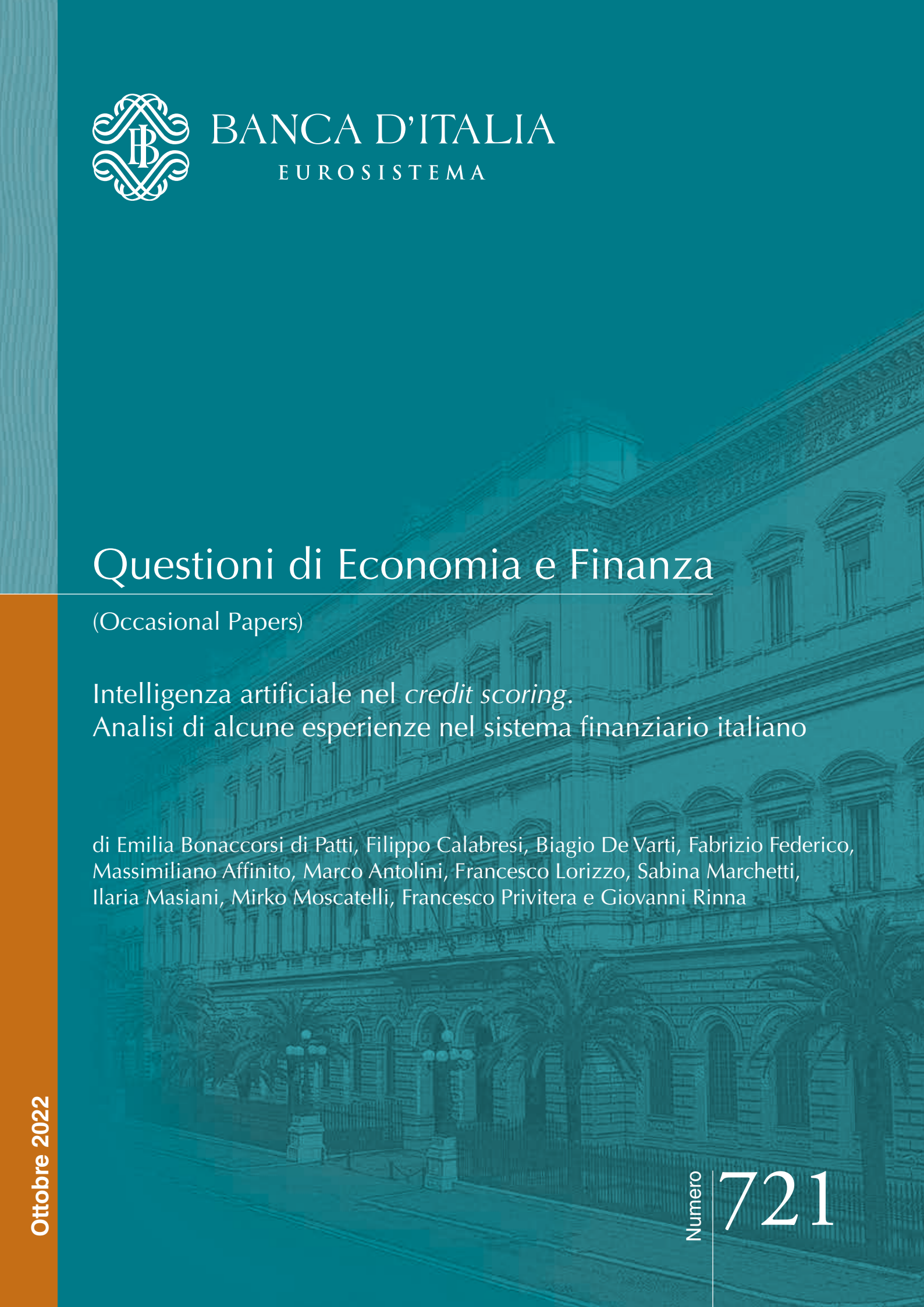
BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Intelligenza artificiale nel *credit scoring*. Analisi di alcune esperienze nel sistema finanziario italiano

di Emilia Bonaccorsi di Patti, Filippo Calabresi, Biagio De Varti, Fabrizio Federico, Massimiliano Affinito, Marco Antolini, Francesco Lorizzo, Sabina Marchetti, Ilaria Masiani, Mirko Moscatelli, Francesco Privitera e Giovanni Rinna



Ottobre 2022

Numero 721

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Intelligenza artificiale nel *credit scoring*. Analisi di alcune esperienze nel sistema finanziario italiano

di Emilia Bonaccorsi di Patti, Filippo Calabresi, Biagio De Varti, Fabrizio Federico, Massimiliano Affinito, Marco Antolini, Francesco Lorizzo, Sabina Marchetti, Ilaria Masiani, Mirko Moscatelli, Francesco Privitera e Giovanni Rinna

Numero 721 – Ottobre 2022

La serie Questioni di economia e finanza *ha la finalità di presentare studi e documentazione su aspetti rilevanti per i compiti istituzionali della Banca d'Italia e dell'Eurosistema. Le* Questioni di economia e finanza *si affiancano ai* Temi di discussione *volti a fornire contributi originali per la ricerca economica.*

La serie comprende lavori realizzati all'interno della Banca, talvolta in collaborazione con l'Eurosistema o con altre Istituzioni. I lavori pubblicati riflettono esclusivamente le opinioni degli autori, senza impegnare la responsabilità delle Istituzioni di appartenenza.

La serie è disponibile online sul sito www.bancaditalia.it .

ISSN 1972-6627 (stampa)
ISSN 1972-6643 (online)

Stampato presso la Divisione Editoria e stampa della Banca d'Italia

INTELLIGENZA ARTIFICIALE NEL *CREDIT SCORING*. ANALISI DI ALCUNE ESPERIENZE NEL SISTEMA FINANZIARIO ITALIANO

di Emilia Bonaccorsi Di Patti[*], Filippo Calabresi[**], Biagio De Varti[§], Fabrizio Federico[§§], Massimiliano Affinito[**], Marco Antolini[§], Francesco Lorizzo[§§], Sabina Marchetti[*], Ilaria Masiani[§], Mirko Moscatelli[*], Francesco Privitera[**], Giovanni Rinna[§]

Sommario

Il lavoro analizza il tema dell'utilizzo di tecniche di intelligenza artificiale e machine learning (AI-ML) a supporto della valutazione del rischio di credito da parte degli intermediari italiani. L'obiettivo è stato quello di verificare le modalità con le quali gli intermediari italiani fanno uso di queste tecniche nella selezione e gestione della clientela nei processi creditizi e il loro livello di consapevolezza circa i peculiari rischi che ne caratterizzano l'utilizzo. Partendo dall'analisi teorica delle determinanti concettuali e tecniche e del contesto normativo/istituzionale dell'AI-ML applicato al *credit scoring*, sono esposti i risultati di una verifica sul campo dell'esperienza maturata dagli intermediari italiani nell'adozione di modelli della specie.

Classificazione JEL: C52, G21, J15, J16, O32.
Parole chiave: intelligenza artificiale, machine learning, credito, *credit scoring*, distorsione, discriminazione.
DOI: 10.32057/0.QEF.2022.0721

Indice

[*] Banca d'Italia, Dipartimento Economia e statistica.
[**] Banca d'Italia, Dipartimento Tutela della clientela ed educazione finanziaria.
[§] Banca d'Italia, Dipartimento Vigilanza bancaria e finanziaria.
[§§] Banca d'Italia, Dipartimento Informatica.

1. Introduzione e principali conclusioni

Il lavoro analizza il tema dell'utilizzo di tecniche di intelligenza artificiale e machine learning (AI-ML) a supporto della valutazione del rischio di credito da parte degli intermediari italiani. L'obiettivo dell'analisi è stato quello di analizzare il livello di consapevolezza degli intermediari circa i rischi specifici che caratterizzano l'utilizzo di tecnologie avanzate nel delicato compito di selezione e gestione dei clienti maggiormente meritevoli di credito. In particolare, si è inteso valutare in che misura adottare modelli di AI-ML per il *credit scoring* comporti un incremento della possibilità che la selezione dei clienti risulti distorta rispetto all'effettiva rischiosità, configurando anche possibili forme di discriminazione, e quanto sia complesso per gli intermediari indagare su tale eventualità a fronte delle difficoltà di ricostruire le logiche seguite dai modelli.

Il lavoro si basa su due modalità di approccio: analisi teorica delle determinanti concettuali e tecniche e del contesto normativo/istituzionale dell'AI-ML applicato al *credit scoring* e verifica sul campo dell'esperienza maturata dagli intermediari italiani nell'adozione di modelli della specie.

Nella parte teorica si è delineato il contesto tecnico/scientifico, normativo e di policy in cui le analisi in materia si stanno svolgendo; è possibile riassumere le principali conclusioni nei punti seguenti:

- *"si diffondono sistemi di apprendimento automatico e tecniche di spiegabilità*" - il capitolo 2 descrive in cenni le determinanti dei profili tecnologici dell'intelligenza artificiale rilevanti per il problema in esame, ovvero il panorama dell'apprendimento automatico, l'utilizzo dei big data e il ricorso a tecniche di *explainable AI*; se ne ricava la convinzione di come la disponibilità crescente di dati e l'affinamento della modellistica per l'apprendimento automatico comportino un livello di complessità tale da richiedere che il progresso della tecnologia si accompagni a forme di presidio che assicurino la necessaria trasparenza sulle logiche seguite;
- "*il trade-off fra accuratezza delle stime e spiegabilità*" - il capitolo 3 analizza in breve gli studi scientifici che mettono a confronto, nella valutazione del merito di credito, l'utilizzo di tecniche di ML rispetto agli approcci statistici tradizionali; in sintesi, il ML porta a stime più accurate, anche perché è in grado di valorizzare informazioni e relazioni tra le stesse dal non immediato significato economico;
- "*i sistemi di ML non imparano necessariamente in modo corretto"* - il capitolo 4 – dopo aver delineato le possibili forme di distorsione (bias) dei sistemi di *credit scoring* (non corretta differenziazione del rischio e discriminazione nei confronti di individui o gruppi sociali) – analizza in dettaglio i bias che possono manifestarsi nei sistemi di AI-ML applicati in questo campo; analogamente ai sistemi statistici tradizionali, queste distorsioni possono verificarsi nelle fasi di raccolta dei dati, nella specificazione del modello e nell'analisi degli output, ma nei sistemi a apprendimento automatico la presenza di bias tende a generare un pericoloso ciclo di feedback in cui la distorsione può essere confermata e rinforzata;

- "*la normativa prudenziale in essere consente di presidiare larga parte dei rischi associati ai modelli di AI-ML"* - il paragrafo 5.1 presenta la normativa prudenziale rilevante per gli intermediari che adottino sistemi di AI-ML per la gestione del credito; non essendo previste prescrizioni specifiche, valgono le previsioni generali sull'efficacia dei meccanismi di governo, gestione e controllo del rischio contenute nella normativa di vigilanza e, per le banche, nelle EBA Guidelines su concessione e monitoraggio del credito; in caso di modelli di ML utilizzati ai fini del calcolo dei requisiti patrimoniali, viene in rilievo la regolamentazione specifica sugli stessi. Anche in assenza di prescrizioni specifiche per i modelli AI-ML, la compiuta applicazione dei principi generali contenuti nelle norme dovrebbe consentire di mitigare larga parte dei rischi specifici di tali modelli;
- "*la non discriminazione nei rapporti con la clientela è un principio non compiutamente declinato nelle norme"* - il paragrafo 5.2.1 mostra che nelle disposizioni di trasparenza sono contenuti pochi riferimenti di carattere generale al principio della non discriminazione; la ricerca del giusto equilibrio fra correttezza nei rapporti con la clientela e libertà di impresa degli intermediari in materia di concessione del credito riceve nuove sollecitazioni con l'adozione di tecniche AI-ML;
- "*nuove sfide dal GDPR sul fronte dell'informativa ai clienti selezionati con ricorso al ML"* - il paragrafo 5.2.2 presenta la rilevanza attribuita dalla normativa europea in materia di privacy (GDPR) ai temi connessi con la profilazione, il consenso del cliente e il diritto a un'informazione rafforzata in caso di processo decisionale automatizzato; l'utilizzo di tecniche AI-ML rende più complesso per gli intermediari rappresentare al cliente che ne fa richiesta la logica seguita dal modello nel prendere o proporre una decisione di mancato affidamento;
- "*convergenza internazionale sui principi ma difficile traduzione in norme e prassi concrete"* - il paragrafo 5.3 riassume alcune valutazioni espresse di recente da autorità e organismi nazionali e internazionali in materia di governo dell'AI; ampia è la convergenza sui principi da salvaguardare, quali l'imprescindibilità del controllo umano sulla macchina, la trasparenza e la non discriminazione, mentre appare meno agevole la traduzione di questi principi in norme, nella sfida di conciliare promozione dell'innovazione e tutela dei diritti dei clienti.

La verifica sul campo si è basata sul confronto con l'esperienza concreta degli intermediari italiani che utilizzano tecniche AI-ML per il *credit scoring* al fine di verificare lo stato di sviluppo del comparto, le soluzioni adottate, i benefici attesi, la consapevolezza dimostrata nell'affrontare i nuovi rischi e i presidi assunti per farne fronte. Il capitolo 6 riporta in dettaglio le principali evidenze emerse nella verifica empirica effettuata con gli intermediari selezionati. Di seguito le principali conclusioni:

a. "*mercato in crescita*" - il fenomeno del ML applicato al *credit scoring* è ancora limitato ma in crescita, trainato da attese di maggior accuratezza nelle stime;
b. "*concentrazione nelle scelte tecnologiche*" - in un panorama di soluzioni tecnologiche praticabili ampio, le scelte degli intermediari si concentrano su pochi modelli di ML e

tecniche di spiegabilità che presentano i minori costi di implementazione, affermandosi come standard di mercato;

c. "*progressi evidenti nei test di accuratezza*" - le evidenze da letteratura relative a una maggiore accuratezza nelle stime dei modelli di ML rispetto ai modelli statistici tradizionali sono confermate dall'esperienza degli intermediari selezionati;
d. "*qualche passo verso una maggiore inclusione finanziaria*" – gli intermediari cominciano a sfruttare il potenziale in termini di leva inclusiva che i sistemi di AI-ML possono esprimere accedendo a fonti dati alternative e sviluppando modelli in grado di trattarli più efficacemente: si osservano prime esperienze di allargamento dell'offerta di credito a fasce di clienti, tradizionalmente esclusi per la loro scarsa storia creditizia;
e. "*poco diffuse le analisi sulla distorsione da discriminazione*" - l'adozione di tecniche di analisi e di eventuale mitigazione della distorsione derivante da discriminazione è risultata ancora limitata in quanto il rischio di discriminazione è percepito come remoto e non attuale;
f. "*percezione di molti benefici e pochi rischi*" – Il feedback degli intermediari mostra un generale ottimismo: l'utilizzo delle tecniche di AI comporterebbe molti benefici e pochi rischi incrementali rispetto alle tecniche tradizionali;
g. "*governance e controlli poco tarati rispetto all'utilizzo delle tecniche AI*" - i meccanismi di governance relativi ai modelli di ML appaiono in molti casi poco focalizzati sui nuovi rischi connessi con l'adozione di tecniche AI-ML anche quanto a adeguatezza della reportistica, a coinvolgimento e formazione del personale, a gestione del - talvolta assai ampio - ricorso all'outsourcing.

2. L'Intelligenza artificiale e le principali tecniche in uso

Il termine "Intelligenza artificiale" (di seguito per brevità AI, acronimo di *Artificial Intelligence*), coniato negli anni '50, identifica in generale diverse teorie, metodologie e tecniche che consentono di progettare soluzioni informatiche in grado di riprodurre a vario titolo l'intelligenza umana. Questo ramo dell'informatica si è dovuto confrontare, negli anni, con alcune intrinseche limitazioni dell'evoluzione tecnologica, in termini di capacità di calcolo, di elaborazione di grandi quantità di dati, di maturità degli algoritmi. La rapida evoluzione del comparto tecnologico ha reso progressivamente disponibili le necessarie capacità elaborative in grado di sostenere gli algoritmi di intelligenza artificiale, che quindi – grazie anche all'ampia disponibilità di dati a disposizione degli stessi – hanno iniziato a rilasciare le loro potenzialità nel tessuto produttivo (si pensi ad es. all'applicazione di tecniche di *image detection* nelle filiere produttive ai fini di controllo di qualità) ma anche nella sfera individuale (ad es. con l'ampia diffusione di interfacce conversazionali, quali *chatbot, voicebot* o assistenti virtuali, anche su dispostivi mobili)[1].

[1] Per una ricognizione dell'uso dell'AI nel settore bancario cfr. CIPA-ABI (2021), *Rilevazione sull'IT nel settore bancario italiano.*

Queste tecniche si differenziano significativamente dall'approccio tradizionale della programmazione, nel quale un programmatore definisce un algoritmo che sia in grado, attraverso la codifica di una serie di operazioni, di trasformare deterministicamente dei dati di input nei dati di output desiderati. Nell'AI, invece, anziché concentrarsi sulla formalizzazione di un algoritmo deterministico si costruisce un modello (secondo un approccio induttivo o deduttivo – cfr. *infra*) in grado di catturare le informazioni necessarie per derivare "conoscenza" in modo automatico a partire dai dati disponibili. In questo ambito, le competenze di sviluppo delle applicazioni vengono integrate con quelle tipiche del profilo di *data scientist*, figura professionale orientata all'analisi dei dati e con significative conoscenze del dominio di business.

L'AI si può sinteticamente ripartire in due differenti approcci, da cui derivano altrettante tecniche algoritmiche: l'approccio **induttivo** e l'approccio **deduttivo**. Nell'approccio induttivo, la macchina sintetizza la propria conoscenza sulla base dell'osservazione empirica dei dati, imparando da questi tramite un processo di generalizzazione; in quello deduttivo, a partire da una rappresentazione formale della conoscenza operata tramite linguaggi di *knowledge representation and reasoning* (KRR), la macchina produce nuova conoscenza dai dati in input in un processo di inferenza. L'approccio induttivo è tipicamente conosciuto come apprendimento automatico, o **Machine learning (ML)**, l'approccio deduttivo è noto come ragionamento automatico, o **Automated reasoning (AR)**.

Nella figura e nel box seguenti vengono rappresentate in modo sintetico e con delle esemplificazioni le principali differenze tra l'approccio induttivo e quello deduttivo.

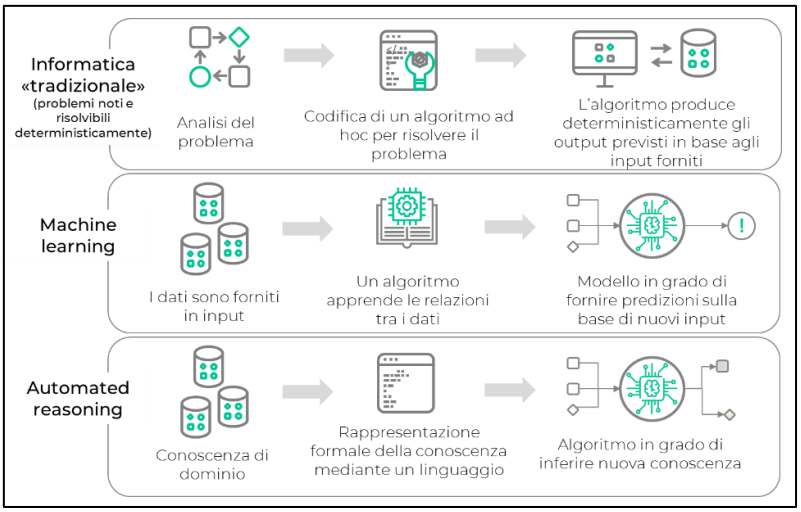


Figura 1. Confronto tra approcci nello sviluppo di algoritmi deterministici ad hoc e l'adozione di tecniche di AI.

