POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

Tesi di Laurea Magistrale

Valutazione del Rischio di Credito nel settore dei trasporti

Relatore
Prof. Franco Varetto

Candidato
Alberto Lupone

Anno Accademico 2020/2021

Sommario

Capitolo 1: Il rischio di credito 1

L'analisi delle perdite 3

Linee guida del Comitato di Basilea 11

Capitolo 2: Modelli di credit-scoring 17

Analisi discriminante lineare 18

Modelli di regressione 20

Regressione lineare 20

Regressione logistica 22

Reti neurali 27

Support Vector Machines 28

Separazione lineare 29

Separazione non lineare 32

Kernel 33

Modelli di portafoglio 35

CreditMetrics 35

Credit Risk Plus 37

Capitolo 3: Analisi del settore 39

I trasporti terrestri 45

I trasporti su strada 45

Rete ferroviaria 47

Trasporti marittimi 49

Trasporti aerei 51

Capitolo 4: Applicazione dei modelli di scoring 57

Applicazione del modello logistico 62

Logit Primo modello – Flag per società 63

Logit Secondo Modello – Flag per anno 67

Applicazione del modello Support Vector Machines 69

Selezione delle Feature 75

Random forest 75

Recursive feature elimination 76

SVM Primo modello - Flag per società 78

SVM Secondo modello - Flag per anno 80

Analisi delle performance e degli errori dei modelli 82

Conclusioni 89

Bibliografia 91

Sitografia 92

Capitolo 1: Il rischio di credito

Il rischio di credito è considerato il principale fattore alla base delle crisi finanziarie, pertanto negli ultimi anni è stato oggetto di analisi approfondite sia da parte di istituzioni creditizie, sia da parte delle autorità di vigilanza nazionali ed internazionali, soprattutto a seguito della crisi finanziaria globale del 2008, scaturita dello shock del mercato immobiliare statunitense.

Secondo il direttore della Swiss Institute of Banking and Finance, Ammann Manuel, il rischio di credito è "l'eventualità che una delle parti di un contratto non onori gli obblighi di natura finanziaria assunti, causando una perdita per la controparte creditrice".

In realtà, tale definizione non è esaustiva in quanto il rischio di credito non riguarda la sola possibilità di insolvenza, ma anche il deterioramento dell'affidabilità economica e finanziaria della controparte. Poiché il rischio di credito possa configurarsi è necessario che la variazione della posizione creditizia sia inattesa poiché è possibile stimare le variazioni attese e proteggersi ex-ante dalle stesse attraverso accantonamenti prudenziali.

Un'ulteriore problematica che contribuisce ad accrescere la difficoltà di argomentazione è legata alla molteplicità delle forme con cui il rischio di credito può presentarsi: gli strumenti finanziari soggetti a tale rischio sono principalmente titoli di debito e strumenti fuori bilancio. Tra i titoli di debito si classificano i titoli di Stato, i titoli di debito emessi da enti pubblici, le obbligazioni emesse da società private, mutui, finanziamenti e crediti al consumo. Tra le posizioni fuori bilancio rientrano i titoli derivati, tali posizioni riguardano spesso attività illiquide di cui non è possibile consultare i prezzi di quotazione e il cui valore può essere solo stimato.

Una valutazione corretta del rischio di credito è fondamentale al fine di una sana gestione di tutte le attività di un'impresa, ed è utile per individuare correttamente i parametri necessari a stimare il prezzo di obbligazioni e prestiti.

La gestione del rischio di credito è messa in atto mediante due differenti modalità che sono in grado di individuare le componenti del rischio: l'approccio model-based tra cui si distinguono i modelli strutturali e i modelli in forma ridotta, e l'approccio tradizionale basato sui dati storici delle insolvenze.

I modelli strutturali sono basati sullo studio dell'evoluzione dello stato patrimoniale, attraverso l'analisi e interpretazione di dati storici sulla struttura dell'attivo, al fine di determinare la probabilità di insolvenza e il tasso di recupero; i modelli in forma ridotta trattano invece l'insolvenza come un evento completamente esogeno e indipendente dalla struttura patrimoniale della società.

La distinzione tra i due diversi approcci non è così marcata ed è possibile utilizzare modelli ibridi che combinano insieme elementi di entrambi i modelli. L'elemento distintivo è dato dal default time, cioè dalla caratterizzazione del momento nel quale si manifesta l'insolvenza. Nei modelli strutturali si ipotizza di disporre delle stesse informazioni del manager, quindi il default è una variabile aleatoria prevedibile, mentre nei modelli in forma ridotta si prescinde dalla conoscenza delle attività e passività della società, assumendo che l'informazione disponibile sia la stessa conosciuta dal mercato.

In generale, è possibile fornire una definizione più esaustiva del rischio di credito solo attraverso la trattazione delle componenti che contribuiscono alla sua determinazione e ne costituiscono le cause, trattate nel dettaglio nei paragrafi seguenti.

L’analisi delle perdite

Ai fini dell’analisi del rischio di credito è molto interessante e strategicamente rilevante la gestione delle probabilità di perdita, in quanto questa non incide solo sul risultato economico d’esercizio, ma anche sulla stessa sopravvivenza dell’intermediario, intesa come permanenza sul mercato e mantenimento delle quote di mercato. Per poter analizzare nel dettaglio la probabilità di perdita è necessario procedere con una distinzione tra perdite attese e inattese:

- la *perdita attesa* si manifesta mediamente in un intervallo temporale di un anno su ogni esposizione, è una perdita che è possibile prevedere e stimare, pertanto solitamente non comporta ulteriori problematiche a livello gestionale;

- la *perdita inattesa* è definita come “la possibilità che una variazione inattesa del merito creditizio di una controparte generi una corrispondente variazione inattesa del valore di mercato della posizione creditoria”.[1]

- [1] *La gestione del rischio e allocazione del capitale nelle banche – Sironi A.*

Perdite attese

La perdita attesa rappresenta la perdita che un istituto di credito si aspetta di dover sostenere per l'esposizione creditizia nei confronti di un determinato portafoglio. A fronte delle perdite attese, banche e intermediari sono tenuti ad effettuare accantonamenti a conto economico, al fine di neutralizzare l'effetto di tali perdite. Questi accantonamenti dovranno essere almeno pari al valore della perdita attesa, poiché devono garantire all'intermediario la sopravvivenza anche in caso in cui tali eventi dovessero verificarsi.

Per determinare il valore monetario della perdita che si verificherebbe in caso di mancata riscossione di un credito, occorre analizzare il problema della variabilità delle perdite potenziali attorno al valore medio seguendo l'approccio in termini contabili e a valori di mercato.

Nell' approccio in termini *contabili*, la perdita attesa (EL) è il risultato del prodotto di tre componenti:

$$\mathbf{EL} = EAD \cdot PD \cdot (LGD\%) = EA \cdot PD \cdot [1 - E(RR)]$$

In cui EAD = esposizione attesa in caso di insolvenza;

PD= probabilità di insolvenza attesa;

LGD%= percentuale attesa di credito non recuperabile in caso di insolvenza;

E(RR) recovery rate atteso o tasso di recupero.

Nell'approccio alternativo, in cui si valuta il portafoglio a *valori di mercato*, il valore di ogni esposizione è ottenuto dal valore attuale dei flussi di cassa futuri attesi, scontati per la probabilità di insolvenza e il tasso recupero atteso.

Probabilità di default

Procedendo con l'analisi di ciascun fattore che influisce nel determinare il valore della perdita attesa, sicuramente è necessario studiare la **Probabilità di Default (PD)**. Questa identifica la probabilità che la controparte non sia in grado di restituire il capitale prestato o gli interessi maturati sullo stesso. La Probabilità di Default può essere determinata attraverso diverse metodologie, tra cui rientrano:

- il mercato dei capitali;
- l'utilizzo di modelli analitico-soggettivi che comprendono sia aspetti quantitativi, attraverso l'utilizzo di indici economico-finanziari, sia aspetti qualitativi, come prospettive evolutive del settore e qualità del management;
- rating creditizi che possono essere calcolati esternamente da agenzie specializzate, oppure internamente dall'istituto creditizio.

L'obiettivo dell'analisi è determinare il grado di financial risk, che può assumere una forma dicotomica (Default/Non-Default) oppure essere espresso attraverso classi omogenee alle quali è associata una denominazione alfabetica o numerica (Figura 1.1). Queste forme consentono la determinazione di condizioni di tasso che risultano strettamente collegate alle condizioni creditizie dell'impresa oggetto di valutazione.

L'assegnazione delle classi di rating è affidata ad agenzie esterne come Moody's, S&P, Fitch, oppure internamente dalla stessa banca, sfruttando informazioni sia qualitative sia quantitative quali le prospettive di guadagno future, i cash-flow, la struttura patrimoniale, il livello di liquidità e di indebitamento, il settore industriale e la qualità della classe dirigente.

La valutazione soggettiva dell'impresa consente di considerare adeguatamente sia variabili di natura quantitativa sia qualitativa, ma di contro può condurre a risultati anche molto diversi per analisi effettuate da soggetti differenti.

I modelli di natura statistica o di scoring invece forniscono una valutazione del merito creditizio attraverso l'analisi di diversi indici contabili, attribuendo ad ognuno una ponderazione mediante opportune tecniche statistiche. Questa tecnica garantisce

l'oggettività delle valutazioni e la possibilità di ottenere valutazioni consistenti. Le agenzie di rating hanno elaborato una scala di valori, assegnando per ciascun valore una probabilità di insolvenza:

Categoria	S&P	Moody's	Fitch	Probabilità default	Rischio
Investment grade (o categoria *investimento*)	AAA	Aaa	AAA	0,01%	Minimo
	AA+	Aa1	AA+	0,02%	Modesto
	AA	Aa2	AA	0,03%	
	AA-	Aa3	AA-	0,04%	
	A+	A1	A+	0,05%	Medio-basso
	A	A2	A	0,07%	
	A-	A3	A-	0,09%	
Investment grade inferiore	BBB+	Baa1	BBB+	0,13%	Accettabile
	BBB	Baa2	BBB	0,18%	
	BBB-	Baa3	BBB-	0,30%	
Non investment grade (o categoria *speculativa*)	BB+	Ba1	BB+	0,50%	Accettabile con attenzione
	BB	Ba2	BB	0,90%	
	BB-	Ba3	BB-	1,60%	
Non investment grade inferiore	B+	B1	B+	2,60%	Attenzione specifica con monitoraggio continuo
	B	B2	B	4,50%	
	B-	B3	B-	7,50%	
	CCC+	Caa1	CCC	13,00%	
	CCC	Caa2	CC	16,00%	Sotto stretta osservazione/esito dubbio
	CCC-	Caa3	C	20,00%	
	CC	Ca	DDD	26,00%	
	SD	C	DD	33,00%	
	D		D	——	

FIGURA 1.1-CREDIT RATING ESTERNO

Loss Given Default

La **Loss Given Default (LGD)** rappresenta la perdita subita a causa dell'insolvenza della controparte, è influenzata dalla qualità del management, dalle condizioni economico finanziarie dell'impresa debitrice e dalle prospettive di evoluzione del settore. Per stimare la LGD si possono utilizzare diverse tipologie di approcci:

- *market LGD* è praticabile solamente con riferimento a strumenti liquidi poiché si fonda sulla conoscenza di dati di mercato. L'approccio del *market LGD* si basa sulla stima del Recovery Rate attraverso l'utilizzo dei prezzi degli strumenti di debito quotati in default. Ci sono almeno due varianti di tale approccio:
    - *emergence LGD* che prevede la stima del Recovery Rate in base al prezzo di mercato (attualizzato al momento dell'insolvenza) dei nuovi strumenti finanziari offerti in sostituzione di quelli in default divenuti inesigibili;
    - *implied market LGD* che utilizza gli spread sui corporate bond non in default e la probabilità di default specifica del debitore per ricavare la LGD implicita. [2]
- l'approccio del *workout LGD* è applicabile ai prodotti illiquidi, prevede che l'intermediario disponga di uno storico che tenga traccia di tutte le caratteristiche delle esposizioni terminate in default in passato. La conoscenza di tali informazioni consente di classificare le diverse esposizioni in classi che potranno essere impiegate per la stima delle LGD future.

[2] *Determination of Default Probability by Loss Given Default- M. Misankova, E. Spuchľakova, K. Frajtova*

Exposure at Default

"L'Exposure at Default stima il valore effettivo del credito al verificarsi dello stato di insolvenza, è una variabile stocastica la cui variabilità dipende dal tipo di finanziamento concesso al debitore."[3]

L'esposizione creditizia può avere un valore certo, di cui si conosce l'ammontare esatto, o un valore incerto. Nel caso di incertezza, il debitore beneficia di un'opzione che gli consente di variare l'ammontare del prestito, seppur rispettando determinati limiti, quindi l'istituto creditizio non può quantificare immediatamente l'ammontare del prestito, ma solo al verificarsi dell'insolvenza.

Per stimare EAD è necessario conoscere sia la parte utilizzata, detta Drawn Portion (DP), sia la parte non utilizzata, Undrawn Portion (UD), ma anche un'ulteriore variabile che rappresenta la quota di UD che potrebbe essere utilizzata dal debitore in prossimità del default, detta fattore di conversione del credito (CCF). Analiticamente quindi si ottiene:

$$\boldsymbol{EAD} = DP \cdot UP \cdot CCF$$

Includere all'interno della formula analitica anche la quota non utilizzata porta a sovrastimare la perdita e conseguentemente un prezzo maggiore per il prestito, ma tale aumento non deve essere interamente trasferito allo spread applicato alla Drawn Portion. La maggiore perdita attesa collegata all'utilizzo dell'UP è coperta da una quota proporzionale alla stessa parte inutilizzata, detta commitment fee. In alcuni Paesi invece si preferisce ridurre al minimo il rischio di esposizione emettendo prestiti revocabili, assumendo quindi un atteggiamento più prudenziale.

[3] *La gestione del rischio e allocazione del capitale nelle banche – Sironi A.*

Perdite inattese

Per "Perdita Inattesa" o **Unexpected Loss (UL)**, si intende "la perdita eccedente la perdita attesa, per tale tipologia di perdita non è possibile prevedere un accantonamento ad hoc, quindi il miglior modo per cautelarsi da tali perdite resta quello di stabilire un livello adeguato di patrimonio di vigilanza, superiore alla soglia minima prevista dalla regolamentazione."[4] Di contro, l'incremento degli accantonamenti oltre il livello di perdita attesa non è considerato favorevole da parte di shareholders e stakeholders degli intermediari bancari e finanziari, in quanto determinano una riduzione del conto economico adducendo effetti negativi al risultato di gestione.

All'interno di un portafoglio di crediti, la perdita inattesa può essere ridotta attraverso la diversificazione del rischio di credito, aggiungendo al portafoglio dei prestiti le cui perdite inattese presentano una bassa correlazione con quelli già presenti. La perdita attesa invece non può essere ridotta diversificando il portafoglio, essendo questa pari alla somma delle perdite attese sui singoli prestiti.

Analiticamente, nel caso in cui il tasso di perdita LGD sia deterministico, la perdita inattesa si definisce come:

$$UL = LGD \cdot \sqrt{PD \cdot (1 - PD)}$$

Altrimenti, se si tiene anche conto della volatilità del tasso di perdita:

$$UL = \sqrt{PD * (1 - PD) \cdot LGD^2 + PD \cdot \sigma^{2}_{LGD}}$$

[4] *La gestione delle perdite attese e delle perdite inattese per gli intermediari bancari e finanziari - D'Auria C., Moderari*

La distinzione tra perdita attesa e perdita inattesa è fondamentale anche da un punto di vista economico, in quanto per le perdite attese è previsto un accantonamento a riserva, registrato come costo in conto economico, mentre per le perdite inattese è prevista una copertura attraverso il capitale della banca, che dovrà quindi essere sopportato dagli azionisti.

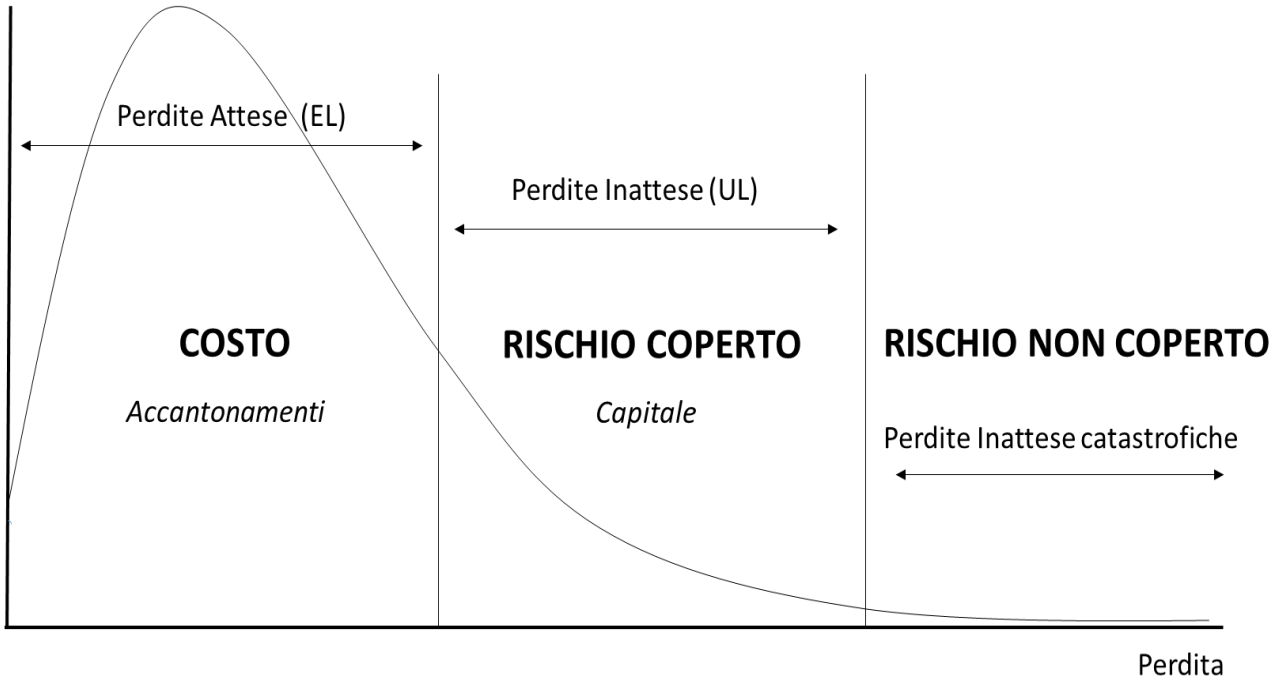


FIGURA 1.2- DISTRIBUZIONE DELLE PERDITE

Linee guida del Comitato di Basilea

Il Comitato di Basilea è un organismo internazionale di cooperazione, composto dai rappresentanti delle banche centrali ed autorità di vigilanza dei Paesi del G10.
Le decisioni del Comitato non hanno un valore giuridico sotto il profilo formale, ma influenzano in modo determinante le legislazioni dei vari Paesi, allo scopo di perseguire la stabilità monetaria e finanziaria, di rafforzare la solidità e la solvibilità del sistema bancario internazionale attraverso l'introduzione di requisiti di capitale minimo obbligatori, finalizzati alla riduzione delle crisi bancarie.

Nel 1988 il Comitato di Basilea ha proposto l'Accordo di Basilea, al fine di regolamentare le Autorità di Vigilanza e stabilire i requisiti richiesti dalle banche, in funzione dei rischi creditizi dalle stesse assunti. Il Comitato ha definito tre elementi attraverso i quali ha strutturato i requisiti di capitale per le istituzioni bancarie: il Patrimonio di Vigilanza, il Rischio, e il Rapporto minimo tra capitale e rischio.

Con il termine "Patrimonio di Vigilanza" si intende la quantità di denaro che la banca è obbligata a detenere a fronte delle sue attività di rischio, generalmente rappresentate dai prestiti. La scelta del livello ottimale di patrimonializzazione è fondamentale e richiede un'attenta ponderazione dei vantaggi e degli svantaggi che ne derivano. Il patrimonio di Vigilanza viene suddiviso a sua volta in patrimonio base (tier1), patrimonio supplementare (tier2), e patrimonio supplementare "discrezionale" (tier3) utilizzabile solo per fronteggiare il rischio di mercato. Ipotizzando un innalzamento del patrimonio di vigilanza, si potrebbero osservare due possibili effetti:

- Positivo in quanto la maggiore quantità di capitale detenuto internamente alla banca garantisce più stabilità al sistema finanziario;
- Negativo a causa della riduzione di redditività per l'istituto bancario.

Infatti, a fronte di una maggiore percentuale di patrimonio detenuto, si registra una diminuzione del patrimonio che la banca può investire in modo profittevole. La minore redditività rende la singola istituzione meno appetibile per gli shareholders e diminuisce la competitività di tutto il settore.

Il Patrimonio di Vigilanza si calcola sottraendo alla somma algebrica del patrimonio di base e supplementare, la somma algebrica di:

- partecipazioni in enti creditizi e/o finanziari superiori al 10% del capitale della società partecipata, nonché gli strumenti ibridi di patrimonializzazione e i prestiti subordinati verso tali enti con la caratteristica di essere computati nel patrimonio dell'ente emittente.
- le partecipazioni in titoli nominativi di Società di investimento a capitale variabile superiore a 20.000 azioni;
- le partecipazioni inferiori al 10% del capitale sociale in enti creditizi e/o finanziari, prestiti subordinati e strumenti ibridi di patrimonializzazione (diversi dal punto 1) per la parte eccedente il 10% dei fondi propri della banca partecipante.

Per ogni tipologia di esposizione è assegnato un determinato grado di rischio, attraverso la creazione di un coefficiente di ponderazione. Questi coefficienti sono utili per individuare il grado di insolvenza del debitore, le garanzie ricevute e l'eventuale *rischio Paese* presente nel rapporto creditizio. "L'Accordo ha stabilito due requisiti minimi di adeguatezza del capitale di una banca:

1. Il rapporto tra attività totali e capitale deve essere inferiore a 20 volte.
2. Cooke Ratio: il capitale deve essere almeno pari all' 8% delle attività ponderate per il rischio. Almeno il 50% del capitale deve essere Tier1."[5]

[5] *http://www00.unibg.it/dati/corsi/60012/39658-2010%2008.%20Basilea%202.pdf*

Ad ogni attività in bilancio è assegnato un "risk weight", che riflette il rischio di quella categoria:

Risk weight	Categorie di attivi
0%	Cassa o lingotto d'oro, prestiti a governi OECD
20%	Prestiti a banche OECD
50%	Prestiti garantiti da immobili
100%	Altri prestiti

Queste ponderazioni però hanno creato incentivi a costruire degli arbitraggi regolamentari per alterare i portafogli bancari con l'obiettivo di massimizzare il valore per gli azionisti. L'Accordo del 1988 è stato oggetto di critiche sin dalla sua introduzione; si è sostenuto che tale disciplina è stata incapace di considerare adeguatamente le diversità di mercato creditizio delle controparti, dando uguale importanza a banche, imprese e Stati con diversa rischiosità. La scadenza dei crediti non è considerata un fattore di rischio, mettendo sullo stesso piano prestiti a breve, medio e lungo termine e inoltre il principio di diversificazione è del tutto trascurato. I fenomeni in oggetto hanno spinto il Comitato all'elaborazione di un comitato all'elaborazione di un nuovo Accordo di Basilea (Basilea II).[6]

[6] *Strumenti di Controllo e Analisi del Rischio – Albergo F.*

Basilea II

L'obiettivo di Basilea II è elaborare un ordinamento più sensibile al rischio, incentivando il monitoraggio e la gestione da parte degli intermediari. Il nuovo Accordo individua due nuove categorie di rischio: il Rischio Operativo e il Rischio di Tasso d'interesse nel portafoglio bancario. I pilastri fondamentali su cui si basa sono:

1. Requisiti patrimoniali minimi;
2. Il controllo prudenziale sull'adeguatezza patrimoniale delle banche valutato in rapporto alle loro specificità operative e organizzative;
3. La disciplina del mercato.

