolare l'importazione su Python. La definizione del validation set invece avverrà in seguito tramite la definizione di un parametro all'interno di una specifica funzione.

Codice 6.9: importazione dei dataset di training e di test

```
# Loading datasets
df_train = pd.read_csv("train_dataset_norm.csv", delimiter = ";")

df_test = pd.read_csv("test_dataset_norm.csv", delimiter = ";")
df_train.info()
```

Nel seguente blocco di codice sono stati definiti gli indicatori di input della rete e la label di riferimento, ossia la colonna flag "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)":

Codice 6.10: definizione delle variabili di input e della colonna flag di riferimento (flag=1+2 SOC)

```
n_feat = 7

feat_names = [
    'ROE',
    'UtileCorrente/ricavi',
    'DebitiTotali/EBITDA',
    'RiservePIUutile/AN',
    'LnRIC',
    'serviziEsterni/costiOperativi',
    'DebFinanziariStimati/Ric'
]

my_features = [0] * n_feat
j = 0
for i in feat_names:
    my_features[j] = np.array(df_train[i])
    j += 1

my_label = np.array(df_train['FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)'])
```

A questo punto si hanno tutti gli strumenti necessari per passare alla creazione dell'architettura della rete neurale. In particolare, nel seguente blocco di codice sono state definite due funzioni che permettono la creazione (*create_model*) della rete e il suo addestramento (*train_model*):

Codice 6.11: definizione delle funzioni di creazione e di addestramento della rete neurale tramite la Functional API di TensorFlow

```
1  def create_model(learning_rate, metrics, n_features):
2      # Defining the architecture
3      input1 = tf.keras.layers.Input(shape=(1,))
4      input2 = tf.keras.layers.Input(shape=(1,))
5      input3 = tf.keras.layers.Input(shape=(1,))
6      input4 = tf.keras.layers.Input(shape=(1,))
7      input5 = tf.keras.layers.Input(shape=(1,))
8      input6 = tf.keras.layers.Input(shape=(1,))
9      input7 = tf.keras.layers.Input(shape=(1,))
10
11     merged = tf.keras.layers.Concatenate(axis=1)([input1, input2, input3, input4, input5, input6, input7])
12
13     dense1 = tf.keras.layers.Dense(5, input_dim=n_features, activation=tf.keras.activations.sigmoid, name='Hidden1')(merged)
14     dense2 = tf.keras.layers.Dense(10, activation=tf.keras.activations.sigmoid, name='Hidden2')(dense1)
15     dense3 = tf.keras.layers.Dense(4, activation=tf.keras.activations.sigmoid, name='Hidden3')(dense2)
16
17     output = tf.keras.layers.Dense(1, activation=tf.keras.activations.sigmoid, name='Output')(dense3)
18
19     model = tf.keras.Model(inputs=[input1, input2, input3, input4, input5, input6, input7], outputs=output)
20
21     # Building the model
22     model.compile(optimizer=tf.keras.optimizers.RMSprop(learning_rate),
23                   loss=tf.keras.losses.BinaryCrossentropy(),
24                   metrics=metrics)
25     return model
26
27
28 def train_model(model, features, label, epochs, batch_size=None):
29     history = model.fit(
30         x=features,
31         y=label,
32         batch_size=batch_size,
33         epochs=epochs,
34         shuffle=True,
35         validation_split=0.2)
36
37     epochs = history.epoch
38
39     hist = pd.DataFrame(history.history)
40
41     return epochs, hist
```

Il funzionamento della funzione *create_model()* si fonda sulla *Functional API* di TensorFlow, ossia un pacchetto di soluzioni che permette la creazione di architetture anche molto complesse con la creazione di strati nascosti anche non necessariamente collegati del tutto allo strato precedente. Tale API, presente tra le funzionalità offerte da TensorFlow, ha permesso in particolare la gestione di reti con molteplici variabili in input.

La funzione *create_model* richiede in ingresso i seguenti tre parametri:

- learning rate;
- metrics: vettore definito nel seguito prima di richiamare le funzioni di creazione e allenamento della rete che raccoglie tutte le metriche che vogliono essere calcolate in quando il modello verrà fatto allenare;
- n_features: valore che specifica il numero di neuroni di input scelti per la rete e che è necessario passare quando si definisce il primo strato nascosto della rete (riga 13) in corrispondenza del parametro input_dim, con il quale viene specificato qual è la numerosità degli input.

In generale, il funzionamento della funzione in esame può essere schematizzato nelle seguenti componenti principali:

1. **Determinazione dei neuroni di input (righe 3-9)**: per ogni indicatore di bilancio identificato viene definita una specifica variabile *input(i)*, ossia un neurone di input della rete;
2. **Determinazione dello strato di input (riga 11)**: le variabili di input definite al punto precedente sono condensate in un unico strato di input che permetterà l'alimentazione della rete neurale;
3. **Determinazione degli strati nascosti e dello strato di output (righe 13-17)**: ogni strato nascosto, identificato dalla variabile *dense(i)*, è creato lanciando la funzione *tf.keras.layers.Dense*, la quale permette di settare diversi parametri per ogni strato come ad esempio il numero di neuroni nascosti (nel caso in figura 5, 10 e 4), la funzione di attivazione dei neuroni appartenenti allo specifico strato e l'eventuale nome dello strato nascosto. In particolare, l'indicazione delle variabili riportate fra partentesi al termine di ogni riga serve per indicare all'API a quale strato si desidera collegare lo strato nascosto che si sta definendo. Ad esempio, il primo strato nascosto (*dense1*) è collegato allo strato di input in cui sono stati condensati gli indicatori scelti (*merged*).

   Infine, con la medesima funzione si definisce lo strato di output, il quale è composto da un solo neurone essendo il problema in esame di classificazione binaria;
4. **Formalizzazione del modello (riga 19)**: il modello definito sin ora viene formalizzato all'interno di una specifica variabile al fine di poter effettuare nel seguito la compilazione dello stesso;
5. **Compilazione del modello (righe 22-24)**: tramite il metodo *.compile()*, il modello formalizzato nella variabile *model* viene compilato, ossia viene definito:
   a. L'algoritmo con il quale avverrà l'aggiornamento dei pesi della rete: in questo caso si è scelto l'*algoritmo RMSprop*[10] (Root Mean Squared Propagation), ossia un'evoluzione del comune algoritmo di Gradient Descent che permette un'accelerazione del processo di ottimizzazione della rete;
   b. La loss function: essendo il problema in esame un problema di classificazione binaria, si è deciso di adottare la Binary Cross Entropy.

[10] L'algoritmo è stato sviluppato da Geoffrey Hinton, professore presso la University of Toronto e conosciuto nell'ambito del Computer Science e Data Science per i suoi contributi nell'ambito delle reti neurali.

c. Le metriche che si desidera vengano calcolate: si tratta del gruppo di metriche passato come parametro nel blocco di codice descritto nel seguito (blocco di codice 6.12) e in cui si richiama la funzione *create_model()*.

Alla seguente figura 6.12 è possibile osservare l'output della funzione *tf.keras.utils.plot_model*, la quale permette di osservare la struttura della rete appena sviluppata.