Un esempio di applicazione dei due approcci all'AI

Gli approcci dell'AI discussi riflettono un diverso orientamento alla soluzione dei problemi. Per illustrarne la differenza, è conveniente considerare un cosiddetto problema di decisione: vogliamo cioè affidare alla macchina la responsabilità di decidere in merito ad una certa situazione (ad esempio svoltare ad un'uscita autostradale o proseguire il viaggio; concedere o negare un credito) non nota a priori.

Adottando un metodo induttivo, la macchina, precedentemente addestrata su un ampio numero di coppie problema-decisione (ad esempio un insieme di scelte di viaggio o di decisioni su specifiche richieste di credito), sarà in grado di prendere la decisione astraendo dai casi concreti analizzati, ad esempio grazie a modelli statistici in grado di approssimare le dinamiche della decisione sulla base delle caratteristiche della nuova situazione.

Con un metodo deduttivo, invece, la macchina, precedentemente istruita sulle dinamiche del dominio di interesse (come ci si comporta in autostrada all'approssimarsi della destinazione o secondo quali criteri si concede o si nega un credito), applicherà al caso specifico le nozioni apprese, fornendo la risposta.

2.1 Il machine learning e i diversi approcci all'apprendimento automatico

Nel machine learning esistono diversi approcci all'apprendimento automatico, che determinano le caratteristiche degli algoritmi e i requisiti richiesti (ad es. in termini di numerosità dei dati necessari per l'addestramento). Gli approcci più noti sono i cosiddetti supervisionato, non supervisionato, semi-supervisionato, con rinforzo, nei quali rispettivamente:

- ***supervisionato***: l'algoritmo apprende il modello delle relazioni tra gli input e gli output mediante un insieme di dati preventivamente etichettati da un essere umano (cosiddetto *labeled dataset*);
- ***non supervisionato***: l'algoritmo apprende autonomamente il modello dal *dataset*, senza necessità che questo sia anticipatamente processato per l'attribuzione delle *label*;
- ***semi-supervisionato:*** caso intermedio tra i due precedenti, dove il *dataset* è solo parzialmente etichettato (ad es. perché il costo di *labeling* di un *dataset* è solitamente alto);
- ***con rinforzo:*** l'algoritmo compie azioni in modo da massimizzare progressivamente una funzione di profitto, che assegna un valore positivo o negativo ad ogni azione. È una tecnica che si potrebbe definire "*trial and error*".

L'approccio supervisionato viene spesso usato nei problemi di classificazione, dove si beneficia di *dataset* già provvisti di una corretta assegnazione in classi che l'algoritmo può apprendere ed applicare sui nuovi dati. Queste tecniche riescono a produrre modelli efficaci con volumi di dati tendenzialmente inferiori rispetto all'approccio non supervisionato, usato quando si vuole apprendere la struttura inerente dei dati senza partire da una conoscenza formalizzata degli stessi; tipica applicazione è il *clustering*, dove l'algoritmo è in grado di partizionare i dati in insiemi contenenti informazioni simili tra loro (ad es. raggruppando immagini di volti separatamente da immagini di gatti).

Si citano nel seguito due famiglie di algoritmi, applicabili trasversalmente negli approcci sopracitati, particolarmente significativi per la loro diffusione o le loro intrinseche caratteristiche[2]:

- **Ensemble learning***:* Le tecniche di *ensemble learning* prevedono, per lo svolgimento dell'esercizio di previsione, l'utilizzo di insiemi di modelli: la previsione finale è poi generalmente ottenuta come la previsione media o maggioritaria dei modelli. Un caso molto diffuso è quello in cui i singoli modelli sono alberi decisionali (cfr. esempio in figura 2) che operano tramite una partizione progressiva dello spazio delle soluzioni. Le tecniche possono essere suddivise in due sotto-categorie basate su *bagging* (apprendimento simultaneo di modelli indipendenti tra loro, ciascuno caratterizzato dallo sfruttamento di porzioni dell'informazione complessiva) e su *boosting* (sequenze di modelli che vanno

[2] Meno diffuse appaiono le applicazioni di ML basate su: i) Algoritmi Evolutivi, processi di apprendimento dinamico per la ricerca di soluzioni ottimali all'interno di popolazioni di parametri; e ii) *Federated learning*, insieme di tecniche per l'apprendimento in un contesto di collaborazione tra parti, secondo un approccio distribuito.

progressivamente a raffinare il processo di apprendimento). Due popolari esempi per gli approcci citati, nel contesto dell'apprendimento supervisionato, sono, rispettivamente, il *random forest* e il *gradient boosting*.

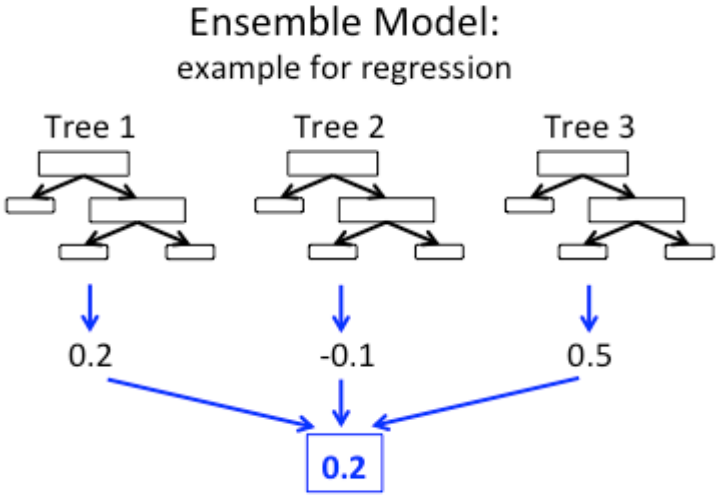


Figura 2. Esempio di *ensemble learning* con *random forest* (con tre alberi e predizione basata sulla media degli output)

- **Deep learning**: si tratta di una famiglia di algoritmi di ML i cui processi elaborativi, ispirandosi al comportamento dei neuroni del cervello umano, si basano su reti che interconnettono nodi organizzati in livelli successivi (cosiddette reti neurali). Ad ogni livello della rete corrisponde una fase dell'apprendimento di concetti via via sempre più complessi.

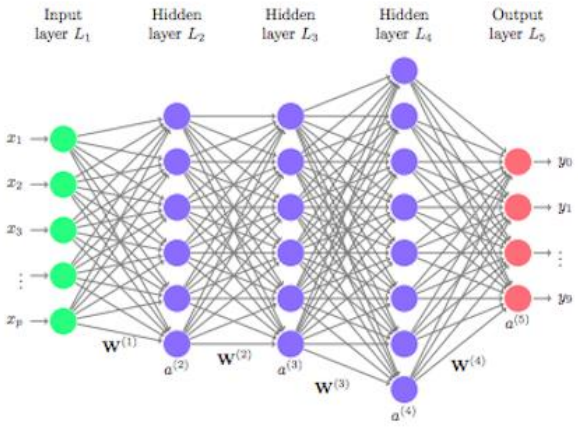


Figura 3. Esempio di rete neurale.

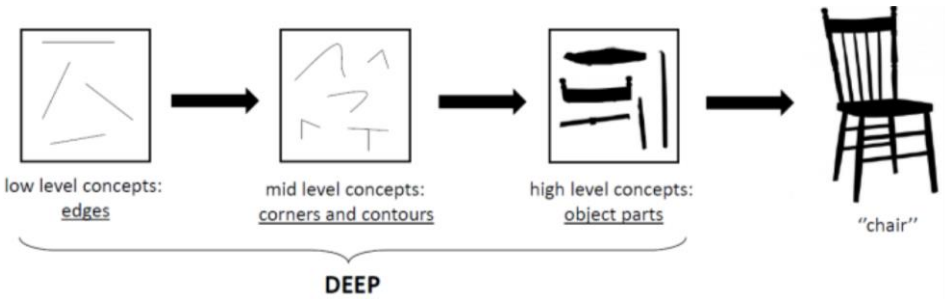


Figura 4. Esempio di apprendimento progressivo di concetti in un approccio *deep learning*.

2.2 Big data, analytics e AI

Il dominio dell'intelligenza artificiale spesso si accomuna all'uso dei cosiddetti **big data**, altrettanto rilevante ma che si pone su un piano sostanzialmente diverso pur presentando punti di contatto. Con il termine big data solitamente si individuano sia i dati contraddistinti da alcune caratteristiche descritte nel seguito, sia l'insieme di algoritmi, tecnologie e soluzioni informatiche in grado di offrire servizi di raccolta, gestione e analisi degli stessi. Il paradigma associato ai big data è quello delle cosiddette "5 V", che individuano tre caratteristiche distintive e due prassi attese:

- **Volume**: i dati disponibili per attività di analisi spesso si attestano nell'ordine dei *terabyte* o superiori;
- **Varietà**: ai più tradizionali dati di natura strutturata, in questo paradigma si affiancano anche quelli semi-strutturati (come file XML) o non strutturati (come documenti testuali o immagini);
- **Velocità**: i dati sono prodotti a ritmi estremamente elevati, pertanto occorre dotarsi di tecnologie in grado di processarli con adeguata velocità ovvero tecniche adeguate alla loro analisi in tempo reale;
- **Veridicità**: poiché i dati possono essere affetti da inattendibilità, dovuta alla natura dei processi di generazione e di raccolta delle osservazioni, occorre fare in modo che gli stessi rappresentino quanto più possibile fedelmente la realtà sottostante;
- **Valore**: occorre essere in grado di trasformare il dato in informazione utile al business.

Pertanto, si definiscono big data insiemi di osservazioni che presentino almeno una tra le caratteristiche di alto volume (nel numero di osservazioni o nel numero di attributi), alta varietà (di contenuto o formato) e velocità di produzione o raccolta, tali da implicare il ricorso a strumenti e tecniche non tradizionali. A seconda dei casi, possono afferire alla categoria di big data sia grandi volumi di transazioni e pagamenti, connotati da alta granularità di informazioni, che dati in formato testuale, come ad esempio le causali di spesa o bonifico, come pure i dati di fonte social network e quelli connessi alla navigazione su internet.

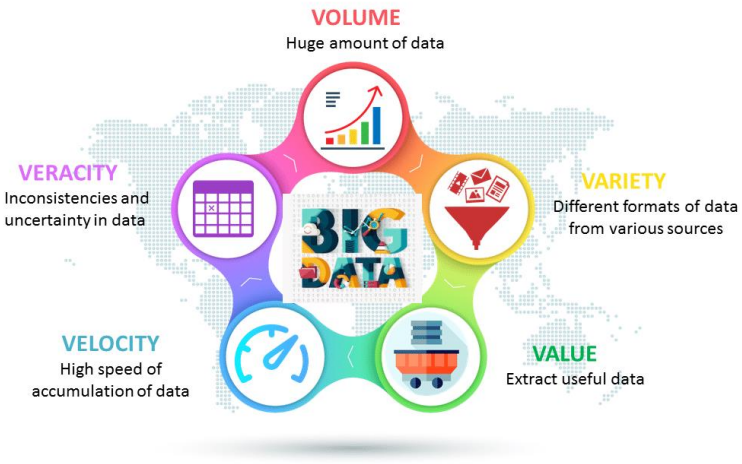


Figura 5. Le cinque "V" dei big data.

Il legame tra il dominio dei big data e le tecniche di AI è piuttosto immediato: la disponibilità di insiemi di dati di dimensioni molto rilevanti, e delle tecnologie per elaborarli in modo efficiente, consente ad esempio agli algoritmi di apprendimento automatico di "imparare" meglio. Tuttavia, è possibile addestrare tali algoritmi anche in assenza dei big data (a patto ovviamente di disporre di *dataset* di dimensioni e caratteristiche comunque adeguate allo scopo), pertanto la presenza di big data non si prefigura come un prerequisito per l'applicazione di tecniche di AI. Neanche la prospettiva complementare prevede un vincolo di dipendenza: esistono difatti numerose tecniche (che ricadono nel dominio a cui ci si riferisce comunemente col termine *analytics*) applicabili ai big data che non prevedono l'impiego di algoritmi di AI. I due domini, pertanto, hanno un rapporto di complementarietà, beneficiando reciprocamente dei progressi compiuti.

2.3 Explainable AI (XAI)

L'applicazione di tecniche di AI pone all'attenzione alcuni aspetti che tipicamente non si affrontano nello sviluppo di soluzioni informatiche tradizionali: tra questi, assume particolare rilievo la spiegabilità dei risultati prodotti dall'algoritmo, che si pone alla base della capacità, da parte dell'analista, di illustrare agli *stakeholder* del processo le motivazioni a supporto delle decisioni prese (o suggerite) dall'algoritmo.

Con il termine *eXplainable AI* (abbreviato in XAI) si indica l'insieme di strumenti applicati dall'analista volti ad integrare il risultato primario dell'algoritmo di ML con un insieme di interpretazioni e spiegazioni (*explanation*) circa il funzionamento del modello.
Da un punto di vista concettuale, l'XAI può riferirsi a due approcci principali:

- **interpretabilità** (*interpretability*), ovvero tracciare quantitativamente i meccanismi che governano il comportamento del modello;
- **spiegabilità** (*explainability*), ovvero formulare valutazioni qualitative (giustificazioni) sui risultati ottenuti, con l'obiettivo di spiegare le logiche di funzionamento del modello.

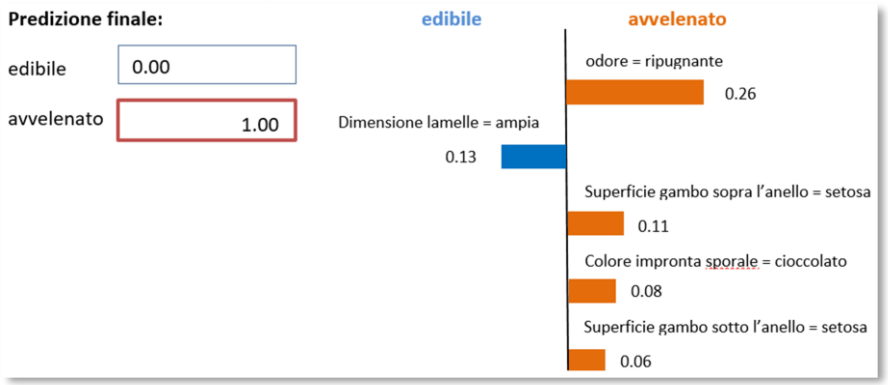
Il grado di spiegabilità intrinseca dei diversi algoritmi che compongono il panorama delle tecniche di machine learning tende di norma a diminuire con l'aumento della capacità previsiva, presentando un rapporto che in generale è di proporzionalità inversa: ad esempio le reti neurali offrono performance molto elevate e possono essere rese in qualche misura spiegabili solo adottando

tecniche XAI; sul fronte opposto un modello intrinsecamente spiegabile, come l'albero decisionale, presenta di norma minore capacità previsiva.

Su un piano realizzativo, le tecniche di XAI si distinguono tra quelle che possono essere applicate solo a specifiche classi di modelli (la cosiddetta *model-specific explainability*), come la *feature importance* degli alberi decisionali, e quelle applicabili a qualsiasi modello a prescindere dalla sua forma funzionale (cosiddetta *model-agnostic explainability*), come SHAP e LIME (cfr. riquadri).

Shapley Additive exPlainations (SHAP): tecnica di *explainabilty model-agnostic* basata sui cosiddetti *Shapley values*, usati nella teoria dei giochi per determinare quanto ogni giocatore (in un contesto collaborativo) abbia contribuito all'*output* prodotto complessivamente dal gruppo. Questi valori sintetizzano il contributo delle *feature* (proprietà individuali misurabili) del modello sull'output dello stesso, come nel grafico sottostante:

LIME (Local Interpretable Model-Agnostic Explainations): tecnica di *explainabilty model-agnostic* che intende spiegare il comportamento del modello mediante la variazione dei valori delle *feature* e l'osservazione degli impatti sugli *output* (in termini di contributo che ogni *feature* apporta alla predizione finale):



L'implementazione di tecniche di XAI può essere radicata in fasi diverse del ciclo di vita dei modelli e più precisamente: prima dello sviluppo del modello, con riguardo alla comprensione dei dati usati per addestrarlo (*pre-modeling explainability*); durante lo sviluppo del modello, per favorirne l'interpretabilità (*explainable modeling*); a valle dello sviluppo, per spiegare a posteriori la logica decisionale del modello (*post-modeling, o post hoc explainability*).

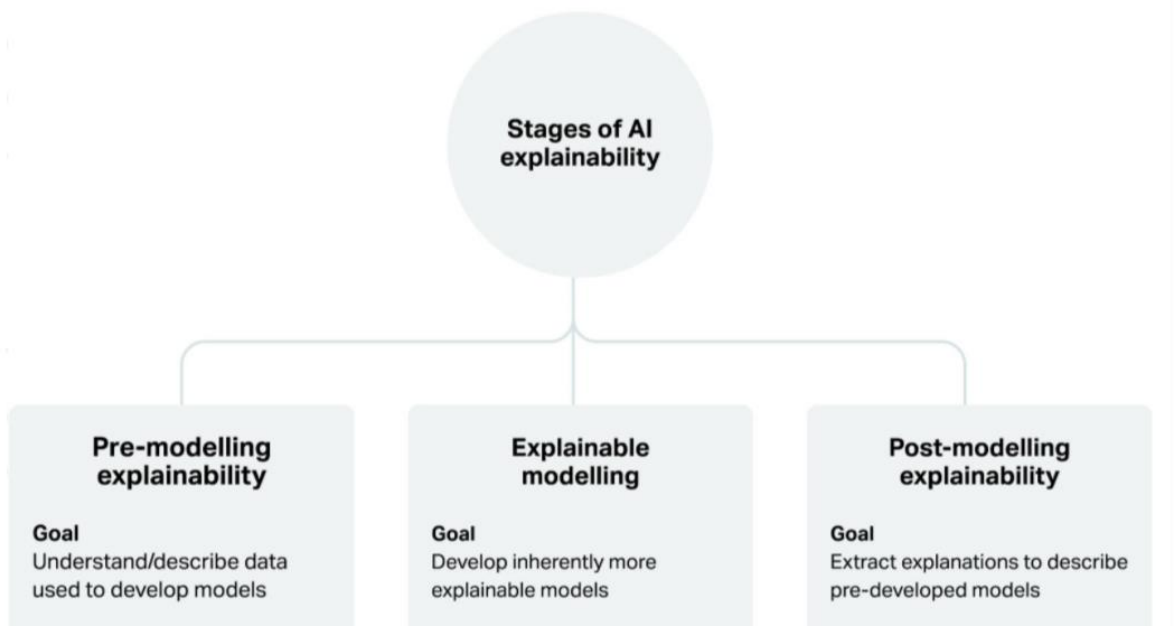


Figura 6. Differenti fasi di applicazione delle tecniche di XAI.

3. **L'AI nella valutazione del merito creditizio: benefici e rischi, presidi**

3.1 Il machine learning *applicato alla valutazione del merito creditizio*

Uno degli ambiti in cui sono state esplorate le proprietà delle tecniche di ML è la valutazione del rischio di credito. Rispetto all'approccio statistico tradizionale, basato sulla stima econometrica della probabilità di default (ad esempio mediante modelli logistici), si evidenziano alcune differenze fondamentali. Nei modelli econometrici la determinazione delle variabili rilevanti e la specificazione delle forme funzionali delle relazioni vengono generalmente guidate dalla teoria economica, mentre nel ML gli algoritmi selezionano in maniera automatica le variabili rilevanti identificando relazioni anche non lineari e di difficile interpretazione tra le stesse.