Il Rischio Operativo è definito come "il rischio di perdite dirette o indirette risultanti dall'inadeguatezza o dalla disfunzione di procedure, risorse umane e sistemi interni oppure da eventi di origine esterna".[7]

Sono diversi i metodi di misurazione del rischio di credito, nell'approccio standard, la ponderazione del rischio è attribuita sulla base del rating assegnato alla controparte da agenzie esterne che soddisfano criteri minimi di obiettività, trasparenza, indipendenza, credibilità e tali agenzie sono denominate "Agenzie esterne di valutazione del merito di credito" (External Credit Assessment Institution, ECAI).

In alternativa a tale metodo per la misurazione del merito creditizio, l'intermediario può applicare un sistema di rating interno, denominato Internal Rating Based (IRB) suddiviso in due categorie: approccio di *base* e *avanzato*, le due versioni differiscono per il numero di parametri che il modello deve stimare. Il Sistema di rating interno deve rispettare alcuni requisiti minimi: i crediti devono essere distribuiti tra le varie classi di rating senza concentrazione in una specifica di esse, il rating va rivisto periodicamente e deve essere assegnato ai debitori prima della concessione del prestito.

[7] *art.158 di Basilea II*

Le Autorità di Vigilanza dovranno validare in ogni caso il sistema di rating interno adottato. Gli intermediari dovrebbero fornire la documentazione riguardante la struttura di capitale, l'adeguatezza patrimoniale e le attività di gestione rischio interna al fine di ottenere un mercato più trasparente.

I limiti di Basilea II si sono palesati a seguito della crisi finanziaria del 2008, infatti nonostante le banche rispettassero i requisiti di capitale minimo, si è riscontrato un'insufficiente quantità e qualità del capitale. Inoltre, non si è tenuto conto del problema della pro-ciclicità del capitale: la misura dei rischi si attenua nella fase ascendente del ciclo e tende a crescere nei momenti di crisi. La stessa quantità di patrimonio sostiene un maggiore volume di attività nella fase di crescita e un volume minore durante la fase di declino. Queste motivazioni hanno portato a rivedere gli accordi di Basilea II, dando vita a un nuovo accordo: Basilea III.[8]

[8] *Strumenti di Controllo e Analisi del Rischio – Albergo F.*

Basilea III

La riforma del sistema finanziario internazionale, approvata nel 2010 prevede come principale obiettivo il progetto di rafforzamento delle banche attraverso requisiti di capitale più stringenti.

Le novità introdotte da Basilea III riguardano:

- *Capitale*: il target minimo per il Patrimonio complessivo non cambia, è sempre l'8% delle attività ponderate per il rischio ma il requisito per il Patrimonio di qualità primaria viene reso esplicito e portato al 4.5%. Viene richiesto alle banche di dotarsi di un cuscinetto di capitale aggiuntivo (buffer), pari al 2.5% delle attività ponderate per il rischio. Tale buffer potrebbe anche aumentare fino al 5% nei momenti di turbolenza. È introdotto un ulteriore cuscinetto, il Countercyclical capital buffer finalizzato a garantire che le banche accumulino risorse patrimoniali nelle fasi di eccessiva crescita del credito aggregato.
- Introduzione di una soglia massima al leverage fissata al 3%, il coefficiente è definito come rapporto tra il volume delle attività e delle esposizioni fuori bilancio e il Capitale.
- Introduzione di due coefficienti di liquidità:
  - Liquidity Coverage Ratio al fine di evitare situazioni di stress tra scadenze dell'attivo e quelle del passivo;
  - Net stable funding ratio, parametro di controllo di medio periodo sull'equilibrio tra attività e passività.

Con il nuovo provvedimento si è tentato di rafforzare la capacità delle banche di assorbire shock derivanti da tensioni finanziarie ed economiche. Le nuove regole si inseriscono nell'ambito della vigilanza bancaria e riguardano sia la regolamentazione micro-prudenziale, a livello di singola banca, sia quella macroprudenziale, a livello di mercato. Gli accordi mirano ad uniformare il sistema finanziario e ad assicurare una disciplina omogenea su aspetti fondamentali nell'esercizio dell'attività bancaria. Il pacchetto normativo contiene anche specifiche disposizioni volte al raggiungimento di una maggiore trasparenza informativa.[9]

[9] *https://www.startingfinance.com/approfondimenti/basilea-iii/*

Capitolo 2: Modelli di credit-scoring

La maggior parte dei modelli di valutazione del rischio si basano sulla stima della probabilità di insolvenza, adottando alcune ipotesi semplificatrici per spiegare la relazione tra probabilità di insolvenza e recovery rate.

Tra i modelli più utilizzati per prevedere il default, vi è una classe di modelli statistici, generalmente noti come modelli di valutazione del credito. Si tratta di modelli multivariati che utilizzano come input i principali indicatori economici e finanziari di un'azienda, attribuendo un peso a ciascuno di essi, che riflette la sua importanza relativa nella previsione del default. Il risultato è un indicatore creditizio espresso come punteggio numerico, che misura indirettamente la probabilità di insolvenza. I diversi approcci di credit scoring si differenziano per la tipologia di rischio considerato, modalità di determinazione dell'esposizione, fattori determinanti la probabilità di migrazione o di insolvenza, classificazione del livello di rischio della controparte o per la determinazione del tasso di insolvenza e di perdita.

Le tecniche alla base dei modelli di credit scoring sono state ideate negli anni '30 da Fisher e Durand, ma la spinta decisiva allo sviluppo e alla diffusione di questi modelli è arrivata circa trent'anni più tardi con gli studi di Beaver, il quale ha introdotto un modello di valutazione del rischio di credito basandosi su un'analisi univariata degli indicatori di bilancio. Spesso tale approccio induce a un'interpretazione fuorviante dei risultati, per cui tale modello è stato modificato in seguito da Altman, il quale ha introdotto il primo approccio basato su un'analisi multivariata.

Di seguito verranno trattati alcuni dei modelli utilizzati per valutare il rischio di credito, questi si differenziano tra loro da un punto di vista formale e concettuale: l'analisi discriminante e i modelli di regressione utilizzano un approccio deduttivo volto a spiegare la probabilità di default, altri modelli come le reti neutrali invece si basano su un approccio puramente empirico.

Analisi discriminante lineare

L'analisi discriminante lineare è basata sull'identificazione di variabili, ottenute da un campione di aziende, che consentono di tracciare il confine separatore tra il gruppo di aziende sane e il gruppo di aziende insolventi, cioè caratterizzate da particolari situazioni quali per esempio liquidazione, ristrutturazione finanziaria o società con debito dubbio.

Il punteggio generato dalla combinazione delle due variabili originali viene visualizzato sull'asse z. L'analisi discriminante lineare costruisce il punteggio di z come combinazione lineare delle variabili indipendenti come indicato analiticamente dalla seguente formula:

$$\mathbf{Z}=\sum_{j=1}^{n} y_j\, x_j$$

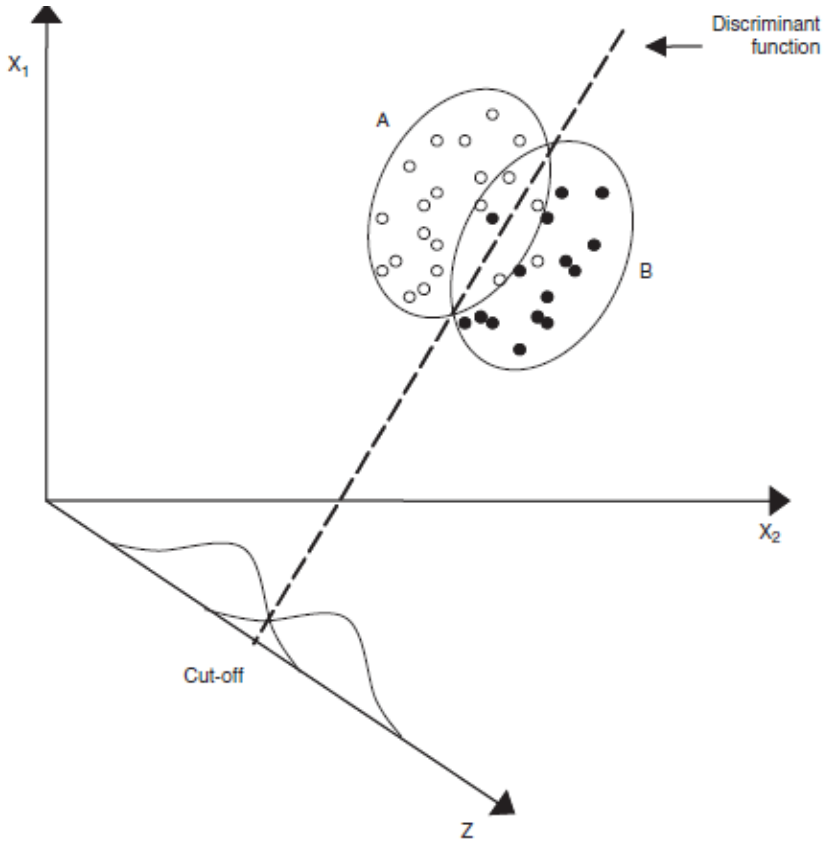


FIGURA 2.1- ANALISI DISCRIMINANTE

I coefficienti y_j di questa combinazione lineare sono scelti in modo da ottenere un punteggio z che distingua in modo più chiaro possibile i gruppi di aziende sane da quelle anomale in modo tale da massimizzare la distanza tra i due gruppi di aziende z_a e z_b (chiamati "centroidi"). L'obiettivo dell'analisi è fare in modo che i valori delle aziende sane siano il più possibile vicini tra loro e il più possibile distanti da quelli delle aziende anomale.

L'analisi discriminante viene utilizzata per stimare la probabilità di default associata alle singole imprese analizzate. Se le variabili indipendenti sono distribuite secondo una distribuzione normale multivariata, la probabilità di default è calcolata come:

$$\mathbf{PD} = p(B|\mathbf{x}_i) = \frac{1}{1+\frac{1-\pi_B}{\pi_B}e^{z_i-\alpha}}$$

In cui π_B rappresenta la "probabilità a priori di default", una misura della qualità media del portafoglio crediti della banca, che non dipende dalle caratteristiche del singolo cliente ma dalle caratteristiche generali del mercato, mentre α è definito come il punto di mezzo tra due centroidi.

Modelli di regressione

Regressione lineare

I modelli di regressione sono alla base di ogni analisi di dati riguardante la descrizione della relazione esistente tra una variabile di risposta e una o più variabili esplicative, diventando quindi il metodo di riferimento in ogni campo, per l'analisi di tali situazioni. Lo scopo della regressione è quello di trovare il miglior fitting per descrivere la relazione causale tra una variabile dipendente quantitativa, e un set di variabili indipendenti quantitative, chiamate anche variabili covariate. Così come accade nel caso dell'analisi discriminante, è necessario individuare un campione di imprese, che saranno suddivise in aziende sane e non sane, attraverso una variabile binaria che può assumere quindi solo valori pari a 0 o 1.

Poiché non è possibile indagare sull'intera popolazione la relazione tra le variabili considerate, si verificano le ipotesi estraendo un campione rappresentativo della popolazione in modo da descrivere la relazione tra le variabili considerate in base a tale campione. In questa fase è necessario selezionare un campione sufficientemente grande in modo da garantire la presenza di un numero di imprese insolventi elevato, affinché i risultati della regressione siano statisticamente significativi.
Successivamente sono selezionate delle variabili da assegnare a ciascuna impresa, di solito sono individuati degli indicatori economico-finanziari riportati nel bilancio aziendale prima che si verifichi l'evento di default. Infine, viene applicato il modello che sarà usato per stimare la probabilità di default delle imprese che richiedono finanziamenti bancari attraverso la formula seguente:

$$Y_i = \alpha + \sum_{j=1}^{n} \beta_j x_{i,j} + \varepsilon_i$$

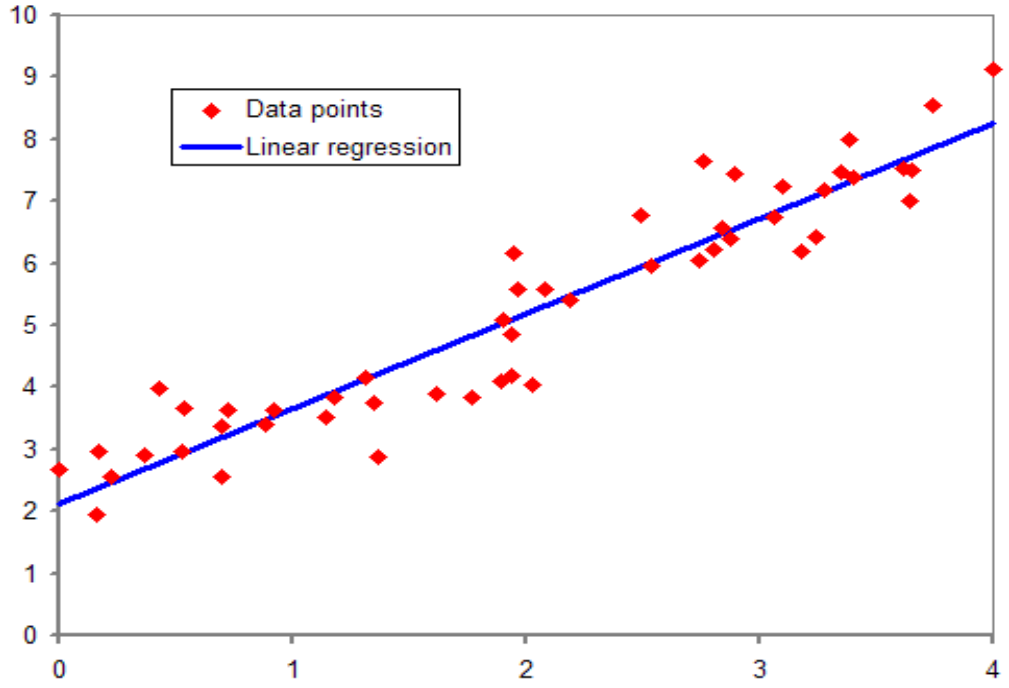


FIGURA 2.2- REGRESSIONE LINEARE

Come si può notare dal grafico, nel modello di regressione lineare, il legame tra le variabili che portano all'identificazione della probabilità di default è espresso attraverso una funzione di tipo lineare, pertanto si parla di regressione lineare multipla. Se la relazione tra le variabili invece non è lineare, è possibile effettuare delle apposite trasformazioni, come nel caso della *regressione logistica*, in cui si fa ricorso ai logaritmi per trasformare la relazione.[10]

[10] *Regressione Multipla e Logistica - Sense V.*

Regressione logistica

L'utilizzo della regressione logistica si è intensificato durante gli ultimi anni, diffondendosi in molti campi, tra cui quello delle ricerche biomediche, per predire il rischio di sviluppare malattie sulla base di determinate caratteristiche osservate del paziente, nel campo dell'ingegneria per la previsione della probabilità di fallimento di un processo, nel mondo del business per predire la probabilità di un acquirente di comprare un prodotto, e nell'ambito finanziario.

Il modello di regressione logistica viene utilizzato al fine di analizzare la relazione causale che intercorre tra una variabile dipendente dicotomica e una o più variabili indipendenti. Nella regressione logistica la variabile dipendente è utilizzata per evidenziare l'appartenenza a un determinato gruppo. La rappresentazione grafica della funzione logistica è una curva a S che partendo da qualsiasi numero reale è in grado di mapparlo in un valore compreso tra 0 e 1.

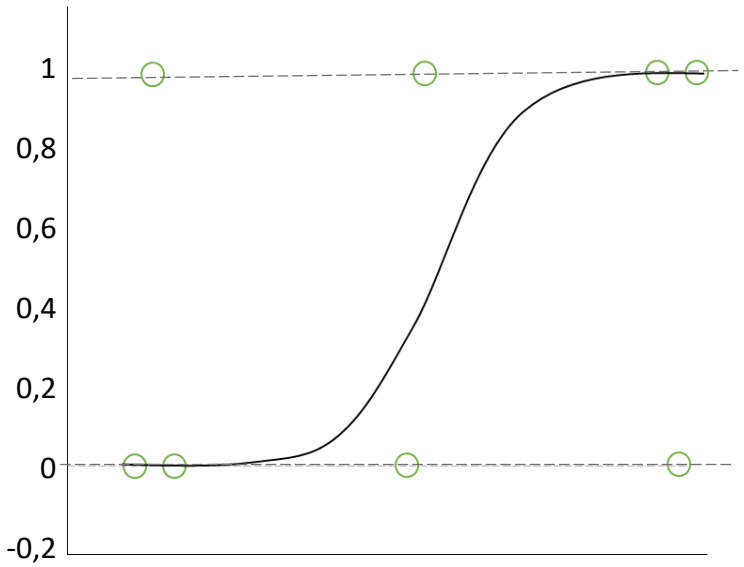


FIGURA 2.3- REGRESSIONE LOGISTICA

Si assume vi sia un'unica popolazione di imprese, per ciascuna delle quali è determinato il rischio di default attraverso una variabile latente non osservabile che indicheremo con y*= β'x+μ. Si è in grado di osservare solo una realizzazione di tale variabile dicotomica:

$$f\ (x) = 1,\ se\ y^* > 0\ se\ \mu > -\alpha - \beta x$$

$$f\ (x) = 0,\ se\ y^* \leq 0 se\ \mu \leq -\alpha - \beta x$$

La probabilità che y_i=1=Prob (u_i> $-\alpha - \beta x_i$)=1-F ($-\alpha - \beta x_i$), in cui F rappresenta la funzione di distribuzione cumulativa di u_i. Assumendo la distribuzione logistica della probabilità si ha:

$$\text{P}\ (y_i{=}1) = \text{Pi} = 1\text{-F}\ (-\alpha - \beta x_i) = \frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

$$\text{P}\ (y_i{=}0) = 1\text{- Pi} = \frac{1}{1+e^{(\alpha+\beta x_i)}}$$

Si ha quindi che $\frac{\text{Pi}}{1-\text{Pi}} = e^{(\alpha+\beta x_i)}$, da cui è possibile ricavare $\alpha + \beta x_i = ln\left[\frac{Pi}{1-Pi}\right]$, equivalente a:

$$\boldsymbol{ln\left[\frac{f_A(X)}{f_S(X)}\right] = \alpha + \beta x_i}$$

in cui f rappresenta la funzione densità di probabilità delle popolazioni sane e anomale.
È possibile sostituire la probabilità con l'odds, cioè un modo di esprimere la probabilità stessa mediante il rapporto tra le frequenze osservate in una categoria e le frequenze osservate nell'altra.

$$odds = \frac{prob(y_i \mid x_i)}{1 - prob(y_i \mid x_i)}$$

Una volta calcolato l'odds, è possibile calcolare il suo logaritmo naturale, il logit:

$$Ln(odds) = logit(prob(y_i|x_i)) = Ln\left[\frac{prob(y_i|x_i)}{1-prob(y_i|x_i)}\right] = \beta' x_i$$

L'odds e il logit sono modi differenti di esprimere la stessa informazione, la trasformazione in logit serve solo a garantire la correttezza matematica dell'analisi. La funzione logit funge da "collegamento" in quanto permette di associare le probabilità (comprese tra 0 e 1) all'intero intervallo di numeri reali.

Valutazione della significatività del modello

Nell'analisi della regressione logistica l'interpretazione della relazione tra variabili indipendenti e variabile dipendente avviene mediante la valutazione dei parametri del modello, attraverso l'algoritmo di massima verosimiglianza (maximum likelihood). Tale algoritmo stima i parametri in modo da massimizzare la funzione log-verosimiglianza, che indica quanto è probabile ottenere il valore atteso della variabile dipendente, dati i valori delle variabili indipendenti.

La funzione di verosimiglianza è:

$$L = \prod_{i=1}^{n} \left(\frac{1}{1+e^{\beta' xi}}\right)^{(1-y_i)} \left(\frac{e^{\beta' xi}}{1+e^{\beta' xi}}\right)^{y_i}$$

La funzione di log-verosimiglianza è quindi:

$$\ln(L) = \sum_{i=1}^{n} \{ y_i \ln[prob(y_i \mid x_i] + (1 - y_i) \ln[1 - prob(y_i \mid x_i]\} =$$

$$= \sum_{i=1}^{n} \left\{ y_i \ln\left[\frac{prob(y_i \mid x_i)}{1 - prob(y_i \mid x_i)} \right] + \ln[1 - prob(y_i \mid x_i)] \right\} =$$

$$= \sum_{i=1}^{n} \left\{ y_i \beta' x_i + \ln\left[1 - \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} \right] \right\} =$$

$$= \sum_{i=1}^{n} \left\{ y_i \beta' x_i + \ln\left[\frac{1}{1 + e^{\beta' x_i}} \right] \right\} = \sum_{i=1}^{n} \left\{ y_i \beta' x_i - \ln\left[1 + e^{\beta' x_i} \right] \right\}$$

Il sistema di equazioni di massima verosimiglianza si ottiene derivando Ln(L) rispetto ai parametri e ponendo la derivata pari a zero.

$$\frac{\partial Ln(L)}{\partial \beta_1} = \sum_{i=1}^{n} \left\{ y_i - \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}} \right\} = 0$$

$$\frac{\partial Ln(L)}{\partial \beta_2} = \sum_{i=1}^{n} \left\{ y_i x_i - \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}} x_i \right\} = 0$$

Indicando con p_i la probabilità dell'evento:

$$p_i = \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}}$$

Si può affermare che la somma delle probabilità dell'evento corrisponde alle somme degli eventi osservati:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p_i$$

$$n_a = \sum_{i=1}^{n} p_i$$

Per valutare la significatività di un coefficiente, si valuta il modello con, e successivamente, senza la variabile in questione e tale confronto si basa sul rapporto di verosimiglianza:

$$G = -2\ln\left[\frac{likelihood\ del\ modello\ senza\ la\ variabile\ x}{likelihood\ del\ modello\ con\ la\ variabile}\right] = -2\ln\left[\frac{\left(\frac{n_a}{n}\right)^{n_a}\left(\frac{n_s}{n}\right)^{n_s}}{\prod_{i=1}^{n} prob(y_i \mid x_i)^{y_i}\left[1 - prob(y_i \mid x_i\right]^{(1-y_i)}}\right]$$

L'ipotesi nulla è che l'inserimento della variabile x non apporti un contributo positivo alla funzione di verosimiglianza. Per verificare tale ipotesi è necessario confrontare il p-value corrispondente al valore calcolato di G, con il livello di significatività α scelto (G si distribuisce asintoticamente come una variabile chi-quadro).[11]

[11] *Applied Logist Regression - Hosmer D., Lemeshow S.*

Regressione Multipla e Logistica - Sense V.

https://it.qaz.wiki/wiki/Logistic_regression#Logistic_function,_odds,_odds_ratio,_and_logit

Reti neurali

I modelli analizzati finora seguono un approccio *strutturale:* si basano su ipotesi e ne cercano la conferma in un campione di dati empirico, solitamente si tratta di indici di bilancio o informazioni qualitative riguardanti l'età dell'azienda, la posizione geografica, ecc. I modelli strutturali sono basati su solidi algoritmi che utilizzano test inferenziali per verificare la reale significatività dei coefficienti stimati.[12]

Le reti neurali invece si basano su un approccio induttivo: se a partire da un campione di dati si analizza una certa regolarità empirica, viene utilizzata questa regolarità per prevedere futuri default di altre società. I modelli induttivi sono utili per individuare regole che governano un determinato fenomeno, anche quando risulta impossibile trovare delle regole deduttive. Sono utilizzati per trovare rapidamente un risultato, ma potrebbero risultare delle scatole nere in cui le regole sottostanti potrebbero non essere comprese completamente. Le reti neurali tentano di imitare i meccanismi di apprendimento della memoria umana, catturando alcuni aspetti che non potrebbero essere portati alla luce da un algoritmo di calcolo. La rete neurale è costituita da un numero di elementi o neutroni, organizzati in strati in modo tale che gli elementi più esterni ricevano degli input, li elaborino e inviino le informazioni allo strato più interno. Dopo che l'informazione è passata anche dagli strati più interni, la rete genera il risultato finale.

