Dall'Etica per l'AI all'AI per l'Etica

Paolo Giudici

Dipartimento di Scienze Economiche ed Aziendali, Universita' di Pavia
Coordinatore dei progetti Europei H2020 FIN-TECH e PERISCOPE

EU AI act: un approccio etico per l'AI

L'AI Act proposto dalla commissione Europea, in corso di discussione, distingue tre tipi di applicazioni dell' AI:

- Applicazioni proibite: Esempio: identificazione biometrica.
- Applicazioni ad alto rischio: permesse ma soggette ad un modello di risk management che ne valuta i rischi nel tempo. Esempio: merito di credito.
- Applicazioni a rischio limitato: permesse e soggette solo ad eventuali obblighi informativi. Esempio: chatbot.

SAFE-AI: valutare l'etica delle decisioni

- In alcuni recenti pubblicazioni di ricerca (Giudici e Raffinetti, 2023; Giudici, Centurelli e Turchetti, 2023) abbiamo proposto modelli di AI per valutare l'eticitá delle decisioni (in particolare di quelle dell'AI).
- Dall'AI Act abbiamo dedotti quattro principi misurabili, al fine di rendere operativi i modelli di AI risk management.
- Il modello é stato denominato S.A.F.E. da: Sustainability, Accuracy, Fairness, Explainability.

SAFE-AI: il modello

- Sustainability: un sistema di AI dovrebbe raggiungere un livello adeguato di robustezza e sicurezza cyber, ed essere resiliente ad anomalie interne ed attacchi esterni. (AI Act, Articoli 15.1, 15.3,15.4);
- Accuracy: un sistema di AI dovrebbe raggiungere un adeguato livello di accuratezza (predittiva) (AI Act, Articoli 15.1, 15.2);
- Fairness: i dati che alimentano un sistema di AI dovrebbe essere rilevanti, rappresentativi e completi, con particolare riferimento alle persone ed ai gruppi di popolazione i cui dati vengono utilizzati (AI Act, Articolo 10);
- Explainability: un sistema di AI dovrebbe produrre risultati (output e decisioni) comprensibili, interpretabili e gestibili da persone umane. (AI Act, Articolo 14).

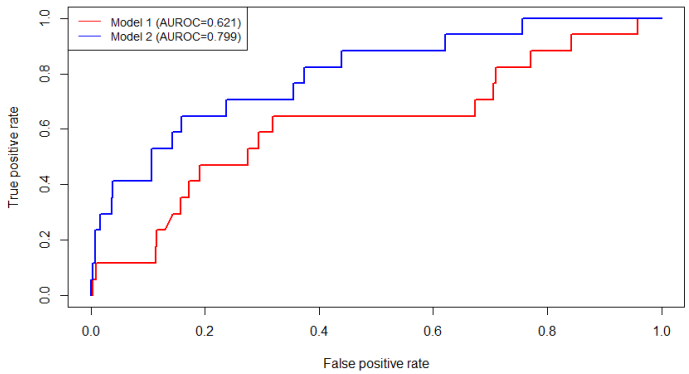
SAFE AI: accuratezza

Esempio 1 - previsione (risposta continua)