Numerosi studi hanno confrontato la capacità delle tecniche di ML di prevedere il default delle imprese e dei privati con quella di approcci più tradizionali. Questi studi mostrano che l'accuratezza del modello di ML nell'individuare le insolvenze è generalmente migliore rispetto a quella dei modelli econometrici (Fantazzini and Figini, 2009; Khandani et al., 2010; Kruppa et al., 2013; Yuan, 2015; Barboza et al., 2017; Bachman and Zhao, 2017; Fuster et al., 2020; Albanesi and Vamossy, 2019; Moscatelli et al., 2020). Il miglioramento in termini di accuratezza delle tecniche di ML deriva in primo luogo dall'estensione della gamma di forme funzionali e relazioni tra le diverse variabili valutate dal modello. Una recente survey mostra che l'utilizzo di tecniche di ML avanzate comporta un miglioramento delle previsioni del default rispetto ai modelli statistici tradizionali per lo più compreso tra 2 e 10 punti percentuali ma in alcuni casi anche superiore (Alonso and Carbó, 2020)[3].

Un altro beneficio del ML è quello di gestire ed elaborare grandi quantità di dati in volume (numero di osservazioni) e ricchezza (numero di variabili, tipi di dato), sfruttando l'aumentata potenza di

[3] Un recente studio su dati italiani (Moscatelli et al. (2020)) mostra che l'entità del miglioramento in termini di accuratezza rispetto a modelli logistici varia in ragione del tipo di modello e del numero di variabili considerate.

calcolo resasi disponibile negli anni più recenti. In generale, nella valutazione del merito di credito si è verificata una graduale estensione delle fonti di dati utilizzate (fenomeno che ha invero riguardato sia i modelli econometrici che le tecniche di ML): dai dati finanziari strutturati (indicatori patrimoniali ed economico-finanziari, sull'andamento dei conti e dei pagamenti, di mercato) a dati non finanziari strutturati (dati di tipo socio-demografico ottenuti anche da fonti terze), dati non strutturati finanziari (analisi delle informazioni transazionali e di quelle derivate da open banking) e dati non strutturati non finanziari (dati di navigazione, *digital footprint*[4], informazioni conferite sui *social network*) (Tobback and Martens, 2019; Óskarsdóttir et al., 2019; Berg et al., 2020; Roa et al., 2021).

Il ricorso a modelli di ML, sfruttando le citate fonti di dati alternative, permette di considerare variabili che non hanno una chiara interpretazione/relazione economica, che non verrebbero considerate in un modello tradizionale, grazie al peculiare processo di addestramento e calibrazione di questi modelli. Lo sfruttamento di informazioni alternative ovvero complementari a quelle in uso in modelli tradizionali può aumentare l'accuratezza previsiva e al contempo estendere la possibilità di valutare il merito di credito di soggetti altrimenti esclusi o penalizzati nell'accesso al credito. Inoltre, può accrescere la concorrenza in questo mercato poiché, sfruttando le nuove fonti dati, anche soggetti diversi dagli intermediari finanziari tradizionali potrebbero offrire finanziamenti a costi più bassi, con tempi di approvazione più brevi, richiedendo minori garanzie e documentazione (Jagtiani and Lemieux, 2017; Bazarbash, 2019).

Dal punto di vista dei rischi, le applicazioni di ML sono caratterizzate da un trade-off che richiede di bilanciare l'incremento dell'accuratezza con il costo in termini di minore spiegabilità del modello. Inoltre, l'utilizzo del ML nella selezione della clientela può comportare rischi legali e reputazionali derivanti dall'opacità dei processi e dei meccanismi che consentono di raggiungere una migliore performance. Il ricorso a metodi di XAI può rendere meno stringente il trade-off tra spiegabilità e accuratezza (Cascarino et al., 2022).

Infine, gli algoritmi di ML possono sì produrre modelli estremamente precisi nella previsione della variabile di interesse all'interno del campione utilizzato, ma potrebbero non essere generalizzabili a contesti diversi da quello in cui sono stati elaborati, nei quali potrebbero avere scarsa capacità predittiva (cosiddetto rischio di *overfitting*). L'affidabilità di un modello di ML deve essere verificata in base a requisiti di validità esterna, ovvero di stabilità della qualità predittiva per popolazioni (validità di popolazione), contesti, luoghi o tempi (validità ecologica) diversi da quelli considerati in fase di sviluppo.

3.2 L'automated reasoning applicato alla valutazione del merito creditizio

Gli approcci di AR applicato alla valutazione del merito creditizio si basano sulla modellazione formale della conoscenza posseduta dall'esperto di business tramite formalismi logico-matematici,

[4] Definita come "insieme unico di attività, azioni, contributi e comunicazioni digitali tracciabili manifestati su Internet o su dispositivi digitali". Fonte: Wikipedia.

spesso in grado di trattare congiuntamente aspetti quantitativi e qualitativi. I sistemi informatici per l'AR automatizzano poi l'applicazione di tale conoscenza alle singole istanze di richiesta di credito.

Esistono una serie di studi che hanno storicamente definito teoria e pratica di utilizzo dell'AR per valutare il merito creditizio. Gli elementi tecnico-teorici di base sono stati definiti già con l'adozione dei primi sistemi esperti (ad es. Zocco (1985)) e hanno efficacemente tracciato il legame tra la decisione di concedere un credito e le tecniche di programmazione logica (Iwasieczko et al. (1986)). Agli studi teorici si sono presto affiancati sistemi dedicati, spesso nati con lo specifico obiettivo di ottimizzare prestazioni (*throughput*) e accuratezza nella concessione del credito. In letteratura è noto il caso di un primario operatore di carte di credito, tra i primi a dotarsi di un sistema basato su AR per la valutazione in tempo reale delle richieste di credito "inusuali" da parte di possessori di carte. Con tale approccio, si è dimostrato di poter incrementare sensibilmente l'accuratezza delle decisioni, precedentemente manuali (Piketty (1987)).

In contesti più recenti, le tecniche di AR vengono utilizzate per la concessione di credito in settori specifici, dove la conoscenza dell'ambito di business è altamente specializzata[5].

La recente introduzione di linguaggi di rappresentazione della conoscenza più maturi e semplici, quali quelli adottati per i *Knowledge Graph* (Gottlob et al. (2015), Hogan (2021), Bellomarini et al. (2018)) e l'adozione di hardware più efficiente hanno portato ad un crescente interesse per l'applicazione del ragionamento automatico in ambito finanziario. I *Knowledge Graph* in particolare sono adottati per la valutazione del merito di credito in molteplici contesti di ricerca (Beydoun et al. 2020).

4. I problemi di non corretta differenziazione e discriminazione nella valutazione del merito di credito

4.1 Definizioni

Un tema ampiamente discusso nella letteratura relativa alle applicazioni di algoritmi per la classificazione di unità o individui rispetto a una variabile di interesse è quello della possibile presenza di distorsioni. Tali distorsioni nell'ambito della valutazione del merito di credito possono dare luogo a fenomeni di non corretta differenziazione del rischio e di discriminazione del potenziale cliente.

In generale, per non corretta differenziazione del rischio si intende la mancata capacità di un modello di ordinare correttamente i clienti per livello di merito di credito, non consentendo di conseguenza il raggiungimento di un livello ottimale di efficienza allocativa delle risorse. L'utilizzo di un modello che non differenzia in maniera appropriata il rischio può condurre a distorsioni sia nella scelta dei clienti da affidare, sia nel *pricing*.

La discriminazione, in termini generali, indica la presenza di pregiudizi o favoritismi nei confronti di specifici individui, o gruppi sociali, identificati da attributi considerati "sensibili" (Mehrabi et al.,

[5] Un caso interessante in questo senso è costituito da ALEES (Bryant (2001)), un sistema di AR per la concessione del credito ad operatori dell'agricoltura, sviluppato dalla Griffith University in Australia.

2021). Essa si determina quando sono presenti distorsioni nei processi decisionali che determinano svantaggi relativi ai danni di un soggetto o di una categoria. La discriminazione può essere diretta o indiretta. Nel primo caso, lo svantaggio è determinato da processi decisionali o condotte che tengono in considerazione attributi che identificano dei soggetti come vulnerabili. Nel secondo caso l'individuo non viene identificato esplicitamente da attributi "sensibili" ma si realizza *de facto* una condizione di disuguaglianza, ad esempio tramite l'uso di variabili correlate con l'appartenenza al gruppo vulnerabile (Hajian et al., 2012)[6]. La discriminazione operata da modelli tradizionali o di ML può essere inoltre riconducibile alla scelta di considerare esplicitamente determinate caratteristiche (*disparate treatment*), ovvero risultare da forme di distorsione algoritmica che sfuggono al controllo dell'analista ed essere pertanto involontaria (*disparate impact*).

Al fine di valutare quantitativamente e prevenire il verificarsi di forme di discriminazione, è possibile ricorrere alle definizioni di *fairness* disponibili in letteratura (Dwork et al., 2012). Il ricorso al concetto di *fairness* consente di misurare e valutare qualitativamente l'assenza di meccanismi discriminatori all'interno di un sistema complesso. Al riguardo, il comportamento di un modello può essere esaminato a livello aggregato adottando una definizione di *fairness* di gruppo (ad es. comunità o categorie vulnerabili) o individuale. Tra le definizioni maggiormente diffuse di *fairness* di gruppo rientrano: i) *fairness through unawareness*, che consiste nello scartare dalla base dati gli attributi sensibili; e ii) *statistical parity*, secondo la quale non vi è discriminazione quando vi è indipendenza statistica tra le decisioni formulate dal modello, condizionatamente all'appartenenza degli individui ad un gruppo. Tra le definizioni di *fairness* individuale si cita la *counterfactual fairness*, che ricorre al ragionamento probabilistico per valutare la performance del modello a fronte di variazioni di specifiche caratteristiche individuali. Una volta fissata[7] la definizione di *fairness* maggiormente appropriata al caso d'uso, questa può essere impiegata per la valutazione a valle del processo di stima o all'interno di quest'ultimo, in forma di vincolo sulla funzione obiettivo ottimizzata dall'algoritmo di apprendimento.

Poiché l'adozione di una specifica definizione di *fairness* tipicamente implica la violazione di quanto prescritto dalle altre, l'analista è tenuto, caso per caso, a valutare quale possa essere maggiormente rilevante al fine di mitigare la presenza di eventuali meccanismi di discriminazione.

La letteratura economica ha ampiamente investigato la presenza di discriminazione nella valutazione del merito di credito, prevalentemente negli Stati Uniti, dove è prevista una normativa che esplicitamente vieta la discriminazione nel mercato del credito (*Fair Lending*). Le analisi misurano la discriminazione in termini di probabilità di accedere al credito e di differenziali nei prezzi praticati a parità di caratteristiche osservate della clientela, riscontrando evidenze di

[6] Un classico esempio è il cosiddetto *redlining*, secondo il quale individui appartenenti ad un certo gruppo sociale geograficamente rintracciabile in un quartiere popolare di una grande città subiscono forme di discriminazione dovute all'uso del codice di avviamento postale nei modelli decisionali.

[7] L'adozione della definizione di *fairness* appropriata può dipendere da una molteplicità di fattori. Tra questi figura l'orientamento dell'analista a perseguire un approccio di tipo punitivo o inclusivo. Nel primo caso si valutano negativamente decisioni favorevoli per i gruppi o i soggetti giudicati non meritevoli, mentre nel secondo si premia l'assenza di decisioni sfavorevoli nei confronti di quelli virtuosi.

discriminazione etnica e, in taluni casi, di genere. Alcuni lavori analizzano specificatamente il legame tra l'utilizzo di metodi quantitativi di valutazione del merito di credito basati su modelli statistici e la presenza di discriminazione e i risultati non sono concordi. Per una rassegna della letteratura si veda l'Appendice 2.

4.2 Il bias nel merito di credito: peculiarità derivanti dall'uso di AI

Nell'utilizzo di un sistema di AI a supporto delle decisioni la presenza di meccanismi di non corretta differenziazione o di discriminazione, oltre ad avere effetti sulla bontà della performance, tende a generare un ciclo di feedback in cui la distorsione è confermata e rinforzata. Nel tempo, ad esempio, la negazione sistematica del credito causata da un modello non correttamente specificato a danno di specifici gruppi sociali, contribuisce a determinare un bias storico nei dati, che si rifletterà nei campioni estratti dalla popolazione, sulla base dei quali verrà aggiornato lo stesso modello distorto, secondo un circolo vizioso.

Le distorsioni possono emergere all'interno delle diverse fasi del ciclo di sviluppo di un algoritmo di AI: nella raccolta dei dati, nella specificazione del modello e nell'apprendimento ed infine nell'analisi degli output. Nel caso di tecniche di ML tali distorsioni possono comparire trasversalmente alle fasi sopracitate; nel contesto dell'*automated reasoning*, le forme di distorsione possono essere causate prevalentemente dall'incorretta specificazione del problema ovvero dalla sua incompleta formalizzazione[8].

Nella fase di raccolta dei dati occorre distinguere tra il caso in cui le tecniche di ML siano applicate a dati tradizionali e quello in cui si utilizzino big data.

Nel primo caso, l'adeguata rappresentatività del campione è in linea di principio garantita se si adottano approcci statistici consolidati per il campionamento da popolazioni finite. Ciononostante, possono permanere fonti di distorsione legate al fatto che i dati sono generati, raccolti e/o elaborati da esseri umani (Ntoutsi et al., 2020), che determinano la sovra- o sotto-rappresentazione di specifici gruppi. Assume rilevanza in questo ambito la possibilità di bias storico, ossia la distorsione dovuta al fatto che soggetti appartenenti ad uno specifico genere, gruppo etnico o sociale a cui in passato era stato negato l'accesso al credito in ragione di meccanismi discriminatori, vengano sotto rappresentati nel campione in favore di categorie a cui storicamente è stato concesso credito[9]. Laddove la scarsa rappresentazione di soggetti caratterizzati da specifici attributi socio-demografici dipenda dalla presenza di meccanismi espliciti di discriminazione sistematica, il processo di stima di un modello risentirà di una forma di distorsione detta bias istituzionale[10]. Infine, la distorsione può essere indotta dall'omissione di caratteristiche rilevanti nella definizione del campione (*omitted*

[8] La formalizzazione del problema richiesta da un sistema di AR può aiutare a superare il rischio di bias indotti da una deviazione nella comprensione dello stesso.

[9] Cfr. al riguardo Appendice n. 2 per evidenze in materia nella letteratura economica.

[10] Nell'ambito della valutazione creditizia, un esempio può essere rappresentato dalla presenza - nella base dati di addestramento - di caratteristiche dei richiedenti sistematicamente connotate come favorevoli o indispensabili per la concessione del credito (come l'aver maturato un numero minimo di anni di esperienza di lavoro, attributo correlato con l'età del richiedente). Tale distorsione sistematica dei dati può comportare forme di discriminazione, "istituzionalmente" incorporate nelle logiche/regole assunte dagli intermediari nella concessione del credito.

variable bias) oppure dal ricorso a attributi, *proxy* di caratteristiche non direttamente osservabili degli individui, che non riflettono correttamente il fenomeno sotto osservazione (*measurement bias*).

Se la tecnica di ML è applicata a big data, non è possibile formulare né verificare ipotesi statistiche sul processo di generazione dei dati e sulla loro raccolta, e quindi applicare correttivi statistici. In linea di principio l'utilizzo dei big data dovrebbe consentire di limitare le distorsioni in virtù dell'accesso ad informazioni caratterizzate da alto livello di granularità (dettaglio delle informazioni, frequenza nel tempo) e ricchezza (numero di attributi disponibili per ciascun individuo), riferite all'intera popolazione e non limitate ad un campione potenzialmente non rappresentativo o affetto, ad esempio, da bias storico (Óskarsdóttir et al., 2019). Nella realtà, tuttavia, i processi di raccolta dei big data sono soggetti comunque ad una gamma di possibili effetti distorsivi. In particolare, la tendenza dei soggetti ad auto-selezionarsi (s*elf-selection bias*)[11] e l'attitudine naturale degli individui ad adottare comportamenti e linguaggi differenti a seconda dei contesti (*Behavioral e Content Production bias*) possono gravemente compromettere la validità del modello, nonché la bontà dei risultati, fino a tradursi in meccanismi di natura discriminatoria.

Il ricorso ai big data rende inoltre l'analisi suscettibile a forme di distorsione da auto-selezione derivanti dal divario digitale tra individui[12], il cui accesso a internet e alle nuove tecnologie concorre a definire l'intensità del processo di generazione continua di dati (*digital footprint*). Ne consegue che taluni soggetti generano una *digital footprint* che ne permette l'analisi basata su profilazione da parte di sistemi di ML, mentre altri sono scarsamente rappresentati nel mondo digitale (*thin-file)* se non del tutto invisibili (*no-file)*. Quando i modelli di ML sono addestrati a partire da dati provenienti da fonti nelle quali è significativa la presenza di soggetti caratterizzati da *thin-file* e *no-file,* la selezione resta esposta a rischi di non corretta differenziazione del merito di credito e di discriminazione. Inoltre, il ricorso a dati comportamentali espone la concessione del credito a dinamiche potenzialmente manipolatorie da parte dei consumatori con elevate abilità digitali (Freeman et al., 2017; Calo, 2013).

Anche qualora il processo di raccolta dei dati fosse immune da distorsioni, l'incompleta specificazione del problema potrebbe determinare forme di distorsione di tipo "algoritmico", a partire dal tipo di modello utilizzato e dal suo processo di apprendimento (Baeza-Yates, 2018).

La distorsione algoritmica può, nelle sue diverse manifestazioni, essere generata da molteplici fattori, tra cui:

i) la specificazione di una forma funzionale inadeguata a descrivere il fenomeno oggetto di studio, con conseguente *underfitting*;

[11] Ad esempio, l'analisi delle abitudini di spesa a partire dall'insieme di transazioni registrate su una determinata applicazione mobile sarà rappresentativa della popolazione che fa uso dello smartphone per effettuare i pagamenti. Le caratteristiche socio-demografiche di tale gruppo non sono riferibili alla popolazione nel suo complesso né tantomeno alle abitudini di spesa generali.

[12] Noto anche come *Big Data Exclusion*.

ii) la mancata o inadeguata considerazione delle diverse tipologie di errore nei risultati del modello[13];

iii) l'omissione di attributi strumentali alla comprensione del fenomeno, in grado di invertirne la valutazione e l'interpretazione (*Simpson's paradox*)[14];

iv) l'inclusione di variabili significative per il fenomeno, *proxy* per l'identificazione del soggetto come vulnerabile (*included variable bias*).

Ulteriori forme di distorsione nell'uso di algoritmi di ML possono essere causate da conclusioni erronee tratte in fase di analisi dell'output, ad esempio in conseguenza della scarsa spiegabilità del modello. L'analista può: i) essere indotto a trarre conclusioni distorte con riferimento a gruppi la cui identificazione si basa su attributi non correttamente identificati (*aggregation bias*)[15]; ii) formulare considerazioni di natura causale a partire da coorti distinte della popolazione osservate trasversalmente (*longitudinal data fallacy*)[16]; iii) interpretare la rilevanza di un attributo ai fini della derivazione dell'output come evidenza in favore di una relazione causale tra i due (*cause-effect bias*).

Si rimanda all'appendice 1 per una breve rassegna delle tecniche applicabili per l'identificazione dei bias e all'appendice 2 per una survey sulle analisi e gli approfondimenti sul tema delle distorsioni nelle applicazioni di *credit scoring* basate su modelli di ML riscontrate in letteratura.

5. Quadro normativo/istituzionale

In questa sezione si illustra il quadro istituzionale e normativo nel quale si inserisce il ricorso a modelli di AI/ML per il *credit scoring* da parte degli intermediari. Nel contesto, rilevano sia la normativa prudenziale sia quella di tutela della clientela, inclusa quella più generale di tutela della riservatezza dei dati. Sono inoltre riportati i principali orientamenti emersi nei consessi internazionali e nazionali sull'applicazione di tecniche di AI/ML al *credit scoring*.