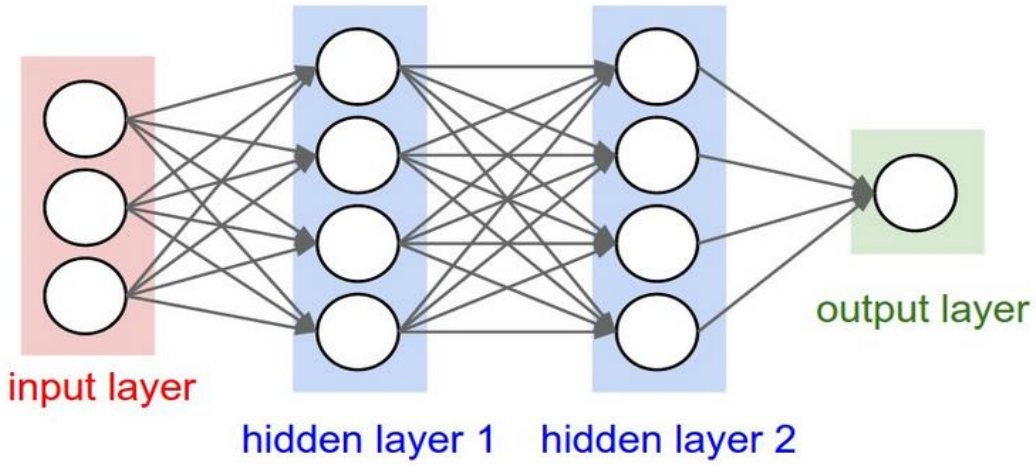
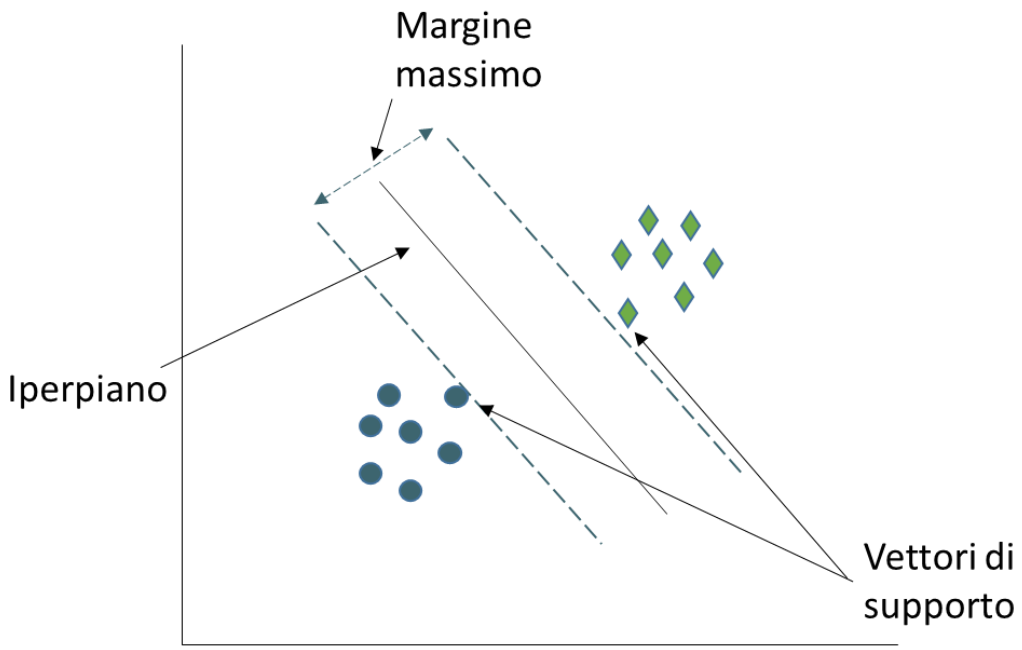


Figura 2.4-RETI NEUTRALI

[12] *La gestione del rischio e allocazione del capitale nelle banche – Sironi A*

Support Vector Machines

Il Support Vector machine o SVM è un algoritmo di apprendimento automatico supervisionato che ottiene la massima efficacia nei problemi di classificazione binari, ma che trova larga applicazione anche nel riconoscimento vocale e delle immagini. L'idea introdotta per la prima volta da Vladimir Naumovich Vapnik nei primi anni '90, si pone come obiettivo quello di trovare un iperpiano che sia in grado di dividere un dataset in due classi distinte. Operando in due dimensioni, l'iperpiano è una linea che separa e classifica un insieme di dati, mentre in tre dimensioni è identificato da un piano. I vettori di supporto sono i punti posizionati più vicini al piano, e sono considerati gli elementi critici del set di dati. La distanza tra i vettori di supporto di due classi differenti più vicini all'iperpiano è chiamata margine.



Separazione lineare

Innanzitutto, SVM cerca un iperpiano linearmente separabile in modo da dividere i valori di una classe dall'altra. Se ne esistono diversi sceglie quello che garantisce un'accuratezza migliore, che generi il minimo errore di classificazione su una nuova osservazione, mentre, se questo iperpiano non esiste il modello utilizza una mappatura non lineare.

L'iperpiano ottimale può essere definito come un prodotto scalare:

$$\vec{\omega}\vec{x} + b = 0$$

Dove $\vec{w}$ è il vettore peso, $\vec{x}$ è il vettore di caratteristiche di input e b è il bias.

Prendendo in considerazione due dimensioni, si ha:

$$b + w_1x_1 + w_2x_2 = 0$$

I limiti dei margini delle classi sono:

$$b + w_1x_1 + w_2x_2 \geq 1$$

$$b + w_1x_1 + w_2x_2 \leq -1$$

Che possono essere scritti in forma compatta come:

$$y_i \lfloor \vec{\omega}\vec{x}_i + b \rfloor \geq 1 \qquad i=1, \ldots, 1$$

L'iperpiano ottimo è quello che separa i vettori nelle due differenti classi $y \in \{+1,-1\}$ con la più piccola norma di coefficienti, cioè con margine massimo. Per trovare la distanza tra i vettori di supporto è possibile utilizzare uno tra i campioni che si trova sui vettori di supporto, scelto casualmente.

Indicando con ||w|| la norma di $\vec{w}$, allora la dimensione del margine massimo si trova come:

$$\min \frac{1}{2} ||\vec{w}||^2$$

Si tratta di un problema di ottimizzazione quadratica, che può essere risolto facendo ricorso alla funzione di Lagrange:

$$L(w,b,\alpha) = \frac{1}{2}\left|\left|\vec{w}\right|\right|^2 - \sum_{i=1}^{l} \alpha_i \{y_i \ [\vec{w}x_i + b] - 1\}$$

in cui α_i rappresenta i moltiplicatori di Lagrange. La ricerca di un punto di sella ottimale (w_0, b_0, α_0) è necessario perché la lagrangiana L deve essere minimizzata rispetto a w e b, e deve essere massimizzata rispetto a α_i non negativi. Si sfruttano le condizioni di Karush-Kuhn-Tucker (KKT) per individuare il punto di ottimo.

Al punto di sella (w_0, b_0, α_0), le derivate della lagrangiana L rispetto alle variabili primali dovrebbero essere uguali a zero:

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \qquad w_0 = \sum_{i=1}^{l} \alpha_i \, y_i \ x_i$$

$$\frac{\partial \mathcal{L}}{\partial b_0} = 0 \qquad \sum_{i=1}^{l} \alpha_i \, y_i = 0$$

Sostituendo i risultati appena ottenuti nella lagrangiana si ottiene:

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i \ y_j \ \alpha_i \alpha_j \overrightarrow{x_i x_j}$$

Per trovare l'iperpiano ottimale, il lagrangiano (α) deve essere massimizzato rispetto ad α_i non negativo.

Un problema di ottimizzazione quadratica standard può essere espresso in una notazione matriciale e formulato come segue:

$$L_d(\alpha) = -0.5\alpha^T \mathbf{H}\alpha + \mathbf{f}^T\alpha \ ,$$

$$\mathbf{y}^T\alpha = 0 \ ,$$
$$\alpha_i \geq 0, \quad i = 1, \ l \ ,$$

in cui H= $y_i \ y_j \ x_i^T x_j$, ed f è un vettore unitario.

La soluzione del problema di ottimizzazione sopra determina i parametri w_o e b_o dell'iperpiano ottimo come segue:

$$\mathbf{w}_o = \sum_{i=1}^{l} \alpha_{0i} y_i \mathbf{x}_i \ ,$$

$$b_o = \frac{1}{N_{SV}} \left(\sum_{s=1}^{N_{SV}} \left(\frac{1}{y_s} - \mathbf{x}_s^T \mathbf{w}_o \right) \right.$$

$$= \frac{1}{N_{SV}} \left(\sum_{s=1}^{N_{SV}} \left(y_s - \mathbf{x}_s^T \mathbf{w}_o \right) \ , \quad s = 1, \ N_{SV} \ . \right.$$

N_{vs} indica il numero di support vectors.

Infine, dopo aver calcolato i parametri w_o e b_o, otteniamo un iperpiano decisionale d(x) e una funzione indicatore $i_F = 0$.

$$D(\mathbf{x}) = \sum_{i=1}^{l} w_{0i} x_i + b_o = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i^T \mathbf{x} + b_o \ , \quad i_F = 0 = \text{sign}(d(\mathbf{x})) \ .$$

Per i dati del campione separabili linearmente tutti i vettori di supporto si trovano sul margine e, generalmente, sono solo una piccola parte di tutti i dati di addestramento.

Separazione non lineare

Quando i dati da analizzare non sono lineari è necessario rilassare i vincoli introducendo delle variabili di slack positive ξ_i, i=1, .., l

$$b + w_1 x_1 + w_2 x_2 \geq 1\text{-}\xi$$

$$b + w_1 x_1 + w_2 x_2 \leq \text{-}1+\xi$$

I vincoli possono essere espressi anche come:

$$y_i \ \lfloor \vec{\omega}\vec{x}_i + b \rfloor \text{ -}1+ \xi \geq 0 \qquad\qquad \xi \geq 0 \text{ i=1, ..., l}$$

Assegnando un costo agli errori ξ, la funzione da minimizzare diventa:

$$\min \frac{1}{2} \|\vec{w}\|^2 + \text{C } (\sum \xi_i^k)$$

C rappresenta un parametro corrispondente alla penalità assegnata agli errori. La funzione lagrangiana addesso vale:

$$L_p = \frac{1}{2}\|\vec{w}\|^2 + C\left(\sum_{i=1}^{l} \xi_i^k\right) - \sum_{i=1}^{l} \alpha_i \left[y_i \left(\vec{w}\,\vec{x_i} + b\right) - 1 + \xi_i\right] - \sum_i \mu_i \xi_i$$

in seguito alle condizioni di KKT, si ottiene:

$$\text{L}(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i \ \ y_j \ \ \alpha_i \alpha_j \vec{x_i}\vec{x_j}$$

Con $\sum_{i=1}^{l} y_j \ \ \alpha_i$=0 e $0 < \alpha_i < C$. La differenza con il caso lineare è che adesso i moltiplicatori lineari sono limitati superiormente dal parametro C.[13]

[13] *Support Vector Machines- Sciandrone M.*

https://lorenzogovoni.com/support-vector-machine/

https://www.developersmaggioli.it/blog/support-vector-machine/

Kernel

In alcuni casi, i problemi che non sono separabili nello spazio bidimensionale possono diventare tali facendo ricorso al trucco del *kernel:* la soluzione consiste nell'aggiungere una terza dimensione al piano di separazione in modo da ottenere uno spazio tridimensionale.

Supponiamo di mappare i dati iniziali non linearmente separabili in uno spazio di dimensione superiore usando una funzione di mapping φ in cui i dati adesso siano linearmente separabili. L'algoritmo di apprendimento dipende dai dati solo tramite il prodotto delle loro immagini attraverso φ. È possibile evitare di eseguire il prodotto esplicito tra le immagini dei vettori attraverso la funzione Kernel, definita come:

$$K(\vec{x_i}, \vec{x_j}) = \varphi(\vec{x_i})\varphi(\vec{x_j})$$

"L'estensione a superfici di decisioni complesse avviene mappando la variabile in input x in uno spazio di dimensione maggiore e lavorando poi con una classificazione lineare in questo nuovo spazio"[14]:

$$x \rightarrow \varphi(x) = (a_1 \varphi_1(x), a_2 \varphi_2(x), \ldots)$$

dove a_i sono numeri reali e φ_i sono funzioni reali. La funzione di decisione con il mapping diventa:

$$f(x) = \text{sign}\left(\sum_{i=1}^{l} y_i \alpha_i \cdot \varphi(x) \cdot \varphi(x_i) + b_0\right)$$

Sostituendo i prodotti scalari si ottiene:

$$f(x) = \text{sign}\left(\sum_{i=1}^{l} y_i \alpha_i \cdot K(x, x_i) + b_0\right)$$

[14] *Support Vector Machines- Maniezzo V.*

La scelta della funzione del kernel può influire notevolmente sulle prestazioni del modello. Tra le funzioni Kernel più diffuse ci sono:

- kernel *lineare*, definito come: K (x_i, x_j)= $\vec{x}_i \vec{x}_j$;

- kernel *polinomiale:* K (x_i, x_j)= $(x_i x_j + c)^d$, contiene una costante C e un grado di libertà g;

- kernel *radial basis function* (RBF) o gaussiano: K (x_i, y_j)=$e^{\left(-\frac{\left\|x_i - y_j\right\|^2}{2\sigma^2}\right)}$: se la differeza tra x e y tende a **0**, la funzione tenderà ad **1**, invece se la differenza tra **x** e **y** è alta, la funzione tenderà a **0**. Il parametro **σ** modifica la forma della campana, rendendola più sottile per valori piccoli e più distesa per valori grandi.

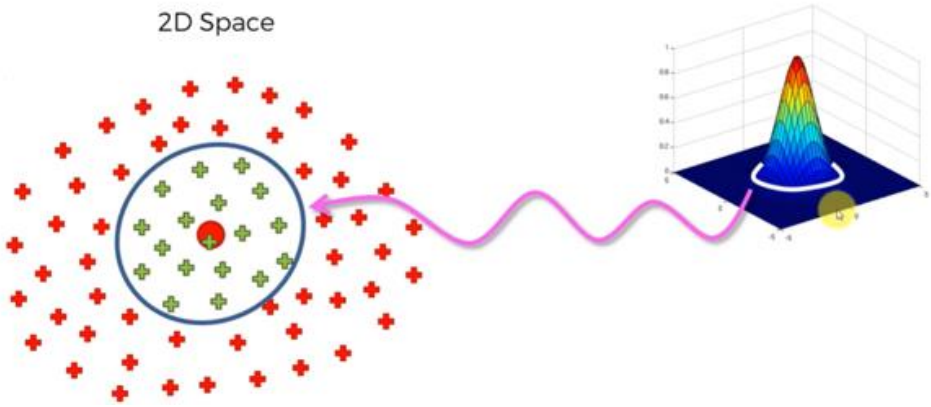


FIGURA 2.5-SUPPORT VECTOR MACHINES

Affinché una funzione sia un Kernel valido deve essere verificata una condizione necessaria e sufficiente, di solito espressa come teorema di Mercer. Questo teorema assicura che la funzione Kernel possa essere esprimibile come prodotto interno di due vettori nello spazio trasformato.
I vantaggi derivanti dall'utilizzo di SVM sono originati dall'efficacia negli spazi ad alta dimensione, efficienza nella memoria, in quanto nell'effettivo processo decisionale è utilizzato solo un sottoinsieme dei punti di allenamento, e ovviamente la possibilità di variare i parametri del kernel in modo da ottenere una maggiore performance di classificazione.

Modelli di portafoglio

Il problema della valutazione del rischio di credito diventa più complesso se dalla stima di una singola posizione si passa a considerare un intero portafoglio crediti. Di seguito si presenteranno i principali modelli per portafogli di esposizioni creditizie CreditMetrics e Credit Risk Plus. Il primo modello si basa sull'approccio alla Merton, mentre il secondo è un modello di tipo attuariale.

CreditMetrics

CreditMetrics (Gupton, Finger, & Bhatia, 1997) considera sia il rischio di default sia il rischio di deterioramento del merito creditizio, applicando una logica di valutazione a valori di mercato, secondo cui l'insolvenza dei titoli di debito emessi può avvenire esclusivamente alla loro scadenza.

CM incorpora una matrice di transizione che mostra la probabilità di un debitore di passare da un grado di credito a un altro, sulla base di dati storici. Le tabelle della probabilità di transazione sono fornite da valutatori come Moody's e Standard & Poor's.

Il modello calcola il valore di mercato del prestito, cedola inclusa, con orizzonte temporale di un anno. Le probabilità di transazione nella tabella vengono moltiplicate per il valore di mercato del prestito per ottenere una probabilità ponderata. Basandosi sulle tabelle, il VaR si ottiene calcolando la varianza del portafoglio ponderata in base alla probabilità e la deviazione standard (σ), quindi si calcola il VaR utilizzando una distribuzione normale (ad esempio 1,645σ per un livello di confidenza del 95%).

Con Creditmetrics è possibile anche utilizzare il metodo Monte Carlo come alternativa per il calcolo del VaR. Il modello sostiene che esiste una serie di valori patrimoniali che determinano il rating di un'azienda. Se il valore patrimoniale di una società scende o aumenta rispetto a un certo livello, alla fine di quel periodo, il suo nuovo valore patrimoniale determinerà la nuova valutazione in quel momento. Queste fasce di valori delle attività sono indicate da CreditMetrics come soglie di asset. Si presume che le

variazioni percentuali degli asset siano normalmente distribuite e utilizzando le probabilità dalla tabella della matrice di transizione, le soglie di probabilità (Pr) di asset Z_{DEF}, Z_{CCC} ..., possono essere calcolate come segue:

$$\Pr(_{Default}) = \Phi(Z_{DEF}/\sigma)$$

$$\Pr(_{CCC}) = \Phi(Z_{CCC}/\sigma) - \Phi(Z_{DEF}/\sigma)$$

$$Z_{DEF} = \Phi^{-1}\sigma$$

CreditMetrics applica le soglie di asset alla modellazione Monte Carlo utilizzando tre passaggi: "in primo luogo, le soglie di rendimento delle attività devono essere generate per ciascuna categoria di rating. In secondo luogo, scenari di asset di ritorno devono essere generati utilizzando una distribuzione normale. Il terzo passaggio consiste nel mappare i rendimenti degli asset nel passaggio 2 con gli scenari di credito della Fase 1. Normalmente vengono generati migliaia di scenari dai quali vengono calcolati la distribuzione del portafoglio e il VaR.".[15]

[15] *Credit risk measurement methodologies-Allen D*

Credit Risk Plus

Credit Plus si caratterizza per un approccio di tipo attuariale al problema, considera unicamente il rischio di insolvenza e si pone l'obiettivo di stimare la probabilità di deterioramento del merito creditizio di portafoglio sulla base dell'andamento corrente dell'economia identificata da una serie di variabili macroeconomiche, che tendono a influenzare le probabilità di transazione e di insolvenza.

"Credit risk concentra la propria attenzione sulla stima delle perdite future. Il valore attuale di un credito è funzione dei flussi di cassa attesi e della curva dei tassi utilizzata per scontarli, questa curva dei tassi incorpora uno spread che è funzione del grado di liquidità del mercato e del merito creditizio del debitore. La distribuzione delle perdite cattura il secondo fattore ma non il primo. Se uno shock sui tassi dovesse raddoppiare, lo spread richiesto dal mercato per prestiti a una certa classe di debitori rischiosi, la distribuzione delle perdite non ne recherebbe traccia mentre il valore di mercato dei prestiti registrerebbe consistenti minusvalenze*".*[16]

Credit Risk Plus, lavorando sulla distribuzione delle perdite, giunge a una stima del rischio che prescinde da eventuali shock negli spread di mercato. Si suppone, inoltre, che un credito evolva in modo binomiale, al termine di un dato arco temporale il prestito può aver dato luogo a una perdita oppure essere solvibile. I possibili stati del mondo si riducono quindi a solvente e insolvente, escludendo ogni possibile situazione intermedia.

Credit Risk Plus ipotizza l'indipendenza condizionale dei singoli crediti, assumendo che per ogni possibile stato del mondo i crediti presenti nel portafoglio di una banca siano non correlati, e che quindi il fallimento di un debitore non dipenda da quello degli altri.

Si ipotizza di conoscere la probabilità di insolvenza di un soggetto che richiede un prestito per un arco temporale di un anno. È possibile rappresentare l'insolvenza del debitore i-esimo con una variabile casuale discreta, uguale a 1 (con probabilità p_i) in caso di default

[16] *La gestione del rischio di credito con modelli di derivazione attuariale: il caso di CreditRisk+-Resti A.*

e 0 altrimenti. La distribuzione di probabilità di questa variabile casuale può essere sintetizzata da:

$$F_i(z) = z^0(1-p_i) + z^1 p_i = 1 + p_i(z-1)$$

Se volessimo considerare più di un debitore, ipotizzando che l'insolvenza di ogni debitore sia indipendente da quella degli altri, allora la funzione generatrice diventa:

$$F(z) = \prod_{i=1}^{m} F_i(z) = \prod_{i=1}^{m} 1 + p_i(z-1) = e^{\sum_{i=1}^{m} \log[1+p_i(z-1)]}$$

Se riteniamo la probabilità di insolvenza dei debitori molto bassa allora:

$$F(z) \cong e^{\sum_{i=1}^{m} p_i(z-1)} = e^{\mu(z-1)}$$

in cui u=$\sum_{i=1}^{m} p_i$ rappresenta il numero totale di default attesi.

Si considera ora l'espansione in serie della serie di McLaurin della F(z):

$$F(z) = F(0) + F'(0)z + F''(0)\frac{z^2}{2} + \cdots + F^{(n)}(0)\frac{z^n}{n!} + \cdots =$$

$$= e^{-\mu} + e^{-\mu} \cdot \mu \cdot z + e^{-\mu}\frac{\mu^2 z^2}{2} + \cdots + e^{-\mu}\frac{\mu^n z^n}{n!} + \cdots = \sum_{n=0}^{\infty} \frac{e^{-\mu}\mu^n z^n}{n!}$$

Il termine **p(n)=** $\boldsymbol{\frac{e^{-u}u^n}{n!}}$ rappresenta la probabilità che si verifichino n insolvenze.

Capitolo 3: Analisi del settore

Le attività che riguardano il settore dei *trasporti* occupano una posizione rilevante nel sistema economico italiano: gli spostamenti dei beni e delle persone costituiscono uno dei motori dello sviluppo economico, si avvalgono di una rete di circa 6.943 km di autostrade, incrementata del 7,2% negli ultimi vent'anni, di 142 km di strade regionali e provinciali, di una linea ferroviaria che accoglie ogni anno 865 milioni di passeggieri, di un traffico marittimo che sposta circa 475 milioni di tonnellate di merci e di una rete aerea che accoglie oltre 175 milioni di passeggeri all'anno. Attualmente però il settore dei trasporti in Italia sta attraversando un periodo di crisi: il rallentamento della crescita a livello macroeconomico, la necessità di rinnovare il modello di business e la crescente competizione con i Paesi esteri influiscono negativamente sulla profittabilità dello stesso.[17]

All'interno del settore si sono sviluppate strutture produttive molto differenti: alcuni comparti come quello ferroviario o aereo presentano un elevato grado di concentrazione, con un numero limitato di grandi imprese e con un'elevata quota di lavoratori dipendenti; altri, come il trasporto merci su strada, sono caratterizzati dalla presenza di piccole imprese, anche a carattere familiare, con un'alta percentuale di lavoratori autonomi.