Quando viene lanciata la funzione *create_model()*, il modello che viene fornito al termine dell'esecuzione (riga 25) viene memorizzato in una specifica variabile che viene poi passata come parametro della funzione di addestramento della rete *train_model()*. In particolare, tale funzione di addestramento richiede in ingresso i seguenti parametri:

- model: variabile in cui è stato salvato il modello creato e compilato tramite la funzione *create_model()*;
- features: vettore *my_features* definito nel blocco di codice riportato al blocco di codice 6.10 e contenente tutti gli indicatori in input del modello;
- label: vettore *my_label* definito nel blocco di codice riportato al blocco di codice 6.10 e contenente i dati relativi alla colonna "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)";
- epochs;
- batch_size.

L'addestramento della rete neurale avviene tramite l'esecuzione della funzione *.fit()*, la quale richiede la definizione dei seguenti parametri:

- x: viene passato il vettore *my_features*;
- y: viene passato il vettore *my_label*;
- batch_size
- epochs
- shuffle: parametro booleano che se impostato su "True" genera un mescolamento dei dati di input (*x*) coi quali si esegue l'addestramento del modello;
- validation split: valore compreso tra 0 e 1 con cui si definisce la porzione del training set che si desidera allocare al validation set.

Ad ogni iterazione dell'addestramento (epoch) le informazioni relative al valore della loss function e alle metriche specificate sono salvate all'interno della variabile *history*.

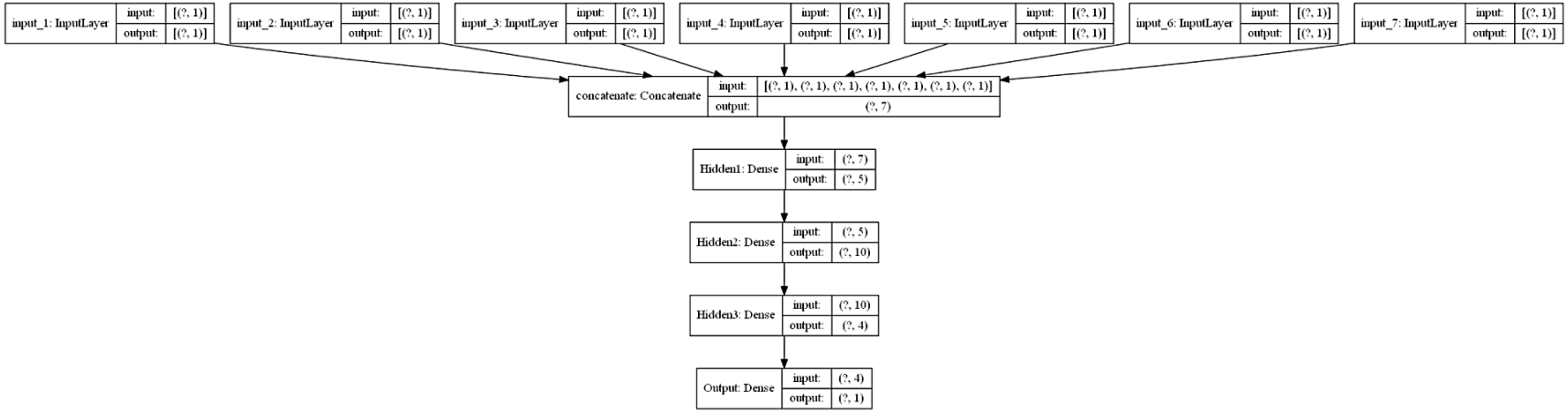


Figura 6.12: architettura della rete neurale definita al blocco di codice 6.11

A questo punto, definite le funzioni di creazione e addestramento della rete, è possibile eseguire il seguente blocco di codice con il quale si richiamano tali funzioni:

Codice 6.12: definizione degli iperparametri, delle metriche e richiamo delle funzioni di creazione e addestramento della rete

```
learning_rate = 0.01
epochs = 200
batch_size = 64
threshold = 0.123821397958389

MY_METRICS = [
    tf.keras.metrics.BinaryAccuracy(name='accuracy',
                                    threshold=threshold),
    tf.keras.metrics.Precision(thresholds=threshold,
                               name='precision'
                               ),
    tf.keras.metrics.Recall(thresholds=threshold,
                            name="recall"),
    tf.keras.metrics.AUC(num_thresholds=1000,
                         name='auc'),

    tf.keras.metrics.TrueNegatives(thresholds=threshold,
                                   name="tn"),
    tf.keras.metrics.TruePositives(thresholds=threshold,
                                   name="tp"),
    tf.keras.metrics.FalseNegatives(thresholds=threshold,
                                    name="fn"),
    tf.keras.metrics.FalsePositives(thresholds=threshold,
                                    name="fp"),
]

my_model = create_model(learning_rate, MY_METRICS, n_feat)

epochs, hist = train_model(my_model, my_features, my_label, epochs, batch_size)
```

```
Epoch 1/200
161/161 [==============================] - 1s 8ms/step - loss: 0.3868 - accuracy: 0.3765 - precision: 0.1553 - recall: 0.8944 -
auc: 0.6557 - tn: 2713.0000 - tp: 1152.0000 - fn: 136.0000 - fp: 6265.0000 - val_loss: 0.3020 - val_accuracy: 0.8387 - val_prec
ision: 0.3835 - val_recall: 0.6179 - val_auc: 0.7968 - val_tn: 1967.0000 - val_tp: 186.0000 - val_fn: 115.0000 - val_fp: 299.00
00
Epoch 2/200
161/161 [==============================] - 0s 3ms/step - loss: 0.3015 - accuracy: 0.8395 - precision: 0.4091 - recall: 0.6289 -
auc: 0.7807 - tn: 7808.0000 - tp: 810.0000 - fn: 478.0000 - fp: 1170.0000 - val_loss: 0.2866 - val_accuracy: 0.8434 - val_preci
sion: 0.3932 - val_recall: 0.6179 - val_auc: 0.7988 - val_tn: 1979.0000 - val_tp: 186.0000 - val_fn: 115.0000 - val_fp: 287.000
0
Epoch 3/200
161/161 [==============================] - 0s 3ms/step - loss: 0.2967 - accuracy: 0.8385 - precision: 0.4071 - recall: 0.6297 -
auc: 0.7917 - tn: 7797.0000 - tp: 811.0000 - fn: 477.0000 - fp: 1181.0000 - val_loss: 0.2851 - val_accuracy: 0.8344 - val_preci
sion: 0.3740 - val_recall: 0.6113 - val_auc: 0.8010 - val_tn: 1958.0000 - val_tp: 184.0000 - val_fn: 117.0000 - val_fp: 308.000
```

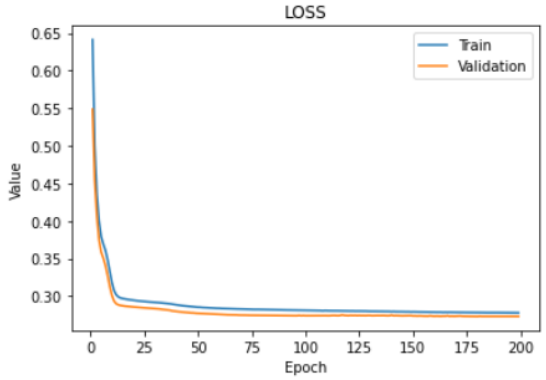
Come si può osservare dalla figura, al lancio della funzione di addestramento vengono generati dei blocchi di testo al termine di ogni epoch e in cui è possibile consultare tutti i valori memorizzati all'interno della variabile *history* e determinati al temine del calcolo di uno specifico ciclo.