5.1 La non corretta differenziazione del rischio dei modelli di credit scoring nel quadro normativo prudenziale

[13] Ad esempio, la funzione di costo in termini di errata valutazione degli inadempienti deve riflettere adeguatamente se i falsi positivi sono da considerarsi "più gravi" dei falsi negativi o viceversa.

[14] L'omissione di attributi fortemente correlati alle variabili di cui si voglia analizzare il comportamento può impedirne la corretta valutazione. Ad esempio, l'omissione dell'informazione relativa all'età, può comportare la valutazione distorta dei meccanismi descritti dal modello quando si considera lo storico dei pagamenti estratto da una popolazione a larga prevalenza anziana (D'Alessandro et al., 2017).

[15] Ad esempio, l'individuazione di una maggior predisposizione al deterioramento del credito tra le imprese appartenenti ad una determinata regione geografica può condurre a valutazioni distorte laddove non si consideri, ancorché presente, l'indicazione dell'attività economica che potrebbe essere in relazione significativa con la variabile output del modello.

[16] Ad esempio, il deterioramento del credito generalizzato per un insieme di soggetti può essere attribuito alla riduzione del potere di acquisto da parte degli stessi nel tempo, anziché correttamente ricercato in caratteristiche del gruppo da rilevarsi trasversalmente.

Nella normativa prudenziale non sono previste prescrizioni specificamente riferite all'utilizzo dell'intelligenza artificiale e del machine learning, né preclusioni all'applicazione di suddette tecniche. Le indicazioni ivi contenute tuttavia sono applicabili e valide indipendentemente dalle tecniche di stima dei modelli utilizzate dagli intermediari.

Con riferimento al rischio di credito, il quadro normativo è rappresentato, in primis, dalle previsioni sul governo societario, controlli interni e gestione dei rischi contenute nella disciplina prudenziale[17] che indicano i principi da seguire per garantire l'efficacia dei processi in materia. Con riferimento ai criteri di erogazione, è previsto che gli intermediari, durante la fase istruttoria, raccolgano tutte le informazioni necessarie per valutare il merito di credito mediante l'utilizzo di sistemi di scoring o rating. Non sono presenti specifici requisiti sulle caratteristiche che tali sistemi devono avere, purché essi "forniscano indicazioni circostanziate sul livello di affidabilità del cliente".

Le Linee Guida dell'EBA su *loan origination and monitoring*, recentemente recepite (20 luglio 2021) nel quadro regolamentare nazionale quali orientamenti di vigilanza[18], definiscono in modo più specifico indicazioni[19] ai fini dell'utilizzo di modelli "automatizzati" per la valutazione del merito di credito, tra cui figurano: i) la comprensione delle assunzioni utilizzate nel modello; ii) politiche e procedure interne volte a rilevare e prevenire le distorsioni e ad assicurare la qualità dei dati inseriti; iii) la necessità di valutare l'adeguatezza (in termini di tracciabilità, verificabilità, robustezza e resilienza) di input e output dei modelli; iv) la presenza di policy che garantiscano che la qualità dei risultati dei modelli sia verificata regolarmente; v) la presenza di meccanismi di valutazione del risultato prodotto ai fini di un suo eventuale "*override*" che incorpori il giudizio di esperti; vi) la disponibilità di un'adeguata documentazione sulla costruzione dei modelli e il loro utilizzo[20].

Un quadro normativo più articolato e prescrittivo è invece definito per gli intermediari che chiedono all'autorità di Vigilanza l'autorizzazione all'utilizzo per il calcolo dei requisiti patrimoniali dei modelli interni per la misurazione del rischio di credito[21].

Le relative prescrizioni, per quanto non vincolanti per modelli che non sono destinati al calcolo dei requisiti patrimoniali, possono rappresentare comunque un utile riferimento per individuare migliori prassi da seguire anche per lo sviluppo e la gestione dei modelli di AI/ML.

[17] Circ. 285 e 288 della Banca d'Italia.

[18] "Gli intermediari compiono ogni sforzo per conformarsi agli Orientamenti" (cfr. Nota n. 13 del 20 luglio 2021 - *Attuazione degli Orientamenti dell'Autorità bancaria europea in materia di concessione e monitoraggio dei prestiti* (EBA/GL/2020/06))

[19] Le indicazioni elencate di seguito sono incluse in una sezione degli Orientamenti applicabile solo per le banche e non per gli intermediari finanziari.

[20] Cfr. paragrafi 53, 54 e 55 delle *EBA-GL-2020-06 on loan origination and monitoring*.

[21] A tal riguardo, l'EBA ha pubblicato l'11 novembre 2021 un *Discussion paper*, nel quale vengono proposte alcune raccomandazioni in merito all'utilizzo di sistemi di AI nell'ambito dei modelli interni per il rischio di credito utilizzati ai fini del calcolo dei requisiti patrimoniali (IRB).

Il riferimento normativo principale è la *Capital Requirement Regulation* (CRR - Regolamento Europeo 575/2013) cui si accompagnano norme tecniche e Linee Guida predisposte dall'EBA. Con riferimento alle banche significative, la BCE ha inoltre pubblicato una guida sui modelli interni[22] nella quale chiarisce come intende applicare il quadro normativo definito dal CRR e dal *framework* EBA.

Le menzionate norme prevedono requisiti sia per quanto riguarda gli aspetti più generali e di processo, sia per gli aspetti quantitativi dei modelli interni.

Gli aspetti generali afferiscono alle caratteristiche di integrità del processo di assegnazione dei rating, all'impiego effettivo dei modelli nei processi aziendali, alla loro documentazione, alla conservazione dei dati, alla istituzione di un processo di validazione delle stime interne e alla presenza di un appropriato assetto di governance e controlli interni.
Per loro natura tali requisiti tendono ad essere trasversali rispetto alle tecniche impiegate per la stima dei parametri di rischio e quindi dei modelli, applicandosi a prescindere dalle loro caratteristiche e dalle tecnologie impiegate per la loro implementazione.

Requisiti generali dei modelli interni

Tra le prescrizioni che hanno una valenza più diretta in relazione alle caratteristiche dei modelli di ML si segnala l'esigenza che l'intermediario disponga di "un processo per vagliare i dati immessi nel modello di previsione che contempli una valutazione dell'accuratezza, completezza e pertinenza dei dati" (art. 174 CRR). Di rilievo anche il richiamo all'esigenza che l'intermediario combini "il modello statistico con la valutazione e la revisione umana in modo da verificare le assegnazioni effettuate in base al modello e da assicurare che i modelli siano utilizzati in modo appropriato". L'intento è che l'intermediario sia in condizioni di scoprire e limitare gli errori derivanti da carenze del modello.

In relazione alla tematica dei dati, è rilevante quanto si richiede in tema di documentazione dei sistemi di rating (art. 175 CRR): la stessa deve in ogni caso assicurare "una descrizione dettagliata della teoria, delle ipotesi e delle basi matematiche ed empiriche su cui si fonda l'assegnazione delle stime a classi, singoli debitori, esposizioni o aggregati, nonché le fonti dei dati, una o più, utilizzate per costruire il modello".

Con riferimento alla tematica dei modelli forniti da terze parti si ribadisce poi l'esigenza che i principi relativi alla documentazione dei sistemi di rating siano in ogni caso soddisfatti anche ove il fornitore "rifiuti o limiti l'accesso dell'ente ad informazioni relative alla metodologia di tale sistema di rating o modello, o ai dati di base utilizzati per elaborare tale metodologia o modello, vantando un diritto di proprietà su tali informazioni".

I più rilevanti requisiti relativi agli aspetti quantitativi dello sviluppo dei modelli, ossia alla determinazione della procedura che differenzia il rischio tra differenti debitori, riguardano cinque ambiti principali: i) qualità dei dati; ii) utilizzo dei driver appropriati; iii) filosofia di rating; iv)

[22] ECB Guide to Internal Models, EGIM.

rappresentatività dei dati; v) capacità di differenziare il rischio e valutazione delle performance del modello.

Aspetti quantitativi dei modelli interni.

i) Qualità dei dati di input dei modelli. – L'intermediario deve verificare che i dati utilizzati siano appropriati, accurati e completi (art. 174(b) CRR). Questo requisito generale assume particolare rilevanza per i modelli di ML che si basano generalmente su una mole rilevante di dati, sia provenienti da fonti interne, sia fornite da provider esterni.

ii) Utilizzo di determinanti del rischio appropriati. - I valori dei parametri di rischio devono essere plausibili e basati su variabili rilevanti, verificando che i parametri stimati siano in linea con il loro significato economico atteso (art. 179(1)(a) CRR). Un'analisi di questo tipo è particolarmente importante per i modelli di ML che basano le loro stime su relazioni non lineari fra numerosi fattori e che presentano problematiche di spiegabilità.

iii) Filosofia di rating. - La banca, per ciascun modello, deve stabilire le proprietà dinamiche attese delle classi di rating (*Point-in-time - PiT vs Through-the-cycle - TTC*) e assicurare che i modelli statistici utilizzati consentano di rispettare l'obiettivo prefissato (para. 66-68 EBA-GL-2017-16). Generalmente, i modelli di ML tendono ad introdurre elementi *PiT* nelle stime che potrebbero risultare incoerenti con gli obiettivi di filosofia di rating identificati dall'intermediario; tale contaminazione potrebbe avvenire in modo non controllato a motivo della scarsa spiegabilità del ML.

iv) Rappresentatività dei dati. – La normativa richiede che i parametri di rischio siano stimati sulla base di un database rappresentativo delle caratteristiche dei clienti dell'intermediario (art. 174(c) CRR). In particolare, la capacità del modello di differenziare il merito di credito dei debitori non deve essere indebolita dalla scarsa rappresentatività dei dati utilizzati. Le caratteristiche dei modelli di ML, in particolare se prevedono il ricorso a big data, rendono difficile garantire che un'eventuale scarsa rappresentatività dei dati non sia causa di non corretta differenziazione dei prenditori.

v) Differenziazione del rischio e verifica delle performance del modello. - Il modello deve riflettere le caratteristiche rilevanti del debitore e delle transazioni, e deve garantire una buona differenziazione del rischio, ossia deve associare livello di rischio più basso a debitori che hanno un buon merito di credito (art. 170 CRR). In tale ambito, nel processo di revisione delle stime è richiesto che l'intermediario valuti la capacità del modello di differenziare il rischio, comparando le performance misurate nel campione di stima del modello con quelle ottenute su portafogli più recenti (para. 218 EBA-GL-2017-16). I sistemi di ML sono esposti al rischio di distorsioni da *overfitting* che possono minare la capacità discriminante del modello.

5.2 La discriminazione nel quadro giuridico europeo e nazionale e nella normativa di settore

Un principio fondamentale in materia di eguaglianza e non discriminazione è espresso dall'art. 3 della Costituzione: "Tutti i cittadini hanno pari dignità sociale e sono eguali davanti alla legge, senza distinzione di sesso, di razza, di lingua, di religione, di opinioni politiche, di condizioni personali e sociali". Nell'ordinamento italiano si trovano inoltre alcune definizioni specifiche di discriminazione, quali quelle di genere, razziale, o verso persone con disabilità, che traggono spunto da quanto stabilito dalla Direttiva Europea sull'Uguaglianza razziale (2000/43/CE) e sul Lavoro (2000/78/CE)[23].

5.2.1 Normativa di trasparenza

Nel Testo Unico Bancario e nelle disposizioni di trasparenza sono contenuti dei riferimenti generali alla discriminazione che, pur limitati ai temi dell'accesso ai servizi di pagamento, introducono una definizione di ciò che può essere reputato discriminatorio nel contesto bancario-finanziario e riaffermano la connessione fra l'inclusione finanziaria e la non discriminazione[24].

Con specifico riferimento alle normative di trasparenza in materia di credito immobiliare ai consumatori, l'art. 120-undecies del TUB, relativo alla verifica del merito creditizio, afferma al comma 1 che "…La valutazione del merito creditizio è effettuata sulla base delle informazioni sulla situazione economica e finanziaria del consumatore necessarie, sufficienti e proporzionate e opportunamente verificate", ponendo dunque dei requisiti alle caratteristiche dei dati utilizzati dall'intermediario ai fini del *credit scoring*.

Il comma 5 del medesimo articolo stabilisce inoltre che "Quando la domanda di credito è respinta, il finanziatore informa il consumatore senza indugio del rifiuto e, se del caso, del fatto che la decisione è basata sul trattamento automatico di dati", evidenziando un diritto del consumatore ad essere informato qualora il rifiuto della sua domanda di credito immobiliare sia stato basato sull'utilizzo di forme di *algorithmic credit scoring*[25].

In generale, come avviene anche negli altri settori dell'economia, esiste la necessità di coniugare la non-discriminazione con la libertà a contrarre di imprese private, quali sono gli intermediari in

[23] Di rilievo nel nostro contesto appare il Testo Unico sull'Immigrazione che all'art. 43, comma 2, specifica che compie un atto di discriminazione "chiunque imponga condizioni più svantaggiose o si rifiuti di fornire beni o servizi offerti al pubblico ad uno straniero soltanto a causa della sua condizione di straniero o di appartenente ad una determinata razza, religione, etnia o nazionalità".

[24] Nel TUB (comma 2 dell'art. 126-noviesdecies) è detto che "Tutti i consumatori soggiornanti legalmente nell'Unione europea, senza discriminazioni e a prescindere dal luogo di residenza, hanno diritto all'apertura di un conto di base …". Le disposizioni di trasparenza in materia di servizi di pagamento (par. 5.2 della sezione VI) recitano: "Tutte le modifiche dei tassi di interesse o di cambio sono applicate e calcolate in modo tale da non creare discriminazioni tra clienti. Le modalità di applicazione e di calcolo di queste modifiche si presumono non discriminatorie quando gli intermediari le adottano sulla base di criteri oggettivi e motivati che applicano a tutti i clienti, a parità di condizioni".

[25] La proposta di revisione della Consumer Credit Directive, pubblicata a luglio 2021 dalla Commissione Europea, introduce delle previsioni simili anche in materia di credito al consumo. L'obbligo di informazione ai consumatori sarebbe altresì presente anche in caso di offerte personalizzate basate su tecniche di profilazione.

materia di concessione, o rinnovo, del credito. Quest'ultimo principio è stato più volte ribadito anche dall'Arbitro Bancario Finanziario, che ha affermato che "non sussiste, in via generale, un obbligo per gli intermediari di concedere credito o di rivedere le condizioni alle quali è stato concesso, fatto salvo il dovere, in fase di valutazione e riscontro di eventuali richieste di rinegoziazione, di rispettare il principio di correttezza nei rapporti contrattuali…la valutazione del merito creditizio rientra infatti nell'autonomia gestionale degli intermediari"[26].

5.2.2 Privacy

La normativa europea in materia di privacy (GDPR) tratta temi che assumono particolare rilievo per l'utilizzo di sistemi di AI-ML per il *credit scoring*: la profilazione, il consenso al trattamento dei dati e il diritto di accesso alla logica di una decisione automatizzata.

Nell'articolo 4 viene definita la profilazione[27], mentre l'articolo 9 sancisce un generale divieto di trattare i dati sensibili[28] per scopi decisionali automatizzati, a meno che non vi sia un consenso esplicito dell'interessato al trattamento per una o più finalità specifiche, o vi sia un motivo di interesse pubblico.

A fronte di un generale divieto di sottoporre un individuo a processi decisionali automatizzati, l'articolo 22, paragrafo 2, stabilisce che, oltre al caso di un esplicito consenso al trattamento, il trattamento automatizzato è possibile se necessario per la conclusione o l'esecuzione di un contratto tra l'interessato e il titolare.

Un diritto all'informazione più esteso di quello contenuto nel comma 5 dell'art. 120-undecies del TUB è presente nell'articolo 15 del GDPR inerente il diritto di accesso dell'interessato, che al paragrafo (1)(h) sancisce il diritto alle informazioni circa "l'esistenza di un processo decisionale automatizzato, compresa la profilazione […] e, almeno in tali casi, informazioni significative sulla logica utilizzata, nonché l'importanza e le conseguenze previste di tale trattamento per l'interessato" . Secondo alcuni commentatori, la richiesta di "informazioni significative sulla logica utilizzata" implica un obbligo degli intermediari a fornire spiegazioni ai richiedenti cosiddette "locali", ovvero inclusive del dettaglio delle principali variabili che hanno contribuito a determinare lo specifico punteggio (cfr. Hacker e Passoth, 2021).

[26] Relazione Annuale ABF 2018, pag. 53.

[27] È "*qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica".*

[28] Ivi definiti quali dati inerenti all'origine razziale o etnica; le opinioni politiche; le convinzioni religiose o filosofiche; l'appartenenza sindacale; genetici; biometrici, relativi alla salute; relativi alla vita sessuale o all'orientamento sessuale della persona; giudiziari.

5.3 La posizione delle istituzioni nazionali e internazionali in tema di AI e la proposta europea di AI Regulation

Al di fuori delle indicazioni presenti nella normativa positiva sopra delineate, sul tema dell'utilizzo dell'AI si registrano le posizioni di numerose autorità nazionali e internazionali.

Tra le altre, meritano di essere menzionate con riferimento al tema in esame:

1) la Commissione Europea, che ha espresso le proprie valutazioni in comunicazioni - COM(2018) 237 e COM(2018) 795 - ed elaborato due documenti specifici (le *Ethic guidelines for trustworthy AI* [29] e le *Policy and investment recommendations for trustworthy AI* [30]) e report rilevanti in materia (*Report of Expert Group on Regulatory Obstacles to Financial Innovation* - ROFIEG [31]);
2) il Financial Stability Board, che ha analizzato le possibili implicazioni per la stabilità finanziaria dell'uso di AI-ML nei servizi finanziari nel report FSB, 2017[32];
3) l'EBA, che ha realizzato un approfondimento sull'uso dei Big Data e Advanced Analytics (BD&AA) nel settore bancario nel contesto della propria FinTech Roadmap[33];
4) il MISE, che nella Strategia Nazionale per l'Intelligenza Artificiale (MISE, 2020)[34] ha delineato i principi guida per l'introduzione dell'AI nei vari settori dell'economia italiana.

Anche le banche centrali hanno nel corso degli ultimi anni avviato riflessioni e ricognizioni in merito all'uso dell'AI da parte dell'industria finanziaria. Tra i contributi delle diverse autorità (ACPR[35], Bafin[36], Bank of England[37]), si segnalano quelli di:

1) la Monetary Authority of Singapore (MAS) che nel 2018 ha definito 4 principi sull'uso dell'AI e dei "data analytics" declinati secondo un approccio denominato FEAT (*fairness*, *ethics*, *accountability and transparency*)[38]; a partire da questo schema logico nel 2020[39] e nel 2022[40] sono stati elaborati spunti per una metodologia volta a consentire la verifica dell'allineamento a questi principi;

2) la Banca centrale olandese (De Nederlandsche Bank - DNB) ha aggiunto due ulteriori principi allo schema proposto dalla MAS elaborando un approccio con 6 step di alto livello

[29] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[30] https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.
[31] https://ec.europa.eu/info/files/191113-report-expert-group-regulatory-obstacles-financial-innovation_en.
[32] https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service.
[33] https://www.eba.europa.eu/eba-publishes-its-roadmap-on-fintech.
[34] https://www.mise.gov.it/index.php/it/strategia-intelligenza-artificiale/contesto.
[35] https://acpr.banque-france.fr/en/governance-artificial-intelligence-finance;
https://acpr.banque-france.fr/en/acpr-tech-sprint-explainability-artificial-intelligence.
[36] https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html;
https://www.bafin.de/SharedDocs/Downloads/EN/Aufsichtsrecht/dl_Prinzipienpapier_BDAI_en.html.
[37] https://www.bankofengland.co.uk/research/fintech/ai-public-private-forum.
[38] https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT.
[39] https://www.mas.gov.sg/schemes-and-initiatives/veritas.
[40] https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions.

denominato SAFEST (*soundness*, *accountability*, *fairness*, *ethics*, *skills and transparency*), accompagnato da 17 indicazioni per rendere applicabili i suddetti principi[41];

3) la Hong Kong Monetary Authority ha sviluppato alcune *supervisory guidelines* per le banche che intendano applicare tecniche di AI nei loro modelli di business[42], articolate su tre aree: *model risk management*[43], *consumer protection*[44] e *cyber security*[45].

