POLITECNICO DI TORINO

Corso di Laurea Magistrale

in Ingegneria Gestionale

Tesi di Laurea Magistrale

Confronto tra Logit e SVM per la valutazione del rischio di credito: il caso del settore delle Costruzioni

Relatore

Prof. Franco Varetto

Candidato

Mauro Siciliano

Anno Accademico 2017/2018

Indice

Introduzione

Nel panorama finanziario internazionale degli ultimi anni, il Credit Scoring ha acquisito un ruolo cruciale in quello che viene definito come processo di Analisi del Rischio di Credito. L'obiettivo finale dei modelli di Credit Scoring è riuscire a classificare in maniera corretta chi fa domanda per l'accesso ad un canale di credito, al fine di ridurre il rischio legato alla possibile insolvenza del richiedente. In seguito alla crisi globale scoppiata negli Stati Uniti per effetto delle insolvenze sui mutui sub prime, gli istituti finanziari e creditizi hanno dovuto fronteggiare nell'ultimo decennio gravi perdite, alle quali è susseguito un lungo periodo di recessione. Tutto ciò ha posto inevitabilmente sotto la lente d'ingrandimento tutto il processo di analisi del rischio di credito, con particolare enfasi verso le tecniche di Credit Scoring.

Il credit scoring ha subìto una crescita significativa soltanto negli ultimi decenni. Prima di questi sviluppi, i finanziamenti venivano concessi a seguito di valutazioni per lo più soggettive, sulla base dei rapporti personali che si venivano a creare tra richiedente e analista del credito dell'ente erogatore. Nella maggior parte dei casi, quest'ultimo era rappresentato proprio dalla banca alla quale veniva richiesto il finanziamento. Tuttavia, la crescita sempre più veloce delle richieste di finanziamento e la moltitudine di prodotti finanziari disponibili, hanno reso necessario un adattamento del sistema di gestione del credito. In particolare, si è cercato di applicare, sempre di più, tecniche automatiche e oggettive che permettessero di effettuare l'analisi del potenziale cliente in maniera più veloce ed efficiente. L'output di queste tecniche prevede appunto l'adozione di uno "Score" che possa classificare in maniera affidabile il merito creditizio del cliente. Nel corso degli anni, le regolamentazioni più rigorose e stringenti, aventi come scopo quello di istruire riguardo l'oggettività di trattamento nella concessione del credito, hanno spinto gli istituti finanziari e creditizi verso lo sviluppo e l'utilizzo di sistemi informativi più elaborati e complessi, capaci di immagazzinare ed aggiornare correttamente la moltitudine di dati appartenenti ai clienti e potenziali tali. La base di dati a disposizione, con la sua completezza e idoneità, è la risorsa indispensabile a garantire l'ottimo funzionamento di qualsivoglia modello di credit scoring. La legislazione ha introdotto inoltre dei parametri standard nella valutazione del credito, al fine di garantire un controllo efficace del rischio, ed un impiego efficiente dei fondi d'investimento. "È in questa ottica che il Comitato di Basilea, composto

dai governatori ` delle Banche Centrali dei maggiori paesi industrializzati, ha emanato il Secondo accordo sul Capitale, noto anche come Basilea 2. L'accordo, entrato in vigore alla fine del 2006, definisce, fra l'altro, i criteri per la determinazione del patrimonio di vigilanza delle banche e degli intermediari finanziari, ovvero il capitale da porre a copertura dei rischi che scaturiscono dal loro normale funzionamento."[1] Nel fornire i principi per la gestione del rischio di credito, inoltre, "il Comitato ha evidenziato l'urgenza di adottare i sistemi di rating e richiede di utilizzarli non solo per il calcolo dei requisiti patrimoniali ma anche nella gestione operativa del credito cioè nei processi di selezione/revisione, nel calcolo del pricing e nella determinazione degli accantonamenti."[2]

Risulta evidente come le tecniche di Scoring siano entrate prepotentemente a far parte del processo di gestione e valutazione del rischio di credito. Per poter garantire uno sguardo d'insieme il più possibile completo, si è deciso di trattare, all'interno del presente studio, gli aspetti teorici alla base del Credit Scoring e i fondamenti dei modelli più adottati. Tra questi, si è scelto di sviluppare due modelli caratterizzati da approcci teorici diversi e ben delineati, in modo da affiancare alla trattazione teorica esempi applicativi capaci di descrivere in maniera più efficace i principi delle tecniche di Scoring.

In questo lavoro di tesi varrà trattata l'applicazione dei modelli di credit scoring nell'ambito della gestione del rischio di credito. Particolare enfasi verrà posta su due tecniche di credit scoring affiancate all'approccio machine learning, il Logit e l'SVM. La trattazione sarà strutturata in sei capitoli. Il primo capitolo affronterà le tematiche legate all'analisi del rischio di credito, la sua funzione all'interno del processo di concessione del credito, le sue componenti, la regolamentazione internazionale, e la sua evoluzione prevista dagli accordi di Basilea. Il secondo capitolo analizzerà più nel dettaglio le funzioni del Credit Scoring e gli obiettivi che intendono ottenere gli istituti finanziari e creditizi con il loro utilizzo. Verrà descritto l'approccio del Machine Learning inteso come apprendimento automatico dei dati, e la sua applicazione nei diversi modelli elencati. In questo senso verrà fornita una panoramica delle tecniche più utilizzate, da quelle più tradizionali e quelle che invece si sono imposte con forza negli ultimi anni come le neural network e le SVMs. Nell'ambito del caso applicativo proposto in questo studio verrà descritto nel corso del terzo capito il

[1] Introduzione ai metodi statistici per il credit scoring, Elena Stanghellini
[2] Valutazione strategica e previsione finanziaria nel rating interno delle imprese, Giovanni Felisari

Settore delle Costruzioni in Italia. Le informazioni e i dati alla base di questo studio appartengono infatti a questo background economico. Entrando nel dettaglio verrà spiegato come si è costruito il campione di dati contenenti le informazioni di partenza, com'è lo stesso è stato elaborato per risultare affine alla lettura del programma statistico, i singoli passaggi e le criticità incontrate. Il quarto capitolo sarà una rapida infarinatura del programma statistico R, utilizzato per costruire ed effettuare le simulazioni dei modelli di classificazione. In particolare, verranno descritte le principali funzioni e i packages utilizzati. Per poter usufruire di questo strumento, il periodo di esecuzione della tesi ha previsto una fase preliminare dedicata alla conoscenza teorica e all'apprendimento del software. Sono state seguite video-lezioni, alle quali sono state alternate esercitazioni pratiche, per poter padroneggiare gli strumenti indispensabili a costruire i modelli di credit scoring ed effettuare le relative simulazioni. Negli ultimi due capitoli, infine, si affronteranno tutti gli aspetti legati al modello Logit e a quello SVM. Nel quinto capitolo verrà spiegato l'utilizzo dell'analisi della regressione e le sue diverse varianti, tra cui soprattutto quella logistica. Verrà fornita una descrizione delle basi matematiche a supporto del modello di regressione e ci si soffermerà sull'implementazione del modello in R e sulle principali tecniche utilizzate. Il sesto capitolo ripercorrerà anch'esso le strutture alla base del modello proposto. Ad una trattazione teorica sulle Support Vector Machines seguirà una sezione dedicata alle nozioni base di questa tecnica, prima di addentrarci nell'implementazione vera e propria del modello all'interno del software. Le conclusioni chiuderanno questo percorso e verrà dato spazio alle riflessioni e le digressioni su quanto trattato fino a questo punto.

L'obiettivo di questo lavoro di tesi è stato dunque quello di sviluppare, applicare e analizzare due diversi modelli di Credit Scoring. Attraverso i modelli scelti per questo studio, vale a dire il modello di regressione logistica "Logit" e quello SVM, si è cercato di immergere la trattazione delle tecniche di scoring all'interno del contesto pragmatico e reale di cui fanno parte. Si sono messi a confronto due metodi profondamente differenti, con lo scopo di rilevare quale dei due offre una maggiore efficienza di classificazione. L'idea è stata quella di verificare se un algoritmo SVM, passato al clamore scientifico negli ultimi anni per la sua capacità di apprendimento dei dati, potesse sopraffare un modello Logit, superandolo in termini di fitting dei dati e precisione di classificazione. Entrambi i modelli sono stati sviluppati utilizzando moderne tecniche di machine learning. Le tecniche di

machine learning, eseguibili tramite l'utilizzo di appositi applicativi software, permettono di "istruire" i modelli in fase di sviluppo. Ciò vuol dire usare un campione di dati adeguatamente selezionato per adattare le capacità di previsione del modello sviluppato a quest'ultimo. Per fare ciò, nel caso oggetto di studio, si è utilizzato il programma open source "R-Studio", che permette, tramite il linguaggio di programmazione R, di sviluppare e analizzare complessi modelli statistici, avvalendosi di una vasta libreria di tools adatti a molteplici tipi di analisi.

Il campione scelto per svolgere questo lavoro di tesi è stato quello delle aziende italiane legate al Settore dell'Edilizia privata. Questo settore ha risentito in maniera importante della crisi finanziaria globale, e ha fornito all'interno del campione un numero cospicuo di aziende insolventi o in difficoltà finanziarie, rendendo di fatto questo campione adatto allo studio oggetto di questo lavoro.

Una parte importante del lavoro svolto, è stata la fase di costruzione del database da utilizzare come base per l'analisi statistica effettuata. Verrà spiegato in seguito, ed in maniera più approfondita, come l'elaborazione del set di dati risulti di fondamentale importanza al fine della costruzione dei suddetti modelli, descrivendo inoltre le principali difficoltà legate a questa prima fase.

Nei capitoli che seguono, verranno affrontati approfonditamente tutti gli aspetti che competono lo sviluppo, la costruzione e l'analisi dei modelli di Credit Scoring, mostrando le potenzialità e i limiti del loro utilizzo.

1 Analisi Rischio di credito

Come abbiamo detto, l'analisi e la gestione del rischio di credito hanno assunto negli ultimi anni sempre maggiore rilevanza nel panorama finanziario nazionale e internazionale. A seguito della crisi finanziaria globale, infatti, scoppiata a seguito di insolvenze su mutui con basso merito di credito (sub prime) negli Stati Uniti, le istituzioni finanziarie e bancarie hanno concentrato gran parte delle proprie risorse nel miglioramento del processo di gestione del rischio. In particolare, con rischio di credito si indica la possibilità che una variazione non prevista del merito creditizio di una controparte generi una corrispondente variazione nel valore attuale della relativa esposizione creditizia. Sono diverse le attività di un istituto creditizio o di una banca che vengono influenzate dal rischio di credito. Dal prestito al recupero crediti in caso di insolvenza, dagli investimenti in titoli obbligazionari fino alla loro emissione. Una corretta valutazione del rischio di credito è indispensabile per poter ottenere una gestione di successo in tutte le attività di una banca. La necessità di accrescere l'accuratezza nella valutazione del rischio da parte degli intermediari finanziari ha dato il via allo sviluppo, aggiornamento e scoperta di nuovi modelli e tecniche per il calcolo delle probabilità di default e dei tassi di recupero attesi. Un'analisi puntuale sull'evoluzione del merito creditizio del debitore ed una stima corretta dei parametri che influenzano il rischio di credito sono indispensabili inoltre per risolvere anche i problemi legati al pricing dei prestiti e delle obbligazioni. Tra i vari modelli e approcci provati, ci sono quelli più moderni che sfruttano le nuove capacità tecnologiche di immagazzinamento, analisi e addestramento dei dati alla teoria statistica in modo da identificare in maniera sempre più accurata le componenti del rischio di credito. Un approccio alternativo è identificato dall'analisi del mercato, a partire dalla struttura dei tassi di interesse a termine per fornire una valutazione del rischio di credito fondata sugli spread del mercato, e che vengono definiti come modelli in forma ridotta.

Come accennato precedentemente, una gestione ottimale del rischio di credito passa da una stima accurata delle componenti che influenzano il rischio. In quest'ottica, l'accordo di Basilea 2, definisce e regola le condizioni standard per il calcolo di alcune di esse. "L'accordo Basilea 2 è un patto internazionale che definisce i **requisiti patrimoniali delle banche** in relazione ai rischi (di credito, di mercato e operativi) assunti dalle stesse. Secondo l'accordo

Basilea 2 le banche dei paesi aderenti dovranno classificare i propri clienti in base alla loro rischiosità, attraverso **procedure di rating"[3].**

Quando si considera il rischio di credito, bisogna tener presente che vi sono principalmente due tipologie di rischio, vale a dire il rischio di insolvenza o controparte, e quello di emittente. Con il primo si fa spesso riferimento a posizioni in crisi di liquidità, e caratterizza soprattutto le transazioni in strumenti derivati, caratterizzati da un alto livello di rischio, in particolare quelli non scambiati su mercati regolamentati (over-the-counter), e viene in genere valutato su un orizzonte temporale medio/lungo. Il rischio di emittente invece, riguarda principalmente l'emissione di titoli obbligazionari, ed è caratterizzato dalla possibilità che l'emittente non riesca a ripagare gli interessi e il debito al creditore. Quando quest'ultimo (impresa, banca, Stato, ecc.) non è in grado di liquidare gli interessi e/o di restituire il capitale, il debitore si definisce insolvente.

1.1 Gli Accordi di Basilea

Il Comitato di Basilea, costituito nel 1975 e formato dai Governatori delle banche centrali dei Paesi del G10, ha avuto sin dall'inizio il compito di fornire delle raccomandazioni politiche e tecniche necessarie per la Vigilanza Bancaria. Nel 1988 è nato l'accordo sul Capitale, noto come Basilea 1, con l'intento di uniformare dal punto di vista normativo i requisiti patrimoniali delle banche. Esso prevedeva l'adozione di metodologie per valutare il rischio di credito delle banche basato su attività ponderate per il rischio e imponeva requisiti minimi patrimoniali per mantenere le banche solvibili durante i periodi di stress finanziario. Per le banche facenti parti degli Stati dell'Unione europea, le indicazioni del Comitato sul capitale minimo hanno avuto carattere d'obbligatorietà in seguito al recepimento della Direttiva 647/1989. Basilea I è stata seguita poi da Basilea II nel 2004, che era in fase di attuazione quando si è verificata la crisi finanziaria del 2008. L'obiettivo di questo accordo era quello di risolvere delle mancanze presenti nel precedente protocollo, soprattutto in merito alla gestione poco prudente del rischio di credito. Una delle novità principali di Basilea II riguardava la possibilità per gli istituti creditizi di adottare, oltre alle fonti di rating esterne, dei sistemi di rating interni sviluppati ad hoc dagli istituti stessi, in modo da fornire a questi ultimi maggiori strumenti per la gestione del rischio. Infine, sottoscritto nel settembre 2009, nasce l'accordo Basilea III. Questo accordo punta, come i suoi predecessori, a correggere gli errori nati nel tempo riguardo

[3] http://www.ratinglab.eu/accordo_basilea2.html

soprattutto il calcolo dei rischi. Questo sembra infatti possa aver contribuito alla crisi globale nata in America a causa delle insolvenze sui mutui sub prime. Con questo aumentano ancora le percentuali di attività che le banche devono detenere in forme più liquide e inoltre viene imposto a queste ultime di finanziarsi utilizzando più capitale , piuttosto che debito.

1.1.2 L'Accordo sul Capitale

L'accordo di Basilea sul capitale prevede una maggiore sensibilità dei requisiti patrimoniali al rischio di credito insito in portafogli di prestiti bancari. Questa maggiore sensibilità solleva la questione se, e in che misura, i nuovi standard patrimoniali intensificheranno i cicli economici. Il comportamento delle banche nel processo di concessione del credito può influenzare la ciclicità delle attività economiche, per motivi legati anche ai requisiti del capitale, come la percezione e l'avversione al rischio di credito. Inoltre, l'erosione del capitale economico delle banche a causa di perdite può ridurre la fornitura di nuovi prestiti anche in assenza di regolamentazione patrimoniale. Il regolamento, tuttavia, può rendere i vincoli sul capitale più stringenti. Elevati coefficienti di adeguatezza patrimoniale, con un'attuazione più severa di quella che i mercati stessi potrebbero imporre, potrebbe indurre alla ciclicità offerta di prestito nella misura in cui i rapporti diventano vincolanti durante le recessioni aziendali. Un regime di regolamentazione può quindi avere effetti ciclici a causa del livello di capitale richiesto, così come la sensibilità del capitale stesso alle variazioni del rischio. Con il regime regolamentare di tale accordo, il livello del capitale richiesto si basa sulla ponderazione del rischio degli asset, e tutti i prestiti commerciali e industriali (C & I) comportano un fattore di ponderazione del rischio del 100%, indipendentemente alle variazioni del rischio di credito tra i prestiti e nel tempo per lo stesso prestito. In base all'accordo Basilea I, i prestiti vengono ripartiti tra classi di rischio, ognuna delle quali con un proprio peso. Un prestito che diventa più rischioso, a causa di un deterioramento della qualità del credito del mutuatario, può essere spostato in una classe di rischio a più alta richiesta di capitale. Se il clima economico si deteriora in maniera significativa, un certo numero di prestiti potrebbero dover essere riassegnati ad una categoria caratterizzata da un fattore di rischio più elevato facendo aumentare sensibilmente dunque i requisiti patrimoniali necessari.

Come conseguenza di questo accordo, le banche possono avere la necessità di raccogliere capitali o eventualmente vendere asset per evitare azioni normative. Poiché l'emissione di nuove azioni può essere difficile nel breve termine, le banche potrebbero spostare le attività lontano dalle categorie di prestiti a rischio più elevato. Di conseguenza, sotto il regime di regolamentazione dell'accordo le condizioni di concessione del credito potrebbero diventare più restrittive quando

l’economia si trova in fase recessiva rispetto a quanto accadeva precedentemente. Per legare più efficacemente i requisiti patrimoniali al rischio sottostante i portafogli di prestito bancari, l’accordo Basilea I, consente due principali approcci per la valutazione del rischio di credito.

Le banche con sistemi di valutazione del rischio sufficientemente sviluppati possono utilizzare un metodo basato sui rating interni per stimare il rischio di credito del proprio portafoglio prestiti. Tuttavia, la mancanza di dati storici rende difficile valutare empiricamente i potenziali effetti ciclici di tale approccio. Questo accordo ha il vantaggio di consentire a qualsiasi banca di utilizzare un "approccio standardizzato" per la valutazione del rischio. L’approccio prevede la valutazione dei prestiti alle imprese tramite l’impiego di rating sulle emissioni di debito non garantito, fornite da agenzie esterne come Moody's, Standard & Poor’s e Fitch. In base a questo approccio, i prestiti alle società sono ripartiti tra quattro categorie di rischio: viene corrisposto un fattore di ponderazione del rischio del 20% per i sinistri con rating Aaa o Aa, del 50 percento per richieste con rating A, 100 percento per valutazioni Baa o Ba e 150 percento per reclami con rating B o inferiore. Le richieste prive di rating contro le società rischierebbero peso del 100 percento.

L’accordo di Basilea sul Capitale, contiene quindi la prima definizione e standardizzazione dei requisiti sul capitale minimo bancario accettate a livello internazionale. Ogni esposizione, in seguito a tale accordo, deve prevedere una quota di riserva di capitale utile a coprirsi dai rischi correlati. La riserva in questione, viene calcolata confrontando il patrimonio di vigilanza e il valore degli Asset bancari impiegati per la concessione dei prestiti ponderando per il rischio di credito. L’accordo impone, in particolare, alle banche di accantonare l’8% delle attività creditizie ponderate per il rischio di credito, definito come coefficiente di solvibilità. Diventato inadeguato rispetto alle nuove frontiere dei mercati finanziari, e alle nuove tecniche di gestione dei rischi, si è deciso di scrivere un nuovo Accordo. Questo è stato pensato per risolvere le problematiche legate alla stima, in genere sottostimata, del rischio di credito, e migliorare la gestione dei rischi in mercati particolari, come quello immobiliare, evitando così le opportunità di arbitraggio che erano nate sfruttando le carenze della regolamentazione.

1.1.3 L'accordo Basilea II

Per poter comprendere meglio il quadro normativo introdotto dall'accordo Basilea II, introduciamo i tre pilastri portanti di questo accordo. In particolare, queste politiche riguardano:

- la definizione dei requisiti patrimoniali minimi
- il processo di controllo prudenziale
- la disciplina di mercato.

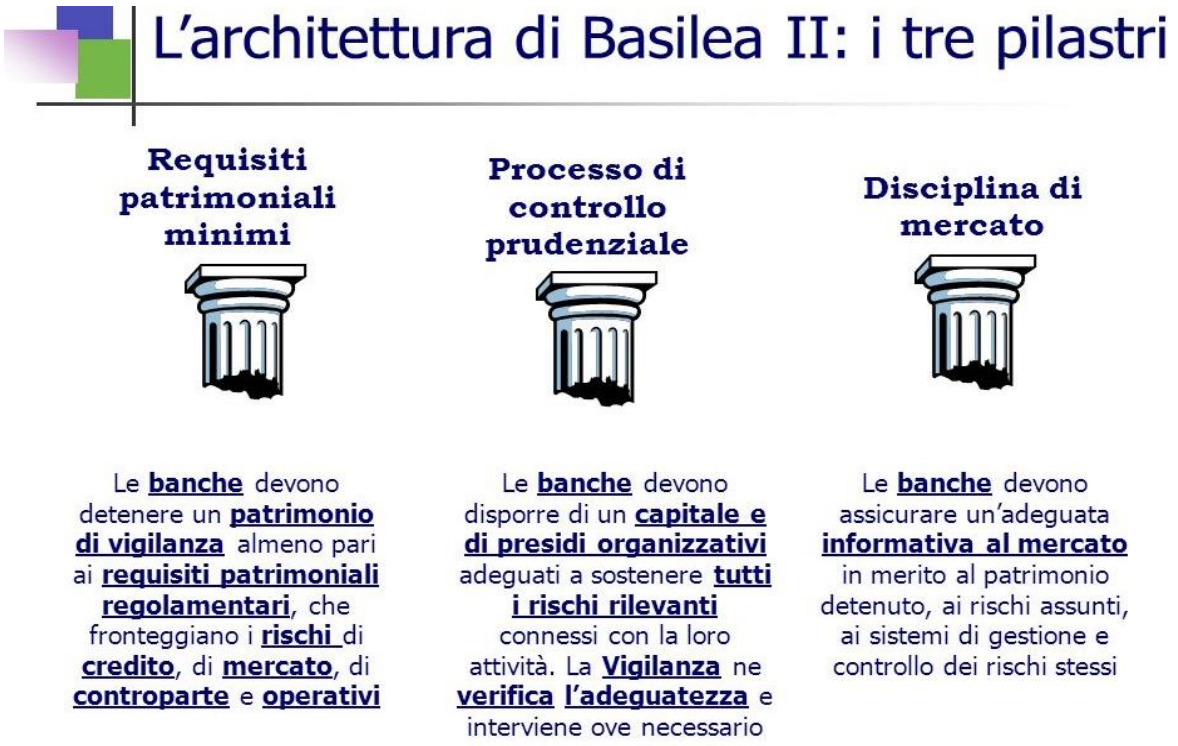


Figura 1.1, I tre pilastri dell'Accordo Basilea II

È in questo momento che viene introdotto il principio che i requisiti minimi patrimoniali debbano coprire le eventuali perdite inattese dovute al rischio di credito, al rischio di mercato e al rischio operativo. In particolare, per quanto riguardo il rischio di credito, si cerca di migliorare la stima di questa variabile, attraverso l'introduzione di nuove metodologie di rating.