Al termine del processo di addestramento, il quale ha richiesto per tutti i modelli sviluppati circa 2-3 minuti per essere completato, sono stati plottati i grafici relativi alla loss function e all'AUC tramite l'esecuzione dei seguenti blocchi di codice:

Codice 6.13: plot dei valori della training loss e della validation loss

```
# Plot train and validation loss
plt.figure()
plt.title('LOSS')
plt.xlabel("Epoch")
plt.ylabel("Value")

x1 = hist['loss']
x2 = hist['val_loss']
plt.plot(epochs[1:], x1[1:], label='Train')
plt.plot(epochs[1:], x2[1:], label='Validation')

plt.legend()

#hist
```

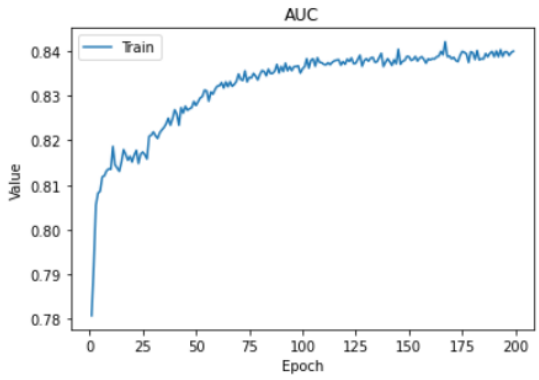
```
<matplotlib.legend.Legend at 0x1410ee609d0>
```



Codice 6.14: plot dei valori dell'AUC

```
# PLot train and validation AUC
plt.figure()
plt.title('AUC')
plt.xlabel("Epoch")
plt.ylabel("Value")

x1 = hist['auc']
#x2 = hist['val_auc']
plt.plot(epochs[1:], x1[1:], label='Train')
#plt.plot(epochs[1:], x2[1:], label='Validation')

plt.legend()
```

```
<matplotlib.legend.Legend at 0x293a43a3610>
```



Al blocco di codice 6.13 si possono osservare gli andamenti della loss function dichiarata precedentemente in fase di compilazione del modello, ossia la *binary cross entropy*, sul training set e sul validation set. Durante la fase di apprendimento della rete l'obiettivo che si cerca di perseguire è duplice: in primo luogo si cerca di ottenere il massimo decremento della loss function, ossia le migliori capacità discriminatorie minimizzando gli errori di classificazione; in secondo luogo si cerca di perseguire un buon livello di fit tra le loss function dei due set per valutare le capacità predittive del modello su dati mai analizzati prima durante l'apprendimento, capacità che al trascorrere delle epochs dovrebbero aumentare e far assumere un andamento della loss function del validation set simile a quella del training set e quindi determinare una certa stabilità predittiva. In particolare, si possono distinguere tre casistiche principali:

- **Underfitting**: termine con il quale ci si riferisce ad una situazione in cui il modello presenta delle difficoltà nell'apprendere le informazioni intrinseche nel train set tra i dati in input e quelli in output, determinando quindi una situazione in cui le capacità predittive producono una loss function troppo elevata. Un esempio ti tale situazione è riportato alla seguente figura:

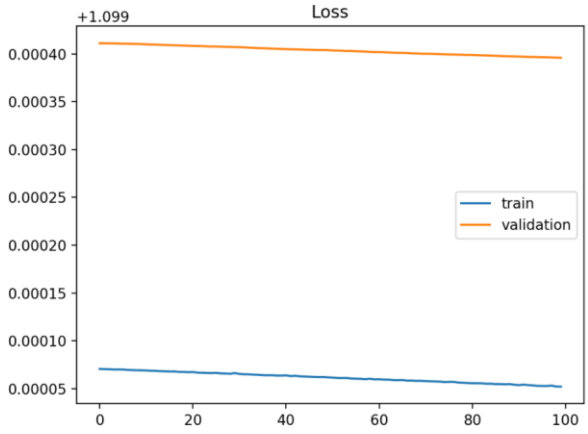


Figura 6.13: esempio di training loss e validation loss in caso di underfitting (1)

Inoltre, una seconda casistica in cui si rileva underfitting è quella riportata alla seguente figura in cui la training loss è decrescente e al termine dell'ultima epoch continua ad assumere un andamento decrescente, facendo pensare che il training sia stato troncato troppo presto e che il modello abbia ancora margini di miglioramento.

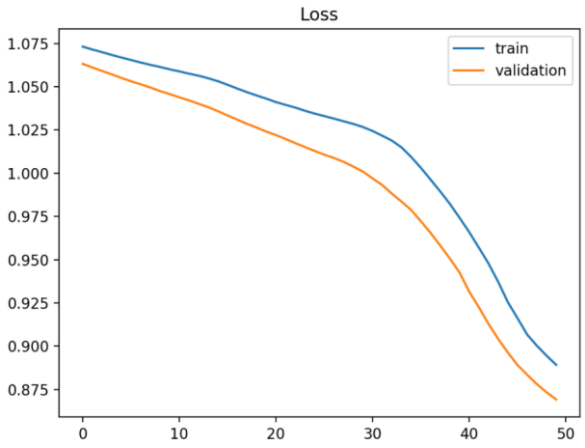


Figura 6.14: esempio di training loss e validation loss in caso di underfitting (2)

- **Overfitting**: termine con il quale si indica una situazione in cui il modello ha appreso eccessivamente bene le relazioni tra input e output della rete neurale tale da non permettergli di generalizzare adeguatamente nuovi valori che vengono forniti dal validaiton set. Una situazione grafica caratteristica di questo comportamento è riportata alla seguente figura:

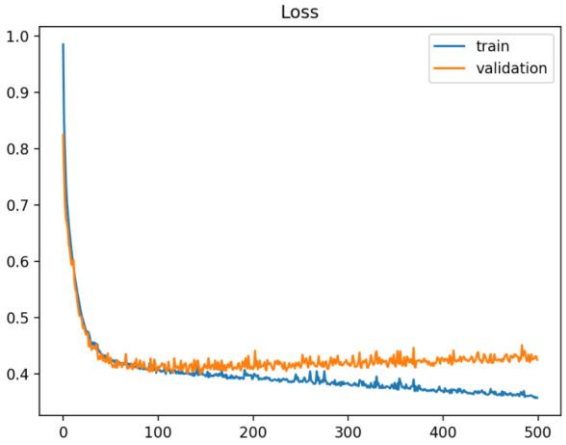


Figura 6.15: esempio di training loss e validation loss in caso di overfitting

- **Good fit**: costituisce l'obiettivo dell'algoritmo di apprendimento ed è approssimativamente una situazione a metà tra l'overfitting e l'underfitting. In generale, il grafico delle loss function può dire che presenti un good fit quando:
    a. Il grafico della training loss decresce fino ad arrivare in un punto di stabilità in cui sono assenti grandi variazioni;
    b. Il grafico della validation loss decresce fino ad un punto di stabilità in cui sono assenti grandi variazioni e presenta un piccolo gap con il grafico della trining loss.