In ambito europeo merita specifica segnalazione la bozza del "Regolamento sull'approccio europeo per l'intelligenza artificiale" (Artificial Intelligence Act) che, seppur solo all'inizio del suo percorso legislativo[46], fornisce una prima definizione normativa di intelligenza artificiale, delinea i presidi necessari nella gestione e nella verifica dei sistemi di AI, introduce importanti punti di contatto con il GDPR. Di particolare interesse ai fini di questo lavoro risulta la menzione specifica dei sistemi AI di *credit scoring* e del connesso rischio di effetti discriminatori.

A livello generale, la bozza ha l'obiettivo di regolamentare l'AI al fine di consentirne lo sviluppo e l'utilizzo ordinato nell'UE, tutelando, al contempo, i cittadini dall'emergere di possibili pratiche lesive dei loro diritti[47]. Alcune applicazioni o pratiche basate sulle tecnologie di AI sono considerate dal Regolamento inaccettabili e dunque esplicitamente vietate (ad esempio, sistemi che adottino tecniche manipolative o attribuiscano punteggi sociali in base al comportamento delle persone), mentre altri utilizzi dell'AI sono considerati ad alto rischio e dunque sottoposti alla previsione di una serie di presidi e verifiche. I sistemi AI di *credit scoring* delle persone fisiche rientrano in questa categoria per via del loro impatto sulla vita degli individui e il rischio di introdurre o perpetuare dinamiche di discriminazione nella valutazione dell'affidabilità creditizia delle persone.

I fornitori di sistemi AI ad alto rischio devono rispettare una serie di requisiti (cfr. infra) e istituire un sistema di controlli che garantisca nel tempo la qualità del servizio offerto e la gestione e mitigazione dei rischi[48]. I fornitori di tali sistemi AI dovranno inoltre fornire all'autorità nazionale competente identificata dallo Stato Membro, dietro apposita richiesta, tutte le informazioni e la documentazione necessaria per dimostrare la conformità del sistema al Regolamento.

Fra i requisiti richiesti per i sistemi AI ad alto rischio si citano:

1. l'utilizzo di *dataset* pertinenti, rappresentativi, completi, privi di errori e dotati di adeguate proprietà statistiche;
2. la presenza di metodologie e pratiche di gestione dei dati, di addestramento e di convalida dei modelli che assicurino la valutazione di possibili distorsioni;
3. dati che tengano conto delle caratteristiche dello specifico contesto geografico, comportamentale o funzionale all'interno del quale il sistema AI deve essere usato;

[41] https://www.dnb.nl/media/voffsric/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf.
[42] https://www.aof.org.hk/docs/default-source/hkimr/applied-research-report/airep.pdf.
[43] https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191101e1.pdf.
[44] https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191105e1.pdf.
[45] https://www.hkma.gov.hk/eng/key-functions/international-financial-centre/fintech/research-and-applications/cybersecurity-fortification-initiative-cfi.
[46] La bozza presentata dalla Commissione è attualmente all'attenzione del Parlamento e del Consiglio Europeo.
[47] La disciplina si applicherà anche ad aziende non europee che utilizzano sistemi di AI nei confronti di utenti europei.
[48] Per gli intermediari vigilati tale presidio dovrà integrarsi con il sistema dei controlli interni normativamente previsto.

4. la tracciabilità dei risultati ottenuti, garantita mediante log generati automaticamente;
5. l'adeguatezza della documentazione e idonee forme di trasparenza nei confronti degli utenti che devono poter avere a disposizione informazioni concise, complete, accessibili e comprensibili, al fine di consentire loro di interpretare l'output del sistema;
6. adeguato livello di accuratezza, robustezza e sicurezza cibernetica;
7. la presenza di un livello di controllo umano sui sistemi AI, effettuata da parte di individui competenti ed esperti, in grado di comprendere, controllare e laddove necessario intervenire sui sistemi, eventualmente decidendo di ignorare l'output da loro prodotto.

6. Analisi dei risultati dell'indagine

Sono stati selezionati alcuni intermediari vigilati per poter approfondire l'adozione delle tecniche di AI nell'ambito della gestione del rischio di credito all'interno del panorama bancario e finanziario italiano.

L'insieme si compone di 10 intermediari, bancari e non, di diversa dimensione e vocazione di business, che - sulla base delle informazioni disponibili - stanno sperimentando, sviluppando o utilizzando modelli basati su tecniche AI-ML nel processo creditizio.

Oltre all'analisi delle risposte ai questionari, i risultati dell'indagine sintetizzano considerazioni qualitative emerse nel corso di incontri bilaterali con gli intermediari.

Le principali evidenze rilevate dall'analisi sono le seguenti:

- Il ricorso a metodi di AI nella valutazione del rischio di credito non è ancora largamente diffuso ma in espansione: i 10 intermediari intervistati hanno indicato di avere sviluppato in totale 38 modelli di cui circa il 60% erano già in uso al momento delle interviste. La maggioranza dei modelli è rivolta alla clientela corporate/PMI.
- Nella quasi totalità dei casi i punteggi prodotti dai modelli vengono forniti a supporto della valutazione del merito creditizio da parte degli analisti, che sono i responsabili della decisione finale. Alcuni intermediari hanno tuttavia dichiarato di avere intenzione in futuro di ridurre progressivamente l'intervento umano all'interno del processo di concessione.
- Il principale beneficio atteso che ha spinto gli intermediari a passare da metodi tradizionali a metodi di intelligenza artificiale è il miglioramento in termini di accuratezza delle previsioni.
- Fra gli altri benefici citati da alcuni intermediari vi è la possibilità di realizzare processi di *instant lending* e quella di sfruttare fonti dati alternative, facilitata dai modelli di ML, che consentirebbe di poter selezionare efficacemente clienti con limitata storia creditizia, ampliando la potenziale clientela.
- Nella maggioranza dei casi i modelli su cui è stato riferito nell'indagine utilizzano dati di tipo finanziario derivanti da fonti interne o acquistati da fornitori di *analytics*; è diffuso il ricorso a dati sulle movimentazioni del conto corrente, provenienti anche da *open banking*. Risulta estremamente limitato, invece, l'uso di dati tratti dal *web* e da social media.

- Circa il 90% dei metodi sviluppati si basa su combinazioni di alberi (*Gradient Boosting Trees, Random Forests*). La scelta, secondo gli intermediari, si deve alla maggiore semplicità di implementazione e all'ottimizzazione del trade-off tra accuratezza e spiegabilità.
- La quasi totalità degli intermediari ha adottato o intende adottare tecniche di spiegabilità per rendere più trasparente la logica decisionale dei modelli. Le tecniche più diffuse sono risultate gli *Shapley Values* e la *Feature Importance*, tecniche di spiegabilità *post hoc* che forniscono spiegazioni circa la logica di un modello già addestrato e calibrato.
- Si è rilevato il mancato ricorso a tecniche esplicitamente mirate alla riduzione della distorsione, quali il bilanciamento del *dataset*, il controllo del bias storico e l'analisi causale dei risultati.
- È stata adottata o si intende adottare una definizione di *fairness* in poco meno della metà dei modelli; tale quota sale a due terzi nel caso di modelli destinati a clientela retail. Tutti gli intermediari che hanno adottato una definizione, hanno optato per la *fairness through unawareness*, che consiste nella rimozione dalla base dati di analisi degli attributi esplicitamente considerati sensibili (come il genere o l'età dei clienti retail), ma non tiene conto degli effetti indotti sul modello dalla presenza di attributi con quelli potenzialmente correlati.
- Tutti gli intermediari hanno riferito di aver istituito o di volere istituire un sistema di governance del modello, assistito da specifica reportistica, ma il processo di monitoraggio copre aspetti afferenti alla qualità e integrità dei dati di input in poco più della metà dei modelli analizzati.
- Si è riscontrata una frequente indisponibilità interna delle competenze di sviluppo, manutenzione e controllo dei rischi dei modelli di ML e il ricorso anche a forme di completa esternalizzazione.

6.1 Panoramica dei modelli

I modelli di intelligenza artificiale analizzati sono in totale 38. Nella totalità dei casi l'approccio di intelligenza artificiale adottato è risultato di tipo induttivo, prevedendo l'utilizzo di tecniche di machine learning, in taluni casi associate a tecniche di *natural language processing*. Le tecniche di *automated reasoning* di tipo deduttivo non trovano riscontro nelle esperienze descritte dal campione di intermediari coinvolto nell'indagine. La preferenza per le tecniche di ML può essere ascritta alla loro ampia diffusione e alla disponibilità sul mercato di soluzioni tecnologiche (*training* di modelli e, in taluni casi, soluzioni pre-addestrate) e servizi professionali (competenze nello sviluppo dei modelli e conoscenza degli strumenti a supporto).

Gli algoritmi sono applicati in diverse fasi del processo di gestione del rischio di credito: il 58 per cento nella fase di concessione, il 26 per cento nel monitoraggio dei crediti e il restante 16 per cento in altre attività (definizione del *pricing*, determinazione delle rettifiche di valore, recupero dei crediti o altre attività svolte da funzioni di controllo interno). Sei modelli sono stati sviluppati in vista di un utilizzo ai fini del calcolo dei requisiti prudenziali.

Il 61 per cento dei modelli risulta già in uso presso gli intermediari, il 13 per cento è in fase di sperimentazione, mentre il 26 per cento ancora in studio o sviluppo; la quasi totalità di questi ultimi verrà impiegata nella fase di concessione del credito (cfr. fig. 7).

Nella maggioranza dei casi analizzati i punteggi prodotti dai modelli vengono forniti a supporto della valutazione del merito creditizio da parte degli analisti, i quali risultano ancora essere responsabili ultimi della decisione finale. Due eccezioni sono rappresentate da un primo caso in cui il contributo umano si attiva solo in presenza di contestazione del cliente e da un secondo caso, in cui il modello è finalizzato allo sviluppo di un processo di *instant lending* interamente automatizzato, seppure per la concessione di finanziamenti di importo molto contenuto. Alcuni intermediari hanno indicato tuttavia di avere intenzione in futuro di ridurre progressivamente l'intervento umano all'interno del processo di concessione del credito, una volta concluse le fasi di sperimentazione e di primo utilizzo, favorendo, ad esempio, la concessione in forma automatizzata in caso di valutazione positiva da parte del modello.

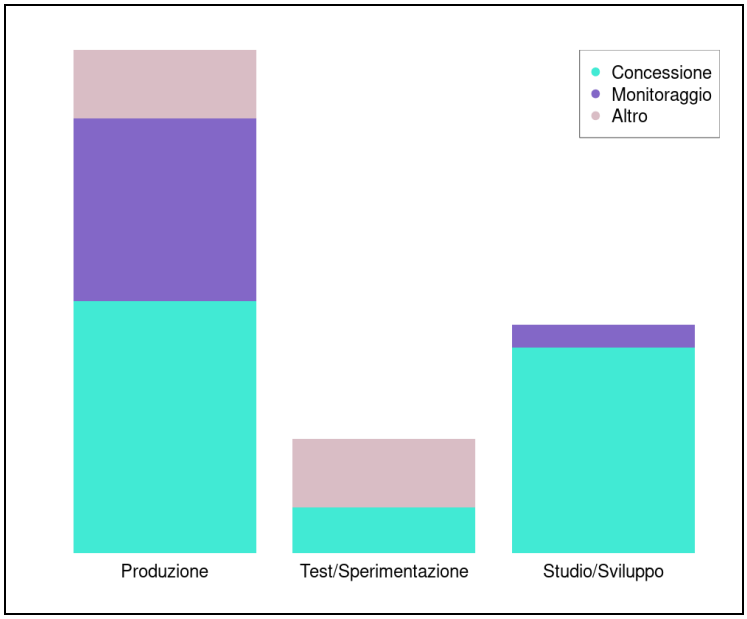


Figura 7. Stato di avanzamento dei modelli per fase di gestione del processo di credito

Due terzi dei modelli sono al servizio di processi creditizi orientati verso clientela corporate o PMI mentre un terzo a quella retail; concentrando l'attenzione solo sulla più critica fase di concessione dei finanziamenti, la metà dei modelli sviluppati è indirizzato a clienti del segmento retail, di cui 5 di questi attualmente in produzione, mentre l'altra metà al segmento Corporate/PMI (cfr. fig. 8).

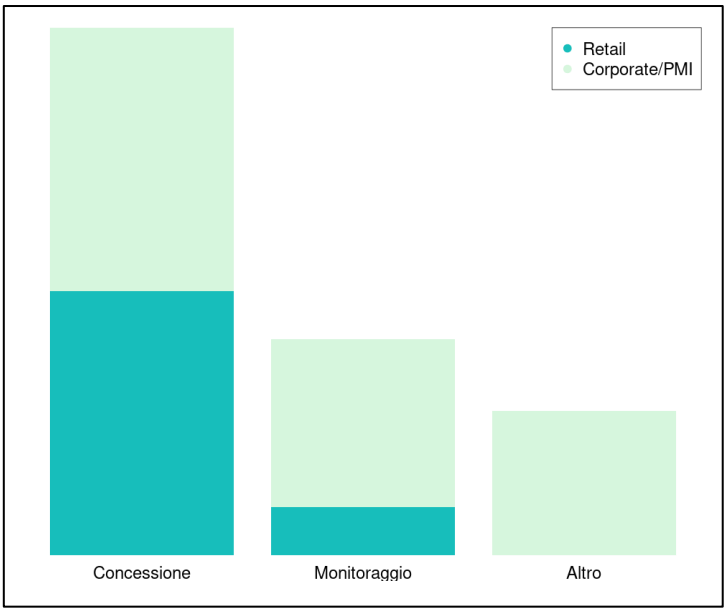


Figura 8. Fase di gestione del processo di credito dei modelli suddiviso per tipologia di clientela

6.2 Specificazione e sviluppo del modello

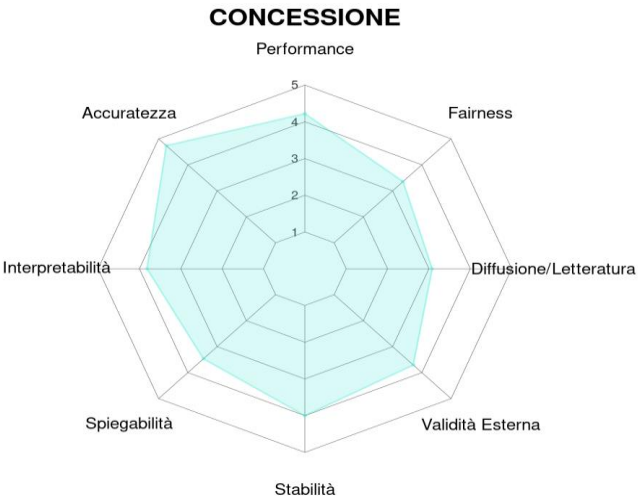
6.2.1 Fonti dati utilizzate

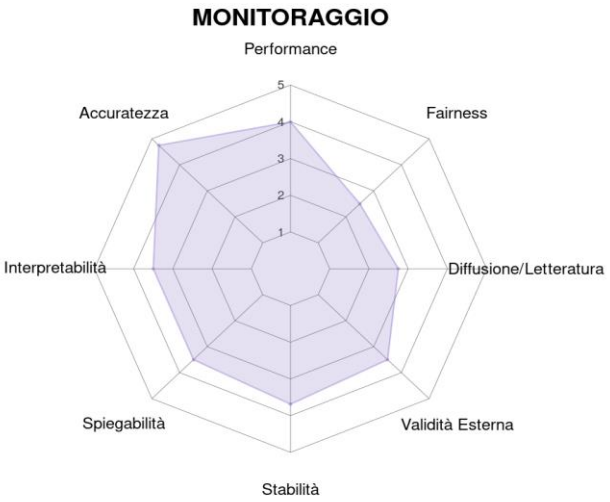
Ogni modello è costruito utilizzando in media quattro fonti di dati diverse, con un minimo di una e un massimo di 12; si tratta nella maggior parte dei casi di dati finanziari di tipo strutturato (ad es. indicatori patrimoniali ed economico-finanziari). Circa un terzo dei modelli sfrutta anche dati non finanziari, quali: informazioni socio-demografiche sulla clientela retail (disponibili internamente o dichiarate dal cliente in fase di richiesta del fido); relative al business di appartenenza o alle relazioni con le altre aziende, per la clientela corporate. Risulta limitato l'uso di fonti dati di tipo non strutturato, ovvero di testi; alcuni intermediari interpretano il testo presente all'interno delle causali di bonifico mediante tecniche di ML di classificazione.

Nella maggioranza dei casi i dati risultano di provenienza interna ovvero acquistati da fornitori di *analytics* attivi nel mercato del credito. La metà degli intermediari si avvale dei dati sulle movimentazioni del conto corrente, provenienti anche da *open banking*. Solo due intermediari stanno sperimentando l'uso di dati provenienti da web o social media: il primo li ha impiegati per alcune sperimentazioni che non sembra intenzionato a proseguire; il secondo ha realizzato un modulo di ML per elaborare le recensioni di imprese presenti sulle piattaforme *web* allo scopo di affinare le valutazioni prodotte dal modello di rating; benché al momento le previsioni presentino bassi livelli di accuratezza, l'intermediario ritiene che in prospettiva il modulo possa dare un contributo più significativo grazie alla crescente disponibilità di dati.

6.2.2 Selezione e sviluppo del modello

Il tipo di modello è stato generalmente selezionato tra due e cinque classi di modelli candidati. Nel processo di selezione assumono particolare rilevanza l'accuratezza delle previsioni e i criteri statistici di performance, soprattutto nei casi di modelli al servizio della concessione del credito. Per i modelli utilizzati nel monitoraggio risulta indispensabile anche la stabilità della performance nel tempo. La spiegabilità delle decisioni, l'interpretabilità del modello e la sua validità esterna sono considerate rilevanti per la scelta tra i modelli candidati, quale che sia la fase del processo del credito nella quale il modello viene utilizzato; al contrario, l'assenza di meccanismi di natura discriminatoria e la diffusione del modello in letteratura sono stati ritenuti in genere elementi poco rilevanti.





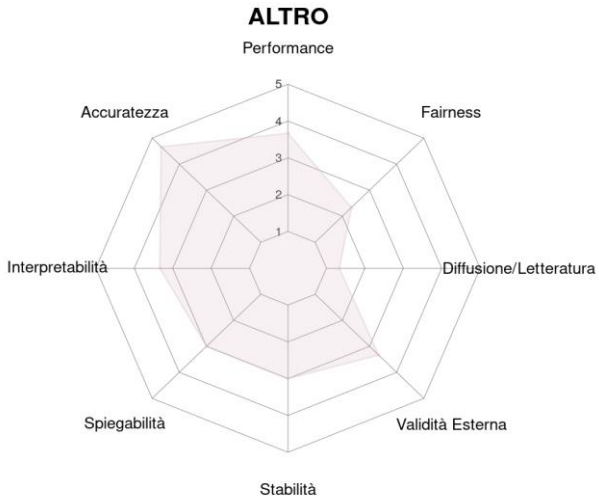


Figura 9. Criteri per la selezione dei modelli (scala da 1 – superfluo, a 5 – indispensabile)

Con riferimento al criterio dell'accuratezza, il ricorso a modelli di ML ha consentito secondo gli intermediari di aumentare sensibilmente la qualità della performance rispetto agli approcci tradizionali. In alcuni casi è stata indicata l'entità del miglioramento ma gli indicatori utilizzati non sono omogenei quindi non sono direttamente confrontabili. I valori indicati sono in linea con i risultati nella letteratura accademica.