Come accennato precedentemente dunque, uno dei pilastri centrali del quadro normativo Basilea II è il concetto dei requisiti patrimoniali basati sul rischio. Sotto il cosiddetto approccio IRB (internal ratings based), l'importo del capitale che una banca deve tenere come protezione da una data

esposizione sarà una funzione del rischio di credito stimato di quest'ultima. La stima del rischio di credito è considerata funzione predeterminata di quattro parametri:

- probabilità di default (PD)
- perdita in caso di default (LGD)
- esposizione al default (EAD)
- maturità (M).

Le banche che adottano la variante "Avanzata" dell'approccio IRB hanno la responsabilità di fornire tutti e quattro questi parametri, basandosi sui propri modelli interni. Le banche che operando utilizzando la cosiddetta variante "Foundation" dell'approccio IRB devono solo fornire il parametro PD, con gli altri tre parametri che devono essere definiti esternamente.

Rispetto all'approccio "one-size-fits-all" contenuto nell'accordo Basilea I, i requisiti patrimoniali basati sul rischio dovrebbero ridurre le distorsioni dei prezzi attraverso le categorie di prestito, nonché gli incentivi che accompagnano le banche a impegnarsi in varie forme di arbitraggio del capitale regolamentare. Allo stesso tempo, questo nuovo approccio alla regolamentazione del capitale solleva alcune preoccupazioni. Una preoccupazione che è stata ripetutamente espressa, ma che ha è stata soggetta ad un'analisi formale relativamente piccola, è che i nuovi standard di capitale avrebbero potuto inasprire le fluttuazioni del ciclo economico. In breve, l'idea è che in una fase di recessione, quando la base di capitale di una banca viene probabilmente erosa da perdite su crediti, i debitori della banca, declassati dai relativi modelli di rischio di credito, costringono la banca a detenere più capitale rispetto al suo attuale portafoglio di prestiti. Pare ovvio come in situazioni di crisi possa risultare difficoltoso per la banca raccogliere nuovi capitali. In queste situazione allora sarà costretta a ridurre la sua attività di prestito, contribuendo in questo modo a peggiorare ulteriormente le condizioni di partenza.

Il secondo pilastro, invece, mira a rafforzare il ruolo delle autorità di vigilanza nazionali nel garantire l'efficacia dell'accordo. Con il secondo pilastro, il Comitato sembra evolvere il processo di revisione prudenziale in modo da renderlo più sensibile al rischio.
Il terzo pilastro infine, punta a integrare i requisiti minimi patrimoniali (pilastro 1) e il processo di revisione prudenziale (Pilastro 2). Il Comitato in questa sezione dell'accordo, vuole incoraggiare la disciplina di mercato sviluppando una serie di requisiti di informativa che consentano agli operatori

di mercato di valutare le informazioni chiave sul campo di applicazione, capitale, esposizioni al rischio, processi di valutazione del rischio e, quindi, adeguatezza patrimoniale dell'istituzione. Basilea II quindi, rappresenta un aumento della sensibilità al rischio della regolamentazione e della vigilanza bancaria e, di conseguenza, riduce il rischio delle banche rispetto i costi di supervisione.

Nel 2017 infine, è stato raggiunto l'accordo finale sulle regole di Basilea III, capitolo terminale che dovrebbe permettere di archiviare il sofferto periodo post crisi finanziaria. L'accordo, come ha spiegato il presidente della BCE Draghi, prevede infatti un approccio più omogeneo alla valutazione dei rischi interni delle banche, limitando «i benefici» dei modelli interni, cioè quelli finora adottati in Europa, e, «deve essere ora trasposto nelle leggi nazionali e recepito dai regolatori». Il panorama della regolamentazione bancaria e finanziaria è in costante movimento per rispondere con sempre più prontezza ai cambiamenti del mercato. Come si evince dalla trattazione, il rischio di credito in questo senso è sempre più centrale nelle tematiche affrontate dagli organi regolatori, sottolineando la centralità del suo ruolo nei sistemi di valutazione del rischio.

1.2 Le componenti del Rischio di credito

Quando un'istituzione finanziaria effettua un prestito, si espone al rischio di credito. Il creditore è esposto al rischio di non vedersi rimborsato, del tutto o parzialmente, del prestito concesso. Considerando questa eventualità dunque, il creditore può stimare in base alle possibilità di accadimento le perdite attese (EL – Expected Loss). Legata al rischio di credito però c'è anche una componente che non può essere prevista ed è definita come perdita inattesa (UL – Unexpected Loss). Di queste due componenti, quella che costituisce la vera fonte di rischio è la perdita inattesa, caratterizzata dall'imprevedibilità che si verificarsi di un deterioramento improvviso della qualità del credito. Uno dei motivi principali per cui le banche detengono il capitale, infatti, è creare una protezione contro picchi di perdite che superano i livelli attesi, tenere solo il le riserve legate alle perdite attese potrebbe non essere appropriato. Picchi di perdite, anche se si verificano abbastanza di rado, possono essere molto grandi quando si verificano. Pertanto, una banca dovrebbe anche riservare denaro per le cosiddette perdite impreviste (UL). Le perdite attese, infatti, per loro stessa natura, vengono considerate in partenza negli accantonamenti e nelle attività di pricing da parte dell'istituzione creditizia. In questo modo chi esercita il credito si copre dal rischio rappresentato da questa componente, trasferendolo sotto forma di costo al debitore. Nell'immagine seguente viene fornita una rappresentazione grafica (effettuata da MPS) della distribuzione delle due componenti principali del rischio.

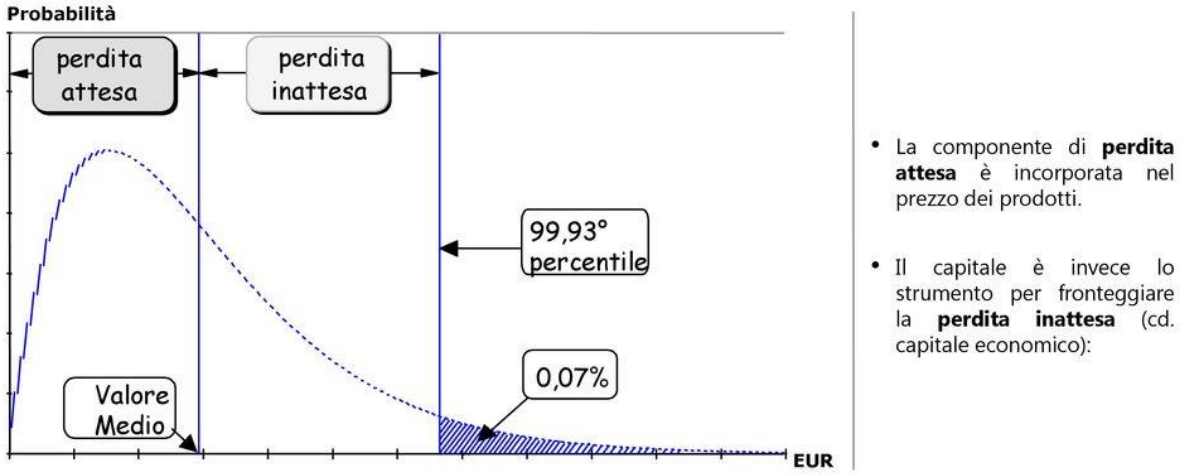


Figure 1.2, Distribuzione del rischio di credito_MPS

1.2.1 La perdita attesa

Si definisce perdita attesa, il "valore medio della distribuzione delle perdite che un'istituzione creditizia si attende di subire su un portafoglio prestiti"[4]. La perdita attesa "rappresenta una misura espressa a priori, che tiene conto sia della dimensione della perdita, sia della probabilità che questa si verifichi. In quanto media ponderata in base alla probabilità dei due possibili risultati, la perdita attesa è un valore intermedio tra questi"[5]. Per poter determinare la perdita attesa o Expected Loss (EL), l'istituzione creditizia deve stimare tre variabili per ogni singola esposizione:

• La Probability of default (PD)
• La Loss given default (LGD)
• L' Exposure at default (EAD)

Con la stima di queste variabili, si determina dunque il valore monetario della perdita che si
verificherebbe in caso di mancata riscossione di un credito. La relazione che intercorre tra l'EL e i tre parametri sopra elencati è la seguente:

$$EL = PD \times EAD \times LGD \tag{1.1}$$

La Probability of Default (PD) è la probabilità d'insolvenza, e identifica il rischio associato al destinatario del credito. La PD può essere determinata attraverso diverse metodologie in base all'approccio che si vuole adottare. Si distinguono, infatti, tre principali metodologie per la stima della PD. Al fine della determinazione della PD è possibile fare riferimento al mercato dei capitali, e ai dati estrapolati da quest'ultimo. Un secondo modo di agire considera, invece, l'utilizzo di modelli analitico/soggettivi che inglobano al loro interno aspetti sia quantitativi (condizioni economico-finanziarie del debitore e prospettive reddituali future) che qualitativi. La terza metodologia, invece, fa riferimento ai rating

[4] www. portalino.it, Rossi Mariano, (2003), Perdita attesa, perdita inattesa e diversificazione

[5]https://web.uniroma1.it/dip_management/sites/default/files/allegati/Capital%20Management%20-%20cap%20%20678.pdf

creditizi che possono essere formulati esternamente, da agenzie specializzate (Standard&Poor's, Fitch Ratings, Moody's), oppure internamente, dall'istituto creditizio stesso, mediante l'utilizzo di modelli di rating che utilizzano basi statistiche e tecniche di calcolo più moderne come le machine Learning, fino ad arrivare alle Neural networks. Sulla base di queste considerazioni, verranno analizzati e comparati, in seguito nella trattazione, un Modello Logit basato sulla regressione logistica ed un modello SVM (support vector machines).

La Loss Given Default (LGD) rappresenta la perdita che l'istituto di credito subisce a causa dell'insolvenza della controparte, nel momento in cui questa diventa effettiva. Pertanto essa non è mai prevedibile ex-ante, ma si manifesta solo quando termina l'operazione di recupero del credito.

Nella definizione di questa variabile bisogna tenere in considerazione gli aspetti oggettivi e non, rilevanti nel processo di valutazione del credito. Rientrano in questo contesto, le specifiche contrattuali e della struttura del credito, e le valutazioni anche soggettive relative al merito del debitore. La LGD dunque, è influenzata da una moltitudine di fattori e caratteristiche proprie del debitore, e conseguenti al tipo di credito che viene considerato. Per poter valutare correttamente questa variabile, bisogna effettuare un'analisi tecnica delle caratteristiche del contratto, di cui è oggetto il credito, così come delle garanzie fornite dal debitore. Queste dipendono ovviamente dalla situazione finanziaria ed economica di quest'ultimo, dalla tipologia e dalla condizione della sua attività commerciale. Per effettuare questo tipo di analisi, si ricorre spesso alla tecnica di valutazione dei flussi di cassa futuri che si ritiene possano essere generati dall'attività del debitore.

Un focus viene effettuato anche su altri aspetti esterni alla condizione del debitore. Ogni istituto finanziario infatti, stima la LGD tenendo presente la propria capacità di monitorare l'andamento del debito, così come la possibilità di effettuare un recupero dei crediti. Infine, a tutti questi fattori, si aggiungono le variabili esterne, macroeconomiche e del mercato, che possono incidere sulla condizione del merito di credito del debitore, così come sulla capacità dell'istituto di continuare ad attuare le proprie strategie.

Il calcolo della Loss Given Default, o LGD, è quindi strettamente correlato alla corretta gestione dei crediti. In questo senso, è lo stesso accordo di Basilea a disciplinare il

comportamento degli istituti di credito verso la stima di questo parametro. Le indicazioni derivanti da questo patto sono chiare, le banche devono stimare la LGD di lungo periodo per ogni esposizione, utilizzando un orizzonte di osservazione minimo di sette anni. È chiaro come per una banca il calcolo della LGD rappresenti un lavoro accurato e dispendioso, basato su un'analisi statistico-econometrica fondata sui record storici di ogni cliente. La determinazione di questo parametro, spesso, non è sufficiente per definire l'esatto ammontare delle perdite future. La sua accentuata volatilità, dipendente anche dalla natura ciclica dei tassi di recupero, può portare a delle sottostime sensibili delle perdite effettive. L'adozione di modelli di stima stocastici, che permettono di valutare l'evoluzione nel tempo delle variabili che influenzano la LGD possono risolvere in maniera parziale questo problema. Per ottenere un risultato soddisfacente è necessario che ogni istituto creditizio spenda le proprie energie nello sviluppo di un modello che possa stimare la LGD, adattandosi il più possibile agli schemi adottati per l'individuazione della probabilità di insolvenza.
Per stimare la LGD si possono utilizzare diversi approcci, a seconda del contesto in cui si opera.

Uno degli approcci che si può seguire è il Market LGD. Si utilizzano le informazioni a disposizione sul mercato finanziario per stimare il tasso di recupero. Questo metodo è quindi adatto nel caso di scambio di strumenti derivati, negoziati all'interno del mercato finanziario. La metodologia prevede di verificare la variazione del prezzo dello strumento a seguito della dichiarazione del default, e di dedurre da quest'ultima la LGD. Si suppone infatti, che il mercato abbia scontato il tasso di recupero sul prezzo dello strumento al momento della dichiarazione, rendendo di fatto quantificabile la LGD.
Alcune agenzie di rating utilizzano invece, una variante più complessa di questo metodo, definita come Emergenze LGD. Alla base di questo metodo vi sono i prezzi dei nuovi strumenti finanziari che i debitori emettono come fonte di rimborso nei confronti dei propri creditori, nel momento in cui i loro crediti diventano inesigibili.
Il Workout LGD è il metodo più diffuso, e utilizza come base analitica un Database contenente una serie storica di dati sempre aggiornata. In particolare, in base al tipo di debito, alle caratteristiche del debitore, al settore di appartenenza e altre caratteristiche, si va ad osservare i flussi di cassa effettivamente realizzati e recuperati nei periodi successivi

ai diversi default. Si costruisce dunque, una matrice contenente le classi di default relative alle diverse procedure e caratteristiche individuate, e si utilizza per la stima del nuovo tasso di recupero, e calcolare poi la LGD.

Per risalire al valore della LGD, bisogna prima considerare tutte le componenti che consentono di calcolare il tasso di recupero. Queste le possiamo elencare come segue:

• tasso di sconto "i" da applicare per attualizzare i flussi di cassa recuperati in seguito al default;

• tutti i costi amministrativi (CA) connessi alla procedura di recupero;

• le rettifiche contabili sul valore nominale dell'esposizione che si verificano durante la procedura di recupero (RL)

La formula che definisce il tasso di recupero (recovery rate), dalla quale si può risalire alla LGD come complemento ad 1 di quest'ultimo, è definita come:

$$RR\,(\%) = \frac{RNS}{EAD} = \frac{RL}{EAD} \times \frac{RL - CA}{RL} \times (1+i)^{-T} \qquad (1.2)$$

$$LGD\,(\%) = 1 - RR$$

Di seguito definiamo i valori presenti all'interno della formula. Questi sono:

RNS = recupero netto scontato, ovvero il valore attualizzato al momento del default degli importi recuperati, al netto di tutti i relativi costi.

EAD = esposizione al momento del default

RL = recupero lordo, ovvero il valore nominale degli importi recuperabili così come riportato nelle scritture contabili della banca

CA = totale dei costi amministrativi e legali connessi alla procedura di recupero

i = tasso di sconto

T = durata del processo di recupero.

Come si può evincere dalla formula, vi è quindi una relazione inversa tra il default e il tasso di recupero. Più è probabile il default quindi, più il recovery rate si riduce.

"EAD è l'ammontare della perdita che una banca può subire a causa del default. Poiché l'inadempienza si verifica in una data futura sconosciuta, questa perdita è condizionata dall'importo al quale la banca è stata esposta al mutuatario al momento del default. Questo è comunemente espresso come esposizione di default (EAD). Nel caso del prestito a

termine normale, il rischio di esposizione può essere considerato piccolo a causa del suo programma di rimborso fisso. Ciò non vale per tutte le altre linee di credito (ad es. Garanzia, scoperto, lettera di credito, ecc.). Il mutuatario può attingere a queste linee di credito entro un limite stabilito dalla banca come e quando sorgono esigenze di finanziamento.

L'utilizzo della linea di credito ha caratteristiche cicliche, cioè l'uso aumenta nelle recessioni e diminuisce nelle espansioni. Il tasso di utilizzo aumenta monotonamente quando il debitore diventa più rischioso e si avvicina al rischio di insolvenza. Le banche in quanto creditori devono monitorare attentamente la potenziale esposizione per valutare il rischio di credito in modo più prudente. È in questo senso che la stima dell'EAD è assolutamente necessaria per il calcolo del capitale sia regolamentare che economico."6

Per calcolare l'EAD le istituzioni creditizie utilizzano un modello basato sulla seguente formula:

$$EAD = DP + UP * CCF \tag{1.3}$$

DP = Drawn portion, la quota dell'ammontare disponibile effettivamente utilizzata;

UP = 1 – DP = Undrawn portion, la quota disponibile non utilizzata;

CCF = Credit conversion factor, fattore stimato internamente che, rappresenta la percentuale di UP che la banca si attende venga utilizzata prima dell'insolvenza.

L'applicazione di tale modello comporta una stima prudenziale della perdita attesa e che si può tradurre nell'applicazione di un tasso d'interesse maggiore al debitore.

[6] Bandyopadhyay, A. (2016). Esposizione a default (EAD) e Loss Given Default (LGD). Nella *gestione del rischio di credito del portafoglio nelle banche* (pagine 137-185). Cambridge: Cambridge University Press. doi: 10,1017 / CBO9781316550915.005

1.2.2 La perdita inattesa

Una delle ragioni principali per la quale le banche conservano del capitale è creare una protezione contro i picchi di perdite improvvise, che quindi eccedono rispetto ai valori attesi. Per questo motivo non è mai appropriato tenere riserve solo per le perdite attese. Anche se le perdite inattese possono verificarsi raramente, l'entità di queste ultime può essere davvero importante. Questo è il motivo per il quale le banche, e gli istituti creditizi in generale, devono prevedere e riservare capitale per le perdite inattese. La perdita inattesa, detta anche "Unexpected Loss", è definita come "la variabilità della perdita attorno al suo valore medio, cioè attorno alla EL"[7]. La variabilità della perdita dal valore atteso EL è generalmente misurata come una media delle deviazioni standard della variabile stessa. Perciò, la perdita inattesa UL si definisce come:

$$U_L = \sqrt{\sigma_{EDF}^2 \times LGD + EDF \times \sigma_{LG}^2} \tag{1.4}$$

dove

σ_{EDF}^2, rappresenta la varianza del tasso di insolvenza

σ_{LGD}^2, rappresenta la varianza del tasso di perdita

LGD, è il tasso di perdita atteso in caso di insolvenza

EDF, è il tasso di insolvenza atteso

La formula espressa per la stima della perdita inattesa vale sotto l'ipotesi di indipendenza tra il tasso di insolvenza attesa (EDF) e il tasso di perdita atteso (LGD). Se questa ipotesi non è verificata bisogna includere nella formula la covarianza tra le variabili per esprimere il grado di correlazione tra esse.

[7] Resti A. & Sironi A., (2008), Rischio e valore nelle banche, Milano, Egea, p.356

2 Credit scoring

Il banking nel mondo si basa sul credito o sulla fiducia nell'abilità del debitore di adempiere ai propri obblighi. Confrontandosi con la crescente pressione dei mercati e dei regolatori, le banche hanno basato la loro fiducia, in maniera sempre più ampia, su tecniche statistiche per la stima della bancarotta aziendale (corporate) conosciuta come rating o scoring. Il loro obiettivo principale è di stimare la situazione finanziaria di una società e, se possibile, la probabilità che una società finisca in default (fallisca), entro un certo periodo.

L'applicazione di modelli statistici alla Corporate Bankruptcy divenne popolare dopo l'introduzione dell'analisi discriminante (DA) di Altman (1968). Più tardi, furono suggeriti i modelli Logit e Probit da Martin (1977) e Ohlson (1980). Tutti questi modelli appartengono alla classe dei "Modelli Lineari Generalizzati" (GLM) e potevano anche essere interpretati usando una variabile latente (score=punteggio). Il loro elemento decisionale principale è una funzione di punteggio (score) lineare (graficamente rappresentata come un iperpiano in uno spazio multidimensionale) che separa le società di successo da quelle in fallimento. Il punteggio di una società è calcolato come un valore di questa funzione. Nel caso del modello Logit e Probit il punteggio è, per via di una funzione di collegamento, direttamente trasformato in una probabilità di default (PD). Lo svantaggio principale di questi approcci così popolari è nella forzata linearità della funzione di collegamento (Logit e Gaussiana) tra le probabilità di Default e la combinazione lineare dei predittori.

Le procedure di credit scoring permettono di valutare il merito creditizio dei richiedenti nel processo di concessione del credito. Il risultato del credit scoring è un modello capace di identificare, tramite l'utilizzo di appropriati strumenti, quali per esempio indici di bilancio e dati finanziari, le società sane e quelle che con più probabilità si dimostreranno insolventi. Il vantaggio principale sottostante l'utilizzo di tali modelli è la diminuzione del rischio di credito, legato alla minore propensione all'erogazione di prestiti nei confronti di società ritenute non meritevoli. L'uso di questi modelli garantisce inoltre l'omogeneità di trattamento dovuta all'obiettività dell'analisi che sta alla base dello scoring, riducendo altresì i tempi necessari allo studio di ogni singolo caso.
Un buon modello di credit scoring consente non solo di prevedere in maniera affidabile il livello di perdite probabili, ma anche di controllare in maniera costante il livello di rischio.

Queste condizioni permettono evidentemente di pianificare l'attività e la redditività delle operazioni future.