Per descrivere l'andamento del settore dei trasporti è possibile mettere in atto un parallelismo con lo studio del ciclo economico del PIL: il ciclo mostra una profonda crisi nel biennio 2008-2009 ripetutasi nel 2012, con accenni di miglioramento per il settore a partire dal 2013. L'indicatore relativo al trasporto merci segue da vicino l'andamento del PIL evidenziando nel quinquennio 2011-2015 una flessione negativa, mentre a partire dal 2016 si registra una ripresa, in concomitanza con l'andamento del ciclo. La dinamica del trasporto passeggeri invece riflette con un certo ritardo l'andamento del PIL e risente in misura minore della crisi economica. L'indice infatti raggiunge un picco nel 2009 in controtendenza con l'andamento della crisi economica e diminuisce fino al 2012, mentre la ripresa si realizza in un periodo di pieno ristagno dell'economia, registrando tassi di crescita nettamente superiori rispetto a quelli del PIL stesso.

[17] *Dati estratti da "Trasporti e Telecomunicazioni – Annuario Statistico Italiano"*

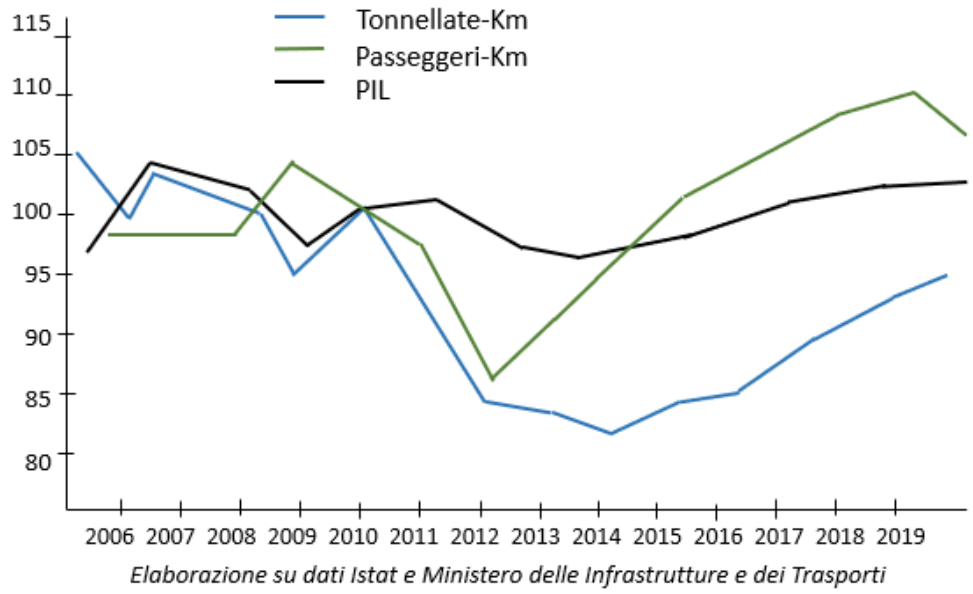


FIGURA 3.1- EVOLUZIONE DELLA DOMANDA DI TRASPORTO E DEL PIL - INDICI DI BASE 2010=100

Il numero delle imprese attive nel settore dei trasporti terrestri si è ridotto negli anni, subendo nel complesso una diminuzione dell'1,6%, tra il 2017 e il 2019, ma la percentuale delle società di capitali è in aumento. Queste subiscono un incremento di oltre un punto percentuale e in particolare, nell'autotrasporto si passa dal 20,4% del primo semestre del 2017 al 23,3% del 2019. Il decremento del numero di imprese, tuttavia, non sembrerebbe essere collegato alla situazione di crisi generale, in quanto le imprese rimaste attive nel settore non si sono limitate a sopravvivere ma hanno intrapreso un percorso di consolidamento, incrementando i loro profitti.

Anno	Trasporto terrestre e mediante condotte	Trasporto marittimo e per vie d'acqua	Trasporto aereo
2017	117.784	2.187	200
2018	116.562	2.287	195
2019	115.764	2.293	193

FIGURA 3.2-NUMERO DI IMPRESE ATTIVE

Di seguito è presentata un'analisi dei principali indicatori di redditività, al fine di comprendere la rilevanza economica del settore dei trasporti. Nel 2018 si registra una variazione positiva del fatturato rispetto all'anno precedente per il comparto del trasporto aereo (+3,9%) e del trasporto terrestre (+1,4%), mentre la variazione del fatturato per il trasporto marittimo e per vie d'acqua è negativa (-2,2%). Per due anni consecutivi i comparti del trasporto aereo e terrestre confermano la maggiore capacità di generare profitti grazie a un'efficiente gestione dei costi. Le imprese, infatti, si sono organizzate attraverso contratti di rete, specializzandosi in settori circoscritti e offrendo un servizio rapido e flessibile, movimentazioni sostenibili e tracciabili, e una presenza capillare sul territorio nazionale.

ANNI	Trasporto terrestre e trasporto mediante condotte		Trasporto marittimo e per vie d'acqua		Trasporto aereo	
	Indici	Variazioni % sull'anno precedente	Indici	Variazioni % sull'anno precedente	Indici	Variazioni % sull'anno precedente
2016	100,0	0,0	95,7	-4,3	95,6	-4,4
2017	104,3	4,3	101,5	6,1	101,9	6,6
2018	105,8	1,4	99,3	-2,2	105,9	3,9

Fonte: Istat, Rilevazione trimestrale sul fatturato dei servizi (R)

TABELLA 3.3-INDICI DEL FATTURATO A PREZZI CORRENTI (BASE 2015=100)

Nell'ultimo biennio l'indicatore di *redditività del capitale proprio* (ROE) ha subito un decremento, passando dal 14% del 2016 al 12% del 2018, mentre il Return on Investment (ROI), l'indice di *redditività del capitale investito*, si posiziona intorno al 7%. L'indicatore di *redditività delle vendite* (ROS) invece si attesta intorno al 3%, in calo rispetto all'anno precedente. I valori degli indicatori di redditività del settore sono in diminuzione nel biennio 2016-18 ma si attestano comunque su valori discreti se paragonati con quelli degli anni immediatamente successivi alla crisi economica del 2008.

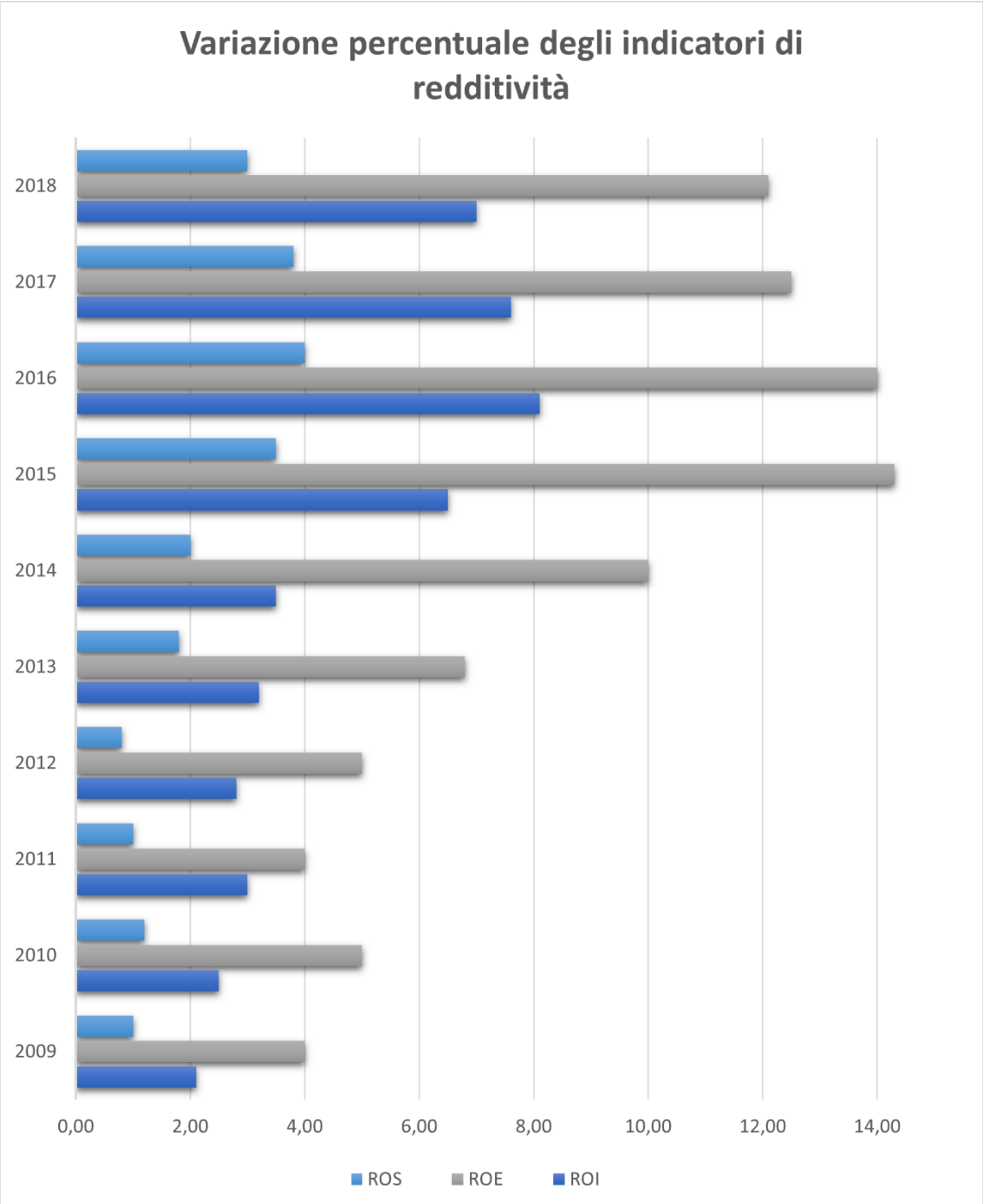


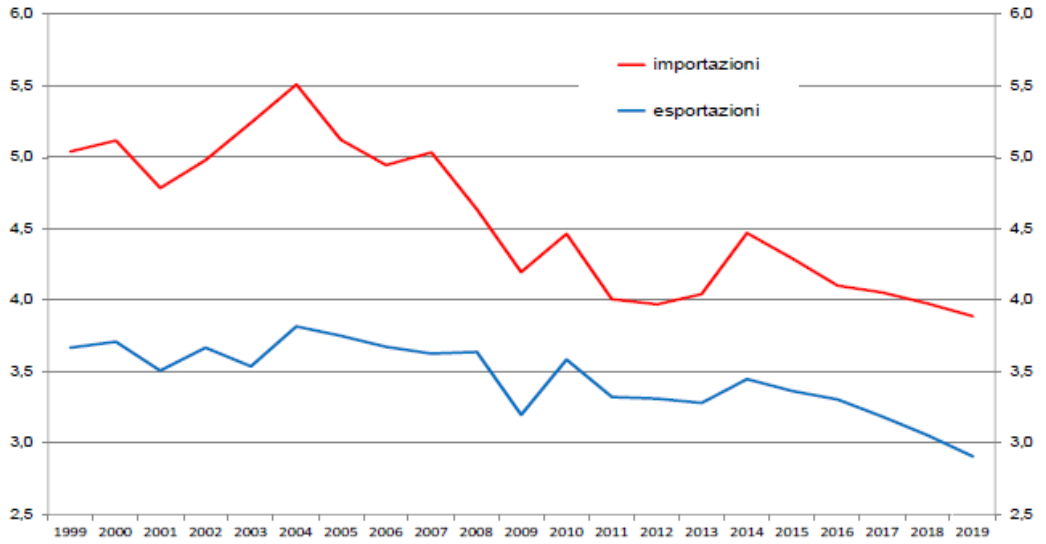
FIGURA 3.4-VARIAZIONE % INDICATORI DI REDDITIVITÀ

Dall'analisi dell'incidenza dei costi di trasporto sul valore delle importazioni ed esportazioni si conferma la diminuzione rispetto agli anni precedenti: i costi di trasporto incidono per il 2,9% sulle importazioni e per 3,8% sulle esportazioni.

Proprio grazie a questa variazione, la tendenza negativa evidenziata tra questo decennio e il precedente, che tende a convergere il traffico commerciale verso l'Europa dell'Est, sembra finalmente accennare a una svolta. L'Italia ha visto calare il volume di merci trasportate dalle proprie imprese dal 2008 al 2018, con una perdita del 40%, ma nello stesso 2018, i principali Paesi dell'Europa occidentale hanno rivisto crescere il loro giro d'affari, compresa l'Italia che attualmente occupa il sesto posto, dietro la Polonia, per tonnellaggio di merci trasportate.

Incidenza dei costi del trasporto sul valore delle importazioni e delle esportazioni dell'Italia

(percentuali)



In Italia, la gran parte dei trasporti terrestri avviene all'interno del territorio nazionale, mentre i trasporti internazionali interessano soprattutto l'Austria, la Francia e la Svizzera. Nel panorama europeo, il settore dei trasporti è una colonna portante in quanto costituisce oltre il 9% del valore aggiunto lordo dell'UE. I tre quarti dei trasporti terrestri Europei viaggiano su strada e un quinto su ferrovia, la quota restante invece si concentra sulle vie d'acqua interne.[18]

[18] *https://www.bancaditalia.it/pubblicazioni/indagine-trasporti-internazionali/index.html*

La politica dell'UE mira a individuare e risolvere i principali problemi che caratterizzano il sistema di trasporto. Gli obiettivi a medio-lungo termine prefissati sono: ridurre la congestione sia terrestre sia aerea, rendere più sostenibile i trasporti attraverso la riduzione delle emissioni di CO_2, migliorare la qualità dell'aria (entro il 2050 l'UE ha l'obiettivo di ridurre le emissioni prodotte dai trasporti del 60% rispetto ai livelli del 1990, e continuare a ridurre l'inquinamento prodotto dai veicoli).

Comparando la politica di incentivazione all'utilizzo del trasporto ferroviario dell'Italia con quella dei principali competitori europei del settore, emergono delle differenze sostanziali. La Svizzera, ad esempio, ha raggiunto risultati concreti: negli ultimi anni sono stati effettuati investimenti dedicati all'ammodernamento della rete ferroviaria, disincentivando il trasporto su ruote attraverso restrizioni orarie e tasse sul trasporto di merci pesanti con tariffe proporzionali a distanza, peso ed emissioni. Sia Svizzera sia Austria vantano la percentuale più alta in Europa per quanto riguarda i trasporti su rotaie, evidenziando una maggiore sensibilità alle tematiche ambientali.

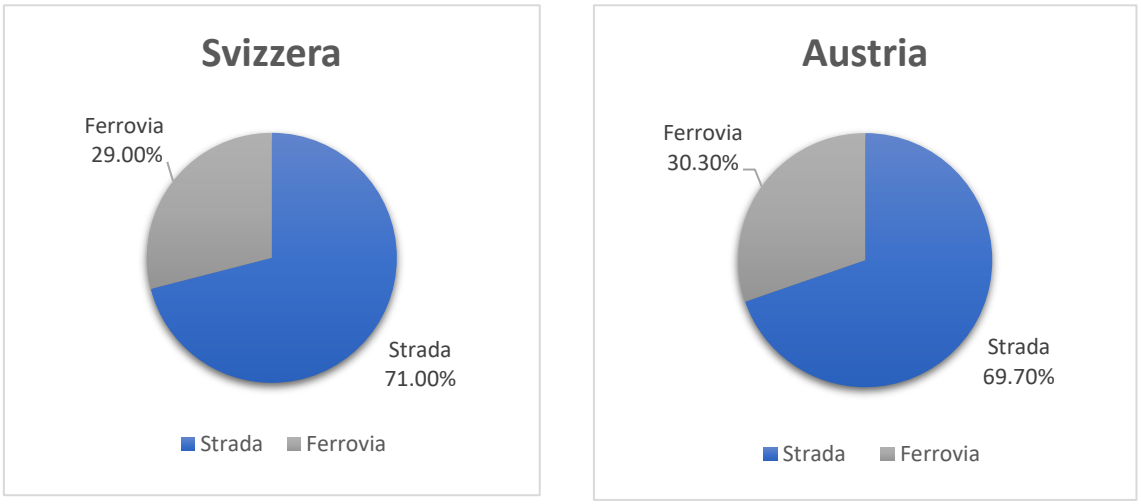


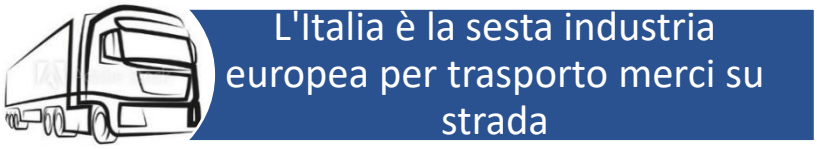
FIGURA 3.5- COMPARAZIONE TRASPORTI IN SVIZZERA E AUSTRIA

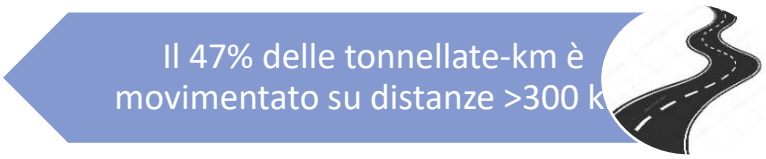
I trasporti terrestri

I trasporti su strada

In base a un'indagine messa in atto da Confetra su un panel di centinaia di imprese tra le più rappresentative dei vari settori, è emerso che nel 2018 si è registrato un rallentamento della crescita in ogni comparto del settore trasporto.

Il rallentamento si è accentuato nella seconda metà dell'anno ripercorrendo l'andamento dell'indice della produzione industriale rilevato da Istat. I comparti che meno hanno risentito di questa flessione sono quello terrestre e, in particolar modo, quello del servizio corrieristico, la cui crescita risulta comunque rallentata. Il trasporto stradale ha fatto registrare un +2,4% nel groupage e un +2,5% nel trasporto internazionale. Negli ultimi anni si registra un incremento dei trasporti affidati a società terze a discapito del trasporto effettuato in proprio, indice della propensione delle imprese a ricorrere sempre di più alla esternalizzazione del servizio, ma anche della maggiore concentrazione del settore a causa della riduzione del numero di imprese di autotrasporto.[19]





[19] *https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Passenger_transport_statistics/it*

Le principali tipologie merceologiche che percorrono tratte brevi, fino a 50 km, sono soprattutto minerali metalliferi, materie prime secondarie e rifiuti, prodotti delle lavorazioni di minerali non metalliferi e prodotti alimentari. Nella fascia superiore i 50 km invece si ha una predominanza di prodotti agricoli e selvicoltura.
I trasporti su percorsi superiori a 300 km sono effettuati più di metà delle volte su strada, nel 2018, secondo la rilevazione ISTAT il trasporto su strada ha movimentato 920,7 milioni di tonnellate di merci, le tonnellate-km movimentate sono state invece 124,9 miliardi.

La vicinanza dell'Italia a Paesi competitivi come Croazia, Slovenia e Romania ha fatto registrare un calo del volume di affari tra il 2008 e il 2016. Le motivazioni di tale flessione sono imputabili tra le altre al forte aumento del prezzo del carburante, che incide pesantemente sui costi di esercizio, contribuendo alla contrazione dei margini degli operatori in tale settore. Il passaggio a carburanti alternativi al diesel, come per esempio il gas naturale liquefatto, potrebbe essere una soluzione a tale questione, ma ciò implicherebbe di dover sostenere ingenti costi per l'acquisto di veicoli alimentati a GNL, che potrebbero essere ammortizzati in alcuni anni in virtù del minor costo del carburante. In Italia gli impianti di erogazione di GNL sono carenti e concentrati soprattutto nel Centro-Nord, pertanto il passaggio a un'alimentazione alternativa è attualmente accantonato.
A influire sui profitti delle imprese contribuiscono anche l'aumento del costo di manodopera e i rincari sui pedaggi, nonostante il blocco agli aumenti attutato dal Ministero dei Trasporti sul 90% della rete autostradale le limitazioni imposte ai veicoli pesanti che non soddisfano lo standard Euro 4 e il divieto di circolazione intermittente.[20]

[20]*https://www.repubblica.it/economia/rapporti/energitalia/mobilita/2019/08/05/news/trasporto_merci_in_italia_piu_costi_meno_competitivita_-232451032/*

Rete ferroviaria

La distinzione tra il trasporto passeggeri e il trasporto merci a media e lunga percorrenza è fondamentale al fine di analizzare in modo accurato il settore dei trasporti ferroviari. Il gruppo piccole e medie imprese ferroviarie ha trasportato quasi 16 milioni di passeggeri nell'ultimo anno e il percorso medio di ciascun passeggero è risultato pari a 28,6 km, contro i 62,2 km delle grandi imprese. Le merci trasportate dalla rete ferroviaria coincidono all'incirca con 94 milioni di tonnellate. Negli ultimi cinque anni i costi medi ferroviari mostrano una tendenza stabile e in diminuzione, grazie ai miglioramenti di efficienza operativa e all'effetto di misure incentivanti che contribuiscono a ridurre i costi di trasporto.

Il mercato del trasporto ferroviario italiano è molto più vicino all'assetto monopolistico a causa di diversi aspetti che limitano ancora la concorrenza delle imprese, quali per esempio l'elevata quota di capitale necessaria per avviare l'attività, oppure ulteriori barriere regolatorie che influiscono negativamente sull'efficienza del settore. Negli ultimi anni il Ministero delle Infrastrutture e dei Trasporti ha avviato una politica volta al rilancio del trasporto ferroviario delle merci al fine di accrescere la competitività nel Paese, allineandosi con gli standard europei. L'obiettivo è quello di individuare i principali deficit strutturali, i quali contribuiscono a frenare la competitività del trasporto ferroviario. Tra questi rientrano la lunghezza massima consentita dei treni, portata da 550m a 750m, il peso massimo trainabile e la sagoma delle gallerie che impediscono il trasporto dei trailer e dei container high cube.

L'indicatore più efficiente per descrivere il trasporto passeggeri è il passeggeri-km, ottenuto dal prodotto tra numero di viaggiatori che hanno usato il treno per i chilometri percorsi da ciascuno di essi.

PASSEGGERI TRASPORTATI	2016	2017	Variazioni % 2017/2016
Passeggeri	869.199.286	864.570.077	-0,5
Passeggeri-km	52.178.065	53.230.628	2,0

Fonte: Istat, Rilevazione del trasporto ferroviario (R)

Si osserva che nel 2017 l'indicatore è stato pari a 53.230 milioni di passeggeri-km, in aumento rispetto all'anno precedente del 2%, nonostante il numero di passeggeri complessivo sia diminuito lievemente, effetto di un incremento delle distanze percorse in treno.

Analizzando il trasporto delle merci si evidenzia un aumento delle tonnellate dell'1,4% rispetto al 2016 e un decremento delle tonnellate-chilometro, che passano da 22.712 a 22.334 migliaia, si segnala ancora una volta l'incremento delle distanze percorse in treno.

