POLITECNICO DI TORINO

Collegio di Ingegneria Gestionale

Corso di Laurea Magistrale
in Ingegneria Gestionale

Tesi di Laurea Magistrale

Interpretabilità dei risultati ottenuti dall'applicazione di algoritmi di Machine Learning ai Sistemi di Rating Interni: analisi del modulo cash-flow del modello SME Retail di Intesa Sanpaolo

Relatore

prof. Franco Varetto

Candidato

Angelo Tripodi

Anno Accademico 2018/2019

Sommario

Indice delle Figure

Indice delle Tabelle

Indice degli Output

Abstract

Il presente lavoro di tesi si pone l'obiettivo di valutare l'interpretabilità dei risultati ottenuti con gli algoritmi di machine learning applicati nella stima delle probabilità di default.

La suddetta probabilità rappresenta il rischio che una controparte, nei confronti della quale esiste un'esposizione, vada in default nell'arco di un determinato orizzonte temporale.

Per calcolare la probabilità di default, è possibile utilizzare metodi statistici, quali ad esempio la regressione logistica, oppure modelli di machine learning.

Tali algoritmi processano i dati in input e ottengono un output sulla base delle features (variabili) fornite: maggiore è il numero di features utilizzato, maggiore è l'informazione contenuta nel modello e quindi le sue performance. All'aumentare del numero di features gli algoritmi si adattano meglio ai dati riuscendo a cogliere l'esistenza di relazioni, anche non lineari, tra le diverse caratteristiche presenti. Sulla base di queste relazioni è generato un output che in questo caso rappresenta la probabilità di default associata a ciascuna controparte.

Data l'elevata numerosità delle caratteristiche e la complessità di calcolo sviluppata, è difficile predire le scelte prese dal modello durante il processo decisionale, pertanto gli algoritmi di machine learning sono intesi come delle black-box.

Negli ultimi anni si sono sviluppati modelli di interpretabilità degli output del machine learning, con l'obiettivo di rendere più comprensibile l'evoluzione del modello e la sua capacità predittiva. Con il termine interpretabilità si intende l'abilità di prevederne il risultato: maggiore è l'interpretabilità di un modello di machine learning, più facile è comprendere per quale ragione sono state prese determinate decisioni.

Il tema dell'interpretabilità è ancora più importante se si applicano queste tecniche al rischio di credito e ai rating interni, in quanto essi devono essere "plausible and intuitive", sia per motivi regolamentari, sia perché il rating è un elemento fondamentale in fase di concessione e il gestore che lo utilizza deve comprenderne le logiche sottostanti.

I modelli di interpretabilità si suddividono in modelli di interpretabilità globale e locale. I primi hanno l'obiettivo di individuare la relazione esistente tra le variabili e l'output del

modello; i secondi individuano le variabili chiave nella determinazione degli output a livello di singola istanza, quindi a livello di singola controparte.

Le principali tecniche di interpretabilità, ad oggi esistenti, sono quattro:
- Local Interpretable Model Explanation (LIME) con l'obiettivo di spiegare la singola previsione, rappresentando le variabili chiave che interagiscono nel modello;
- Partial Dependence Plot (PDP) che rappresenta la relazione media esistente tra una variabile e l'output del modello;
- Individual Conditional Expectation (ICE-box) che coglie la relazione esistente tra una variabile e l'output del modello, facendo variare per ogni istanza del database i valori che la suddetta variabile può assumere rappresentando l'impatto che essa ha con l'output predetto;
- SHAP, che basandosi sulla teoria del valore di Shapley, la quale trae origine dalla teoria dei giochi, associa ad ogni feature (giocatore) un peso o contributo marginale (payoff) nella definizione delle previsioni (gioco cooperativo).

Definito il contesto in cui si opera, nel presente lavoro di tesi è stato analizzato un caso reale di applicazione del machine learning: il modulo cash flow del nuovo modello di rating SME Retail di Intesa Sanpaolo. Esso in stima ha adottato due algoritmi distinti di machine learning, Random Forest e XGBoost. È stato scelto come modello ufficiale quello con potere discriminante migliore, e sulla base dei risultati sono stati applicati i modelli di interpretabilità per valutare meglio la capacità predittiva del modello e la causa delle scelte da esso effettuate.

1 Rischio di credito e contesto normativo

1.1 Rischio di credito

Il rischio di credito rappresenta una voce considerevole dei rischi finanziari a cui un'organizzazione va incontro svolgendo la propria attività di business. Tra i principali rischi si hanno: il rischio di tasso di interesse, il rischio di mercato, il rischio di liquidità, il rischio operativo e il rischio di credito.

Quest'ultimo rappresenta la possibilità che una variazione inattesa del merito di credito di una controparte, nei confronti della quale esiste un'esposizione, generi una corrispondente variazione inattesa del valore di mercato del credito.

Si possono avere due diversi scenari: default (o insolvenza della controparte) oppure deterioramento della qualità del credito, e in questo caso si parla di rischio di migrazione della controparte. In caso di default si ha il rischio di perdita conseguente all'insolvenza, mente in caso di migrazione si ha la perdita conseguente alla diminuzione della qualità del credito.

Una diminuzione della qualità del credito corrisponde ad un aumento della probabilità di insolvenza della controparte, con un conseguente aumento dello spread che serve a remunerare il premio per il rischio di credito.

La variazione del merito di credito inattesa determina un aumento di capitale che la banca deve detenere, mentre la perdita attesa è incorporata al momento della concessione, o del rinnovo, del credito ed inclusa nel pricing del rischio.

Una variazione inattesa implica che le aspettative formulate al momento della concessione del credito si rivelano errate (anche solo in parte), per errori di valutazione o/e per l'insorgere di nuovi eventi economici che influiscono sulla situazione della controparte, o per la volatilità insita nell'operazione

Nella formulazione del rischio di credito si possono riscontrare diverse fonti o cause del rischio, tra le principali si hanno:
- rischio di default: rischio che la controparte sia insolvente;

- rischio di recupero: incertezza relativa all'ammontare che verrà effettivamente recuperato dal creditore al termine delle procedure di contenzioso nei confronti dei debitori insolventi;
- rischio di esposizione: effettivo ammontare del prestito al momento dell'insolvenza;
- rischio di spread: a parità di merito creditizio, cambia il premio per il rischio;
- rischio di concentrazione: possibile correlazione tra una pluralità di soggetti, pertanto il cambiamento del merito creditizio di uno provoca cambiamenti anche agli altri. (1)

1.1.1 Expected Loss e Unexpected Loss

Il rischio di credito, per la complessa natura, è caratterizzato da due parametri: la perdita attesa e la perdita inattesa.
La perdita attesa o Expected Loss (EL) è, a sua volta, calcolata come il prodotto di tre componenti:

$$\boldsymbol{EL = PD \times LGD \times AE} \qquad (1)$$

- PD = Probabilità di Default; indice della misurazione del merito creditizio. Con il termine Default si indica lo stato della controparte nel momento in cui viene meno la capacità di onorare i propri impegni finanziari. È convenzione usare un orizzonte temporale pari ad un anno, per la stima di questo parametro.
  Per la definizione del valore della PD si usano diversi tipi di dati: dati finanziari, quali i bilanci aziendali, integrati spesso con dati andamentali, quali movimentazioni, saldi e indici di rotazione; dati qualitativi, quali il numero di dipendenti, informazioni sui soci etc.; dati anagrafici che costituiscono il supporto informativo più importante nel caso di clienti privati; dati provenienti dalla Centrale dei Rischi in cui sono presenti le posizioni in sofferenza segnalate dalle banche del sistema creditizio.
  Con una buona disponibilità di dati, è possibile definire un sistema di scoring, assegnando ad ogni controparte uno score, o valutazione, per indicare il suo merito creditizio nei confronti della banca.

- LGD = Loss Given Default; rappresenta la percentuale di perdita in caso di insolvenza, che tuttavia non è nota al momento dell'erogazione del credito. È calcolata come segue: $LGD = 1 - RR$;

RR = recovery rate o tasso di recupero; esso indica la percentuale di recupero in caso di default della controparte.

Il Recovery Rate è influenzato da diversi fattori: caratteristiche del finanziamento, caratteristiche della controparte finanziata, caratteristiche dell'ente affidante e fattori esterni.

Per la stima del Recovery Rate, si possono avere due approcci:

- uso dei dati interni dell'istituto bancario, per tipo di esposizione e forme tecniche (mutui, finanziamenti) e categorie di controparti (imprese corporate, PMI, privati). In tal modo si stimano dei tassi medi di recupero;
- uso dei dati di mercato riguardanti i prezzi delle obbligazioni di imprese andate in insolvenza: il prezzo successivo all'insolvenza riflette le aspettative del mercato sull'entità del recupero e sui tempi del recupero per i creditori. Metodo applicabile solo per imprese con debiti quotati sul mercato.

Il calcolo del Recovery Rate (RR) è il seguente:

$$RR = \frac{\sum_{t=1}^{n} \frac{ER_t - AC_t}{(1+i)^t}}{EAD} \qquad (2)$$

- ER = Expected Recovery: importo recuperato nel periodo t;
- AC = costi amministrativi sostenuti nel periodo t;
- i = tasso di attualizzazione che può basarsi su:
  - Tasso interesse di trasferimento di fondi (costo marginale del "funding" della banca);
  - Tasso contrattuale del finanziamento andato in default (tale tasso può non riflettere il rischio del prestito post-default);
  - Tasso congruo per il rischio, tenuto conto dei rischi su ER;
  - Tasso risk-free: nell'ipotesi in cui si assuma che il rischio (sistematico) e il premio per l'avversione ad esso siano già inclusi in altri parametri (downturn LGD, asset correlations implicite nella regulation…).
- EAD = esposizione al momento del default;
- n = periodo di tempo stimato per realizzare il recupero.

- AE = Adjusted Exposure; indica l'ammontare che può essere perso in caso di default. In alcuni casi è di facile determinazione, ad esempio può essere un mutuo: il cliente, in questo caso, non ha alcuna discrezionalità circa il finanziamento e il piano di rimborso. In altri casi, invece, vi è una componente aleatoria, come ad esempio un fido di c/c: la banca mette a disposizione del cliente una certa quantità di fondi, ma è il cliente che decide quando e quanto utilizzarne. Le imprese in situazioni di difficoltà tendono a utilizzare al massimo i fidi ottenuti e spesso vanno in "sconfinamento": rischio di esposizione (opzione implicita a favore dell'impresa). Il calcolo per il calcolo dell'Adjusted Exposure, può essere riassunto in:

$$AE = DP + UP * UGD \qquad (3)$$

- DP = drawn portion: quota utilizzata;
- UP = undrawn portion: quota non utilizzata;
- UGD = Usage Given Default: percentuale che si ritiene verrà utilizzata della quota disponibile.

La perdita attesa non rappresenta la vera e propria perdita in caso di insolvenza della controparte, bensì è il valore medio della distribuzione delle perdite; essa viene inclusa nello spread creditizio al momento dell'erogazione, ed è considerata una componente di costo stabile.

L'altro parametro per la valutazione del rischio di credito è la perdita inattesa o Unexpected Loss (UL) e rappresenta la variabilità della perdita intorno al suo valore medio; è considerata il vero fattore di rischio in ottica di portafoglio, in quanto costituisce la possibilità che i tassi effettivi di insolvenza e di perdita siano più elevati di quelli previsti. La perdita inattesa può essere calcolata come segue:

$$\boldsymbol{UL = LGD \times \sqrt{PD \times (1 - PD)}} \qquad (4)$$

in caso di LGD deterministica;

$$\boldsymbol{UL = \sqrt{LGD^2 \times PD \times (1 - PD) + PD \times \sigma^2_{LGD}}} \qquad (5)$$

In caso di LGD stocastica.

La variabilità della perdita (UL) dipende strettamente dal grado di correlazione tra i singoli crediti.

La perdita attesa non può essere eliminata con la diversificazione del portafoglio crediti (non diversificabile). La perdita inattesa può essere ridotta con una diversificazione di portafoglio (diversificabile: minore è la correlazione tra le perdite inattese, minore è la perdita inattesa di portafoglio).

La forma della distribuzione delle perdite per crediti, o Credit Loss Distribution, rappresentata in Figura 1, è asimmetrica a destra; questa forma presenta, infatti, la coda di destra più lunga rispetto a quella di sinistra, indice dell'incertezza del livello effettivo delle perdite.

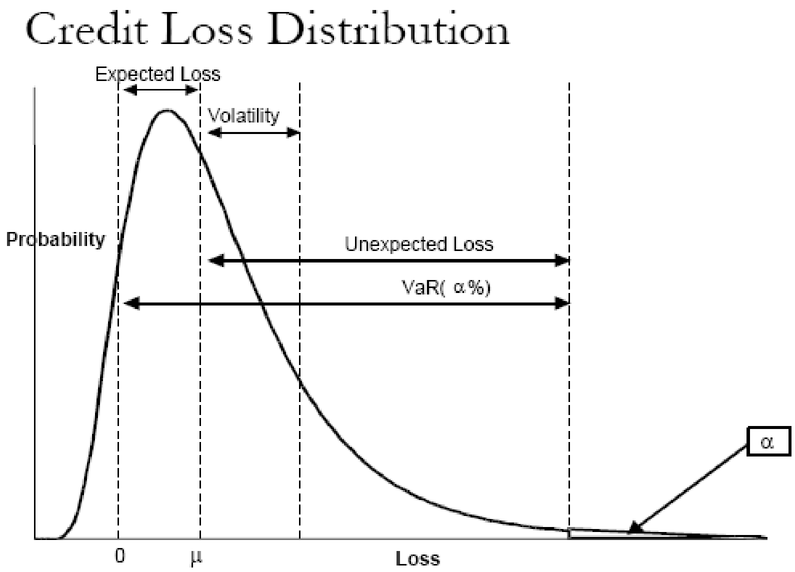


FIGURA 1 - CREDIT LOSS DISTRIBUTION[1]

La figura compara il livello di probabilità sulle ordinate con il livello della perdita sulle ascisse. La prima parte a sinistra rappresenta l'Expected Loss ed è coperta con accantonamenti, o prevista al momento della concessione del credito, quindi inclusa nel pricing. In questa area si verificano gli eventi con maggiore probabilità di accadimento.
La parte centrale rappresenta l'Unexpected Loss, corrispondente alla volatilità rispetto alla media, e la perdita conseguente è coperta con il Capitale Economico.
La parte a destra rappresenta gli scenari peggiori corrispondenti al default della controparte; si tratta di eventi con probabilità di accadimento basse, ma con grande impatto economico.

Il Capitale Economico rappresenta l'ammontare di capitale necessario per far fronte le perdite inattese. Non è sufficiente conoscere il livello della perdita inattesa per stimare l'ammontare necessario, occorre, considerare anche il livello di confidenza desiderato, indice della propensione al rischio. La distribuzione delle perdite su crediti, asimmetrica,

[1] Fonte: F. Varetto, corso di "Economia degli intermediari finanziari", A.A. 2017/2018, Politecnico di Torino.

non è approssimabile con una Variabile Casuale normale e, per calcolare il parametro associato al livello di confidenza voluto, si può operare in due modi:

- non si costruisce la funzione di probabilità, bensì si stimano le perdite storiche ed il relativo tasso di perdita, arrestando l'analisi al percentile desiderato (99%, 95%, etc.) e stimando così il VaR: Value At Risk.

- Si approssima la distribuzione Credit Loss, alla distribuzione Beta (tipica distribuzione asimmetrica). Tale distribuzione è caratterizzata da due parametri, α e β, calcolabili direttamente dai valori di media e varianza della distribuzione. Si utilizzano, pertanto, i valori di EL intesa come media e UL intesa come varianza.

$$\boldsymbol{\alpha} = \frac{\boldsymbol{EL^2 * (1 - EL)}}{\boldsymbol{UL^2}} - \boldsymbol{EL} \qquad (6)$$

$$\boldsymbol{\beta} = \frac{\boldsymbol{EL^2 * (1 - EL)}}{\boldsymbol{UL^2}} + \boldsymbol{EL} - \boldsymbol{1} \qquad (7)$$

Calcolando, con la Beta inversa, il valore del percentile corrispondente a una probabilità pari al livello di confidenza desiderato (99% ad esempio) si ottiene il livello della perdita nella coda della distribuzione per via analitica (UL99%); a partire da questo valore è possibile calcolare il VaR a quel determinato intervallo di confidenza:

$$\boldsymbol{VaR_{99\%}} = \boldsymbol{UL_{99\%}} - \boldsymbol{EL} \qquad (8)$$

1.2 Contesto normativo – Accordi di Basilea e principali regolamentazioni

Gli istituti bancari presentano una regolamentazione fondata sulla solidità patrimoniale, con lo scopo di eliminare, o ridurre il rischio di default. La regolamentazione, adattata alle varie realtà operative nazionali, si basa sulle norme stabilite dal Financial Stability Board e dal Comitato di Basilea per la Vigilanza Bancaria.

Il Comitato di Basilea è un organismo internazionale di cooperazione istituito nel 1974, composto dai rappresentanti delle banche centrali ed autorità di vigilanza dei paesi del G10. Le riunioni del Comitato avvengono nella sede della Banca per i Regolamenti Internazionali a Basilea.

Il comitato non possiede autorità di vigilanza sovranazionale, ma fornisce standard e linee guida alle autorità individuali che le adottano nel proprio contesto. Oggi le legislazioni non riguardano solo i paesi del G10, ma anche gli altri Paesi, che hanno deciso di adottare volontariamente le regole emanate dal Comitato.

Prima della fondazione del Comitato, ogni Paese decideva i propri criteri per determinare l'adeguatezza patrimoniale degli intermediari; Paesi come il Giappone, ad esempio, non avevano alcuna regolamentazione. Tale difformità determinava a livello internazionale profonde disparità concorrenziali tra i vari sistemi bancari.

Gli obiettivi del primo accordo sono stati quelli di rafforzare la solvibilità e la solidità dei sistemi bancari ed in parallelo di ridurre le disparità competitive.

1.2.1Accordi di Basilea I

Il Primo Accordo (1988) riguardava solo le banche internazionali, considerate efficienti e ben gestite dal punto di vista dei rischi e delle perdite attese. L'attenzione del Comitato si è concentrata, quindi, sulla necessità di assicurare la copertura delle perdite inattese, ovvero sui requisiti minimi di capitalizzazione.

Con l'Accordo furono introdotti, per la prima volta, dei requisiti di capitale uniformi e correlati alla rischiosità delle attività che le banche sostenevano.

Il requisito di capitale è stato definito in base a tre elementi principali:

1- Capitale di vigilanza: capitale destinato a difendere i creditori dalla possibilità di perdite. Suddiviso in due blocchi: Patrimonio di base (Tier 1 Capital) e Patrimonio supplementare (Tier 2 Capital); il primo comprende il Capitale Sociale e le riserve palesi, mentre il secondo comprende le riserve di rivalutazione, le riserve occulte e strumenti ibridi di capitale.
2- Rischio: valutate cinque fattori di ponderazione per le esposizioni in base alla loro rischiosità:

   - Rischio nullo 0%: contante e crediti verso governi centrali e banche centrali dei Paesi OCSE;

- Rischio basso 20%: crediti verso banche multilaterali di sviluppo e crediti garantiti e/o emesse da tali istituzioni;
- Rischio medio 50%: mutui assistiti da garanzie reali;
- Rischio pieno 100%: crediti verso imprese private, partecipazioni in imprese private, crediti verso banche e governi non OCSE.

3- Rapporto minimo tra capitale e rischio: l'Accordo prevede che le banche abbiano un livello di capitale almeno pari all'8% delle attività ponderate per il rischio.

L'accordo considerava solo il rischio di credito, ma con un successivo aggiornamento fu introdotto anche il rischio di mercato.

La Regulation di Basilea I può essere così sintetizzata:

$$\frac{Patrimonio\ di\ Vigilanza}{RWA + 12.5 * (Rischio\ di\ mercato)} \geq 8\%$$

$-$ RWA: Risk Weighted Assets $=$ Credito $*$ fattore di ponderazione;

- 12.5: coefficiente per la ponderazione dei rischi di mercato;

Queste ponderazioni hanno, tuttavia, dimostrato di essere troppo semplicistiche, creando incentivi a costruire degli arbitraggi regolamentari per alterare i portafogli bancari con l'obiettivo di massimizzare il valore per gli azionisti, sotto il vincolo della regolamentazione e, soprattutto, non sono risk-sensitive.

Il risultato è stato quello di inserire più rischi nei portafogli bancari, anziché il contrario: poiché non veniva differenziata ai fini regolamentari la concessione di un credito a basso rischio da uno ad alto contenuto di rischio, la banca aveva l'incentivo ad investire in crediti rischiosi, che offrono rendimenti più elevati, rispetto all'investimento in crediti a minore rischio ed a minore rendimento, essendo entrambi soggetti allo stesso accantonamento di capitale. (2)

1.2.2 Accordi di Basilea II

Data la superficialità dei primi accordi internazionali, nel 2004, in seguito a molti anni di consultazioni interne, il Comitato di Basilea promulga gli accordi di Basilea II. La nuova regolamentazione mira a rafforzare la normativa concernente la stabilità delle banche attraverso la definizione dei loro requisiti patrimoniali e il miglioramento dei metodi di misurazione e gestione dei rischi.

La regolamentazione di Basilea II si fonda su tre pilastri[2]:

1. CAPITAL REQUIREMENT: il vincolo dell'8% non muta, ma ora deve coprire anche il rischio operativo; resta invariato il calcolo del rischio di mercato, mentre il rischio di credito viene quantificato in modo più sofisticato, introducendo la possibilità per le banche di ponderare i propri attivi in base ai propri modelli interni (Risk Weight Asset RWD).

$$\frac{Patrimonio\ di\ Vigilanza}{Rischi\ di\ credito + Rischi\ di\ mercato + Rischi\ operativi} \geq 8\%$$

   I metodi per il calcolo delle attività ponderate al rischio sono volti a migliorare la valutazione della rischiosità, e rendere più significativi i coefficienti patrimoniali.

   Tale coefficiente fissa l'ammontare minimo di capitale che le banche devono possedere in rapporto al complesso delle attività ponderate in base al loro rischio creditizio. Sono stati introdotti i concetti di rischio operativo (ad es. frode interna, frode esterna, risarcimenti richiesti da dipendenti, violazione delle norme a tutela della salute e sicurezza del personale, pratiche discriminatorie, responsabilità civile e penale) e di rischio di mercato (definito come il rischio di perdite derivanti da negoziazione di strumenti finanziari sui mercati, indipendentemente dalla loro classificazione in bilancio).

2. SUPERVISORY REVIEW: l'Autorità di Vigilanza bancaria ha la possibilità di imporre, in seguito a dei controlli bancari, ove ritiene necessario, requisiti patrimoniali più elevati di quelli previsti dalla Regulation di Basilea II. Si pone, quindi, un controllo continuo del Capitale bancario per evitare che scenda sotto i minimi, o non salga a sufficienza per sostenere i rischi. È precisato il ruolo degli Organi di Vigilanza che sono tenuti a monitorare costantemente l'adeguatezza dei livelli di capitalizzazione rispetto ai rischi e a valutare la coerenza delle politiche gestionali messe in atto dalle banche per rispettare gli indicatori stabiliti dalla normativa.

[2] Web: http://www.economiaoggi.it/Basilea-2/

3. MARKET DISCIPLINE = il mercato è il miglior giudice per valutare e prezzare il contenuto di rischio di una banca; le banche, inoltre, devono soddisfare criteri di "disclosure", di trasparenza e di informazioni al mercato.

   Sono previste regole di trasparenza per l'informazione al pubblico sui livelli patrimoniali, sui rischi e sulla loro gestione. Il terzo punto introduce cambiamenti nella diffusione di informazioni da comunicare al pubblico sia per quanto attiene il bilancio che le aree di rischio.

   In questo modo, diviene possibile esprimere una valutazione sull'adeguatezza della capitalizzazione della banca e, inoltre, discriminare tra banche che hanno sistemi di gestione appropriati e capitalizzazione adeguata e banche che si trovano, invece, in condizioni diverse.

Il secondo e il terzo pilastro sono di particolare interesse per il settore bancario e per gli operatori finanziari, mentre il primo assume una rilevanza indiscutibile per il soggetto a cui si deve concedere il finanziamento (le imprese).

In merito alla valutazione del rating creditizio, Basilea II introduce delle importanti novità: è possibile, per ogni istituto bancario, ricevere i parametri per la valutazione del rischio di credito da agenzie esterne, oppure stimarli internamente. (3)

Sono stati introdotti, quindi, due approcci:

- STANDARDIZED APPROACH:

  Le ponderazioni di rischio sono determinate in base alla categoria dei debitori (come in BASILEA I del 1988) [paesi sovrani, banche, imprese] ma la ponderazione è attribuita sulla base del rating assegnato alla controparte da Agenzie esterne, denominate ECAI (Eligible External Credit Assessment Institution).

  Nel consentire l'uso del metodo standardizzato il Comitato si è reso conto di autorizzare il ricorso di valutazioni di agenzie di rating esterne (ECAI) per la stima del merito creditizio: ha delegato, quindi, le autorità di vigilanza nazionali ad accertarsi che tali istituzioni soddisfino criteri minimi di obiettività, indipendenza, trasparenza, pubblicità delle informazioni, risorse e credibilità.

  Nella Tabella 1 sono rappresentati i nuovi parametri di ponderazione del rischio.

TABELLA 1 - VALUTAZIONE DEI FATTORI DI PONDERAZIONE AL RISCHIO

	AAA	AAA-	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	Inferiore	Senza Rating	Scaduti
Corporate	20%					50%			100%						150%				100%	150%
Stati sovrani	0%					20%			50%			100%						150%	100%	
Banche	20%					50%						100%						150%	50%	
Banche (Paese d'origine)	20%											150%						150%	100%	
Retail: (privati e PMI)	75%																			150%
Mutui residenziali	35%																			100%
Mutui commericali	Da 100% a 50% a scelta delle Autorità nazionali																			150%

Al 21 luglio 2011 le ECAI riconosciute da Banca d'Italia sono 4: Fitch, Standard&Poor's, Moody's e Cerved Group.

- INTERNAL RATING-BASED APPROACH: IRB

L'approccio IRB, o metodo dei Rating Interni, concede agli istituti bancari la facoltà di stimare i parametri per la valutazione dei rating creditizi internamente. A differenza rispetto al metodo standard, che si basa principalmente sul rating esterno assegnato alla controparte, nei metodi IRB le banche effettuano internamente delle valutazioni sui debitori e stimano il capitale necessario per coprire la massima perdita che potrebbe registrarsi in un dato periodo di tempo con una certa probabilità. Sono dunque calcolati i coefficienti di ponderazione tenendo conto dei seguenti elementi qualitativi:

- l'esposizione al momento del default (Exposure At Default, EAD): il valore delle attività di rischio per cassa e fuori bilancio (garanzie rilasciate e impegni). Per queste ultime si fa ricorso ad uno specifico fattore di conversione creditizia (Credit Conversion Factor, CCF);

- la probabilità di default (Probability of default, PD): probabilità riferita a ogni singolo debitore o ai pool (aggregati di attività) che passi allo stato di insolvenza in un orizzonte temporale di un anno;

- la perdita in caso di default (Loss given default, LGD): valore atteso del rapporto tra la perdita relativa al default e l'importo dell'esposizione al momento del default (EAD). Per perdita si tiene conto dei flussi recuperati e dei costi diretti e indiretti collegati al recupero dei crediti, che devono essere attualizzati utilizzando un opportuno tasso di interesse;

- la scadenza effettiva (Maturity, M): la media, delle durate residue contrattuali, per una data esposizione, ciascuna ponderata per il relativo importo;

- la ponderazione dei rischi (Risk Weighting, RW);

- l'aggiustamento per il grado di frazionamento del portafoglio (granularity, G): correzione da apportare al totale delle attività ponderate per rischio per includere nel sistema di calcolo il livello di diversificazione dell'attivo[3].

Per le classi di attività diverse dalle esposizioni al dettaglio, il metodo dei rating interni si presenta in due tipologie distinte di calcolo in relazione ai parametri di rischio da stimare: Foundation IRB (FIRB) ed Advanced IRB (AIRB). I due metodi si differenziano per la stima dei parametri.

- FIRB: Le banche che adottano il modello FIRB, o Sistema di Rating Interno di base, stimano internamente solo la probabilità di default, mentre gli altri parametri (EAD, LGD, M, RW) sono prefissati esternamente dalle Autorità di vigilanza, pertanto le banche le applicano semplicemente nei loro modelli.

- AIRB: Le banche che adottano, invece, il modello AIRB, o Sistema di Rating Interno avanzato, stimano, oltre la Probabilità di Default, anche tutti gli altri parametri internamente. Tuttavia, prima di poter applicare ufficialmente questi metodi di stima, le banche sono tenute a presentarli e dimostrarne la validità, l'efficacia e la solidità degli stessi, all'Autorità di Vigilanza, che avrà la facoltà di approvarli, o respingerli.
  I modelli di Rating Interni devono rispettare i seguenti requisiti minimi per poter essere utilizzati:
  - valutare separatamente la PD e la LGD;
  - i crediti sono distribuiti tra le varie classi di rating, senza concentrazione in una specifica classe;
  - il rating va assegnato ai debitori prima della concessione del prestito;
  - il rating va rivisto periodicamente;
  - il rating va utilizzato nella gestione dei crediti e nel pricing dei prestiti;

[3] Web: http://www.bankpedia.org/index.php/it/115-italian/m/21091-metodo-dei-rating-interni-irb

- la banca deve disporre di un adeguato sistema di validazione dell’accuratezza e coerenza del SRI;
- vi sono inoltre requisiti di documentazione formale del SRI e del suo funzionamento. (4)

Per la definizione di default le banche possono imporre una condizione soggettiva o una oggettiva: la prima è definita secondo giudizi interni; la seconda è definita quando il debitore è in ritardo di più di 90 giorni in almeno un pagamento.