Nel processo di gestione del credito, lo Scoring di Accettazione rappresenta una delle fasi più delicate supportata dai sistemi di credit scoring. In questa fase infatti, si valuta la probabilità di insolvenza del soggetto richiedente, decidendo se può essere meritevole o meno della concessione del credito. Sono molteplici i modelli statistici analizzati e sviluppati che sono stati applicati in questa fase, e più in generale in ambito finanziario. Si inseriscono in questa fase i modelli di machine learning, tra i quali quelli più tradizionali di regressione e quelli più attuali basati sulle reti neurali e support vector machine. Lo scopo di questi modelli è calcolare le probabilità di insolvenza e perfezionare in tal modo la classificazione dei richiedenti, al fine di ottimizzare il processo di concessione del credito. Lo studio della probabilità, in questi casi, si basa su serie di dati storiche appartenenti al cliente, di tipo quantitativo e qualitativo che possano descriverne il comportamento futuro.

2.1 Modelli machine learning

Il “machine learing” è un metodo di insegnamento dei computer ad analizzare i dati, imparare da essi e quindi fare una determinazione o una previsione sui nuovi dati. Invece che costruire manualmente un codice con una serie di istruzioni specifiche ad eseguire una determinata attività, la macchina viene "addestrata" utilizzando grandi quantità di dati e algoritmi per imparare come eseguire l'attività. L'apprendimento automatico si sovrappone come campo di applicazione all'apprendimento statistico. “Entrambi tentano di trovare e imparare da schemi e tendenze all'interno di dataset di grandi dimensioni per fare previsioni. Il campo dell'apprendimento automatico ha una lunga tradizione di sviluppo, ma i recenti miglioramenti nella memorizzazione dei dati e nella potenza di calcolo li hanno resi onnipresenti in molti campi e applicazioni diversi, molti dei quali sono molto comuni. Uno dei primi usi dell'apprendimento automatico è stata la modellazione del rischio di credito, il cui obiettivo è utilizzare i dati finanziari per prevedere il rischio di default.”[8]

Quando un'impresa richiede un prestito, il prestatore deve valutare se l'azienda può rimborsare in modo affidabile il capitale e gli interessi del prestito. I prestatori usano comunemente misure di redditività e leva per valutare il rischio di credito. Una società redditizia genera abbastanza denaro per coprire gli interessi passivi e il capitale dovuto. Tuttavia, un'impresa più indebitata ha meno risorse disponibili per affrontare gli shock economici. La complessità legata allo stabilire il merito creditizio di una società si moltiplica quando le banche incorporano le molte altre dimensioni che vengono esaminate durante la valutazione del rischio di credito. Queste dimensioni aggiuntive includono in genere altre informazioni finanziarie come il rapporto di liquidità o informazioni comportamentali come il comportamento di pagamento del credito. Riassumere tutte queste diverse dimensioni in un unico punteggio è impegnativo, ma le tecniche di apprendimento automatico aiutano a raggiungere questo obiettivo.

L'obiettivo comune dell'apprendimento automatico e degli strumenti tradizionali di apprendimento statistico è imparare dai dati. Entrambi gli approcci mirano a indagare le

[8] https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling

relazioni sottostanti utilizzando un set di dati di formazione. Tipicamente, i metodi di apprendimento statistico assumono relazioni formali tra variabili sotto forma di equazioni matematiche, mentre i metodi di apprendimento automatico possono imparare dai dati senza richiedere alcuna programmazione basata su regole. Come risultato di questa flessibilità, i metodi di apprendimento automatico possono adattarsi meglio ai modelli nei dati.

I modelli machine learning apprendono quindi da dati contenuti in database opportunamente creati, contenenti per lo più indici di bilancio e finanziari appartenenti a società sane e ad altre anomale o insolventi. Attraverso l'utilizzo di applicativi software i modelli vengono sviluppati in modo da verificare ipotesi sul merito creditizio e ottimizzati in modo da essere più aderenti possibile alla realtà. Il loro scopo è predire, sulla base dei dati analizzati, la probabilità di default del soggetto richiedente il credito. La probabilità di default è un valore fondamentale poi per calcolare il merito creditizio delle società oggetto di analisi. Per selezionare i fattori più rilevanti per l'analisi del merito creditizio i modelli di scoring vengono definiti sulla base di processi di ottimizzazione della funzione di verosimiglianza.

La qualità di un modello machine learning dipende tanto dalla bontà dei metodi statistici e matematici adottati nello sviluppo di questo ultimo, quanto dalla qualità dei dati utilizzati per addestrare e in seguito testare il modello. Spesso infatti i dati mancanti o gli errori contenuti nel database, utilizzato come fonte per l'analisi e la costruzione del modello, sono alla base dello scarso potere predittivo del modello stesso. Più in particolare, affinché un metodo machine learning sia adeguatamente efficace, bisogna garantire che il database di riferimento sia stato costruito nel modo corretto.

Questo significa prima di tutto contenere un numero di osservazioni tale da poter essere ritenuto statisticamente significativo in modo da poter effettuare su di esso analisi statistiche attenibili. Le osservazioni collezionate devono far parte di un campione selezionato adeguatamente per essere rappresentativo della popolazione oggetto di analisi. Nel caso in esame, questo significa selezionare un campione di aziende del settore edilizio che sia rappresentativo dell'intero settore italiano. Questo è indispensabile, in quanto, nel momento in cui venisse sviluppato un modello predittivo istruito su un campione non rappresentativo dell'intera popolazione, qualora lo si utilizzasse per una società con caratteristiche molto diverse da quelle delle società contenute nel campione,

allora i suoi risultati potrebbero essere non attendibili, ed il modello inefficace. In aggiunta alla quantità di osservazione a disposizione è fondamentale che queste osservazioni facciano parte di una serie temporale significativa, in modo da poter associare il cambiamento dei dati nel tempo, al corrispettivo cambiamento del merito creditizio della società.

2.2 Modelli di regressione

A partire dal caso più semplice, la regressione lineare a variabile singola è una tecnica utilizzata per modellare la relazione tra una singola variabile di input indipendente (variabile di funzione) e una variabile dipendente in output utilizzando un modello lineare. Il caso più generale è la regressione lineare a più variabili in cui viene creato un modello per la relazione tra più variabili di input indipendenti (variabili di caratteristiche) e una variabile dipendente. Il modello rimane lineare in quanto l'output è una combinazione lineare delle variabili di input.

Esiste un terzo caso più generale chiamato regressione polinomiale in cui il modello diventa una combinazione non lineare delle variabili di caratteristiche in cui possono esistere variabili esponenziali, seno e coseno, ecc. Ciò richiede tuttavia la conoscenza della relazione tra le variabili indipendenti e quella dipendente in output.

I principali modelli di regressione utilizzati per il Credit Assessment sono **i modelli logit e probit**. Questi modelli, sostanzialmente identici, differiscono sulla base della distribuzione di probabilità, che per il modello logit è logistica, mentre per il modello probit è normale. La funziona di regressione, non lineare, è caratterizzata da una variabile risposta di natura dicotomica o categoriale che può assumere due valori, di tipologia 1/0, che individua società sane (1) e insolventi o anomale (0). Le variabili dipendenti utilizzate più spesso per questi modelli sono rappresentate da indici di bilancio e altri indicatori finanziari che descrivono l'andamento economico e finanziario della società in un determinato periodo. La probabilità che una società sia insolvente o meno è dunque condizionata alla realizzazione delle sue variabili, ovvero degli indicatori economico-finanziari.

I vantaggi legati all'utilizzo di questi modelli sono da ricercare nella velocità di modellazione, e nei casi in cui la relazione da modellare non è estremamente complessa e se non si dispone di molti dati. Per esempio, la regressione lineare è semplice da comprendere e può essere molto utile per le decisioni aziendali.

D'altra parte, per i dati non lineari, la regressione polinomiale può essere abbastanza difficile da progettare, poiché è necessario disporre di alcune informazioni sulla struttura dei dati e sulla relazione tra le variabili delle caratteristiche. Come risultato di quanto detto, questi modelli non sono buoni come altri quando si tratta di dati altamente complessi.

2.2.1 La regressione

L'analisi della regressione è una delle tecniche statistiche maggiormente utilizzate. Il suo compito è quello di spiegare la relazione esistente tra una variabile Y (continua) detta variabile risposta, dipendente, e una o più variabili indipendenti dette regressori (X_1, X_2, … X_k). La funzione di regressione si può esprimere come:

$$Y=f(X_1, X_2, \dots X_k)+\varepsilon \tag{2.1}$$

Che attraverso la componente f (X_1, X_2, … X_k) indica l'esistenza di un legame funzionale tra la variabile dipendente Y e i regressori, e definita come componente sistematica. A questa componente va ad aggiungersi una seconda componente denominata accidentale, o casuale, che a differenza della prima non può essere spiegata dai regressori, e deve essere ricondotta quindi al caso, vale a dire a cause diverse, esterne non considerate dal modello regressivo.

Il legame funzionale è spesso rappresentato da una funzione di tipo lineare e pertanto si parla di regressione lineare multipla o modello lineare può essere formulato come segue:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon \tag{2.2}$$

dove β_0 è detto termine noto, β_1, …, β_k sono i coefficienti di regressione e, insieme alla varianza dell'errore, sono i parametri che caratterizzano il modello e devono essere stimati a partire dalle osservazioni che compongono il campione. Laddove la relazione tra le variabili non appaia lineare, i modelli di regressione permettono di effettuare delle apposite trasformazioni con l'obiettivo di linearizzare la relazione. È questo il caso della regressione logistica, dove per trasformare la relazione vengono utilizzati i logaritmi.

Infine, quando la variabile risposta Y non è continua, come avviene in caso di variabili dicotomiche (**regressione logistica**) o di conteggio (regressione di Poisson), è uso utilizzare una generalizzazione del modello lineare (GLM).

2.2.2 La regressione Logistica

Quando Y è dicotomica, e codificata come 0 -1, la sua distribuzione è binomiale. La stima della variabile dipendente dunque, dovrà variare tra 0 e 1 e non tra - infinito e + infinito come le stime della regressione lineare. In questo caso, allora, si osserva che per valori molto alti di X (o molto bassi se la relazione è negativa) il valore in Y è molto vicino ad 1 e non deve superare tale limite. Lo stesso avviene in prossimità dello 0. In pratica la curva che rappresenta la relazione tra X e Y è di tipo logistico e non lineare.

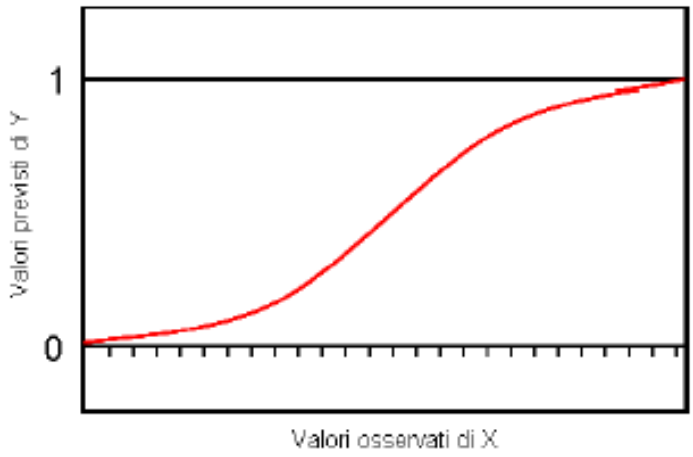


Figure 2.1, funzione che lega la probabilità Y alla combinazione delle variabili indipendenti X

Nella regressione logistica la variabile dipendente identifica l'appartenenza ad un gruppo. I valori che vengono assegnati ai diversi livelli, in modo da identificare un gruppo piuttosto che un altro, sono attribuiti in maniera arbitraria. Il risultato di interesse del modello di regressione logistica, dunque, è la probabilità che un dato soggetto appartenga a meno a uno dei due gruppi. Nonostante questo, i valori assegnati ai livelli, anche se arbitrari, influenzano il risultato dell'analisi, ed è per questo che talvolta può essere utile sostituire la probabilità con l'odds corrispondente.

L'odds è un modo di esprimere la probabilità mediante un rapporto. Si calcola facendo il rapporto tra le frequenze osservate nei diversi livelli. Una volta definite le frequenze, è possibile risalire alle frequenze relative, o percentuali, dei due livelli. Inoltre, dopo aver calcolato l'odds, eseguendo il suo logaritmo naturale si ottiene quello che viene definito come **logit**. "Se confrontiamo la distribuzione delle frequenze, frequenze relative, degli

odds e dei logit possiamo notare come tutte queste statistiche forniscono la stessa informazione sebbene con valori matematicamente differenti. Quando le categorie successi (Y = 1) e fallimenti (Y = 0) sono equiprobabili le frequenze relative sono uguali a 0.5 per entrambe le categorie di Y, gli odds sono uguali ad 1, mentre i logit sono uguali a 0. Quando il numero di successi è maggiore del numero dei fallimenti le frequenze relative assumono valori superiori a 0.5 per la categoria Y = 1 e minori per la categoria Y = 0, gli odds assumono valori superiori ad 1, mentre i logit valori superiori allo 0. Infine, quando il numero di successi è inferiore al numero dei fallimenti le frequenze relative assumono valori inferiori a 0.5 per la categoria Y = 1 e superiori per la categoria Y = 0, gli odds assumono valori inferiori a 1, mentre i logit valori negativi. In pratica, mentre le frequenze relative hanno un range di variabilità che va da 0 a 1, gli odds hanno un range di variabilità che va da 0 a più infinito, mentre i logit possono variare da meno infinito a più infinito."[9]

In questo caso, la non linearità della relazione tra le variabili non permette di applicare il metodo OLS senza applicare prima opposite trasformazioni che rendano lineare la relazione, e in particolare nei termini dei parametri. Bisogna procedere dunque con la trasformazione logaritmica della variabile dipendente.

Per esprimere la relazione tra le variabili indipendenti e quella dipendente in termini lineari si parte dalla seguente formulazione in cui il valore atteso della variabile dipendente è la probabilità:

$$(\hat{Y} = \mu_Y = P_{(Y=1)}), \tag{2.3}$$

per cui la probabilità di `Y = 1` come funzione lineare di `X` diventa:

$$P(Y=1) = \alpha + \beta X \tag{2.4}$$

Come accennato precedentemente, questo modello non è adeguato. I valori della probabilità hanno un range di ammissibilità che va da 0 e 1, mentre il termine α+βX può assumere valori che vanno da meno infinito a più infinito. Per risolvere tale problema si applica come primo passo la trasformazione esponenziale al termine di destra della funzione che diventa:

[9]http://psiclab.altervista.org/MetTecPsicClinica2015/Materiali/Dispensa_Regressione_Lineare_e_Logistica_2014.pdf

$$P(Y=1) = e^{\alpha+\beta X} \tag{2.5}$$

Anche con questa trasformazione il problema non viene risolto del tutto, infatti consente solamente di restringere i valori dell'equazione entro il range 0 + ∞. Per risolvere definitivamente il problema, si applica la trasformazione logistica che consente di controllare i valori e restringerli nel range della probabilità (0; 1):

$$P(Y=1) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \tag{2.6}$$

Nel caso di variabili dicotomiche l'odds diventa:

$$odds\,(Y=1) = \frac{P(Y=1)}{1-P(Y=1)} \tag{2.7}$$

Dove

$$\mathrm{P}(Y=0) = \left[1-P(Y=1)\right] \tag{2.8}$$

esprime la probabilità della seconda categoria come complemento ad uno della prima. Se definiamo in questo modo la probabilità di Y = 0 possiamo calcolare l'odds di Y = 1 che diventa:

$$odds_{Y=1} = \frac{\dfrac{e^{(\alpha+\beta X)}}{1+e^{(\alpha+\beta X)}}}{\dfrac{1}{1+e^{(\alpha+\beta X)}}} = e^{(\alpha+\beta X)} \tag{2.9}$$

Infine, applicando le proprietà dei logaritmi, calcolando il logaritmo dell'odds, al primo membro, si osserva che il logaritmo naturale dell'odds di Y = 1 è funzione lineare della variabile X:

$$\ln\left(odds_{Y=1}\right) = \alpha + \beta X \tag{2.10}$$

Applicando queste trasformazioni, l'equazione della relazione tra le variabili X_k e Y diviene:

$$P(Y=1)=\frac{e^{(\alpha+X_1\beta_1+X_2\beta_2+\ldots+X_i\beta_i+\varepsilon)}}{1+e^{(\alpha+X_1\beta_1+X_2\beta_2+\ldots+X_i\beta_i+\varepsilon)}} \quad (2.11)$$

"È importante sottolineare che la probabilità, l'odds e il logit sono tre differenti modi di esprimere esattamente la stessa cosa. La trasformazione in logit serve solo a garantire la correttezza matematica dell'analisi."[10]

[10]http://psiclab.altervista.org/MetTecPsicClinica2015/Materiali/Dispensa_Regressione_Lineare_e_Logistica_2014.pdf

2.3 Support Vector Machines

Le Support Vector Machines sono modelli predittivi di machine learning basati su algoritmi di apprendimento di dati che analizzano, istruiscono e ricercano dei pattern all'interno dei dati stessi. Questi modelli sono stati introdotti da Vapnik (1995,1998) nella sua pubblicazione sulla teoria della statistica di apprendimento con una particolare enfasi sulle support vector machines. Da allora, il campo del "machine learning" è stato testimone di un'intensa attività nello studio delle SVMs, che si sono diffuse sempre di più all'interno di altri campi e discipline come la statistica e la matematica. In questo momento dunque, ci sono molte comunità che lavorano sulle SVMs e sui metodi "kernel-based" legati ad esse. L'utilizzo principale delle SMVs risiede nella risoluzione dei problemi di classificazione, vale a dire ottenere un modello che restituisce in output una classe di appartenenza. Considerato un set di dati come input, definito come training set, con una variabile di tipo categoriale caratterizzata dalle classi (-1,1), nella forma più semplice, l'algoritmo di apprendimento SVMs restituisce come risultato l'appartenenza ad una delle due classi. Il modello mappa le osservazioni all'interno di uno spazio multidimensionale, e cerca l'iperpiano o gli iperpiani all'interno di esso in modo tale che le due classi di osservazioni siano le più distanti possibili tra loro. I punti che si avvicinano di più all'iperpiano sono quelli che servono ad individuare l'iperpiano di separazione ottimo, e vengono definiti per questo *support vectors*. L'algoritmo SVMs, a differenza dei modelli di regressione più tradizionali, riesce ad individuare anche schemi non-lineari. Il modello utilizza una funzione "Kernel". Quando la separazione lineare delle classi non è possibile, tramite la funzione Kernel, i dati vengono trasformati dal loro spazio originale (*input space*) e mappati in un nuovo spazio (*feature space*), multidimensionale, in modo da riportare il problema, ad un problema di ottimizzazione che segue lo stesso approccio del caso lineare. Questa soluzione, sotto la condizione definita come "Mercer's condition" garantisce che il problema di ottimizzazione converga verso un'unica soluzione. Verrà fornita di seguito un'analisi più puntuale e approfondita delle basi matematiche dei modelli SVMs.

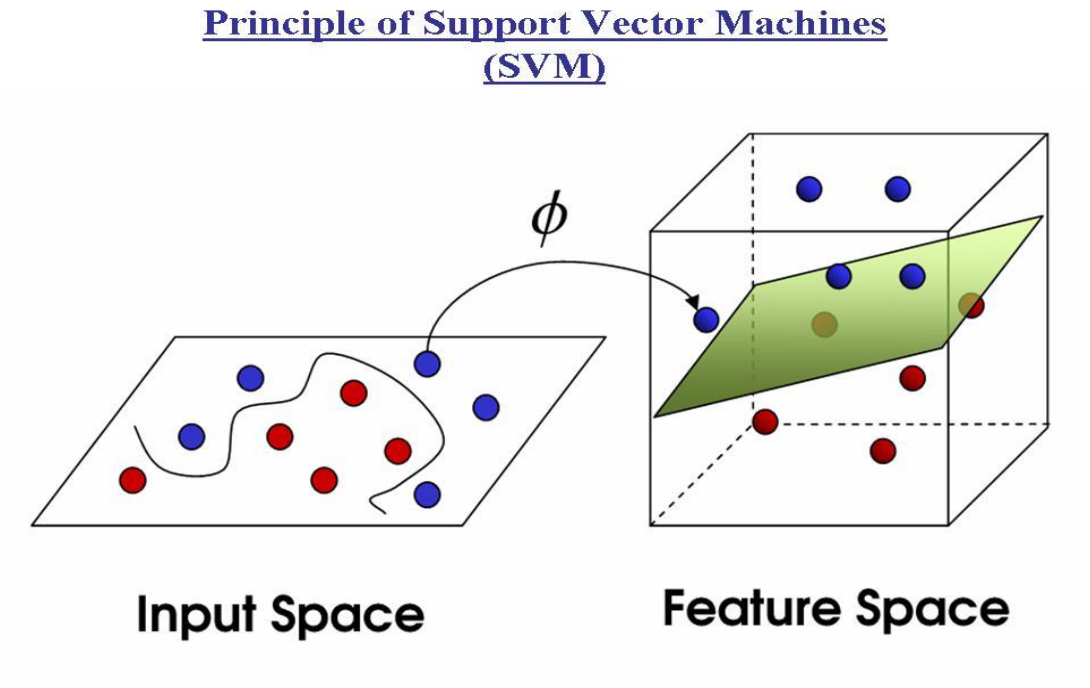


Figure 2.2, funzione di mappatura ϕ dei dati nello spazio trasformato

2.3.1 Metodo Svm

Le support vector machines presentano una combinazione unica di diversi concetti chiave:

1. L'iperpiano di margine massimo per trovare un classificatore lineare attraverso l'ottimizzazione

2. kernel trick per espandersi dalla classificazione lineare a quella non lineare

3. Soft-margin per far fronte al rumore presente nei dati non significativi.

Questo capitolo descrive la derivazione delle support vector machines come iperpiano di massimo margine usando la funzione Lagrangiana per trovare un estremo globale del problema. Innanzitutto, viene descritto un modello SVM con margini rigidi applicabile solo a dati linearmente separabili. Vengono in seguito formulate la forma primaria e duale del problema di ottimizzazione Lagrangiano. Verrà spiegato il motivo del perché la rappresentazione duale rappresenti un vantaggio significativo per l'ulteriore generalizzazione. Inoltre, viene descritta una generalizzazione non lineare tramite l'uso delle funzioni del kernel e la variazione "soft-margin" dell'algoritmo, calcolando infine il miscalssification error.