Oltre il 95 per cento dei modelli si basa su tecniche di apprendimento supervisionato, principalmente metodi di *ensemble learning* che utilizzano combinazioni di alberi decisionali (come ad esempio *Gradient Boosting e Random Forest*), tipicamente disponibili nei pacchetti di software standard e open source; solamente due intermediari hanno sviluppato i loro modelli, finalizzati alla concessione dei finanziamenti, con tecniche di deep learning (reti neurali) supervisionato. I modelli restanti si avvalgono di approcci misti o esclusivamente non supervisionati, utilizzando algoritmi di *clustering* (cfr. fig. 10).

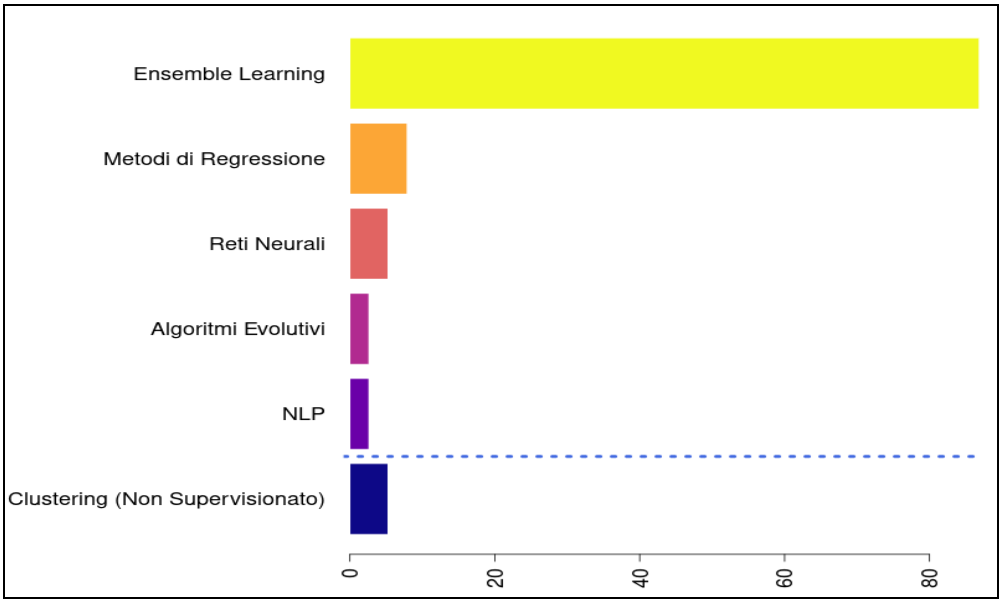


Figura 10. Tipologia di tecniche utilizzate (percentuali) – sopra la linea tratteggiata vengono riportate le tecniche usate con un approccio supervisionato, sotto la barra l'unica tecnica usata con un approccio non supervisionato.

6.3 Mitigazione della distorsione

Come accennato, la distorsione algoritmica può originarsi in diversi punti del processo di sviluppo di un modello basato su machine learning. È stato chiesto agli intermediari di indicare quali fossero le tecniche adottate per prevenire e mitigare la distorsione algoritmica per le tre fasi di raccolta dati, specificazione del modello e analisi delle previsioni (cfr. Sez. 4.2).

Complessivamente, gli intermediari hanno riferito di avere istituito presidi in fase di raccolta e preparazione dei dati per oltre 9 modelli su 10. Tra questi, il trattamento delle osservazioni mancanti (usualmente operato anche nei modelli statistici tradizionali) è stato impiegato nel 95% dei modelli utilizzati per la concessione (91% nel caso di quelli già in produzione) e nella totalità di quelli per il monitoraggio in produzione. Per nessuno dei modelli utilizzati in fase di concessione o monitoraggio è stato indicato l'uso di tecniche per la mitigazione del bias storico.

Con riferimento alla fase di specificazione, addestramento e calibrazione dei modelli, i presidi maggiormente indicati per la mitigazione della distorsione algoritmica sono stati la selezione delle variabili e la definizione del processo di addestramento (scelta della funzione obiettivo), rispettivamente per il 55% e il 45% dei modelli. Per la sola fase di concessione, nel 59% dei casi è stato indicato che nella scelta del modello si è privilegiato quello maggiormente interpretabile a parità di performance.

Infine, le tecniche di mitigazione della distorsione applicate a valle del processo di stima, importanti in particolare quando il modello di ML contribuisce all'automatizzazione delle decisioni, sono state impiegate dal 45 per cento dei modelli (dal 39 per cento nel caso di quelli in produzione). I risultati sono rappresentati dalla figura 11.

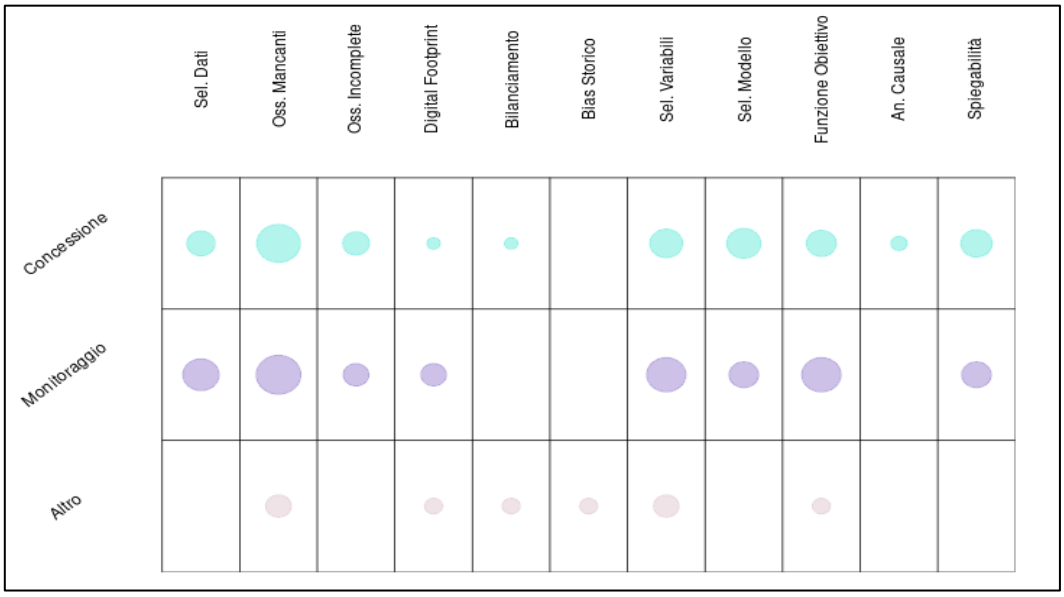


Figura 11. Panoramica dei presidi adottati dagli intermediari per la mitigazione della distorsione. Il diametro dei cerchi riflette la proporzione sul totale dei modelli utilizzati per ciascuna fase.

Nel corso degli incontri bilaterali è emerso come larga parte degli intermediari abbia adottato metodi statistici di trattamento del *dataset* (e.g. imputazione dei valori mancanti), di specificazione del modello (e.g. scelta del modello) e di analisi dei risultati (e.g. tecniche di spiegabilità) per esigenze diverse dalla mitigazione della distorsione algoritmica, ad esempio quella di migliorare l'accuratezza predittiva. Ciò presumibilmente spiega il mancato ricorso ad alcune tecniche mirate alla riduzione della distorsione, quali il bilanciamento del *dataset*, il controllo del bias storico e l'analisi causale dei risultati che sono da considerarsi migliori prassi per la mitigazione delle distorsioni rilevanti nell'ambito dei modelli per la valutazione del merito di credito.

6.4 Fairness

La distorsione algoritmica può tradursi in meccanismi discriminatori di tipo indiretto o comunque involontario (cfr. par. 4.1). Nel contesto della valutazione del merito di credito l'assenza di controllo sull'eventuale presenza di dinamiche discriminatorie è particolarmente rilevante. Complessivamente, per il 46 per cento dei modelli è stata adottata una definizione di *fairness* oppure è stato indicato di avere intenzione di adottarla; come illustrato dalla Figura 12, suddividendo i modelli per tipo di clientela, la percentuale è pari al 61 per quelli orientati alla clientela retail e del 37 per la clientela corporate/PMI.

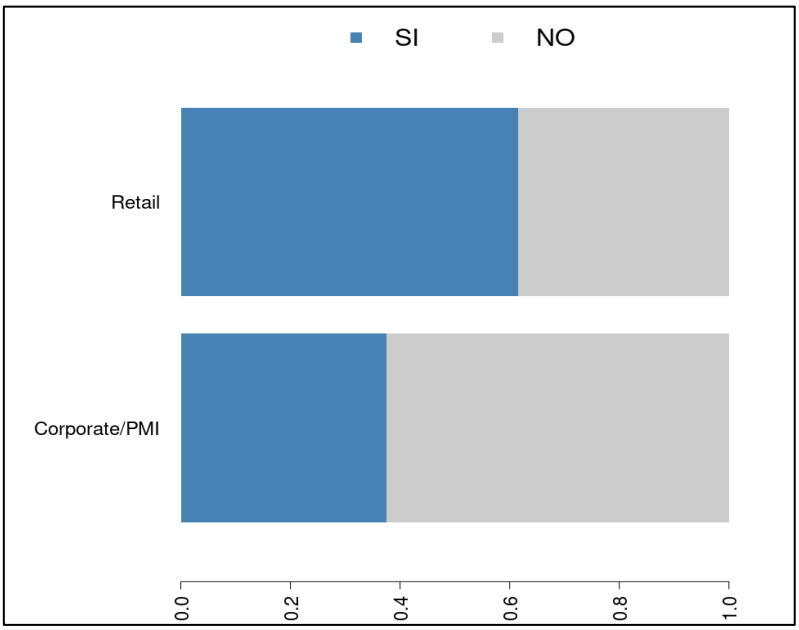


Figura 12. Proporzione di modelli per i quali è stata indicata l'adozione di una definizione di fairness, suddivisa secondo la clientela retail e corporate/PMI.

Rispetto alle diverse possibili definizioni di *fairness*, sia individuali che di gruppo, è stato chiesto agli intermediari di indicare l'approccio seguito. La totalità ha risposto di adottare la definizione di *fairness through unawareness,* che prevede l'esclusione dai modelli di attributi sensibili in grado di identificare l'appartenenza di un individuo ad un gruppo socio-economico vulnerabile, come ad esempio il genere nel caso della clientela retail. In alcuni casi gli intermediari hanno ritenuto di includere attributi potenzialmente sensibili, anche se non definiti tali nella normativa sulla privacy

(ad es. l'età, per la clientela retail), in quanto utilizzati da modelli di tipo tradizionale. Secondo gli intermediari il contributo di questi attributi alla determinazione del merito di credito sarebbe da considerarsi "trattamento differenziato" giustificabile e non discriminazione.

La misurazione della *fairness* è stato oggetto di approfondimenti con gli intermediari. Complessivamente, è emerso che l'attenzione dichiarata al tema della discriminazione non si è tradotta in pratiche di controllo organizzate in una policy aziendale in materia. Solo un intermediario ha una politica strutturata che tiene conto delle indicazioni fornite da un comitato interno di *compliance*. Un solo intermediario ha riferito di effettuare attività di controllo periodico della *fairness* del modello (nello specifico, tramite analisi controfattuale), ancorché non strutturate in un processo codificato. Infine, due intermediari hanno sviluppato modelli orientati a specifiche fasce socio-demografiche, operando una selezione del campione a partire da attributi sensibili, a monte del processo di stima. Per questi modelli, seppure non tecnicamente esplicitato, si è fatto ricorso a nozioni di *fairness* che favorissero l'estensione del credito a categorie di soggetti tipicamente esclusi a causa della loro scarsa rappresentazione nei dati creditizi.

In generale, nel corso degli incontri è emersa la mancanza di un approccio strutturato nell'affrontare la questione della *fairness* e l'eventuale presenza di meccanismi discriminatori. La percezione da parte degli intermediari è che l'uso di dati non tradizionali (altamente granulari) e di modelli intrinsecamente più opachi di quelli tradizionali non si accompagni a rischi effettivi di presenza di meccanismi discriminatori verso la clientela. Tra gli elementi più frequentemente addotti per giustificare tale approccio si citano: la circostanza che i modelli non portino a decisioni automatiche o che siano riferiti a processi diversi dalla concessione del credito; il non aver formulato ipotesi esplicitamente discriminatorie in fase di sviluppo del modello; l'assenza di meccanismi di *re-training* automatico.

Sono stati riscontrati taluni casi nei quali è emerso il ricorso a forme di profilazione di fatto della clientela. In questi casi, a fini di concessione del credito retail, sono state utilizzate informazioni relative alle abitudini di spesa (ad es. spesa in cultura, in beneficenza, in spese mediche), fattori cui il modello attribuiva peraltro limitato peso nella decisione. Gli intermediari hanno giudicato l'impiego di tali informazioni come non ascrivibile a pratiche di profilazione del comportamento, in quanto "dati finanziari". In un caso il modello utilizza attributi non finanziari (ad es. modalità e tempi di navigazione delle pagine web dell'intermediario) per definire il tasso di interesse massimo che può essere applicato al cliente a partire comunque da un valore coerente con il suo merito di credito.

6.5 Spiegabilità del modello

Nel questionario è stato chiesto agli intermediari di indicare se e quali tecniche di spiegabilità abbiano utilizzato e chi siano stati i principali destinatari dei relativi output (ad esempio, sviluppatori, validatori interni, clientela esterna, ecc.).

Nessun intermediario ha dichiarato di utilizzare tecniche di *explainable modelling* per sviluppare modelli intrinsecamente spiegabili, pertanto le tecniche di spiegabilità adottate dagli intermediari sono risultate quelle di *pre-modelling explainability* o di *post hoc explainability* (cfr. sez. 2.3).

Le tecniche di *pre-modeling explainability* sono finalizzate alla comprensione delle caratteristiche del *dataset* utilizzato per addestrare il modello. Gli intermediari intervistati hanno utilizzato per questo fine tecniche statistiche tradizionali, come l'analisi esplorativa dei dati o la creazione e l'uso di variabili interpretabili.

Le tecniche di *post hoc explainability* hanno invece l'obiettivo di spiegare a posteriori la logica di un modello, e sono particolarmente utili nel caso di modelli complessi come quelli di AI. Tali tecniche risultano utilizzate per la quasi totalità dei modelli (92%). Le tecniche di *post hoc explainability* più utilizzate sono risultate gli *Shapley Values* e la *Feature Importance* (indicate rispettivamente nel 71% e 39% dei modelli), seguite da *Partial Dependence Plot*[49] (19%) e LIME (6%). Tali tecniche sono disponibili *off-the-shelf* nei principali linguaggi di programmazione utilizzati per sviluppare modelli di AI (Python, R, ecc.), caratteristica che ha probabilmente contribuito al loro impiego. La frequenza di utilizzo delle tecniche di spiegabilità nei modelli è mostrata nella figura 13.

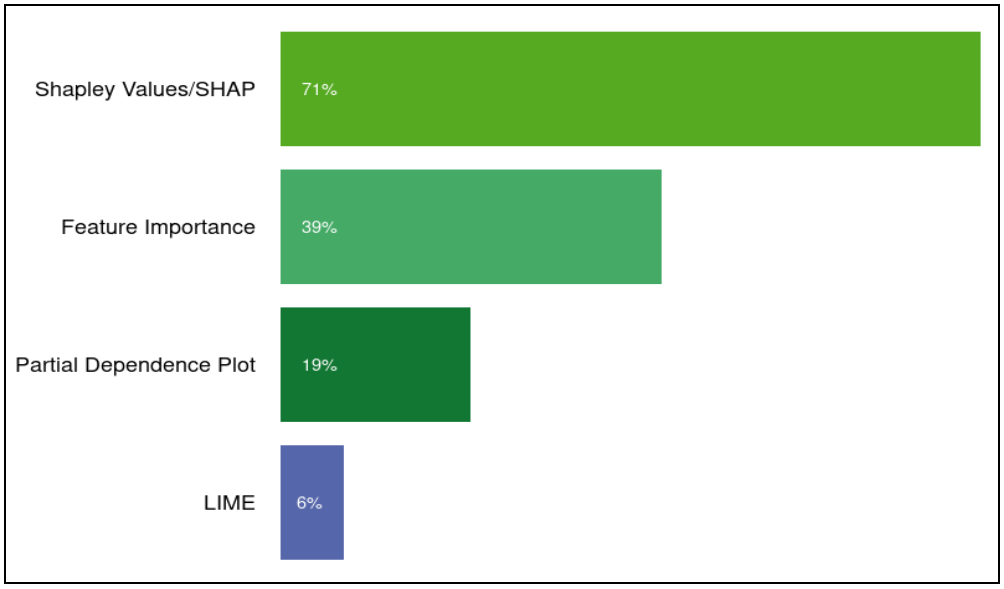


Figura 13. Frequenza di utilizzo delle tecniche di spiegabilità post hoc (% sul totale dei modelli che hanno adottato tecniche di spiegabilità post hoc)

La figura 14 mostra invece i destinatari dei risultati dell'applicazione delle suddette tecniche. Sviluppatori e utenti sono i destinatari principali (indicati rispettivamente nel 84% e 79% dei modelli), seguiti dai validatori interni (68%) e dal top management (61%). Nonostante le previsioni normative ex-GDPR impongano che su richiesta il cliente riceva informazioni sulla logica utilizzata dal modello con cui è stata determinata la valutazione che lo riguarda, al momento nessun intermediario ha indicato tra i destinatari delle tecniche di spiegabilità la clientela esterna. Al riguardo gli intermediari hanno sottolineato che il modello in molti casi è uno strumento di supporto

[49] Il ricorso al *Partial Dependence Plot* consente di valutare l'impatto di variazioni di un singolo attributo, all'interno di un determinato intervallo, sul valore assunto dalla variabile restituita in output dal modello. L'uso di tale strumento ha come obiettivo quello di investigare e visualizzare le possibili non-linearità nelle relazioni tra variabili input del modello e l'output di questo.

che produce una proposta oggetto di validazione da parte dell'analista, che effettua la sua valutazione anche sulla base di altri elementi.

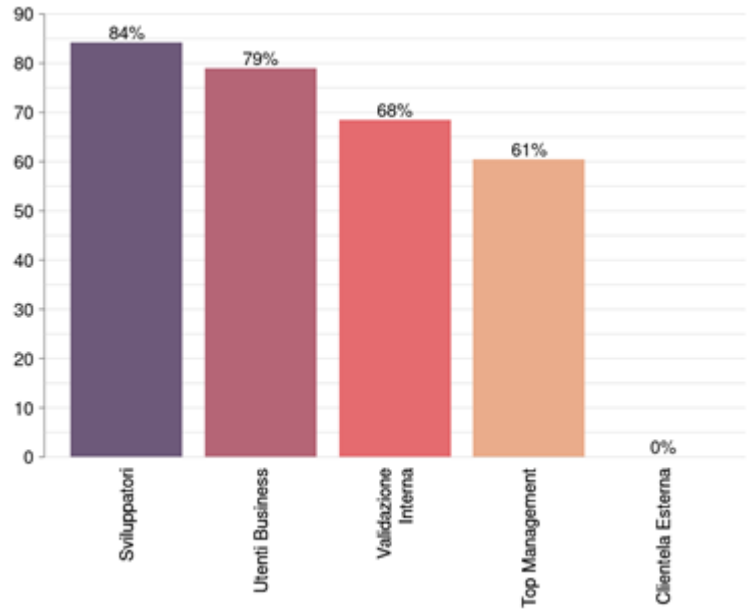


Figura 14. Destinatari delle tecniche di spiegabilità post hoc (dati per modello)

6.6 Governance

Gli aspetti legati alla governance dei processi che utilizzano i modelli sono di fondamentale importanza; questi assumono una rilevanza ancora maggiore nel caso di sistemi di AI a causa dell'opacità della struttura interna e della logica decisionale seguita. È necessario pertanto che gli intermediari istituiscano appropriati presidi per controllare che i sistemi di AI diano un effettivo valore aggiunto nei processi in cui vengono utilizzati e non generino rischi incrementali di difficile gestione. I controlli necessari spaziano dalla verifica della performance dei modelli e della stabilità degli output nel tempo, alla verifica della qualità e integrità dei dati utilizzati[50], fino all'identificazione e al presidio dei rischi connessi all'utilizzo di tali tecniche.