1.2.2.1 Limiti di Basilea II:

La crisi finanziaria 2008 ha mostrato alcuni limiti di Basilea II, anche se il nuovo framework era in fase di avvio e non operativo:

- Si è riscontrato un insufficiente capitale, nonostante formalmente le banche rispettassero i requisiti di capitale (il patrimonio Tier 1 delle banche europee era in media pari all’8% (rispetto al 4% previsto dalla Regulation); inoltre si è riscontrato un insufficiente qualità del capitale. Molte banche, infatti, avevano aumentato il loro capitale con strumenti ibridi e non con strumenti del Core Tier 1; cospicui dividendi avevano contribuito ad impoverire ulteriormente il patrimonio di migliore qualità.
- Sopravalutazione del cosiddetto “tocco leggero” nell’azione di vigilanza, cioè supervisione leggera, basata su un perimetro limitato di regole imposte dall’Autorità Interbancaria, con un ampio affidamento alla razionalità dei comportamenti degli intermediari, considerati meglio in grado di individuare, misurare e gestire i rischi (la regolamentazione fino a poco tempo fa era percepita come un fattore distorsivo dell’efficienza dei mercati finanziari).
- Si è riscontrato il problema della pro-ciclicità: il requisito di capitale è sensibile all’andamento del ciclo economico. La misura dei rischi si attenua nella fase ascendente del ciclo e tende invece a crescere nei momenti di crisi: una stessa quantità di patrimonio bancario quindi sostiene un maggior volume di attività nelle fasi di crescita ed un volume minore durante i momenti di declino, in cui invece l’economia avrebbe bisogno di essere sostenuta. La pro-ciclicità è ampliata anche dall’uso di modelli molto simili da parte delle banche, che generano comportamenti collettivi che amplificano le dinamiche del credito.
- Vi è stata un’inadeguatezza nel prevenire crisi sistemiche: la regulation è di tipo micro-prudenziale ed in quanto tale non può tenere conto delle interazioni tra gli intermediari e tra i mercati e tra questi e le imprese e le famiglie (la stabilità micro non garantisce la stabilità

macro). La percezione dei rischi potenzialmente sistemici avviene tramite l'uso di stress test riguardanti scenari con profili estremi ma plausibili: peraltro l'attuale crisi dimostra che capitalizzazioni anche superiori al minimo non sono in grado di reggere alle perdite che si verificano in crisi sistemiche.

- Si è ipotizzato, inoltre, che la volatilità fosse costante, ma si avrebbe dovuto tenere conto che la volatilità aumenta in misura significativa durante i periodi di instabilità.
- Si sono ipotizzati rendimenti distribuiti secondo la distribuzione normale (simmetrici e con code non grasse) mentre era già ampiamente noto che la distribuzione dei rendimenti è asimmetrica. Il periodo di tempo usato per la stima dei parametri, inoltre, era relativamente corto e stabile (caratterizzato dalla "grande moderazione") e non incorporava a sufficienza episodi di crash di mercato.

Data la crisi del 2008 e i relativi problemi riscontrati si è deciso di rivedere Basilea II con un iter molto lungo che è terminato con l'accordo del 7 dicembre 2017 (Basel 3: Finalising post-crisis reforms), ed entrerà in vigore nel 2022. (5)

1.2.3 Accordi di Basilea III

Gli obiettivi della Regulation di Basilea III possono essere sintetizzati in tre pilastri:

1) Primo Pilastro
    a. Capitale
        i. Qualità e livello del patrimonio di vigilanza:
           Vi è una maggiore attenzione sulle azioni ordinarie e le riserve di utili (common equity). Il requisito minimo (Core Tier 1) è innalzato al 4,5% delle attività ponderate per il rischio.
        ii. Assorbimento delle perdite al punto di non sopravvivenza:
           Gli strumenti del patrimonio di vigilanza devono essere provvisti di una clausola contrattuale che ne consenta la cancellazione o conversione in azioni ordinarie qualora la banca non sia più ritenuta solvibile.
        iii. Buffer di conservazione del capitale:
           Il Common Equity in misura circa del 2,5% delle attività ponderate per il rischio sarà portato al 7% con l'introduzione di un buffer prudenziale.
        iv. Buffer anticiclico:

Quando il common equity è compreso tra lo 0 e il 2,5%, è imposto dalle autorità se ritengono che la crescita del credito generi un accumulo di rischio sistematico.

In Tabella 2 sono rappresentate le principali differenze tra i requisiti di capitale tra Basilea II e Basilea III.

TABELLA 2 - DIFFERENZE TRA I REQUISITI DI CAPITALE TRA BASILEA II E BASILEA III

QUALITA' PATRIMONIO	REQUISITO	BASILEA II	BASILEA III
Common Equity (Core Tier 1)	Minimo	2%	4.50%
	Conservation Buffer		2.50%
	Totale		7%
Tier 1	Minimo	4%	6%
	Totale		8.50%
Total Capital	Minimo	8%	8%
	Totale		10.50%
Requisiti macroprudenziali	Buffer anticiclico		0-2.5%
	Banche SIFI		lavori in corso

b. Copertura dei rischi
   i. Cartolarizzazioni:
      Le banche devono effettuare analisi più rigorose del merito di credito per le posizioni cartolarizzate provviste di rating esterno.
   ii. Portafoglio di negoziazione:
      è necessario un aumento di capitale a fronte di attività di negoziazione e di strumenti derivati, in generale per operazioni complesse. Vi è l'introduzione di un requisito basato sul valore a rischio in condizioni di stress (stressed VaR) per attenuare la pro-ciclicità. Il requisito patrimoniale deve tener conto dei rischi di insolvenza e di migrazione di rating dei prodotti creditizi non cartolarizzati.
   iii. Esposizioni verso controparti centrali (CCP):
      Le esposizioni di negoziazione verso le CCP idonee devono ricevere una ponderazione di rischio del 2% e quelle verso i loro fondi di garanzia (default fund) devono essere trattate secondo il metodo basato sul rischio.

c. Contenimento della leva finanziaria
   i. Indice di leva finanziaria (leverage ratio):

L’indice di leva finanziaria tiene conto delle esposizioni fuori bilancio e serve da complemento ai requisiti patrimoniali basati sul rischio. Contribuisce inoltre a contenere l’accumulo di leva finanziaria a livello di sistema.

2) Secondo Pilastro
    a. Gestione dei rischi e vigilanza
        i. Requisiti supplementari nell’ambito del secondo pilastro:
        I requisiti riguardano la gestione del rischio e la governance a livello di impresa. La rilevazione del rischio connesso alle esposizioni fuori bilancio e le operazioni di cartolarizzazione deve seguire specifici fattori come incentivi per una migliore gestione dei rischi e dei rendimenti di lungo periodo; procedure corrette di remunerazione; pratiche di valutazione; prove di stress.
3) Terzo Pilastro
    a. Disciplina di mercato
        I nuovi requisiti si riferiscono alle esposizioni cartolarizzate e alle operazioni fuori bilancio. Sono richieste maggiori informazioni sulle componenti del patrimonio di vigilanza e una spiegazione approfondita delle modalità di calcolo dei coefficienti patrimoniali regolamentari.

Con Basilea III sono introdotti dei requisiti globali che riguardano la liquidità e il monitoraggio regolamentare. Vi sono anche disposizioni speciali per le istituzioni finanziarie di rilevanza sistemica (SIFI) in quanto sono tenute a dotarsi di una maggiore capacità di assorbimento delle perdite coerentemente con i maggiori rischi che pongono per il sistema finanziario. Al fine di determinare quali banche vadano considerate sistemicamente rilevanti, è stata formulata una metodologia comprendente criteri sia quantitativi sia qualitativi[4]. (6)

[4] Web: https://www.bis.org/bcbs/basel3/b3summarytable_it.pdf

1.2.3.1 Rivalutazione degli accordi di Basilea III

Il Comitato di Basilea nel dicembre 2015 ha emesso un nuovo documento di consultazione per la revisione dell'approccio standardizzato sul rischio di credito. L'obiettivo è quello di ridurre il riferimento alle agenzie di rating esterno. Fra i motivi principali vi è stato il comportamento delle agenzie di rating nel creare condizioni favorevoli per lo scoppio della crisi del 2008. Ed inoltre vi è stato il principio di fondo tale per cui "chi si assume dei rischi deve essere in grado di valutarli, senza delegare a terzi questa funzione essenziale".
Per far ciò, il Comitato ha proposto di affiancare al giudizio delle agenzie di rating un requisito di due-diligence, per ottenere una adeguata comprensione del profilo di rischio e delle caratteristiche delle loro controparti.
Per valutare le ponderazioni delle esposizioni la banca è tenuta a fornire una "ponderazione di base", da affiancare al giudizio delle agenzie.

Anche la revisione della normativa sui modelli interni di rating è in corso di revisione: secondo le proposte in discussione, l'utilizzo dei modelli interni dovrebbe rimanere limitato solo a portafogli di crediti sui quali possono essere stimati in modo sufficientemente robusto i parametri rilevanti per la valutazione del rischio (essenzialmente PD, LGD). Inoltre, verrebbero introdotti dei limiti minimi (input floor) ai valori dei parametri di input ed ai RWA ottenuti dai calcoli (output floor): questo orientamento nasce dall'obiettivo di ridurre la variabilità dei RWA tra i diversi sistemi bancari.

È anche in fase di consultazione la metodologia di calcolo dei rischi operativi: viene eliminata la possibilità di avvalersi di modelli interni e resta una sola metodologia standardizzata, basata sulla combinazione di business (con riferimento a voci del conto economico) e di perdita (riferita alle perdite operative registrate dalla banca). (7)

1.2.4 Accordi di Basilea "IV"

A dicembre 2017 il Comitato di Basilea per la vigilanza bancaria ha emesso un documento di revisione finale della regulation Basilea III, dopo lunghe discussioni e rinvii, causati prevalentemente dalla mancanza di accordo a livello internazionale tra gli Stati Uniti e gli altri paesi sviluppati, Europa soprattutto, sulle nuove regole, per il timore che vengano richiesti ulteriori cospicui aumenti di capitale, maggiori per i sistemi bancari europei e minori per quello americano.

L’obiettivo dichiarato della riforma del 2017, convenzionalmente nota come Basilea IV, è quello di restituire credibilità ai calcoli dei RWA (risk weighted assets) e migliorare la confrontabilità dei ratios patrimoniali delle banche. Ripetute analisi hanno infatti messo in luce una variabilità eccessiva dei RWA tra le diverse banche, apparentemente non spiegabile da differenze nella rischiosità dei loro portafogli.
I modelli interni di valutazione dei rischi, inoltre, se non ben calibrati, possono essere usati, più che per una onesta quantificazione dei rischi, per minimizzare il requisito patrimoniale della banca: vi è infatti un chiaro incentivo a sviluppare modelli interni orientati in tale direzione.
Per questi motivi, la riforma del 2017 introduce una serie di vincoli all’uso dei modelli interni, che entreranno in vigore a partire dal 2022.

A livello europeo, nel frattempo, con l’introduzione del Singol Supervisoy Mechanism (S S M) e dell’European Bank Authority (E B A) è stato avviato un iter molto complesso per ridare ai modelli interni misurazione del rischio maggiore credibilità ed eliminare “l’injustified variability” tra le banche.

2 Modelli di Rating Interni

2.1 Il rating di agenzia e il rating bancario

La Regulation di Basilea II introduce la possibilità, per gli istituti bancari, di calcolare e valutare il merito creditizio dei propri clienti. Le banche possono avvalersi anche del giudizio delle agenzie di rating: enti esterni adibiti alla diffusione dei giudizi riguardanti il merito creditizio di un'impresa; possono anche stimare internamente tali rating e valutare la bontà creditizia dei propri clienti.

Vi è una differenza sostanziale tra il giudizio delle agenzie di rating e i sistemi di rating interni applicati dalle banche: le prime emettono il rating di un'impresa e queste informazioni e valutazioni sono diffuse sul mercato; le banche, invece, effettuano delle valutazioni interne sul merito creditizio delle proprie controparti.

Anche la modalità di emissione del rating è diversa. Le agenzie di rating effettuano una valutazione definita "through the cycle", in quanto il loro obiettivo è dare un giudizio oggettivo, da modificare raramente; il rating è assegnato simulando scenari pessimistici circa le condizioni in cui opera l'impresa per comprendere la capacità di rimborso anche nelle situazioni difficili. In tal modo, le agenzie, cambiano di rado i loro giudizi fornendo una sorta di stabilità agli investitori operanti nel mercato. Questi, avendo una valutazione che resiste anche negli scenari peggiori saranno portati a fidarsi di tale giudizio, anziché vedere le considerazioni cambiate ad ogni scenario e quindi considerate provvisorie e non meritevoli di attenzione.

Le banche, a differenza delle agenzie che hanno l'obiettivo di immettere le info sul mercato, formulano delle valutazioni per le proprie finalità interne. Sono, quindi, interessate ad avere valutazioni che riflettono nel modo più preciso possibile la situazione dell'impresa e le sue prospettive, nei vari momenti. Se le condizioni dell'impresa mutano, il rating delle banche deve reagire immediatamente ai cambiamenti (soprattutto se si tratta di deterioramenti della situazione economico-finanziaria): il rating quindi è definito "point-in-time" o "ibrido".

Ciò che rende i rating assegnati dalle banche più variabili è anche l'orizzonte di valutazione: le operazioni bancarie stimano per normativa una probabilità di default a un anno, pertanto, le previsioni incorporate nel rating coprono quel periodo. Le valutazioni delle agenzie di rating, invece, coprono un periodo più ampio variabile intorno ai cinque anni. Esse non modificano il rating al variare delle condizioni economiche, bensì implicitamente modificano le probabilità di default associate a ciascuna classe di rating. Viceversa, i modelli interni delle banche tendono a modificare le classi di rating ma mantenendo stabili le probabilità di default associate a ciascuna classe. (8)

2.2 I sistemi di rating interni (SRI)

2.2.1 Regole di Basilea sui SRI

La costruzione di un sistema di rating interno comporta il rispetto di diverse regole che la Regolamentazione impone:

- avere una congrua distribuzione delle esposizioni tra i diversi gradi di merito, senza eccessive concentrazioni di debitori e di operazioni; ci devono essere almeno sette gradi di merito per i crediti in bonis ed uno per quelli in default;
- l'orizzonte temporale di stima delle PD è di un anno, ma nell'assegnazione dei rating le banche dovrebbero adottare un orizzonte più esteso;
- il grado di merito deve rappresentare la valutazione della capacità e volontà del debitore di onorare i propri impegni nonostante l'insorgere di condizioni avverse o eventi inattesi. Le condizioni economiche considerate, nell'effettuare la valutazione, devono essere compatibili con la situazione corrente e la sua presumibile evoluzione nell'arco di un ciclo congiunturale nel settore o/e nell'area geografica.
- i modelli statistici (credit scoring o risk differentiation) sono ammissibili come base primaria o parziale per l'assegnazione dei rating, ma è necessaria un'adeguata valutazione per far sì che vengano prese in considerazione tutte le informazioni pertinenti e rilevanti e che il modello sia usato in modo corretto. Le variabili del modello devono formare un insieme ragionevole di indicatori predittivi. La banca deve disporre di procedure per la revisione dei rating assegnati dal modello. Le procedure devono essere orientate alla individuazione ed eliminazione degli errori del modello; attraverso un ciclo regolare di validazione, la banca deve verificare la

performance del modello, la sua stabilità, la verifica delle correlazioni ed il confronto periodico delle risultanze previste del modello con gli esiti effettivi;

- il SRI deve essere adeguatamente documentato. In particolare, per i modelli devono essere stabiliti rigorosi procedimenti statistici per la loro validazione (con campioni di controllo diversi da quelli di stima) e devono essere chiarite le ipotesi su cui si fondano e le circostanze in cui operano in modo inefficace.
- i rating devono essere aggiornati almeno una volta all'anno;
- i SRI devono avere un ruolo essenziale nella gestione del rischio e nell'allocazione interna del capitale.

2.2.2 Costruzione di un Sistema di Rating Interno

Nella costruzione di un Sistema di Rating Interno, bisogna fare in modo che il metodo sia:

- quantitativo e rappresentativo: deve misurare la probabilità che accada un evento di default nel corso dell'intervallo di tempo definito;
- significativo e robusto: deve essere calcolato su fattori economici rilevanti come ad esempio i dati di bilancio, informazioni qualitative e dati andamentali;
- oggettivo: si deve sempre giungere allo stesso giudizio di qualità creditizia della controparte, a parità di dati e algoritmi utilizzati;
- confrontabile: deve essere riconducibile ad un'unica scala di valutazione e, quindi, comparabile con altri profili di rischio calcolati con lo stesso sistema.

Pertanto, tenendo conto di queste considerazioni, la costruzione di un modello di rating si articola in diversi step:

- struttura del modello e caratteristiche principali:
  In questa prima fase, sono definiti la struttura e i confini di applicabilità del modello. La Figura 2 mostra la struttura esemplificativa di un modello di rating; questo utilizzando dati quantitativi, quali dati anagrafici, dati finanziari e dati andamentali, definisce uno score statistico che sarà poi combinato ad uno score qualitativo, ottenuto dai dati di tipo qualitativo. La combinazione tra i due definisce uno score integrato, comprendente entrambi i tipi di caratteristiche, e da questo si ottiene un rating integrato. A questo punto finisce la fase di calibrazione del modello e si passa alla fase di valutazione, che porterà alla definizione del rating finale.

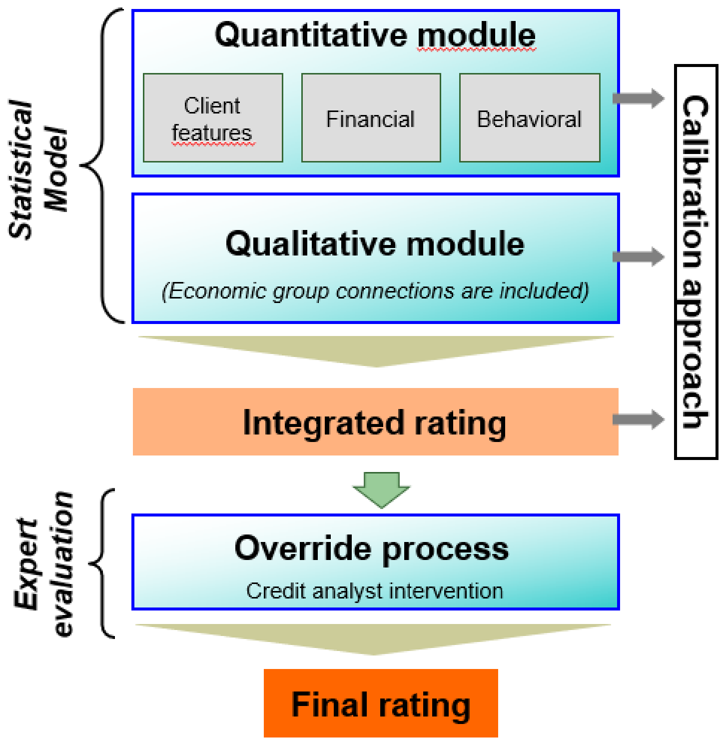


FIGURA 2 - STRUTTURA DI UN MODELLO DI RATING[5]

- Raccolta dati:

  La raccolta dei dati è un processo accurato che deve riflettere le caratteristiche generali del portafoglio (settore, dimensioni, dimensioni geografiche); prima di essere processati, tuttavia, i dati devono essere "ripuliti" da rumori di fondo.

  Occorre definire quale sia la variabile target da utilizzare per la definizione di default e l'orizzonte temporale.

  Vi è la necessità di definire il campione di training del modello (in sample) per lo sviluppo e il campione di stima (out of sample) per la fase di validazione.

  Per poter operare al meglio sono necessarie serie temporali sufficientemente lunghe.

- Analisi univariata:

  A partire da un grande elenco di variabili da testare, dai dati di bilancio agli indicatori comportamentali, definito long-list di indicatori è effettuata un'analisi univariata che ha come obiettivo quello di trasformare gli indicatori e passare da una long-list ad

[5] Fonte: Group Risk Management Intesa Sanpaolo S.p.A.

una short-list, mantenendo solo le variabili che sono realmente rilevanti per il modello. La nuova lista di indicatori, deve tuttavia rispettare i seguenti requisiti:
- potere discriminatorio: avere capacità di suddividere il campione tra buoni e anomali (misurato tramite l'Accuracy Ratio);
- solidità economica: le variabili devono avere un senso economico rilevante;
- numero contenuto di valori mancanti;
- andamento monotono del tasso di default.

In seguito, è effettuata un'analisi di correlazione per individuare ulteriori variabili non significative e creare così la lista degli indicatori da inserire nel modello per ottenere una base di dati di regressione.

- Analisi multivariata:

  L'analisi multivariata ha come obiettivo quello di creare uno score, o variabile target, per stabilire il verificarsi dell'evento default/non default. Tale variabile è definita per ogni controparte. Lo score è ottenuto applicando un'analisi di regressione. Per ottenere un risultato valido è necessario che gli indicatori siano significativi (bassa correlazione) e di facile comprensione.

  Una volta ottenuto lo score per ogni controparte, si possono seguire due approcci:
  - raggruppare gli score in classi di rating, e per ognuno dei quali calcolare la probabilità di default intesa come frequenza di default della classe;
  - calcolare per ogni score una probabilità di default, e definire le classi di rating raggruppando insieme le controparti con simili livelli di rating.

- Calibrazione del modello:

  L'obiettivo di questa fase è l'aggiustamento dell'output (probabilità di default o score) con la media dei tassi di default storici del portafoglio analizzato, la "tendenza centrale", che solitamente comprende un intero ciclo economico. Questo processo è importante perché i modelli, di solito, sono stimati su una popolazione che fa riferimento ad un periodo di tempo breve e non rappresentativo di un ciclo economico: risulta influenzato, quindi, dal preciso momento in cui è stato stimato. (9)

2.2.2.1 Raccolta dati

- Definizione di Default:

Il primo passo per la costruzione di un Sistema di Rating Interno è la definizione di default. L'obiettivo è quello di classificare le controparti e predire il default un anno prima. Il default deve essere: realistico e bilanciato, comune tra le stime e il testing, e rispettare le regole di Basilea.

Vi sono tre definizioni di default:

  - SOFFERENZA: situazione influenzata da eventi significativi che hanno portato alla diminuzione della bontà creditizia del debitore; la controparte diviene quindi in uno stato di insolvenza, anche se non è riconosciuta tale legalmente. Gli eventi caratterizzanti possono essere: la dichiarazione di bancarotta o liquidazione (anche volontaria), azioni legali iniziate dalla banca o da terze parti, cessazione dell'attività di impresa (seppur con delle limitazioni);
  - INADEMPIENZE PROBABILI: stato definito dal Gestore, che percepisce la controparte in una situazione temporanea di difficoltà economica o finanziaria;
  - PAST-DUE: scaduti da oltre 90 giorni e dichiarati inesigibili anche legalmente. Non è prevista nessuna compensazione, né una soglia di recupero. Se la controparte ha più crediti scaduti, è considerato il ritardo maggiore;

É anche possibile definire tre tipi di performance per i crediti:

  - IN BONIS: nessun problema con il pagamento;
  - CREDITI PROBLEMATICI (o in proattivo): possibile deterioramento del credito se non si effettuano azioni di recupero;
  - DEFAULT TECNICO: crediti che per motivi tecnici sono considerati scaduti o in stato sub-standard ma che ritorneranno in bonis in breve periodo senza arrecare perdite

- Segmentazione e definizione dei confini del modello

Tutte le controparti sono suddivise in base alla loro categoria: credito sovrano, crediti verso banche e pubbliche istituzioni, imprese corporate, PMI, retail e altri debiti. Inoltre, i segmenti banche retail e corporate, essendo i segmenti chiave su cui è possibile applicare i modelli di rating interni, sono ancora suddivise in grandi e piccole, a seconda di due parametri: l'exposure e il turnover. L'idea di fondo, che vi è per la segmentazione, è che il trattamento dovrebbe essere uniforme per le controparti in simili situazioni, invece differente verso le altre. Le controparti in situazioni simili dovrebbero quindi ricevere lo stesso rating ed avere la stessa segmentazione.

- Fonti dei dati

Nello sviluppo di un modello di rating interno occorre: prendere in considerazione tutti i dati e tutte le informazioni che potrebbero essere rilevanti; scegliere un metodo attraverso un attento processo di selezione; creare un campione di stima basato sull'esperienza storica ed empirica a lungo termine e non sulla valutazione soggettiva; essere plausibile e facilmente interpretabile (la semplicità è un valore); rivedere l'intero processo ogni volta che sono disponibili nuove informazioni pertinenti o, in ogni caso, almeno una volta all'anno.

Il primo passo si basa sulla selezione delle origini dati. Vi sono due categorie di fonti di dati diverse: fonti interne, se le informazioni sono prodotte direttamente dalla banca (stato default/bonis, segmentazione, dati personali etc.); fonti esterne se le informazioni vengono acquistate da un fornitore esterno (score esterni, dati di mercato, rating di agenzia).

- Requisiti del campione

Prima di iniziare lo sviluppo del modello è necessario che siano rispettate alcune condizioni, per rendere il campione affidabile:

  - Completezza: la lunghezza delle serie storiche dei dati deve essere, il più possibile, rappresentativa delle condizioni attuali e futuri della realtà; la banca dovrebbe collezionare i dati relazionati all'intero ciclo economico per evitare distorsioni nel campione; collezionare dei dati "out of sample" e "out of time" per effettuare dei test di robustezza del modello.

- Accuratezza: i dati da usare nel modello devono essere scremati prima dell'applicazione nel modello per ottenere una certa qualità dei dati;
- Pertinenza: le informazioni collezionate devono essere integrate coerentemente in modo tale che la data del default sia perfettamente riconoscibile; l'impatto del default deve essere riconducibile a determinati valori, i quali dipendono dal tipo di portfolio a cui appartengono.
- Replicabilità: tutte le azioni effettuate per ottenere il campione di training devono essere replicabili;

- Problemi di rappresentazione del portafoglio

I campioni usati sia per lo sviluppo, o training del modello, sia per la stima, devono necessariamente essere rappresentativi della popolazione a cui si riferiscono, altrimenti si avrebbero dei risultati distorti, e soprattutto non affidabili. È necessario misurare la comparabilità tra i campioni e la popolazione tramite test statistici: Indice di stabilità del sistema (SSI) calcolato come segue:

$$\boldsymbol{SSI} = \sum(\boldsymbol{S} - \boldsymbol{P}) * \ln(\boldsymbol{S}/\boldsymbol{P}) \quad (9)$$

Nella quale S rappresenta la distribuzione del campione e P il target della popolazione. Per essere rappresentativo, il campione deve presentare un SSI inferiore a 0,10; per valori di SSI tra 0,10 e 0,25 si ha una bassa criticità; mentre per valori superiori a 0,25 si ha un'alta criticità.

2.2.2.2 Analisi univariata

Dopo aver definito i campioni, si passa ad operare con gli indicatori. Questi racchiudono ogni tipo di informazione sia di tipo finanziario, sia di tipo comportamentale o informazioni personali. Questi dati sono racchiusi all'interno di una long-list di indicatori che sono ritenuti d'interesse e con un potere predittivo per stimare la probabilità di default, tra i quali possiamo trovare indicatori riguardanti l'esperienza di risk-management, ma anche altri di letteratura scientifica, benchmarking etc.

Operare con una long-list di indicatori non è, dal punto di vista computazionale, vantaggioso e semplice, pertanto, questa è ridotta effettuando un'analisi univariata (single factor analysis) che ha l’obiettivo di identificare gli indicatori più performanti e che meglio racchiudono tutta l’informazione necessaria; è prassi considerare i seguenti criteri per formare una buona short-list di indicatori:

- alto potere predittivo;
- senso economico;
- pochi valori mancanti;
- tasso di default monotono.

- Tipo di dati
    a) Dati personali

Negli ultimi anni, si è notato una tendenza ad inserire nel modello anche dati personali, che di per sé non presentano alte performance per discriminare bene tra imprese sane e anomale, ma se inseriti nel modello, insieme ai dati di tipo quantitativo, danno un contributo positivo alle performance del modello. L’informazione racchiusa in questo pacchetto di indicatori riguarda, in generale, il settore di business (finanziario, servizi, operativo…), la forma legale (società di capitali, società di persone…), l’area geografica (nord, sud, centro, regione, città…) e l’anno di fondazione.

b) Dati finanziari

La maggior parte dei dati, usati nel modello, sono di tipo finanziario e provengono dai bilanci aziendali. È importante, quando si passa dalla long-list alla short-list, di conservare la maggior parte delle informazioni disponibili, che si possono raggruppare per aree:

- struttura di capitale e livelli di debiti: gli indicatori maggiormente usati, per definire quest’area sono: equity / (totale asset) e leverage o leva finanziaria;
- liquidità definita dal current-ratio e l’acid test;
- profittabilità misurata da indicatori come il ROE, ROI, ROS;
- capacità di risanare i debiti che si può calcolare come EBITDA/ (tot. costi), (tot. costi) /turnover o EBITDA/ (tot. debiti).

c) Dati comportamentali

Si utilizza questo set di dati per capire l'andamento dell'azienda; l'informazione che si tenta di analizzare riguarda il numero di contratti attivi, il numero delle insolvenze (sia attive che estinte), i pagamenti che avvengono in determinati periodi etc.

È difficile recuperare questo tipo di dati, pertanto si possono recuperare da fonti esterne quale la Centrale dei Rischi: ente nazionale che monitora l'indebitamento delle società o imprese verso banche o istituti finanziari.

La Centrale dei Rischi riceve periodicamente le informazioni da tutte le banche riguardante i crediti attivi verso la propria clientela e in cambio fornisce alle banche le informazioni sul debito totale verso il sistema creditizio di ciascun cliente segnalato.

d) Dati qualitativi

I dati di tipo qualitativo sono integrati nel modello per aumentarne le performance. Questo tipo di dati sono raccolti tramite dei questionari sottoposti ai clienti. Per strutturare al meglio un questionario efficiente è necessario includere il contributo degli esperti e definire delle linee guida per formulare delle domande chiare ed evitare di ricevere risposte ambigue.

La dimensione delle controparti incide sull'impatto e sul peso che il tipo di dati possono avere. Per esempio, controparti di grandi dimensioni presentano dei dati finanziari che sono più rilevanti rispetto ai dati personali o comportamentali; questi infatti vengono estratti da bilanci aziendali consolidati, dai quali è possibile estrarre molte informazioni. Le informazioni comportamentali, invece, sono più rilevanti per le piccole o medie imprese. Infine, per quanto riguarda le informazioni qualitative o il giudizio degli esperti sono rilevanti equamente sia per le piccole e medie imprese sia per le imprese corporate.

- Trattamento dei valori mancanti

Con il termine valori "mancanti" si intendono quei valori che non è possibile trovare nel database. Se vi sono dei valori mancanti, non è possibile assegnare il rating, pertanto il processo si blocca. È necessario, quindi, prima di iniziare la fase di processo, analizzare il

database per verificare che non vi siano valori mancanti. Le cause principali per questa anomalia possono essere: dati non applicabili o non disponibili, dati non noti o valori nulli.
Per il trattamento di questi dati vi sono diversi metodi, dipendenti dalla frequenza di accadimento:

- valori mancanti concentrati in specifiche osservazioni/indicatori e in tal caso è possibile escluderli;
- valori mancanti presenti con una frequenza bassa; in questo caso, se la frequenza è inferiore al 5% del totale è possibile non effettuare alcun intervento, altrimenti è necessario sostituire il valore mancante con il valore medio dell'indicatore.