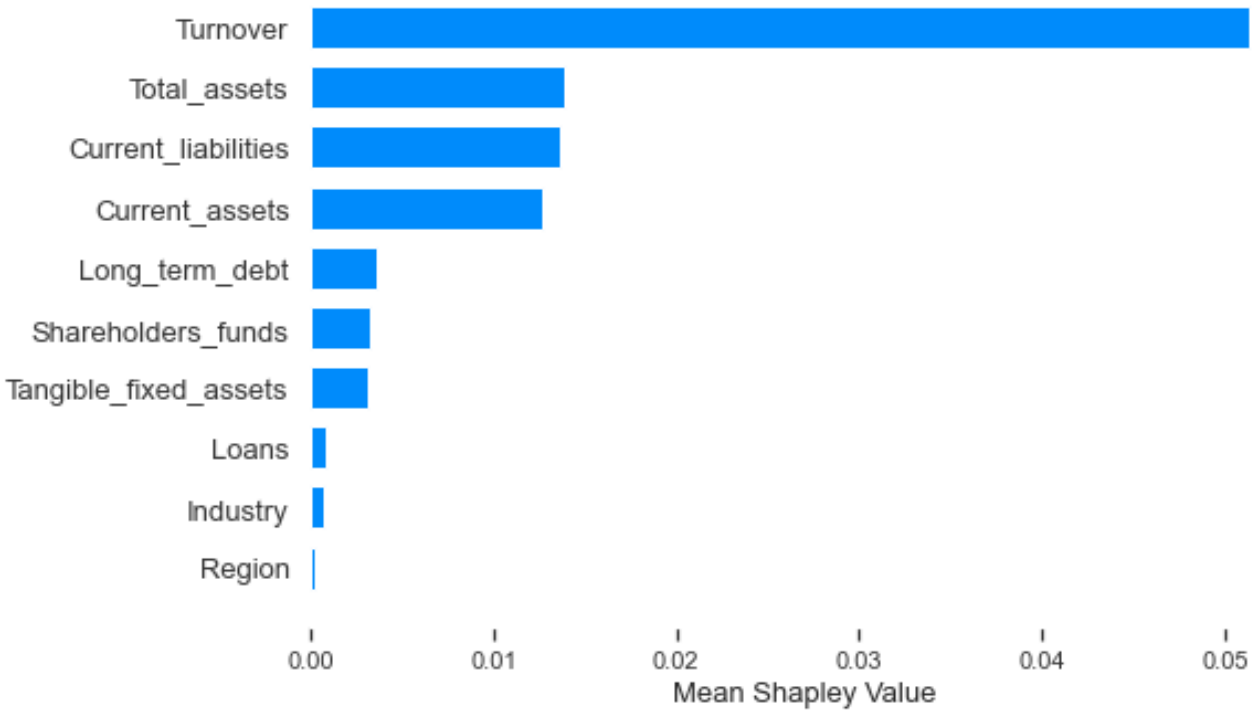
	ROE Effettivo	ROE Previsto
1	3.912	3.679
2	2.171	-0.327
3	-16.649	-0.293
4	10.705	10.325
...		
3453	1.302	2.047
	MSE = 39.14	RMSE = 6.26

Esempio 2: classificazione (risposta binaria)

	Non Default Effettivi	Default Effettivi
Previsti	0	1
0	2949 (TN = 0.868)	9 (FN = 0.153)
1	445 (FP = 0.132)	50 (TP = 0.847)
	3394	59



SAFE AI: spiegabilitá



Valori Shapley delle variabili potenzialmente esplicative del default di 100,000 PMI europee, ottenuti rielaborando l'output di un modello "random forest" (black box ma molto accurato).