  Un esempio di tale situazione è riportato alla seguente figura, oppure nel grafico della training loss riportato precedentemente e relativa al modello sviluppato.

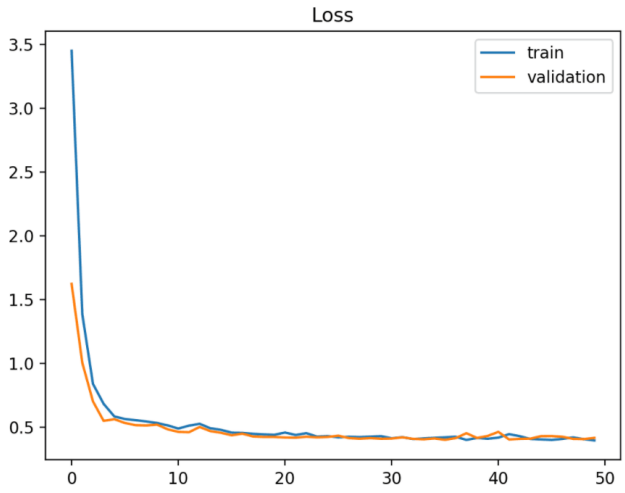


Figura 6.16: esempio di training loss e validation loss in caso di good fit

Per quanto riguarda invece l'andamento dell'AUC, questo risulta crescente fino a circa l'epoch 150 per poi stabilizzarsi intorno a valori poco inferiori a 0,84 fino all'epoch 200 senza mostrare particolari segni di potenziali miglioramenti.

Una volta ottenute delle informazioni relativamente alla qualità del modello sviluppato, si è passati al test della rete neurale costruita sul training set. Il seguente blocco di codice invoca la funzione *.evaluate()*, la quale restituisce come output un blocco di testo contenente le metriche indicate precedentemente in fase di lancio dell'addestramento della rete ma calcolate sul test set:

Codice 6.15: esecuzione della funzione .evaluate()

```
test_features = [0] * n_feat
j = 0
for i in feat_names:
    test_features[j] = np.array(df_test[i])
    j += 1

test_label = np.array(df_test['FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)'])

test_output = my_model.evaluate(x = test_features, y = test_label)
```

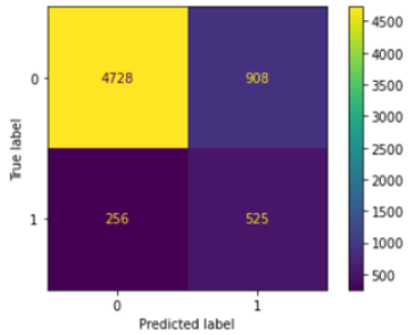
```
201/201 [==============================] - 0s 1ms/step - loss: 0.2766 - accuracy: 0.8220 - precision: 0.3728 - recall: 0.6773 - auc: 0.8363 - tn: 4746.0000 - tp: 529.0000 - fn: 252.0000 - fp: 890.0000
```

Infine, per avere una visualizzazione grafica delle performance ottenute dal modello in uno specifico contesto operativo, e quindi tenendo in considerazione la soglia di classificazione scelta, come per i modelli basati sulla regressione logistica è stata plottata una confusion matrix avvalendosi della medesima libreria utilizzata in precedenza:

Codice 6.16: calcolo della confusion matrix

```
n = len(df_test['FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)'])

a_cnt = 0
for i in range(n):
    if df_test['FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)'][i] == 1:
        a_cnt += 1

y_true, y_pred = [0] * n, [0] * n

for i in range(a_cnt):
    y_true[i] = 1

for i in range(int(test_output[6])):
    y_pred[i] = 1

for i in range(a_cnt, a_cnt + int(test_output[8])):
    y_pred[i] = 1

cm = confusion_matrix(y_true, y_pred)
cm_disp = ConfusionMatrixDisplay(confusion_matrix=cm)
cm_disp.plot()
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1410f634d30>
```



6.2.2 Commenti sulle reti neurali sviluppate

I dati presentati alla precedente tabella 6.11 e relativi a tentativi di architetture sono stati generati a partire dal codice descritto nel capitolo precedente effettuando limitate variazioni al codice. Infatti, quello che si è cercato di fare sin da subito è di creare un codice flessibile e facilmente riadattabile al fine di perdere meno tempo possibile nella modifica del codice e concentrarsi maggiormente sull'evoluzione delle architetture.

Una caratteristica che accomuna tutte le reti neurali costruite è la scelta della numerosità dei neuroni presenti negli strati nascosti rispetto al numero di quelli presenti negli strati di input. Infatti, al fine di sfruttare al massimo la possibilità di avere pressoché infiniti gradi di libertà nella determinazione della struttura delle reti, si è deciso di:

1. Definire un primo strato nascosto con un numero di neuroni leggermente inferiore rispetto ai neuroni dello strato di input al fine di filtrare i dati passati e catturare le informazioni di fondo rilevanti;

2. Definire uno strato nascosto centrale con un numero di neuroni maggiore dei neuroni presenti nello strato di input al fine di sfruttare al massimo le capacità di calcolo messe a disposizione dalle reti neurali, inserendo quindi complessità nel modello e rielaborando le informazioni processate nello strato precedente di filtro. Una variante dello strato nascosto intermedio appena descritto e che è stata applicata per alcuni modelli si basa sulla presenza di due strati nascosti centrali con entrambi un numero di neuroni fra loro uguale e pari o superiore ai neuroni presenti nello strato di input. Anche in questo caso lo scopo di questa architettura è quello di aggiungere complessità al modello al fine di rielaborare le informazioni per determinare dei pattern non lineari tra le variabili del modello;
3. Definire un terzo strato nascosto con un numero di neuroni inferiore rispetto a quelli presenti nel primo strato nascosto al fine di poter filtrare e pulire i dati passati dallo strato nascosto intermedio prima del processamento di questi dall'ultimo neurone presente nello strato di output.

Così come successo nel caso dei modelli relativi alle regressioni logistiche, le differenze in termini di performance sono minime fra i modelli sviluppati prendendo come riferimento la colonna "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)" e quelli che prendono come riferimento la colonna "FLAG soc sana/anom+ in liquidazione - ANNO (flag=1+2)". Tuttavia, per scrupolo sono stati ugualmente calcolati i modelli anche per questa casistica ma solamente per i test più significativi.

Tutte le architetture sviluppate con le reti neurali hanno fornito dei risultati in termini di valore dell'AUC superiori ai modelli sviluppati sulle regressioni logistiche. Pertanto, da un punto di vista delle performance globali si può affermare che lo sviluppo di modelli basati sulle reti neurali ha fornito migliori capacità diagnostiche rispetto ai modelli che hanno fondamenta statistiche e che si basano sulla regressione logistica. In particolare, i modelli migliori che sono stati rilevati sulla base dell'andamento delle loss function sul training set e sul validation set e del valore dell'AUC sono i seguenti:

- Modello 1;
- Modello 6;
- Modello 14;
- Modello 21.