- Weights of Evidence (WoE)

Gli indicatori presenti nel modello, come già discusso, racchiudono diversi tipi di informazione e quindi diversi tipi di dati; avere più informazioni possibili è importante per costruire un buon modello, ma tali dati devono, tuttavia, poter essere confrontabili. Per trasformare quindi, delle variabili, discrete o continue, in variabili ordinali, si utilizza la tecnica WoE. Questa serve anche ad incorporare la struttura del tasso di default nell'indicatore in questione.
Per ogni indicatore, definiti i valori possibili, se ne osserva, quanti di questi vanno in default e quanti invece sono considerati in bonis. La formula da applicare per trasformare i suddetti valori è la seguente:

$$\boldsymbol{WoE = \ln\left(\frac{bonis_j / bonis_{tot}}{default_j / default_{tot}}\right)} \qquad (10)$$

- Bonis $_j$ = bonis nella classe j;
- Bonis $_{tot}$ = bonis nella classe totale;
- Default $_j$ = default nella classe j;
- Default $_{tot}$ = default nella classe totale;

La formula esposta si applica a tutte le classi di un indicatore e il risultato che ne consegue è un numero, che incorpora il tasso di default della classe dell'indicatore. È possibile, quindi, ordinare le classi di un indicatore, in quanto il risultato ottenuto sarà maggiore per le classi con minor tasso di default, e minore per le classi con maggior tasso di default. Il calcolo, fatto per tutti gli indicatori, permette di avere tutte le variabili sotto la stessa scala di misura.

- Dalla long-list alla short-list degli indicatori

Il passaggio dalla long-list alla short list degli indicatori è un processo che può essere analizzato tramite tre componenti:

a) Accuracy Ratio (AR)

L'Accuracy Ratio è un indicatore che ha l'obiettivo di identificare l'accuratezza in un modello; nel caso della costruzione del sistema di rating Interno, l'AR misura l'abilità del modello nel discriminare le controparti sane da quelle anomale. Il valore dell'AR è compreso tra 0 e 1, in cui 0 rappresenta i peggiori e 1 i migliori. La Figura 3 - Accuracy Ratio rappresenta un'applicazione dell'Accuracy Ratio.

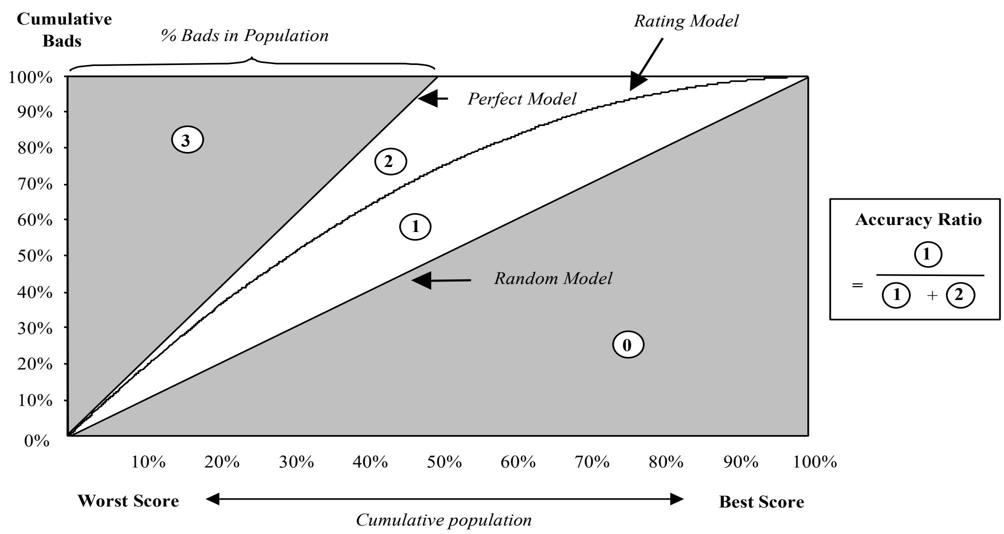


FIGURA 3 - ACCURACY RATIO[6]

Nell'asse x in figura è rappresentata la popolazione, ordinata in base al valore dei suoi fattori, dal peggiore al migliore. Nell'asse y vi è la percentuale di controparti anomali individuati per ogni percentuale del campione completo sull'asse x. La diagonale rappresenta un modello casuale che assegna a tutte le controparti la stessa probabilità di default (non discrimina); Il segmento che racchiude l'area "3" rappresenta il modello perfetto, che individua tutte le controparti anomale perfettamente. Il segmento curvo rappresenta il modello in questione: esso si colloca tra quello perfetto e il random, e tanto si avvicina al modello perfetto, tanto è migliore; l'Accuracy Ratio, che misura il potere discriminatorio, è calcolato come il

[6] Fonte: Group Risk Management, Intesa Sanpaolo S.p.A.

rapporto dell’area compresa tra la diagonale e il modello di scoring effettivo (1) e l’area tra il modello random e il modello perfetto (1+2).

b) Matrice di correlazione

La matrice di correlazione ha come obiettivo di verificare l’indipendenza di un fattore dagli altri. Si calcola la covarianza (K_{XY}) per tutte le coppie degli indicatori come segue:

$$K_{XY} = \frac{1}{n} * \sum_{i=1}^{n}(X_i - \mu_X) * (Y_i - \mu_Y) \qquad (11)$$

La varianza per ogni indicatore è calcolata:

$$\sigma^2{}_X = \frac{1}{n}\sum(X_i - \mu_X)^2 \qquad (12)$$

$$\sigma^2{}_Y = \frac{1}{n}\sum(Y_i - \mu_Y)^2 \qquad (13)$$

Due indicatori sono considerati correlati, a seconda del coefficiente di correlazione lineare che presentano, calcolato come segue:

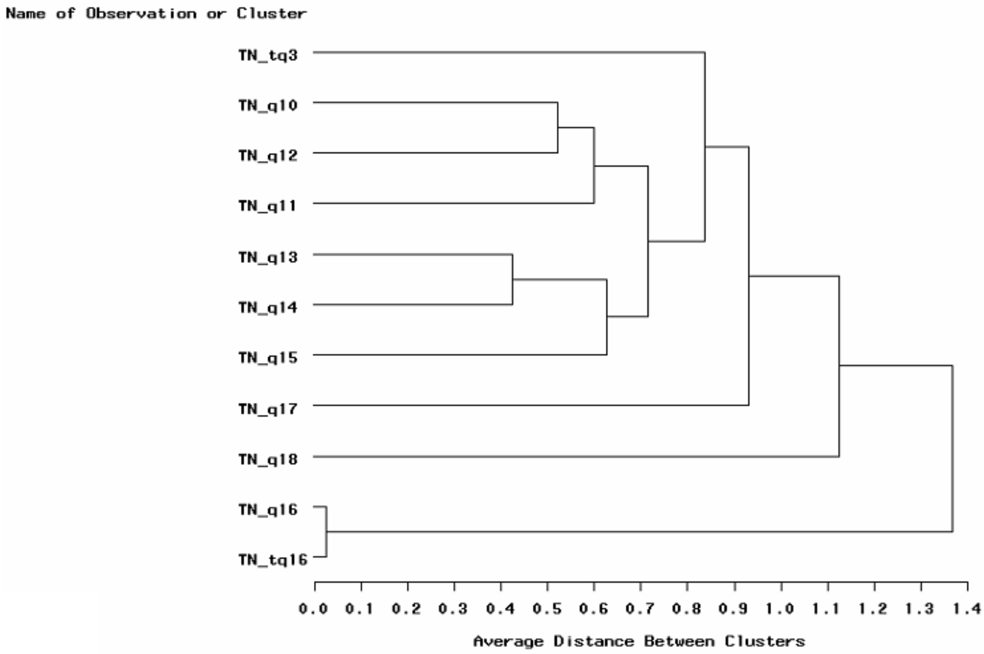
$$\rho_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y} \qquad (14)$$

ρ è un valore compreso tra -1 e +1, in cui -1, o per valori prossimi ad esso, comporta una correlazione negativa tra le variabili; viceversa se ρ è prossimo a +1. Se ρ è uguale a 0 non vi è correlazione, e le variabili sono definite indipendenti.

Una volta definito il coefficiente di correlazione, per tutte le coppie di variabili, si costruisce una matrice in cui si riportano tutti i valori. Avendo definito le soglie di correlazione (<40%, 40-70%, >70%) si decide di eliminare dal modello quelle maggiormente correlate, che presentano un Accuracy Ratio minore.

c) Cluster analisi

L’obiettivo di questa analisi è calcolare la correlazione, non solo lineare, che vi è tra i cluster. Questa procedura serve ad aggregare diversi cluster, riducendone il numero. Sono aggregati i cluster con alti valori di correlazione, come rappresentato, ad esempio, in Figura 4. Per definire la short-list, in base alla correlazione, le variabili si accorpano, eliminando quella con l'Accuracy Ratio minore. Si definisce il numero di variabili da inserire nel modello, e si procede fino a quando non lo si raggiunge. In alcuni casi, può essere difficile decidere tra determinati fattori, quindi gruppi di fattori alternativi vengono selezionati e testati nella regressione multivariata.



[7]

FIGURA 4 - RAGGRUPPAMENTO DEI CLUSTER

- Trasformazione univariata

Diversi fattori possono avere intervalli e dimensioni molto diversi. Se si includono fattori non standardizzati nell'analisi di regressione, non sarà possibile interpretare direttamente i pesi dei diversi indicatori. Assicurando che ogni fattore sia standardizzato su una media di 0 e una deviazione standard di 50, il coefficiente di regressione di ciascun fattore può essere espresso come un peso. La trasformazione dei fattori impone una relazione intuitiva tra il valore del fattore e il rischio di credito e garantisce che gli outliers non abbiano un impatto significativo sulla regressione. Si definisce la Average Default Frequency (ADF) per ogni valore dell'indicatore; si sceglie, quindi, la funzione che meglio trasforma l'indicatore in uno score.

[7] Fonte: Group Risk Management, Intesa Sanpaolo S.p.A.

- Trasformazione logistica

Si effettua una trasformazione logistica per mappare il valore tra 0 e 1:

$$\boldsymbol{valore\ ottenuto} = \frac{1}{1+e^{-score}} \qquad (15)$$

In seguito, per ottenere gli score finali, si settano la media a 0 e la deviazione standard a 50 della variabile trasformata logisticamente.

$$\boldsymbol{T(x)} = \frac{\frac{1}{1+e^{-S}}-\mu}{\sigma} * \mathbf{50} \quad (16)$$

La trasformazione assicura che tutti gli indicatori abbiano la stessa deviazione standard e che i pesi dei fattori siano significativi. Ha anche l'ulteriore vantaggio di fornire un risultato limitato in modo che gli outliers non causino risultati distorti durante la modellazione.

2.2.2.3 Analisi multivariata

Con il termine analisi multivariata si indica quell'insieme di metodi statistici usati per analizzare simultaneamente più caratteri. L'esistenza di molte variabili interagenti l'una con l'altra complica l'analisi rispetto al modello univariato. Le procedure statistiche univariate possono essere generalizzate, ma la complessità aumenta sempre più all'aumentare delle dimensioni del problema. Per effettuare un'analisi multivariata ci sono diversi metodi:

a) Regressione lineare multipla

È stimata una relazione lineare tra la variabile dipendente o target, e le variabili indipendenti. Il risultato è un valore continuo e non può essere considerato (se non trasformato) come una probabilità. La Figura 5 e la Figura 6 rappresentano la formula usata per la sima della relazione e l'output della stessa. L'intercetta rappresenta il punto di partenza della retta; i coefficienti β quanto impattano le variabili sulla variabile target; ε rappresenta il residuo.

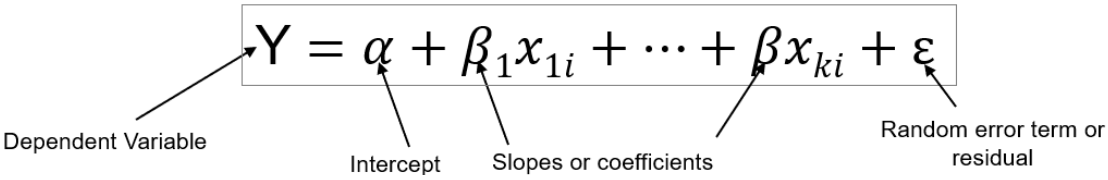


FIGURA 5 - ANALISI DI REGRESSIONE MULTIPLA[8]

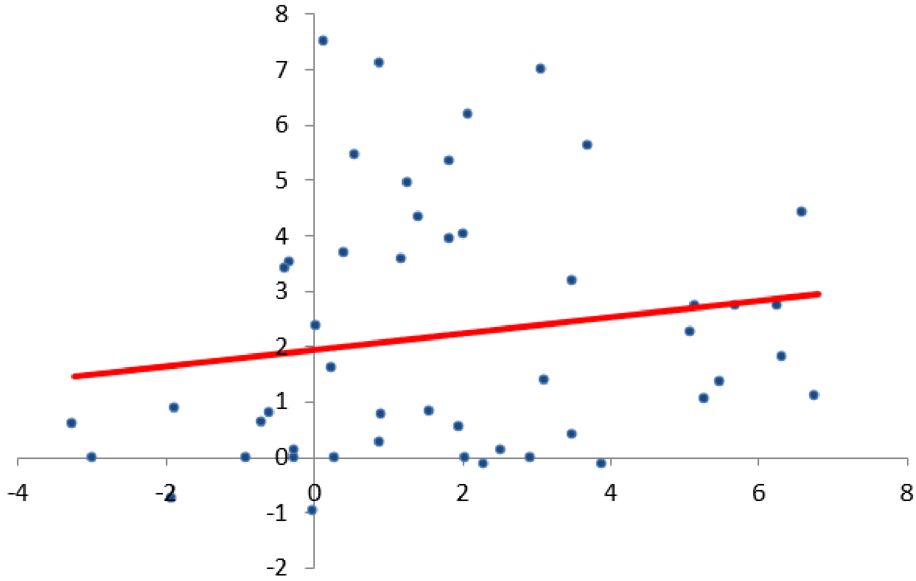


FIGURA 6 - OUTPUT ANALISI DI REGRESSIONE[9]

b) Regressione logistica

"È un metodo discriminatorio, tra due insiemi: 0 (sane) e 1 (anomale) e definisce se la variabile da stimare appartiene ad un insieme o all'altro. Si ipotizza l'esistenza di una variabile (y^*) espressiva dello "stato di salute" dell'impresa; tuttavia y^* non è osservabile, ma lo è una sua realizzazione dicotomica:

$$y = \begin{cases} 1 & se \quad y^* > 0 \\ 0 & se \quad y^* \leq 0 \end{cases} \qquad (17)$$

Indicando con p la probabilità di default si ottiene:

$$p = F(\alpha + \beta X) \qquad (18)$$

[8] Fonte: Group Risk Management, Intesa Sanpaolo S.p.A.
[9] Fonte: Group Risk Management, Intesa Sanpaolo S.p.A.

Nella quale F indica la funzione standard cumulativa logistica;

$$\boldsymbol{F(\alpha + \beta X) = \int_{-\infty}^{\alpha+\beta X} f(h)dh = \frac{1}{1+e^{-(\alpha+\beta X)}}} \quad (19)$$

con f(h) che indica la funzione di densità logistica;

$$\boldsymbol{f(h) = \frac{e^h}{(1+e^h)^2}} \quad (20)$$

Il modello stabilisce la forma della distribuzione della probabilità di default:

$$\boldsymbol{p = \frac{1}{1+e^{-(\alpha+\beta X)}}} \quad (21)$$

Si ha quindi:

$$\boldsymbol{e^{-(\alpha+\beta X)} = \frac{1-p}{p}} \quad (22)$$

Nella quale il termine a destra indica l'odd-ratio, dato dal rapporto tra la probabilità dell'evento e il suo complemento; applicando il logaritmo naturale si ottiene:

$$\boldsymbol{\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X} \quad (23)$$

La differenza tra la regressione lineare e modello logistico sta nel fatto che, nel primo è p ad essere messo in relazione con α+βX, mentre nel secondo è il logaritmo dell'odd-ratio"[10].

La Figura 7 mette in relazione il modello lineare con la regressione logistica. Si nota che i valori ottenuti con il modello lineare sconfinano l'intervallo delle probabilità (0-1) mentre i valori ottenuti con la regressione logistica approssima meglio la natura binaria della variabile target.

[10] Fonte: Niccolò Mangione: "Credit Risk Scoring Model con metodologie di data science", Politecnico di Torino, Corso di laurea magistrale in Ingegneria Gestionale, 2019

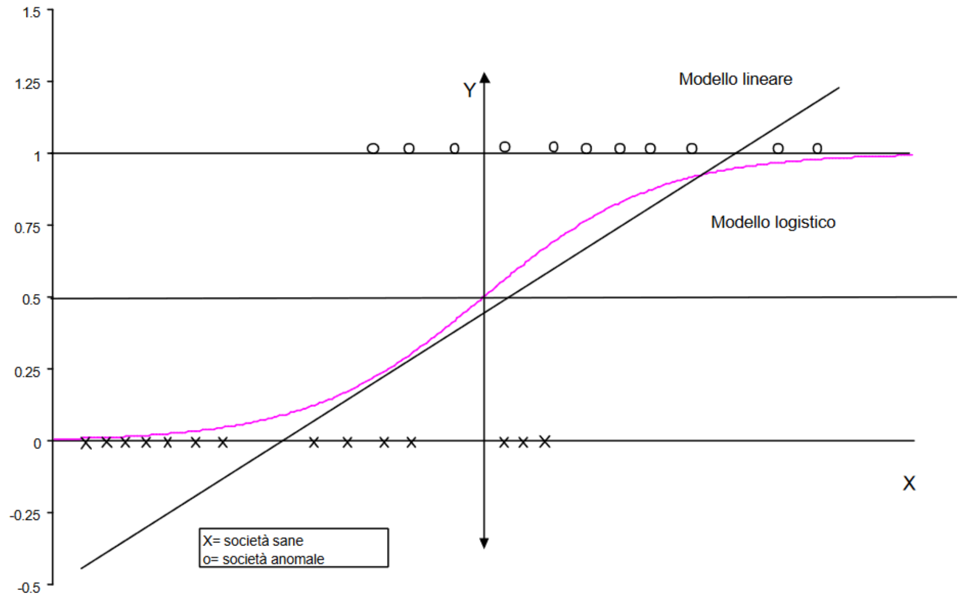


FIGURA 7 - CONFRONTO MODELLO LINEARE E MODELLO LOGISTICO[11]

Vi sono diversi metodi di selezione degli indicatori:

- Forward selection: Le variabili vengono introdotte una alla volta a partire da quella con la statistica F più alta introducendo a mano a mano quelle con il contributo più alto. Ogni variabile introdotta rimane nel modello.

- Backward selection: Tutte le variabili sono nel modello e vengono tolte a una a una in base alle statistiche F parziali che definiscono il contributo di ciascuna variabile. Ogni variabile esclusa, non rientra nel modello.

- Stepwise selection: Compromesso fra le due tecniche. Ad ogni passo viene utilizzata una selezione forward per decidere quale variabile includere e un'eliminazione backward per decidere quale eliminare.

Il metodo di selezione ottimale per il modello non è mai solo automatico, ma un compromesso tra: senso economico, senso logico e stepwise. L'aspetto principale da considerare è l'utilizzo di tutte le informazioni disponibili sia finanziarie, sia qualitative.

[11] Fonte: F. Varetto, corso di "Economia degli intermediari finanziari", A.A. 2018/2019, Politecnico di Torino

2.2.2.4 Calibrazione del modello

L'input di questa fase del modello è lo score calcolato con una delle tecniche di analisi multivariata. Lo score, adesso, è trasformato in probabilità di default per avere la misura delle controparti che vanno in default e di quelle in bonis. La probabilità di default è calibrata alla Tendenza Centrale, che corrisponde al tasso di default storico del portafoglio analizzato e che, solitamente, comprende un intero ciclo economico. L'obiettivo di calibrazione primario è quello di trasformare il punteggio di ciascun debitore in una costante probabilità di default (PD).

Il primo passo da attuare in questa fase è la scelta del campione di calibrazione; esso potrebbe essere anche diverso dal campione di training utilizzato per la stima dello score perché si basa su un arco temporale più recente che include la composizione del portafoglio reale. Si determina, in seguito, il punto di ancoraggio di Tendenza Centrale (TC), definito come tasso di default di lungo periodo. Per essere completa la TC dovrebbe: coprire un ciclo economico completo, essere coerente con la definizione predefinita di Basilea II, essere rappresentativo del portafoglio della banca (compensando l'effetto campione) ed essere uguale alle stime medie della PD.

La Tendenza Centrale rappresenta la misura del tasso di default di lungo periodo; essa è calcolata come la media semplice dei tassi di default calcolati per ogni intervallo di tempo.

$$\boldsymbol{TC} = \sum_{i=1}^{n} \frac{\boldsymbol{TD_i}}{n} \qquad (24)$$

$$\boldsymbol{TD_i} = \frac{\#\boldsymbol{default_i}}{\#\boldsymbol{performing_{i-1}}} \qquad (25)$$

TD_i rappresenta il tasso di default nel periodo i. Il periodo generalmente usato è un anno, ma se non si ha a disposizione una serie storica abbastanza lunga (minimo 5 anni) si aumenta la frequenza di monitoraggio fino ad arrivare, anche, a periodi di tre mesi.

Il tasso di default campionario è corretto per la Tendenza Centrale, che rappresenta il tasso di default di lungo periodo per ancorare il modello al ciclo economico. Per la correzione si utilizza l'aggiustamento bayesiano:

$$\boldsymbol{Adj.TD_{campione}} = \frac{TD_{campione} * \frac{TC}{TD_{popolazione}}}{TD_{campione} * \frac{TC}{TD_{popolazione}} + (1 - TD_{campione}) * \frac{1-TC}{1-TD_{popolazione}}} \qquad (26)$$

L'accuratezza della calibrazione del modello dipende da quanto la probabilità di default, predetta dal modello, si avvicini al tasso di default effettivamente realizzato. È effettuato quindi un confronto tra le probabilità di default e i tassi di default realizzati per classi di rating, settore industriale, etc.

Effettuata la correzione, si raggruppano le PD in classi, per definire una master scale: le PD saranno così assegnate a delle classi di rating con l'obiettivo di raggruppare insieme le controparti con caratteristiche simili e stessi livelli di PD. Le classi di rating devono essere coerenti con le direttive di Basilea II (7 classi per la suddivisione dei crediti in bonis e 1 classe per definire i crediti in default) e il numero delle classi deve essere tale da assicurare che non vi sia un eccesso di debitori in ogni classe. Sono definiti dei limiti di PD (upper e lower) e classificati in questo modo le PD del campione. Ogni segmento creato è confrontato con le classi di rating esterne assegnate dalle agenzie. Sul numero delle classi utilizzate, è importante tener presente che maggiore è il numero delle classi, più difficile è verificare la coerenza delle classi raggruppate; mentre, se il numero delle classi è basso, vi è il rischio che siano raggruppati controparti con caratteristiche diverse.

2.2.2.5 Questionario qualitativo e integrazione

Per la gestione delle informazioni di tipo qualitativo, generalmente, è sottoposto un questionario alle controparti in modo tale da aggiungere al modello più informazione possibile. Il suddetto questionario è suddiviso in molte sezioni: area finanziaria ed economica, business risk, settori e mercati, strategie e business plan, controllo e management, gruppo economico di appartenenza, dati dei clienti etc. Date le diverse sezioni, è possibile, quindi, ricevere diversi tipi di risposte: dicotomiche (sì o no), categoriche (minore e maggiore), in relazione alla media (superiore o inferiore alla media etc.). Per suddividere e classificare le risposte è necessario dare un peso, e quindi un'importanza, alle sezioni e alle domande, che saranno diversi a seconda della controparte: per le grandi imprese, saranno essenziali le sezioni riguardanti i settori e i mercati, mentre per le piccole e medie imprese avranno una maggiore rilevanza il controllo o il management.

Per la classificazione delle risposte si possono usare due approcci diversi. Approccio a peso e approccio a notching:

- Approccio a peso: è così definito, in quanto si assegna alle risposte uno score statistico, in linea con lo score quantitativo. È generalmente caratterizzato da domande che presentano una buona capacità discriminante e una buona distribuzione di risposte.

- Approccio a Notching: è caratterizzato da domande con un forte potere discriminante ma che prevedono una bassa distribuzione di risposte. Il punteggio, o score, assegnato, non è di tipo statistico, bensì di tipo “judgemental” a causa di insufficienza di dati sulla popolazione. (10)

Una volta ottenute le risposte e classificate, se si è utilizzato l’approccio a peso, si costruisce uno score integrato che racchiude l’informazione di tipo quantitativo e quella di tipo qualitativo:

$$\boldsymbol{score_{integrato} = \alpha + \beta_1 * score_{statistico} + \beta_2 * score_{qualitativo}} \qquad (27)$$

3 Big Data e Machine Learning: evoluzioni e prospettive per i modelli di Rating

3.1 Big Data

Con il termine dati si intende l'insieme di quantità, caratteri e simboli su cui vengono eseguite le operazioni da un computer; essi possono essere memorizzati e trasmessi sotto forma di segnali elettrici e registrati su supporti di registrazione magnetici, ottici o meccanici.

I big data sono pur sempre dati ma di dimensioni enormi. Questo termine è usato per descrivere una raccolta di dati di grandi dimensioni che presenta una crescita esponenziale nel tempo. In breve, tali dati sono così grandi e complessi che richiedono nuovi trattamenti rispetto a quelli tradizionali per essere elaborati in modo efficiente.

Per comprendere davvero i big data, è utile un excursus storico. La definizione di Gartner, diffusa intorno ai primi anni 2000, recita: "i big data sono dati che contengono una maggiore varietà, che arriva in volumi crescenti e con velocità sempre maggiore". Da qui, nascono le tre V: varietà, volume e velocità.

- Varietà: si riferisce a fonti eterogenee e alla natura dei dati, ovvero ai diversi tipi di dati disponibili;
- Volume: si riferisce alla grande quantità di informazione che si genera continuamente e che non è possibile archiviare con i sistemi tradizionali;
- Velocità: si riferisce alla rapidità di generazione dei dati. La velocità, con cui i dati sono generati ed elaborati per soddisfare le esigenze, determina il potenziale reale dei dati;

Con il passare degli anni, alla definizione si sono aggiunte altre tre V: veridicità, variabilità e valore.

- Veridicità: si riferisce all'affidabilità che i dati hanno. Questi devono, infatti, non essere distorti per poterli considerare indice di qualità;

- Variabilità: si riferisce al cambiamento del significato dei dati quando essi sono interpretati;
- Valore: si riferisce all'abilità di trasformare i dati in valore, per riuscire a trarne un vantaggio economico.

La Figura 8 rappresenta le cosiddette V dei Big Data.

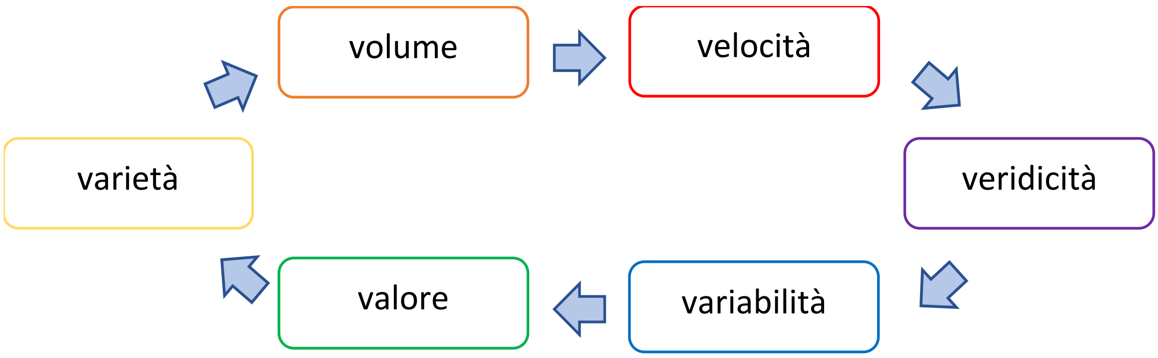


FIGURA 8 - V DEI BIG DATA

Parlando di Big Data, si deve far riferimento alla forma nella quale essi si presentano: strutturata, non strutturata e semi-strutturata. (11)

- Strutturata: tutti i dati che possono essere memorizzati, consultati ed elaborati sotto forma di formato fisso sono definiti come dati strutturati. Nel corso del tempo, il talento nell'informatica ha raggiunto un successo maggiore nello sviluppo di tecniche per lavorare con questo tipo di dati (in cui il formato è ben noto in anticipo). Un esempio di dati strutturati, possono essere le tabelle di un database.
- Non strutturata: tutti i dati con forma, o struttura, sconosciuta sono classificati come dati non strutturati. Oltre alle dimensioni enormi, i dati non strutturati presentano molteplici sfide in termini di elaborazione per ricavarne valore. Un tipico esempio di dati non strutturati è una fonte di dati eterogenea contenente una combinazione di semplici file di testo, immagini, video, etc.
- Semi-strutturata: dati semi-strutturati possono contenere entrambe le forme di dati. Si possono considerare come strutturati nella forma ma in realtà non sono definiti. Un esempio di dati semi-strutturati è un dato rappresentato in un file XML.

La Figura 9 raffigura il tipo di dati e la loro evoluzione nel tempo.