SAFE AI: equitá

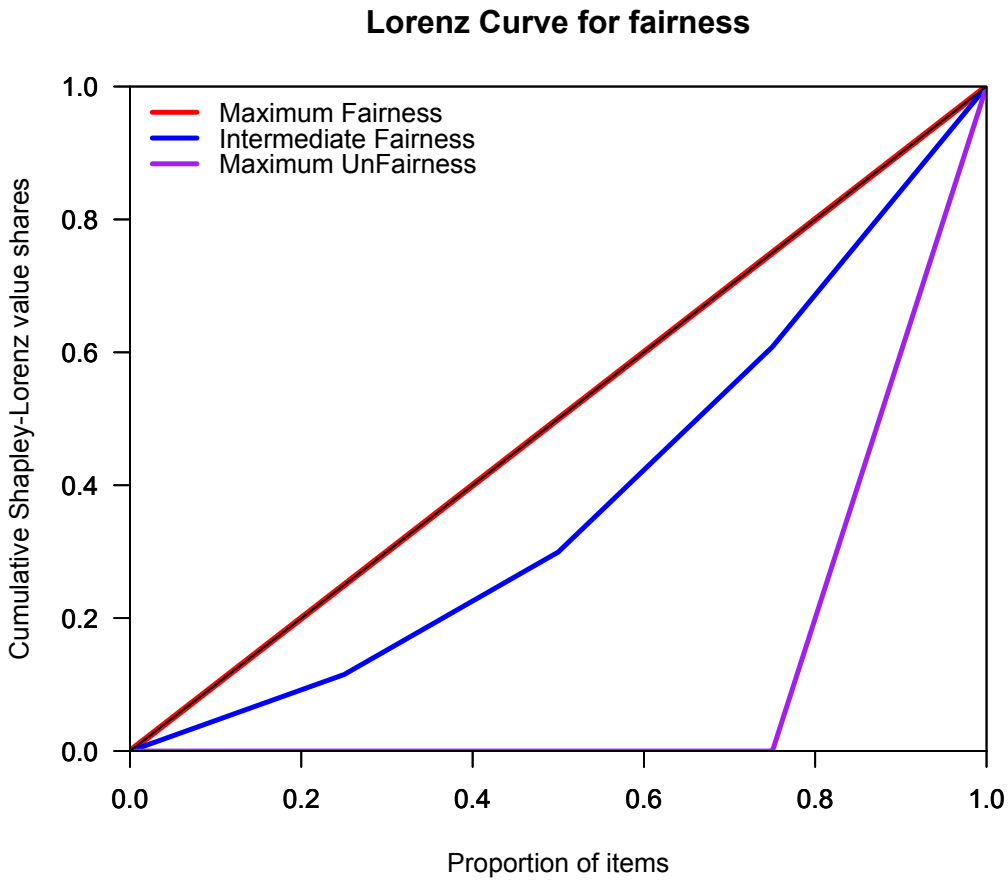


Figure: Variabilitá nella stima del rating creditizio di un insieme di imprese appartenenti a diversi settori. Curva rossa: massima equitá, uguaglianza dei rating fra i settori. Curva viola: minima equitá, discriminazione fra i settori. Curva blu: situazione intermedia.

SAFE AI: sostenibilitá

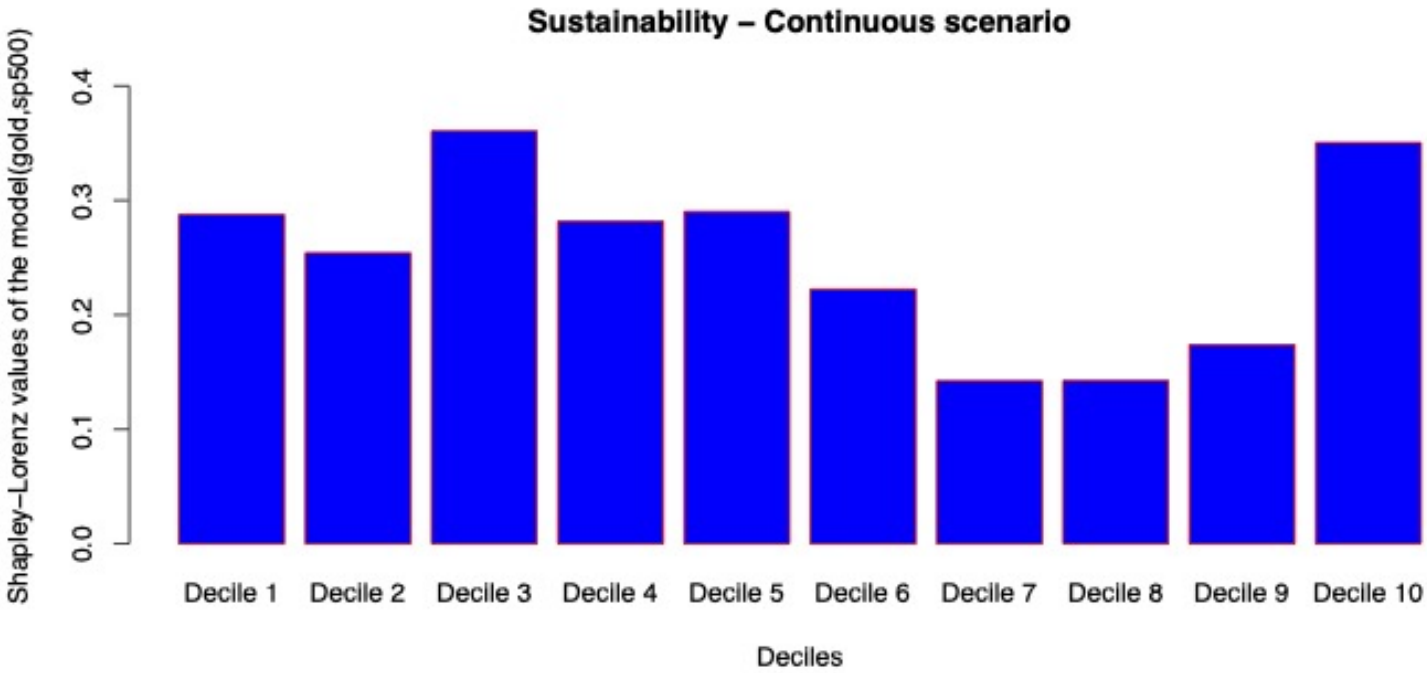


Figure: Variabilitá delle previsioni ottenute in corrispondenza di osservazioni "normali" (verso sinistra) ed "anomale" (verso destra)

Riferimenti bibliografici


Giudici, P., Raffinetti, E. (2023) SAFE Artificial Intelligence in Finance. *Finance Research Letters*, 56, 104088.


Giudici P., Centurelli, M., Turchetta, S. (2023). Artificial Intelligence risk measurement *Expert Systems With Applications*, 235, 121220.