2.3.1.1 L'iperpiano di margine massimo

Partiamo dal presupposto che il training dataset sia linearmente separabile. Un iperpiano è definito da tutti i punti che soddisfano l'equazione $\mathbf{w} \cdot \mathbf{x} = 0$, caratterizzata da un prodotto interno e in cui w è un vettore perpendicolare all'iperpiano.

Stiamo cercando i parametri dell'iperpiano (w; b), in modo che la distanza tra l'iperpiano e le osservazioni sono massimizzate. Chiameremo la distanza euclidea tra il punto xi e l'iperpiano come margine geometrico.
Il margine geometrico di una singola osservazione xi è data dall'equazione:

$$\delta_i = \frac{y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} \tag{2.12}$$

Si noti che l'equazione sopracitata corrisponde realmente alla distanza perpendicolare del punto, mentre, solo il moltiplicatore yi sembra essere superfluo. Ma si ricorda che yi assegna +1 o -1. Il suo unico scopo nell'equazione dunque, è assicurare che la distanza non sia un numero negativo.

Dato il training set D, il margine geometrico di un iperpiano (w; b) rispetto a D è il più piccolo dei margini geometrici rispetto alle singole osservazioni del training set:

$$\delta = \min_{i=1\ldots m} \delta_i \tag{2.13}$$

Se stiamo cercando l'iperpiano di separazione ottimale, il nostro obiettivo dovrebbe essere quello di massimizzare il margine geometrico dell'addestramento - vale a dire posizionare l'iperpiano in modo tale che la sua distanza dai punti più vicini sia massimizzata. Tuttavia, questo è difficile da fare direttamente in quanto tale problema di ottimizzazione è un problema non convesso. Pertanto, i problemi devono essere riformulati prima di poter andare avanti. Il margine funzionale di una singola osservazione xi è data da:

$$\hat{\delta}_i = y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \tag{2.14}$$

Dato un training set D, il margine funzionale di (w; b) rispetto a D è il più piccolo dei margini funzionali rispetto alle singole osservazioni del training set:

$$\hat{\delta} = \min_{i=1\ldots m} \hat{\delta}_i \tag{2.15}$$

Si noti che c'è una relazione tra il margine geometrico e il margine funzionale:

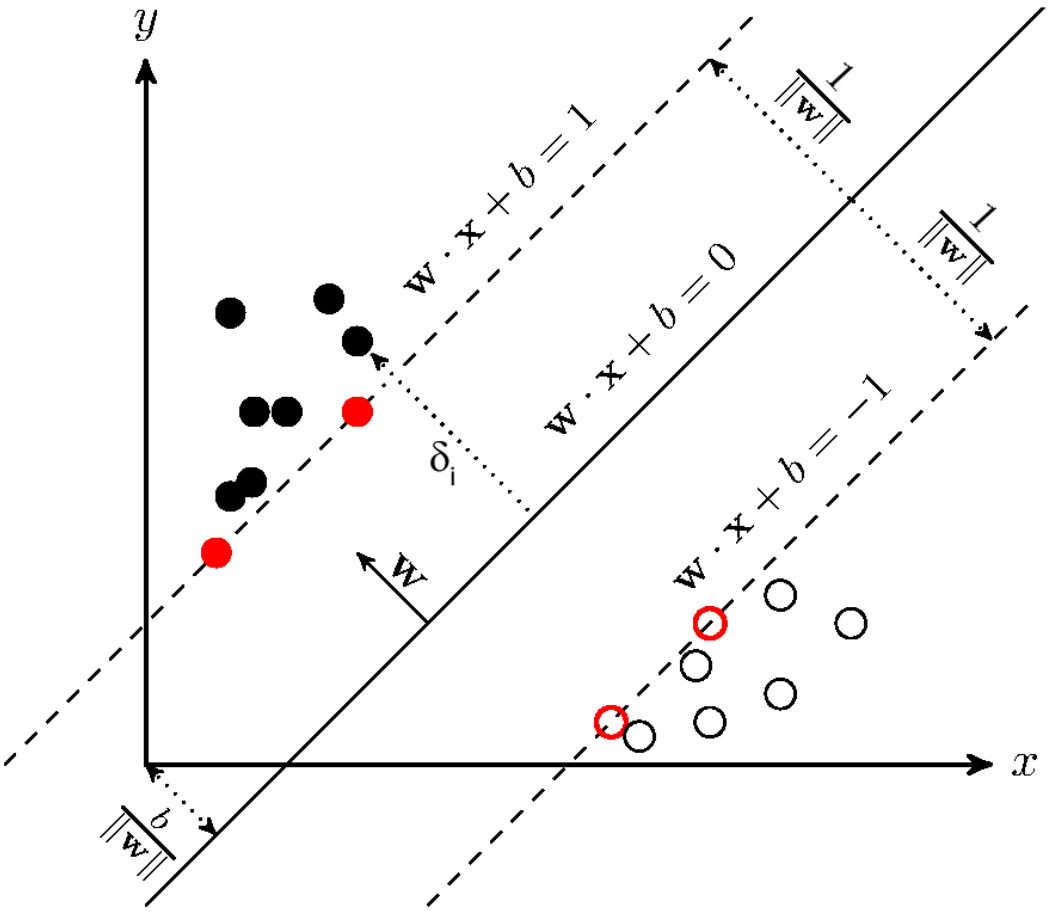
$$\delta = \frac{\hat{\delta}}{\|\boldsymbol{w}\|} \tag{2.16}$$

Si noti inoltre, che il margine geometrico è caratterizzato dall'invarianza di scala, rimpiazzando i parametri (w, b) con (2w, 2b), per esempio, il risultato non cambia. Il contrario è vero per il margine funzionale. Perciò, noi possiamo imporre un vincolo di scala arbitrario su w in modo che si adatti alle nostre esigenze.

Generalmente, si imposta il vincolo di scala in modo che il margine funzionale sia 1, e che il margine sia massimizzato sotto tale vincolo

$$\hat{\delta} = 1, \tag{2.17}$$

Quando vogliamo massimizzare il margine geometrico, dalla (2.16) segue che si debba massimizzare il termine $\frac{1}{\|w\|}$. Questo compito equivale a minimizzare il solo denominatore. Già questo è un compito abbastanza difficile da risolvere.



2.3, Margine geometrico e iperpiano di margine Massimo. La linea separa le due classi (cerchi pieno e cerchi vuoti) ed è definito dal vettore di pesi w che è perpendicolare alla linea, e dall'errore b, che determina la posizione della linea rispetto all'origine. La distanza tra ogni punto e la linea è chiamata margine geometrico ed è definito dalla (2.13).

Tuttavia, è possibile rimpiazzare questo termine con $\frac{1}{2}\|w\|^2$ senza cambiare la soluzione ottima. Il fattore moltiplicativo ½ è aggiunto per convenienza matematica nei prossimi passi.

$$\Phi(\boldsymbol{w}, b) = \frac{1}{2}\, \boldsymbol{w} \cdot \boldsymbol{w} \tag{2.18}$$

È necessario assicurarsi che l'equazione (2.15) rimanga valida quando la funzione obiettivo viene minimizzata. Il margine funzionale di ogni singola osservazione deve essere uguale o maggiore del margine funzionale dell'intero dataset.

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq \hat{\delta}, \ \forall i \tag{2.19}$$

Sostituendo la (2.17), si ottiene

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \ \forall i \tag{2.20}$$

Questa disuguaglianza pone un vincolo al problema di ottimizzazione (2.18).

La soluzione di questo problema sarà una superfice di decisione lineare con il massimo margine possibile (dato dal training set). La funzione obiettivo è chiaramente una funzione convessa, il che implica la possibilità di trovare il suo massimo. Il problema di ottizimizzazione generale consiste in una funzione obiettivo quadratica (2.18) e vincoli lineari (2.20) ed è perciò risolvibile direttamente con l'utilizzo della programmazione quadratica.

2.3.1.2 Metodi Lagrangiani per l'ottimizzazione

Dato che la programmazione risulta spesso inefficiente dal punto di vista computazionale quando si ha a che fare con grandi dataset, si continuerà nella risoluzione di questo problema di ottimizzazione usando i metodi Lagrangiani.

Partiamo in maniera più generale e poi applichiamo la teoria matematica al problema di ricerca dell'iperpiano di margine massimo.

Consideriamo il problema di ottimizzazione (primario).

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f(\boldsymbol{w}) \\ s.t. \quad & g_i(\boldsymbol{w}) \leq 0 \\ & h_i(\boldsymbol{w}) = 0 \end{aligned} \tag{2.21}$$

La minimizzazione è vincolata da un numero arbitrario di vincoli di uguaglianza e disuguaglianza. Un metodo per risolvere un problema del genere consiste nel formulare un problema derivato, che ha la stessa soluzione di quello originale. Definiamo la funzione Lagrangiana generalizzata del (2.21) come:

$$\mathcal{L}(\boldsymbol{w}, \alpha, \beta) = f(\boldsymbol{w}) + \sum \alpha_i g_i(\boldsymbol{w}) + \sum \beta_i h_i(\boldsymbol{w}) \tag{2.22}$$

Definiamo inoltre

$$\theta_P(\boldsymbol{w}) = \max_{\alpha,\beta;\alpha_i \geq 0} \mathcal{L}(\boldsymbol{w}, \alpha, \beta) \tag{2.23}$$

Questo permette di formulare il cosiddetto "problema di ottimizzazione primario", denotato dal pedice P nella seguente riga:

$$p^* = \min_{\boldsymbol{w}} \theta_P(\boldsymbol{w}) = \min_{\boldsymbol{w}} \max_{\alpha,\beta;\alpha_i \geq 0} \mathcal{L}(\boldsymbol{w}, \alpha, \beta) \tag{2.24}$$

Si noti che la funzione θp assume come valore più infinito, in caso non siano verificati i vincoli presenti nella (2.21). Se $g_i(w) > 0$ per ogni i, la (2.23) può essere massimizzata semplicemente ponendo $\alpha = +\infty$. Se d'altra parte $h_i(w) \neq 0$, la (2.23) è massimizzata impostando β come più o meno infinito (dipende dal segno della funzione hi).

Se tutti i vincoli sono rispettati, θp assume lo stesso valore della funzione obiettivo del problema originale, *f*(w). A condizione che entrambi i vincoli reggano, il termine della terza

equazione nella (2.22) è sempre zero e anche il secondo termine è massimizzato quando è azzerato con l'impostazione appropriata dei pesi αi. Quindi abbiamo:

$$\theta_P(\boldsymbol{w}) = \begin{cases} f(\boldsymbol{w}) & \text{if constraints are satisfied} \\ +\infty & \text{otherwise} \end{cases} \tag{2.25}$$

È ovvio che la (2.24) è, invece, uguale al problema originale (2.21).

Con il metodo appena descritto si può risolvere il problema primario, ma le proprietà dell'algoritmo non porterebbero molti benefici. Invece, per il bene delle prossime soluzioni, definiamo un diverso problema di ottimizzazione cambiando l'ordine di minimizzazione e massimizzazione.

$$\theta_D(\alpha, \beta) = \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \alpha, \beta) \tag{2.26}$$

Il pedice D sta per "duale" e sarà usato per formulare il cosiddetto "problema di ottimizzazione duale":

$$d^* = \max_{\alpha,\beta;\alpha_i \geq 0} \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \alpha, \beta) = \max_{\alpha,\beta} \theta_D \tag{2.27}$$

I valore del problema primario *p** e del problema duale *d** non sono necessariamente uguali. Esiste una relazione generale valida tra il massimo della minimizzazione e il minimo della massimizzazione

$$d^* \leq p^* \tag{2.28}$$

La differenza tra la soluzione primaria e quella duale, *p** - *d** è chiamata "*duality gap*". Come segue dalla (2.28), questo valore è sempre maggiore o uguale a zero.

Alcune condizioni sono state formulate che (se soddisfatte) garantiscono la nullità del "duality gap", cioè in tali condizioni, la soluzione del problema duale è uguale alla soluzione del problema originale.

Si suppone che *f* e tutte le funzioni g_i siano convesse e tutte le h_i siano affini. Supponiamo che i vincoli siano rigorosamente rispettati (es che $g_i(w) < 0$ per ogni i). Allora esiste w*, α*, β* tali che w* sia la soluzione del problema primario, α*, β* sia la soluzione del problema duale, $p^* = d^* = \mathcal{L}$ (w*, α*, β*) e w*, α*, β* soddisfino le seguenti condizioni di **Karush-Kuhn-Tucker** (KKT):

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0, \quad i = 1, \ldots, n \tag{2.29}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_i} = 0, \quad i = 1, \ldots, n \tag{2.30}$$

$$\alpha_i^* g_i(\boldsymbol{w}^*) = 0, \quad i = 1, \ldots, n \tag{2.31}$$

$$g_i(\boldsymbol{w}^*) \leq 0, \quad i = 1, \ldots, n \tag{2.32}$$

$$\alpha_i^* \geq 0, \quad i = 1, \ldots, n \tag{2.33}$$

Per la sua importanza, la condizione KKT (2.31) è definita spesso come "condizione complementare".

2.3.1.3 Applicazione iperpiano di margine massimo

Ritorniamo al compito originale definito dalle equazioni (2.18), (2.20). I vincoli devono essere riscritti per adattarsi alla forma della (2.21)

$$g_i(\boldsymbol{w}) = -y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) + 1 \leq 0 \tag{2.34}$$

Il Langrangiano della funzione obiettivo (2.18) soggetto alla (2.20) sarà, in accordo con la (2.22)

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, b) = \frac{1}{2}\boldsymbol{w} \cdot \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i \left(y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1\right) \tag{2.35}$$

Il duale può essere trovato in maniera abbastanza semplice. Per fare questo, abbiamo bisogno di minimizzare $\mathcal{L}$. Con la (2.29) calcoliamo le derivate parziali di $\mathcal{L}$ rispetto a *w* e *b*

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{2.36}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.37}$$

Sostituendo queste equazioni nella (2.35) e semplificando il risultato

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, b) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \; \boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{2.38}$$

Questa equazione è stata ottenuta minimizzando L rispetto a w e b. In accordo con la (2.27), questa espressione deve essere adesso massimizzata. Mettendola insieme alla (2.33) e (2.37), il problema di ottimizzazione duale è:

$$\begin{aligned} \max_{\alpha} \quad & \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \, \boldsymbol{x}_i \cdot \boldsymbol{x}_j \right) \\ s.t. \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \; i = 1, \ldots, l \end{aligned} \tag{2.39}$$

Si può dimostrare, che le condizioni KKT sono soddisfatte e che la (2.39) presenta una soluzione reale al problema originale.

È giusto sottolineare come il vincolo (2.20) rafforzi il fatto che il margine funzionale di ciascuna osservazione debba essere uguale o maggiore del margine funzionale del dataset (che abbiamo impostato uguale a 1 nella (2.17).

Sarà esattamente uguale a 1 solo per i punti più vicini all'iperpiano. Tutti gli altri punti che giacciono lontano perciò avranno un valore del margine funzionale maggiore di 1. Con la "condizione complementare" (2.31) i corrispondenti moltiplicatori di Lagrange α_i saranno uguali a zero per ogni punto.

Perciò la soluzione al problema, il vettore dei pesi w nella (2.36), dipende solo da pochi punti che non presentano zero come α_i, e che giacciono più vicino all'iperpiano di separazione. Questi punti sono chiamati "***support vectors***".

La soluzione del problema duale (2.39) porta ad un vettore ottimo α^*, che può essere usato immediatamente per calcolare il vettore dei pesi dell'iperpiano di separazione w usando la (2.36). Dato che la maggior parte dei coefficienti α_i saranno zero, il vettore dei pesi sarà una combinazione lineare di solamente pochi "support vector".

Avendo calcolato l'ottimo w*, l'errore b* deve essere trovato usando i vincoli primari, in modo che il valore b non appaia nel problema duale:

$$b^* = -\frac{\max_{y_i=-1} \boldsymbol{w}^* \cdot x_i + \min_{y_i=+1} \boldsymbol{w}^* \cdot x_i}{2} \tag{2.40}$$

Questo porta ad una funzione di decisione che può essere usata per classificare le nuove osservazioni x al di fuori del campione:

$$\hat{f}(\boldsymbol{x}) = sgn\left(\sum_{i=1}^{m} \alpha_i^* y_i \boldsymbol{x}_i \cdot \boldsymbol{x} - b^*\right) \tag{2.41}$$

Come detto precedentemente, questo classificatore è completamente determinato dai support vectors. A causa di questa proprietà, è chiamato "**support vector machine**". Per essere più precisi, un support vector machine lineare, in quanto è basato su una superficie di decisione lineare. Questo è capace dunque, di classificare soltanto dati separabili linearmente. D'altra parte, a differenza di altri classificatori, assicura la convergenza e di trovare l'iperpiano di margine massimo. Dato che il problema di ottimizzazione è convesso e le condizioni KKT sono valide, c'è sempre una soluzione unica ed è un estremo assoluto.

Vi è una proprietà della funzione di decisione (2.30) che merita particolare attenzione. Le osservazioni appaiono solo come prodotti tra punti tra i vettori in input. Questo porta all'uso di una tecnica speciale che i ricercatori spesso chiamano come "kernel trick". Grazie a questa tecnica, il support vector machine ricavato in questa sezione può essere esteso e generalizzato per adattarsi a quasi ogni superficie di decisione non lineare.

2.3.1.4 Kernel e Kernel trick

Il problema con i dataset presenti nella vita reale è che raramente la relazione tra le variabili di input è semplice e lineare. Nella maggior parte dei casi sarebbero necessarie alcune superfici di decisione non lineari per separare correttamente le osservazioni in due gruppi. Questa necessità, d'altra parte, aumenta notevolmente la complessita e la difficoltà computazionale di questo compito.

Invece, le support vector machines implementano un'altra idea concettuale: i vettori di input sono mappati da uno spazio di input originale in un nuovo spazio (future space) multidimesionale attraverso alcune funzioni non lineari scelte a priori. Nel "future space" poi viene costruita la superficie di decisione lineare.

La giustificazione sottostante può essere trovata nel teorema di Cover, che può essere qualitativamente espresso come segue: "un problema di classificazione complesso, formulato attraverso una trasformazione non-lineare dei dati in uno spazio ad alta

dimensionalità, ha maggiore probabilità di essere linearmente separabile che in uno spazio a bassa dimensionalità".

Dato che ora vogliamo operare nel "future space" piuttosto che nello spazio di input originale, possiamo leggermente modificare l'equazione originale (2.39) e sostituire i prodotti scalari xi·xj con *ϕ(xi)·ϕ(xj).* Con *ϕ()* indichiamo delle appropriate funzioni scelte per mappare i dati.

Si consideri che le trasformazioni *ϕ(x)* possono essere molto complesse. I risultati dovrebbero appartenere a spazi multidimensionali, in alcuni casi anche a spazi di dimensione infinita. Queste trasformazioni potrebbero essere dunque molto pesanti dal punto di vista computazionale, o addirittura impossibili.

L'idea di base di questo concetto è spiegata dal fatto che per trovare l'iperpiano di margine massimo nel "future space", non c'è bisogno di calcolare esplicitamente i vettori delle trasformazioni ϕ(xi), ϕ(xj) in quanto sono presenti solo come prodotto scalare nell'algoritmo di training.

È sufficiente utilizzare la cosiddetta funzione "Kernel" che è definita come il prodotto scalare tra due vettori nello spazio trasformato e il quale può essere calcolato senza tanta capacità computazionale.

Data un'appropriata funzione di mappatura $\phi : \mathbb{R}^n \to \mathbb{R}^m$ con $m \geq n$, le funzioni di forma

$$K(x, z) = \phi(x) \cdot \phi(z) \tag{2.42}$$

con x e z appartenenti al primo spazio, sono chiamate funzioni Kernel o semplicemente Kernel.

Il processo di mappatura dei dati dallo spazio di input originale al "future space" è chiamato "Kernel trick". Il Kernel trick è un approccio piuttosto universale, la sua applicazione non è limitata al solo utilizzo nelle support vector machines. In generale, ogni classificatore che dipende dal solo prodotto scalare dei dati di input può capitalizzare questa tecnica.

Riformuliamo ora le support vector machines usando le funzioni Kernel. L'algoritmo di training (2.39) può essere espresso ora come:

$$\begin{aligned} \boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\text{argmax}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ s.t. \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \; i = 1, \ldots, l \end{aligned} \tag{2.43}$$

mentre la superficie di decisione (2.41) in termini di funzione Kernel è definita come

$$\hat{f}(\boldsymbol{x}) = sgn\left(\sum_{i=1}^{m} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}) - b^*\right) \tag{2.44}$$

2.3.1.5 Kernel Lineare

Ci sono molte funzioni Kernel note che sono comunemente usate in connessione alle support vector machines. Il primo esempio è rappresentato dal Kernel lineare. Questo è definito semplicemente come

$$K(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{z} \tag{2.45}$$

In questo caso, il "future space" e lo spazio di input sono gli stessi e ritorniamo alla soluzione ricavata in precedenza, quando abbiamo discusso delle support vector machines lineari.

Il Kernel lineare è stato citato per enfatizzare il fatto che la sostituzione con una funzione Kernel non cambia del tutto l'algoritmo originale. Semplicemente lo generalizza.

2.3.1.6 Kernel Gaussiano (Radial basis function)

La funzione Kernel gaussiana è probabilmente la funzione Kernel più utilizzata ed è definita come

$$K(\boldsymbol{x}, \boldsymbol{z}) = exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2}{2\sigma^2}\right\} \tag{2.46}$$

Questo mappa lo spazio input in uno spazio di dimensione infinita. Grazie a questa proprietà, è molto flessibile e permette di ottenere il fitting di un'ampissima varietà di bordi decisionali. Il kernel gaussiano è spesso definito come "radial basis function" (RBF). In alcuni casi è parametrizzato in modo leggermente diverso:

$$K(x,z) = exp\ \{-\gamma\|x-z\|^2\} \tag{2.47}$$

Ci sono molte altre funzioni Kernel ricercate. Molte di loro, comunque, sono usate solo per scopi di ricerca o in speciali circostanze e non hanno alcun ruolo nei modelli di credit scoring. Affinchè una funzione K sia un Kernel valido, deve essere verificata una condizione necessaria e sufficiente. Questa condizione è 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, "future space", e garantisce l'esistenza di una soluzione unica e la convessità del problema.

2.3.1.7 Classificatori "Soft Margin"

Fino ad ora abbiamo assunto che i dati di training siano linearmente separabili, sia direttamente nello spazio di input, sia nello spazio trasformato. In altre parole abbiamo supposto di avere a che fare con dei dataset perfetti. In realtà i dataset sono spesso molto lontani dalla perfezione. I dati reali sono affetti da errori, contengono osservazioni sbagliate, errori casuali ecc.