MERCI TRASPORTATE	2016	2017	Variazioni % 2017/2016
Tonnellate	92.948.907	94.287.070	1,4
Tonnellate-chilometro	22.712.340	22.334.637	-1,7

Fonte: Istat, Rilevazione del trasporto ferroviario (R)

Trasporti marittimi

Nel 2018 il trasporto marittimo di merci nei porti italiani si attesta intorno a 475 milioni di tonnellate, di cui il 64% riguarda merci esportate e la restante parte importate. Il settore dei trasporti marittimi è abbastanza diversificato e caratterizzato dall'interazione tra diversi attori suddivisi in categorie in relazione all'ambito geografico in cui operano:

- operatori marittimi come compagnie di navigazione e aziende che prestano servizi nautici. Il traffico di container e rinfuse liquide è organizzato principalmente in modo oligopolista, mentre la concessione di servizi tecnici è strutturata in modo monopolistico;

- operatori portuali che si occupano dello sbarco e imbarco, stoccaggio e lavorazione delle merci. Gli operatori compiono le loro attività in regime di concessione per alcune infrastrutture portuali o in regime di concorrenza con altri prestatori di servizi;

- operatori terrestri che si occupano del traporto del recupero delle merci e del trasporto verso la destinazione finale.

Tra i principali porti per la movimentazione di merci si elencano il porto di Trieste e quello di Genova, che insieme ricoprono un quarto dei trasporti marittimi di merci in Italia. Il porto di Messina è invece il primo per quanto riguarda il trasporto di passeggeri, seguito da Reggio Calabria e Napoli. I porti caratterizzati invece da un'intensa attività di traffico con l'estero sono situati al Nord-Est e nelle Isole, questi movimentano ogni anno circa 270 milioni di tonnellate di merci. Nelle isole si concentra anche il traffico di prodotti petroliferi che rappresentano la merce più scambiata nel trasporto marittimo.

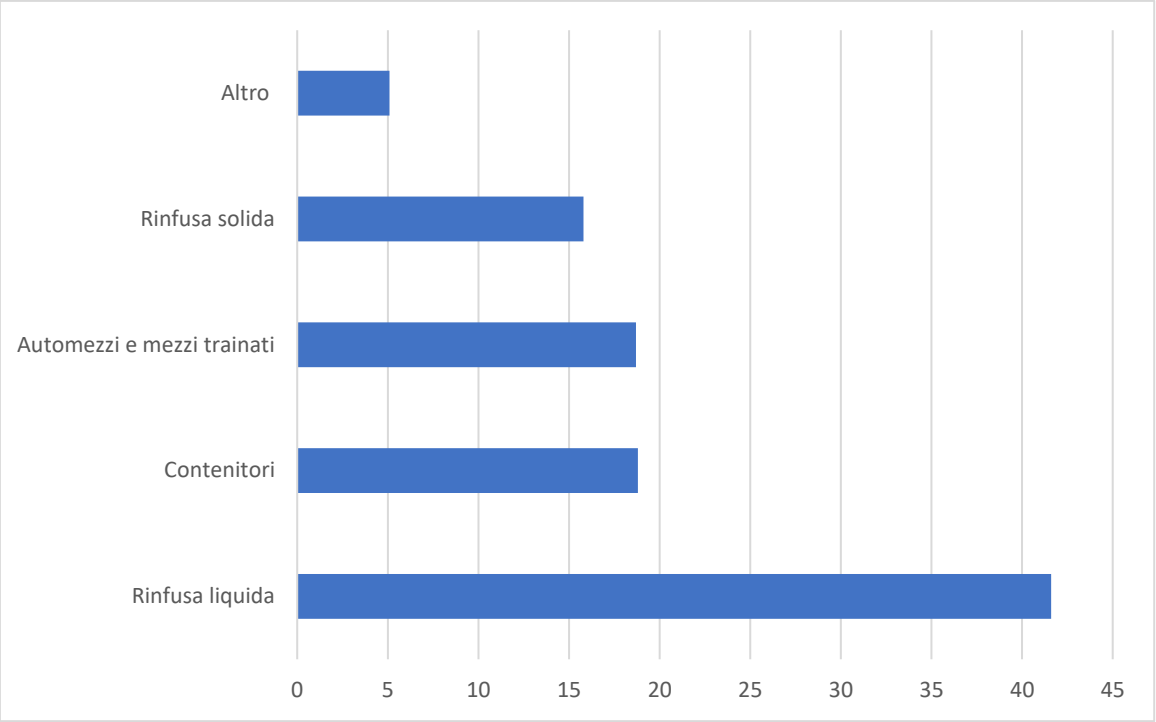


FIGURA 3.6 TIPOLOGIA MERCEOLOGICA TRASPORTATA VIA MARE

I costi del trasporto navale sono classificati per modalità di carico, in modo da considerare le particolarità tariffarie che caratterizzano ciascun segmento di mercato: dopo il 2014 i costi navali legati al trasporto di container hanno mostrato una tendenza in declino ma rimangono su livelli superiori rispetto al biennio della crisi 2008-09. I costi navali bulk (rinfusa liquida e solida) hanno mostrato una tendenza al rialzo, grazie a una politica attenta alla dismissione di navi obsolete, ponendosi su livelli non eccessivamente elevati ma comunque superiori rispetto al minimo storico. I costi di materie prime, tra cui rientrano soprattutto petrolio e derivati, si sono mantenuti stabili nel tempo, consentendo una maggiore profittabilità agli operatori del settore. I costi navali legati al trasporto di macchinari e impianti rimangono su livelli minimi grazie a una ricomposizione geografica degli scambi a favore dei Paesi più prossimi all' Italia.

Trasporti aerei

Il trasporto aereo in Italia costituisce l'1,8% della produzione e dell'occupazione nazionale, rappresentato da 193 imprese attive nel 2019 che fatturano oltre 9 miliardi di euro. Tra il 2010 e il 2018 il numero di passeggeri trasportati in Italia è aumentato del 33%, in particolare nell'ultimo anno i passeggeri transitati nei 39 scali italiani sono stati 193 milioni, con un incremento dell'1,3% rispetto al 2018 per l'aeroporto di Fiumicino e del 16,7% per l'aeroporto di Malpensa.

Variabili e indicatori	Trasporto aereo
Numero di imprese	193
Fatturato (migliaia di euro)	9.351.319
Valore della produzione (migliaia di euro)	9.998.527
Valore aggiunto (migliaia di euro)	1.922.027
Lavoratori dipendenti	19.430

La diffusione del Covid-19 ha interrotto quasi completamente i trasporti nel 2020 ma il comparto che ha subito maggiormente è il trasporto aereo di passeggeri. Le restrizioni applicate per contenere l'avanzamento dell'epidemia hanno azzerato la possibilità di volare, limitandola a ragioni lavorative o di salute. L'analisi dei dati Istat che monitorano il volume di passeggeri in Italia evidenzia un calo di oltre l'85% tra il mese febbraio e quello di marzo, tradendo tutte le aspettative di raggiungere circa 17 milioni di passeggeri per il periodo preso in analisi.

Gli aeroporti europei invece nel mese di marzo hanno registrato una contrazione dei traffici del 59,5% rispetto allo stesso mese dell'anno precedente, con una perdita di 106 milioni di passeggeri. Per avere un termine di paragone, durante la crisi finanziaria del 2009 sono stati necessari 12 mesi per registrare una perdita delle stesse dimensioni.

Con la fine dei voli di rimpatrio, necessari a riportare i cittadini europei nei loro Paesi di origine, il traffico aereo si è limitato al trasporto di merci o servizi sanitari e di altro tipo.

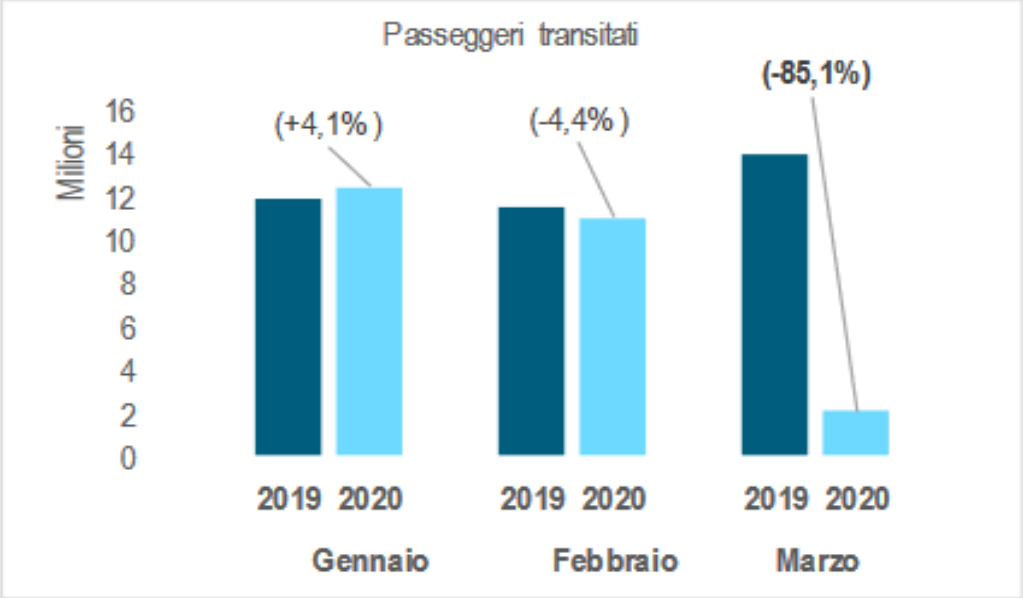


FIGURA 3.7-VOLUME DI PASSEGGERI TRASPORTATO

Con riferimento alla distribuzione dei flussi di passeggeri l'Asia è stata la zona nella quale nel 2019 si è concentrata la quota più alta di traffico, seguita da Europa e Nord America. È significativo segnalare come, a distanza di un anno, la densità di distribuzione sia diminuita notevolmente e i collegamenti tra alcuni Paesi si siano ridotti o annullati.

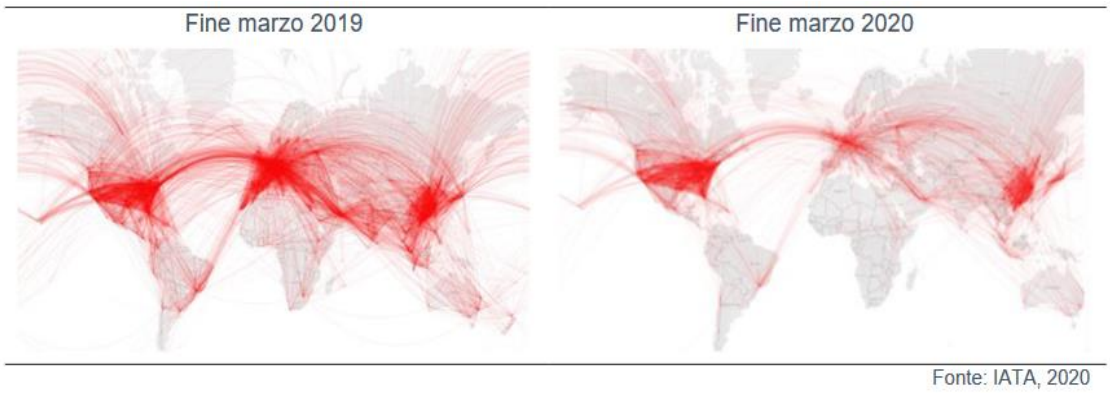
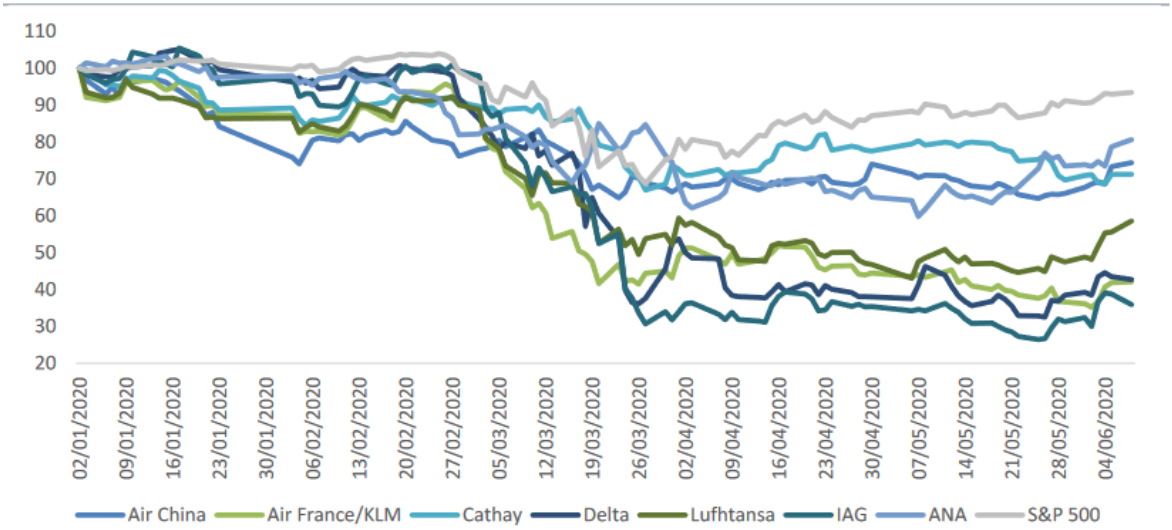


FIGURA 3.8-DISTRIBUZIONE FLUSSO PASSEGGERI

La contrazione delle attività, congiuntamente alle pessime previsioni di ripresa del settore in un breve orizzonte temporale, ha influito negativamente sull'andamento azionario delle principali compagnie aeree internazionali quotate. Dall'inizio del 2020 ai primi giorni di giugno si sono registrate contrazioni comprese tra il 20% di ANA e il 65% del Gruppo IAG, sensibilmente peggiori rispetto all'andamento dell'indice S&P 500 che ha perso meno del 10%.



Fonte: Thomson Reuters, 2020

FIGURA 3.9 -ANDAMENTO AZIONARIO COMPAGNIE AEREE NEL 2020

Il settore del trasporto aereo ha già attraversato periodi di crisi, ma è stato in grado di reagire agli shock in tempi piuttosto contenuti, intraprendendo una scalata di crescita moderata nelle fasi successive. Tuttavia, il brusco calo della domanda, la lenta ripresa della fiducia dei consumatori e l'imposizione di stringenti misure di sicurezza, sembrano rimandare inevitabilmente alle peculiarità della crisi statunitense legata agli attentati dell'11 settembre 2001. In quel caso sono stati necessari circa tre anni e mezzo per recuperare i volumi di traffico precedenti all'attacco alle Torri Gemelle, e oltre cinque per ristabilire lo stesso livello di redditività del settore.

A ridosso della crisi del 2001, le principali compagnie aeree americane hanno registrato perdite per oltre 9 miliardi di dollari; cifra che è cresciuta ulteriormente fino a raggiungere circa 40 miliardi di dollari nel 2007. La perdita di redditività ha dato il via a una serie di eventi senza precedenti e numerosi cambiamenti nel panorama mondiale: il Governo americano ha introdotto un pacchetto di sostegno da circa 50 miliardi di euro che ha permesso alle compagnie aeree di procedere alla ristrutturazione, riducendo la posizione debitoria e riorganizzando l'operatività. In aggiunta alle procedure fallimentari e alle misure di salvataggio attuate il settore dei trasporti aerei ha subito ulteriori trasformazioni, che ne hanno ridefinito l'assetto, attraverso intensi processi di fusione e acquisizione.

Il mercato delle compagnie aeree si presenta molto concentrato, con poche aziende di grandi dimensioni, che negli ultimi dieci anni hanno consolidato i loro bilanci e la loro attività, e un numero più ampio di imprese meno solide. La struttura dei costi delle compagnie aeree è costituita per il 49% da costi fissi e semi fissi e per il 51% da costi variabili. Ad incidere maggiormente sui costi variabili sono soprattutto le spese legate al carburante, mentre i costi di equipaggio e quelli legati a manutenzione e riparazione costituiscono insieme il 20% dei costi fissi.

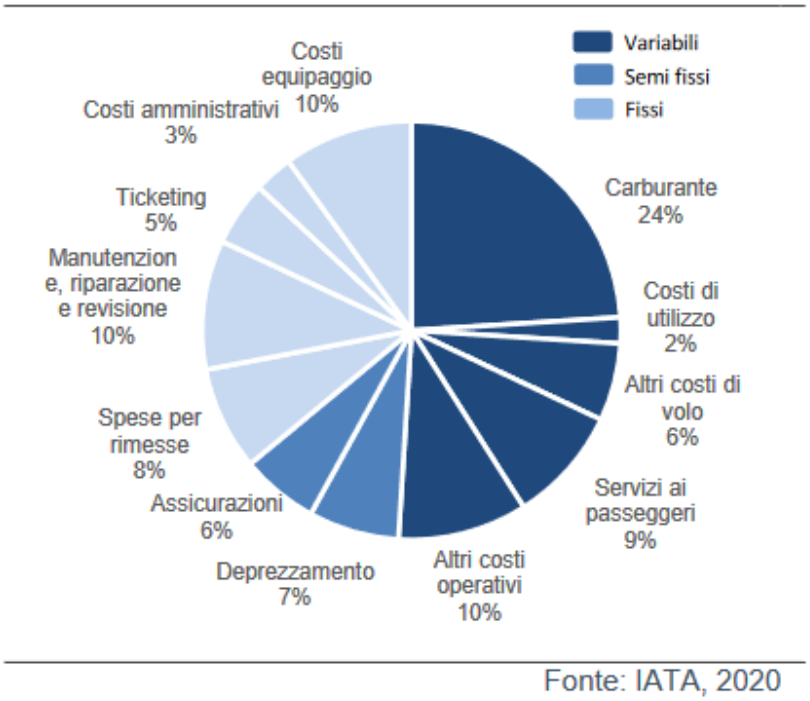
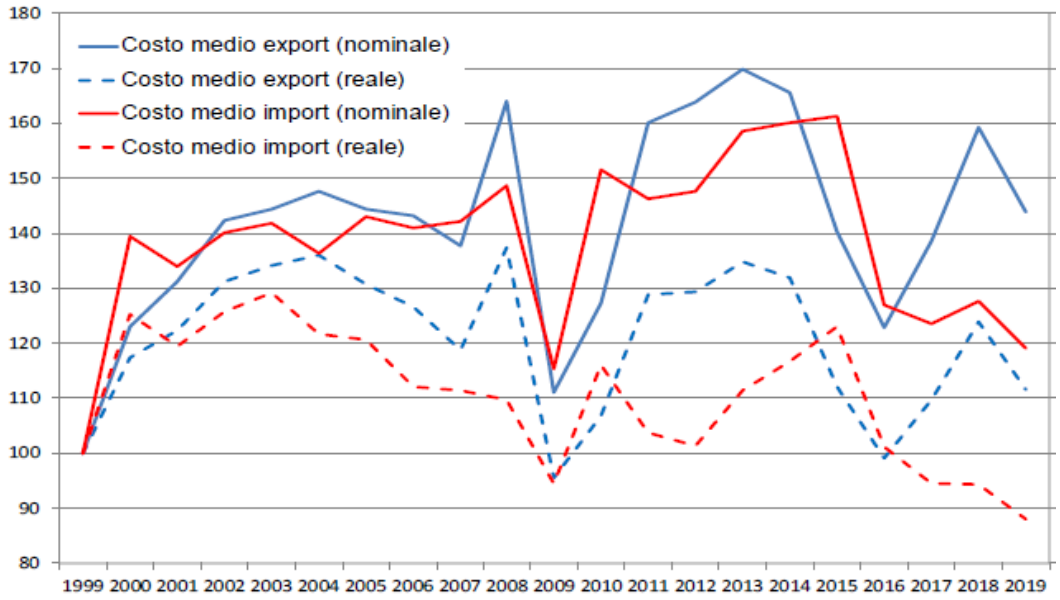


FIGURA 3.10-RIPARTIZIONE DEI COSTI DELLE COMPAGNIE AEREE

Nell'ultimo decennio i costi del trasporto aereo hanno registrato un andamento altalenante, i costi dell'esportazione sono diminuiti in maniera significativa, spesso con tassi intorno al 10%, e attualmente seguono una tendenza in diminuzione a seguito di un incremento durato un biennio. I costi di importazione invece hanno subito un calo drastico già a partire dalla fine del 2014. Il traffico merci aereo nel 2019 ha iniziato a ridursi, come conseguenza del ridimensionamento del commercio mondiale, dovuto alle tensioni tra Stati Uniti e Cina e al ritorno a misure protezionistiche. Nel 2019 si è registrata una riduzione del 3,3% dei volumi globali di trasporto merci via aereo; si tratta della prima variazione negativa dal 2012 e del valore più basso dalla crisi del 2008.

Costi medi del trasporto aereo (1)

(medie ponderate per flusso, indici dei costi misurati in euro/ton, 1999=100)



A seguito delle restrizioni imposte dai Governi, scaturite dalla necessità di dover applicare misure operative e sanitarie di sicurezza idonee a rispettare nuovi standard di distanziamento sociale, sarà necessario effettuare ingenti investimenti per una diversa configurazione degli aeromobili e degli spazi aeroportuali. Tali mutamenti si rifletteranno inevitabilmente sia nel segmento a monte dei produttori di aeromobili sia in quello a valle delle gestioni aeroportuali.

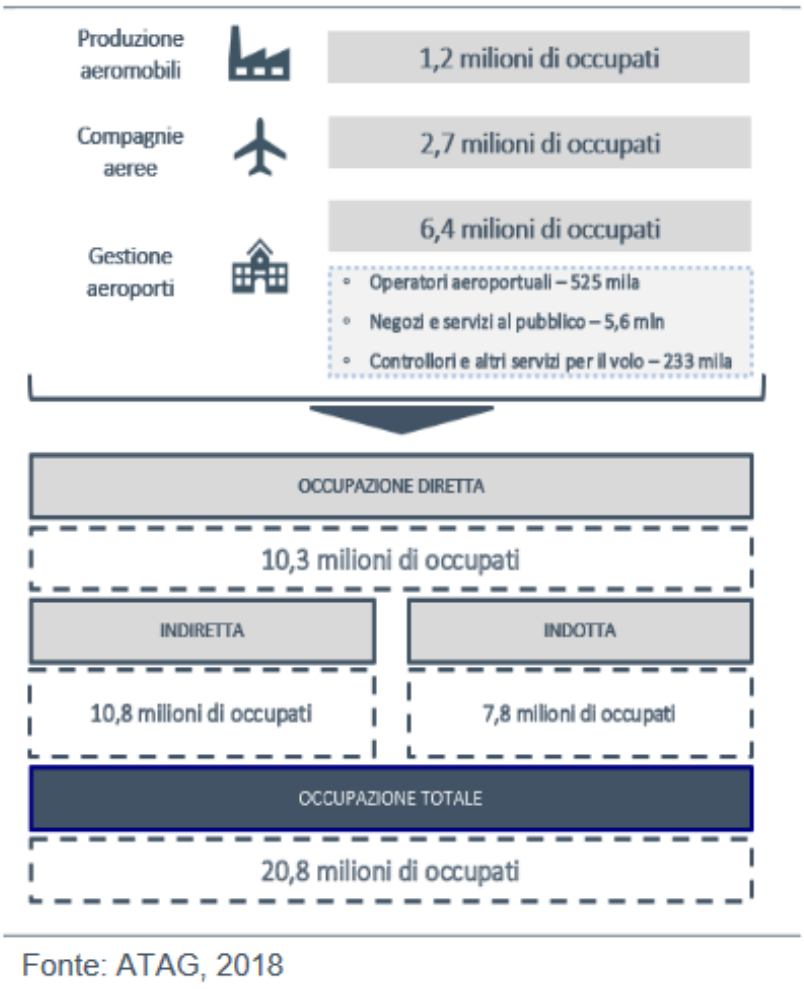


FIGURA 3.11 LA FILIERA DEL TRASPORTO AEREO

A livello mondiale, nel segmento della filiera che produce aeromobili, sono presenti due grandi aziende: Boeing e Airbus, le quali coprono quasi l'intero mercato e hanno un ruolo importante per le catene di fornitura globali; coinvolgono anche compagnie di altri settori produttivi come Rolls-Royce per quanto riguarda i motori, o l'italiana Leonardo per le aerostrutture. Il segmento a valle invece è costituito da diversi comparti e offre lavoro a oltre sei milioni di persone nel mondo, comprende non solo gli operatori aeroportuali ma anche controllori, operatori di volo e l'insieme di attività commerciali e servizi offerti al pubblico a ridosso e a seguito del viaggio. La chiusura di gran parte degli aeroporti per il trasporto di passeggeri a partire da marzo 2020 ha avuto gravi conseguenze sull'intera filiera, ma anche sulle attività a essa connesse.