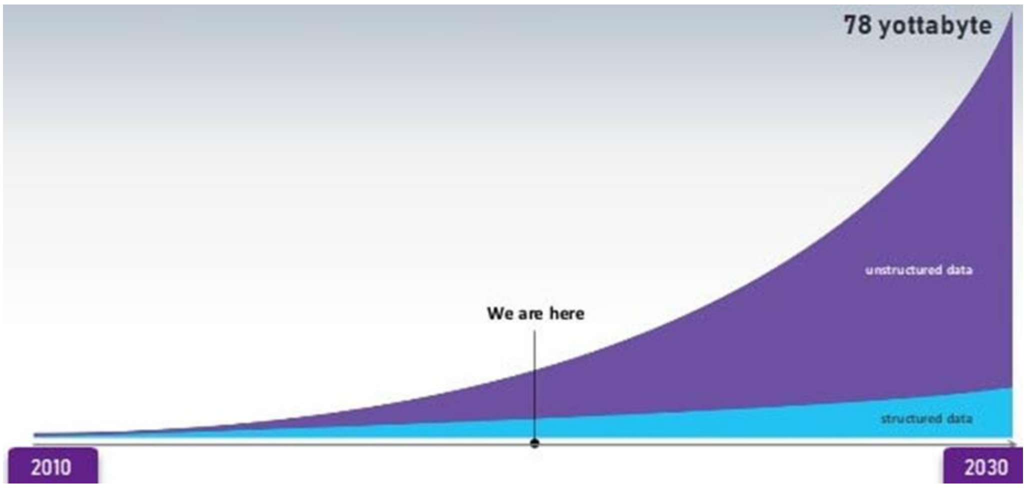


FIGURA 9 - TIPO DI DATI ED EVOLUZIONE NEL TEMPO[12]

Sebbene il concetto stesso di big data sia relativamente nuovo, le origini di grandi insiemi di dati risalgono agli anni '60 e '70, quando il mondo dei dati stava appena iniziando con i primi data center e con lo sviluppo dei database relazionali. In seguito, con l'avvento di Internet, dei social network e altri servizi online, ci si è resi conto dell'ammontare di dati che si genera ogni giorno. Da allora il volume dei big data è salito in modo esponenziale, non solo grazie all'attività umana, ma anche la nascita dell'Internet of Things (IoT), che collega sempre più oggetti e dispositivi, ha contribuito a generare una grande quantità di dati.

I big data hanno il potenziale di fornire alle aziende preziose informazioni sui loro clienti che possono essere utilizzate in svariati settori. Le aziende che utilizzano i big data detengono un vantaggio competitivo (che andrà ad aumentare negli anni), rispetto a coloro che ignorano i dati, poiché hanno la capacità di prendere decisioni aziendali più rapide e informate.

I settori in cui si possono utilizzare i Big Data sono:

- Sviluppo prodotto: aziende come Netflix o Procter & Gamble utilizzano i big data per anticipare la domanda dei clienti. Costruiscono modelli predittivi per nuovi prodotti o servizi, classificando gli attributi chiave e modellando la relazione tra tali attributi e il successo commerciale delle offerte.

[12] Web: https://www.guru99.com/what-is-big-data.html

- Manutenzione predittiva: fattori che possono prevedere guasti meccanici possono essere inseriti nei dati strutturati, come l'anno, la fabbricazione e il modello delle attrezzature, nonché nei dati non strutturati che coprono milioni di voci di registro, quali dati dei sensori e messaggi di errore. Analizzando queste indicazioni di potenziali problemi prima che si verifichino, le imprese possono implementare la manutenzione in modo più economico e massimizzare il tempo di attività di parti e apparecchiature.
- Frode e conformità: scenari di sicurezza e requisiti di conformità sono in continua evoluzione. I big data aiutano a identificare modelli nei dati che indicano frodi e aggregano grandi volumi di informazioni per rendere i report normativi molto più completi.

Questi sono solo alcuni dei settori nei quali le aziende stanno usando maggiormente i big data, tuttavia questi trovano una maggiore applicabilità nell'ambito del "Machine Learning".

Il termine può essere definito come un sistema di elaborazione automatizzata dei dati e degli algoritmi decisionali, progettato per migliorarne il funzionamento in base ai risultati del proprio lavoro.

Nell'ambito dei Big Data, il machine learning è utilizzato per tenere il passo con il flusso di dati in continua crescita ed evoluzione e fornire approfondimenti in continua evoluzione. (12)

Di solito, gli algoritmi di machine learning sono utilizzati per classificare i dati in entrata e riconoscere i modelli in essi contenuti; possono, successivamente, anche essere tradotti in preziose informazioni ed implementati nell'attività aziendale. La Figura 10 presenta le fasi del processamento dei dati.

FIGURA 10 - FASI DEL PROCESSAMENTO DEI DATI

3.2 Machine Learning

Il Machine Learning (o apprendimento automatico) è un'applicazione dell'intelligenza artificiale (AI) che fornisce ai sistemi la capacità di apprendere e migliorare automaticamente dall'esperienza, senza essere programmata esplicitamente. L'apprendimento automatico si concentra sullo sviluppo di programmi che possano accedere ai dati e utilizzarli per poter apprendere da soli.

Il processo di apprendimento inizia con le osservazioni o dati, quali possono essere esempi di esperienza diretta o istruzione, con l'obiettivo di cercare modelli nei dati e prendere decisioni migliori in futuro, sulla base degli esempi che si forniscono. Lo scopo principale è consentire ai computer di apprendere automaticamente senza intervento o assistenza umana e regolare le azioni di conseguenza.

Gli algoritmi di machine learning sono spesso classificati come supervisionati o non supervisionati.

Gli algoritmi di machine learning supervisionato possono applicare ciò che è stato appreso in passato a nuovi dati, utilizzando esempi etichettati per prevedere eventi futuri. A partire dall'analisi di un set di dati di training noto, l'algoritmo di apprendimento produce una funzione dedotta per fare previsioni sui valori di output. Il sistema è in grado di fornire output per qualsiasi nuovo input a seguito di una formazione sufficientemente adeguata. L'algoritmo di apprendimento può anche confrontare il suo output con l'output corretto e previsto e trovare errori per modificare di conseguenza il modello.

Al contrario, gli algoritmi di machine learning senza supervisione sono utilizzati quando le informazioni usate per la formazione non sono né classificate né etichettate. L'apprendimento senza supervisione studia come i sistemi possano dedurre una funzione per descrivere una struttura nascosta da dati senza etichetta. Il sistema non riesce a trovare l'output giusto, ma esplora i dati e può trarre inferenze tra questi per descrivere strutture nascoste da dati senza etichetta.

Gli algoritmi di machine learning semi-supervisionati rientrano tra l'apprendimento supervisionato e quello non supervisionato, poiché utilizzano sia i dati etichettati sia quelli senza etichetta; in genere usano una piccola quantità di dati etichettati e una grande quantità di dati non etichettati. I sistemi che utilizzano questo metodo sono in grado di migliorare considerevolmente l'accuratezza dell'apprendimento. Di solito, l'apprendimento semi-supervisionato è da preferire quando i dati etichettati acquisiti richiedono risorse qualificate e pertinenti al fine di addestrarli/apprendere da essi. In caso contrario, l'acquisizione di dati senza etichetta in genere non richiede risorse aggiuntive.

Il rafforzamento degli algoritmi di machine learning è un metodo di apprendimento che interagisce con il proprio ambiente producendo azioni, errori e ricompense. La ricerca di prove ed errori e la ricompensa ritardata sono le caratteristiche più rilevanti per il rafforzamento dell'apprendimento. Questo metodo consente alle macchine di determinare automaticamente il comportamento ideale in un contesto specifico al fine di massimizzarne le prestazioni. È richiesto un semplice feedback sulla ricompensa affinché la macchina apprenda quale azione sia la migliore; questa tecnica è definita segnale di rinforzo.

L'apprendimento automatico consente l'analisi di enormi quantità di dati. Sebbene in genere fornisca risultati più rapidi e precisi al fine di identificare opportunità vantaggiose o rischi pericolosi, può anche richiedere tempo e risorse aggiuntive per essere addestrato correttamente. La combinazione dell'apprendimento automatico con l'intelligenza artificiale e le tecnologie cognitive può renderlo ancora più efficace nell'elaborazione di grandi volumi di informazioni.

Il machine learning è correlato al data mining, il processo di scoperta di modelli in grandi set di dati. Entrambi i metodi, spesso, utilizzano gli stessi procedimenti ma, mentre il machine learning si concentra sulla previsione, sulla base delle proprietà note apprese dai dati di addestramento, il data mining si concentra sulla scoperta di proprietà sconosciute nei dati. Il Data mining utilizza molti metodi di machine learning, ma con obiettivi diversi; dall'altro lato, l'apprendimento automatico utilizza anche metodi di data mining come apprendimento non supervisionato o come fase di preelaborazione per migliorarne l'accuratezza di apprendimento[13].

[13] Web: http://www.intelligenzaartificiale.it/machine-learning/

3.2.1 Algoritmi di Machine Learning

Gli algoritmi di machine learning sono programmi (matematica e logica) che si adattano per funzionare meglio quando sono esposti a più dati. La parte "apprendimento" del machine learning si riferisce al modo con cui i programmi cambiano il modo di elaborare i dati nel tempo. Quindi un algoritmo di apprendimento automatico è un programma con un modo specifico di adattare i propri parametri, dato il feedback sulle sue prestazioni precedenti facendo previsioni su un set di dati.

3.2.1.1 Alberi decisionali

Un albero decisionale (o decision tree) è un insieme di nodi e archi; un grafico direzionale che inizia alla base con un singolo nodo e si estende ai numerosi "nodi foglia" rappresentanti le categorie che l'albero può classificare. Un albero decisionale può essere paragonato come un diagramma di flusso, nel quale il flusso inizia nel nodo radice e termina con una decisione presa al livello dei nodi foglie. È considerato uno strumento di supporto alle decisioni, utilizzante un grafico ad albero per presentare le previsioni risultanti da una serie di suddivisioni basate sulle diverse funzionalità. La Figura 11 rappresenta un esempio di albero decisionale, su tre livelli.

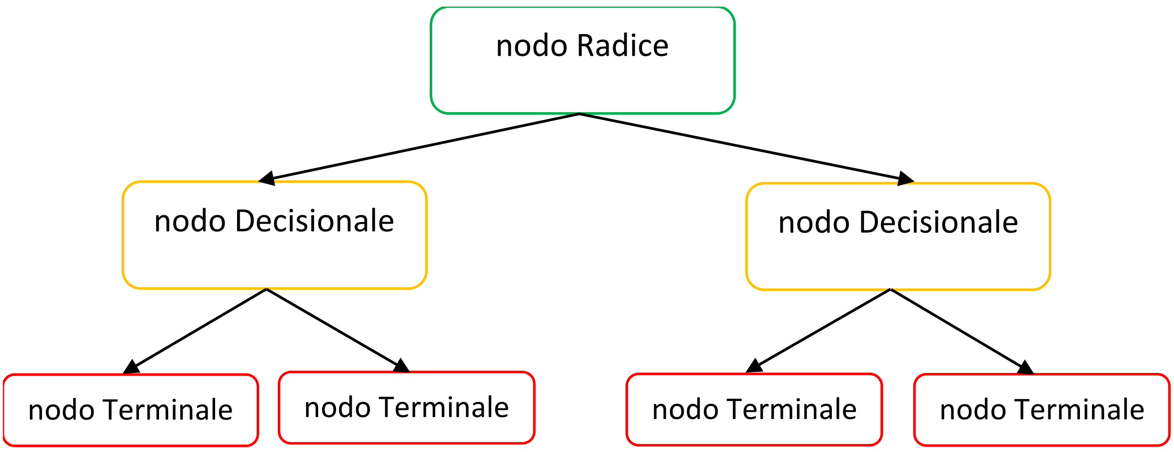


FIGURA 11 - ALBERO DECISIONALE

Vi sono alcuni termini chiave e caratterizzanti gli alberi decisionali:

- Nodo radice: un nodo radice è il punto d'inizio di un albero decisionale. Rappresenta l'intera popolazione analizzata. Dal nodo radice, la popolazione è divisa in base a varie funzionalità, in sottogruppi. Questi, a loro volta, sono suddivisi in ciascun nodo decisionale sotto il nodo radice.

- Suddivisione: è il processo di divisione di un nodo in due o più sotto-nodi.
- Nodo decisionale: è un nodo (secondario) che si suddivide in ulteriori nodi.
- Nodo foglia o Nodo terminale: sono i nodi che non si dividono, ovvero quelli situati ai punti finali dell'albero.
- Taglio: la rimozione dei sotto-nodi. Un albero è ampliato attraverso la scissione e ridotto attraverso il taglio.
- Branch o sottoalbero: rappresenta una sottosezione dell'albero decisionale chiamata, appunto, diramazione o sottoalbero; proprio come una parte di un grafico è chiamata sotto-grafico.
- Nodo padre e nodo figlio: si tratta di termini relativi. Qualsiasi nodo che rientra in un altro nodo è un nodo figlio o sotto-nodo e qualsiasi nodo che precede quei nodi figlio viene chiamato nodo principale

Gli alberi decisionali sono un algoritmo molto popolare per diversi motivi:

- Potere esplicativo: l'output degli alberi decisionali è facilmente interpretabile. Può essere compreso da persone senza background analitici o matematici. Non richiede nemmeno alcuna conoscenza statistica per essere compreso.
- Analisi dei dati esplorativi: i decision tree possono consentire agli analisti di identificare variabili significative e relazioni importanti tra due o più variabili, contribuendo a far emergere il segnale contenuto da molte variabili di input.
- Pulizia minima dei dati: gli alberi decisionali, resistenti ai valori anomali e ai valori mancanti, richiedono meno pulizia dei dati rispetto ad altri algoritmi più complessi.
- Qualsiasi tipo di dati: gli alberi decisionali possono effettuare classificazioni basate su variabili sia numeriche sia categoriche.

Tuttavia, presentano alcuni svantaggi:

- Overfitting: fenomeno che corrisponde a un eccesso di adattamento, ed è un difetto comune degli alberi decisionali. Si verifica, in generale, quando il modello si adatta ai dati osservati a causa del numero eccessivo di parametri rispetto al numero delle osservazioni, e ciò porta l'algoritmo a modellare "troppo bene" i dati in sample. Per ovviare a questo problema, si possono impostare dei vincoli sui parametri del modello (limitazione della profondità) e la semplificazione del modello mediante il taglio; ciò contribuisce a migliorare la capacità di un albero decisionale di generalizzare sul set di test (out of sample).

- Previsione di variabili continue: sebbene gli alberi decisionali possano importare input numerici continui, non sono un modo pratico per prevedere tali valori, poiché le previsioni dell'albero decisionale devono essere separate in categorie discrete, il che comporta una perdita di informazioni quando si applica il modello a valori continui.
- Ingegnerizzazione di funzioni pesanti: il rovescio della medaglia del potere esplicativo dell'albero decisionale è che richiede l'ingegnerizzazione di funzioni pesanti. Quando si trattano dati non strutturati o dati con fattori latenti, ciò rende gli alberi decisionali non ottimali.

Vi è un trade-off, quindi, sulla semplicità concettuale degli alberi decisionali rispetto alla loro implementazione[14].

3.2.1.2 Random Forest

L'algoritmo Random Forest (o foresta casuale) è una tecnica che prevede l'uso di molti alberi decisionali. Ogni albero decisionale, presente nell'insieme, è creato utilizzando un sotto-insieme degli attributi utilizzati per classificare la popolazione: sono creati, così, degli alberi decisionali in modo casuale. Ogni albero genera delle previsioni di classe e la classe con il maggior numero di voti diviene la previsione dell'intero modello. La Figura 12 rappresenta un esempio di Random forest.

[14] Web: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

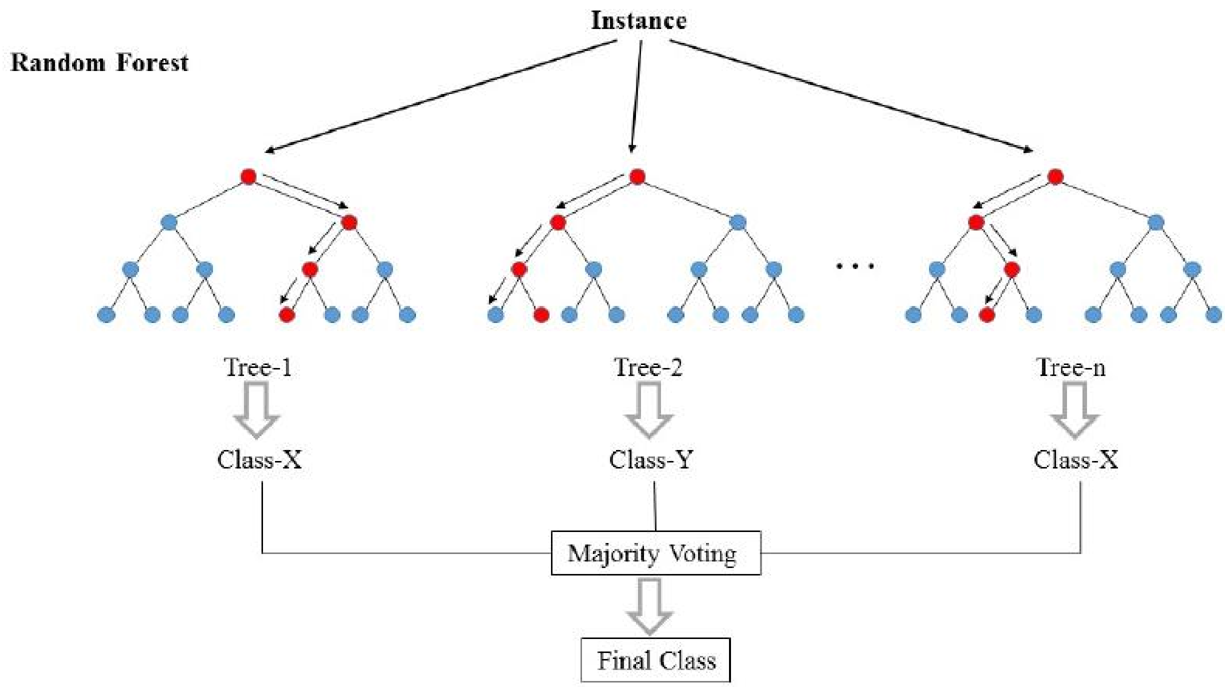


FIGURA 12 - RANDOM FOREST[15]

L'algoritmo è un metodo di classificazione supervisionato. Crea una foresta (molti alberi decisionali), ordina i loro nodi e si divide casualmente. Più alberi sono presenti nella foresta e migliori sono i risultati che può produrre. La questione fondamentale è che vi sia una bassa correlazione tra i modelli inseriti; in tal modo, ogni albero decisionale creato, prende delle decisioni in modo a sé stante. Sono così ridotti anche gli errori individuali, in quanto si crea una sorta di "protezione" tra un albero e l'altro: alcuni alberi potranno riportare dei risultati sbagliati, ma altri che riporteranno dei risultati corretti saranno in grado di condurre il modello nella direzione giusta.

Per evitare la correlazione tra il comportamento di un albero con gli altri, si può ricorrere a due metodi:

- Bootstrap Aggregation (insaccamento): Il metodo si basa sulla sostituzione parziale del dataset. Gli alberi decisionali sono sensibili ai dati sui quali vengono formati; piccole modifiche al set di dati possono comportare strutture di alberi significativamente diverse. Il random forest ne trae vantaggio permettendo a ciascun singolo albero di campionare casualmente dal set di dati con la sostituzione, ottenendo alberi diversi.

[15] Web: https://www.researchgate.net/figure/Classification-process-based-on-the-Random-Forest-algorithm-2_fig1_324517994

Con tale metodo non si suddivide il dataset, facendolo analizzare in sotto-blocchi, bensì si alimenta il modello con un altro dataset di numerosità uguale all'originale, ma sostituendo alcuni dati.

- Caratteristica casuale: in un albero decisionale, al momento di suddividere un nodo, si considera ogni caratteristica possibile e si sceglie quella che produce la massima separazione tra le osservazioni nel nodo sinistro rispetto a quelle nel nodo destro. Al contrario, ogni albero in una foresta casuale può scegliere solo da un sottoinsieme casuale di caratteristiche. Ciò impone una variazione ancora maggiore tra gli alberi nel modello e ciò si traduce in una minore correlazione tra gli alberi e una maggiore diversificazione.

Il metodo Random Forest, quindi, funziona meglio rispetto agli alberi decisionali, ma presenta alcuni svantaggi:

- La precisione dei risultati non aumenta sempre: il modello, con l'aumentare dei campioni disponibili, migliora le proprie prestazioni, tuttavia questo miglioramento presenta un andamento logaritmico, ovvero, aumenta a tassi decrescenti.
- Complessità d'interpretazione: il modello, con struttura più complessa rispetto agli alberi decisionali, presenta una maggiore difficoltà interpretativa per capire come il modello evolve.
- Richiede l'uso di numerosi campioni e un numero elevato di tentativi per migliorare il proprio apprendimento[16].

3.2.1.3 Extreme Gradient Boosting (XGBoost)

Nella modellazione basata sui dati, l'approccio più frequente è la creazione di un solo modello predittivo robusto. Un approccio diverso potrebbe essere quello di creare un modello strong (robusto), a partire da un insieme di modelli weak (deboli). L'obiettivo, in questi casi, è quello di avere una visione d'insieme forte, grazie alla combinazione di tanti modelli deboli.

L'algoritmo fa parte della famiglia di modelli appartenenti al Gradient Boosting Machines, o semplicemente GBM, che prevede una procedura di apprendimento basata sull'aggiunta

[16] Web: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

di nuovi modelli che portano ad avere una stima sempre più accurata della variabile di risposta.

Insieme al modello Random Forest, l'algoritmo può essere considerato come un insieme di alberi decisionali; la differenza tra i due modelli sta nel fatto che il metodo Random Forest modella gli alberi decisionali in modo casuale, mentre gli algoritmi GBM prevedono l'aggiunta sequenziale degli alberi decisionali. Questi lavorano, quindi, in modo iterativo e, grazie a queste ripetizioni, si raggiungono risultati migliori. Il modello, comprendente i precedenti, non è da intendersi come migliorativo di questi ultimi (altrimenti si potrebbe usare solo il nuovo modello), bensì aggiunge caratteristiche in più al modello che apportano maggiori informazioni e permettono di ottenere risultati migliori. Un modello di Gradient Boosting è presentato in Figura 13, indicante un esempio di discriminazione di alcuni punti rispetto agli altri[17].

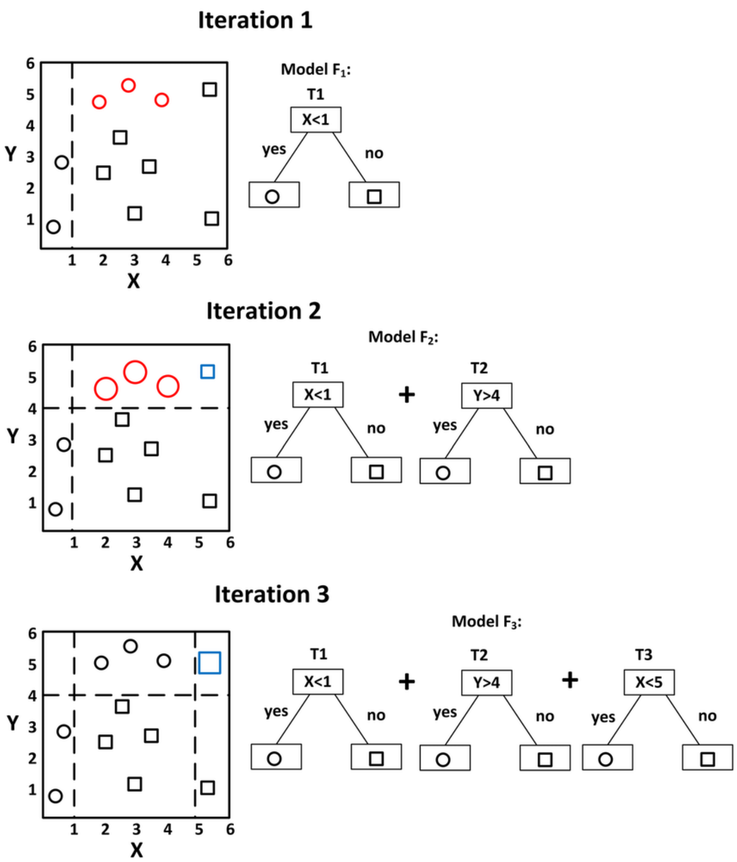


FIGURA 13 - GRADIENT BOOSTING[18]

[17] Web: https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
[18] Web: https://www.researchgate.net/figure/A-simple-example-of-visualizing-gradient-boosting_fig5_326379229

Il modello Gradient Boosting, in generale, è costruito inserendo ad ogni iterazione un modello debole, che aggiunto ai precedenti, definisce un modello più robusto. Le performance generali del modello sono migliorate ad ogni iterazione, in quanto i risultati nuovi partono dai risultati ottenuti agli step precedenti. Il principio di funzionamento del modello si basa sulla minimizzazione del Mean Squared Error (MSE) rappresentante la differenza quadratica tra il valore predetto e il valore osservato. Il modello minimizza l'MSE, inserendo ad ogni iterazione una variabile che massimizza la discriminazione della popolazione.

In particolare, il modello gradient boosting può essere riassunto nei seguenti step:

- è calcolata una prima stima della popolazione pari al tasso medio di default;
- dal valore del tasso medio della popolazione sono calcolati i residui pari alla probabilità osservata meno la probabilità stimata (media);
- è costruito un albero decisionale considerato i residui come variabile target per le iterazioni;
- i residui delle foglie finali sono trasformati in LogOdds secondo la formula:

$$\boldsymbol{LogOdds} = \frac{\sum_{j=1}^{J} \boldsymbol{Residuo}_{\boldsymbol{j}}}{\sum_{j=1}^{J} [\boldsymbol{Probabilit\grave{A}}_{y_i-1,\ j} * (1 - \boldsymbol{Probabilit\grave{A}}_{y_i-1,\ j})]} \qquad (28)$$

  In cui j è il numero di osservazioni presenti nel nodo e la probabilità al denominatore è la probabilità stimata per ogni osservazione fino all'albero precedente. Per la prima iterazione è pari alla probabilità totale per tutte le osservazioni;

- è calcolato il LogOdds Totale come segue:

$$\boldsymbol{LogOdds}_{i,\ T} = \boldsymbol{LogOdds}_0 + \sum_{t=1}^{T} \boldsymbol{LR}_t \boldsymbol{LogOdds}_{i,\ t} \quad (29)$$

  - T è il numero complessivo di alberi decisionali nel modello sino a questa fase dell'algoritmo;
  - $LogOdds_0$ sono i logodds di partenza, ovvero la trasformazione della probabilità totale
  - $LogOdds_{i,\ t}$ è l'output di ogni singolo albero decisionale t per l'i-esima osservazione

   - LR_t è il Learning Rate dell'albero t, ovvero un parametro di scala specifico di ogni albero che permette di ridurre l'overfitting del modello

- I LogOdds totali vengono trasformati in probabilità attraverso la seguente formula:

$$\boldsymbol{P} = \frac{\boldsymbol{e}^{\boldsymbol{LogOdds}}}{1 + \boldsymbol{e}^{\boldsymbol{LogOdds}}} \qquad (30)$$

- Si ripetono iterativamente i passi fino a quando è stato costruito il numero di alberi richiesti.

L'algoritmo Extreme Gradient Boosting (XGBoost), così come i metodi GBM, si fonda sul principio base di miglioramento continuo tramite l'aggiunta di modelli weak, tuttavia, l'XGBoost migliora il framework GBM attraverso l'ottimizzazione dei sistemi e miglioramenti algoritmici.

- Ottimizzazione del sistema
  a) Parallelizzazione: l'algoritmo affronta il processo di costruzione sequenziale di alberi usando un'implementazione parallela. Vi sono due loop che lavorano in parallelo: il primo enumera i nodi foglia di un albero e il secondo ne calcola le caratteristiche.
  b) Ottimizzazione hardware: l'algoritmo è stato progettato per un uso efficiente delle risorse hardware. La cache alloca i buffer interni in ciascun thread per memorizzare le statistiche del gradiente. Ulteriori miglioramenti come l'elaborazione "out-of-core" ottimizzano lo spazio disponibile su disco gestendo frame di dati di grandi dimensioni che non si adattano alla memoria.
  c) Taglio degli alberi: l'algoritmo prevede il taglio degli alberi superflui "all'indietro", andando così a ridurre l'onerosità computazionale, senza peggiorare le prestazioni del modello.

- Miglioramenti algoritmici
  a) Regolarizzazione: penalizza i modelli più complessi attraverso la regolarizzazione di LASSO e Ridge per evitare un eccesso di adattamento.
  b) Gestione dei valori mancanti: il modello riesce ad apprendere dell'esistenza dei valori mancanti e a modellarli efficientemente.

c) Convalida incrociata: l’algoritmo, ad ogni iterazione, effettua una validazione dei risultati, per verificarne l’accettabilità[19].

3.2.1.4 Reti Neurali

Una rete neurale è una rete, o circuito, composta da neuroni o nodi artificiali. Tali nodi sono considerati delle unità elementari e operano ricevendo uno o più input, che sommano per produrre un output. Un nodo combina input dai dati con una serie di coefficienti, o pesi, che amplificano o smorzano quell'input, assegnando così significato agli input in relazione al compito che l'algoritmo sta cercando di apprendere. Questi prodotti di peso in ingresso sono sommati e la somma è passata attraverso la cosiddetta funzione di attivazione di un nodo, per determinare se, e in che misura, tale segnale debba progredire ulteriormente attraverso la rete per l’influenza del risultato finale. Se i segnali passano, il neurone è definito "attivato".

I nodi sono strutturati in layer (livello); i neuroni, ad ogni livello, si accendono o si spengono in base all’input ricevuto. L'output dei nodi ad ogni layer è contemporaneamente l'input dei nodi al layer successivo. Le connessioni sono modellate come pesi. Un peso positivo riflette una connessione in bonis, mentre valori negativi indicano connessioni inibitorie: questa attività è definita come una combinazione lineare. Infine, all’ultimo livello, una funzione di attivazione controlla l'ampiezza dell'output. Ad esempio, un intervallo accettabile di output è generalmente compreso tra 0 e 1, oppure potrebbe essere −1 e 1.