Abbiamo bisogno, quindi, di generalizzare l'algoritmo in modo che possa gestire anche i dati non separabili. Abbastanza sorprendentemente, l'idea di base della generalizzazione è molto semplice. Abbiamo bisogno di permettere ad alcune osservazioni di finire nel lato sbagliato dell'iperpiano di separazione.

La proprietà richiesta è soddisfatta indtroducendo la variabile slack ξ e riformulando i vincoli di partenza (2.20)

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b) \geq 1-\xi_i,\ i=1,\ldots,l \tag{2.48}$$

Dove $\xi_i \geq 0$. In questo modo il margine funzionale di alcune osservazioni sarà minore di 1.

La variabile slack lavora come un termine correttivo nell'equazione. Per le osservazioni localizzate dal lato corretto dell'iperpiano, il corrispondente $\xi_i = 0$ e la (2.48) è uguale alla (2.9). Per le osservazioni errate o classificate in maniera sbagliata, che compaiono quindi nel lato sbagliato dell'iperpiano, il corrispondente $\xi_i > 0$, come mostrato nella figura seguente.

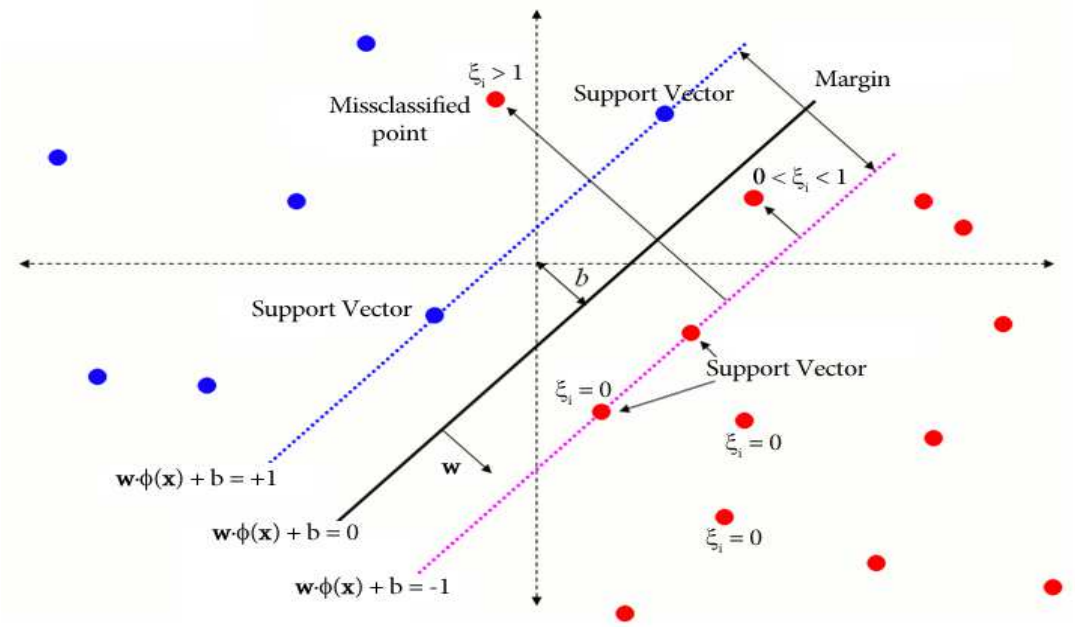


Figure 2.4, Lo scopo della variabile slack spiegato attraverso un semplice esempio.

È necessario che per ogni classificazione errata si paghi un prezzo, in quanto in generale vogliamo massimizzare il margine minimizzando simultaneamente l'effetto avverso delle errate classificazioni. Questo è il motivo per il quale dobbiamo aggiungere alla funzione obiettivo originale un termine di penalità

$$\min_{w,\xi,b}\left(\frac{1}{2}w\cdot w + C\sum_{i=1}^{m}\xi_i\right) \tag{2.49}$$

dove $C > 0$ è il parametro di costo, che determina l'importanza delle osservazioni errate. Piccoli valori di C portano ad una soluzione più "morbida", mentre alti valori di C portano a soluzioni simili ai classificatori "hard margin" visti precedentemente.

2.4 Neural Networs

Le reti neurali sono una famiglia di tecniche di machine learning che si ispira al modo in cui i sistemi nervosi biologici, come il cervello, elaborano le informazioni. L'elemento chiave di questo paradigma è la nuova struttura del sistema di elaborazione delle informazioni. È composto da un gran numero di elementi di elaborazione altamente interconnessi (i neuroni) che lavorano all'unisono per risolvere problemi specifici. I neuroni sono delle funzioni matematiche, dette primitive, delle variabili esplicative. Le reti neurali, con la loro notevole capacità di derivare il significato da dati complicati o imprecisi, possono essere utilizzate per estrarre schemi e rilevare tendenze troppo complesse per essere notate dagli umani o da altre tecniche informatiche. Una rete neurale addestrata può essere quindi utilizzata per fornire proiezioni date nuove situazioni di interesse.

Una rete neurale è dunque uno schema di regressione non lineare a due o più stadi costituito da strati di neuroni che, collegati tra loro, mettono in relazione le variabili di input con quelle di output. La capacità di imparare delle reti neurali, le rende uno strumento molto flessibile e potente. Inoltre non è necessario ideare un algoritmo per eseguire un compito specifico, vale a dire, non c'è bisogno di capire i meccanismi interni di quel compito. Per la loro capacità di avere molti livelli (e quindi parametri) con non linearità, si rendono molto efficaci nella modellazione di relazioni non lineari altamente complesse.

La ricerca inoltre, ha costantemente dimostrato che il semplice fatto di fornire alla rete più dati sull'addestramento, siano essi totalmente nuovi o dall'aumento del set di dati originale, avvantaggia le prestazioni della rete.

Pur avendo diversi vantaggi, questi modelli presentano anche alcuni svantaggi. A causa della loro complessità, questi modelli non sono facili da interpretare e capire. Possono essere abbastanza impegnativi e computazionalmente intensi da addestrare, richiedendo un'accurata sintonizzazione iperparametrica e l'impostazione del programma di velocità di apprendimento. Richiedono molti dati per ottenere prestazioni elevate e sono generalmente sovraperformati da altri algoritmi Machine Learning in casi di piccoli volumi di dati. L’arbitrarietà nella scelta dei parametri, la scarsa accuratezza, e la difficoltà nell’interpretazione dei risultati frenano infatti l’applicazione di questi modelli.

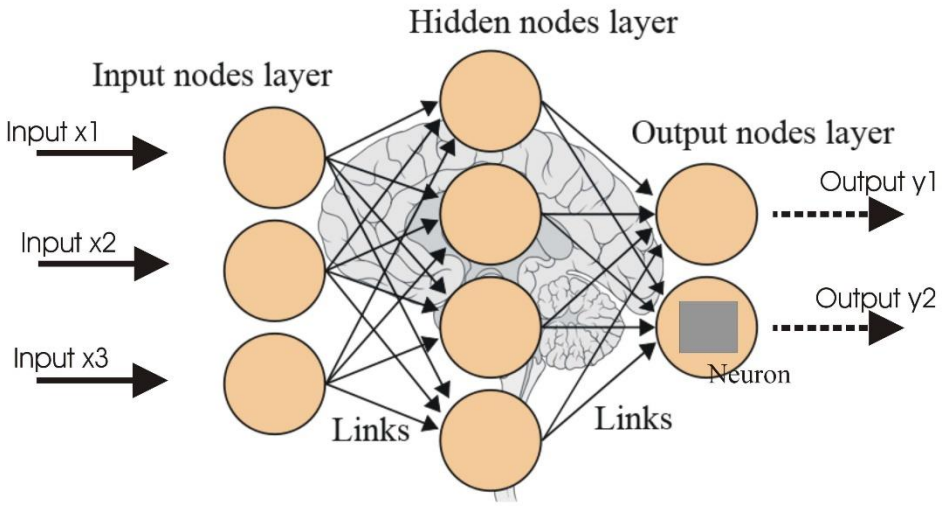


Figure 2.5, Struttura logica delle Neural Networks

3 Il Caso del Settore delle Costruzioni

Per sviluppare i modelli di credit scoring, si è scelto di considerare un campione di società italiane all'interno del settore dell'edilizia "privata". È stato scelto un campione omogeneo di società appartenenti allo stesso settore, e in particolare al comparto delle costruzioni residenziali e non residenziali. La natura delle società selezionate fa sì che queste presentino un processo di concessione del credito molto simile. Queste società hanno fortemente risentito della crisi finanziaria globale scoppiata nel 2007/2008, e un gran numero di queste infatti, presenta nel periodo oggetto di studio una situazione di difficoltà patrimoniale e finanziaria. Ciò ha reso il campione valido per rappresentare sia società sane, sia società insolventi o anomale. Per effettuare questo studio sono stai utilizzati i dati di bilancio di circa 37'000 società per gli anni che vanno dal 2012 al 2016, coprendo un orizzonte di osservazione di cinque anni.

3.1 Il settore dell'edilizia in Italia nel 2012

Il Settore delle Costruzioni in Italia nel 2012 conosceva un forte periodo di recessione. Secondo l'osservatorio congiunturale sull'industria delle costruzioni effettuato dall'ANCE (Associazione Nazionale Costruttori Edili) "in Italia nel corso del 2012, abbiamo vissuto un forte peggioramento della situazione di crisi del settore delle costruzioni. Tutti gli indicatori settoriali disponibili danno evidenza della gravità della situazione del mercato con intensità di cadute simili a quelle registrate nel 2009 e cioè nella fase iniziale della crisi."[11] Preoccupante risulta il basso livello del portafoglio ordini, oltre alla netta prevalenza dei giudizi negativi legati alla percezione del progressivo deterioramento del quadro settoriale: il 73,7% delle aziende valuta bassa la consistenza del proprio portafoglio ordini contro il 24,3% che ne riscontra la normalità e il residuo 1,9% che la ritiene elevata. Molte aziende dichiarano il proprio portafoglio ordini insoddisfacente e, a prova di ciò, la consistenza degli ordinativi risulta in netto declino rispetto alle rilevazioni effettuate nel 2011. "Nel 2012 gli investimenti in costruzioni secondo l'Ance, registrano una flessione del 7,6% in termini reali che risulta più sostenuta di quella rilevata nel 2011 (-5,3%) e peggiorativa rispetto alla stima

[11] http://leg16.camera.it/temiap/temi16/ance_osservatorio_dic2012.pdf

rilasciata nel giugno 2012 (-6,0%).”[12]

INVESTIMENTI IN COSTRUZIONI(*)

	2012(°) Milioni di euro	2008	2009	2010 (°)	2011 (°)	2012 (°)	2013(°)	2008-2012(°)	2008-2013(°)
		Variazioni % in quantità							
COSTRUZIONI	**130.679**	**-2,4%**	**-8,6%**	**-6,6%**	**-5,3%**	**-7,6%**	**-3,8%**	**-27,1%**	**-29,9%**
.abitazioni	69.577	-0,4%	-8,1%	-5,1%	-2,9%	-6,3%	-2,7%	-21,0%	-23,1%
- nuove (°)	*24.757*	*-3,7%*	*-18,7%*	*-12,4%*	*-7,5%*	*-17,0%*	*-13,0%*	*-47,3%*	*-54,2%*
- manutenzione straordinaria(°)	*44.820*	*3,5%*	*3,1%*	*1,1%*	*0,5%*	*0,8%*	*3,0%*	*9,3%*	*12,6%*
.non residenziali	61.102	-4,4%	-9,1%	-8,1%	-7,9%	-9,1%	-5,1%	-33,2%	-36,6%
- private (°)	*36.281*	*-2,2%*	*-10,7%*	*-5,4%*	*-6,0%*	*-8,0%*	*-4,2%*	*-28,6%*	*-31,6%*
- pubbliche (°)	*24.821*	*-7,2%*	*-7,0%*	*-11,5%*	*-10,5%*	*-10,6%*	*-6,5%*	*-38,9%*	*-42,9%*

(*) Investimenti in costruzioni al netto dei costi per trasferimento di proprietà

(°) Stime Ance

Figure 3.1, Elaborazioni Ance su dati Istat relative agli investiomenti nelle costruzioni

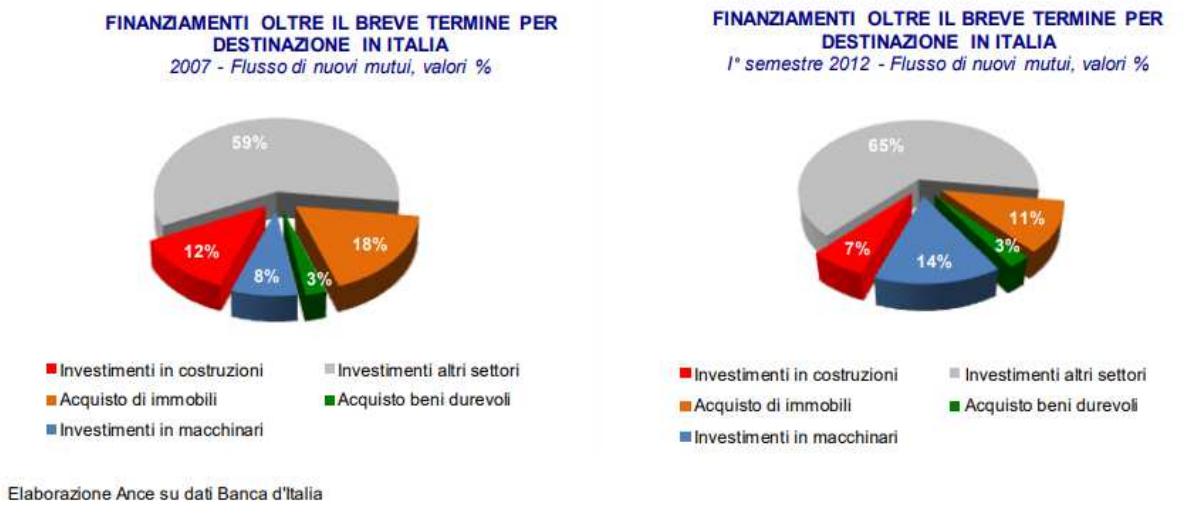
Nel giro di sei anni, dal 2008 al 2013, il settore delle costruzioni perde circa il 30% degli investimenti. In questo scenario sono rari i comparti in cui non si avverte una forte sofferenza; si va dal 54,2% perso dal comparto della produzione di nuove abitazioni, all’edilizia non residenziale privata che registra una riduzione di circa il 32%. L’unica eccezione è rappresentata dal comparto della riqualificazione degli immobili residenziali, che mantiene un segno positivo del 12,6%.

Analizzando l’impatto della crisi nel processo di concessione del credito, si registra in questo periodo un sensibile razionamento del credito a cui va aggiunto il ritardo nei pagamenti da parte della pubblica amministrazione. “A giugno 2012, la stretta del credito nei confronti del settore delle costruzioni ha raggiunto il livello più alto dall’inizio della crisi. La stessa BCE, analizzando le condizioni di accesso al credito (Bank Lending Survey), in base ai dati provenienti dalle singole banche centrali che aderiscono all’Euro, afferma che le condizioni applicate per l’erogazione di finanziamenti alle PMI da parte delle banche italiane sono state generalmente tra le più rigide tra i 17 Paesi aderenti.”[13]

[12] http://leg16.camera.it/temiap/temi16/ance_osservatorio_dic2012.pdf

[13] http://leg16.camera.it/temiap/temi16/ance_osservatorio_dic2012.pdf

Questo razionamento ha causato un calo dei finanziamenti a medio-lungo termine di circa l'8% dal 2007 al 2011. È importante notare come le costruzioni, nel confronto con gli altri settori economici, sono quelle che hanno subito la restrizione maggiore da parte delle istituzioni creditizie. Come si vede facilmente dal grafico sottostante, nel confronto tra il



Elaborazione Ance su dati Banca d'Italia

Figure 3.2, Distribuzione dei finanziamenti a medio lungo termine nel 2007 e primo semestre 2012

periodo 2007-2011 e il primo semestre del 2012, la situazione risulta peggiorata. I finanziamenti a medio lungo termine sono scesi del 9%, mentre nel settore abitativo si è avuto un'ulteriore restrizione del 20%, nel non residenziale il calo è stato del 33%. Gli investimenti in macchinari e attrezzature, invece, nello stesso semestre, hanno visto un ulteriore incremento: +56%.

Questa situazione ha portato inevitabilmente all'aumento della domanda di credito, e ad un continuo deterioramento del rapporto sofferenze-impieghi, soprattutto per il settore delle costruzioni. Un tale razionamento da parte delle banche ha portato nel comparto dell'edilizia privata ad una situazione di crisi indotta con gravi sofferenze da parte delle imprese.

3.2 L'evoluzione del settore

La situazione dell'edilizia in Italia oggi si trova in fase di stallo, in controtendenza con l'andamento generale dell'economia. Sembra esserci infatti una sorta di crisi permanente dovuta tra l'altro alla mancanza di investimenti pubblici nel settore. Tuttavia, il settore costituisce una voce importante nel bilancio dello Stato, andando a generare nel 2017 il 4,5% del valore aggiunto italiano, pari circa a 64 miliardi di euro, e rappresentando il 6,1% dell'occupazione. Questi numeri descrivono un ridimensionamento significativo rispetto ai valori pre-crisi del 2007, in cui il settore aveva registrato il più alto valore aggiunto degli ultimi 15 anni, con oltre 94 miliardi di euro.

La crisi economica ha portato ad un crollo della domanda per le costruzioni residenziali a causa anche delle restrizioni avvenute nella concessione del credito e alla contrazione dei redditi. Il razionamento del credito nell'ambito privato e le politiche economiche restrittive riguardanti il settore dell'edilizia hanno avuto portato le costruzioni non residenziali a dover fronteggiare un ciclo economico sfavorevole. Tuttavia secondo i dati più recenti, dal 2015 l'economia italiana ha segnato un graduale, seppur lieve, miglioramento, e questo ha portato degli effetti anche al settore edilizio, soprattutto nel 2017.

Secondo il centro studi ANCE, "gli investimenti privati in costruzioni non residenziali hanno segnato, nel 2016, un andamento lievemente migliore alle previsioni: +0,8% in termini reali rispetto al +0,2% previsionale. Il dato è dovuto al migliorato contesto economico del Paese e all'accresciuta quantità di Permessi a costruire per edilizia non residenziale privata che, nel 2015, avevano raggiunto un +14,1% rispetto all'anno precedente. Miglioramento nel numero dei permessi a costruire che invece, nei primi 9 mesi 2016, sono tornati a decrescere, reagendo poi con un lieve aumento nel terzo trimestre che non è stato però sufficiente a compensare i cali dei mesi precedenti"[14].

[14] http://www.ceramicworldweb.it/filealbum/Home/materialicasa/pdf/tile-italia/2017/1/040_043%20Edilizia%20ANCE%20CORR_1.pdf

Figure 3.3 Analisi degli investimenti nel comporto costruzioni non residenziali

Il miglioramento citato, non ha fermato però la tendenza negativa del settore. Sebbene l'indice della produzione nazionale si sia stabilizzato, dopo anni di contrazione, non ha ancora mostrato un segno di sensibile ripresa. Dalle analisi inoltre, risultano importanti differenze tra il comparto edilizio residenziale e quello non residenziale.

La moderata ripresa avvertita nel settore delle costruzioni non è di certo trainata dall'edilizia residenziale, che rimane ferma al milione di metri quadri, mentre quella non residenziale prosegue la tendenza alla crescita iniziata nel 2015.

L'analisi degli investimenti nel settore edilizio mostra due dinamiche di grande rilevanza. Tra il 2013 e il 2016, il valore degli investimenti nel settore è diminuito, ma con un'intensità via via minore, facendo registrare nel 2017 un lievissimo incremento. Gli studi effettuati hanno confermato "la ripresa dell'edilizia non residenziale, a fronte del perdurare della crisi di quella residenziale; anche se, nell'ambito della seconda, radicalmente opposta è la tendenza registrata tra nuove abitazioni e manutenzione straordinaria. Gli investimenti in manutenzione straordinaria sono infatti nel 2017 più alti di circa il 40% rispetto al 2000, mentre quelli in nuove edificazioni sono più bassi del 70% nel 2017 rispetto al picco del 2007"[15].

In questo contesto, a fronte degli ultimi anni di forte crisi, il 2017 è sicuramente un anno positivo per il settore edilizio. A far registrare questo risultato hanno contribuito, seppure in modo parziale, la ripresa del mercato immobiliare, le politiche di incentivazione, la ripresa del comparto non residenziale e l'aumentare delle concessioni di mutui alle famiglie. La crisi economica ha avuto effetti devastanti su questo settore che in questo

[15] http://www.rassegna.it/articoli/costruzioni-un-difficile-percorso-di-uscita-dalla-crisi

momento cerca di riprendersi, conservando comunque notevoli fragilità dal punto di vista economico e occupazionale.

3.3 Il campione

Nel panorama economico italiano di piena crisi, descritto nei paragrafi precedenti, si collocano le società del settore edilizio che sono alla base di questo studio. Per ottenere le informazioni e i dati economici appartenenti alle società si è sfruttato l'accesso alla banca dati digitale Aida. Aida contenente i bilanci riclassificati di oltre 700.000 società di capitale italiane, e permette selezioni per settore di attività, area geografica, elaborazione dei dati societari, e altre attività. Le società di cui si sono prelevati i dati sono identificate dal codice "Ateco 2007" 412 che identifica il settore delle Costruzioni di edifici residenziali e non residenziali, filtrando le stesse con il vincolo del numero di dipendenti minimo uguale a uno. Nel nostro caso si sono selezionate come variabili da esportare, i dati anagrafici, i dati di bilancio, e eventuali procedure in corso. Alcuni esempi delle procedure considerate sono: "liquidazione volontaria", "scioglimento e liquidazione", "concordato preventivo" ecc. Alle società in condizioni particolari o anomale è stato assegnato successivamente un flag "1" che indicasse l'anomalia. Le informazioni così ottenute sono state inserite all'interno di una matrice in un foglio di calcolo Excel per poter proseguire con le operazioni necessarie di elaborazione e controllo dei dati. Partendo dai dati di bilancio appartenenti a ciascuna società si sono ricavati, pero ognuna di essa, più di 40 indicatori finanziari. Prima di fare ciò però, si è reso necessario controllare ed eventualmente correggere errori contenuti all'interno del set di dati scaricato dalla banca dati.