Con riferimento ai risultati dell'indagine, tutti gli intermediari hanno istituito o intendono istituire un processo di monitoraggio e reportistica dei sistemi AI utilizzati. I report in uso sono risultati concentrati per la quasi totalità dei casi sull'accuratezza degli output (97%); per una gran parte (74%) viene anche verificata la stabilità delle stime nel tempo. Meno presenti le metriche relative alla

[50] Il mancato monitoraggio della qualità e integrità dei dati di input potrebbe comportare problemi di *data drift* (variazione delle relazioni ipotizzate o stimate derivanti dalla modifica nel tempo dei dati forniti in fase di addestramento) e richiedere pertanto opportuni riaddestramenti.

qualità e integrità dei dati utilizzati dai modelli, che risultano oggetto di reportistica per poco più della metà dei casi[51].

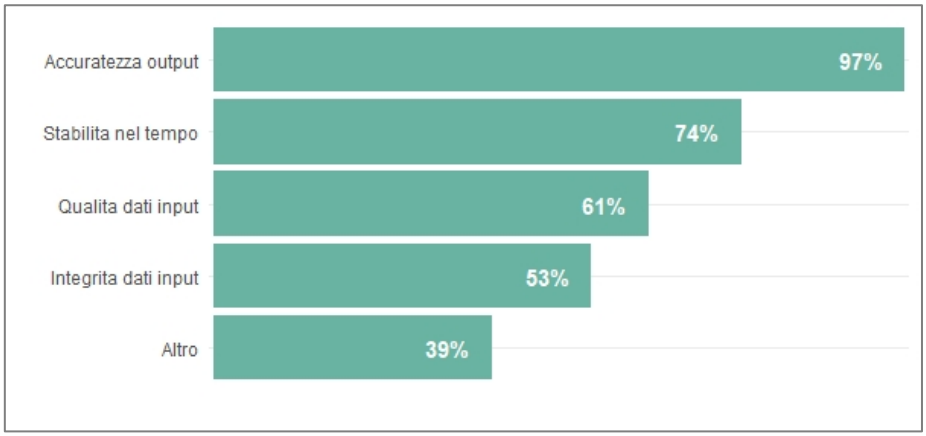


Figura 15. Metriche oggetto del reporting (dati per modello)

Quanto alla strategia di aggiornamento dei sistemi AI, nel 66% dei casi vengono applicati sistemi di addestramento semi-automatici (ovvero parzialmente supervisionati da un intervento umano) e nel 29% sistemi di *re-training* manuale. Solamente un intermediario risulta fare uso, per due modelli, di una procedura di aggiornamento interamente automatica. Secondo quanto riferito, addestrare il modello mensilmente migliora la performance in termini di accuratezza; non è ritenuta rilevante la possibile discontinuità tra le successive versioni dei modelli in quanto questi sono utilizzati per il monitoraggio delle posizioni.

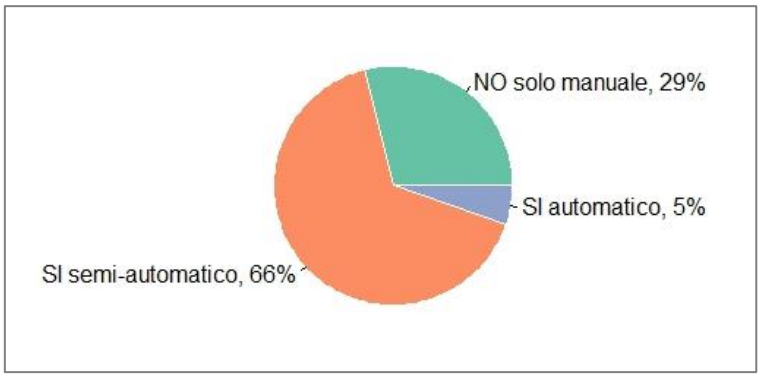


Figura 16. Sistema di re-training del modello

Con riferimento ai benefici e ai rischi connessi all'utilizzo delle tecniche di AI, il feedback degli intermediari mostra un generale ottimismo: l'utilizzo di tali tecniche comporterebbe molti benefici e pochi rischi rispetto alle tecniche tradizionali. Il maggior beneficio è ritenuto essere l'incremento dell'accuratezza delle previsioni, considerato elemento rilevante da tutti gli intermediari indipendentemente dalle finalità di utilizzo dei sistemi di AI. Ulteriori benefici ritenuti generalmente rilevanti riguardano la possibilità di utilizzare fonti alternative di dati, la maggiore efficienza nei

[51] Sono inoltre state dichiarate come oggetto di reporting (sotto la voce "Altro" in fig. 15) altre metriche fra le quali si citano le verifiche sulla coerenza fra le distribuzioni dei dati correnti con quelli usati nel training, i controlli sulla distribuzione degli score statistici di monitoraggio e forme di *backtesting*.

processi e l'accesso al credito da parte di una clientela più ampia. Meno rilevante è generalmente ritenuto il beneficio percepito in tema di stabilità di performance del modello nel tempo.

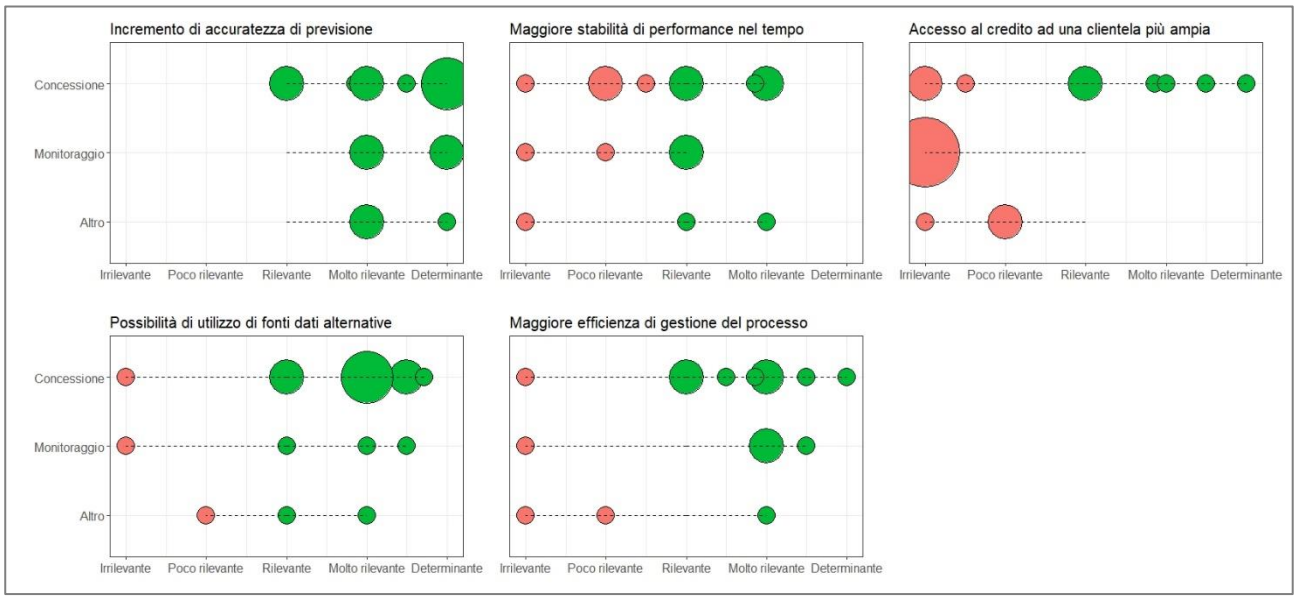


Figura 17. Benefici percepiti connessi all'utilizzo di sistemi AI (dati per intermediario, media dei modelli attivi per ciascuna fase). Il diametro dei cerchi riflette il numero di intermediari indicanti lo stesso livello di rilevanza.

Con riferimento ai rischi incrementali percepiti rispetto alle tecniche tradizionali l'evidenza della rilevazione mostra che l'attenzione degli intermediari è soprattutto concentrata sui rischi operativi: per i modelli di concessione 5 intermediari su 9 ritengono che i rischi operativi siano rilevanti o molto rilevanti (2 su 4 per i modelli di monitoraggio). Gli altri rischi, tra cui quelli reputazionali, legali e della sicurezza IT sono ritenuti nella maggior parte dei casi irrilevanti o poco rilevanti.

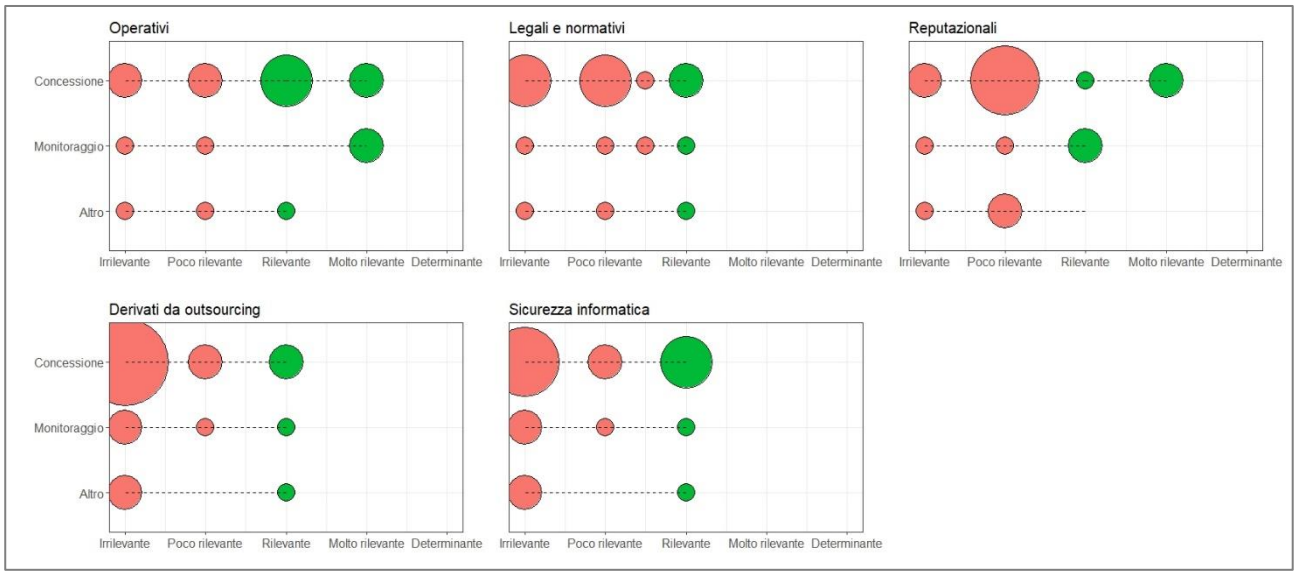


Figura 18. Rischi percepiti connessi all'utilizzo di sistemi AI (dati per intermediario, media dei modelli attivi per ciascuna fase). Il diametro dei cerchi riflette il numero di intermediari indicanti lo stesso livello di rilevanza.

Per quanto riguarda il ricorso all'outsourcing nello sviluppo e gestione dei modelli, è necessario adottare presidi utili a consentire agli intermediari di mantenere il pieno governo della variabile

modellistica, con particolare riferimento ad alcune fasi del processo (es. monitoraggio della performance e dell'assenza di bias nella selezione, riaddestramento, trasparenza nei confronti del cliente).

Sul tema il 47 per cento dei modelli ha visto l'impiego anche di personale esterno nello sviluppo; solo l'8 per cento dei modelli è stato completamente sviluppato da personale esterno. In quest'ultimo ambito si sono rilevate due forme di completa esternalizzazione:

- l'outsourcing totale presso una società esterna, la quale ha sviluppato in autonomia il modello e fornisce direttamente lo score finale all'intermediario;
- il caso di "doppio outsourcing" dove il fornitore mette a disposizione di banche e finanziarie il risultato di un modello di *credit scoring* sviluppato in collaborazione con un soggetto terzo.

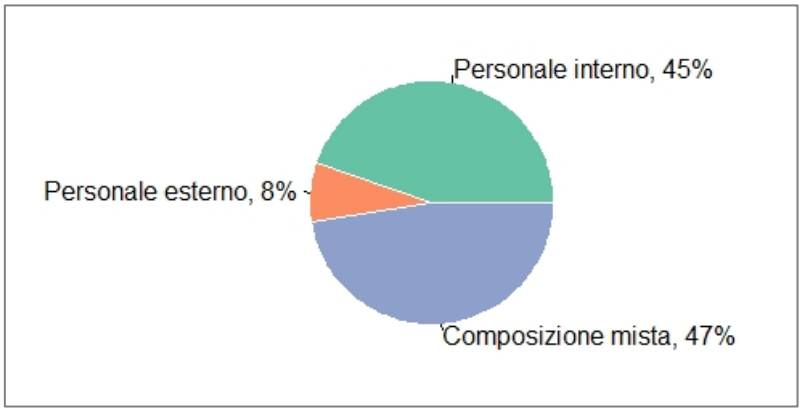


Figura 19. Composizione dei gruppi di lavoro (dati per modello)

Gli utenti di business che utilizzano i modelli sono risultati sufficientemente coinvolti nella fase di disegno e sviluppo (82 per cento a coinvolgimento medio o alto), così come elevata sarebbe la loro comprensione delle assunzioni poste alla base dei modelli (76 per cento ad un livello medio-alto). Sensibilmente meno elevato risulta invece il livello di conoscenza generale delle tecniche di AI da parte degli utenti di business, configurabile come di livello medio-basso.

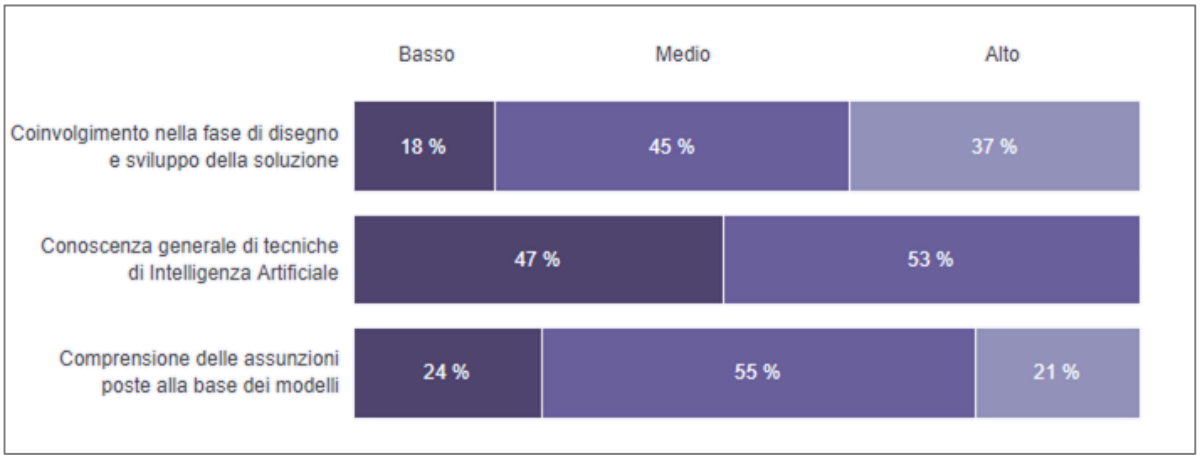


Figura 20. Coinvolgimento utenti di business in tema di AI (dati per modello)

È stato infine chiesto agli intermediari quali informazioni vengono fornite ai clienti in caso di richiesta circa la logica utilizzata per la decisione assunta con un modello di ML, in ottemperanza ai

dettami dell'art. 15 GDPR. Le informazioni fornite ai clienti sono generalmente scarne, anche in considerazione dell'assunto che il modello è utilizzato ordinariamente a supporto di una decisione creditizia rimessa ad un analista.

Appendice 1 – Tecniche per la rilevazione dei bias in caso di ricorso a modelli AI-ML

Di seguito sono riportate alcune tecniche che permettono di contribuire alla rilevazione delle distorsioni che possono emergere all'interno delle diverse fasi del ciclo di sviluppo di un algoritmo di AI-ML: nella raccolta dei dati, nella specificazione del modello e nell'apprendimento e nell'analisi degli output.

Raccolta dati

Le forme di distorsione la cui origine è da ricercarsi nella fase di raccolta e preparazione dei dati sono molteplici, e sono tipicamente rilevate per mezzo di un'approfondita analisi esplorativa dei dati. Includono:

i) i meccanismi di selezione del campione, ad esempio non bilanciamento delle classi per i *dataset* tradizionali e *self-selection bias* per i big data;

ii) i meccanismi di *missingness* non casuali e relativa gestione dei dati mancanti;

iii) la presenza di osservazioni incomplete rispetto alla variabile target, tra cui *thin-file* e *no-file*, *digital footprint* non valide;

iv) il bias storico nei dati e bias istituzionale, esogeni rispetto al processo di modellizzazione.

La rilevazione della distorsione in questa fase avviene tipicamente mediante processi di elicitazione[52] (Manzi et al., 2019) e di sviluppo collettivo (*crowdsourcing*).

La mitigazione della distorsione può avvenire, a seconda delle applicazioni e del quadro normativo che le regola, mediante la selezione mirata di sottoinsiemi di osservazioni o l'attribuzione di pesi che garantiscano l'adeguata rappresentatività del *training set*.

Specificazione del modello e apprendimento

In fase di definizione del modello, eventuali forme di distorsione vengono rilevate sotto forma di deviazioni rispetto ai formalismi ontologici per il ragionamento. Questi sono dei costrutti definiti dall'analista per esplicitare l'insieme di ipotesi sottostanti la formulazione del modello, e le

[52] L'elicitazione della distorsione è un processo che richiede ad un gruppo di esperti di valutare tecnicamente un dato modello nelle sue diverse assunzioni e componenti, al fine di identificare i diversi punti critici. Il ricorso a tale esercizio è tipicamente associato all'impiego di modelli econometrici tradizionali ed è stato successivamente esteso al ML. Operativamente, consiste nel valutare i benefici, ovvero i rischi, apportati dall'inclusione di un attributo, rilevato secondo una determinata scala, o di singole unità di osservazione.

implicazioni derivanti dalle relazioni tra le stesse, secondo un approccio analogo alla già citata elicitazione della distorsione.

Tra gli aspetti critici, da controllare, si citano il bilanciamento tra distorsione causato da *omitted variable* e da *included variable,* la scelta del modello (ad esempio favorendo un modello maggiormente interpretabile a parità di performance) e della funzione di costo/obiettivo in fase di addestramento (ad esempio, ricorrendo all'inclusione di termini di regolarizzazione o penalizzazione per il controllo della distorsione algoritmica)[53].

Inoltre, verificato il bilanciamento ottimale tra validità esterna ed interna di un modello, è buona norma elaborare dei presidi di monitoraggio della stabilità di queste, nel tempo o nei diversi contesti d'applicazione. Tali presidi sono tipicamente mirati a prevenire e mitigare degradazioni di performance in presenza di dinamiche di deriva (*drift*), causate da variazioni nel fenomeno oggetto di analisi o in parte delle sue caratteristiche. Tra gli altri, si citano gli schemi di aggiornamento periodico (*re-training*) o continuo (*online learning*) del modello, nonché di selezione dinamica delle *features* impiegate (*feature dropping*), volti a prevenire e mitigare meccanismi di deriva.