Le reti neurali possono essere utilizzate per la modellazione di problemi di intelligenza artificiale. L'autoapprendimento derivante dall'esperienza avviene all'interno delle reti, che può trarre conclusioni da un insieme complesso e apparentemente non correlato di informazioni. In una rete neurale, cambiare il peso di una qualsiasi connessione, ha un effetto su tutti gli altri neuroni nei livelli successivi. La Figura 14 raffigura un esempio di rete neurale. (13)

[19] Web: https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5

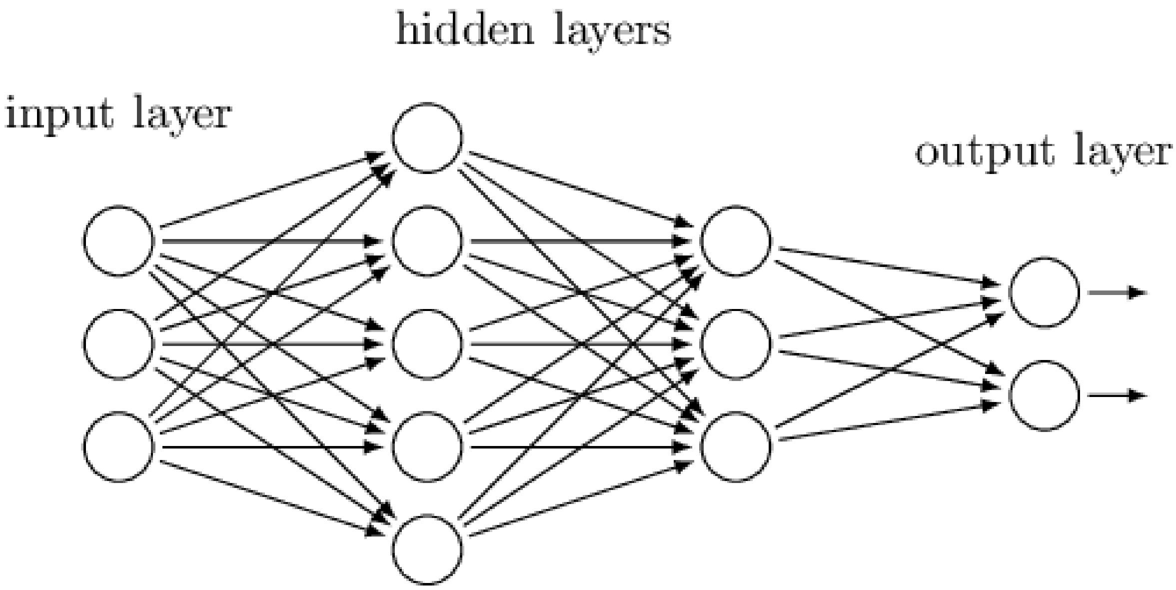


FIGURA 14 - RETE NEURALE[20]

3.3 Applicazione algoritmi di machine learning ai modelli di Rating Interni

Negli ultimi anni, grazie all'avvento dei Big Data, si ha avuto una grande applicazione degli algoritmi di machine learning in diversi settori: la sanità, il marketing, l'agricoltura, la finanza, etc. Nell'ambito dei servizi finanziari, importanti segmenti di applicazione hanno incluso il rischio di credito e l'individuazione di riciclaggio di denaro e frodi.

Studi effettuati dall'Institute of International Finance (IIF), che ha analizzato le applicazioni del machine learning relative al rischio di credito da parte degli istituti finanziari, attraverso vari sondaggi e documenti di ricerca, hanno messo in evidenza cinque argomenti generali quali: stato di maturità, area di applicazione, benefici, sfide e impegno regolamentare. (14)

L'adozione del machine learning nella modellizzazione e nella gestione del rischio di credito è aumentata in modo significativo, e con essa anche l'ampiezza dell'applicazione tra i segmenti di clientela ha registrato progressi significativi. L'adozione di queste tecniche offre

[20] Web: http://www.ce.unipr.it/people/medici/geometry/node107.html

numerosi vantaggi, tra cui una migliore accuratezza del modello, il superamento di carenze e incoerenze dei dati e la scoperta di nuovi segmenti o modelli di rischio.

La scelta della tecnologia presenta anche nuove sfide, in particolare quelle incentrate sulla comprensione da parte dei supervisori o del consenso all'utilizzo di nuovi processi.

3.3.1 Livello di maturità

In termini di livelli di maturità, si riscontra un aumento significativo del numero di aziende che utilizzano modelli di machine learning in produzione o progetti attivi.

Si evince che l'adozione non è esclusiva per le economie sviluppate o per le grandi imprese, piuttosto, la maturità continua ad essere allineata alla strategia aziendale e al processo di innovazione.

L'uso di machine learning per il rischio di credito è aumentato nettamente in tutte le aree geografiche nell'ultimo anno. Un caso particolarmente interessante è quello del Giappone che nell'ultimo anno ha visto un drastico aumento del numero di progetti di prova.

Come si vede in Figura 15, raffigurante il livello di maturità del ML, le imprese sono state suddivise per il loro totale degli asset e si notano le differenze ai vari livelli di maturità: il 42% degli strumenti finanziari utilizza già tecniche di machine learning in produzione, con un ulteriore 45% in progetti di prova e il 10% prevede di iniziare a utilizzarlo nei prossimi 6-12 mesi. Solo il 3% non ha in programma di adottare il machine learning nella funzione di rischio di credito nel prossimo futuro.

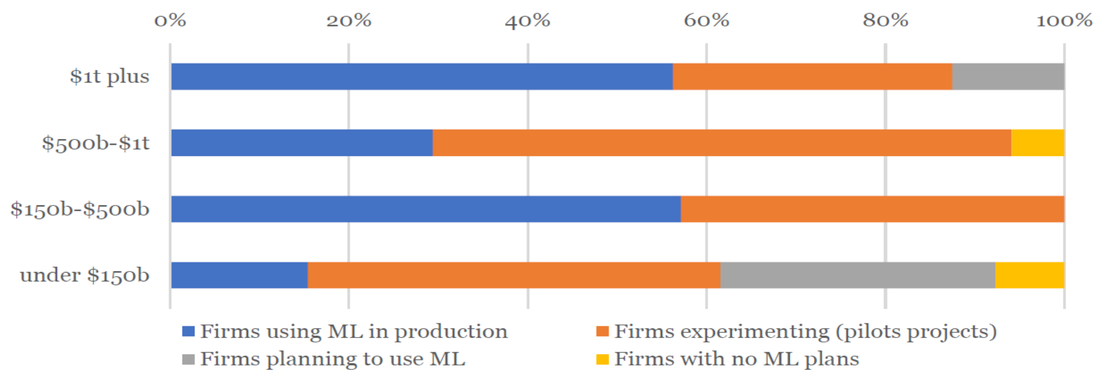


FIGURA 15 - LIVELLO MATURITÀ MACHINE LEARNING[21]

[21] Fonte: Institute of International Finance (IIF). Machine Learning in credit risk report. 2019

3.3.2 Applicazione al rischio di credito

L'applicazione principale delle tecniche di machine learning, che si hanno nell'ambito del rischio di credito, riguardano i metodi di scoring e la valutazione del merito creditizio delle controparti. Gli intermediari finanziari si sono allontanati dall'uso delle tecniche di ML per le aree normative come capitale, stress test e provisioning, focalizzando l'applicazione in aree come il monitoraggio del credito, la ristrutturazione e il recupero. Molte banche hanno specificato che i requisiti normativi esistenti non sempre si allineano con l'applicazione diretta del ML poiché i modelli regolamentari devono essere semplici, mentre i modelli ML possono essere più difficili da interpretare e spiegare. La Figura 16 presenta l'applicazione dei modelli di ML nelle aree aziendali.

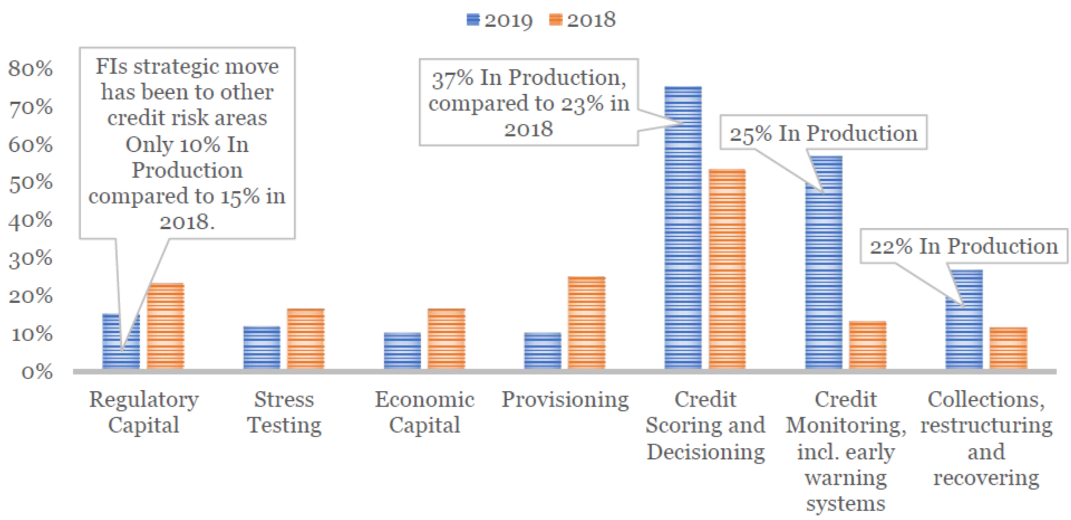


FIGURA 16 - APPLICAZIONE ML NELLE AREE AZIENDALI[22]

Gli utilizzatori dei modelli di machine learning si possono raggruppare in due categorie principali: coloro che li utilizzano su più segmenti di mercato e di prodotto e coloro che li utilizzano per segmenti ristretti. Inoltre, un'altra suddivisione è effettuata sulla sofisticazione e complessità del modello utilizzato: coloro che utilizzano il ML solo per una funzione nel processo di sviluppo (machine learning di livello inferiore) e coloro che lo usano per lo sviluppo dell'intero processo (machine learning di livello completo). In generale, quelli che utilizzano il ML su più segmenti, presentano un livello di maturità nell'uso degli algoritmi e una conoscenza maggiore, frutto di maggiori anni di applicazione, mentre quelli che li utilizzano in determinate fasi, sono ad un livello di maturità inferiore.

[22] Fonte: Institute of International Finance (IIF). Machine Learning in credit risk report. 2019

Vi sono inoltre istituti finanziari che applicano le tecniche di machine learning in diversi processi riguardanti il rischio di credito quali la segmentazione dei dati, lo sviluppo del modello, ma anche la pulizia dei dati e la fase di validazione del modello.

Si riscontra anche che gli istituti di credito, che presentano una corretta suddivisione e diversificazione del portafoglio crediti, presenta una migliore performance riguardo l'utilizzo dei modelli machine learning esistenti, dato il fatto che possiedono volumi consistenti di dati standardizzati e di alta qualità. La Figura 17 rappresenta i progressi effettuati dagli intermediari finanziari nell'applicazione degli algoritmi, in base alla suddivisione dei segmenti di clientela.

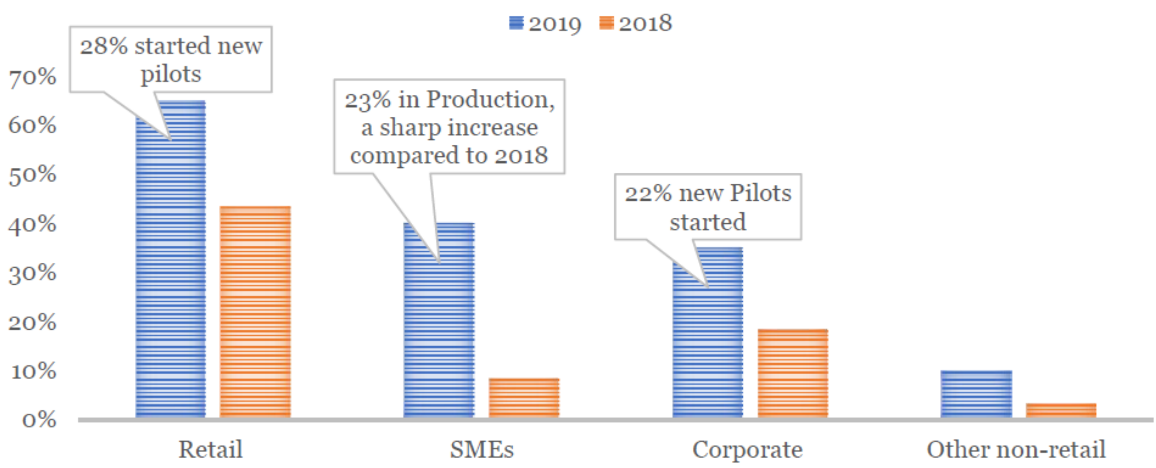


FIGURA 17 - APPLICAZIONE ML AI SEGMENTI DELLE CONTROPARTI[23]

3.3.3 Benefici e sfide future

Il vantaggio principale dell'utilizzo delle tecniche di machine learning è la possibilità di utilizzare molte variabili, anche in modo olistico, consentendo quindi l'estrazione di maggiori approfondimenti predittivi. Inoltre, a differenza dei modelli tradizionali, che per operare richiedono l'uso di dati già filtrati e aggiustati, utilizzando gli algoritmi di machine learning è possibile operare con una grande quantità di dati, anche se si presentano in forma incompleta o incoerente. Le banche che utilizzano il ML, possono quindi estrarre approfondimenti da insiemi di dati di grandi dimensioni, nonostante essi siano fortemente correlati o distorti. Utilizzando il machine learning, gli istituti finanziari sono in grado di sviluppare una moltitudine di modelli, in termini di obiettivi e costrutti, che gli consentono

[23] Fonte: Institute of International Finance (IIF). Machine Learning in credit risk report. 2019

di acquisire una maggiore comprensione degli schemi che governano i segmenti di popolazione. La Figura 18 mette in evidenza i benefici del ML rispetto ai modelli tradizionali in termini di accuratezza del modello, scoperta di ulteriori fattori prima sconosciuti, maggiore efficienza, superamento dei problemi di dati correlati, abilità di condurre analisi olistiche sui dati.

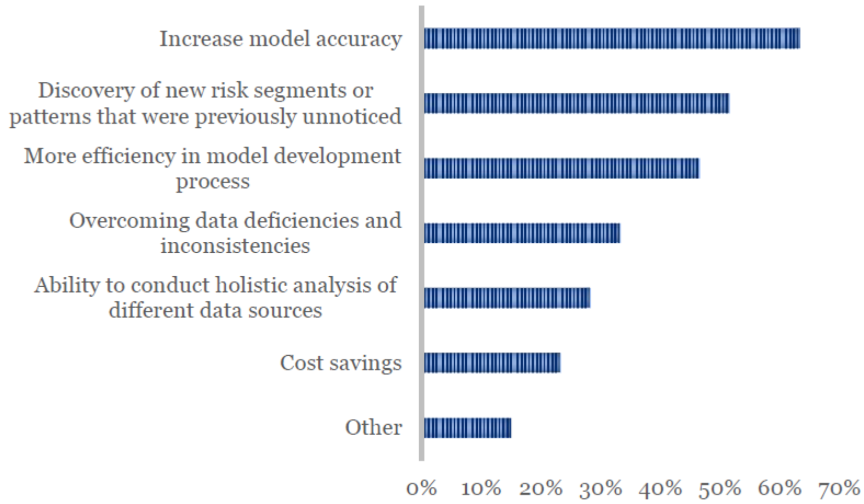


FIGURA 18 - BENEFICI DERIVANTI DALLE APPLICAZIONI ML[24]

La scoperta dei benefici che emergono dall'uso degli algoritmi di machine learning è stata graduale, invece la percezione delle sfide a cui gli istituti di credito vanno incontro si è presentata, man mano che essi hanno preso familiarità e conoscenza con tali tecniche. Come si evince dalla Figura 19 la sfida principale riscontrata riguarda l'interpretabilità dei risultati, seguita dalla difficoltà di spiegare i risultati, dal costo di implementazione delle tecnologie, dai problemi legati alle installazioni informatiche, dalla difficoltà di validare i risultati, dai vincoli governativi e dalla mancanza di supporto da parte degli stakeholders.

[24] Fonte: Institute of International Finance (IIF). Machine Learning in credit risk report. 2019

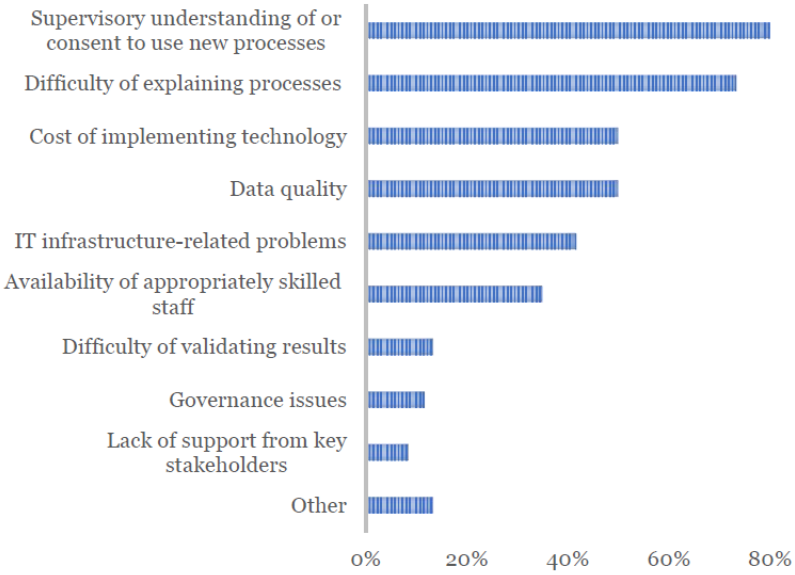


FIGURA 19 - SFIDE FUTURE NELL'UTILIZZO DEGLI ALGORITMI ML[25]

La comprensione da parte dei supervisori o il consenso all'uso dei nuovi processi è una grande sfida per gli istituti di credito. Si riscontra una sorta di collaborazione tra i due gruppi per superare questo problema, dato il crescente impiego delle tecniche di machine learning nei servizi finanziari e la maggiore curva di apprendimento della tecnologia.

Da parte delle banche è stato evidenziato all'autorità di vigilanza che manca una linea guida riguardo il livello di spiegazione che il modello dovrebbe avere per poter essere approvato in produzione. Si parla infatti di problemi per quanto riguarda la trasparenza, la verificabilità e l'interpretazione dei risultati da parte di revisori e supervisori. Questi problemi possono essere in parte superati dal momento che le banche (e i loro supervisori) sviluppano una maggiore maturità con le tecniche, aiutati da una combinazione di giudizio di esperti, strumenti di convalida come grafici di dipendenza parziale e ponendo maggiore enfasi al controllo dei dati di input.

La qualità dei dati rimane un importante ostacolo all'utilizzo delle tecniche ML per il rischio di credito. I problemi chiave associati alla qualità dei dati includono molteplici fonti e formati di dati e scarsa qualità dei dati, come dati incompleti o imprecisi. Del resto, tale issue vale per tutti i modelli.

Anche i problemi relativi all'infrastruttura IT sono stati considerati tra le principali sfide. I dirigenti hanno spiegato che nel corso degli anni molte banche hanno visto i loro sistemi IT

[25] Fonte: Institute of International Finance (IIF). Machine Learning in credit risk report. 2019

aggiornati in modo progressivo. Si parla in questi casi di sistemi legacy, ovvero quei sistemi retrodatati e obsoleti che non sono adatti all'applicazione degli algoritmi di machine learning; ad esempio questi sistemi, spesso, non supportano i linguaggi di codifica moderni ed è necessaria una traduzione.

Infine, la disponibilità di talenti con adeguate abilità, comprese le elevate competenze relative alla programmazione, alla matematica e allo sviluppo di software, e una forte comprensione degli algoritmi e delle ipotesi alla base di essi, è uno degli ostacoli più importanti per istituti di credito, ma anche di tutti coloro che si affacciano su questo ambito. Esiste una concorrenza significativa per attrarre e trattenere talenti dal ristretto pool di candidati con competenze ML. La carenza di talenti influisce direttamente sulla velocità di adozione del ML, con le banche consapevoli che i nuovi strumenti sono utili solo se accompagnati da utenti qualificati.

Poiché la competizione per i talenti esterni con queste specifiche competenze è così competitiva, molti operatori finanziari stanno investendo in programmi di miglioramento del personale interno per rafforzare il capitale umano che già possiedono.

Dato il suo potere e il suo impatto, il machine learning richiede uno sforzo più collaborativo tra gli istituti bancari e la comunità di supervisione per garantire che innanzitutto i clienti siano tutelati e che l'adozione dei modelli non sia ostacolata per evitare di bloccare l'innovazione nel settore finanziario. L'innovazione è vantaggiosa, tenendo conto, ovviamente, della consapevolezza del rischio che definisce il settore finanziario; pertanto, i progetti pilota che esplorano e testano nuove innovazioni meritano incoraggiamento da parte dei responsabili politici e dei supervisori. L'autorità di regolamentazione, quindi, dovrà affrontare la notevole sfida di allineare le diverse prospettive giuridiche, tecniche e politiche, garantendo al contempo che queste non ostacolino l'innovazione.

4 Interpretabilità degli algoritmi machine learning

Gli algoritmi di machine learning continuano il loro progressivo impiego nell'ambito del rischio di credito; questi permettono, infatti, alle banche di operare in modo più efficiente. L'adozione di questi metodi da parte di un numero sempre crescente di banche ha portato loro numerosi benefici in termini di: miglioramento dell'accuratezza dei modelli, superamento delle inconsistenze e mancanze nei dati e, a volte, anche la scoperta di nuovi parametri di rischio, prima non considerati.

L'autorità di Vigilanza, per evitare di bloccare il processo innovativo in ambito machine learning che stanno vivendo le banche, ma anche per rispettare gli obblighi giuridici, tecnici e politici dei quali è garante, si ritrova a dover decidere l'approvazione, o meno, di tali algoritmi innovativi. Il problema generale a cui l'autorità di vigilanza, ma non solo, anche chiunque si affacci a questa realtà, va incontro è legato anche all'interpretabilità dei risultati.

Il tema riguarda il modo in cui è possibile interpretare, e quindi predire, i risultati ottenuti con gli algoritmi di machine learning. Riuscendo a predire gli algoritmi, si riuscirebbe a capire come il modello evolve, pertanto si supererebbe l'idea che tali metodi siano paragonabili a delle "black box", o scatole nere. In alcuni casi, l'interpretabilità è considerata come sinonimo di capire come i modelli funzionano. Sono considerati dei modelli comprensibili quei modelli i quali presentano una trasparenza riguardo il loro funzionamento.

Una risposta generale a questo problema, che valga per tutti gli algoritmi di machine learning, ad oggi non esiste. Esistono tuttavia dei modelli, i quali tentano di descrivere diversi aspetti del problema, con i loro limiti e le loro potenzialità.

In questo capitolo, sarà presentato e spiegato il tema dell'interpretabilità del machine learning e le tecniche che si possono utilizzare per tentare di risolvere tale problema.

4.1 Definizione di interpretabilità

Il termine interpretabilità può avere diversi significati, dipendenti dal contesto. In alcuni casi può essere considerata sinonimo della comprensione del funzionamento dei modelli. Si parla, in questo caso, di trasparenza. Trasparenza che può essere correlata a tre livelli diversi: a livello di modello, a livello di componenti o a livello di allenamento dell'algoritmo.

- Trasparenza a livello di modello: si riferisce in questo caso alla comprensione del modello in sé; il modello, in genere, è di semplice comprensione come ad esempio i decision tree.
- Trasparenza a livello di componenti: si riferisce alla comprensibilità di ogni parte del modello inclusi i dati di input, i parametri e i processi svolti; un esempio può essere il singolo nodo di un albero decisionale.
- Trasparenza a livello di allenamento (training) dell'algoritmo: si intende che i fattori influenzanti l'algoritmo siano comprensibili anche dalle persone.

In altri casi, con il termine interpretabilità si intende la casualità e la trasferibilità, cioè la spiegazione di "cosa il modello riesce a dire". In questo contesto è opportuno distinguere la differenza tra modelli esplicativi e modelli predittivi. Spiegazione e previsione causali sono spesso confuse, ma ognuna gioca un ruolo diverso. La modellazione esplicativa è intesa come l'uso di modelli statistici per testare spiegazioni causali. I modelli di previsione sono descritti come il processo di applicazione di un modello a dei dati allo scopo di prevedere osservazioni nuove o future. Il focus della modellazione esplicativa è la causalità; il focus della modellazione predittiva è l'associazione. Ne consegue che il ruolo principale della modellazione predittiva è generare previsioni accurate di nuove osservazioni. La Figura 20 rappresenta le tecniche di apprendimento usate oggi e il trade-off che vi è tra l'accuratezza predittiva e l'interpretabilità di un modello. Il termine accuratezza predittiva si riferisce alla capacità di un modello di effettuare delle predizioni sulle osservazioni; con il termine interpretabilità ci si riferisce alla capacità di comprendere il modello. Modelli con alto potere predittivo possono essere le reti neurali, mentre quelli ad alta interpretabilità possono esserlo i modelli di regressione lineare.

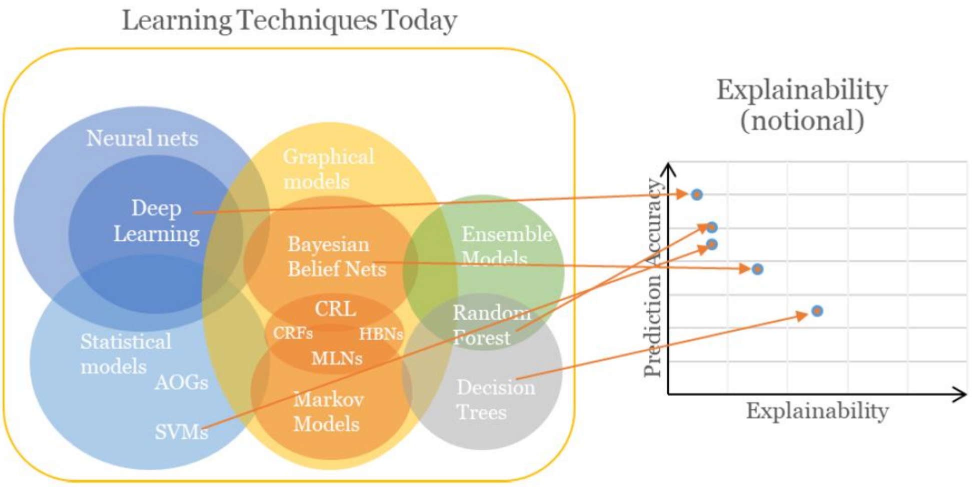


FIGURA 20 - TECNICHE DI APPRENDIMENTO VS INTERPRETABILITÀ

Si definisce l'interpretazione come la misura in cui un essere umano può comprendere le scelte prese dai modelli nel loro processo decisionale (come, perché e cosa). Ciò non significa spiegare in dettaglio come funziona un modello, ma piuttosto fornire informazioni utili.

La statistica ha tradizionalmente testato le teorie esplicative e utilizza modelli basati sulla correlazione come la regressione o un modello di percorso per catturare l'associazione per decidere se il modello causale è confutato o meno. I modelli costruiti a scopo esplicativo, quindi, sono diversi da quelli costruiti per la previsione: essi hanno obiettivi diversi. In genere, la modellazione esplicativa si fonda su un modello teorico che verifica un'ipotesi causale e, quindi, i dati si adattano ai parametri in un modello predefinito. La modellazione predittiva, invece, non prevede l'esistenza di un modello teorico, ma applica un algoritmo ai dati allo scopo di prevedere osservazioni nuove. (15)

Gli algoritmi machine learning hanno in comune con i modelli statistici la convalida incrociata dei risultati. La popolazione è ripartita, anche in questo caso, in un campione di training e un campione di stima, rispettivamente per l'applicazione e per la validazione del modello. Questo processo di suddivisione dati avverrà all'inizio, durante la fase di esplorazione e preparazione dei dati. Durante la fase di training, che in questo caso corrisponde all'addestramento del modello, i set di dati vengono tenuti separati e il campione di validazione non è utilizzato. L'idea è di valutare un diverso set di dati rispetto a quello utilizzato per addestrare il modello.

Una volta completata questa fase inizia il processo di addestramento della rete; a questo stadio il modello svolge un lavoro iterativo, durante il quale, letteralmente, impara dal passato per predire il futuro, basandosi su correlazioni e associazioni tra input e output predittivi, non su relazioni causali. Esaminando in modo accurato le previsioni, il machine learning può fornire approfondimenti su possibili spiegazioni, collegare i risultati a teorie esistenti e fornire idee per nuovi. L'uso di tecniche di machine learning a fini predittivi non impedisce l'utilizzo di modelli lineari per l'interpretazione. Tuttavia, è importante notare che gli algoritmi di apprendimento automatico non garantiscono la causalità.

Per ottenere un buon modello è necessario identificare e definire il tipo di problema che si intende risolvere all'inizio; ad esempio per la modellazione predittiva ci si dovrebbe concentrare sulla correlazione al rischio, anziché sui fattori di rischio. I modelli machine learning non spiegano il "perché", ma basandosi sulle correlazioni o associazioni tra i dati, formulano delle possibili previsioni future.

4.2 Importanza dell'interpretabilità

I modelli machine learning sono considerati come una "black box", in quanto non si ha una comprensione chiara sull'evoluzione del modello e sulla sua predizione. Con interpretabilità si intende l'abilità di prevederne il risultato. Maggiore è l'interpretazione di un modello di machine learning, più facile è capire perché sono state prese determinate decisioni o previsioni. Un modello è meglio interpretabile di un altro modello se le sue decisioni sono più facili da comprendere per un essere umano rispetto alle decisioni dell'altro. Se un algoritmo machine learning funziona bene, perché non ci si fida solo del modello e si ignora il perché abbia preso una determinata decisione? Quando si tratta di modellazione predittiva, bisogna tener conto di due aspetti: si vuole sapere solo cosa è previsto? Oppure si vuol conoscere il perché la previsione è stata fatta, capendo quindi anche qualche aspetto in più del problema?

Il tema dell'interpretabilità dei modelli machine learning nasce dalla curiosità umana. L'uomo ha il bisogno di trovare un significato nel mondo, pertanto vi è la necessità di interpretare il modello. L'altro fattore da tener conto è il rischio che si è disposti a correre; vi sono algoritmi di machine learning che non richiedono di essere interpretati, come quei modelli per la scelta dei film, o la raccomandazione dei prodotti acquistati su amazon, mentre altri, come quelli applicati in ambito credit risk o nella sanità, richiedono un più alto livello di interpretazione. Essa è un utile strumento di debug per poter rilevare errori nei modelli di apprendimento automatico. I modelli di apprendimento automatico possono solo essere sottoposti a debug e verificati quando possono essere interpretati. Un'interpretazione per una previsione errata aiuta a capire la causa dell'errore. Per verificare che il modello spieghi le decisioni, bisogna controllare i seguenti tratti:

- Equità: garantire che le previsioni siano imparziali e non discriminino implicitamente o esplicitamente alcuni gruppi.
- Privacy: garantire che le informazioni sensibili nei dati siano protette.
- Affidabilità o robustezza: assicurare che piccoli cambiamenti nell'input non portino a grandi cambiamenti nella decisione.
- Causalità: verificare che vengano rilevate solo le relazioni causali.
- Fiducia: è più facile per l'uomo fidarsi di un sistema che spiega le sue decisioni rispetto a una "black-box".