Inoltre, oltre ai modelli sviluppati secondo le logiche specificate all'inizio del capitolo corrente è stata sviluppata una seconda classe di modelli basati sulle reti neurali per

entrambe le colonne flag (SOC e ANNO) che hanno utilizzato come neuroni nello strato di input gli stessi indicatori utilizzati nei modelli logistici individuati come migliori (modello 7 per entrambe le colonne flag SOC e ANNO). La scelta di sviluppare questa seconda classe di modelli risiede nella possibilità di valutare un confronto ancora più diretto fra le performance delle due tipologie di modelli di credit scoring dal momento che sono state sviluppate con le medesime variabili. In questo caso, i modelli migliori sono risultati essere.

- Modello 16;
- Modello 18.

Di seguito sono riportate le confusion matrix dei modelli appena citati relativamente alla colonna "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)" unitamente ai valori delle cost function che sono state calcolate utilizzando i medesimi pesi impiegati per le regressioni logistiche (agli errori *false negative* è stato attribuito un peso di 20 mentre agli errori *false positive* un peso pari a 1). Anche in questo caso lo scopo della definizione delle funzioni di costo è quello di agevolare il confronto tra i modelli dal momento che una comparazione sulla base dell'AUC risulta proibitiva essendo i valori molto simili fra loro.

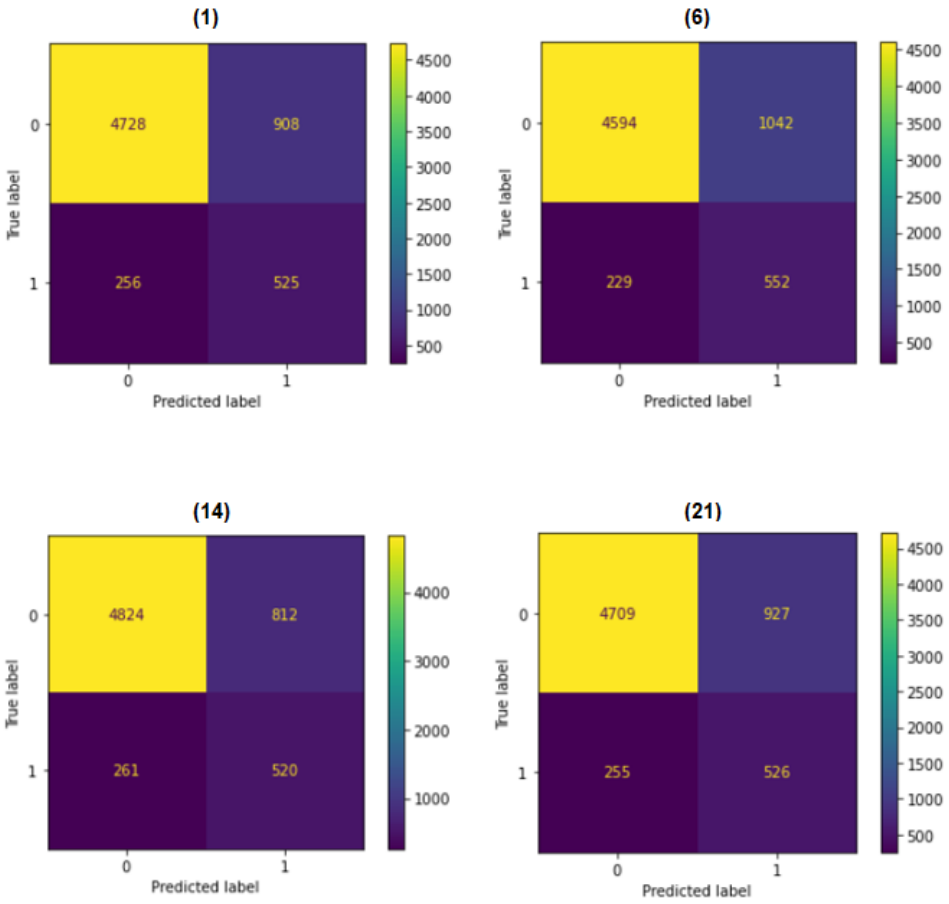


Figura 6.17: confusion matrix dei quattro modelli più significativi relativi alla colonna flag=1+2 SOC

Le corrispondenti cost function sono le seguenti:

- Modello 1: $908 \cdot 1 + 256 \cdot 20 = 6.028$
- Modello 6: $1042 \cdot 1 + 229 \cdot 20 = 5.622$
- Modello 14: $812 \cdot 1 + 261 \cdot 20 = 6.032$
- Modello 21: $927 \cdot 1 + 255 \cdot 20 = 6.027$

Pertanto, il modello migliore risulta essere il modello numero 6 dal momento che presenta la funzione di costo inferiore.

Inoltre, le confusion matrix relative alle reti neurali sviluppate a partire dagli stessi indicatori di input dei modelli di regressione logistica sono le seguenti:

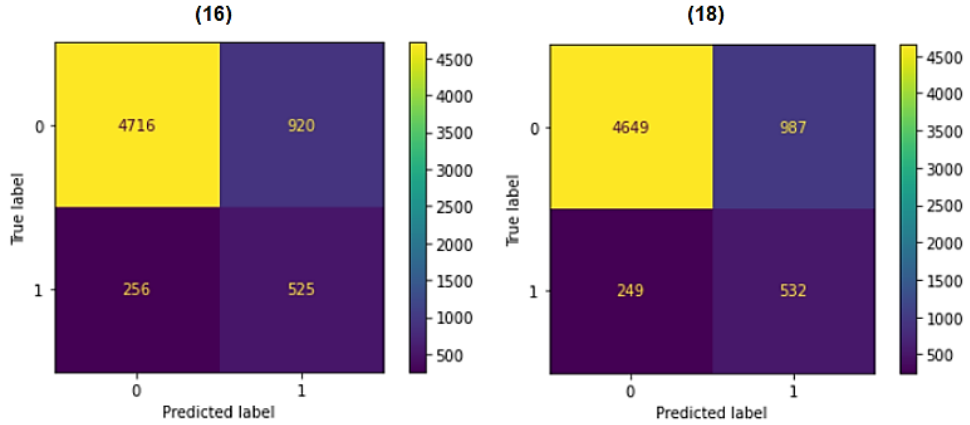


Figura 6.18: confusion matrix dei due modelli più significativi relativi alla colonna flag=1+2 SOC aventi come variabili di input gli stessi regressori del modello di riferimento sviluppato con la regressione logistica

Le relative cost functions sono:

- Modello 16: $920 \cdot 1 + 256 \cdot 20 = 6.040$
- Modello 18: $987 \cdot 1 + 249 \cdot 20 = 5.967$

Pertanto, il modello migliore fra quelli calcolati per questa seconda classe di reti neurali risulta essere il modello 18. Tuttavia, a livello generale è preferibile il modello numero 6 dal momento che presenta una cost function nettamente inferiore a quella del modello 18, un risultato probabilmente dovuto al fatto che la maggior complessità del modello 6 permette alla rete neurale di cogliere maggiori relazioni intrinseche fra i dati in input e quelli in output, garantendo in questo modo capacità diagnostiche superiori.