Un aspetto che in questi casi è sempre molto critico, è quello relativo alla pulizia dei dati. I dati così come vengono scaricati infatti, sono affetti da errori, imprecisioni e in alcuni casi anche da parti del tutto mancanti. È importante dunque che il set di dati iniziale venga adeguatamente preparato prima di poter essere utilizzato come base di un'analisi statistica. Questa operazione è dunque necessaria al fine di rendere i dati significativi dal punto di vista statistico, e quindi pronti per essere utilizzati per analisi e simulazioni specifiche. Il processo è spesso lungo e dispendioso, in quanto la pulizia deve essere effettuata in modo accurato per non compromettere il set di informazioni racchiuse

all'interno dei dati. Nel caso in esame si è proceduto con specifiche fasi di controllo, sequenziali, che hanno permesso di individuare e correggere degli errori presenti nel database iniziale.

Gli indicatori, utilizzati come base per l'analisi statistica, sono stati calcolati utilizzando gli strumenti logici e matematici forniti dal foglio di calcolo, e si possono suddividere per ambito di interesse in:

- Indici di redditività
- Indici di produttività e struttura operativa
- Indici di Liquidità e struttura finanziaria.

I risultati ottenuti sono stati inseriti in una seconda matrice, la quale è servita per effettuare ulteriori fasi di controllo. La formazione di questi indicatori, a partire dai dati inziali, ha fatto emergere infatti, in alcuni casi, ulteriori criticità. Gli indicatori finanziari sono stati utilizzati per formare il Database che è stato utilizzato come input all'interno del programma statistico per produrre le simulazioni dei modelli di credit scoring.

Per poter utilizzare un Database all'interno di un software di programmazione, è fondamentale che questo rispetti tutti i criteri necessari alla codifica all'interno del programma. Per ottenere questo risultato è stato necessario svolgere un lavoro di rifinitura affinché non ci fosse nessun campo del Database vuoto e che avesse invece un valore significativo leggibile correttamente all'interno del codice.

Tornando alle variabili costitutive del modello, come detto precedentemente, ad ogni società del campione è stato assegnato un "flag" di tipo 0/1, solvente/insolvente, rappresentativo dello status della stessa. La variabile "flag" è quella che viene identificata all'interno del modello come variabile risposta, o variabile dipendente. Il set di indicatori derivati dai dati di bilancio e conto economico del set di partenza invece, costituiscono l'insieme delle variabili indipendenti, che spiegano la variabile "flag".

Uno dei problemi alla base di questi modelli è costituito dalla scelta ottima delle variabili da includere nel modello. Non tutte le variabili sono ugualmente rilevanti all'interno dei modelli di analisi di credit scoring. In particolare, in presenza di molti indicatori fortemente correlati tra loro, anche un piccolo gruppo di essi può contenere molte informazioni utili alla classificazione. In generale dunque, aggiungere al modello variabili fortemente correlate con quelle già inserite precedentemente non aggiunge informazioni significative, quanto piuttosto aumenta gli elementi di disturbo, riducendo la performance globale del

modello. L'identificazione delle variabili rilevanti per ciascun modello è un compito svolto nella procedura di selezione delle variabili. Nel caso oggetto di studio, è stata effettuata un'analisi preventiva rispetto alla selezione delle le variabili che ha visto il susseguirsi delle seguenti fasi. Si è calcolato su Excel l'accuracy ratio di ogni indicatore rispetto ad un modello ideale, collezionando i risultati in ordine decrescente all'interno di un vettore. Successivamente si è costruita una matrice di correlazione tra i vari indicatori, andando a rilevare in ordine di accuracy quelli più correlati con gli indicatori precedenti. Nella figura successiva si può osservare una piccola parte della matrice di correlazione costruita in Excel.

	EBITDA/Rica	EBIT/ricavi	corrente/n	netto ret	EBITDA/AN	ROA	O-ROI	ROE	ante impagg	Oper/Rmi/costi o	terni/costi oro/	costi materiali/c	cal Agg/ITN
EBITDA/Ri	1												
EBIT/ricav	0,449212	1											
Utile corre	0,850529	0,400981	1										
Risultato r	0,277162	0,759455	0,317754	1									
EBITDA/AI	0,659795	0,370255	0,666428	0,301568	1								
ROA	0,653408	0,394236	0,69185	0,314088	0,953251	1							
O-ROI	0,432307	0,295987	0,488646	0,252299	0,667024	0,691024	1						
ROE	0,510058	0,270635	0,571835	0,247028	0,594282	0,630513	0,464679	1					
ROE ante i	0,518641	0,280495	0,591785	0,254219	0,629124	0,661481	0,51822	0,970318	1				
Val Agg O	0,80908	0,286162	0,711209	0,175672	0,500493	0,482226	0,321643	0,386704	0,402426	1			
Consumi/	0,15089	0,073312	0,226372	0,073295	0,14727	0,156104	0,118412	0,164751	0,16898	0,063886	1		
servizi est	-0,20523	-0,06445	-0,25342	-0,06544	-0,16884	-0,1353	-0,10817	-0,14317	-0,15481	-0,42833	-0,65779	1	
costo lavo	0,069873	0,002745	0,147167	0,014126	0,090348	0,065886	0,068277	0,054346	0,075888	0,506637	-0,15691	-0,50903	1
ammortar	0,142652	0,028754	-0,07511	0,002983	0,035875	-0,06939	-0,09579	-0,01644	-0,04446	0,076007	-0,22649	-0,00115	-0,09034
Val Agg/IT	0,285692	0,154598	0,376098	0,12297	0,32287	0,358411	0,302494	0,233203	0,275689	0,380659	0,116048	-0,21652	0,299479

Figure 3.4, Particolare della matrice di correlazione delle variabili economico-finanziarie

È possibile intuire la complessità del processo sopra descritto, a causa dell'ampiezza della matrice di correlazione e della quantità di dati e altrettanti confronti da effettuare.

Successivamente, si sono selezionati gli indicatori corrispondenti ad un accuracy ratio più alto, escludendo via via quegli indicatori che sono risultati più correlati. Le variabili ottenute sono state inserite, partendo dal modello nullo, nella fase di modellazione all'interno del software statistico, valutando di volta in volta la qualità del risultato ottenuto da quest'ultimo. In questo studio si è scelto di servirsi per del software statistico R, anche per via della sua relativa facilità di approvvigionamento, e ampiezza di tools e pacchetti che mette a disposizione. Nei capitoli successivi verrà trattata nel dettaglio la fase di implementazione dei dati e della costruzione dei modelli in R.

4 R

Lo strumento utilizzato per la modellazione e la stima dei modelli SVM e Logit oggetto di questo studio è stato il programma open source R. R è un linguaggio e un ambiente per il calcolo statistico e la grafica. R può essere considerato come una diversa implementazione di S. Esistono alcune differenze importanti, ma gran parte del codice scritto per S può essere eseguito inalterato in R.

R fornisce un'ampia varietà di tecniche statistiche (modellizzazione lineare e non lineare, test statistici classici, analisi di serie temporali, classificazione, clustering, ...) e tecniche grafiche ed è altamente estensibile. Il linguaggio S è spesso il veicolo di scelta per la ricerca nella metodologia statistica e R fornisce una via Open Source alla partecipazione a tale attività.

Uno dei punti di forza di R è la facilità con cui è possibile produrre trame di qualità di pubblicazione ben progettate, compresi simboli matematici e formule dove necessario. È stata prestata molta attenzione alle impostazioni predefinite per le scelte di progettazione minori nella grafica, ma l'utente mantiene il controllo completo.

4.1 L’ambiente R

R è una suite integrata di funzioni software per la manipolazione dei dati, il calcolo e la visualizzazione grafica. Include

- Un efficace sistema di gestione e stoccaggio dei dati,
- una suite di operatori per calcoli con vettori, e matrici,
- una raccolta ampia, coerente e integrata di strumenti intermedi per l'analisi dei dati,
- strutture grafiche per l'analisi dei dati e visualizzazione su schermo e
- un linguaggio di programmazione ben sviluppato, semplice ed efficace che include condizionali, cicli, funzioni ricorsive definite dall'utente e strutture di input e output.

Il termine "ambiente" intende caratterizzarlo come un sistema completamente pianificato e coerente, piuttosto che un accrescimento incrementale di strumenti molto specifici e inflessibili, come spesso accade con altri software di analisi dei dati.

R, come S, è progettato intorno a un vero linguaggio informatico e consente agli utenti di aggiungere funzionalità aggiuntive definendo nuove funzioni. Gran parte del sistema è essa stessa scritta nel dialetto R di S, che rende facile per gli utenti seguire le scelte algoritmiche fatte. Per le attività ad

alta intensità di calcolo, il codice C, C ++ e Fortran possono essere collegati e richiamati in fase di esecuzione. Gli utenti esperti possono scrivere codice C per manipolare direttamente oggetti R.

R è pensato come un ambiente nel quale vengono implementate tecniche statistiche. R può essere esteso (facilmente) tramite pacchetti. Ci sono circa otto pacchetti forniti con la distribuzione R e molti altri sono disponibili attraverso la famiglia di siti Internet CRAN che copre una vasta gamma di statistiche moderne. Per costruire i modelli Logit e SVM, al centro di questo studio, sono stati utilizzati i due R packages “Caret” e “e1071” che saranno brevemente descritti di seguito.

4.2 Caret Package

Il pacchetto Caret (abbreviazione di _C_lassification _A_nd _RE_gression _T_raining) è un insieme di funzioni che tentano di semplificare il processo per la creazione di modelli predittivi. Questo pacchetto è stato utilizzato nell’ambito della costruzione del modello predittivo Logit. Il pacchetto contiene strumenti per:

- divisione dei dati
- pretrattamento
- selezione delle funzionalità
- modellazione del modello usando il ricampionamento
- stima dell’importanza delle variabili

così come altre funzionalità.

Esistono molte funzioni di modellazione diverse in R. Alcune hanno una sintassi diversa per l'addestramento e / o la previsione del modello. Il pacchetto è stato pensato come un modo per fornire un'interfaccia uniforme alle funzioni stesse, nonché un modo per standardizzare le attività comuni (tale regolazione dei parametri e importanza variabile). Le principali funzioni che sono state utilizzate all’interno di questo pacchetto sono: Train, Predict, e Glm.

La funzione GLM (Generalized Linear Model) è usata per adattarsi a modelli lineari generalizzati, specificati fornendo una descrizione simbolica del predittore lineare e una descrizione della distribuzione dell'errore.

La codifica tipo con gli elementi che compongono questa funzione è:

```
glm(formula, family = gaussian, data, weights, subset,

start = NULL, etastart, mustart, offset,

control = list(...), model = TRUE, method = "glm.fit",

x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)
```

La funzione TRAIN è utilizzata per mettere a punto i modelli selezionando i relativi parametri ottimali. Per ogni modello, viene creata una griglia di parametri (se esiste) e il modello viene addestrato su dati leggermente diversi per ciascuna combinazione di parametri di sintonizzazione candidati. Attraverso ciascun set di dati, viene calcolata la performance dei campioni estratti. Per ogni combinazione viene rappresentata la media e la deviazione standard. La combinazione ottimale viene scelta come modello finale e l'intero training set viene utilizzato per adattarsi al modello finale. I predittori in x possono essere la maggior parte degli oggetti purché la funzione di adattamento del modello possa gestire la classe dell'oggetto. La funzione è stata progettata per funzionare con matrici semplici e dataframe in input. La codifica della funziona "train" si presenta come segue:

```
train(x, y, method = "rf", preProcess = NULL, ...,

 weights = NULL, metric = ifelse(is.factor(y), "Accuracy", "RMSE"),

 maximize = ifelse(metric %in% c("RMSE", "logLoss", "MAE"), FALSE, TRUE),

 trControl = trainControl(), tuneGrid = NULL,

 tuneLength = ifelse(trControl$method == "none", 1, 3))
```

La funzione PREDICT crea un oggetto ricavandolo dalle previsioni di un modello che è stato già soggetto al fitting (ad esempio, con lm, glm). Essa produce i valori previsti, ottenuti valutando la funzione di regressione nel frame newdata (model.frame(object)). Per il modello costruito in questo lavoro si è utilizzata questa funzione per ricavare i valori delle probabilità di default stimate dal Logit. Il metodo di codifica di questa funzione è:

```
predict (object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
```

interval = c ("none", "confidence", "prediction"),

level = 0.95, type = c ("response", "terms"),

terms = NULL, na.action = na.pass,

pred.var = res.var / weights, weights = 1, ...)

4.3 e1071 Package

Le SVMs sono state sviluppate da Cortes & Vapnik (1995) per la classificazione binaria. Il loro approccio prevede di ottenere una separazione di due classi. In particolare, si cerca l'iperpiano di separazione ottimale tra le due classi massimizzando il margine tra i punti delle classi. I punti che si trovano al confine delle due classi, sull'iperpiano di separazione sono chiamati vettori di supporto, e il centro del margine è il nostro iperpiano di separazione ottimale.

L'interfaccia R per la modellizzazione svm si trova nel pacchetto e1071. La funzione SVM (), è stata progettata per essere il più intuitivo possibile. I modelli, e le previsioni sono eseguite come al solito, implementando nel modello il dataframe di input. Come previsto per le altre funzioni statistiche di R, anche in questo caso il motore cerca di essere intelligente sul modo di scegliere la tipologia del modello. In particolare, se la variabile dipendente Y è un fattore, il motore passa alla modalità di classificazione, altrimenti si comporta come una macchina di regressione.
Il metodo di codifica della funzione SVM è:

svm(x, y = NULL, scale = TRUE, type = NULL, kernel = "radial",

degree = 3, gamma = if (is.vector(x)) 1 else 1 / ncol(x),

coef0 = 0, cost = 1, nu = 0.5, class.weights = NULL, cachesize = 40,

tolerance = 0.001, epsilon = 0.1, shrinking = TRUE, cross = 0,

probability = FALSE, fitted = TRUE, ..., subset, na.action = na.omit)

Per ricercare I parametri di modellizzazione ottimi, si può usare all'interno di questo pacchetto la funzione TUNE. Per il modello SVM sviluppato in questo studio questa funzione di ricerca ottima è stata utilizzata per identificare i parametri ottimi gamma e C. Il metodo di codifica di tale funzione è:

Tune (method, train.x, train.y = NULL, data = list(),

validation.x = NULL, validation.y = NULL,

ranges = NULL, predict.func = predict,

tunecontrol = tune.control(), ...) best.tune(...)

5 Implementazione del modello Logit in R

Dopo aver descritto nel capitolo 2 i componenti fondamentali di un modello Logit, ci concentriamo adesso sugli strumenti utili per la sua implementazione all'interno del software R. In particolare, la stima dei parametri attraverso l'utilizzo di diverse tecniche di scelta per l'ottenimento del modello ottimo, la descrizione della "confusion matrix", matrice costruita in R utile a ricavare l'accuracy di un modello, e infine l'implementazione vera e propria del modello.

5.1 La stima dei parametri del modello in R

Nei modelli di regressione logistica, la valutazione dei parametri del modello permette l'interpretazione della relazione che lega le variabili indipendenti con quella dipendente. Non è un'operazione semplice, poiché prevede che all'interno del modello vengano considerate solamente le variabili cosiddette esplicative, la cui variazione sia effettivamente significativa per la variazione della variabile risposta. Normalmente, tanto più è maggiore il numero dei regressori all'interno del modello, più diminuisce la devianza dei residui. Tuttavia, potrebbe accadere per pura casualità che determinate variabili esplicative vengano comprese nel modello solo perché risultanti statisticamente significative. Al contrario, variabili esplicative che sarebbero invece logicamente fondamentali potrebbero non comparire nel modello, perché considerate invece statisticamente non significative. Ne deriva che ottenere un modello che possa definirsi ottimo non è semplice; pertanto una soluzione sarebbe quella di prendere in considerazione più modelli che sono risultati in egual maniera statisticamente significativi e scegliere tra questi il più idoneo. La scelta dipende anche da aspetti relativi al particolare fenomeno da analizzare ed alla sua interpretazione.

È possibile esplicare la modalità di scelta delle variabili facendo riferimento ad alcune fasi fondamentali. Prima di tutto, è necessario determinare le variabili che fanno parte del più vasto insieme dei k repressori, ed al suo interno individuare uno o più sottoinsiemi di variabili (p) che possano adeguatamente giustificare la variabile di risposta. Dopodiché, si rende utile l'applicazione di una regola d'arresto che permetta di determinare il numero di varianti esplicative da utilizzare.

Gli algoritmi di scelta delle variabili più utilizzati sono diversi. Primo tra questi è la *Backward elimination,* che prevede di considerare inizialmente il modello che include la totalità delle variabili a disposizione. Successivamente, viene fissato un livello di significatività, che permette di eliminare la variabile con il coefficiente di regressione meno significativo basandosi sul test t. Vengono nuovamente calcolate le stime dei coefficienti delle restanti variabili, per poi ripetere le stesse operazioni finchè non appaiono più covariate non significative al livello prefissato.

Un secondo algoritmo di scelta delle variabili è la *Forward selection*, la quale prevede invece di partire solamente con la covariata che ha ottenuto, a seguito del test t, una maggiore correlazione con la variabile risposta. Viene quindi fissato un livello di significatività, che serve per determinare la seconda variabile da inserire. Quest'ultima deve infatti presentare il coefficiente di correlazione parziale più elevato e significativo. Dopodiché si inserisce una successiva variabile dipendente, e si prosegue finché il coefficiente di correlazione parziale dell'ultima variabile inserita non risulta più significativo rispetto al livello prefissato.

Un ulteriore criterio di scelta delle variabili è rappresentato dalla *Stepwise regression,* la quale può essere definita come una combinazione degli altri due criteri. Infatti, le covariate da includere nel modello vengono selezionate esattamente come nella *forward selection*. Successivamente viene integrata una nuova variabile, e così facendo i coefficienti di regressione delle variabili già presenti nel modello potrebbero risultare singolarmente non significativi, poiché fortemente correlati con la nuova variabile. Per questo, emulando il procedimento nella *backward elimination,* dopo l'inserimento di ogni nuova variabile si riconsidera il modello per poter verificare la presenza di eventuali variabili da eliminare.

5.2 Confusion Matrix

Una matrice di confusione è una tecnica per riassumere la performance di un algoritmo di classificazione. Calcolare una matrice di confusione può dare una migliore idea di come il modello sta lavorando e quali tipi di errore sta eventualmente commettendo.

Una matrice di confusione riassume i risultati della previsione effettuata all'interno di un problema di classificazione. Il numero di previsioni corrette, e non, viene suddiviso per ogni classe. Questo non solo dà un'idea degli errori fatti dal classificatore, ma anche del tipo di errore commesso, superando il limite che si avrebbe considerando solo l'accuracy ratio.

Per costruire la matrice ci si serve del training e del testing set. Dopo aver addestrato i dati utilizzando il training set, si esegue il fitting del modello, ottimizzando la scelta delle variabili con appositi algoritmi. Fino a questo punto dunque, si utilizza esclusivamente il training set. Una volta definito il modello, lo si mette alla prova con i dati appartenenti al testing set. Terminata la previsione, i risultati ottenuti possono essere paragonati al risultato effettivamente contenuto nel testing set. Ogni riga della matrice corrisponde ad una classe stimata, il cui valore è il risultato della previsione dal modello, mentre ogni colonna corrisponde ad una classe effettiva. Nella matrice è contenuto dunque il numero delle classificazioni corrette ed errate.

Il numero totale delle previsioni corrette per una determinata classe si colloca nelle righe del valore ottenuto, in corrispondenza della colonna che contiene la stessa classe. Nello stesso modo il numero totale delle previsioni incorrette, si trova nella riga della classe ottenuta, ma in corrispondenza della colonna con il valore della classe opposta (in presenza di due classi).

Con il rapporto tra la somma dei valori presenti nella diagonale principale, che rappresentano il numero di classificazioni corrette, e la somma di tutti i valori presenti nella matrice, che rappresentano la totalità delle classificazioni, si ottiene l'accuracy del modello di classificazione.

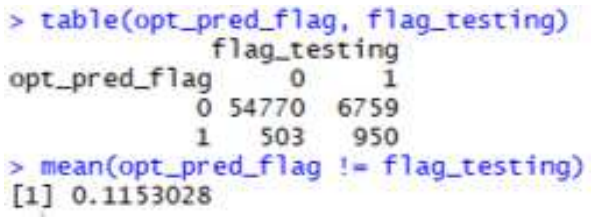


Figure 5.1, Esempio di Confusion Matrix

Questa tecnica è stata utilizzata per il caso oggetto di questo studio. Osservando una matrice costruita come esempio durante uno dei fitting effettuati, si può notare come il rapporto (54770 + 950) / (54770 + 950 + 6759 + 503) fornisca un'accuracy dell'88% circa, che corrisponde al complemento a 1 del misclassification error rate, visualizzato a valle della matrice.

5.3 Il modello Logit per la classificazione in R

Il modello Logit è stato costruito tramite l'utilizzo del programma R. Partendo dai dati in input presenti nel dataset dal nome "Databaseeee1", formato dai valori degli indicatori economici delle società analizzate, è stato scritto un codice per eseguire l'addestramento dei dati caricati, modellizzare il Logit, selezionare i parametri ottimi e successivamente valutare l'accuracy del modello. Viene descritto nella trattazione che segue il percorso step by step che ha condotto dal dataset iniziale al modello finito.