L'interpretazione non è richiesta se il modello non ha un impatto significativo e nemmeno quando il problema è ben studiato. (16) Mentre in ambiti come il rischio di credito nel quale il modello influenza in modo sempre più determinante la decisione creditizia è richiesto, ed è comprensibile, che sia il gestore debba essere in grado di comprenderne le logiche sottostanti sia i valutatori e l'autorità di vigilanza che devono controllare e permetterne l'utilizzo.

4.3 Classificazione dei metodi di interpretazione

I metodi per l'interpretazione degli algoritmi machine learning possono essere classificati in base a vari criteri.

- Intrinseco o post hoc:

  Questo criterio distingue se l'interpretazione si ottiene limitando la complessità del modello machine learning (intrinseco) o applicando metodi che analizzano il modello dopo la fase di training (post hoc). L'interpretabilità intrinseca si riferisce ad algoritmi che sono considerati interpretabili a causa della loro struttura semplice, come alberi decisionali o modelli lineari. L'interpretabilità post hoc si riferisce all'applicazione di metodi di interpretazione dopo l'addestramento del modello.

- Risultato del metodo di interpretazione:

I vari metodi di interpretazione possono essere approssimativamente differenziati in base ai loro risultati.

a) Statistica delle funzioni:
alcuni metodi forniscono statistiche di riepilogo per ciascuna funzione. Possono restituire un singolo numero per funzione, indice dell'importanza della funzione, oppure un risultato più complesso, come i punti di forza dell'interazione di una funzione tra una coppia di variabili.

b) Visualizzazione delle funzioni:
è possibile visualizzare dei grafici indicanti l'impatto e il risultato medio previsto di una variabile.

c) Punto dati:
Per spiegare la previsione di un'istanza di dati, il metodo trova un punto di dati simile modificando alcune delle funzionalità per le quali il risultato previsto cambia in modo pertinente (ad esempio, un ribaltamento nella classe prevista). Un altro esempio è l'identificazione di prototipi di classi previste. Per essere utili, i metodi di interpretazione che generano nuovi punti dati richiedono che i punti dati stessi possano essere interpretati. Funziona bene con immagini e testi, ma è meno utile per i dati tabulari con centinaia di funzioni.

- Specifico per il modello o indipendente dal modello:
Gli strumenti di interpretazione specifici del modello sono limitati a classi di modelli specifici. Esempi validi di metodi ad interpretazione specifica possono esserlo l'interpretazione dei pesi di regressione in un modello lineare oppure gli strumenti che funzionano per l'interpretazione di reti neurali.
Gli strumenti indipendenti dal modello possono essere utilizzati su qualsiasi modello di apprendimento automatico e vengono applicati dopo che il modello è stato addestrato (post hoc). Questi metodi agnostici di solito funzionano analizzando le coppie di input e output delle caratteristiche. Per definizione, questi metodi non possono avere accesso ai modelli interni come pesi o informazioni strutturali. (17)

4.3.1 Interpretabilità globale dell’algoritmo

4.3.1.1 Visione olistica

Il modello è definito interpretabile a livello globale se si riesce a comprendere la sua evoluzione nel tempo. Per spiegare l'output del modello globale, è necessario avere il modello già addestrato, la conoscenza dell'algoritmo e dei dati. Questo livello di interpretabilità riguarda la comprensione del modo in cui il modello prende le decisioni, sulla base di una visione olistica delle sue caratteristiche e di ciascuno dei componenti appresi come pesi, altri parametri e strutture. L'interpretazione del modello globale aiuta a comprendere la distribuzione dei risultati target in base alle funzionalità.

Tuttavia, l'interpretazione del modello globale è molto difficile da raggiungere, in quanto è improbabile che qualsiasi modello superante una piccola quantità di parametri o pesi si adatti alla memoria dell'essere umano (ad esempio è impossibile immaginare un modello lineare con un numero di caratteristiche superiore a 3, si creerebbe uno spazio multidimensionale). Di solito, quando si cerca di comprendere un modello, si prendono in considerazione solo parti di esso, come i pesi nei modelli lineari.

4.3.1.2 Visione modulare

Mentre l'interpretazione dei modelli globali è generalmente fuori portata, ci sono buone possibilità di comprendere almeno alcuni modelli a livello modulare. Non tutti i modelli sono interpretabili a livello di parametro. Per i modelli lineari, i moduli in considerazione, quindi le parti interpretabili, sono i pesi mentre per gli alberi sarebbero le divisioni del nodo foglia. Modelli lineari, ad esempio, sembra che possano essere perfettamente interpretati a livello modulare, ma l'interpretazione di un singolo peso può essere correlata agli altri. Bisogna, quindi, tenere anche conto di questo aspetto.

4.3.2 Interpretabilità locale dell’algoritmo

4.3.2.1 Singola previsione

Il termine interpretabilità locale, a livello della singola previsione, si riferisce alla comprensione di piccole parti dell’intero algoritmo, andando a spacchettare i diversi moduli, fino ad arrivare anche al livello delle singole previsioni. È possibile ingrandire una singola istanza ed esaminare le previsioni del modello per un determinato input e spiegare il perché il modello evolve verso una certa direzione. Se si osserva una previsione individuale, il comportamento del modello, altrimenti complesso, potrebbe comportarsi in modo più semplice e comprensibile. A livello locale, la previsione potrebbe dipendere solo linearmente o monotonamente da alcune funzionalità, piuttosto che avere una dipendenza complessa da esse.

4.3.2.2 Insieme di previsioni

Le previsioni del modello per più istanze possono essere spiegate con metodi di interpretazione globale del modello (a livello modulare) o con spiegazioni di singole istanze. I metodi globali possono essere applicati prendendo il gruppo di istanze, trattandole come se il gruppo fosse il set di dati completo e usando i metodi globali con questo sottoinsieme. I singoli metodi di spiegazione possono essere utilizzati su ciascuna istanza e quindi elencati o aggregati per l'intero gruppo. (18)

4.4 Valutare l’interpretazione

Non esiste un vero consenso su ciò che l'interpretazione rappresenta in ambito machine learning, né è chiaro come misurarla. Ma ci sono alcune ricerche iniziali su questo tema e un tentativo di formulare alcuni approcci per la valutazione.

- Valutazione a livello di applicazione (compito reale):

  Inserire la spiegazione nel modello e far effettuare dei test all'utente finale. Ciò richiede una buona configurazione sperimentale e una comprensione di come valutare la qualità. Una buona base, in questi casi, è sempre quanto un essere umano sarebbe bravo a spiegare la stessa decisione. Si parla in questo caso di predire il modello, riuscendo a capire la sua evoluzione; si tratta, quindi, di anticipare, in un certo senso, le scelte che il modello riesce a prendere.

- La valutazione a livello umano (compito semplice):

  Si tratta di una valutazione a livello di applicazione semplificata. La differenza è che questi esperimenti non vengono condotti da esperti. Questo rende gli esperimenti più economici ed è più facile trovare più tester. Un esempio potrebbe essere quello di mostrare all'utente diverse spiegazioni e lasciare poi all'utente la scelta la migliore.

- La valutazione a livello di funzione (attività proxy):

  Il metodo non richiede l'intervento dell'uomo. È inserita una funzione proxy del modello che spiega alcune peculiarità dello stesso. Un esempio può essere l'inserimento nel modello di alberi decisionali allo scopo di "spiegarlo" o renderlo più chiaro; in questo caso, un proxy per la qualità della spiegazione può essere la profondità dell'albero. Gli alberi più corti otterrebbero un punteggio migliore.

Per spiegare le previsioni di un modello di apprendimento automatico si fa affidamento su un metodo di spiegazione, ovvero un algoritmo che genera spiegazioni. Questo metodo mette in relazione i valori delle caratteristiche di un'istanza con la previsione del suo modello in modo umanamente comprensibile. (19)

I metodi di spiegazione possono essere classificati in base alle seguenti proprietà:

- Potere espressivo:

  Rappresenta il linguaggio o la struttura delle spiegazioni che il metodo è in grado di generare. Un metodo di spiegazione potrebbe generare regole IF-THEN, alberi decisionali, una somma ponderata o anche un linguaggio naturale.

- Specializzazione:

  Descrive quanto il metodo di spiegazione dipende dall'analisi del modello machine learning. I metodi di spiegazione basati su modelli intrinsecamente interpretabili, come il modello di regressione lineare, sono altamente traslucidi. I metodi che si basano solo sulla manipolazione degli input e sull'osservazione delle previsioni non hanno specializzazione. A seconda dello scenario, potrebbero essere desiderabili diversi livelli di specializzazione. Il vantaggio di un'elevata specializzazione è che il metodo può fare affidamento su più informazioni per generare spiegazioni, mentre il vantaggio di una bassa specializzazione è che il metodo di spiegazione è più versatile.

- Versatilità:

  Il modello di spiegazione è definito versatile se è possibile adattarlo a più modelli machine learning. Versatilità e specializzazione sono contrari in questo caso, in quanto i modelli che presentano un'alta versatilità hanno una bassa specializzazione e considerano il modello come una black-box.

- Complessità algoritmica:

  Descrive la complessità computazionale del metodo che genera la spiegazione. Questa proprietà è importante da considerare quando il tempo di calcolo è un bottleneck nel generare spiegazioni.

- Precisione:

Rappresenta un indice di quanto un modello esplicativo riesca a predire il modello. L'elevata precisione è particolarmente importante se la spiegazione viene utilizzata per le previsioni al posto del modello machine learning. Una bassa precisione può essere utile se l'obiettivo è spiegare cosa fa il modello black-box.

- Fedeltà:

Rappresenta la misura di quanto la spiegazione approssimi la previsione del modello machine learning. Spiegazioni con bassa fedeltà sono inutili per spiegare i modelli machine learning. Precisione e fedeltà sono strettamente correlate. Se il modello di scatola nera ha un'alta precisione e la spiegazione ha un'alta fedeltà, anche la spiegazione ha un'alta precisione. Alcune spiegazioni offrono solo fedeltà locale, il che significa che la spiegazione si avvicina bene alla previsione del modello per un sottoinsieme dei dati.

- Coerenza:

Indica quanto differisce una spiegazione riferita a due modelli che sono stati addestrati sullo stesso set di dati. Due spiegazioni, riferite a due modelli machine learning che producono risultati simili, si definiscono coerenti se danno risultati affini sui due modelli presi in considerazione.

- Stabilità:

Misura quanto sono simili le spiegazioni per casi simili. Mentre la coerenza confronta le spiegazioni tra i modelli, la stabilità confronta le spiegazioni tra istanze simili per un modello fisso. Elevata stabilità significa che lievi variazioni nelle caratteristiche di una variabile non cambiano la spiegazione. Una mancanza di stabilità può essere il risultato di un'elevata varianza del metodo di spiegazione. (20)

4.4.1 Importanza delle interazioni e delle permutazioni

Caratteristiche comuni ai modelli di interpretabilità sono l'importanza delle interazioni tra le features, che potrebbero dare dei risultati distorti, e delle permutazioni dei valori per stimarne l'impatto sui modelli:

- Interazione tra le features

Quando le funzionalità interagiscono tra loro in un modello di previsione, la previsione non può essere espressa come la somma degli effetti della funzione, poiché l'effetto di una funzione dipende dal valore dell'altra. Se un modello machine learning effettua una previsione basata su due funzioni, si può scomporre la previsione in quattro termini: un termine costante, un termine per la prima funzione, un termine per la seconda funzione e un termine per l'interazione tra le due features. L'interazione tra due funzionalità è la modifica della previsione che si verifica variando le variabili dopo aver considerato gli effetti delle singole funzionalità: in un certo senso, il calcolo è analogo alla legge di composizione della varianza.

- Permutazioni dei valori
  L'importanza della funzione di permutazione misura l'aumento dell'errore di predizione del modello dopo la perturbazione dei valori della funzione, interrompendo la relazione tra la funzione e il risultato reale.

Si misura l'importanza di una funzione calcolando l'aumento dell'errore di predizione del modello dopo averla permutata. Una funzione è "importante" se cambiando i suoi valori aumenta l'errore del modello, perché in questo caso il modello si è basato sulla funzione per la previsione. Una funzione è "non importante" se il cambiamento dei suoi valori lascia invariato l'errore del modello, poiché in questo caso il modello ha ignorato la funzione per la previsione.

4.5 Modelli di interpretabilità

Una volta esposte le caratteristiche che possono avere i metodi di spiegazione degli algoritmi machine learning, sono adesso presentati i modelli utilizzati per la loro valutazione in termini di validità, interpretabilità e affidabilità.
I modelli di interpretazione che prevedono la separazione delle spiegazioni dall’algoritmo machine learning sono definiti modelli agnostici. Hanno come obiettivo quello di interpretare i modelli, senza preoccuparsi del tipo di algoritmo utilizzato, in quanto possono essere applicati a svariati algoritmi. Tali modelli valutano le relazioni che esistono tra i dati in input, valutano l’impatto delle caratteristiche sui dati e formulano delle ipotesi sui dati in output. La caratteristica principale di questi modelli indipendenti dagli algoritmi applicati è la loro versatilità. Gli sviluppatori di machine learning sono liberi di utilizzare qualsiasi modello che preferiscono in quanto i metodi di interpretazione possono essere applicati a qualsiasi algoritmo. In generale, molti tipi di modelli machine learning sono valutati per

risolvere un'attività e, quando si confrontano i modelli in termini di interpretabilità, è più facile lavorare con spiegazioni indipendenti dal modello.

Un'alternativa ai metodi di interpretazione indipendenti dal modello è quella di utilizzare solo modelli interpretabili, che tuttavia hanno il grande svantaggio di perdere prestazioni predittive rispetto ad altri algoritmi. L'altra alternativa è utilizzare metodi di interpretazione specifici del modello ma lo svantaggio è che ci si fossilizza su un tipo di modello e sarà difficile applicarlo ad altri (bassa versatilità).
Nelle sezioni successive saranno presentati i principali strumenti di interpretabilità dei risultati degli algoritmi machine learning.

4.5.1 Partial Dependence Plot (PDP)

Partial Dependence Plot (PDP), o diagramma di dipendenza parziale, è un indicatore dell'effetto marginale che una o due caratteristiche hanno sul risultato previsto dal modello. Il PDP mette in evidenza se la relazione tra il risultato e una caratteristica è lineare, monotona o più complessa. Ad esempio, quando applicati a un modello di regressione lineare, i grafici di dipendenza parziale presentano sempre una relazione lineare.

La metodologia Partial Dependence Plot prende in considerazione solo la relazione tra alcune features "target", in generale una o due, emarginando il contributo di tutte le altre definite "complemento". Le features target sono, solitamente, tra le più importanti, ovvero quelle che hanno un impatto maggiore. L'assunzione che vi è alla base è l'assenza di correlazione tra le features target e quelle complemento, anche se nella realtà, quest'assunzione è spesso violata. (21) La Figura 21 presenta degli esempi di grafici mettendo in evidenza la relazione tra le features (ascisse) e il prezzo delle case (risultato ottenuto con un algoritmo machine learning). La figura in alto a sinistra rappresenta l'effetto lineare che ha il reddito sul prezzo delle case. Si possono anche ottenere dei grafici, che mettono in relazione due features e le loro interazioni. La figura in basso a destra rappresenta quale impatto hanno le variabili HouseAge (età della casa) e AveOccup (occupanti medi per famiglia). Si nota come per numero di occupanti superiori a 2, si ha un'indipendenza dall'età mentre per valori inferiori si riscontra una forte dipendenza.

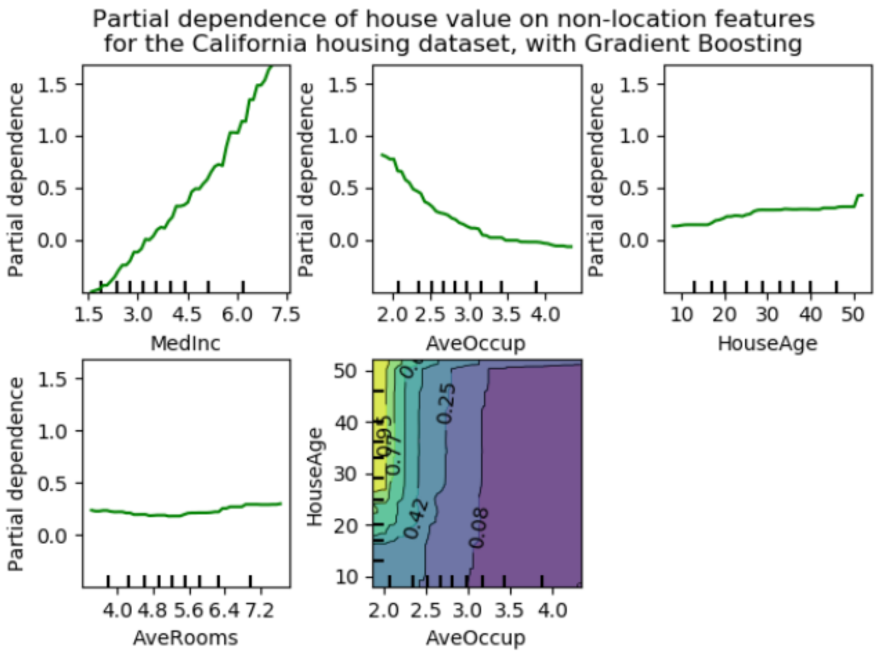


FIGURA 21 - PARTIAL DEPENDENCE PLOT[26]

Il calcolo dei grafici di dipendenza parziale è intuitivo; la funzione di dipendenza parziale rappresenta la relazione tra la feature in considerazione e l'output dell'algoritmo machine learning. Ha come svantaggio, tuttavia, l'uso massimo di due variabili per volta.

4.5.2 Individual Conditional Expectation (ICE)

Individual Conditional Expectation (ICE), o Aspettativa condizionale individuale, rappresenta una riga per istanza (dato) che mostra il cambiamento della previsione dipendente dalla funzione in considerazione. Il diagramma di dipendenza parziale (PDP) rappresenta l'effetto medio di una funzione, non si concentra cioè sulle istanze specifiche, ma su una media complessiva. L'equivalente di un PDP per singole istanze di dati è chiamato grafico delle aspettative condizionate individuali (ICE).

Il diagramma ICE visualizza la dipendenza della previsione da una funzione per ciascuna istanza separatamente, risultante in una riga per istanza. Un PDP è la media delle linee di un diagramma ICE. I valori per una linea (e un'istanza) possono essere calcolati mantenendo tutte le altre features uguali, creando delle varianti sostituendo il valore della funzione con i valori di una griglia e facendo previsioni con il modello machine learning per le nuove

[26] Web: https://scikit-learn.org/stable/modules/partial_dependence.html

istanze create. I diagrammi di dipendenza parziale possono oscurare una relazione eterogenea tra le features e come si presenta la relazione media tra una feature e la previsione; funziona bene solo se le interazioni tra le features target, per le quali viene calcolato il PDP, e le altre features (complemento) sono deboli. In caso di interazioni, il diagramma ICE fornirà molte più informazioni. (22) Un esempio di diagramma ICE è presentato in Figura 22, rappresentante, per ogni dato, una linea indicante la probabilità prevista di cancro (output), in base all'età (feature); la linea centrale in giallo corrisponde alla media delle diverse osservazioni, rappresentante, quindi, un diagramma PDP.

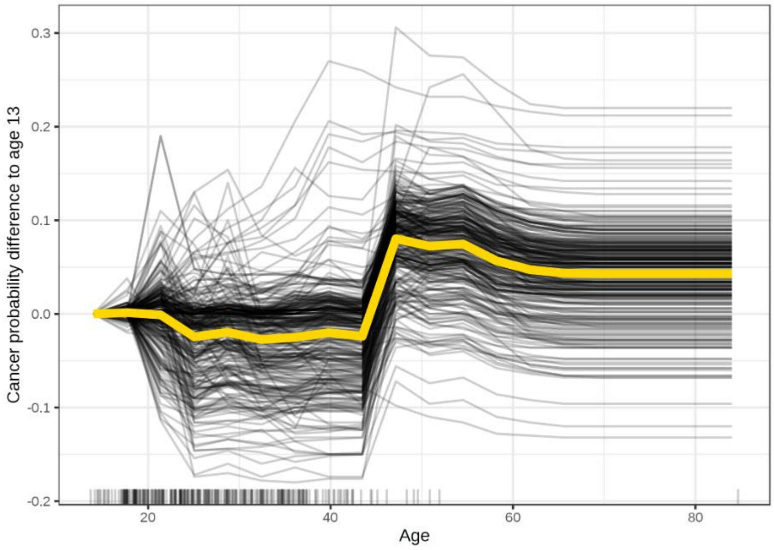


FIGURA 22 - ICEBOX[27]

I diagrammi ICE sono ancora più intuitivi da comprendere rispetto ai grafici PDP. Una riga rappresenta le previsioni per un'istanza se si varia la caratteristica di interesse. A differenza dei grafici PDP, le curve ICE possono scoprire l'esistenza di relazioni eterogenee. Le curve ICE possono visualizzare, tuttavia, solo una caratteristica in modo significativo. In più, se la feature in considerazione, è correlata ad un'altra, si potrebbe avere un risultato distorto. (23)

4.5.3 Global Surrogate Models

Con il termine metodi surrogati globali si identificano i metodi di facile comprensione e a basso costo computazionale che possono essere utilizzati per predire le scelte del modello machine learning realmente applicato. Anch'essi sono algoritmi di machine learning con la

[27] Web: https://christophm.github.io/interpretable-ml-book/lime.html

peculiarità di essere facilmente interpretabili. I modelli surrogati sono anche usati in altri ambiti di tipo ingegneristico (come le simulazioni) soprattutto se il risultato ha un costo elevato. I vincoli da tenere in considerazione in questi casi sono: approssimazione il più possibile accurata al modello reale e che il modello sia, appunto, di facile interpretazione; il modello utilizzato deve essere indipendente dal modello reale, ovvero non contenere nessuna informazione riguardante tale modello; per la sua implementazione è necessario solo l'accesso ai dati. Un metodo utilizzato per la valutazione del modello surrogato è l'R^2, indice di quanta varianza è spiegata nel modello. Quanto esso è più vicino a 1 e tanto il modello surrogato approssima bene il risultato del modello reale.

Come vantaggi presenta la possibilità di adattare diversi modelli allo stesso dataset e fornire diversi risultati, ognuno interpretabile in una certa misura (lineare, albero decisionale etc.); l'approccio è molto intuitivo e diretto, con facile implementazione e interpretazione; l'R^2 è un'ottima misura della bontà del modello.
Si deve avere la consapevolezza tuttavia che si traggono informazioni sul modello, e non sui dati o le caratteristiche; non è chiaro quale deve essere il miglior valore di R^2 per essere sicuri che il modello surrogato sia vicino al modello reale. Può accadere che si ha una buona approssimazione per un particolare sottoinsieme di dati, ma divergente per un altro sottoinsieme.

4.5.4 Local Interpretable Model-agnostic Explanations (LIME)

LIME è un modello surrogato localmente interpretabile per spiegare le singole previsioni dei modelli machine learning. Invece di formare un modello surrogato globale, LIME si concentra sulla formazione di modelli locali per spiegare le previsioni individuali.

L'idea su cui si fonda è molto intuitiva: capire perché il modello machine learning ha fatto una certa previsione. LIME verifica ciò che accade alle previsioni quando si forniscono variazioni dei dati nell'algoritmo. Si genera un nuovo set di dati costituito da campioni permutati e dalle corrispondenti previsioni del modello inteso come una black-box. Su questo nuovo set di dati LIME addestra quindi un modello interpretabile, che è ponderato dalla vicinanza delle istanze campionate all'istanza di interesse. Il nuovo modello può essere di qualsiasi tipo, anche un albero decisionale, e deve contenere una buona approssimazione locale del modello machine learning usato. Non è importante che vi sia un’approssimazione globale, invece è essenziale la fedeltà locale. La metodologia per l'addestramento di modelli surrogati locali è la seguente:

- selezione dell’istanza d’interesse per la quale si desidera avere una spiegazione della previsione del modello black-box;
- perturbazione del dataset e ottenimento dei risultati per la nuova previsione;
- ponderazione di nuovi campioni in base alla loro vicinanza all'istanza di interesse;
- Addestramento di un modello ponderato e interpretabile sul set di dati con le variazioni;
- Spiegazione della previsione interpretando il modello locale.

Per procedere, si sceglie un numero di variabili K da inserire nel modello interpretabile; vi è un trade-off nella scelta del numero di variabili, in quanto più il numero è basso e meglio è interpretabile il modello, per contro più è alto e meglio si approssima al modello reale. Per la scelta delle variabili, si possono considerare delle logiche differenti: forward o backward (inserendo o togliendo le variabili). Per la variazione dei dati, dipende dal tipo di dati in considerazione: testo, immagini, dati tabulari. Per dati testuali o immagini basta attivare o disattivare alcuni elementi quali parole o pixel; per i dati tabulari LIME crea nuovi campioni perturbando individualmente ogni feature. La Figura 23 rappresenta un esempio di un output ottenuto con il modello LIME; in particolare, sono presentate due istanze di set di dati riguardanti un servizio di noleggio biciclette. Le feature prese in considerazione sono la temperatura e le condizioni meteorologiche. Una temperatura più calda e le buone condizioni meteorologiche hanno un effetto positivo sulla previsione. L'asse x mostra l'effetto della funzione: il peso moltiplica il valore effettivo della funzione.

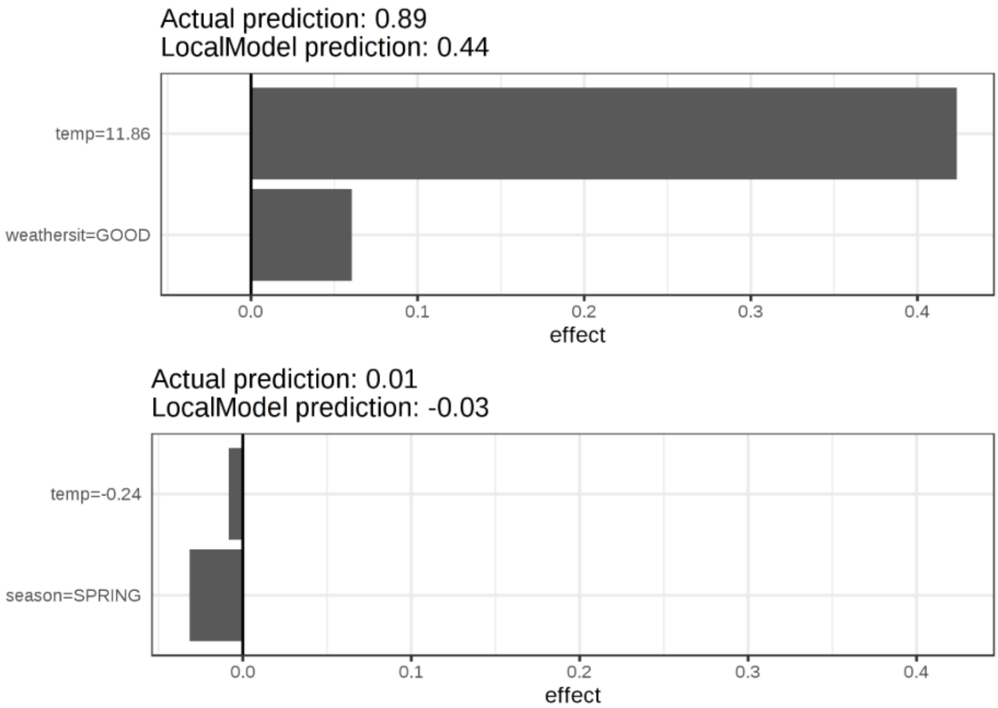


FIGURA 23 - LIME[28]

Matematicamente, i modelli surrogati locali con vincolo di interpretabilità possono essere espressi come segue:

$$\boldsymbol{\varepsilon(x)} = \boldsymbol{arg\,min_{g \in G} L(f, g, \pi_x) + \omega(g)} \quad (31)$$

- $\varepsilon(x)$ rappresenta la spiegazione del modello nell'intorno dell'istanza x;

- g rappresenta il modello interpretabile utilizzato; - L rappresenta la perdita da minimizzare, indice della vicinanza tra interpretazione e modello originale;

- f rappresenta la previsione del modello originale;

- G rappresenta l'insieme dei modelli interpretabili;

- π_x definisce quanto è grande l'intorno dell'istanza considerata;

- ω rappresenta la complessità del modello interpretabile.

Il primo termine misura l'errore di g nell'approssimazione di f (modello originale) nell'intorno limitato. Il secondo termine è una misura della complessità del modello della

[28] Web: https://christophm.github.io/interpretable-ml-book/lime.html

spiegazione g (modello interpretabile). Ad esempio, se il modello usato è un albero decisionale, si può riferire alla profondità dell'albero o, nel caso di modelli di spiegazione lineare, può essere il numero di pesi diversi da zero.

LIME ottimizza solo la parte di perdita locale nell'intorno dell'istanza. I vantaggi che offre LIME sono molti: creando un nuovo algoritmo machine learning, più interpretabile dell'originale, è possibile avere una descrizione del problema sotto diversi punti di vista, prima non considerati, e comprendere la convenienza nell'uso di un modello machine learning piuttosto che un altro. Quando si usano modelli brevi (le variabili saranno più selettive) si possono avere spiegazioni contrastanti. Questi modelli forniscono interpretazioni a misura d'uomo, dividendo il problema generale in sotto-problemi con una maggiore comprensione. LIME è uno dei pochi metodi che si presta all'interpretazione di modelli che lavorano sia con dati tabulari, sia testo e sia immagini. (24) (25)

4.5.5 SHAP

La tecnica SHAP si fonda sulla teoria dei valori di Shapley che trae la sua origine nella teoria dei giochi. Ogni istanza del database corrisponde a un "giocatore" in un gioco, nel quale la previsione rappresenta il payoff. A differenza della teoria dei giochi, nella quale il payoff è assegnato ai giocatori sulla base delle scelte che essi prendono, in questo caso il payoff ottenuto è frutto della combinazione delle variabili che caratterizzano il database. Ad ogni variabile, quindi, è assegnato un peso, o contributo marginale, ed esso rappresenta il valore di Shapley.