Di seguito sono riportati i medesimi calcoli effettuati per i modelli sviluppati in riferimento alla colonna flag “FLAG soc sana/anom+ in liquidazione - ANNO (flag=1+2)”, i quali hanno portato ai medesimi risultati ottenuti con gli sviluppi descritti finora.

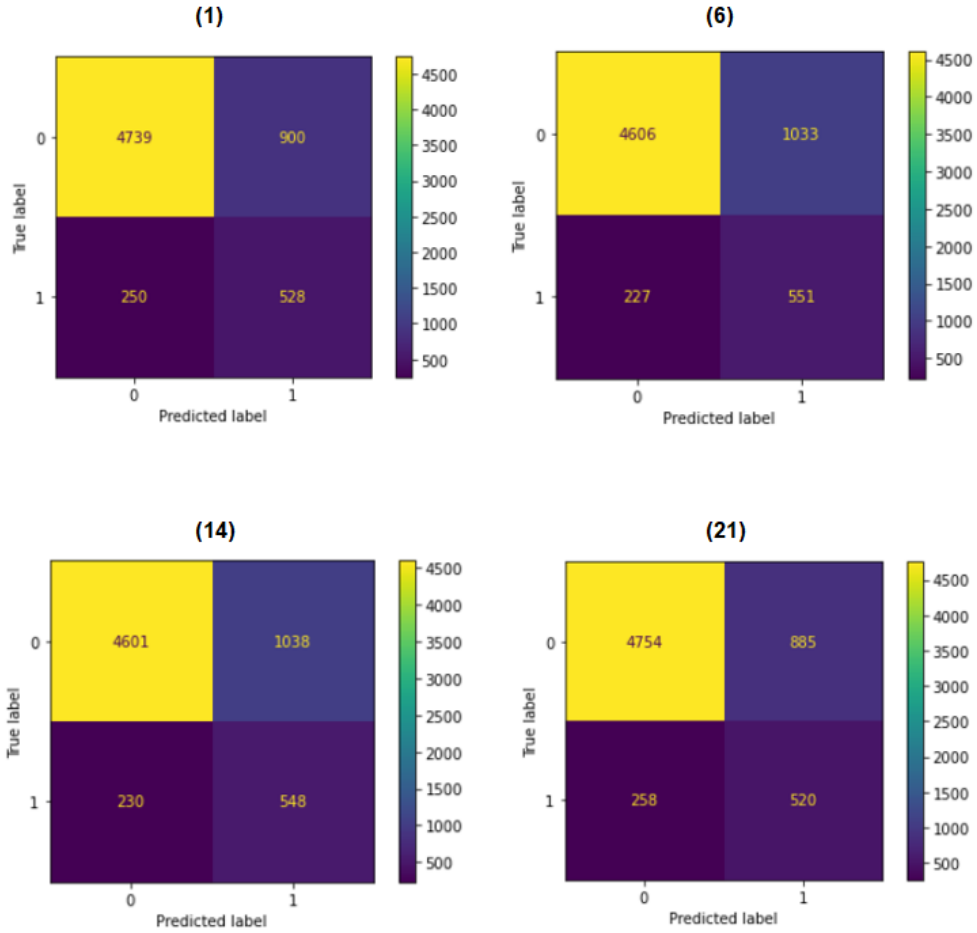


Figura 6.19: confusion matrix dei quattro modelli più significativi relativi alla colonna flag=1+2 ANNO

Di seguito le relative funzioni di costo:

- Modello 1: $908 \cdot 1 + 250 \cdot 20 = 5.908$
- Modello 6: $1033 \cdot 1 + 227 \cdot 20 = 5.573$
- Modello 14: $1038 \cdot 1 + 230 \cdot 20 = 5.638$
- Modello 21: $885 \cdot 1 + 258 \cdot 20 = 6.045$

Anche in questo caso il modello migliore risulta essere il numero 6.

Vengono riportate di seguito le confusion matrix relative ai modelli migliori fra quelli sviluppati partendo dai medesimi indicatori di input dei modelli di riferimento relativi alla regressione logistica.

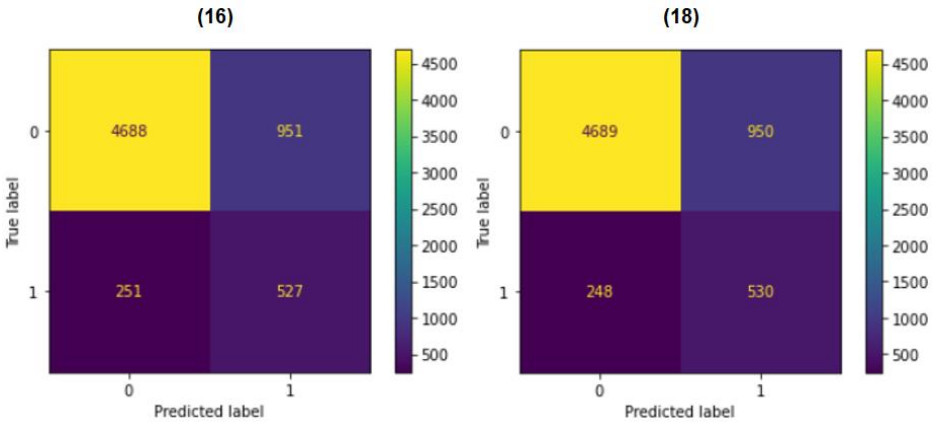


Figura 6.18: confusion matrix dei due modelli più significativi relativi alla colonna flag=1+2 ANNO aventi come variabili di input gli stessi regressori del modello di riferimento sviluppato con la regressione logistica

Le relative funzioni di costo sono:

- Modello 16: $951 \cdot 1 + 251 \cdot 20 = 5.971$
- Modello 18: $950 \cdot 1 + 248 \cdot 20 = 5.910$

Pertanto, anche in questo caso il modello migliore risulta il 18, tuttavia a livello globale, come verificato nel caso dei modelli sviluppati con la colonna flag "FLAG soc sana/anom+ in liquidazione - SOC (flag=1+2)", il modello globalmente preferibile risulta essere il numero 6 dal momento che presenta una cost function inferiore a quella del modello 18.

In conclusione, per entrambe le classi di modelli risulta migliore il modello 6 per via del minor valore assunto dalla cost function dal momento che il valore della metrica AUC con cui viene misurata la performance globale risulta essere molto simile e, quindi, poco discriminante fra un modello e l'altro.

6.3 Confronto dei risultati

Terminati gli sviluppi di entrambe le classi di modelli e determinati i migliori per ciascuna classe, è possibile effettuare un confronto diretto fra questi per determinare quale tipologia di modello di credit scoring sia più performante al fine di diagnosticare potenziali situazioni di default per le controparti di un istituto di credito.