1. Carichiamo il Database e dividiamo il set in training e testing set, che verranno usati rispettivamente per "allenare" i dati, effettuare il fitting del modello e per testarlo successivamente. I due set sono stati individuati in modo da avere in entrambi i set circa la stessa percentuale di default. Nel caso in esame questa è di circa il 12,235.

```
#Import Databaseeee1

attach(Databaseeee1)

Databaseeee1$flag = as.factor(Databaseeee1$flag)

library("caret", lib.loc="~/R/win-library/3.5")

training <- Databaseeee1[1:84529,]
```

testing <- Databaseeee1[84530:157457,]

Abbiamo usato una codifica di tipo letterale per alleggerire la scrittura del codice in R, rinominando dunque le colonne del Database con le lettere corrispondenti a ciascuna variabile.

d1	Riserve+utile/AN	o	costo lavoro/costi operativi
h	ROE	u	AC/PC
i	ROE ante imposte	p	ammortam materiali/costi operativi
a	EBITDA/Ricavi	r	Ricavi/AN
a2	Debiti totali/EBITDA	b1	Cap Circ/AN
e	EBITDA/AN	g	O-ROI
q1	Autof Lordo/AN	d2	Ln(AN)
r1	Auto Lordo-comp straord/AN	t	gg magazz
f	ROA	v1	Deb finanziari (stimati)/VA
c1	Patr netto/AN	u1	PC/ric
c	Utile corrente/ricavi	z	Liq/PC
e1	Patr netto tang/AN	t1	Deb totali/Ric
l1	OFN/EBITDA	v	AC-mag/PC
f1	Patr netto tan/Debiti tot+PN	s	O-Circ Oper (mag+cl-for)/Ric
m1	OFN/EBIT	i1	O-Patr.netto/deb fin-liq+PN
g1	Patr netto tan/Debiti tot-Liq+PN	a1	Liq/AN
c2	Ebitda-servizio debito/AN	s1	Deb totali/VA
b2	Debiti finanziari/EBITDA	z1	Deb finanziari (stimati)/Ric
l	Val Agg Oper/Ricavi	d	Risultato netto rettif/ricavi
o1	OFN/Autof lordo	h1	O-Patr.netto/deb fin+PN
m	Consumi/costi operativi	n1	OFN/AN
q	Val Agg/ITN	b	EBIT/ricavi
n	servizi esterni/costi operativi	p1	O-OF/Deb Fin

Figure 5.2, Trasposizione Letterale delle singole variabili per la codifica in R

2. Utilizziamo il metodo di training per “addestrare” i nostri dati sul modello selezionato

mod_fit <- train(flag ~ d1 + h + i + a + a2 + e + q1 + r1 + f + c1 + c + e1 + l1 + f1 + m1 + g1 + c2 + b2 + l + o1 + m + q + n + o + u + p + r + b1 + g + d2 + t + v1 + u1 + z + t1, data=training, method="glm", family="binomial")

3. Per la scelta ottima delle variabili partiamo dal modello che considera una singola variabile (mod_fit_zero). Scegliamo quella che ha l’acuracy più alta, e ne aggiungiamo una alla volta, togliendo quelle che sono fortemente correlate con le precedenti. Per ogni modello verifichiamo la significatività delle variabili, andando ad osservare il loro p-value.

mod_fit_zero<- glm(flag ~ d1 , data=training, family="binomial")

```
Console ~/
> mod_fit_zero <- glm(flag ~ d1, data=training, family="binomial")
> summary(mod_fit_zero)

Call:
glm(formula = flag ~ d1, family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0430  -0.5500  -0.4637  -0.3072   2.7567

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.7055507  0.0109583 -155.64   <2e-16 ***
d1          -0.0345238  0.0005373  -64.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62743  on 84528  degrees of freedom
Residual deviance: 58199  on 84527  degrees of freedom
AIC: 58203

Number of Fisher Scoring iterations: 5
```

Figure 5.3, Evidenza dei parametri del modello Logit di partenza

4. Al termine di questa procedure, vengono individuate come variabili ottime: d1, h, a, a2, m, n, d2, t. Ora Includiamo queste variabili nel modello ottimo.

mod_fit_opt <- glm(flag ~ d1 + h + a + a2 + m + n + d2 + t, data=training, family="binomial")

5. Visualizziamo il risultato del modello stimato richiamando il comando summary(). Nella schermata si osservano i coefficienti dei regressori, con gli errori standard corrispondenti. Nell'ultima colonna sono rappresentati i p-values corrispondenti ai coefficienti stimati.

summary(mod_fit_opt)

```
> mod_fit_opt<-glm(flag ~ d1 + h + a + a2 + m + n + d2 + t ,data=training, family="binomial")
> summary(mod_fit_opt)

Call:
glm(formula = flag ~ d1 + h + a + a2 + m + n + d2 + t, family = "binomial",
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5410  -0.4701  -0.3889  -0.3205   2.8025

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.869e+00  6.204e-02 -30.125  < 2e-16 ***
d1          -1.563e-02  6.597e-04 -23.688  < 2e-16 ***
h           -3.028e-03  1.879e-04 -16.114  < 2e-16 ***
a           -2.675e-03  7.108e-04  -3.763 0.000168 ***
a2           6.300e-04  4.044e-05  15.579  < 2e-16 ***
m           -2.804e-03  7.541e-04  -3.718 0.000201 ***
n            8.828e-03  5.623e-04  15.699  < 2e-16 ***
d2          -9.240e-02  7.751e-03 -11.921  < 2e-16 ***
t           -2.225e-04  3.391e-05  -6.561 5.35e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62743  on 84528  degrees of freedom
Residual deviance: 55222  on 84520  degrees of freedom
AIC: 55240

Number of Fisher Scoring iterations: 5
```

Figure 5.4, Evidenza dei parametri del modello Logit ottimo

Ai fini dell'interpretazione, si ricorda che l'ipotesi nulla è che le varianze siano uguali fra di loro, e che dunque la variabile indipendente non produca effetti su quella dipendente. La probabilità che sia vera l'ipotesi nulla è indicata dal valore Pr. Nel caso in esempio, le relazioni sono significative, infatti i p-values dei coefficienti sono tutti inferiori al valore critico 0,05, per cui si può rigettare l'ipotesi nulla.

Di seguito I valori dei coefficienti stimati dal modello, codificati per indicatore:

		Coefficienti
	Intercetta	-1,869
d1	Riserve+utile/AN	-0,01563
h	ROE	-0,003028
a	EBITDA/Ricavi	-0,002675
a2	Debiti totali/EBITDA	0,00063
m	Consumi/costi operativi	-0,002804
n	servizi esterni/costi operativi	0,008828
d2	Ln(AN)	-0,0924
t	gg magazz	-0,0002225

Figure 5.5, Variabili ottime e rispettivi coefficienti nel modello Logit ottimo

6. Vogliamo prevedere quale sarà il valore della variabile risposta “flag”, dunque individuo con “flag_testing” solo i valori del testing set, in modo da usarli alla fine per testare quello che il modello ha previsto.

```
flag_testing<-testing$flag
```

7. Usiamo la funzione “predict(..,type=response)” per ottenere la probabilità che la variabile dipendente sia uguale a 1 per ogni osservazione; questo perché la regressione logistica non dà come risultato una classificazione, ma solo delle probabilità.

```
opt_pred_probs <- predict(mod_fit_opt, testing, type="response")
```

8. Applichiamo un metodo per classificare le società utilizzando le probabilità calcolate precedentemente. Creiamo un vettore della stessa lunghezza di quello testing, e ripetiamo “0” per tutti gli elementi.

```
opt_pred_flag <- rep("0", 72928)
```

9. Infine cambiamo I valori da 0 a 1 solo quando la probabilità delle previsioni all’interno del vettore delle probabilità è superiore a 0,12235. Questo valore è stato indivuato dopo aver considerato la distribuzione di frequenza delle probabilità di default predette dal modello.

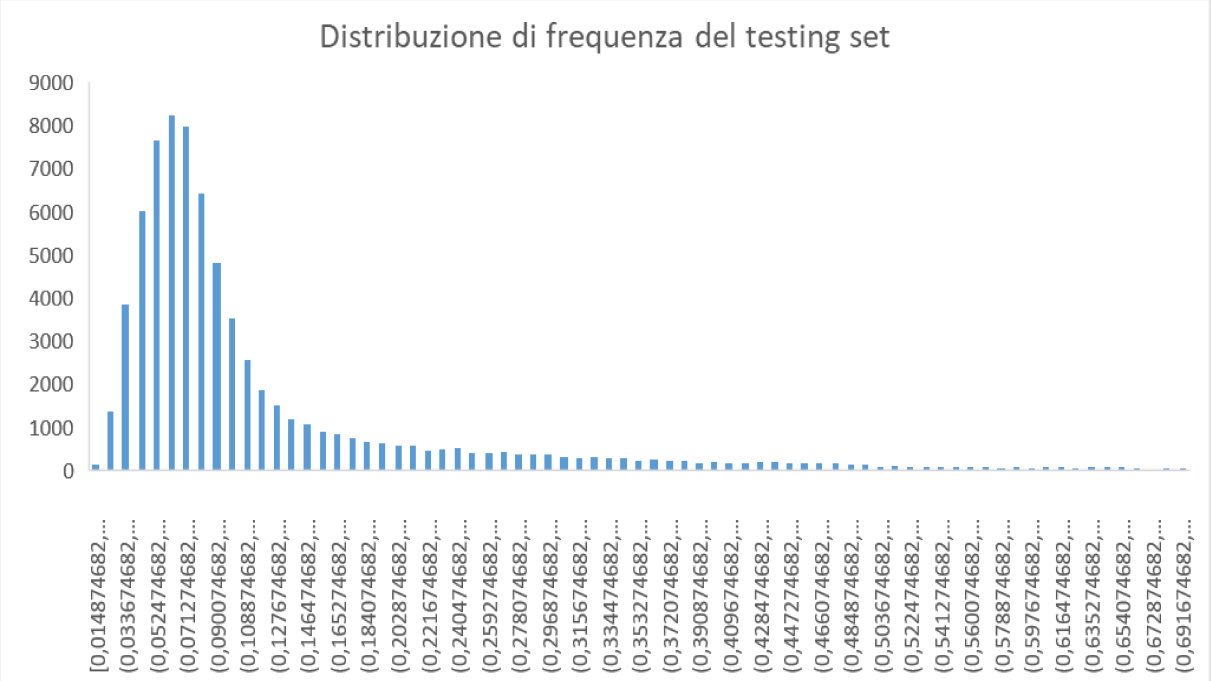


Figure 5.6, Distribuzione di frequenza delle Probabilità di Default all’interno del testing set

Considerando il valore medio della distribuzione, che si attesta intorno al 7,5%, è stato scelto un valore di soglia del 12,24% per identificare le società anomale, effettuando una scelta ottimistica sul numero di osservazioni anomale considerando la distribuzione ottenuta. Si nota inoltre, come il modello abbia ottenuto valori di probabilità generalmente bassi, sottolineando l'indecisione del definire un'osservazione come appartenente ad una classe piuttosto che un'altra. Continuando con l'implementazione del modello, riportiamo la scrittura in codice dell'operazione appena descritta.

```
opt_pred_flag[opt_pred_probs>0.12235] <- "1"
```

10. Creiamo la confusion matrix per calcolare successivamente l'accuracy, e il tasso di errata classificazione o misclassification error rate.

```
table(opt_pred_flag, flag_testing)

mean(opt_pred_flag != flag_testing)
```

```
> flag_testing<-testing$flag
> opt_pred_probs <- predict(mod_fit_opt, testing, type="response")
> opt_pred_flag <- rep("0", 72928)
> opt_pred_flag[opt_pred_probs>0.12235]<- "1"
> table(opt_pred_flag, flag_testing)
             flag_testing
opt_pred_flag     0     1
            0 49792  3664
            1 14185  5287
> mean(opt_pred_flag != flag_testing)
[1] 0.2447482
```

Figure 5.7, Risultato espresso tramite misclassification error

Dalla matrice di confusione, risulta che l'accuracy del modello è 75,53%, calcolato come complemento ad uno del misclassification error rate.

6 Implementazione modello SVM

6.1 Importanza del rischio di classificazione

L'accordo di Basilea (Basilea II) ha messo pressione alle aziende e alle banche da più punti di vista. Le banche devono chiedere un premio di rischio in accordo con le probabilità di default specifiche del debitore. Dunque, più alto è il rischio di default della società, più alto sarà il premio di rischio che la banca richiederà. In secondo luogo, Basilea II chiede alle banche di tenere un equity buffer specifico per cliente. L'entità di questi Buffer è determinata dal peso della funzione di rischio definita dall'accordo di Basilea e considera un coefficiente di solvibilità dell'8%. La funzione mappa le probabilità di default in pesi di rischio. Oltre i determinanti di base del rischio come la probabilità di default (PD), la loss given default (LGD), i pesi del rischio dipendo anche dal tipo del prestito e al fatturato annuo. Basilea II ha un impatto su tutte quelle aziende che hanno bisogno di accedere ad una fonte di finanziamento esterno. Considerato che sia il premio di rischio che i costi di credito sono determinati dal rischio di default, il rating delle aziende ha un profondo impatto economico sulle banche così come sulle aziende stesse come mai prima d'ora. Perciò il metodo di rating è di fondamentale importanza dopo Basilea II. Per evitare frizioni di grande entità, il metodo impiegato deve andare incontro a determinate condizioni. Da un lato, la procedura di rating deve cercare di tenere il misclassification error rate più basso possibile, e inoltre, deve essere più semplice possibile. Le support vector machines hanno il potenziale per soddisfare entrambe queste richieste. La procedura è semplice da implementare, in modo tale che ogni società può generare il suo metodo di rating. Il metodo è adatto per stimare una sola probabilità di default per ogni società. La stima del rating eseguita con le SVM è trasparente e non dipende da giudizi di esperti. Questa proprietà implica obiettività e un alto grado di robustezza contro i cambiamenti dell'utente. Inoltre, un metodo SVM adeguatamente "allenato" permette di individuare l'impatto di tutti i determinanti del rating sulla classificazione. L'utilizzo di un modello SVM può aiutare dunque a rispettare Basilea II in modo più semplice.

6.2 Svm come Tecnica di credit scoring

L'obiettivo di questo studio è analizzare e sviluppare uno strumento per l'analisi del Corporate default basato sulle Support Vector Machines (SVM). Le SVM hanno già dato prova del loro potere predittivo in aree che non sono collegate alla finanza. L'applicazione delle SVM ai problemi economici in generale e alle tematiche finanziarie in particolare è ancora rara. In questo studio si cerca di adattare la tecnica SVM non parametrica alla previsione di Bancarotta e al rating delle società. Per fare ciò utilizziamo gli stessi indici finanziari usati per il modello di regressione logistica, in modo da comparare l'accuratezza della previsione con la tecnica SVM rispetto a quella dell'approccio Logit. Si propone un metodo che colleghi il risultato della classificazione del modello SVM alle classi di rating future delle società considerate. In questa parte dello studio introduciamo un nuovo modo di valutare il merito di credito delle società. La metodologia di rating proposta si basa su un modello machine learning fondato sul metodo di classificazione non lineare SVM, e su una tecnica non parametrica per mappare il punteggio di rating in probabilità di default. Il modello SVM si basa sul principio di una separazione sicura delle società solventi e insolventi in modo tale che la distanza tra le classi sia massimizzata mentre, l'errata classificazione di una società viene penalizzata proporzionalmente alla distanza dalla propria classe. Il metodo permette di usare le tecniche di Kernel e quindi, superfici di separazione non lineari, in contrasto rispetto al modello Logit che usa superfici lineari. Questo risulta evidente da alcuni esempi grafici, come per esempio la figura successiva.

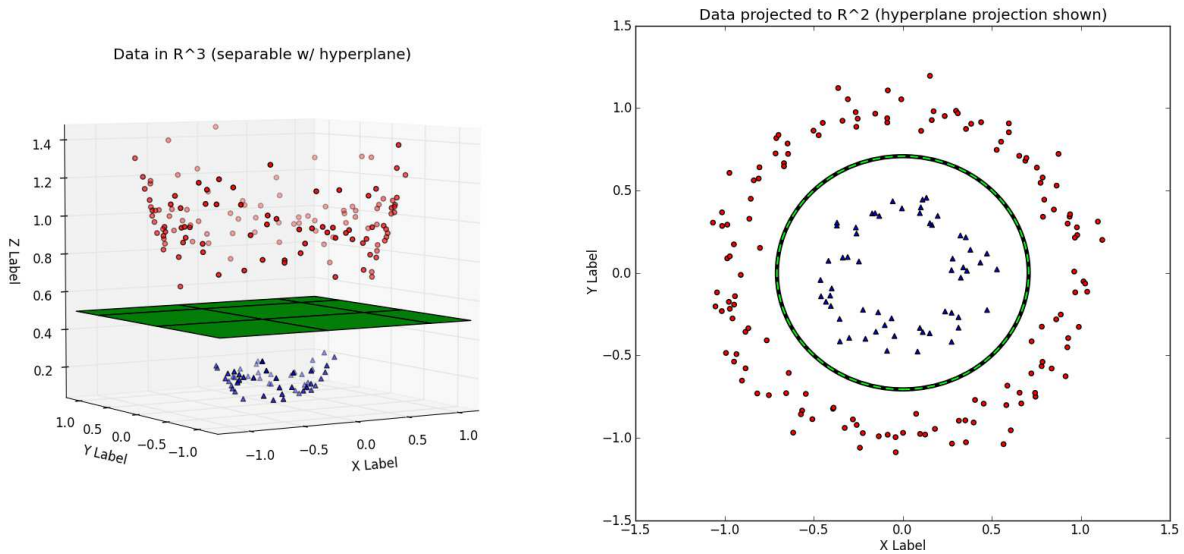


Figure 6.1, Mappatura dei dati nello spazio trasformato (Future Space)

L'SVM può fornire dunque una migliore separazione delle classi, che si traduce in una migliore performance di classificazione. Un'altra importante caratteristica del SVM è che identifica automaticamente la forma delle superfici di separazione.

Esaminiamo empiricamente in questo studio se l'SVM porta ad una più accurata stima degli eventi di default rispetto al modello Logit. Si dimostra che la non – monotonicità e la non - linearità nei dati influenza in maniera significativa l'accuratezza. Le dipendenze non-lineari nei dati, confermate dalla letteratura (Fernandes 2005, Manning 2004) e valutate nei modelli di mercato (Falkestein, Boral, and Carty 2000), sono la ragione per cui vengono considerate tecniche non-lineari come alternative.

L'SVM è una tecnica statistica non lineare che in molte applicazioni, come diagnosi mediche o previsione del carico elettrico, mostra un livello di accuratezza molto alto. Ha come soluzione una funzione di classificazione flessibile che viene controllata regolando solo pochi parametri. La soluzione SVM è stabile, vale a dire cambia lentamente in risposta ad un cambiamento lento dei dati, in quanto il metodo è basato su un problema di ottimizzazione convergente. La sua buona performance e flessibilità, fa dell'SVM un valido modello per le società di rating.

Dal momento che anche un piccolo miglioramento dell'accuracy del modello di credit scoring può tradursi in notevoli risparmi futuri, il problema principale che è stato affrontato nella maggior parte degli studi passati, ha riguardato l'aumento dell'accuracy delle decisioni sul credito. Per le tecniche statistiche convenzionali di classificazione, deve essere assunto il modello di probabilità sottostante il calcolo delle probabilità future, sulla base del quale vengono prese le decisioni di classificazione.

Le tecniche di data mining più recenti come le reti neurali, la programmazione genetica (GP) e le support vector machines (SVM) possono eseguire la classificazione senza avere tale limitazione. Inoltre, questi metodi di intelligenza artificiale hanno anche ottenuto prestazioni migliori rispetto ai metodi statistici tradizionali. Le Support vector machines (SVM) sono state inizialmente proposte da Vapnik (1995) e sono state recentemente utilizzate in una gamma di problemi inclusi il riconoscimento di schemi (Pontil e Verri, 1998), nella bioinformatica e categorizzazione del testo (Joachims, 1998). Huang, Chen, Hsu, Chen e Wu (2004) hanno ottenuto previsioni, con una precisione intorno all'80% sia

con le reti neurali che con metodi SVM, per analisi legato al mercato statunitense. Quando si utilizza l'algoritmo SVM, ci si confronta con due problemi: come scegliere il sottoinsieme ottimo in input per l'SVM e come impostare i migliori parametri del kernel. Questi due problemi sono cruciali in quanto la scelta del sottoinsieme in input influenza la scelta dei parametri del kernel e viceversa (Fro¨ hlich & Chapelle, 2003). La selezione del dataset è un aspetto molto importante nella creazione dei sistemi di classificazione. È vantaggioso limitare il numero di osservazioni in input in un classificatore per avere un modello meno intensivo dal punto di vista computazionale e con una buona base predittiva. Con un piccolo set di osservazioni, la spiegazione della logica di classificazione può essere più facilmente compresa. In aggiunta alla selezione della funzione, la corretta impostazione dei parametri del modello può migliorare sensibilmente l'accuratezza della classificazione SVM. I parametri che dovrebbero essere ottimizzati includono un parametro di penalità C e i parametri della funzione kernel come la gamma (γ) per la funzione radiale Kernel (RBF). Per progettare un SVM, è necessario scegliere una funzione kernel, impostare i parametri del kernel e determinare un margine con costante C. Una valida alternativa per individuare i parametri ottimi C e γ quando si usa la funzione Kernel RBF è l'algoritmo "grid". Prima di affrontare la descrizione del modello proposto, viene fornita una veloce trattazione sui concetti matematici alla base dell'algoritmo che caratterizza le SVMs.

6.3 Il modello SVM per la classificazione in R

Il modello SVM, come il modello Logit, è stato costruito utilizzando i tools del programma R. Il database di partenza, con i dati di input è, come per il modello Logit, "Databaseeee1" contenente i valori degli indicatori economici delle società analizzate. Anche in questo caso è stato sviluppato un codice per eseguire l'addestramento dei dati caricati, modellizzare l'SVM, ed è stata eseguita una funzione per la selezione dei parametri ottimi. Successivamente si è valutata l'accuracy del modello. Di seguito il percorso che ha condotto dal dataset iniziale al modello finito, con l'analisi del codice sviluppato.

1. Carichiamo il Database e dividiamo il set in training e testing set, che verranno usati rispettivamente per "allenare" i dati, effettuare il fitting del modello e per testarlo successivamente. I due set sono stati individuati in modo da avere in entrambi i set la stessa percentuale di default. Nel caso in esame questa è di circa il 12,235. È stato caricato per questo modello il pacchetto "e1071" contenente le funzioni logiche per la modellizzazione svm. La variabile dicotomica "flag" è stata identificata come un fattore in quanto espressione dei soli due valori 0/1. Eseguiamo la funzione attach in modo da fissare l'intero dataset tra gli oggetti di ricerca, per poter richiamare gli oggetti contenuti in esso senza fare riferimento in ogni occasione al dataset stesso. La funzione view(data) è stata richiamata per visualizzare il dataset. La prima parte del codice è stata così espressa:

```
library(readxl)
Databaseeee1 <- read_excel("C:/Users/Desktop/Databaseeee1.xlsx")
View(Databaseeee1)
attach(Databaseeee1)
library(e1071)
Databaseeee1$flag<-as.factor(Databaseeee1$flag)
training <- Databaseeee1[1:84529,2:48]
testing <- Databaseeee1[84530:157457,2:48]
```

Il programma presenta a questo punto una visualizzazione del Dataset e nel “workplace”, che definisce gli oggetti costituiti durante l’esecuzione del codice, I tre set che si sono definiti.