Analisi degli output

La distorsione può intervenire a valle del processo di stima e calibrazione di un modello, in assenza di adeguata comprensione dei meccanismi che hanno prodotto i risultati. A tal fine, sono critici i criteri per l'analisi:

i) dei meccanismi di causalità e delle relazioni di dipendenza, al fine di arginare il *cause-effect bias* o l'*aggregation bias*;

ii) della giustificabilità delle decisioni determinate dal valore assunto da attributi sensibili;

iii) della spiegabilità del modello, ricorrendo a tecniche di *eXplainable* AI in fase di sviluppo.

[53] L'inclusione di vincoli in grado di fornire garanzie a tutela della *fairness* può in alcuni casi comportare una riduzione dell'accuratezza; ad es. Hardt et al., (2016); Hickey et al., (2020).

Appendice 2 - Letteratura economica sulla discriminazione nel mercato del credito e sul contributo dei metodi quantitativi e dei modelli ML

La letteratura economica sulla discriminazione nel mercato del credito si è occupata prevalentemente di verificare la presenza o meno di discriminazione connessa con l'appartenenza a determinati gruppi etnici, particolarmente nel mercato nord-americano e britannico, e di discriminazione di genere.

Vi sono diversi lavori secondo cui le imprese di proprietà di alcuni gruppi (principalmente gli afroamericani) sono discriminate generalmente sotto forma di un maggiore razionamento della quantità di credito (ad esempio Cavalluzzo e Cavalluzzo (1998), Cavalluzzo et al. (2002) e Blanchflower et al. (2003), Fraser (2009a)).

Alcuni studi riscontrano la presenza di discriminazione etnica nel mercato dei prestiti alle famiglie. In particolare, il tasso di rifiuto di una domanda di mutuo risulta più elevato per alcuni gruppi etnici rispetto alla popolazione di origine caucasica, a parità di caratteristiche del mutuatario e del prestito (Munnell et al. (1996); Ross e Yinger, 2002; Ross e Tootell (2004)). Il risultato è confermato anche quando si tiene conto delle caratteristiche della zona di residenza dei richiedenti, inclusa la prevalenza o meno in essi di gruppi etnici svantaggiati (Tootell, 1996).

Altre analisi forniscono evidenze sui tassi di interesse praticati: Edelberg (2007) mostra che i tassi sul credito al consumo e sui mutui presentano un'ampia eterogeneità non spiegata, particolarmente prima del 1995, attribuibile potenzialmente a discriminazione. Evidenze dirette di un costo del credito più elevato per alcuni gruppi etnici sono presenti anche in Bayer et al. (2018), Ghent et al. (2014) e Cheng et al. (2015). Tuttavia, negli Stati Uniti il costo del credito è costituito da diverse componenti quindi potrebbe non essere sufficiente confrontare i tassi di interesse per dimostrare la presenza di discriminazione. In un recente lavoro si mostra infatti che i gruppi etnici svantaggiati pagano tassi di interesse più elevati ma tendono a selezionare contratti di mutuo con costi fissi iniziali più contenuti (Bhutta e Hizmo, 2021). Lo studio non approfondisce le ragioni delle differenze osservate ma suggerisce che potrebbero riflettere diverse preferenze oppure divari nelle disponibilità liquide. Inoltre, inefficienze nel mercato dei mutui, che non favoriscono la clientela nella ricerca del contratto più vantaggioso, potrebbero avere effetti differenziati tra categorie di famiglie in ragione delle diverse competenze in materia finanziaria (Woodward, 2008; Woodward e Hall, 2012).

La discriminazione tra gruppi potrebbe inoltre manifestarsi nella qualità dei servizi offerti. Hanson et al. (2016) mostrano, attraverso un esperimento condotto sempre negli Stati Uniti, che il tasso di risposta ai potenziali mutuatari che chiedono informazioni riguardo ai contratti offerti è inferiore per alcuni gruppi etnici e che a questi gruppi vengono fornite meno informazioni dal personale degli intermediari.

Nella letteratura vi sono diverse evidenze che documentano la presenza di discriminazione di genere. La discriminazione sarebbe "*taste-based*", cioè basata sul pregiudizio, piuttosto che di tipo statistico, ossia risultante dalla correlazione tra genere e caratteristiche rilevanti per l'identificazione del merito di credito ma non osservabili (Hisrich e Brush, 1984; Buttner e Rosen 1988, 1989). Le analisi mostrano che le imprenditrici hanno maggiori difficoltà nell'accesso al mercato del credito rispetto agli imprenditori (Fay e Williams, 1993; Cavalluzzo et al., 2002; Fraser, 2009b), soprattutto nella fase di start-up (Orser et al., 2000); devono inoltre conferire maggiori garanzie e pagano tassi di interesse più elevati (Coleman, 2000). Alcune evidenze basate su più paesi suggeriscono, d'altra parte, che le differenze di accesso al credito a discapito delle imprenditrici sarebbero più accentuate nei contesti dove il sistema finanziario è meno sviluppato (Muravyev et al., 2009).

Evidenze a sostegno della presenza di discriminazione, sia etnica sia di genere, sono disponibili anche per l'Italia, con riferimento al credito alle imprese. In generale, gli studi italiani ricorrono a basi dati molto dettagliate che consentono di inserire nelle analisi un ampio insieme di caratteristiche relative alle imprese, alla tipologia contrattuale, all'intermediario; in tal modo è maggiormente possibile verificare la presenza di differenze sistematiche nell'accesso al credito tra gruppi di prenditori a parità di condizioni. Albareto e Mistrulli (2011) mostrano che, tra il 2004 e il 2008, i tassi di interesse applicati dalle banche italiane ai prestiti alle micro-imprese di proprietà di immigrati sono superiori di circa 70 punti base a quelli pagati dagli imprenditori italiani, a parità di caratteristiche dell'impresa, ma che il differenziale si riduce con l'allungamento della storia creditizia degli imprenditori. Secondo gli autori, da un lato la conoscenza acquisita dalle banche nel corso della relazione creditizia permette di superare le maggiori asimmetrie informative presenti ex ante, dall'altro gli imprenditori migranti migliorano la loro conoscenza del mercato bancario italiano e riescono a ottenere condizioni migliori con il passare del tempo.

Bellucci et al. (2010), analizzando le linee di credito concesse da una grande banca italiana a ditte individuali in due province tra il 2004 e il 2006, pongono in evidenza che le imprenditrici hanno maggiori difficoltà di accesso al credito e, anche quando non pagano tassi di interesse più elevati, hanno una probabilità significativamente maggiore di dover conferire garanzie reali. Evidenze simili risultano dallo studio di Calcagnini et al. (2015), basato sui dati relative alle linee di credito concesse da tre banche italiane nel periodo 2005-2008; nel campione analizzato le imprese femminili sono significativamente più piccole e giovani di quelle di proprietà di uomini e tendono a ricorrere a prestiti di minore importo e meno garantiti; il minore accesso al credito è solo in parte spiegato da queste caratteristiche poiché, anche considerando tale fattore, permane una differenza a svantaggio delle imprese di proprietà di donne, che devono conferire più frequentemente garanzie.

Evidenza di discriminazione di genere è presente anche nel lavoro di Alesina, Lotti e Mistrulli (2013), che utilizza dati tratti dalla Centrale dei rischi su finanziamenti concessi a microimprese nel periodo 2004-2007. L'analisi, che tiene conto di un insieme molto ampio di caratteristiche delle aziende e del mercato locale del credito, mostra che le imprenditrici pagano tassi di interesse più elevati, anche quando non sono più rischiose di quelle di proprietà di uomini. Il differenziale di tasso è

contenuto ma non scompare se si tiene conto anche di differenze nella storia creditizia. Sempre con dati a livello banca-impresa tratti dalla Centrale dei rischi, Cesaroni et al. (2013) mostrano che, controllando per tutte le caratteristiche osservabili disponibili, le aziende di proprietà di donne hanno dovuto affrontare nel periodo 2007-2009 un inasprimento più pronunciato nelle condizioni di offerta di credito rispetto alle altre imprese.

Le analisi citate confrontano le condizioni praticate tra categorie di clienti ma non approfondiscono l'eventuale contributo alla discriminazione dei metodi quantitativi di valutazione del merito di credito. Alcune analisi hanno tentato di gettare luce sull'impatto del *credit scoring* negli Stati Uniti, dove la diffusione dell'utilizzo dei metodi statistici di *credit scoring* è stata accompagnata da un ampio dibattito sugli effetti sull'accesso al credito (cfr. *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*, 2007). I risultati non sono concordi e dipendono dai dati utilizzati e dalle metodologie di investigazione. Avery et al. (2009 e 2012) concludono che l'adozione dei metodi quantitativi diminuirebbe la discriminazione mentre altre analisi suggeriscono che alcune caratteristiche, come l'età, possano determinare differenze nella classificazione della clientela non giustificate da effettive differenze nel merito di credito.

Più recentemente, alcuni lavori di ricerca analizzano l'effetto sulla discriminazione dell'adozione di tecniche di valutazione del merito di credito basate su algoritmi di AI nel confronto con i modelli tradizionali di *credit scoring*. Bartlett, Morse, Stanton e Wallace (2021) stimano i differenti effetti sulla discriminazione che derivano dall'uso di tecniche di valutazione del merito creditizio tradizionali e innovative nel mercato dei mutui degli Stati Uniti; la loro analisi mostra che i debitori latino-americani e afroamericani, con caratteristiche equivalenti agli altri clienti, pagano tassi più elevati di circa 8 punti base (4 per il rifinanziamento di mutui in essere) e che nei prestiti concessi da imprese FinTech le disparità tariffarie sono inferiori di circa un terzo e non vi sono differenza nei tassi di diniego. Sempre con riferimento ai mutui immobiliari americani, Fuster, Goldsmith-Pinkham, Ramadorai e Walther (2020) confrontano in un esercizio di simulazione la capacità di prevedere il default di modelli tradizionali e modelli di ML, nonché i tassi di interesse che verrebbero applicati sulla base del rischio di credito. La loro evidenza suggerisce che i modelli di ML sono più accurati nel prevedere il default delle tecnologie meno sofisticate e che tendono ad ampliare l'accesso al mercato, riducendo la dispersione tra tassi di accettazione delle richieste tra diversi sottogruppi della popolazione. I modelli di ML determinano anche una maggiore differenziazione nel costo del credito tra clienti; considerando l'insieme dei clienti che avrebbero credito sulla base di entrambe le tecnologie, tradizionale e innovativa, risulterebbero penalizzati quelli più avversi al rischio all'interno dei gruppi degli afroamericani e degli ispanici.

Un'altra simulazione, di Bono, Croxon e Giles (2021), ricorre a un ampio insieme di dati molto dettagliati per il Regno Unito per verificare gli effetti di un ipotetico passaggio da un sistema di valutazione con modello tradizionale di *credit scoring* a un modello di ML. L'analisi conferma la maggiore accuratezza del modello di ML e suggerisce che questo non aggravi né elimini potenziali distorsioni nei confronti di gruppi caratterizzati da attributi sensibili, come l'etnia e il genere della clientela.

Nel complesso, l'evidenza finora disponibile non indicherebbe che il ML abbia il potenziale di amplificare la discriminazione, se presente, rispetto ai metodi statistici tradizionali.

Riferimenti bibliografici

Relativi al testo e all'appendice 1

Albanesi, S., & Vamossy, D. F. (2019). Predicting consumer default: A deep learning approach (No. w26165). National Bureau of Economic Research.

Alonso, A., & Carbó, J.M. (2020). Machine Learning in Credit Risk: Measuring the Dilemma between Prediction and Supervisory Cost, Banco de España WP 2032.

Bacham, D., & Zhao, J. (2017). Machine learning: challenges, lessons, and opportunities in credit risk modeling. Moody's Analytics Risk Perspectives, 9, 30-35.

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54-61.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Bazarbash, M. (2019). Fintech in financial inclusion: machine learning applications in assessing credit risk. International Monetary Fund.

Bellomarini, L., Gottlob, G., & Sallinger, E. (2018). The vadalog system: Datalog-based reasoning for knowledge graphs. arXiv preprint arXiv:1807.08709.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845-2897.

Beydoun, G., Suryanto, H., Guan, C., Guan, A., & Sugumaran, V. Unlocking Knowledge Graphs (KG) Potentials in Support of Credit Risk Assessment.

Bryant, K. (2001). ALEES: an agricultural loan evaluation expert system. Expert systems with applications, 21(2), 75-85.

Calo, R. (2013). Digital market manipulation. Geo. Wash. L. Rev., 82, 995.

Cascarino, G., Moscatelli, M., & Parlapiano, F. (2022). Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning. Bank of Italy Occasional Paper, (674).

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. Big data, 5(2), 120-134.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. Methodology and computing in applied probability, 11(1), 29-45.

Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: lessons from peer-to-peer lending. International Journal of Industrial Organization, 51, 185-222.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2020). Predictably unequal? the effects of machine learning on credit markets. (October 1, 2020).

Gottlob, G., & Pieris, A. (2015, June). Beyond SPARQL under OWL 2 QL entailment regime: Rules to the rescue. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

Hacker, P. & Passoth J. H. (2021). Varieties of AI Explanations Under The Law. From the GDPR to the AIA, and Beyond (August 25, 2021).

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7), 1445-1459.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 3315-3323.

Hickey, J. M., Di Stefano, P. G., & Vasileiou, V. (2020). Fairness by Explicability and Adversarial SHAP Learning. arXiv preprint arXiv:2003.05330.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge, 12(2), 1-257.

Iwasieczko, B., Korczak, J., Kwiecień, M., & Muszyńska, J. (1986). Expert system in financial analysis. IFAC Proceedings Volumes, 19(17), 113-120.

Jagtiani, J., & Lemieux, C. (2017). Fintech lending: Financial inclusion, risk pricing, and alternative information.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125-5131.

Manzi, G., & Forster, M. (2019). Biases in bias elicitation. Communications in Statistics-Theory and Methods, 48(18), 4656-4674.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. Expert Systems with Applications, 161, 113567.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, 74, 26-39.

Piketty, L. (1987, April). The authorizer's assistant: a large commercial expert system application. In Proceedings of the AI and advanced computer technology conference. Long Beach, CA.

Roa, L., Rodríguez-Rey, A., Correa-Bahnsen, A., & Valencia, C. (2021). Supporting Financial Inclusion with Graph Machine Learning and Super-App Alternative Data. arXiv preprint arXiv:2102.09974.

Tobback, E., & Martens, D. (2019). Retail credit scoring using fine-grained payment data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4), 1227-1246.

Yuan, D. (2015). Applications of machine learning: consumer credit risk analysis (Doctoral dissertation, Massachusetts Institute of Technology).

Zocco, D. P. (1985). A framework for expert systems in bank loan management. Journal of Commercial Bank Lending, 67(2), 47-55.

Relativi all'appendice 2

Albareto, G., & Mistrulli, P. E. (2011). Bridging the gap between migrants and the banking system. Bank of Italy Temi di Discussione (Working Paper) No, 794.

Alesina, A. F., Lotti, F., & Mistrulli, P. E. (2013). Do women pay more for credit? Evidence from Italy. Journal of the European Economic Association, 11(suppl_1), 45-66.

Avery, R. B., Brevoort, K. P., & Canner, G. B. (2009). Credit scoring and its effects on the availability and affordability of credit. Journal of Consumer Affairs, 43(3), 516-537.

Avery, R. B., Brevoort, K. P., & Canner, G. (2012). Does Credit Scoring Produce a Disparate Impact?. Real Estate Economics, 40, S65-S114.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2021). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics.

Bayer, P., Ferreira, F., & Ross, S. L. (2018). What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders. The Review of Financial Studies, 31(1), 175-205.

Bellucci, A., Borisov, A., & Zazzaro, A. (2010). Does gender matter in bank–firm relationships? Evidence from small business lending. Journal of Banking & Finance, 34(12), 2968-2984.

Bhutta, N., & Hizmo, A. (2021). Do minorities pay more for mortgages?. The Review of Financial Studies, 34(2), 763-789.

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. J. (2003). Discrimination in the small-business credit market. Review of Economics and Statistics, 85(4), 930-943.

Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. Oxford Review of Economic Policy, 37(3), 585-617.

Buttner, E. H., & Rosen, B. (1988). Bank loan officers' perceptions of the characteristics of men, women, and successful entrepreneurs. Journal of Business venturing, 3(3), 249-258.

Buttner, E. H., & Rosen, B. (1989). Funding new business ventures: Are decision makers biased against women entrepreneurs?. Journal of Business Venturing, 4(4), 249-261.

Calcagnini, G., Giombini, G., & Lenti, E. (2015). Gender differences in bank loan access: An empirical analysis. Italian Economic Journal, 1(2), 193-217.

Cavalluzzo, K. S., & Cavalluzzo, L. C. (1998). Market structure and discrimination: The case of small businesses. Journal of Money, Credit and Banking, 771-792.

Cavalluzzo, K. S., Cavalluzzo, L. C., & Wolken, J. D. (2002). Competition, small business financing, and discrimination: Evidence from a new survey. The Journal of Business, 75(4), 641-679.

Cesaroni, F. M., Lotti, F., & Mistrulli, P. E. (2013). Female Firms and Banks' Lending Behaviour: What Happened during the Great Recession?. Bank of Italy Occasional Paper, (177).

Cheng, P., Lin, Z., & Liu, Y. (2015). Racial discrepancy in mortgage interest rates. The Journal of Real Estate Finance and Economics, 51(1), 101-120.

Coleman, S. (2000). Access to capital and terms of credit: A comparison of men-and women-owned small businesses. Journal of small business management, 38(3), 37.

Edelberg, W. (2007). Racial dispersion in consumer credit interest rates.

Fay, M., & Williams, L. (1993). Gender bias and the availability of business loans. Journal of Business Venturing, 8(4), 363-376.

Fraser, S. (2009a). Is there ethnic discrimination in the UK market for small business credit?. International Small Business Journal, 27(5), 583-607.

Fraser, S. (2009b). Small firms in the credit crisis: Evidence from the UK survey of SME finances. Warwick Business School, University of Warwick.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2020). Predictably unequal? the effects of machine learning on credit markets. The Effects of Machine Learning on Credit Markets (October 1, 2020).

Ghent, A. C., Hernandez-Murillo, R., & Owyang, M. T. (2014). Differences in subprime loan pricing across races and neighborhoods. Regional Science and Urban Economics, 48, 199-215.

Hanson, A., Hawley, Z., Martin, H., & Liu, B. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. Journal of Urban Economics, 92, 48-65.

Hisrich, R., & Brush, C. (1984). The woman entrepreneur: Management skills and business problems. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.

Munnell, A. H., Tootell, G. M., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. The American Economic Review, 25-53.

Muravyev, A., Talavera, O., & Schäfer, D. (2009). Entrepreneurs' gender and financial constraints: Evidence from international data. Journal of comparative economics, 37(2), 270-286.

Orser, B. J., Hogarth-Scott, S., & Riding, A. L. (2000). Performance, firm size, and management problem solving. Journal of small business management, 38(4), 42.

Ross, S. L., & Tootell, G. M. (2004). Redlining, the Community Reinvestment Act, and private mortgage insurance. Journal of Urban Economics, 55(2), 278-297.

Ross, S. L., & Yinger, J. (2002). The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement. MIT press.

Tootell, G. M. (1996). Redlining in Boston: Do mortgage lenders discriminate against neighborhoods?. The Quarterly Journal of Economics, 111(4), 1049-1079.

Woodward, S. E., & Hall, R. E. (2012). Diagnosing consumer confusion and sub-optimal shopping effort: Theory and mortgage-market evidence. American Economic Review, 102(7), 3249-76.

Woodward, S. E. (2008). A study of closing costs for FHA mortgages. US Department of Housing and Urban Development, Office of Policy Development and Research.