Sia a livello di performance globale (AUC), sia a livello di performance di classificazione impostando un determinato valore di cut-off, i modelli sviluppati con le reti neurali per entrambe le colonne flag hanno sempre presentato delle capacità diagnostiche superiori a quelle espresse dai modelli statistici basati sulle regressioni logistiche. In particolare, la classificazione realizzata con la soglia pari a 0,123821 nel caso delle reti neurali ha

permesso di ottenere una cost function inferiore rispetto a quella calcolata sui modelli statistici per entrambe le colonne flag, come riassunto alla seguente tabella 6.3:

	Modello 6 ANN	Modello 7 RL	Differenza
SOC	5.622	6.561	-939
ANNO	5.573	6.498	-925

Tabella 6.3: confronto dei valori delle cost function per i modelli di riferimento delle regressioni logistiche (modello 7) e delle reti neurali (modello 6) per entrambe le colonne flag=1+2

Tuttavia, da un punto di vista delle performance globali c'è da dire che una volta giunti al termine degli sviluppi dei modelli statistici e prima di cimentarsi nella costruzione delle reti neurali, l'aspettativa era quella di ottenere delle performance di molto superiori rispetto a quelle ottenute con le regressioni logistiche, mentre quello che si è ottenuto sono valori di AUC di non molto superiori: per le regressioni logistiche i valori di AUC erano inclusi approssimativamente fra l'intervallo 0,70-0,74; per le reti neurali si sono rilevati valori di AUC compresi approssimativamente nell'intervallo 0,82-0,84. Pertanto, i pressoché infiniti gradi di libertà offerti dalle reti neurali e la possibilità di determinare architetture anche molto complesse per estrarre il più possibile le informazioni intrinseche che relazionano i dati in input con quelli in output hanno determinato un incremento delle performance globale limitata a circa 0,10 punti. Ci si è quindi interrogati sul perché di tale risultato e le spiegazioni che si sono trovate sono le seguenti:

- **Scarsa qualità dei dati grezzi**: infatti, come descritto nel capitolo 5.2.1, a seguito dell'estrazione da Aida dei dati di bilancio si sono rese necessarie numerose operazioni di pulizia e correzione dei dati. Questo fa pensare che sulla totalità delle informazioni scaricate ci possano essere ulteriori errori, non rilevabili tramite le colonne di controllo, in grado di compromettere la fase di addestramento delle reti neurali;
- **Varietà dei dati limitata**: seppur i record scaricati abbiano una certa numerosità (19,250 records), è anche vero che tali osservazioni riguardano un numero ristretto di imprese (2.613 imprese). Pertanto, ne deriva che la varietà dei dati dai quali le reti neurali possono apprendere informazioni è limitata e questo comporta una limitata capacità di ottenere performance superiori a quelle espresse;
- **Scarsa numerosità di imprese anomale**: un ulteriore elemento che può aver compromesso un apprendimento ottimale è dato dal numero limitato di imprese anomale secondo la classificazione "flag=1+2" presenti nel campione. Infatti, minori sono le imprese anomale, minori sono le possibilità per l'algoritmo di

apprendere le relazioni e i pattern fra le variabili in input e quelle di output che determinano gli eventi di default.

In conclusione, nonostante i possibili margini di miglioramento ottenibili tramite l'utilizzo di un database più vasto e qualitativamente superiore, le performance globali possono dirsi discrete e comunque nettamente superiori a quelle che presenterebbe un modello che effettuerebbe predizioni totalmente casuali.

7. Considerazioni finali

Il presente lavoro di tesi ha permesso di mettere in evidenza sia le potenzialità che le criticità di due metodologie di costruzione di modelli di credit scoring molto differenti fra loro, ossia le regressioni logistiche e le reti neurali. In conclusione, la risposta alla domanda che ha guidato lo sviluppo del presente elaborato, ossia quale metodologia offrisse migliori performance discriminatorie nell'ambito del credit scoring per il settore metallurgico italiano, risulta essere il modello di riferimento sviluppato con la rete neurale (modello 6) costituita da 7 neuroni in ingresso, quattro strati nascosti costituiti da 6,9,9,4 neuroni nascosti e 1 neurone nello strato finale di output per determinare un output binario.

Nel capitolo 6.3 è stato messo in evidenza come il modello di riferimento basato sulle reti neurali si sia rivelato migliore rispetto a quello sviluppato sulla regressione logistica, anche se si attendevano performance decisamente superiori a quelle effettivamente riscontrate. In particolare, sia durante la fase di preparazione dei dati antecedente agli sviluppi in cui si sono preprocessati i dati grezzi estratti dal database Aida per renderli fruibili e rielaborabili da Python, che durante quella di produzione vera e propria dei modelli tramite programmazione sul codice, si è acquisita sempre più consapevolezza degli strumenti che si stavano analizzando. Infatti, al termine della produzione del codice, studiando e analizzando i risultati ottenuti sono state fatte delle riflessioni in merito ai seguenti aspetti:

- Qualità dei dati di partenza;
- Semplicità di costruzione dei modelli;
- Capacità computazionale richiesta;
- Condivisione dei modelli;
- Identificazione e risoluzione degli errori;
- Gradi di libertà nella costruzione delle architetture.

Per quanto riguarda la **qualità dei dati di partenza**, nel capitolo 6.3 è stato messo in evidenza come uno dei motivi per cui i modelli basati sulle reti neurali non abbiamo ottenuto risultati superiori potrebbe essere l'incompletezza dei dati esportati dal database Aida. Infatti, si sono rese necessarie numerose operazioni di correzione dei valori delle poste di bilancio errate, come ad esempio le somme delle varie voci che costituiscono gli aggregati intermedi, e di sostituzione di valori *non machine-readable* in valori fruibili ed elaborabili dai modelli costruiti. Queste operazioni si sono rivelate indispensabili per proseguire con lo sviluppo dei modelli e hanno necessariamente compromesso la qualità dei dati che sono stati poi forniti in input ai modelli. Pertanto, si ha avuto modo di riflettere

come la qualità dei dati forniti in input a queste tipologie di modelli sia di fondamentale importanza per ottenere dei risultati consistenti, soprattutto per quanto riguarda aspetti delicati come il rischio di credito dove degli errori di valutazione potrebbero produrre conseguenze anche catastrofiche.

Un secondo aspetto che è stato oggetto di riflessione è la **semplicità di costruzione dei modelli**. Per costruire i modelli basati sulla regressione logistica è stato sufficiente specificare i regressori e richiamare la specifica funzione della libreria statsmodels per eseguire automaticamente il calcolo della funzione di massima verosimiglianza per definire i valori dei rispettivi coefficienti e dell'intercetta. Un procedimento simile viene seguito anche da altri software statistici come Stata. Pertanto, si può dire che i meccanismi sottostanti alla regressione logistica sono così conosciuti e matematicamente consistenti che definire un modello logistico a livello operativo è un'operazione piuttosto semplice. Al contrario, per definire una rete neurale non esistono funzioni che ottimizzano un'architettura o gli iperparametri della rete, quindi ogni elemento deve essere specificatamente dichiarato e modellato, come ad esempio il numero di neuroni in input e in output, il numero degli strati nascosti con i relativi neuroni, le funzioni di attivazione per i neuroni di uno specifico strato, i collegamenti tra gli strati, ecc. Pertanto, la costruzione di un modello basato sulle reti neurali risulta molto più dispendiosa a livello di tempistiche di produzione. Tuttavia, tale aspetto negativo può essere in parte mitigato avendo premura durante la prima produzione del codice di rendere agevoli le successive modifiche delle architetture e degli iperparametri, rendendo quindi necessarie solamente piccole modifiche in specifici punti per cambiare le strutture del modello. Questo accorgimento che è stato preso durante la produzione del codice ha infatti permesso di abbattere i tempi di rimodellamento delle reti.