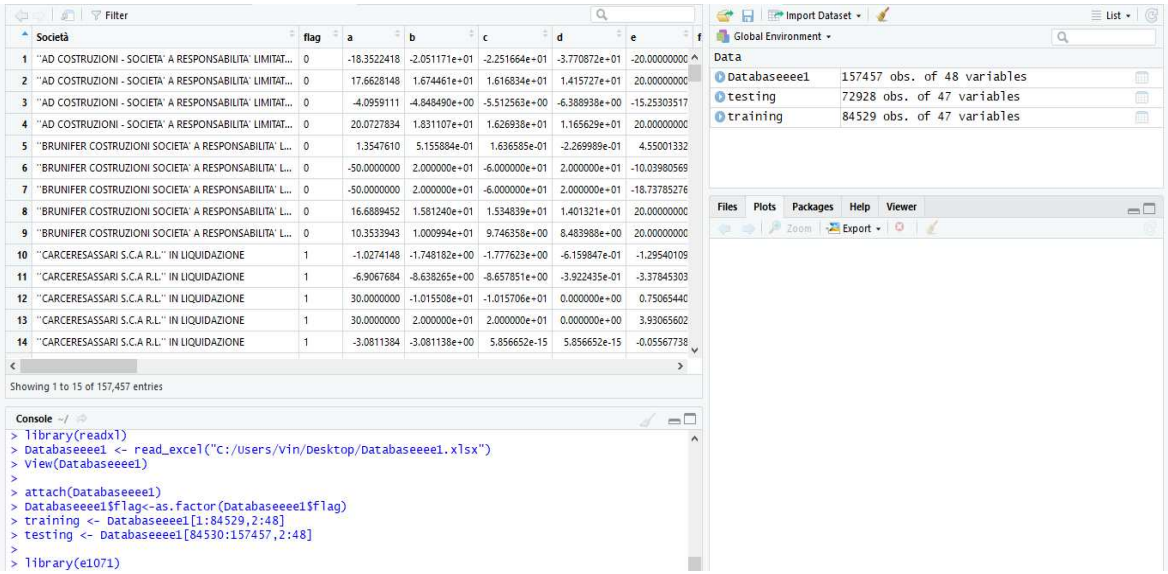


Figure 6.2, Panoramica Dataset e workplace environment

2. Eseguiamo ora la funzione “svm” per il nostro modello, andando a riempire i campi richiesti. In particolare la variabile dipendente, che nel nostro caso è identificata da “flag”, il dataset con i dati in input, e indichiamo la tipologia del modello che è nel nostro caso di classificazione. Per ottenere un confronto più accurato con il modello Logit, utilizziamo le stesse variabili presenti nel modello Logit ottimo. Il modello è stato chiamato inizialmente “svm_linear” in quanto ipotizziamo in partenza che il modello sia lineare. Sarà poi il risultato della simulazione a definire se il Kernel sarà lineare o radiale. Eseguiamo quindi la funzione print (), con la quale ci viene mostrato il risultato ottenuto con la simulazione del modello SVM. Continuiamo la compilazione del codice come segue:

svm_linear<-svm(flag ~ d1+h+a+a2+m+n+d2+t, data=training, type="C-classification", Kernel=”linear”, scale=FALSE)

print(svm_linear)

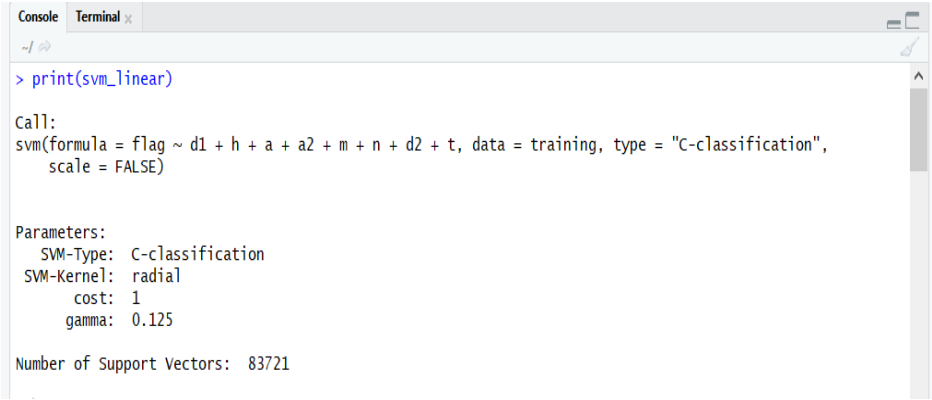


Figure 6.3, Parametri del Modello Svm

Il risultato mostrato evidenzia la tipologia della funzione Kernel "radial" che meglio descrive il modello, e i parametri di modellizzazione individuati, che sono C=1 e gamma=0,0125. Inoltre, viene riportato il numero di support vector individuati nella simulazione: 83721. Questi support vectors rappresentano le osservazioni che giacciono sull'iperpiano di separazione ottimo. L'elevato numero evidenzia una separazione non ben definita delle classi di appartenenza delle osservazioni.

Per questioni legate all'ampiezza dei dati analizzati non è stato possibile attuare una visualizzazione grafica dei risultati della simulazione.

3. Per la selezione dei parametri ottimali, C e gamma, si effettua ricerca ottima utilizzando la funzione "tune.svm". Ancora una volta andiamo a riempire i campi necessari, indicando l'algoritmo svm, la variabile dipendente, il dataset, la funzione kernel, e in questo caso anche il range da considerare per la ricerca dei parametri ottimi. In questo caso si è usato 10^(-3:1) sia per gamma che per C. Eseguendo la funzione summary() con oggetto la funzione "tune.svm" possiamo visualizzare il risultato della funzione eseguita precedentemente. In questa schermata vengono

presentati i valori di C e gamma, l'errore e la dispersione. I parametri ottimi corrispondono a quelli con il valore dell'errore più piccolo.

```
> tuned=tune.svm(flag~., data=training, gamma=10^(-3:1), cost = 10^(-3:1), kernel="radial")
> summary(tuned)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
  0.01   10

- best performance: 0.1105

- Detailed performance results:
   gamma  cost  error  dispersion
1  1e-03 1e-03 0.1215 0.007763876
2  1e-02 1e-03 0.1215 0.007763876
3  1e-01 1e-03 0.1215 0.007763876
4  1e+00 1e-03 0.1215 0.007763876
5  1e+01 1e-03 0.1215 0.007763876
6  1e-03 1e-02 0.1215 0.007763876
7  1e-02 1e-02 0.1215 0.007763876
8  1e-01 1e-02 0.1215 0.007763876
9  1e+00 1e-02 0.1215 0.007763876
10 1e+01 1e-02 0.1215 0.007763876
11 1e-03 1e-01 0.1215 0.007763876
12 1e-02 1e-01 0.1215 0.007763876
13 1e-01 1e-01 0.1215 0.007763876
14 1e+00 1e-01 0.1215 0.007763876
15 1e+01 1e-01 0.1215 0.007763876
16 1e-03 1e+00 0.1215 0.007763876
17 1e-02 1e+00 0.1139 0.008372574
18 1e-01 1e+00 0.1145 0.007276293
19 1e+00 1e+00 0.1196 0.006866990
20 1e+01 1e+00 0.1213 0.007660142
21 1e-03 1e+01 0.1185 0.007735201
22 1e-02 1e+01 0.1105 0.009033887
23 1e-01 1e+01 0.1231 0.005724218
24 1e+00 1e+01 0.1190 0.006666667
25 1e+01 1e+01 0.1211 0.007680422
```

Figure 6.4, Ricerca ottima, tune.svm function

I valori corrispondenti ai "best parameters" per gamma e C individuati dalla simulazione sono rispettivamente 0,01 e 10, corrispondenti ad un valore di errore di 0,1105 e una dispersione dello 0,009.

4. Adesso andiamo a inserire i valori dei parametri ottimi all'interno del modello, e andiamo a predire il valore della variabile utilizzando il testing dataset e eseguendo la funzione predict(), nella quale andiamo a specificare il type"class" per indicare che si tratta di un problema di classificazione. Con la funzione plot andiamo a visualizzare, con l'ausilio di un grafico a barre, le previsioni ottenute dal modello.

```
svm_linear<-svm(flag ~ d1+h+a+a2+m+n+d2+t, data=training, type="C-classification",
Kernel="linear", scale=FALSE, cost=10, gamma=0.01)

predictions=predict(svm_linear, testing, type="class")

plot(predictions)
```

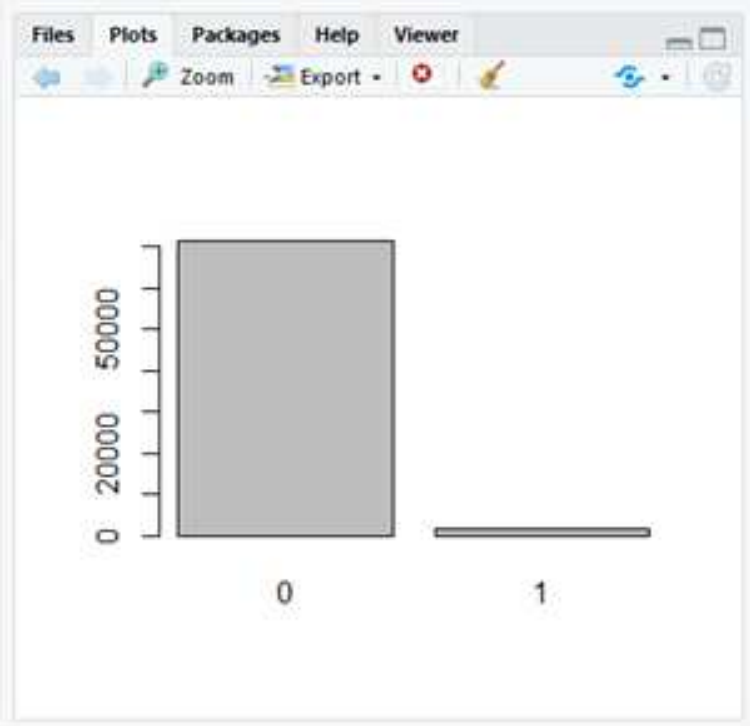


Figure 6.5, Predictions plot in R

Come si può verificare, vi è una netta prevalenza di una classe rispetto all'altra. Tuttavia, questo è un risultato atteso in quanto all'interno del campione la percentuale di società sane è prevalente. Bisogna verificare adesso, in che misura le previsioni della simulazione siano coerenti con le osservazioni reali.

5. A questo punto si può visualizzare il risultato riportato nella "confusion matrix" che mostra le previsioni eseguite durante la simulazione, incrociate con le osservazioni reali del testing set. Partendo da questa matrice possiamo calcolare l'accuracy della classificatione calcolando il valore percentuale delle classificazioni corrette eseguite.

table(predictions,testing$flag)

mean(predictions==testing$flag)

```
> table(predictions,testing$flag)

predictions     0     1
          0 63294  7980
          1   683   971
> mean(predictions==testing$flag)
[1] 0.8812116
```

Figure 6.6, Risultato Simulazione tramite livello di accuracy

Il modello di classificazione svm presenta un livello di accuracy dell'88,12%.

7 Confronto Logit/SVM

Dopo l'implementazione dei modelli Logit e SVM all'interno del software R e dopo aver approfondito i principi teorici di questi due modelli, andiamo ad effettuare un confronto più puntuale sui due metodi.

Il Support Vector Machine (SVM) è un algoritmo utilizzato per problemi di classificazione simili alla Regressione logistica (LR). Il Logit e il modello SVM con Kernel lineare infatti generalmente funzionano in modo comparabile nella pratica.

Dato un problema di classificazione binaria, l'obiettivo è trovare la linea di separazione "migliore" che abbia la massima probabilità di classificare correttamente i punti invisibili. In base a come si definisce questa nozione di "migliore", si possono avere diversi modelli come l'SVM e Logit.

Per ottenere ciò, l'SVM tenta di massimizzare il margine tra i support vectors più vicini mentre la Regressione Logistica ricerca la probabilità di appartenenza ad una classe. Pertanto, l'SVM trova una soluzione che sia il più adatta possibile per le due categorie, mentre il Logit non ha questa proprietà, come si evince dall'immagine sottostante.

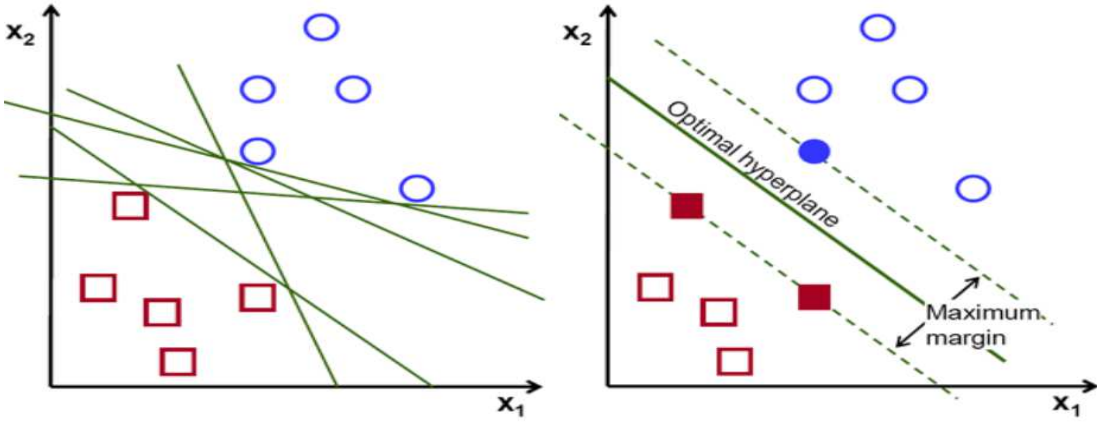


Figure 7.1, Confronto SVM LR, metodo di separazione delle classi

Il Logit è più sensibile ai valori anomali rispetto a SVM perché la funzione di costo del logit diverge più rapidamente di quella di SVM. Quindi con l'inserimento di un outlier, vale a dire un'osservazione molto lontana dalla maggior parte dei dati, i grafici Logit e Svm risponderebbero in maniera molto differente, come appare nell'immagine qui sotto:

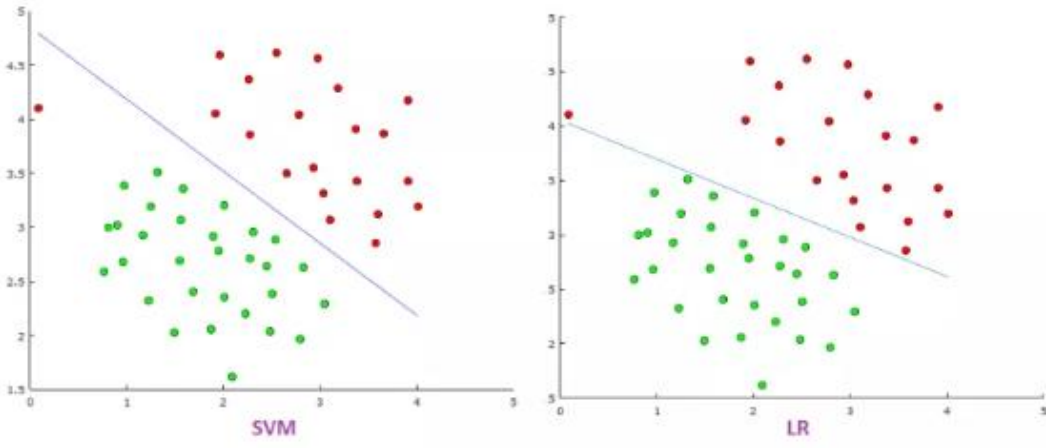


Figure 7.2, Confronto SVM LR, sensibilità agli outliers

La regressione logistica produce valori probabilistici mentre l'SVM produce 1 o 0. Quindi in poche parole il Logit non fa previsioni assolute e non presuppone che i dati siano sufficienti per dare una decisione finale. Questa potrebbe essere una buona proprietà quando ciò che vogliamo è una stima o non abbiamo una grande fiducia nei dati. Di solito, per ottenere valori discreti 1 o 0 per il Logit, possiamo dire che quando il valore probabilistico è maggiore di una certa soglia, classifichiamo come 1 e quando il valore è inferiore alla soglia, classifichiamo come 0, che è quello che è stato fatto in questo lavoro.

Il modello SVM tenta di trovare il margine di separazione più ampio possibile, mentre la regressione logistica ottimizza la funzione di probabilità di log, con probabilità modellate dalla funzione sigmoide.

Il modello SVM amplia il proprio campo di applicazione utilizzando i trucchi del kernel, trasformando i dataset in un nuovo spazio, in modo che i problemi complessi possano essere affrontati nella stessa modalità "lineare" nell'iper-spazio trasformato.

Il lavoro applicativo svolto ha mostrato che in generale, l'accuracy della classificazione dell'SVM con Kernel RBF è nettamente migliore rispetto alla classificazione ottenuta con il modello Logit. Mentre l'SVM ha dato un valore vicino all'88%, il Logit si è fermato al 75%. Si ricorda, come spiegato precedentemente, che anche per questo Logit si è usato un criterio di classificazione affiancato alla stima delle probabilità ottenute dal modello.

Inoltre, una differenza notevole tra i due modelli è dovuta all'onere computazionale richiesto dal modello SVM rispetto al Logit, soprattutto per quanto riguarda la funzione di ricerca ottima dei parametri Kernel, che può richiedere, come nel caso in questione, diversi giorni di calcolo.

Conclusioni

All'interno del presente lavoro sono stati introdotti gli elementi che permettono di definire due metodi di credit scoring differenti, il Logit e l'SVM. Attraverso una prima trattazione teorica, sono state descritte le strutture e i fondamenti dei suddetti modelli. Le simulazioni al software hanno invece fornito un esempio pratico con conseguenti risultati da valutare. L'obiettivo del lavoro è stato quello di descrivere e confrontare i metodi Logit e SVM nell'ambito di un problema di classificazione binaria.

I primi due capitoli sono stati dedicati agli studi sull'Analisi del Rischio di Credito e ad una panoramica generale sui principali metodi e modelli di credit scoring, dopodiché è stato introdotto il caso del settore delle costruzioni, soffermandosi sul settore dell'edilizia in Italia nel 2012 e sulla sua evoluzione, per poi illustrare il campione selezionato e oggetto d'analisi. Infatti, dopo aver selezionato un campione di società italiane appartenenti al settore delle Costruzioni, si sono utilizzati i dati dei bilanci per ottenere le variabili economico finanziarie da sottoporre come oggetto delle simulazioni, avvalendosi prima del modello Logit e poi del modelo SVM. I due modelli hanno portato a risultati differenti. Il modello Logit ha prodotto le probabilità di default utilizzate alla base del criterio di classificazione adottato. L'evidenza statistica ha mostrato come le probabilità ottenute siano risultate abbastanza inefficaci per la determinazione della classe di appartenenza, sottolineando la difficoltà legata all'individuazione di una relazione precisa tra le variabili indipendenti e quella dipendente al centro della regressione logistica. Nonostante queste difficoltà, ricorrendo all'estrapolazione di un criterio di classificazione si è ottenuto un valore di accuracy del modello del 75%. A differenza del Logit, il modello SVM non richiede la conoscenza della relazione che intercorre tra le variabili considerate. Inoltre, l'obiettivo delle Support Vector Machines è quello di classificare le osservazioni, senza passare per la determinazione delle probabilità di default. Attraverso l'utilizzo di funzioni di modellizzazione e ricerca ottima più complesse e onerose dal punto di vista computazionale, il modello SVM ha individuato l'intercorrere di una relazione di separazione di tipo non lineare tra le variabili. Per fare ciò si è servito delle funzioni Kernel, alla base della generalizzazione del modello lineare. Il modello in questione ha avuto come risultato un incremento del livello di accuracy significativo, di circa il 13%, rispetto al modello Logit. Entrambi questi risultati non hanno potuto prescindere dalla qualità dei dati

in possesso, che hanno costituito la base di questo studio. Nonostante una gran parte del lavoro effettuato sia stata dedicata alla fase di pulizia e controllo dei dati, è evidente la difficoltà che sussiste nell'eliminare del tutto gli errori da cui sono affetti. Questo è il motivo per il quale, considerando inoltre l'utilizzo di un software open source con un range di tools ampio ma non specializzato, e la qualità delle informazioni in input, il risultato dell'accuracy del modello SVM dell'88% deve considerarsi soddisfacente. Al contrario, il Logit ha mostrato tutti i suoi limiti nell'interpretare le relazioni esistenti in dataset di grandi dimensioni e con schemi non ben definiti.

Al termine di questo studio risulta chiaro come nel processo di analisi del rischio di credito, i modelli di credit scoring siano soggetti a continui studi ed evoluzioni. Individuare un criterio univoco di classificazione dei soggetti che richiedono l'accesso ad un canale di credito è un compito estremamente complesso. La raccolta delle serie storiche, l'individuazione degli schemi logici e l'interpretazione dei risultati ottenuti sono tutti aspetti indispensabili per la corretta costruzione di un modello efficiente di scoring. Per l'elaborazione della moltitudine di dati numerici e non, che intervengono in questo processo, è ragionevole pensare come l'evoluzione delle tecnologie e potenzialità di calcolo possano influire in modo determinante in futuro nell'applicazione ed efficacia di questi metodi, rendendo sempre più il credit scoring una scienza esatta e oggettiva.

Sitografia e Bibliografia

https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#What%20is%20a%20Neural%20Network

https://books.google.it/books?hl=it&lr=&id=HUnqnrpYt4IC&oi=fnd&pg=PP7&dq=support+vector+machine&ots=g9gEAu0tTe&sig=_lBvYQYp0XzltHiZI2lLZ3_D-00#v=onepage&q=support%20vector%20machine&f=false

http://nlg.csie.ntu.edu.tw/~cjwang/paper/Credit%20Card%20Scoring%20with%20a%20Data%20Mining%20Approach%20Based%20on%20Support%20Vector%20Machine.pdf

https://machinelearningmastery.com/confusion-matrix-machine-learning/

https://cran.r-project.org/web/packages/e1071/e1071.pdf

https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf

https://cran.r-project.org/web/packages/caret/caret.pdf

http://www.rassegna.it/articoli/costruzioni-un-difficile-percorso-di-uscita-dalla-crisi

https://www.panorama.it/economia/edilizia-ecco-perche-il-settore-non-riparte/

https://cran.r-project.org/doc/contrib/Ricci-regression-it.pdf

http://psiclab.altervista.org/MetTecPsicClinica2015/Materiali/Dispensa_Regressione_Lineare_e_Logistica_2014.pdf

https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f

Predicting Bankruptcy with Support Vector Machines; Wolfgang Härdle, Rouslan A. Moro, Dorothea Schäfer

Capital Requirements, Business Loans, and Business Cycles: An Empirical Analysis of the Standardized Approach in the New Basel Capital Accord_Seth B. Carpenter, William Whitesell, and Egon Zakraj_ek Board of Governors of the Federal Reserve System* November 13, 2001

Cyclical Implications of the Basel-II Capital Standards_Anil K Kashyap Graduate School of Business, University of Chicago Jeremy C. Stein Department of Economics, Harvard University November 2003

From Basel I To Basel Ii: An Analysis Of The Three Pillars_Abel Elizalde CEMFI Working Paper No. 0704, June 2007