Capitolo 4: Applicazione dei modelli di scoring

Nei paragrafi successivi verranno esposti i risultati ottenuti dall'applicazione dei modelli Logit e Support Vector Machines a un campione rappresentativo di aziende appartenenti al settore dei *trasporti.* Per mettere in atto questo studio sono stati utilizzati i dati di bilancio di 32.476 società, estratti dalla banca dati AIDA (Analisi Informatizzata delle Aziende Italiane), prendendo in considerazione il periodo che va dal 2008 al 2019.

La selezione del campione è stata effettuata in base al codice alfanumerico ATECO, che identifica un'attività economica. Il codice è composto da 6 cifre, di cui le prime due identificano le categorie dei settori: il 49 è associato al trasporto terrestre e mediante condotte, 50 è utilizzato per il trasporto marittimo e per vie d'acqua, 51 per il trasporto aereo.

Il campione di aziende appartenente al settore dei trasporti è stato selezionato perché si presta molto bene a descrivere la situazione patrimoniale di società sane, anomale e in default. Infatti, un numero rilevante di aziende appartenenti a tale settore ha risentito fortemente degli shock finanziari globali che si sono susseguiti a partire dal 2008.

Il campione è costituito quasi esclusivamente da società di capitali e di persone, che insieme costituiscono oltre il 95% delle osservazioni.

Forma giuridica	Numero di imprese	Peso percentuale
Società di capitali	25.400	78,2%
Società di persone	5.985	18,4%
Società cooperative	142	0,4%
Società consortili	937	2,8%
Società estere	12	0,1%

Dalla banca dati AIDA sono stati esportati in un file Excel i dati anagrafici: ragione sociale, anno di costituzione, regione e provincia di appartenenza, forma e stato giuridico, ed eventuali procedure in corso (liquidazione, fallimento, cancellazione, scioglimento, scissione ...); tutti i dati di bilancio di stato patrimoniale e conto economico, indicatori di solvibilità, redditività ed equilibrio economico, liquidità e struttura patrimoniale.

I bilanci delle società sono stati raggruppati per anno crescente e ragione sociale, associando un numero a ogni osservazione e a ciascuna azienda oggetto di analisi. Una prima pulizia dei dati di bilancio è avvenuta attraverso l'esclusione di tutte le società riportanti attivo nullo, e la sostituzione delle voci di bilancio contrassegnate con il valore n.d. con lo zero.

In seguito, si è assegnato un flag a ciascuna di esse per poterne descrivere lo stato:

- **Flag= 0** società sana**;**
- **Flag=1** società anomala in concordato preventivo, chiusa per fallimento, liquidazione giudiziaria, stato di insolvenza, bancarotta;
- **Flag=2** società sana in condizioni particolari**:** liquidazione volontaria, scioglimento, cessazione;
- **Flag 3** fusione, scissione, trasferimento all'estero;
- **Flag 4** cancellata dal registro impresa, trasferimento in altra provincia, cancellata d'ufficio, cessata.

Poiché il numero delle società con flag=0 è molto più grande di quello con flag=1, si è deciso di includere all'interno dell'analisi anche i bilanci di società a cui è stato assegnato flag=2. L'assegnazione dei flag è avvenuta secondo due distinti criteri: se l'azienda ha subito uno degli eventi presenti sotto la voce di flag=1 e flag=2, si è assegnato valore uguale a 1 per la classificazione della società come anomala. In accordo con il primo criterio (stima del **modello per società**), il valore 1 è stato assegnato all'anno di bilancio in cui si è verificato l'evento di default, contrassegnando con il valore 1 anche tutti gli anni precedenti e successivi all'evento; il secondo criterio di classificazione (stima del **modello per anno**) prevede l'assegnazione del valore 1 solo in corrispondenza dell'anno di bilancio in cui si verifica l'anomalia, mentre tutti gli altri anni sono considerati sani.

Tutti i bilanci delle società senza eventi sono considerati sani, quindi i flag per società e per anno sono inizializzati con valore pari a zero.
Per l'elaborazione dei modelli di credit scoring sono stati presi in considerazione solo alcuni dei dati di bilancio e degli indicatori, estratti dalla banca dati AIDA, tra cui:

- Totale attivo
- Patrimonio netto
- Valore della produzione
- Utile/Perdita
- Ebitda
- Capitale circolante netto
- Margine di struttura
- Flusso di cassa
- Indicatori di redditività
- Indicatori di produttività
- Indicatori di liquidità e struttura finanziaria

Diversi indicatori presentano valore nullo, derivante dall'assenza di denominatore, per tal motivo prima di procedere con l'elaborazione dei dati è stato necessario effettuare un lavoro preliminare di correzione, inserendo delle colonne che presentano flag uguale a 1 nei casi in cui *valore della produzione*=0, *ebitda* nullo o negativo, *patrimonio netto* nullo o negativo.

Per non compromettere i risultati dei modelli oggetto di analisi, includendo all'interno dei modelli valori troppo distanti tra loro, gli outliers sono stati gestiti calcolando il quinto e novantacinquesimo percentile di ciascun indicatore, per poi successivamente allineare al quinto percentile i valori inferiori e al novantacinquesimo i valori superiori.

Uno degli aspetti fondamentali da tenere in considerazione, per ottimizzare il risultato dei modelli, è la scelta delle variabili. Pertanto, si è calcolata la correlazione tra ogni coppia di indicatori per valutare il grado di sovrapposizione dei segnali: avere ben chiara la correlazione tra indicatori è utile per evitare di inserire all'interno dei modelli variabili fortemente correlate, in quanto queste potrebbero non apportare nessun miglioramento o addirittura ridurre la performance globale del modello.

Correlazione tra indicatori	TOTALE ATTIVO migl EUR	TOTALE PATRIMONIO NETTO migl EUR	TOT. VAL. DELLA PRODUZIONE migl EUR	UTILE/PERDITA DI ESERCIZIO migl EUR	EBITDA migl EUR	Capitale circolante netto migl EUR	Margine di struttura migl EUR	Flusso di cassa di gestione migl EUR	Indice di liquidità	Indice corrente	Indice di indebitam. a breve %	Indice di indebitam. a lungo %	copertura delle immob. (patrimoniale) %
EBITDA/Vendite (%) %	0.01648	0.01179	0.0099	0.04045	0.04093	-0.01135	-0.023	0.03859	0.18103	0.17694	-0.14244	0.1689	0.05854
Redditività del totale attivo (ROA) (%) %	-0.00562	-0.00441	-0.0037	0.02334	0.00916	0.00156	0.00452	0.00792	0.2449	0.24869	0.05078	-0.03831	-0.07624
Redditività delle vendite (ROS) (%) %	0.00857	0.01067	0.00409	0.02935	0.01632	0.00402	-0.00265	0.01755	0.21078	0.21525	-0.05029	0.06661	-0.01153
Redditività del capitale proprio (ROE)													

FIGURA 4.1- CORRELAZIONE TRA INDICATORI

Inoltre, si è calcolata la correlazione tra indicatore e flag per avere un'indicazione sul segno che devono assumere le variabili di bilancio:

- segno negativo se al crescere dell'indicatore diminuisce la probabilità di default;
- segno positivo se al crescere dell'indicatore aumenta la probabilità di default.

	Indice di liquidità	Indice corrente	Indice di indebitam. a breve	Indice di indebitam. a lungo	Indice di copertura delle immob.	Rapporto di indebitamento
*Correlazione flag per **società***	-0.0896439	-0.103580	0.0307978	**-0.02844**	-0.07845	0.013511
Accuracy	0.20807550	0.211135	0.08356186		0.297658	0.010291
*Correlazione flag per **anno***	-0.0180223	-0.022531	0.0026712	**-0.00148**	-0.03016	**-0.0161**
Accuracy	0.2501679	0.260571	0.1049983		0.490455	

FIGURA 4.2-CORRELAZIONE TRA INDICATORI E FLAG

Le variabili che presentano segno economico errato sono state evidenziate in rosso (Figura 4.2) ed escluse del dataset utilizzato per i modelli. Per selezionare invece gli indicatori più adatti è stata calcolata l'accuracy di ciascun indicatore, impostando un file Excel come segue:

- si inseriscono rispettivamente nella prima, seconda e terza colonna del file il numero di osservazione, il valore dell'indicatore di cui si vuole conoscere la precisione, e il flag associato all'osservazione;

- si utilizza una colonna per il confronto tra le osservazioni di bilancio e il modello perfetto, ottenuto inserendo solo valori pari a 0 e 1, in modo che nelle ultime posizioni siano presenti tanti flag uguali a 1 quante sono le osservazioni anomale;

- si ordinano contemporaneamente le prime tre colonne: dal maggiore al minore le variabili con segno economico positivo, e dal minore al maggiore quelle con segno economico negativo. Dopo aver calcolato il rapporto tra [% anomale indicatore - % totale] e [% anomale modello perfetto - % totale], si ottiene l'accuracy di ciascun indicatore.

Applicazione del modello logistico

Il modello Logit è stato sviluppato utilizzando il software statistico R: si tratta di un software opensource che fornisce un'ampia varietà di tecniche che possono essere implementate attraverso l'utilizzo di pacchetti, in base alle esigenze dell'utente. In particolar modo, la regressione logistica è stata sviluppata attraverso la funzione glm().

L'obiettivo dell'analisi è stimare la probabilità di default di un'impresa, sfruttando i dati e gli indicatori di bilancio. Per poter sviluppare il modello logistico, bisogna importare su R un file Excel in formato CSV contenente tutte le osservazioni di bilancio e i flag, associati a ogni società, attraverso il comando:

- Index <- read.csv2("Indicatori.csv")

Successivamente, si crea l'oggetto *model* che contiene i risultati della procedura glm sui dati *Index.* L'opzione family=binomial produce un modello di regressione logistica.

A seguito di questi comandi, si visualizzano i risultati della regressione in termini di coefficienti, intercetta e significatività di ciascuna variabile inserita:

- model <- glm (flag~indicatore1+indicatore2+…+indicatore n, data=Index, family="binomial")

- summary(model)

Per valutare l'accuratezza del modello logistico si utilizza la stessa tecnica adottata per la valutazione dell'accuracy degli indicatori: il Logit in questo caso è ottenuto dalla combinazione lineare degli indicatori utilizzati in R:

$$\textbf{\textit{Logit}} = intercetta + indicatore1 \cdot \beta_1 + indicatore2. \cdot \beta_2 + \ldots$$

Partendo dall'indicatore che presenta l'accuracy più elevata per ogni famiglia di indicatori si aggiungono a ogni iterazione nuovi indicatori al modello. Il processo di inserimento termina quando l'accuracy non cresce più, oppure quando uno o più indicatori già inclusi diventano non significativi.

Troppe caratteristiche possono influire negativamente sulle prestazioni del modello, sia perché alcune di esse possono essere tra loro positivamente correlate, sia perché potrebbero aggiungere poche informazioni al dataset. Le procedure descritte sono state ripetute sia per il modello flag per società sia per il modello con flag per anno. Di seguito si descrivono i risultati più rilevanti ottenuti per il modello Logit.

Logit Primo modello – Flag per società

Il modello con flag per società che ha ottenuto l'accuracy più elevata è costituito da 9 indicatori e da 3 variabili dummy (*ebitda, patrimonio netto* e *valore della produzione* nulli o negativi). Prima di giungere a tale risultato, sono stati effettuati diversi tentativi, variando per ognuno di essi la famiglia di indicatori da cui iniziare l'algoritmo. È fondamentale inserire all'interno del modello solo variabili significative (significatività del 99% contrassegnata da 3 asterischi) e con il corretto segno economico. L'accuracy più alta raggiunta per il modello Logit con flag per società è del **45,53%**:

```
Call:
glm(formula = flag ~ ebitda + vp + pn + X20. + X16. + X18. +
    X11. + X13. + X19. + X14. + X29 + X30, family = "binomial",
    data = Index)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6970  -0.5352  -0.4287  -0.3530   2.7933

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.470e+00  4.436e-02 -55.682  < 2e-16 ***
ebitda       6.584e-01  1.971e-02  33.402  < 2e-16 ***
vp           3.224e-01  2.894e-02  11.138  < 2e-16 ***
pn           1.243e+00  2.268e-02  54.817  < 2e-16 ***
X20.        -1.250e-01  1.018e-02 -12.278  < 2e-16 ***
X16.         1.274e-01  6.508e-03  19.578  < 2e-16 ***
X18.        -1.152e-01  1.526e-02  -7.548 4.41e-14 ***
X11.         5.806e-01  4.134e-02  14.045  < 2e-16 ***
X13.        -3.790e-02  7.175e-03  -5.282 1.28e-07 ***
X19.        -1.001e-01  1.630e-02  -6.137 8.41e-10 ***
X14.         1.682e-02  6.206e-04  27.102  < 2e-16 ***
X29         -2.778e-02  7.392e-03  -3.759 0.000171 ***
X30         -6.973e-04  5.443e-05 -12.813  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le variabili utilizzate per il modello sono le seguenti:

Indicatore	Variabile
X20	Rotazione capitale circ. lordo
X16	Oneri finanziari su fatturato
X18	Grado di indipendenza da terzi
X11	Indice di indebitamento a breve
X13	Indice di copertura delle immob.
X19	Rotazione capitale investito
X14	Rapporto di indebitamento
X29	Indice corrente*Indice cop. immob.
X30	ROS*Indice indip. Finanziaria
ebitda	Dummy Ebitda nullo o negativo
vp	Dummy Valore Produzione nullo
pn	Dummy Pat. Netto nullo o negativo

Indicatori della gestione corrente

- *Rotazione capitale circolante lordo* = $\frac{Ricavi\ vendite\ e\ prestazioni}{Attivo\ circolante}$

Verifica la velocità di trasformazione in liquidità dell'attivo circolante e dei flussi finanziari legati ai cicli di trasformazione/vendita. Più elevato è il valore del rapporto, maggiore è la capacità dell'azienda di reperire internamente i mezzi finanziari per far fronte ai pagamenti.

- Rotazione capitale investito = $\frac{Ricavi\ vendite\ e\ prestazioni}{Totale\ attivo}$

È un indicatore di efficienza importante per la gestione aziendale, segnala la capacità dell'impresa di trasformare il capitale investito in ricavi. Esso indica il numero di volte in cui il capitale investito ruota per effetto delle vendite (turnover). Un innalzamento di questo indice misura un miglioramento per l'azienda e una maggiore efficienza nella gestione del capitale.

Indicatori finanziari

- *Oneri finanziari su fatturato* = $\frac{Totale\ oneri\ finanziari}{Ricavi\ vendite\ e\ prestazioni+Altri\ ricavi} \cdot 100$

Indica l'assorbimento dei ricavi prodotti dagli oneri finanziari. Più elevato è il valore di questo indicatore, maggiore è la debolezza economica dell'azienda.

- *Grado indipendenza da terzi* = $\frac{Totale\ Patrimonio\ netto}{Totale\ Debiti}$

Misura l'autonomia dell'imprese nel coprire il totale dell'esposizione nei confronti dei terzi attraverso il proprio patrimonio netto. All'aumentare del valore dell'indicatore diminuisce l'instabilità finanziaria dell'impresa.

- Indice di indebitamento a breve = $\frac{Debiti\ a\ breve}{Debiti\ a\ breve+Debiti\ a\ oltre}$

Misura il livello di rischio legato a fonti di finanziamento esterne a breve termine, è calcolato come una media pesata tra i debiti a breve termine e la sommatoria dei debiti dell'impresa. Maggiore è il valore dell'indicatore, maggiore sarà la probabilità di default.

- Indice di copertura delle immobil. = $\frac{Totale\ immobilizzazioni\ materili}{Totale\ Patrimonio\ netto}$

L'indice di copertura delle immobilizzazioni monitora la struttura patrimoniale. Misura se la società è in grado di coprire le immobilizzazioni mediante fonti di finanziamento a lungo termine.

- Rapporto di indebitamento = $\frac{Totale\ attivo}{Totale\ Patrimonio\ netto}$

Dal punto di vista finanziario il rapporto di indebitamento riflette la dipendenza della gestione dall'indebitamento. Viene di solito indicato in termini unitari, il suo campo di variabilità spazia da zero a uno (totale attivo = patrimonio netto) e da uno in poi (attivo via via più elevato del patrimonio netto).

Variabili congiunte

- *Current Ratio · Indice di copertura immobilizzazioni*

$$Current\ Ratio = \frac{Attività\ correnti}{Passività\ correnti}$$

Il *current ratio*, o indice corrente a breve, è un indicatore utilizzato per misurare la capacità di un'impresa di convertire i propri beni tangibili in beni monetari per estinguere il proprio debito a breve termine. Il *current ratio* è utile per stabilire la situazione di liquidità dell'azienda.

- ROS · *Indice di Indipendenza finanaziaria*

$$ROS = \frac{Risultato\ operativo}{Ricavi\ vendite\ e\ prestazioni + Altri\ Ricavi}$$

Il ROS misura quanto rendono percentualmente le vendite. Esprime sinteticamente la capacità remunerativa dei ricavi tipici dell'azienda in esame: un miglioramento di questo rapporto implica una maggiore redditività delle vendite e margini più elevati.

Il modello *Logit* derivante da tali indicatori è:

Logit = -2.47 + 0.6584 · ebitda + 0.3224 · vp + 1.243 · pn - 0.125 · Rotaz.cap.circ.lordo +0.1274·Oneri fin.su fatt.-0.1152·Grado indip da terzi+0.5806·Indice indeb. a breve-0.0379·Indice copertura immob.-0.1001·Rotaz. capitale inv.+0.01682·Rapporto indebit.-0.02778·(Indice corrente·Indice cop.Immob) -0.0006973·(ROS·Indice indip. Fin.)

Logit Secondo Modello – Flag per anno

In accordo con il secondo modello, si attribuisce il flag di società anomala solo nell'anno in cui si verifica uno degli eventi riportati sotto la voce di flag=1 e flag=2, mentre si assegna flag pari a zero per tutti gli altri anni. Il secondo modello include sei indicatori e tre variabili dummy, raggiungendo una precisione del **64,44%**:

```
Call:
glm(formula = flag ~ ebitda + vp + pn + X20. + X16. + X19. +
    X13. + X17. + X24., family = "binomial", data = Index)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3742  -0.0841  -0.0571  -0.0431   4.1191

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.182151   0.141377 -36.655  < 2e-16 ***
ebitda       1.090203   0.098237  11.098  < 2e-16 ***
vp          -0.867337   0.124396  -6.972 3.12e-12 ***
pn           1.078934   0.118939   9.071  < 2e-16 ***
X20.        -0.298364   0.059352  -5.027 4.98e-07 ***
X16.         0.109262   0.025846   4.227 2.36e-05 ***
X19.        -0.327553   0.092182  -3.553  0.00038 ***
X13.        -0.172595   0.039086  -4.416 1.01e-05 ***
X17.        -0.007524   0.002503  -3.007  0.00264 **
X24.        -0.009411   0.002931  -3.210  0.00133 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10070.2  on 181629  degrees of freedom
Residual deviance:  8747.6  on 181620  degrees of freedom
AIC: 8767.6