Nell'ambito della teoria dei giochi, il valore di Shapley[29] è un sistema per distribuire il payoff ottenuto (ricompensa) dalla coalizione tra i componenti del gioco e lo scopo che si prefigge è di distribuire tale ricompensa in modo proporzionale al contributo che ogni giocatore apporta alla coalizione. Una possibile soluzione a tale calcolo consiste nel fare una media di tutti i contributi marginali del giocatore su tutti gli ordinamenti possibili dei giocatori presenti nella coalizione. Matematicamente è possibile rappresentare il concetto secondo la formula (32).

$$\emptyset(\boldsymbol{i},\boldsymbol{v}) = \frac{1}{|N|!}\sum_{\pi \in \prod N} \boldsymbol{v}[\,\boldsymbol{B}(\boldsymbol{\pi},\boldsymbol{i})\,\boldsymbol{U}\,\{\boldsymbol{i}\}\,] - \boldsymbol{v}\,[\,\boldsymbol{B}(\boldsymbol{\pi},\boldsymbol{i})\,] \qquad (32)$$

[29] Web: https://it.wikipedia.org/wiki/Valore_di_Shapley

Nella quale i termini indicano:

- $\emptyset(i, v)$: payoff ricevuto dal giocatore i;

- v : funzione caratteristica rappresentante l'utilità per ogni coalizione di giocatori;

- N : insieme di giocatori;

- $\prod N$: è l'insieme di tutti gli ordinamenti possibili degli elementi N;

- $B(\pi, i)$: insieme dei giocatori che precedono il giocatore i nell'ordinamento preso in considerazione;

Nell'ambito del machine learning invece, il contributo marginale di ogni variabile è calcolato considerando tutte le possibili interazioni con le altre presenti nel modello. Si valuta, quindi, quanta informazione è contenuta in ogni combinazione, stimando il valore aggiunto che ogni variabile apporta nella previsione. Ad ogni variabile è associato un contributo marginale sulla base dell'incremento dell'accuratezza della previsione.

Nello specifico, sono testate molte combinazioni, dipendenti dal numero di features da inserire nel modello. Per ogni variabile si calcolano tutte le combinazioni con le altre attribuendo in una prima fase il peso previsto; successivamente si calcola la stessa previsione senza considerare la variabile in questione e, sulla base della differenza di previsione ottenuta, si attribuisce il contributo marginale alla variabile. Questo processo è effettuato per ogni variabile del modello su tutte le istanze del database per avere il valore medio del contributo marginale di ogni variabile.

Il valore di Shapley potrebbe essere l'unico metodo per fornire una spiegazione completa, in quanto in situazioni in cui la legislazione richiede risultati spiegabili, il valore di Shapley risulterebbe legalmente conforme, poiché basato su una solida teoria scientifica. Tuttavia, a livello computazionale è molto oneroso in quanto il calcolo per le possibili combinazioni richiede molto tempo di elaborazione e adeguati sistemi software. Anche riducendo il valore delle features (n), per diminuirne il tempo di calcolo, si raggiungerebbe una soluzione approssimativa, in quanto la varianza del valore Shapley aumenta. Il numero n dovrebbe essere abbastanza grande da stimare con precisione i valori di Shapley, ma abbastanza piccolo da completare il calcolo in un tempo ragionevole.

In Figura 24 è rappresentato un output della tecnica SHAP: il problema in questione è la ricerca dell'influenza che hanno le variabili (ordinata sinistra), sulla variabile target (qualità del vino); dall'ampiezza dei segmenti è possibile valutare l'impatto della variabile, mentre dal colore è possibile capire sia il verso sia l'effetto che assume sulla previsione. Considerando la prima caratteristica, il livello di alcool, si nota che quando essa assume valori elevati (rosso) ha un contributo maggiore sulla predizione della variabile target.

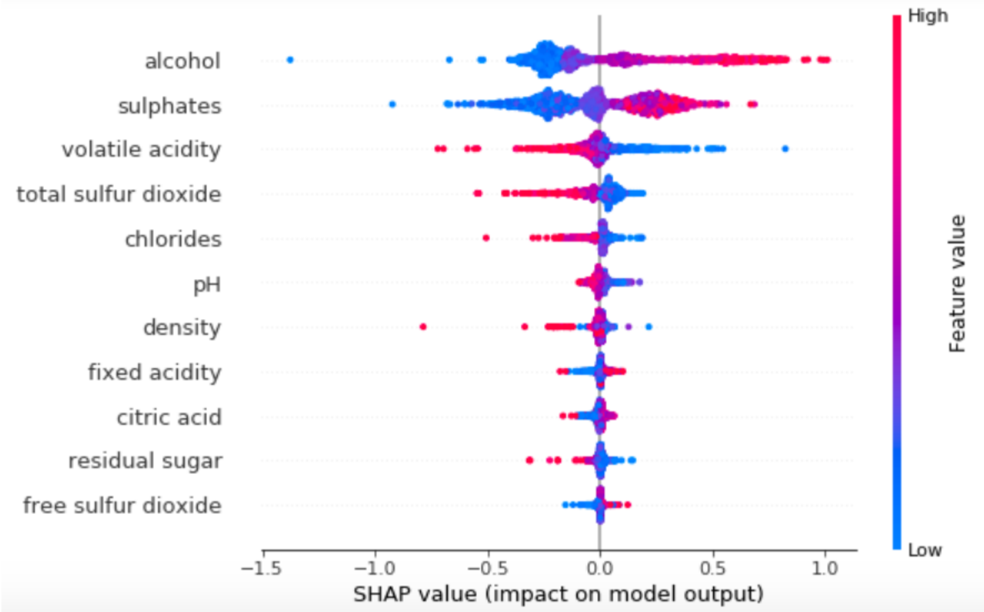


FIGURA 24 - SHAP[30]

4.6 Il futuro dell'interpretabilità

Per poter capire quale potrebbe essere il futuro dell'interpretabilità, bisogna considerare i possibili sviluppi del machine learning e dell'intelligenza artificiale. Negli ultimi periodi le aziende, come già discusso, stanno investendo risorse e capacità sull'intelligenza artificiale. Un problema riscontrato a livello generale può essere l'integrazione dei modelli innovativi con la cultura aziendale, che ad oggi in molte realtà non risulta essere pronta. L'apprendimento automatico quindi da un lato offre potenzialità, ma dall'altro richiede sistemi all'avanguardia per poterle cogliere pienamente.

L'evoluzione del machine learning si potrebbe rilevare con una crescita lenta ma costante, in quanto sempre più spesso aumentano le applicazioni dall'ambito scientifico e di test,

[30] Web: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

all’ambito aziendale dei processi, prodotti e servizi. L’evoluzione potrebbe essere, invece, su larga scala seguendo il principio “ciò che può essere automatizzato, sarà automatizzato” e le attività saranno formulate come problemi di ottimizzazione e risolti con il machine learning. Molte attività attualmente svolte dall'uomo potranno essere sostituite dall'apprendimento automatico, come già successo con le rivoluzioni industriali. L’interpretazione dei risultati può catalizzare l’adozione del machine learning, in quanto gli algoritmi non possono essere perfettamente specificati e in molte aree o settori non si adotteranno per la mancanza di spiegazione. Oltre alla mancata specificazione del problema, molte industrie richiedono interpretabilità, sia per motivi legali, a causa dell'avversione al rischio o per ottenere informazioni sull'attività sottostante.
L’utilizzo, pertanto, di modelli interpretabili per colmare il divario tra l’interpretazione del modello black-box e gli obiettivi raggiunti potrebbe quindi rendere attraente l’implementazione del machine learning su larga scala in ambito aziendale.

Per quanto riguarda, quindi, i possibili sviluppi dell’interpretabilità dei risultati, essa dipende molto dall’evoluzione del machine learning. In tendenza vi sono gli algoritmi agnostici che utilizzano modelli disaccoppiati dal modello originale, molto versatili e applicabili in diversi contesti, allo scopo di verificare la bontà e l’affidabilità dell’algoritmo originale. Si possono avere anche dei modelli integrati con le spiegazioni allo scopo di rendere più comprensibili le attività svolte dall’algoritmo: calcolo dell’importanza delle feature, diagrammi di dipendenza parziale sono solo alcuni dei possibili esempi.

In conclusione, l’applicazione di tools di interpretabilità accompagna l’adozione delle tecniche machine learning. La relazione delle tecniche di interpretabilità però non è completamente dipendente dal machine learning; è sbagliato pensare “senza il machine learning, non sarà necessaria l’interpretabilità”, in quanto quest’ultima può influenzare positivamente l’adozione del machine learning su largo impiego. Avendo a disposizione delle tecniche in grado di poter spiegare le decisioni del modello, o le influenze dei dati sul modello, si riuscirebbe ad “aprire la black-box” per eliminare lo scetticismo che vi è riguardo questo tema e favorirne un maggiore sviluppo. (26)

5 Caso Pratico: modulo “cash-flow” del rating SME Retail di Intesa Sanpaolo

Nel seguente capitolo sarà presentata un’applicazione degli algoritmi di machine learning e delle tecniche di interpretabilità dei risultati, presentati nei capitoli precedenti, ad un caso pratico. Nello specifico si tratta del modulo “cash-flow” del modello di Rating per l’intero segmento SME Retail del portafoglio crediti di Intesa Sanpaolo. Il modello è strutturato secondo una logica modulare, tale per cui ogni area informativa è esplorata a parte e poi integrata alle altre. Il modulo è stato stimato sulla base di un campione rappresentativo in cui sono presenti tutte le controparti che hanno Intesa Sanpaolo come banca principale, aventi quindi un conto corrente definito “attivo” secondo dei criteri interni stabiliti dalla banca. Il modulo specifico, fa parte del segmento Small Medium Enterprises Retail (abbreviato in SME Retail) comprendente le piccole e medie imprese, definite tali da criteri interni riguardanti i livelli di esposizione o di fatturato.

Il capitolo sarà suddiviso in una prima parte riguardante l’esplorazione del dataset utilizzato per le analisi, nella quale sarà presentato il database e le principali operazioni di data cleaning e data preparation. La seconda parte presenterà l’applicazione degli algoritmi di machine learning con i relativi studi di accuracy per la definizione del modello migliore e a seguire, nella terza parte del capitolo, saranno applicate le tecniche di interpretabilità dei risultati ottenuti dal modello più performante.

5.1 Analisi del dataset ed esplorazione dei dati

Il dataset analizzato è composto da clienti che presentano una linea di credito attiva nei confronti della banca e aventi le caratteristiche sopra elencate. Le controparti facenti parte del campione esaminato sono circa 300.000 caratterizzate da un codice univoco per ognuna di esse. Il dataset è stato costruito su cinque punti di osservazione al 31/12 di ogni anno tra il 2013 e il 2017, nel quale la situazione finanziaria della controparte è stata esaminata nel caso dei 12 mesi precedenti.

Ad ogni controparte è stato associato un altro parametro (variabile target) rappresentante l’evento default nel corso dei 12 mesi successivi: 0 nel caso di non default e 1 nel caso di default. Ogni controparte del campione, quindi, può ripetersi per più anni, ottenendo un totale di istanze (righe) dell'intero dataset di circa 820.000. Ogni controparte presenta, quindi, la situazione al 31/12 per gli anni 2013, 2014, 2015, 2016 e 2017 con annessa la variabile default (0,1) per ogni anno.

Ogni istanza del dataset presenta una variabile che può assumere due valori categorici "SVIL" o "TEST". Questa variabile discrimina la popolazione del dataset in due campioni differenti: SVIL definisce il campione di sviluppo (o training), utilizzato per addestrare il modello di machine learning; TEST definisce il campione di testing, utilizzato per il testare il modello. I due campioni sono tenuti separati durante l'esecuzione del modello ML, per evitare di avere risultati distorti o un modello che "impara" dai dati in maniera troppo specifica (fenomeno di overfitting). Il campione di training rappresenta l'80% della campione totale, mentre il campione di stima il 20%. La popolazione è stata suddivisa nei due campioni in modo casuale.

Definite le variabili indicanti lo stato delle controparti, si passa ad analizzare le features discriminatorie che entrano nel modello machine learning e che saranno indice della valutazione della probabilità di default per ogni controparte per ogni anno. Il numero totale delle variabili che entrano nel modello è di 66. Tali features rappresentano indicatori della situazione finanziaria delle controparti, costruiti sulla base dell'osservazione delle informazioni di c/c. Le suddette variabili riguardano diverse aree informative e sono state analizzate mensilmente o nei 12 mesi precedenti. Esse riguardano:
- movimentazioni di conto corrente;
- variazione della liquidità;
- variazione degli incassi;
- totale degli afflussi di cassa lordi e netti;
- totale dei deflussi di cassa lordi e netti;
- Saldo del conto bancario;
- rapporto tra il saldo e l'afflusso totale di denaro;
- stipendi pagati;
- confronti tra il saldo e le entrate di conto corrente;
- etc.

Le diverse features, essendo indici di misure diverse, hanno valori diversi; alcune possono essere dei ratio, altri dei valori numerici interi in migliaia. Tutte le features, prima di essere processate dagli algoritmi machine learning, hanno subito una fase di data cleaning e data preparation, necessaria per renderle confrontabili tra di loro ed avere una misura unica ed uniforme per tutto il dataset. Per far ciò, ad ogni feature è stata applicata la metodologia Weight Of Evidence (WoE) secondo la formula (10). La trasformazione delle variabili, discrete o continue, in variabili ordinali permette di incorporare la struttura del tasso di default nell'indicatore in questione e di renderle quindi confrontabili tra loro.

In una prima fase la variabile è suddivisa in classi ordinali (se ad esempio la variabile è lo stipendio, si identificano delle fasce quali 0-500, 500-1000, 1000-1500, 1500-2000 e così via) e per ogni classe si identificano il numero di controparti in bonis e il numero di controparti in default. Applicando la formula (10), si ottiene un numero, che può assumere valori limitati portando i diversi valori sotto un'unica scala di misura. Ci si aspetta, tuttavia, che i valori ottenuti, procedendo dalle classi peggiori alle migliori, seguono un andamento crescente in modo monotono; se così non fosse, allora le classi che presentano un andamento non monotono sono accorpate insieme e il valore di WoE è ricalcolato per la nuova classe ottenuta. Il processo si ripete fino a quando non si raggiunge l'andamento monotono desiderato. La monotonicità è essenziale in questo caso, poiché valori più bassi di WoE sono indici di probabilità di default (PD) più alte, mentre da valori alti di WoE derivano PD più basse. Questo procedimento è effettuato per tutte le variabili in questione. Per assicurare la monotonicità di tutte le variabili, il processo di accorpamento delle classi può avvenire più volte, fino a ridurre il numero di classi anche in termini di due o tre. Le feature che alla fine della procedura presentano più valori hanno una monotonicità più regolare e risultano essere più predittive di quelle che hanno subito numerosi processi di accorpamento. Un esempio è riportato in Tabella 3, nella quale è rappresentato il secondo valore di WoE non monotono come gli altri, pertanto è accorpato al primo e ne è rappresentata la nuova classe in Tabella 4; infine sono rappresentate le nuove classi e i rispettivi valori di WoE in Tabella 5.

TABELLA 3 - CLASSI DELLA VARIABILE E CALCOLO WOE

classi stipendio originali	numero bonis	numero default	WoE
0-500	1000	360	-0.3086
500-1000	1500	600	-0.4140
1000-1500	1400	400	-0.0775

1500-2000	2000	200	0.9723
totale	5900	1560	

TABELLA 4 - ACCORPAMENTO CLASSE 1 E 2 E CALCOLO NUOVO WOE

nuova classe	numero bonis	numero default	WoE
0-1000	2500	960	-0.3732

TABELLA 5 - CLASSI E WOE FINALI

classi stipendio corretto	numero bonis	Numero default	WoE
0-1000	2500	960	-0.3732
1000-1500	1400	400	-0.0775
1500-2000	2000	200	0.9723
totale	5900	1560	

Una volta effettuata questa procedura per tutte le variabili in questione, il dataset risulta essere pronto per il suo utilizzo nei modelli machine learning e si presenta nella forma della Tabella 6, in cui è rappresentata ogni controparte, con il proprio codice univoco, la data analizzata, i valori del flag Default/bonis, il discriminante SVIL/TEST e tutte le variabili calcolate con la tecnica WoE.

TABELLA 6 – DATABASE

CODICE UNIVOCO	DATA	SOGLIA DEFAULT	FLAG SVIL/TEST	FEATURE 1	FEATURE 2	…	FEATURE 66
1	31/12/2013	0	SVIL	…	…	…	…
1	31/12/2014	0	TEST	…	…	…	…
1	31/12/2015	0	SVIL	…	…	…	…
1	31/12/2016	0	SVIL	…	…	…	…

1	31/12/2017	1	SVIL	…	…	…	…
2	31/12/2013	0	SVIL	…	…	…	…
2	31/12/2014	1	SVIL	…	…	…	…
2	31/12/2015	0	SVIL	…	…	…	…
2	31/12/2016	0	TEST	…	…	…	…
2	31/12/2017	0	SVIL	…	…	…	…
3	31/12/2015	0	SVIL	…	…	…	…
3	31/12/2016	1	SVIL	…	…	…	…
…	…	…	…	…	…	…	…
N	31/12/2013	0	SVIL	…	…	…	…
N	31/12/2014	0	SVIL	…	…	…	…
N	31/12/2015	0	SVIL	…	…	…	…
N	31/12/2016	0	SVIL	…	…	…	…
N	31/12/2017	0	TEST	…	…	…	…

Le variabili presenti nel modello, d'ora in avanti saranno denominate: "woe_Numero_Variabilie" e una descrizione del significato economico sarà affiancata nel testo, senza rivelare il significato di tutte le variabili in questione per motivi di privacy per la banca e per i clienti.

5.2 Applicazione degli algoritmi Machine Learning

Una volta strutturato il database, è stato possibile lanciare gli algoritmi di machine learning per determinare le probabilità di default per le diverse controparti in esame. Gli algoritmi utilizzati sono stati il Random Forest e l'XGBoost. Entrambi gli algoritmi fanno parte della famiglia tree-based, poiché fondano le loro scelte sul principio di funzionamento degli alberi decisionali. Un albero decisionale elabora ad ogni nodo una feature alla volta dividendo la popolazione in diversi sotto-campioni, fino ad arrivare agli ultimi nodi contenenti la popolazione perfettamente discriminata.

L'algoritmo Random Forest crea e combina tanti alberi decisionali non correlati modellando le diverse caratteristiche in modo casuale e la predizione finale sarà una media delle predizioni dei singoli alberi decisionali.
L'algoritmo XGBoost, invece, crea degli alberi in modo iterativo, minimizzando l'errore, tra una previsione e la precedente. La predizione finale sarà il prodotto congiunto di tutte le predizioni ottenute.

Gli algoritmi di machine learning sono governati da diversi parametri, i quali possono essere preimpostati per avere delle soluzioni personalizzate. I parametri in questione cambiano in base all'algoritmo usato; per quanto riguarda il random forest i parametri principali sono i seguenti:
- n_estimators: individua il numero di alberi da utilizzare nel modello;
- criterion = "Gini": misura la qualità della divisione nei vari nodi;
- max_depth: massimo numero di nodi verticali in un albero;
- min_samples_split: Il numero minimo di campioni richiesti per dividere un nodo interno; se è rappresentato da un 'float' è una frazione e rappresenta il numero minimo di campioni per ciascuna divisione.
- min_samples_leaf: Il numero minimo di campioni richiesto per essere in un nodo foglia.

I parametri, invece, principali per l'algoritmo XGboost sono:
- max_depth: profondità degli alberi;
- learning_rate: tasso di apprendimento del modello;
- n_estimators: numeri di alberi da creare;
- reg_lambda: termine di regolarizzazione sui pesi.

Utilizzando il linguaggio di programmazione Python, sono stati definiti diversi valori per i parametri sopra citati ed è stata applicata una funzione Grid-Search allo scopo di individuare la combinazione di parametri migliori. In tal modo, in base ai dati utilizzati, è possibile definire i parametri del modello, che meglio si adattano al dataset. La tecnica di ottimizzazione è automatizzata per i parametri al fine di individuarne la combinazione che ne consenta di massimizzare le performance. Sono, infatti, costruiti iterativamente diversi modelli caratterizzati da differenti valorizzazioni di parametri, scegliendo infine la configurazione ottimale per la particolare analisi effettuata. Dalla funzione Grid-Search gli algoritmi migliori da utilizzare sono risultati i seguenti, caratterizzati dai parametri sotto indicati:

- Random Forest:
  { max_depth = 25 ; min_samples_leaf = 50 ; min_samples_split = 0.01 ; n_estimators=200 }
- XGBoost:
  { learning_rate = 0.2 ; n_estimators = 200 ; reg_lambda = 0.8 }

Ottenuti i parametri migliori per entrambi gli algoritmi, essi sono stati addestrati sul dataset di training ed utilizzati per fornire le predizioni delle probabilità di default.

A questo punto, per ogni controparte sono state ottenute due probabilità di default, ottenute dai due modelli effetto di assunzioni diverse su cui si basano. Per definire il modello migliore da adottare è stato considerato l'Accuracy Ratio (AR), che insieme alla curva ROC, è un indicatore di accuratezza del modello. L'AR è stato calcolato per entrambi i modelli e confrontato in termini di intera popolazione, campione di sviluppo e campione di test. I risultati sono riportati in Tabella 7.

TABELLA 7 - ACCURACY RATIO

Modulo CASH FLOW	AR		
	SVILUPPO	TOT POPOLAZIONE	TEST
XGBoost	0.776	0.773	0.764
Random Forest	0.740	0.738	0.730

Il modello che ha presentato un Accuracy Ratio maggiore è stato l'algoritmo XGBoost, pertanto questo è stato scelto come modello ufficiale per la stima delle probabilità di default.

5.3 Interpretabilità dei risultati

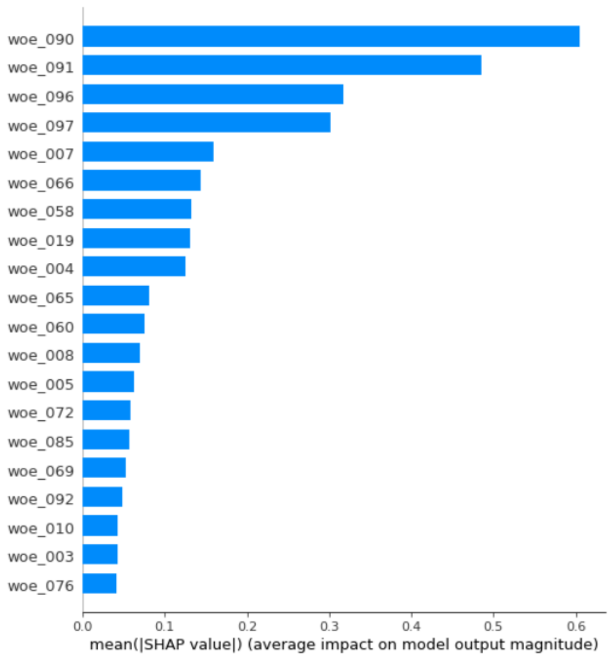
Scelto l'algoritmo, sorge il problema dell'interpretabilità dei risultati. Il tema dell'interpretabilità, già discusso nel capitolo precedente, è un argomento che sta crescendo di pari passo con l'implementazione dei modelli di machine learning. Il focus sull'interpretabilità dipende, ovviamente, dal problema affrontato; utilizzando un algoritmo per scopi personali (scelta del film da vedere ad esempio) o per scopi aziendali è ben diverso e si hanno impatti differenti. Data la natura del problema affrontato in questo studio, si è scelto di implementare sul software Python le tecniche di interpretabilità precedentemente discusse. Saranno presentate, in ordine, le tecniche di interpretabilità globale con l'obiettivo di individuare le variabili rilevanti e la loro relazione con l'output del modello. In seguito, sarà implementata la tecnica di interpretabilità locale LIME, per individuare le variabili rilevanti, che a livello di singola controparte, hanno dato un contributo maggiore nella definizione della probabilità di default. Di seguito saranno riportati gli output delle tecniche ottenuti dall'implementazione sul software Python.

5.3.1 Applicazioni tecniche interpretabilità al caso in esame

5.3.1.1 Applicazione SHAP

La tecnica SHAP, presentata nel paragrafo 4.5.5, è una tecnica di interpretabilità globale da utilizzare allo scopo di individuare le variabili rilevanti per il modello. L'applicazione della tecnica prevede l'implementazione di un modello machine learning da cui trarne le spiegazioni. Pertanto, il modello è stato allenato sul medesimo campione di stima impostando gli stessi parametri ottenuti dalla "grid-search" utilizzata nella versione ufficiale. È stato usato come algoritmo di riferimento l'XGBoost, per avere una visualizzazione grafica esclusiva del modello utilizzato.

- SHAP VALUES



OUTPUT 1 - IMPORTANZA VARIABILI SHAP

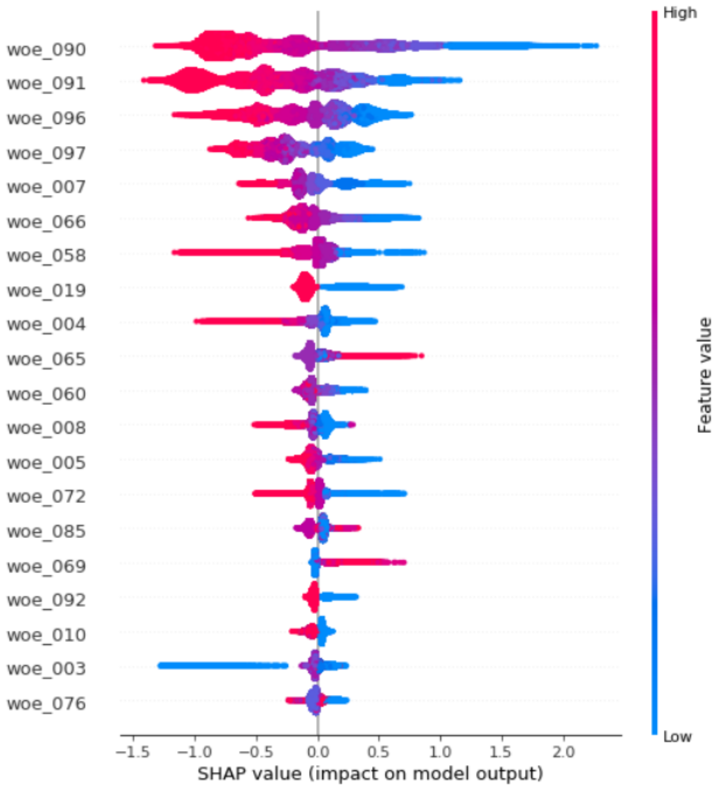
La tecnica SHAP individua le variabili più rilevanti, assegnando un peso a ciascuna di esse, come rappresentato dall'Output 1.

L'output raffigurato individua le variabili con peso maggiore e le rappresenta in ordine di importanza, riportando il contributo marginale per ogni variabile sull'asse delle ascisse. Le dieci variabili più rilevanti sono le seguenti, accompagnate da una breve descrizione riguardo il loro significato economico:

- woe_090: Saldo conto bancario ultimi 6 mesi / afflusso denaro ultimi 6 mesi;
- woe_091: Saldo conto bancario ultimi 12 mesi / afflusso denaro ultimi 12 mesi;
- woe_096: Pagamento tasse / deflusso denaro ultimi 6 mesi;
- woe_097: Pagamento tasse / deflusso denaro ultimi 12 mesi;
- woe_007: Variazione entrate lordo mensile ultimi 6 mesi;
- woe_066: Totale afflussi lordi ultimi 6 mesi;
- woe_058: Totale Afflussi di cassa ultimi 6 mesi;
- woe_019: Flag ritardi di pagamento ultimi 12 mesi;
- woe_004: Variazione liquidità ultimi 12 mesi;
- woe_065: Spese in stipendi ultimi 12 mesi.

- Output impatto variabili SHAP

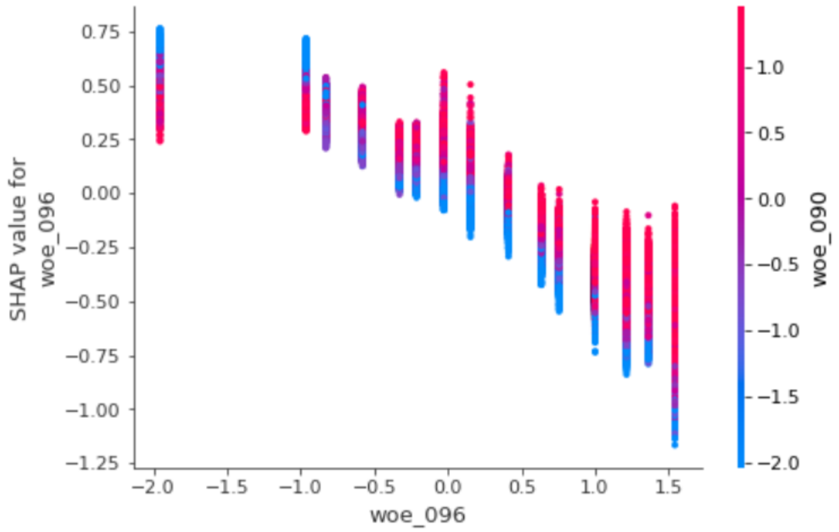
Una rappresentazione migliore e più completa è visualizzato nell' Output 2, il quale raffigura le variabili e i loro contributi marginali, specificandone, tramite il colore, il loro impatto.



OUTPUT 2 - IMPATTO VARIABILI SHAP

Come mostrato dalla scala di colore sulla destra, valori alti delle feature sono raffigurati in rosso (valori positivi per come sono stati costruiti i WoE), mentre i valori bassi sono in blu (negativi). Sull'asse delle ordinate, sono rappresentate le variabili in ordine di impatto, mentre sull'asse delle ascisse il valore del loro contributo (SHAP value). L'output combina l'impatto delle variabili con i loro effetti, a seconda del valore assunto. Ad esempio, per la variabile 90 ("woe_090") è possibile notare che essa ha un impatto positivo, ovvero contribuisce all'aumento della probabilità di default, quando assume valori negativi (blu); mentre ha impatto negativo, contribuendo all'abbassamento della probabilità di default, quando assume valori alti (rosso). Dalla forma dell'output per ogni variabile è possibile anche notare la distribuzione della medesima. Nel leggere l'output occorre anche tener conto che le variabili non sono indipendenti le une dalle altre, e nella rappresentazione incide anche il peso delle variabili. Nel grafico le feature in alto sono le più predittive e la loro rappresentazione è coerente con la loro costruzione (valori negativi di woe sono indici di alta PD, mentre valori positivi di bassa PD); data questa definizione può sembrare fuorviante il comportamento della variabile "woe_065" che presenta un andamento opposto rispetto alle altre, tuttavia bisogna tener conto del suo peso e del suo Accuracy Ratio che risultano essere bassi e prevaricati dai valori di WoE delle variabili più importanti.