Un altro aspetto più di carattere tecnico su cui si è riflettuto è la **capacità computazionale della macchina** su cui vengono eseguiti i modelli. Infatti, l'ottenimento dei risultati per il modello logistico è stato quasi istantaneo dopo il lancio della funzione di calcolo, anche grazie all'efficienza dell'algoritmo di massima verosimiglianza sulla base del quale si fondano i calcoli dei coefficienti e dell'intercetta. Invece, l'addestramento di ogni rete neurale ha richiesto circa 2-3 minuti di calcolo per 200 epochs. Pertanto, l'utilizzo delle reti neurali, soprattutto per l'esecuzione di modelli più complessi e con più dati da elaborare, richiede che tali modelli vengano lanciati su macchine dotate di componentistica hardware di un certo livello al fine di abbattere i tempi di calcolo, dal momento che nella

porzione di tempo in cui la macchina è in utilizzo essa risulta praticamente inutilizzabile se la sua capacità di calcolo è utilizzata al 100%.

Un altro aspetto rilevante su cui è stata fatta una riflessione riguarda l'eventuale **condivisione dei modelli** fra ricercatori. La regressione logistica si basa su forti fondamenta teoriche statistiche, pertanto le metodologie con cui si costruiscono e si analizzano i modelli sono comuni fra i ricercatori e sono basate su procedure universalmente conosciute, rendendo quindi la condivisione dei modelli estremamente agibile dal momento che la lettura dei modelli risulta immediata. Al contrario, non esistono fondamenta teoriche per la costruzione delle reti neurali e le procedure seguite per la loro costruzione sono quasi a completa discrezione del programmatore, pertanto l'eventuale condivisione dei modelli potrebbe non risultare agevole.

L'**identificazione degli errori o delle problematiche** costituisce un elemento molto differenziante fra le due tipologie di modelli sviluppati. Nella regressione logistica eventuali errori o problematiche, come ad esempio l'esistenza di un coefficiente non statisticamente significativo o il segno di un coefficiente non coerente col suo significato economico, sono immediatamente identificabili. Pertanto, in questo caso si riesce ad analizzare il problema in modo più strutturato avendo la possibilità di risalire in modo preciso alla fonte del problema, costruendo quindi delle ipotesi e provando ad implementare delle soluzioni. Al contrario, nelle reti neurali non si ha pieno controllo delle elaborazioni negli strati nascosti, ed è infatti per questa ragione che gli strati intermedi delle reti vengono considerate come delle *black box*. Ne deriva che eventuali problematiche o le cause di alcune inefficienze non siano immediatamente identificabili, rendendo quindi necessario la maggior parte delle volte l'attuazione di un procedimento per tentativi e fondato sull'intuizione per risolverle, modificando l'architettura delle reti o gli iperparametri settati e cercando in questo modo di risolvere la problematica riscontrata.

Infine, un ultimo aspetto su cui è stata posta l'attenzione riguarda la **libertà nella definizione delle architetture**. La costruzione dei modelli basati sulla regressione logistica è dettata da rigide condizioni statistiche che devono necessariamente essere soddisfatte affinché il modello possa considerarsi consistente. In particolare, ogni modello deve necessariamente presentare, pena il suo rifiuto, tutti i coefficienti:

1. Statisticamente significativi;
2. Con un segno coerente con il significato economico della variabile del modello a cui sono associati.

In particolare, se anche un solo coefficiente non rispetta tali condizioni, allora l'intero modello deve essere necessariamente rigettato. Ne deriva che le possibilità di scegliere la struttura del modello è piuttosto limitata e circoscritta da forti regole statistiche. Al contrario, le reti neurali presentano il grandissimo vantaggio di avere infiniti gradi di libertà nella definizione delle architetture delle reti, tanto che la modellazione è totalmente a discrezione del programmatore/analista e senza vincoli di alcun tipo. Tale libertà permette infatti la costruzione di reti con gradi di complessità molto elevati, con la conseguente possibilità di effettuare rielaborazioni profonde dei dati, cogliendo fenomeni di non linearità tra le variabili inserite nel modello e facendo emergere relazioni intrinseche fra di esse che difficilmente sarebbero individuabili con le regressioni logistiche.

Dall'analisi dei modelli sviluppati quello che si può osservare è che le reti neurali, essendo uno strumento molto sofisticato, si rivelano generalmente una soluzione più complessa da sviluppare e adattare a contesti operativi rispetto alle regressioni logistiche. Tuttavia, gli svantaggi riscontrati possono considerarsi poco rilevanti e, comunque, i vantaggi ottenibili in termini di complessità computazionale e libertà di definizione delle architetture che si ha avuto modo di osservare hanno permesso il raggiungimento di performance superiori rispetto ai modelli basati sulle regressioni logistiche. Infatti, quello che è emerso è che le reti neurali hanno sempre mostrato performance superiori sia in termini globali, tramite la lettura della metrica AUC, sia in specifici contesti operativi tramite i confronti messi a punto con la costruzione delle funzioni di costo a seguito della definizione di una specifica soglia di classificazione.

Bibliografia e sitografia

[1] Materiale didattico del professor Varetto Franco condiviso durante il corso "*Mercati, rischi e strumenti finanziari*" A.A. 2020/2021.

[2] Andrea Resti Andrea Sironi, *"Rischio e valore nelle banche. Misura, regolamentazione, gestione."*, EGEA, Milano, II edizione, 13 agosto 2008.

[3] Vincenzo Paolo Senese, *"Regressione Multipla e Regressione Logistica: concetti introduttivi ed esempi"*, I Edizione, ottobre 2016.

[4] Christine Bolton, *"Logistic regression and its application in credit scoring"*, University of Pretoria, 2009.

[5] Giacomo di Tollo, *"Reti neurali e rischio di credito: stato dell'arte e analisi sperimentale"*, 17 novembre 2005.

[6] Sito web Federacciai: http://federacciai.it/

[7] Sito web Bank of International Settlements: https://www.bis.org/

[8] Sito web Istat: https://www.istat.it/it/

[9] Marco Ferfoglia, *"BASILEA4: IL FRAMEWORK NORMATIVO"* https://www.riskcompliance.it/news/basilea4-il-framework-normativo/

[10] Sito web *"Google Machine Learning Crash Course"* https://developers.google.com/machine-learning/crash-course

[11] Andrea Provino, *"Precision and Recall con F1 Score | Precisione e Recupero"* https://andreaprovino.it/precision-and-recall-precisione-e-recupero/

[12] Sito web Bureau Van Dijk – Moody's Analytics Company: https://www.bvdinfo.com/it-it/

[13] Sito web TensorFlow: https://www.tensorflow.org/

[14] Sito web Statsmodels: https://www.statsmodels.org/stable/index.html

[15] Jason Brownlee, *"How to use Learning Curves to Diagnose Machine Learning Model Performance"* https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/