Number of Fisher Scoring iterations: 9
```

Le variabili utilizzate sono le seguenti:

Indicatore	Variabile
X20	Rotazione capitale circ. lordo
X13	Indice di copertura delle immob.
X16	Oneri finanziari su fatturato
X19	Rotazione capitale investito
X17	Indice di indip.finanziaria
X24	ROE
ebitda	Dummy Ebitda
vp	Dummy Valore Produzione
pn	Dummy Patrim. Netto

In aggiunta alle variabili già elencate e descritte precedentemente, si menzionano:

- *Indice di indipendenza finanziaria* = $\frac{Totale\ Patrimonio\ netto}{Totale\ attivo}100$
- *ROE* = $\frac{Utile\ (perdita\ )netto}{Patrimonio\ netto}$

Il ROE (*Return on Equity*) è utilizzato per verificare il tasso di remunerazione del capitale di rischio, cioè quanto rende il capitale conferito dai soci all'azienda. L'indicatore valuta come il management è in grado di gestire i mezzi propri per aumentare gli utili aziendali. Il ROE non è solo determinato dalle scelte compiute nell'ambito della gestione caratteristica, ma anche dalle decisioni inerenti alla gestione finanziaria e patrimoniale.

Il modello Logit derivante da tali indicatori è:

Logit = -5.182151 +1.09·ebitda- 0.86·vp + 1.078·pn - 0.298·Rotaz.cap.circ. lordo +0.109·Oneri fin.su fatt. -0.1725·Indice copertura immob.-0.0075·Indice di indip.finanziaria – 0.09411

Applicazione del modello Support Vector Machines

Il modello SVM è stato sviluppato sfruttando il linguaggio di programmazione Python. In particolar modo, per lo sviluppo del modello SVM è stato utilizzato l'ambiente di sviluppo PyCharm, che integra diverse funzionalità utili nel development. PyCharm fornisce supporto in termini di assistenza e analisi della codifica, inoltre consente la facile installazione di pacchetti aggiuntivi e qualsiasi libreria richiesta appropriata per la propria piattaforma di elaborazione.

Prima di approcciarsi allo sviluppo del modello SVM, è stato necessario gestire lo sbilanciamento del dataset, in quanto la classe di società aventi bilanci anomali è molto inferiore rispetto alla classe di società senza anomalie. La gestione dello sbilanciamento è fondamentale in quanto a seguito dell'applicazione di un algoritmo di classificazione la classe maggiormente rappresentata avrà una buona accuratezza, mentre la capacità predittiva della classe meno rappresentata può essere molto inferiore, sebbene in genere siano le prestazioni della classe di minoranza ad essere più importanti.

Le modalità utilizzate per risolvere questo problema sono i sampling method, ossia, metodi di campionamento che permettono di modificare il dataset originale in modo da riequilibrarlo. Tra i metodi di ricampionamento si possono distinguere due categorie:

- ***Undersampling method***: prevedono la creazione di un sottoinsieme di dati ribilanciato, ottenuto eliminando alcuni campioni della classe più popolata. I vantaggi derivanti da tale metodologia sono la rapidità di elaborazione e la bassa probabilità di overfitting.
- ***Oversampling method***: questo metodo consiste nella creazione di un dataset in cui sono stati inserite delle copie della classe meno rappresentata o sono state creati nuovi valori. Poiché nessun valore è rimosso, non c'è perdita di informazione, ma i tempi di esecuzione dell'algoritmo possono diventare eccessivamente elevati.

NearMiss

Uno degli algoritmi più utilizzati per bilanciare il set di campionamento, appartenente alla categoria dell'Undersampling technique, è il *NearMiss*. Questo algoritmo permette di eliminare gli elementi appartenenti alla classe maggiormente rappresentata, finché non si raggiunge il rapporto di bilanciamento 1:1 tra le due classi. Per non incorrere nel rischio di perdita dati a seguito dell'applicazione l'algoritmo NearMiss opera in questo modo:

1. L'algoritmo trova la distanza tra tutti i campioni della classe maggiormente rappresentata e quella della classe meno popolata;
2. Successivamente vengono selezionati *n* elementi della classe più rappresentata, che hanno una relazione particolare con gli elementi della classe minore;
3. L'algoritmo continua finché il numero degli elementi appartenenti a ciascuna classe è bilanciato.

Per determinare gli elementi più vicini della classe maggiormente rappresentata si possono utilizzare tre tecniche diverse:

- NearMiss 1: i dati sono bilanciati calcolando la distanza minima media tra la distribuzione più grande e le tre distribuzioni più piccole più vicine.
- NearMiss 2: i dati vengono bilanciati calcolando la distanza minima media tra la distribuzione più grande e le tre distribuzioni più piccole più distanti.
- NearMiss 3: vengono considerate le istanze di classi più piccole e vengono memorizzati *m* vicini. Quindi viene presa la distanza tra questi e la distribuzione maggiore ed eliminata la distanza maggiore.[21]

[21] *https://analyticsindiamag.com/using-near-miss-algorithm-for-imbalanced-datasets/*

SMOTE

L'algoritmo SMOTE appartiene alla famiglia degli Oversampling method, questo metodo non si limita a replicare i campioni della classe meno rappresentata, ma ne introduce di nuovi, i quali sono ottenuti dall'interpolazione di vari elementi realizzati dalla classe che si vuole riprodurre. Anche in questo caso si sfrutta una funzione che tiene conto della distanza tra i vari punti del dataset.

SMOTE (tecnica di sovracampionamento minoritario sintetico o **Synthetic Minority Oversampling TEchnique**) è uno dei metodi di sovracampionamento più utilizzati per risolvere il problema dello sbilanciamento delle classi. SMOTE risolve questo problema andando a incrementare le osservazioni della classe minoritaria, in modo casuale.

"Nella pratica si sfrutta un algoritmo K-nearest neighbors per creare dati sintetici: si inizia scegliendo osservazioni casuali dalla classe meno popolata, successivamente vengono impostati i k vicini più prossimi a tali osservazioni. I dati sintetici sono creati tra i dati casuali e il vicino k più prossimo, selezionato casualmente tramite interpolazione lineare.

Per una data osservazione di minoranza x_i, viene generata una nuova osservazione "sintetica" interpolando tra uno dei k vicini più prossimi, x_{zi}:

$$x_{new} = x_i + \lambda\,(x_{zi} - x_i)$$

dove λ è un numero casuale compreso nell'intervallo [0,1]."[22]

Questa interpolazione crea un campione sulla linea tra x_i e x_{zi}. La procedura di sovracampionamento viene ripetuta finché i due campioni non sono bilanciati.

[22] *https://lorenzogovoni.com/algoritmo-smote/*

Elaborazione del modello SVM

Lo sviluppo del modello di credit-scoring SVM è avvenuto attraverso l'utilizzo della libreria open source *Scikit-learn*. Questa contiene algoritmi di classificazione, macchine a vettori di supporto, ed è progettata per operare con librerie NumPy.

La fase di sviluppo di SVM è stata preceduta dalla selezione della tecnica adatta per costruire un buon modello predittivo, a partire da un dataset fortemente sbilanciato. Innanzitutto, l'analisi del modello ha avuto inizio suddividendo il dataset iniziale, il quale comprende l'intero campione dei bilanci delle società appartenenti al settore dei Trasporti. La ripartizione del campione iniziale è stata possibile grazie all'utilizzo del pacchetto *train_test_split*, ottenendo un sottogruppo di training pari al 70%, e un sottogruppo utilizzato invece come test, pari al 30%. La procedura di suddivisione del test di addestramento è appropriata quando si dispone di un set di dati molto grande e si richiede una buona stima delle prestazioni del modello. Tutte le osservazioni sono state standardizzate attraverso StandardScaler: gli stimatori, infatti, possono comportarsi in modo errato se le singole feature non sono normalmente distribuite.

Per la manipolazione e l'analisi dei dati di bilancio contenuti all'interno del file Excel in formato CSV è stata caricata la libreria software Pandas. Di seguito sono riportati i comandi implementanti su PyCharm:

#importo i pacchetti

```
1. from sklearn.model_selection import train_test_split
2. from sklearn.preprocessing import StandardScaler
3. from sklearn.svm import LinearSVC, SVC
4.  import pandas as pd
```

Tra le librerie utilizzate vi è anche Linear SVC (*Support Vector Classifier*), che è in grado di individuare un iperpiano "best fit" che divide e classifica i dati, in modo da distinguere in modo chiaro i bilanci delle società anomale da quelli delle società sane.

Le operazioni descritte di seguito sono state ripetute sia per il Primo Modello che utilizza flag per società, sia per il Secondo Modello che utilizza flag per anno, sfruttando in prima battuta lo stesso set di indicatori individuato nel precedente modello Logit, per poi modellarlo al fine di comprendere come si modifica l'accuracy al variare del numero di feature introdotte. Si riportano di seguito, le stringhe che hanno permesso di ottenere la migliore performance predittiva del modello SVM, le quali saranno poi analizzate nel dettaglio:

#import dataset

```
dataset =pd.read_csv ("/Users/ASUS/Documents/Tesi/SVM/primoRFE.csv",delimiter= ";" )
X = dataset.drop("flag", axis=1)
Y = dataset["flag"]
from imblearn.under_sampling import NearMiss
nm = NearMiss()
X_res, Y_res = nm.fit_resample(X, Y)
X_train, X_test, Y_train, Y_test = train_test_split(X_res, Y_res, test_size=0.3, random_state=0)
scaler = StandardScaler()
X_train_std = scaler.fit_transform(X_train)
X_test_std = scaler.transform(X_test)
```

 # creazione della SVM e allineamento

```
svm = SVC(kernel="rbf")
svm.fit(X_train_std, Y_train)
```

#stampa dei risultati

```
print("Accuracy Train Set:", svm.score(X_train_std, Y_train))
print("Accuracy Test Set:", svm.score(X_test_std, Y_test))
print()
```

Il problema dello sbilanciamento del dataset è stato risolto applicando due differenti algoritmi: si è comparata la tecnica di Oversampling SMOTE con la tecnica di Undersampling NearMiss. A questo punto, si è applicato SVM ai due nuovi dataset con le tecniche di ribilanciamento. Fondamentale ai fini di raggiungere una migliore capacità predittiva è la scelta del kernel che si intende utilizzare: per determinare con quale delle due tecniche si ottiene il migliore risultato si è testata la capacità di ciascun algoritmo, applicando inizialmente un kernel lineare ad entrambi i modelli.

Si espongono di seguito i risultati ottenuti:

- *Undersampling Nearmiss:* l'algoritmo di sotto campionamento si è dimostrato essere quello con la maggiore capacità predittiva, per il Primo Modello (flag per società) si raggiunge un accuracy del 67,6%, mentre per il Secondo Modello (flag per anno) l'accuracy è pari al 95%.
- *Oversampling Smote:* l'algoritmo di sovra campionamento, invece, evidenzia dei risultati caratterizzati da una minore capacità di previsione, con accuracy pari al 62.8% per il Primo Modello e del 90% per il Secondo Modello.

La tecnica di sotto campionamento ha evidenziato risultati decisamente superiori, impiegando tempi di elaborazione ridotti rispetto all'Oversampling, in quanto risulta più efficiente ridurre il numero di società sane piuttosto che creare nuove istanze per pareggiare il numero di bilanci in default.

Alla luce di quanto appena emerso, il passo successivo è stato quello di applicare l'algoritmo NearMiss allo stesso campione, sostituendo il kernel lineare con quello RBF. Abbinare il kernel RBF alla tecnica NearMiss ha permesso di incrementare ulteriormente la capacità predittiva dei modelli, che con tale combinazione si attesta al 73% per il modello con flag per società e al 95,7% per il modello con flag per anno.

Selezione delle Feature

I modelli di classificazione appena descritti portano a risultati piuttosto soddisfacenti se paragonati con il Logit, ma modellando il dataset è ancora possibile incrementare la capacità predittiva. Per raggiungere tale obiettivo risulta fondamentale la fase di selezione delle feature: per la costruzione di un buon modello si possono infatti eliminare le variabili che non portano alcuna informazione aggiuntiva, al fine di raggiungere una maggiore accuratezza. Anche in questo caso si è preferito utilizzare due approcci differenti per poter confrontare i risultati ottenuti. Sono stati implementati due algoritmi di selezione delle variabili ottime: l'algoritmo Random forest e Recursive feature elimination, di cui si descrivono in seguito le modalità di funzionamento.

Random forest

Random forest è uno degli algoritmi di machine learning più popolari in quanto fornisce buone performance previsionali e basso overfitting. La tecnica Random forest prevede l'utilizzo di 400-1200 alberi decisionali, ognuno dei quali costruito secondo un'estrazione casuale di osservazioni dal dataset e un'estrazione casuale di feature. Per evitare di estrarre alberi correlati tra loro, ciascuno di essi non considera tutte le features e tutte le osservazioni, ma, ogni albero è una sequenza di domande Si/No basate su una singola o una combinazione di caratteristiche. Ad ogni nodo l'albero si divide in due rami, ognuno dei quali ospita osservazioni simili tra loro e diverse da quelle dell'altro. L'importanza di ciascuna variabile deriva da quanto è "puro" ciascun ramo. Per implementare su Python l'algoritmo Random forest si utilizzano i seguenti comandi:

```
1. from sklearn.ensemble import RandomForestClassifier
2. from sklearn.feature_selection import SelectFromModel
3. sel = SelectFromModel(RandomForestClassifier())
4. sel.fit(X_train, Y_train)
5. sel.get_support()
6. selected_feat= X_train.columns[(sel.get_support())]
7. len(selected_feat)
```

Recursive feature elimination

Recursive feature elimination (RFE) o Eliminazione delle caratteristiche ricorsive è un algoritmo di selezione delle feature efficace nel selezionare quelle caratteristiche rilevanti nella previsione della variabile obiettivo. La RFE è una tecnica che rimuove in modo ricorsivo le variabili, che può sfruttare diversi strumenti per eseguire la selezione delle caratteristiche, classificabili in:

- **Metodo Wrapper**: complesso dal punto di vista computazionale e poco pratico in caso di ricerca esaustiva, mira a trovare il modello con le prestazioni migliori, attraverso la migliore combinazione possibile di caratteristiche.
- **Metodo Filtro**: permette di selezionare le feature da un dataset in modo indipendente, basandosi solo sulle caratteristiche di queste variabili. Questo è invece un metodo potente e semplice in quanto rimuove rapidamente le variabili che non apportano valore aggiunto al modello.
- **Metodo Embedded**: completa il processo di selezione delle feature all'interno della costruzione dell'algoritmo di apprendimento automatico stesso. Nello specifico, esegue la selezione delle caratteristiche durante l'addestramento del modello, motivo per cui tali algoritmi sono chiamati "incorporati".

In Scikit-Learn esistono due opzioni di configurazione per l'utilizzo di RFE**:** la scelta del numero di variabili da selezionare e la scelta dell'algoritmo per selezionare le feature. Il funzionamento di questo algoritmo può essere sintetizzato nel seguente modo:

1. Innanzitutto, adatta il modello al set di dati di allenamento [metodo fit()];
2. Registra le metriche di punteggio corrispondenti per il modello;
3. Determina quale caratteristica è la meno importante nel formulare previsioni sul dataset di test e la elimina;
4. Se il set di feature è composto da più di una variabile, si ritorna al passaggio 1 (con una feature in meno), altrimenti si prosegue al passaggio 5;
5. In fine, l'algoritmo seleziona il set di feature che fornisce la metrica di punteggio più alta.

Per l'applicazione di RFE si utilizzano i seguenti comandi:

```
1. rfe= RFE(model,n_features_to_select=5, step=1)
2. rfe.fit(X,Y)
3. for i in range (X.shape[1]):
4.  if rfe.support_[i] ==True:
        Print ('Colonna scelta' + str(i) + '('+str(X.columns[i])+'),
        punteggio: ' + str(rfe.ranking_[i]))
```

Sia per il *Primo modello* sia per il *Secondo modello* si è deciso di ridurre il numero di feature, scelte tra il set di variabili di bilancio e le tre variabili dummy, in modo da evitare la distorsione derivante da variabili correlate tra loro, che influisce negativamente sulla capacità predittiva del modello. Infatti, il numero di variabili economiche inserite all'interno di entrambi i modelli SVM è stato portato a sei.

Dopo aver applicato entrambi gli algoritmi di selezione delle feature, si evince che Recursive feature elimination fornisce risultati più soddisfacenti, utilizzando un dataset limitato rispetto a quello del modello Logit. Random forest e RFE sono stati applicati al Primo e Secondo modello, comparando i risultati ottenuti utilizzando un kernel lineare con quelli ottenuti sfruttando invece un kernel RBF.

SVM Primo modello - Flag per società

A seguito dell'applicazione dell'algoritmo *Random forest* al Modello per società, utilizzando un kernel lineare, emerge che l'accuracy è pari al 67.8%, attestandosi a livelli leggermente superiori a quelli ottenuti utilizzando l'intero dataset. Se invece si sostituisce il kernel lineare con quello RBF, la precisione del modello cresce fino al 72,03%. Le variabili selezionate dall'algoritmo Random forest sono le seguenti:

Indicatore	Variabile
X20	Rotazione capitale circ. lordo
X16	Oneri finanziari su fatturato
X13	Indice di copertura delle immob.
X19	Rotazione capitale investito
X14	Rapporto di indebitamento
X29	Indice corrente*Indice cop. immob.

Con l'applicazione dell'algoritmo *RFE*, emerge invece che utilizzando un kernel lineare, la capacità di previsione si attesta intorno al 58.2%, mentre questa cresce notevolmente fino a raggiungere il **74.5**% se si utilizza un *kernel RBF*. In definitiva, si evidenzia che l'accuracy più alta si raggiunge proprio in quest'ultimo caso in cui si implementa la tecnica di sotto campionamento NearMiss e algoritmo di selezione delle feature RFE. Le variabili utilizzate nel caso appena descritto sono le seguenti:

Indicatore	Variabile
X18	Grado di indipendenza da terzi
X11	Indice di indebitamento a breve
X19	Rotazione capitale investito
ebitda	Dummy Ebitda nullo o negativo
vp	Dummy Valore Produzione nullo
pn	Dummy Pat. Netto nullo o negativo

L'unica variabile in comune con l'algoritmo Random forest è la Rotazione del capitale investito, mentre in aggiunta sono presenti le tre variabili dummy per *ebitda, patrimonio netto* e *valore della produzione* nullo o negativo.

Python permette di visualizzare graficamente il modello appena descritto e tale visualizzazione è utile per chiarire come avviene la suddivisione delle società sane dalle anomale: i puntini in verde individuano i bilanci delle società flaggate con valore 0, mentre in rosso sono individuati i bilanci delle società anomale.

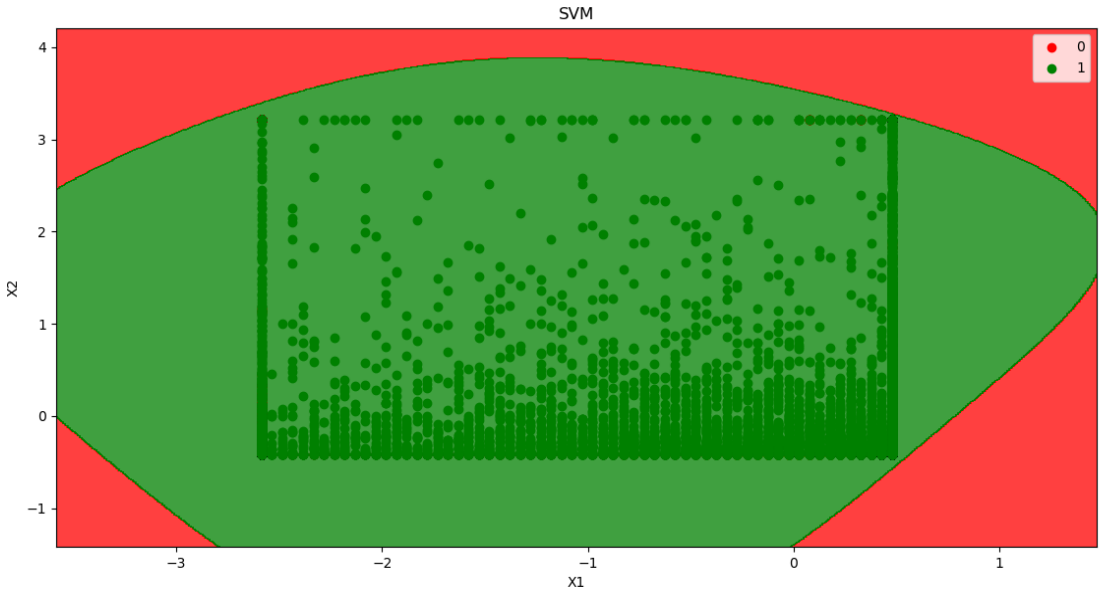
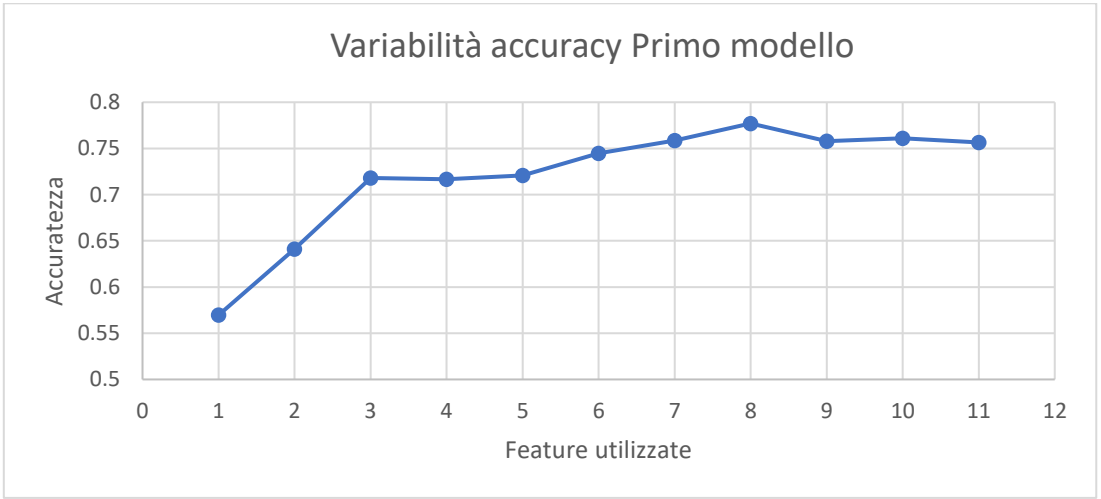


Figura 3.3-RAPPRESENTAZIONE GRAFICA SVM – MODELLO PER SOCIETA'

Per individuare il numero di feature che permette di ottenere il modello più performante, si è calcolata la variabilità dell'accuracy al variare del numero delle caratteristiche inserite all'interno del modello. Partendo da 1 indicatore, ad ogni iterazione è introdotta una feature in più, fino a calcolare l'accuratezza con tutte le feature. L'accuratezza più elevata si raggiunge considerando otto variabili di bilancio, ed è pari al **77.06%,** mentre l'aggiunta di ulteriori variabili causa una diminuzione della performance di previsione.



SVM Secondo modello - Flag per anno

Il secondo modello presenta una precisone che si attesta su livelli già elevati, ma anche in questo caso è possibile modellare le variabili per ottenere la massima precisione possibile. A seguito dell'applicazione dell'algoritmo *Random forest* con kernel lineare, si ottiene una capacità di previsione del 94.88%, che aumenta fino al 95.73% se si applica un kernel RBF. Le variabili in gioco selezionate dall'algoritmo sono le seguenti:

Indicatore	Variabile
X20	Rotazione capitale circ. lordo
X13	Indice di copertura delle immob.
X16	Oneri finanziari su fatturato
X19	Rotazione capitale investito
X17	Indice di indip.finanziaria
X24	ROE

Utilizzando l'algoritmo di selezione *RFE*, si ottiene un'accuracy pari al **95.94%** in caso di *kernel lineare* e del 95.1% in caso di kernel RBF. Il Modello che utilizza flag per anno mostra l'accuracy più elevata in assoluto quindi con la tecnica NearMiss e algoritmo RFE, con kernel lineare, impiegando le seguenti variabili:

Indicatore	Variabile
X20	Rotazione capitale circ. lordo
X19	Rotazione capitale investito
X17	Indice di indip.finanziaria
X24	ROE
ebitda	Dummy Ebitda
vp	Dummy Valore Produzione

Rispetto all'agoritmo Random forest, RFE utilizza le stesse feature ma sostituisce l'Indice di copertura delle immobilizzazioni e Oneri finanziari su fatturato con le variabili dummy per Ebitda e Valore della produzione, raggiungendo una capacità predittiva superiore.

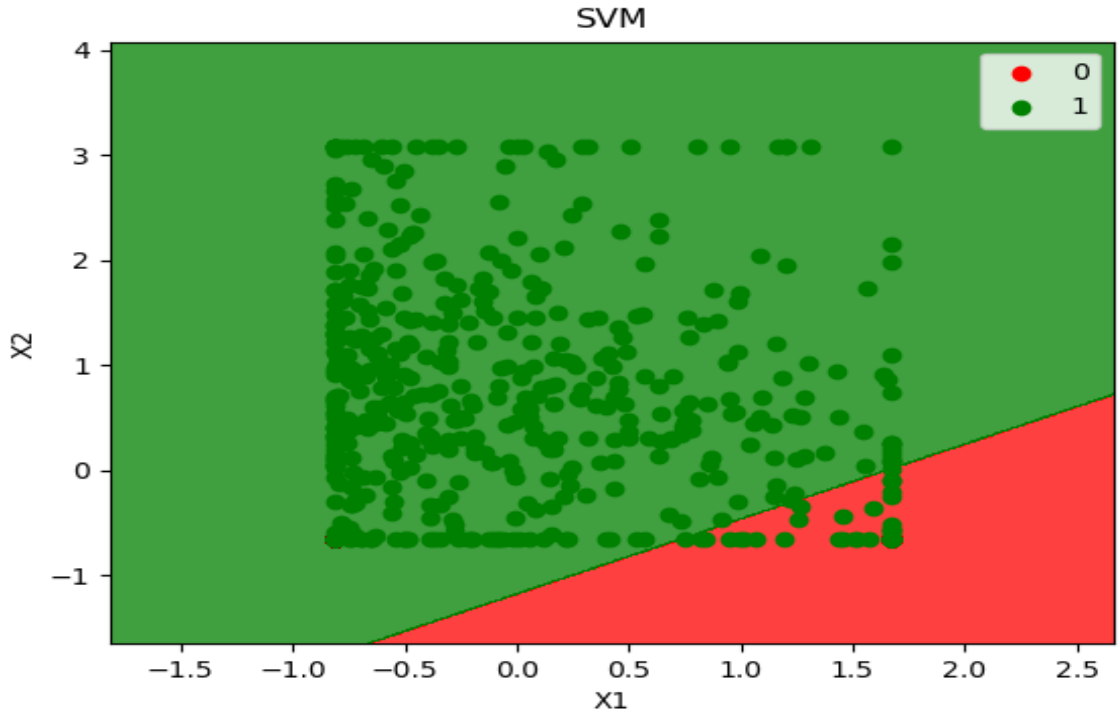
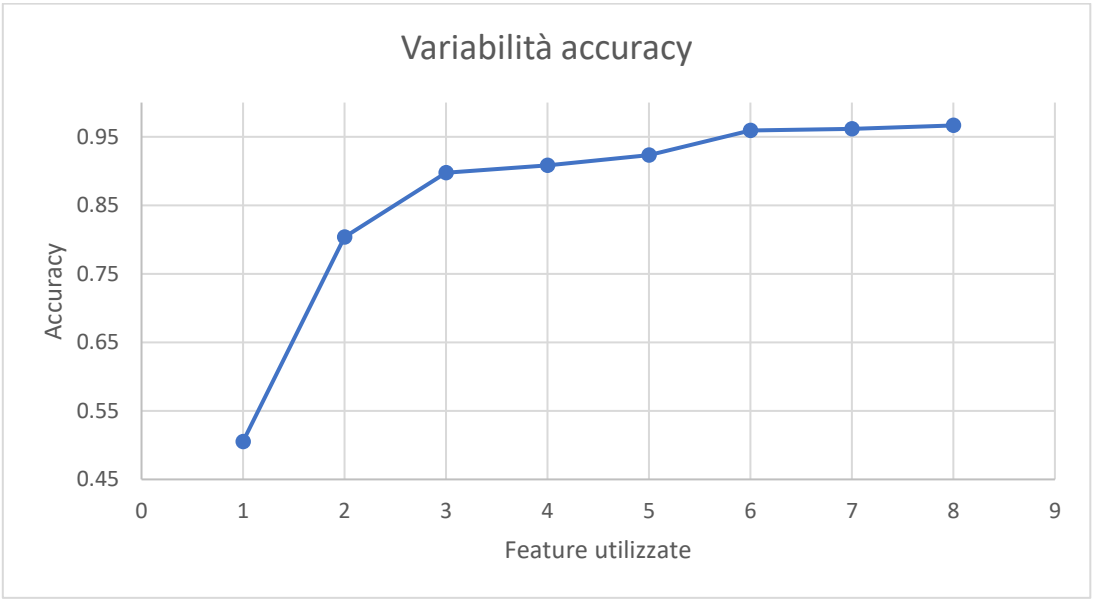


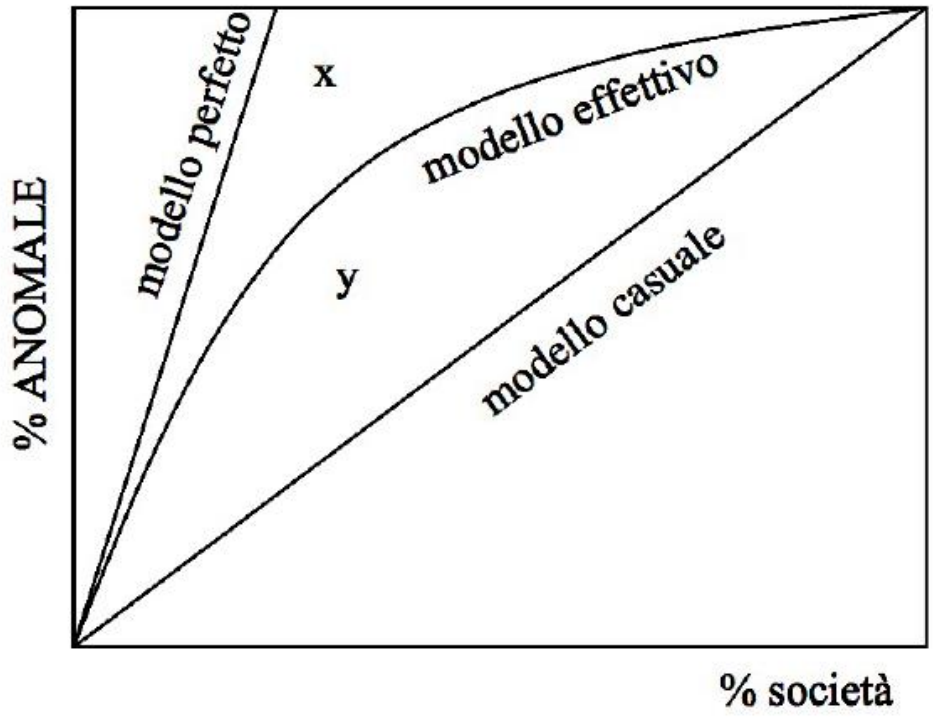
FIGURA 4.4-RAPPRESENTAZIONE GRAFICA SVM MODELLO PER ANNO

In questo secondo grafico si nota una maggiore capacità predittiva con solo quattro indicatori di bilancio. L'accuracy più alta si ottiene utilizzando otto feature, raggiungendo il **96.58%,** ma già con sei indicatori di bilancio si raggiunge un accuracy di circa il 96%.



Analisi delle performance e degli errori dei modelli

Per analizzare la performance dei modelli possono essere impiegate diverse tecniche che permettono di mettere in evidenza le società classificate in modo non corretto. Uno dei metodi impiegati per la valutazione della precisione del modello è l'accuracy, o curva di Gini (Power curve o Lorenz Curve).



Il modello perfetto identifica correttamente tutte le società anomale, mentre quello casuale attribuisce a tutte le imprese la stessa capacità di default, non diagnosticando alcuna informazione utile. Il coefficiente di Gini è calcolato come rapporto tra le osservazioni rientranti nella zona y e il totale delle osservazioni appartenenti alla zona x e y (valore compreso tra 0 e 1):

$$\textbf{Gini coefficient} = \frac{x}{x+y}$$

Di seguito sono riportate le curve di accuracy del modello che utilizza flag per società (Figura 4.5) e flag per anno (Figura 4.6), di cui si analizzeranno gli errori di classificazione.

La curva blu rappresenta il modello casuale, quella arancio il modello perfetto, mentre in grigio il modello reale:

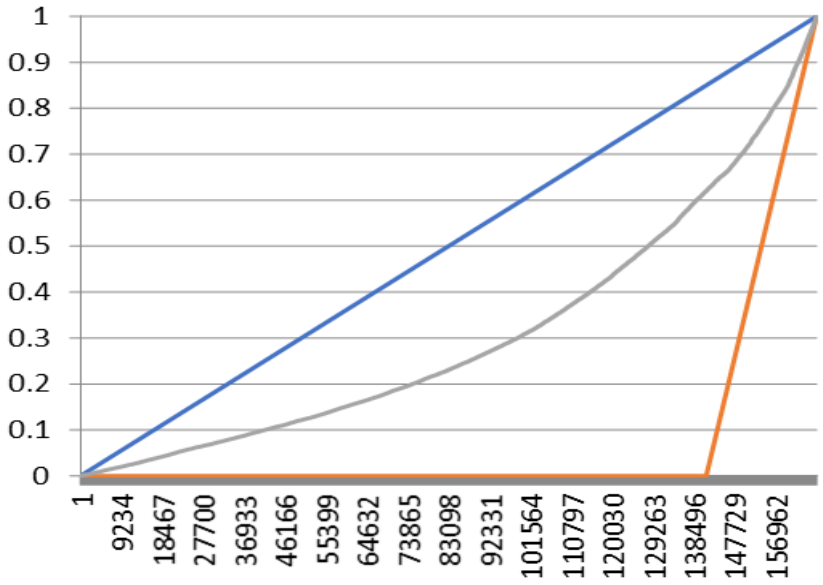


FIGURA 4.5 CURVA DI ACCURACY MODELLO PER SOCIETÀ

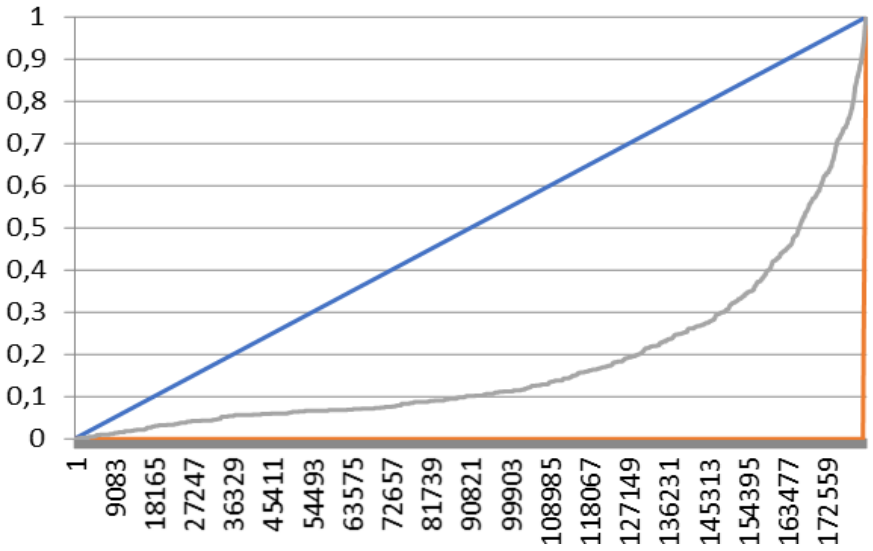
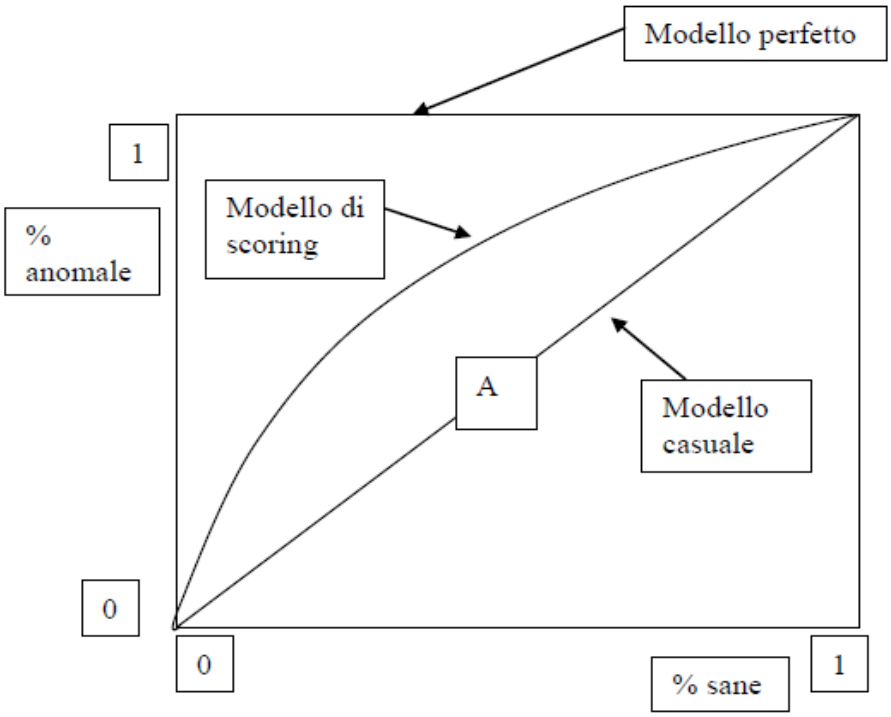


FIGURA 4.6-CURVA DI ACCURACY-MODELLO PER ANNO

Un altro strumento adoperato per calcolare l'accuratezza dei modelli è la curva ROC (*Receiving Operating Characteristic*), di cui si riporta una rappresentazione grafica:

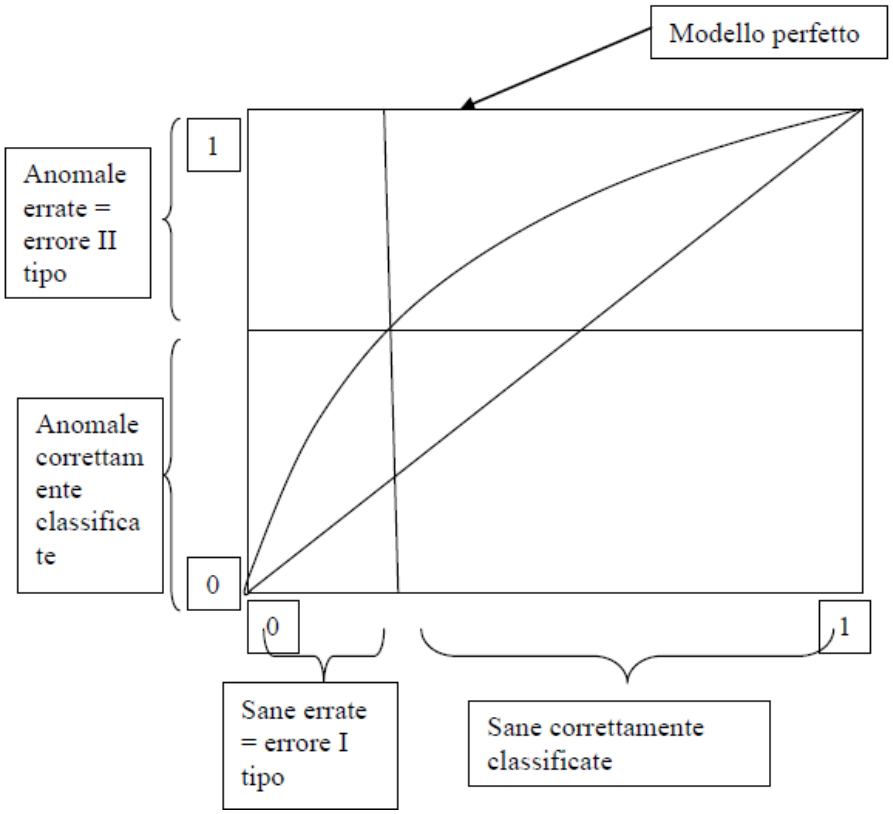


Attraverso l'analisi delle curve ROC si valuta la capacità del classificatore di distinguere i campioni sani e anomali, all'interno di un campione rappresentativo, calcolando l'area sottesa alla curva. Le curve ROC hanno inoltre due condizioni che rappresentano due curve limite:

- la diagonale rappresenta il caso del classificatore casuale e l'area sottesa è pari a 0,5.
- la seconda curva che rappresenta il classificatore perfetto è determinata dal segmento che dall'origine sale al punto (0,1), e da quello che congiunge il punto (0,1) a (1,1) avendo un'area sottesa di valore pari a 1.

La relazione che collega l'accuracy al ROC è:

$$\textbf{Accuracy}=2\cdot ROC-1=2\cdot(ROC-0.5)$$



La classificazione degli errori nel modello ROC avviene attraverso l'individuazione di quattro cluster. Le società sane correttamente classificate si trovano nel riquadro a destra del grafico, le anomale correttamente classificate e le sane errate (errore di I tipo) nel riquadro in basso a sinistra, e le società anomale classificate in modo errato (errore II tipo) si trovano nel riquadro in alto a sinistra.

Valutazione degli errori nei modelli

Per valutare gli errori dei modelli si sono suddivise le osservazioni in quattro cluster: società sane correttamente classificate, società sane erroneamente classificate, società anomale correttamente classificate e infine quelle anomale erroneamente classificate. La matrice di classificazione è impiegata per individuare le classi è la seguente:

		Classificazione	
		Sana	**Anomala**
Situazione effettiva	**Sana**	Corretta	Errore Secondo tipo
	Anomala	Errore Primo tipo	Corretta

Per errore di *Primo tipo* si intende la classificazione di una società come anomala quando in realtà questa società risulta sana, mentre per errore di *Secondo tipo* si intende la classificazione di una società come sana quando in realtà è anomala. Sono state utilizzate diverse variabili di bilancio sia nel Primo sia nel Secondo modello, al fine di valutare la differenza tra le medie di società correttamente ed erroneamente classificate.

Analisi errori Primo modello

Analizzando l'indicatore di bilancio "*rotazione del capitale investito*", le società sane classificate correttamente rappresentano il 75% del campione, mentre le società anomale corrette ne costituiscono il 5%. La differenza tra la media delle società sane correttamente classificate (2.141) e la media delle società sane erroneamente classificate è notevole (0.390), ma confrontando il primo risultato con la media delle società anomale si evidenzia come la differenza tra le medie delle due classi sia molto ridotta (2.141064 ≈ 2.107873). Per l'"Indice di copertura delle immobilizzazioni" il numero di bilanci di società sane classificate correttamente è 124.265, mentre i bilanci anomali corretti sono 7.956. La media dell'indicatore per le società sane è pari a 1,5, quella delle società anomale classificate in modo errato è di 1,3472.

Un altro indicatore utilizzato per l'analisi è "*oneri finanziari su fatturato*". La media dell'indicatore per società sane classificate correttamente (120.082 società) è 0,4512, mentre la media delle sane errate è 3,268 (21.161 società). Si espongono di seguito i dati riferiti alle medie di ciascun indicatore:

- *Rotazione del capitale circolante lordo:*

Media sane corrette	2.141064
Media sane errate	0.390036
Media anomale errate	2.107873
Media anomale corrette	0.390012

- *Indice di copertura delle immobilizzazioni:*

Media sane corrette	1.5
Media sane errate	0
Media anomale errate	1.3742
Media anomale corrette	0

- *Oneri finanziari su fatturato:*

Media sane corrette	0.4512
Media sane errate	3.2689
Media anomale errate	0.44764
Media anomale corrette	3.27946

Analisi errori Secondo modello

Per quanto concerne il modello con flag per anno, la variabile "*rotazione capitale circolante*" presenta una media delle società sane corrette pari a 1,8636 per 180.077 bilanci, mentre la media delle anomale classificate in modo errato è di 1,1371, calcolata su 772 bilanci. L'"Indice di copertura delle immobilizzazioni" invece, presenta una media di 1.2248 per le sane calcolate in modo corretto e media zero per le sane calcolate in modo errato. La variabile "Oneri finanziari su fatturato" presenta la media di 0.853 per le società sane calcolate in modo corretto e 3.98 per le sane errate, la media delle anomale calcolate in modo errato è invece più simile alla prima e si accosta a 1,2312.

- *Rotazione capitale circolante lordo*

Media sane corrette	1.86367
Media sane errate	0.39
Media anomale errate	1.13705
Media anomale corrette	0.39

- *Indice copertura immobilizzazioni*

Media sane corrette	1.22488
Media sane errate	0
Media anomale errate	0.4425
Media anomale corrette	0

- *Oneri finanziari su fatturato*

Media sane corrette	0.853664
Media sane errate	3.98
Media anomale errate	1.2312
Media anomale corrette	3.98

Conclusioni

Il presente lavoro di tesi ha l'obiettivo di confrontare tra loro due differenti modelli di credit scoring: la regressione logistica e Support Vector Machines. Dopo aver descritto la teoria alla base di entrambi sono stati analizzati i risultati ottenuti dall'applicazione pratica su un campione rappresentativo di imprese appartenenti al settore dei trasporti.

I due modelli hanno portato a risultati diversi, in quanto il criterio adottato da ciascuno per determinare il confine che separa le società anomale dalle sane è differente: la regressione logistica si basa su un approccio statistico, mentre SVM si basa principalmente sulle proprietà geometriche dei dati. Infatti, SVM tenta di trovare un particolare iperpiano di separazione ottimale in grado di massimizzare il margine tra le classi, e questo riduce il rischio di errore sui dati.

Generalmente si utilizza prima la regressione logistica per studiare come funziona il modello, se i risultati ottenuti sono scadenti si prova il modello SVM con kernel lineare. Questi modelli hanno caratteristiche molto simili ma uno dei due potrebbe portare a risultati più efficienti rispetto all'altro, così come avvenuto nel caso oggetto di analisi.

I risultati ottenuti dal modello logistico sono insufficienti per determinare la classe di appartenenza di ciascuna impresa, il modello SVM, invece, è stato in grado di individuare l'intercorrere di una relazione di separazione non lineare tra le variabili, sfruttando la potenzialità dei kernel.

Confrontando i risultati ottenuti applicando i due modelli, si nota un netto incremento in termine di accuracy dall'applicazione del modello SVM rispetto alla regressione logistica. In particolar modo, per il Primo Modello che utilizza flag per società si è in grado di ottenere una precisione del 74,5%, risultato apprezzabile se paragonato con il 46% ottenuto con la regressione logistica. Per il modello che utilizza invece flag per anno, si passa dal 64% di accuracy del modello Logit al 96% di SVM. Alla luce dei risultati ottenuti si può affermare che il modello che flagga come anomalo solo l'anno di bilancio in cui si verifica l'evento di default (*Flag per anno*) è in grado di ottenere risultati migliori per entrambi i modelli di scoring, e che SVM restituisce in ogni caso risultati più accurati della regressione logistica.

Nel paragonare i risultati ottenuti, bisogna però tenere in considerazione che per il Logit è stato utilizzato il software statistico R, mentre per SVM è stato utilizzato il linguaggio di programmazione in Python.

R è un linguaggio indicato per l'inferenza statistica, cioè la metodologia con cui si rilevano le caratteristiche di un insieme dall'osservazione di una parte di esso. Il principale problema rilevabile in R riguarda la sua documentazione e la fornitura dei pacchetti, che a volte risultano incompleti e inadatti per poter lavorare su un campione costituito da una vastità di dati come quello preso in esame.

Python invece, rappresenta uno strumento ottimale per migliorare l'accuratezza delle previsioni, e infatti è uno degli strumenti più utilizzati al mondo del machine learning ed è stato scelto per l'analisi del modello SVM proprio perché in grado di colmare le lacune di R gestendo in modo più efficiente il campione in esame.

Da un punto di vista pratico invece la regressione logistica richiede l'applicazione di un metodo iterativo per individuare la migliore combinazione possibile di indicatori di bilancio al fine di ottenere una precisione elevata. Ciò implica di dover effettuare diverse iterazioni inserendo in ciascuna di esse indicatori differenti; operazione che necessita di un notevole impiego di tempo. Al contrario, SVM è in grado di fornire un risultato immediato dopo aver importato correttamente i comandi di elaborazione del modello e dopo aver eseguito un'attenta fase di pulizia del campione. Pertanto, il modello di scoring SVM, oltre a fornire risultati più accurati, garantisce una maggiore efficienza computazionale.

Bibliografia

La gestione del rischio e allocazione del capitale nelle banche – Sironi A.

La gestione delle perdite attese e delle perdite inattese per gli intermediari bancari e finanziari - D'Auria C., Moderari

Determination of Default Probability by Loss Given Default- M. Misankova, E. Spuchľakova

Strumenti di Controllo e Analisi del Rischio – Albergo F.

Applied Logist Regression - Hosmer D., Lemeshow S.

Regressione Multipla e Logistica - Sense V.

Support Vector Machines- Sciandrone M.

Support Vector Machines- Maniezzo V.

Credit Risk Measurement Methodologies - Allen D.

La gestione del rischio di credito con modelli di derivazione attuariale: il caso di CreditRisk+-Resti A.

L'autotrasporto italiano tra crisi congiunturale, competizione internazionale e nuovi modelli di business - Contship

Trasporti e Telecomunicazioni – Annuario Statistico Italiano

Sitografia

http://bankpedia.org/index.php/it/125-italian/r/22174-rischio-di-credito

https://economiafinanzaonline.it/rischio-di-credito-definizione-informazioni/guide/

https://www.startingfinance.com/approfondimenti/basilea-iii/

https://www.moderari.com/public/articoli/34/2013-03-Perdite-attese-perdite-inattese.pdf

http://www00.unibg.it/dati/corsi/60012/39658-2010%2008.%20Basilea%202.pdf

https://it.qaz.wiki/wiki/Logistic_regression#Logistic_function,_odds,_odds_ratio,_and_logit

https://lorenzogovoni.com/support-vector-machine/

https://www.developersmaggioli.it/blog/support-vector-machine/

https://www.repubblica.it/economia/rapporti/energitalia/mobilita/2019/08/05/news/trasporto_merci_in_italia_piu_costi_meno_competitivita_-232451032/

https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Passenger_transport_statistics/it

https://www.bancaditalia.it/pubblicazioni/indagine-trasporti-internazionali/index.html

https://analyticsindiamag.com/using-near-miss-algorithm-for-imbalanced-datasets/

https://lorenzogovoni.com/algoritmo-smote/