- Output dipendenze SHAP



OUTPUT 3 - RELAZIONE VARIABILI SHAP

Tramite l'uso della tecnica SHAP è possibile anche creare un grafico, raffigurante la relazione esistente tra due variabili e come interagiscono nella stima della probabilità di default. L'Output 3 rappresenta la relazione esistente tra due variabili; nel caso in esame è stata scelta la variabile "woe_096" e il programma seleziona in automatico un'altra variabile con cui la prima interagisce maggiormente producendo un alto potere predittivo. Alla variabile 96, indicante i pagamenti nei confronti della Pubblica Amministrazione in confronto al totale dei pagamenti, è stata associata la variabile 90, rappresentante il rapporto tra il saldo di conto corrente e l'afflusso di denaro negli ultimi sei mesi. Le variabili in questione sono tra quelle con peso maggiore e dall'Output 3 è possibile cogliere il peso associato alla variabile scelta (96) in base alla relazione con l'altra (90). Le barre verticali indicano i possibili contributi che la variabile 96 assume dall'interazione con la 90. Ogni barra è situata in prossimità dei valori che la variabile 96 può assumere. L'ampiezza delle barre indica la variabilità del contributo marginale associata alla variabile 96. Ad esempio, quando la variabile 96 vale -2.0 è associato il contributo marginale maggiore, rappresentato sulle ordinate a sinistra, quando anche la variabile 90 assume valori negativi, rappresentati in blu; mentre quando la variabile 90 assume valori alti (positivi, in rosso), il contributo marginale della variabile 96 diminuisce, nonostante essa assuma sempre il valore peggiore di -2.0.

In questa chiave, è possibile interpretare ogni barra e il contributo marginale della variabile scelta, in base alla combinazione con l'altra variabile generata. Nella Figura 25 è rappresentato un insieme di output rappresentanti le variabili principali e la loro interazione con quelle che producono un potere predittivo alto. È possibile ottenere questo tipo di output utilizzando qualsiasi variabile, tuttavia, per quelle con un peso basso si avrebbero dei risultati distorti. Questo tipo di output è particolarmente interessante qualora si fosse interessati a capire la relazione che vi è tra due variabili, oppure comprendere il cambiamento del contributo marginale associato loro.

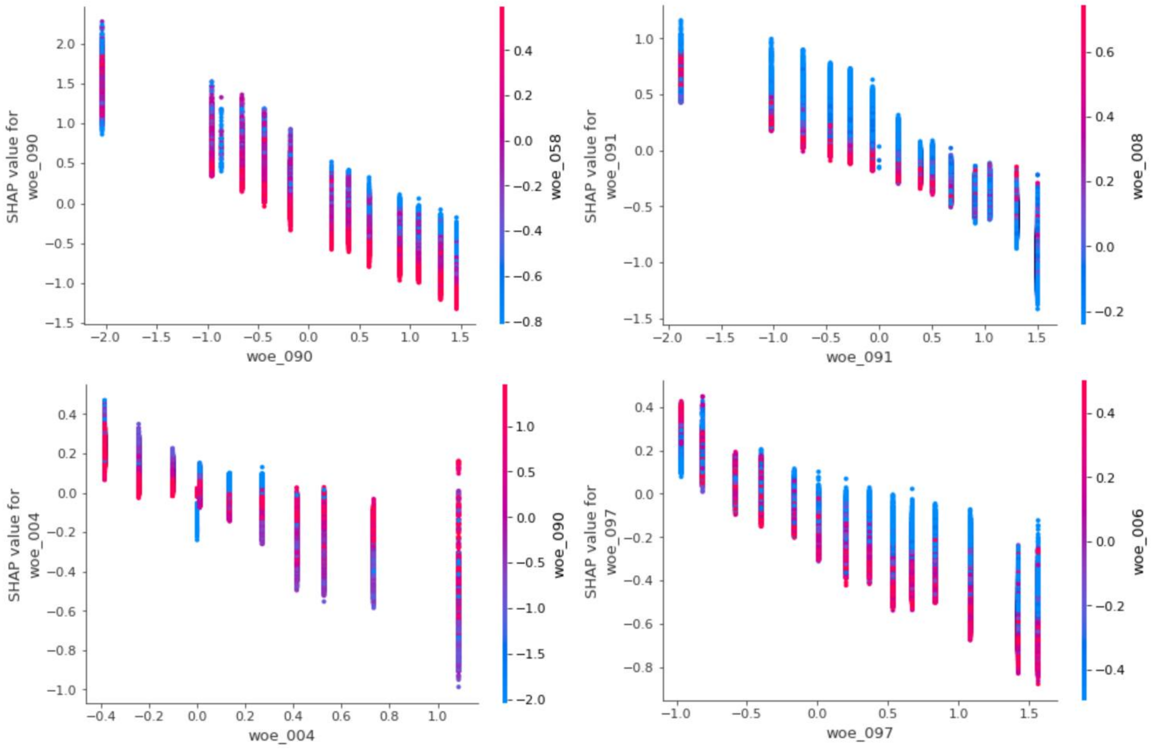


FIGURA 25 - INSIEME OUTPUT SHAP DEPENDENCE

La prima tecnica presentata ha l'obiettivo, come discusso, di individuare le variabili più performanti assegnando loro un peso marginale, rappresentare le medesime in relazione le une con le altre e, cosa più importante, rappresentare il potere predittivo associato ad ognuna di esse.

5.3.1.2 Applicazione Partial dependence plot (PDP)

Dopo aver individuato le variabili più performanti del modello, è possibile utilizzare la seconda tecnica per rappresentare la relazione esistente tra la variabile in esame e l'output del modello, ovvero la probabilità di default. La metodologia PDP, già discussa al paragrafo 4.5.1, è anch'essa una tecnica di interpretabilità globale. Questa tecnica rappresenta il contributo di una variabile osservando come cambia la PD.

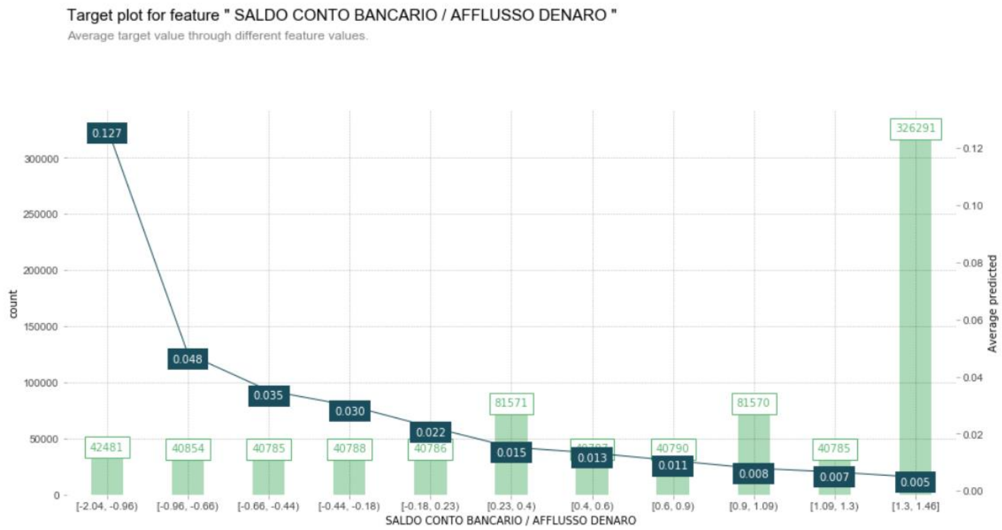
Anche in questo caso è stato utilizzato l'algoritmo XGBoost, pertanto i risultati sono esclusivi del modello implementato, presentano quindi versatilità nulla.

- PDP singola variabile

Come rappresentato nell'Output 4, si individua la relazione esistente tra la variabile 90 e la probabilità di default. Sulle ascisse sono rappresentati i valori che la variabile può assumere;

sulle ordinate a destra sono rappresentati i livelli di PD media; sulle ordinate a sinistra il numero di controparti interessate. L'output rappresenta quindi, il cambiamento della PD al variare dei valori della feature in esame, riportando per ogni valore il livello di PD medio e la numerosità delle controparti.

La relazione esistente tra la variabile e la probabilità di default risulta essere di tipo iperbolico. La probabilità di default più alta è predetta quando la variabile assume valori bassi, e all'aumentare dei valori assunti, diminuisce la PD ma a tassi decrescenti. Ciò è indice dell'alto impatto che assume la variabile quando presenta valori bassi, ma diminuisce il potere predittivo nei casi in cui assume valori via via migliori.

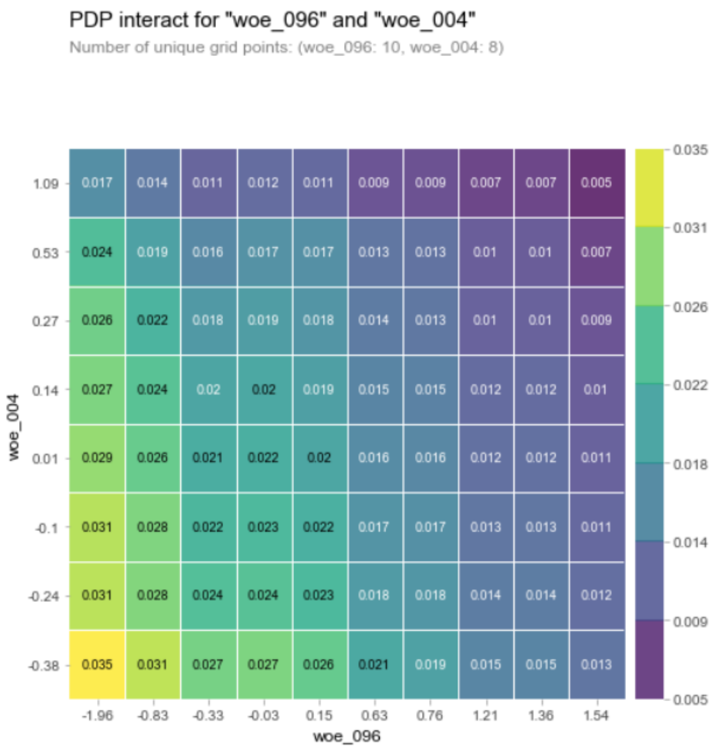


OUTPUT 4 - PDP SINGOLA VARIABILE

La variabile scelta per la rappresentazione è la più performante, ma anche in questo caso è possibile rappresentare gli effetti di tutte le variabili. La relazione decrescente tra variabile e output era già nota, data la natura della costruzione delle variabili, ma il valore aggiunto di questo output è l'andamento iperbolico della variabile indice di un'elevata capacità predittiva nei casi in cui assume valori più bassi.

- PDP due variabili

È possibile, tramite l'uso di questa tecnica, rappresentare l'output del modello considerando l'interazione tra due variabili. Anche in questo caso, il risultato può essere utile nel caso in cui si è interessati a visualizzare l'effetto di due variabili, trascurando le altre.



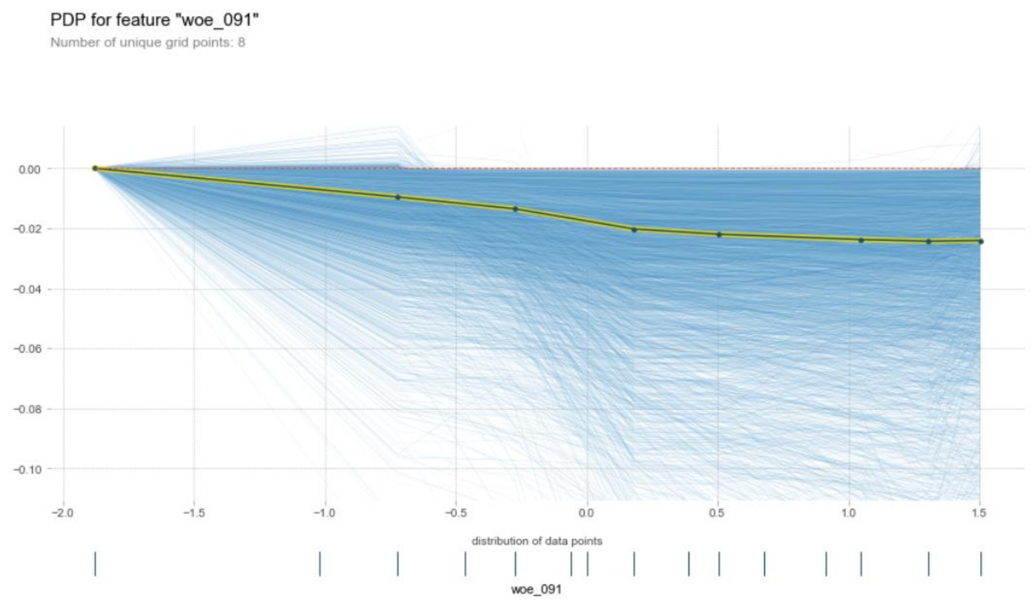
OUTPUT 5 - PDP DUE VARIABILI

L' Output 5 rappresenta l'interazione di due variabili e i livelli di PD medi prodotti. Le variabili in questione, indicano rispettivamente il livello di tasse sulle spese totali e la disponibilità di liquidità. Dal colore è possibile notare il cambiamento del livello medio di PD per le diverse combinazioni di valori delle due features, come rappresentato dalla barra di colore sulla destra. Combinazioni diverse producono PD diverse. Ad esempio, il riquadro in basso a sinistra, che presenta una PD di 0.035, indica la combinazione dei valori peggiori per entrambe le variabili. All'interno ricadono le controparti con bassa liquidità e con un'elevata percentuale di tasse pagate. L'output risulta avere un andamento regolare, in quanto i dati sono stati trattati già prima di essere usati nel modello di machine learning.

5.3.1.3 Applicazione Individual Conditional Expectation (ICE)

La tecnica ICE, presentata al paragrafo 4.5.2, è una tecnica di interpretabilità globale con l'obiettivo di individuare l'impatto che ha la singola variabile sulla stima della PD. Si isola il contributo della variabile dalle altre, rappresentando la variabile come se fosse indipendente da tutte le altre. Anche in questo caso è stato utilizzato come algoritmo l'XGBoost, per avere una rappresentazione esclusiva del modello.

Un esempio della tecnica ICE è presentato, di seguito, dall' Output 6. La variabile in questione è il rapporto tra il saldo e le entrate nel corso dell'ultimo anno.



OUTPUT 6 - ICEBOX

Per ottenere questo output, l'algoritmo simula l'assegnazione di tutti i valori che una variabile può assumere ad ogni controparte presente nel database. Una controparte avrà quindi diversi valori della stessa variabile e l'effetto sarà rappresentato dalla linea in blu. Ad ogni linea blu corrisponde una controparte, ottenuta facendo variare i valori della variabile in questione. Sulle ascisse è rappresentata la fascia di valori che la variabile può assumere, mentre sulle ordinate la variazione di PD provocata da una variazione della variabile. In media, si nota un andamento decrescente raffigurato dalla linea gialla. Si possono notare alcune linee in blu con un andamento crescente, nonostante all'aumentare del valore della variabile ci si aspetterebbe una diminuzione della PD. Ciò può essere provocato dall'effetto di qualche altra variabile che, nonostante il miglioramento della variabile in questione, impatta maggiormente nel modello.

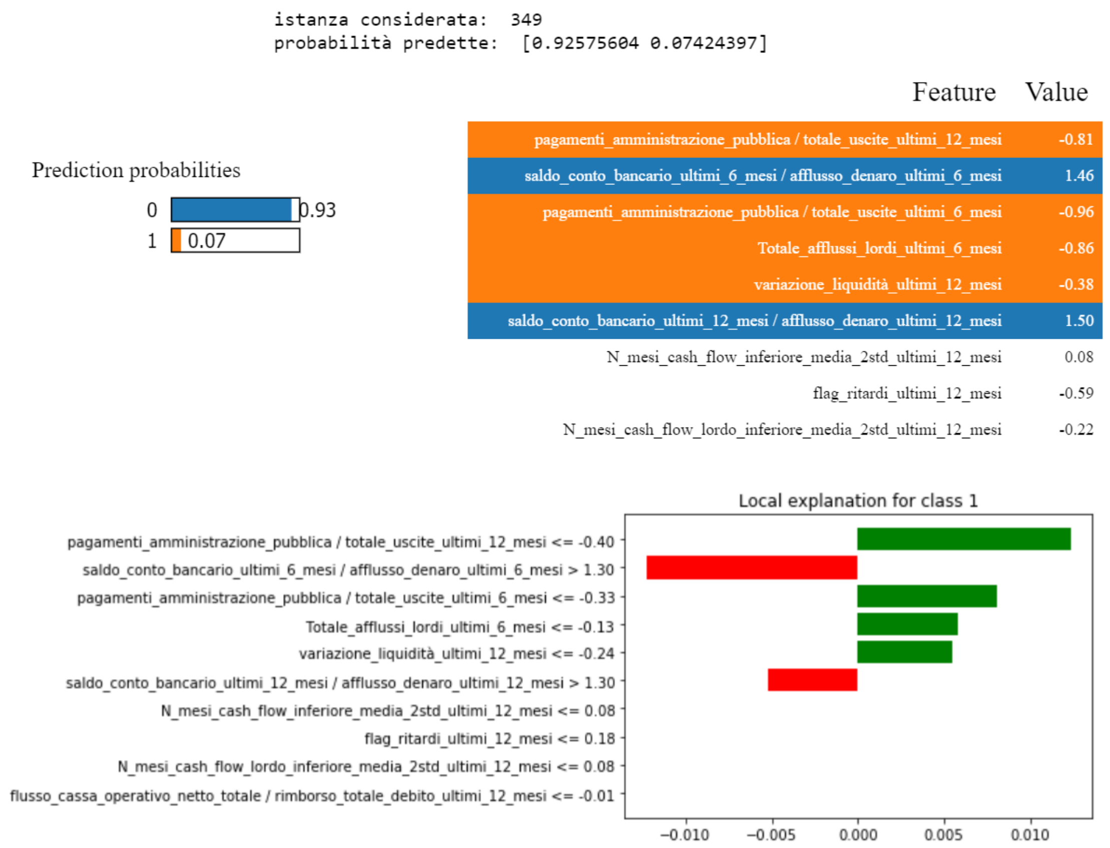
5.3.1.4 Applicazione Local Interpretable Model Explanation (LIME)

Una volta implementate le tecniche di interpretabilità globale, il cui obiettivo è quello di trovare le relazioni esistenti tra le variabili e tra queste ultime e l'output del modello, è stata applicata la metodologia LIME, descritta nl paragrafo 4.5.4. Questa tecnica ha l'obiettivo di individuare le variabili rilevanti per la definizione dell'output del modello a livello di singola controparte. Per ogni istanza del database (singola controparte) l'algoritmo definisce una probabilità di default e questa metodologia aiuta a comprendere il motivo per il quale il modello ha assegnato tale livello di PD.

La letteratura scientifica prevede l'utilizzo di un algoritmo più semplice, in generale un modello Decision Tree, per l'implementazione della tecnica LIME allo scopo di interpretare il risultato prodotto dal modello ufficiale, nel caso in esame XGBoost. È stato scelto, tuttavia, di applicare la metodologia LIME all'algoritmo XGBoost in modo da avere un risultato affidabile e specifico del modello in esame. Usare un modello più semplice per spiegarne uno più complesso ha il vantaggio di risparmiare tempo macchina ma lo svantaggio di produrre risultati approssimativi.

La base su cui si fondano entrambi i modelli sono le medesime, tuttavia le scelte di discriminazione sono leggermente diverse, come presentato nel paragrafo 3.2.1, pertanto i risultati finali saranno diversi, sia in termini di PD sia di variabili rilevanti. In questi casi, occorre considerare la natura del problema e valutare se è necessario ottenere un risultato approssimato e replicabile o esclusivo ma non replicabile. La tecnica, quindi, è stata applicata ad entrambi gli algoritmi ma è stata scelta l'applicazione al modello XGBoost, in modo da comprendere appieno le scelte effettuate ed avere una rappresentazione grafica chiara ed univoca.

Nell'Output 7 è rappresentato un esempio di applicazione della metodologia LIME.



OUTPUT 7 - LIME

In ordine è possibile visualizzare:

- l'istanza del database, corrispondente ad una specifica controparte;

- le probabilità predette: probabilità di essere in bonis e probabilità di default. Nel caso in esame, la controparte ha una probabilità di default pari a 0.07424387. È presente anche una rappresentazione grafica (approssimata alla seconda cifra decimale) delle stesse, indicata con 1 la probabilità di default e caratterizzata dal colore arancione. Il colore blu identifica, invece, la probabilità di restare in bonis.

- Una tabella raffigurante le variabili più importanti e il valore che assumono per la controparte in esame. Le variabili rappresentate in arancione contribuiscono all'aumento della probabilità di default, mentre quelle in blu contribuiscono alla diminuzione della PD.

- Le variabili in ordine di importanza: in verde sono rappresentate le variabili che provocano un aumento della PD (rappresentate in arancione nella tabella precedente), mentre in rosso quelle che provocano una diminuzione della PD (raffigurate in blu nella tabella precedente).

Le variabili sono rappresentate in ordine di contributo marginale, il quale è rappresentato sull'asse delle ascisse.

Un esempio: la variabile "Pagamenti Pubblica Amministrazione / Totale Cash Out negli ultimi 6 mesi", che per la controparte in esame assume un valore di -0.96, ha un contributo positivo nell'aumento della probabilità di default, e tale contributo vale circa 1%.

La metodologia LIME, a differenza delle altre presentate in precedenza, può essere applicata più volte, per tutte le controparti che si desidera indagare. Si avrà quindi una rappresentazione di tutte le controparti esaminate, e qualora fosse necessario capire come si è arrivati a definire quel determinato livello di PD, grazie a questa metodologia, è possibile individuare le variabili più rilevanti e predittive.

È possibile applicare la tecnica LIME anche ad un nuovo campione e fornire le predizioni di quest'ultimo, utilizzando il modello già addestrato, effettuando dei test out of sample e out of time per verificare l'accuratezza dei risultati.

6 Conclusioni

L’adozione degli algoritmi di machine learning per lo sviluppo dei Sistemi di Rating Interni è aumentata in modo significativo negli ultimi anni. L'applicazione di queste tecniche offre numerosi vantaggi, tra cui una migliore accuratezza del modello, il superamento di carenze e incoerenze dei dati e la scoperta di nuove relazioni esistente tra essi. La scelta della tecnologia presenta anche nuove sfide, in particolare quelle incentrate sulla comprensione da parte dei supervisori o del consenso all'utilizzo di nuovi processi.

Il vantaggio principale dell’impiego degli algoritmi di machine learning è la possibilità di lavorare con un numero rilevante di variabili ed ottenere risultati più accurati rispetto a quelli ottenuti con i metodi tradizionali. Tuttavia, se da un lato aumenta l’accuratezza del modello, dall’altro aumenta la complessità relazionale tra le variabili impiegate e di conseguenza si rischia di avere un modello difficilmente governabile. Si potrebbe verificare, quindi, il rischio di avere un modello altamente performante ma non interpretabile e ciò potrebbe vincolare la Regulation ad approvarne l’utilizzo.

Grazie alle tecniche di interpretabilità implementate, è possibile avere una visione più chiara dei modelli machine learning:
- La metodologia SHAP permette di comprendere quali sono le variabili rilevanti, in base al loro contributo marginale e individuare quelle essenzialmente predittive e quelle, invece, che provocano solo rumore nel modello aumentandone la complessità.
- Le metodologie PDP e ICEBOX permettono di valutare l’impatto delle variabili sull’output del modello: la prima rappresenta l’output considerando anche l’effetto delle altre variabili presenti, mentre la seconda produce un output esclusivo del contributo della variabile in esame. È così possibile individuare le variabili più rilevanti e valutare anche un’integrazione tra più features.
- La metodologia LIME permette di individuare a livello di singola istanza (singola controparte) le variabili rilevanti per il modello nella definizione della probabilità di default. Si ha così una rappresentazione delle variabili che hanno influito maggiormente nella predizione dell’output. Per ogni controparte è possibile risalire quindi, a partire dalle variabili risultanti, ai fattori reali che hanno influenzato maggiormente la previsione.

Le tecniche di interpretabilità presentate, dato il caso in esame e la complessità del problema sono state adattate perfettamente all’algoritmo utilizzato. Esse sono quindi dipendenti dal modello e non sono versatili, cioè non sono utilizzabili per altri algoritmi. Tuttavia, la letteratura scientifica prevede l'uso di un modello indipendente dall'algoritmo originale in modo da avere una interpretazione oggettiva e, nonostante inizialmente sia stato scelto questo approccio, in seguito è stato accantonato, data la differenza decisionale esistente tra i modelli e di conseguenza tra i risultati ottenuti. È stato scelto, quindi, di applicare i tools di interpretabilità all'algoritmo originale in modo da avere una spiegazione chiara ed esclusiva del caso in esame.

L'applicazione degli algoritmi machine learning ai Sistemi di Rating Interni è soggetta all'approvazione da parte dell'autorità di Vigilanza, la quale ha come obiettivo primario quello di garantire la solidità degli indici patrimoniali delle banche e di conseguenza di tutelare i clienti del sistema bancario. Per tale ragione l'approvazione degli algoritmi machine learning è spesso vincolata al fatto che i suddetti modelli siano difficilmente interpretabili, o spiegabili, e che essi possano portare a risultati tali da compromettere la tutela dei clienti. Per avere, quindi, un modello interpretabile è stato scelto di utilizzare le tecniche di interpretabilità adattate all'algoritmo utilizzato per la formulazione delle predizioni.

Nell'ambito del rischio di credito, nel quale sempre più istituti bancari stanno applicando gli algoritmi di machine learning nella formulazione dei sistemi di scoring, è necessario non perdere di vista l'obiettivo che una banca persegue: valutare accuratamente e coerentemente la situazione reale delle proprie controparti. A tal proposito, per evitare di commettere errori che possano compromettere la valutazione delle controparti è necessario effettuare, prima dell'applicazione dei modelli, un'attenta fase di data cleaning e data preparation. Dopo questa fase è possibile applicare gli algoritmi e ottenere le previsioni delle probabilità di default delle controparti. Per ovviare al problema dell'interpretabilità dei risultati, inoltre, è possibile affiancare l'applicazione delle sopra citate tecniche per ottenere una migliore spiegazione e rappresentazione riguardo l'impatto delle variabili sia a livello di intero modello sia di singola controparte. Utilizzare, pertanto, algoritmi di machine learning e tecniche di interpretabilità in concomitanza ha il vantaggio di ottenere risultati più accurati e facilmente interpretabili e di comprendere meglio l'evoluzione dei modelli.

L'adozione delle tecniche di interpretabilità permette di spiegare meglio le logiche sottostanti agli algoritmi di machine learning. Essi, come già citato, sono considerati come "black-box" proprio perché sono di difficile interpretazione e spiegazione. Grazie a queste tecniche, ovviamente, non saranno di certo cancellati lo scetticismo o i dubbi riguardanti i temi del machine learning, ma da un loro affiancamento è possibile aumentarne la comprensione e influenzare positivamente l'adozione dei modelli su largo impiego per favorirne un maggiore sviluppo futuro.

Nel caso in esame è stato studiato solo il modulo cash-flow del modello di Rating SME Retail di Intesa Sanpaolo riuscendo a capire quali sono le variabili più impattanti nella stima delle probabilità di default e definendo quelle variabili determinanti a livello delle singole controparti.
Da un estensione dell'applicazione degli algoritmi di machine learning e delle tecniche di interpretabilità all'intero modello di Rating e dalla diffusione del convincimento generale sulla loro efficacia e adeguatezza si potrebbe ottenere un modello ancora più performante e soprattutto interpretabile, in linea con la definizione secondo la quale il rating deve essere "plausible and intuitive". Il Rating, infatti, è un elemento fondamentale in fase di concessione e il Gestore che lo utilizza deve comprenderne le logiche sottostanti. Inoltre, per le banche che utilizzano i modelli interni per il calcolo dei requisiti di capitale, è necessario fornire un'adeguata disclosure al Regulator affinché essi possano essere validati.

7 Bibliografia

1. **Ciby, Joseph.** *Advanced credit risk analysis and management.* s.l. : wiley Finance, 2013.

2. **Balthazar, Laurent.** *From Basel I to Basel III - The integration of state of the art Risk Modeling in banking Regulation.* s.l. : Palgrave Macmillan, 2006.

3. **Bernd Engelmann, Robert Rauhmeier.** *The Basel II risk parameters: estimation, validation stress testing - with application to loan risk management.* s.l. : Springer, 2011.

4. **Banca d'Italia.** *Metodo dei rating interni per il calcolo del requisito patrimoniale a fronte del rischio di credito.* Roma : s.n., 2006.

5. **Benli, Vahit Ferhan.** *A critical assessment of Basel II, Internal Rating Based approach.* s.l. : Haupt, 2010.

6. **Koffer, Timo.** *Basel III - Implications for Banks'capital structure .* Hamburg : Anchor Academic Publishing, 2014.

7. *Basel III: Finalising post-crisis reforms.* **Basel Committee on banking supervision.** s.l. : Bank for International Settlements, 2017.

8. **Langohr, Herwing Langohr Patricia.** *The Rating agencies and thei credit ratings: What they are, how they work and why they are relevant.* s.l. : Wiley Finance, 2008.

9. **Stefan Trueck, Svetlozar T. Rachev.** *Rating Based modeling of credit risk: theory and applications of migration matrices.* s.l. : Elsevier, 2009.

10. **Group Risk Management Intesa SanPaolo S.p.A.** *Risk Academy: Developing Compliant Rating Models .*

11. **Rezzani, Alessandro.** *Big data: architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati.* Milano : Maggioli Editore, 2013.

12. **Mauro, Andrea De.** *Big Data analytics: analizzare e interpretare i dati con il machine learning.* s.l. : Apogeo.

13. **Nielsen, Michael.** *Neural Networks and Deep Learning.* s.l. : Determination Press, 2015.

14. **Institute of International Finance (IIF).** *Machine Learning in credit risk report.* 2019.

15. **Molnar, Christoph.** Scope of interpretability. *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

16. —. Importance of Interpretability. *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

17. —. Taxonomy of Interpretability Methods. *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

18. —. Scope of Interpretability. *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

19. —. Evaluation of Interpretability. *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

20. —. Properties of Explanations. *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

21. —. Partial Dependence Plot (PDP). *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

22. **Alex Goldstein, Adam Kapelnery, Justin Bleichz, Emil Pitkinx.** *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation.* s.l. : The Wharton School of the University of Pennsylvania, 2014.

23. **Molnar, Christoph.** Individual Conditional Expectation (ICE). *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

24. —. Local Surrogate (LIME). *Interpretable machine learning: a guide for making black box models explainable.* s.l. : Leanpub, 2019.

25. **Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin.** *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* 2016.

26. **Molnar, Christoph.** The Future of Interpretability. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* s.l. : Leanpub, 2019.