Valutazione delle Performance dei Modelli AI

⚠️ **Disclaimer** ⚠️

Il presente documento ha finalità esclusivamente informative e descrive le metodologie utilizzate per valutare le performance tecniche dei modelli predittivi di intelligenza artificiale sviluppati all'interno della piattaforma Penelope. Le metriche riportate (Accuracy, Precision, Recall, F1-Score, ecc.) rappresentano strumenti di analisi quantitativa e non costituiscono in alcun modo una garanzia di risultato o una promessa di rendimento futuro.

Le prestazioni passate dei modelli, così come i risultati delle simulazioni o delle metriche predittive, ***non sono indicative né garantiscono performance future****.*

I modelli possono comportarsi in modo diverso in base alle condizioni di mercato, alla volatilità o a fattori esterni non prevedibili.

Penelope non fornisce consulenza finanziaria personalizzata, né raccomandazioni all'investimento.

L'utilizzo delle informazioni contenute in questo documento è a discrezione dell'utente, che rimane l'unico responsabile delle proprie decisioni operative o strategiche.

Metriche per la valutazione dei modelli di AI

Obiettivo del Documento

Questo documento illustra come vengono calcolate e interpretate le metriche di performance per i modelli predittivi sviluppati da Penelope, specificando l'approccio utilizzato nella classificazione dei segnali operativi e la metodologia di validazione.

1. Output del Modello e Riclassificazione

Il modello predittivo restituisce, per ciascun asset e per ogni giorno, un valore continuo compreso tra 0 e 1, interpretato come la probabilità di un movimento rialzista (Long) o ribassista (Short).

Questa probabilità viene trasformata in una classificazione binaria secondo la seguente logica:

Probabilità	Interpretazione
> 0,49	Segnale Long (+1)
< 0,49	Segnale Short (-1)

Nota: non è prevista una zona neutra o di incertezza. Il modello opera sempre una scelta Long o Short.

2. Etichettatura c.d. Ground Truth

Per valutare se la previsione era corretta, si considera la variazione del prezzo dell'asset nel periodo di riferimento (es. 1-3-5 giorni di trading successivi al momento della previsione).

Se la variazione percentuale è positiva, si etichetta come Long (vera).

Se è negativa, si etichetta come Short (vera).

3. Confusion Matrix per Classe

Per ogni classe (Long o Short), le previsioni vengono confrontate con la realtà di mercato per determinare:

Termine	Descrizione
True Positive (TP)	Il modello prevede Long, e il movimento reale del mercato è effettivamente rialzista.

False Positive (FP)	Il modello prevede Long, ma il movimento reale è ribassista (sarebbe stato Short).
True Negative (TN)	Il modello prevede Short, e il mercato si muove effettivamente al ribasso.
False Negative (FN)	Il modello prevede Short, ma il mercato sale (sarebbe stato corretto un segnale Long).

4. Metriche Utilizzate

Accuracy

Cos'è: misura la percentuale complessiva di previsioni corrette sul totale dei casi.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

A cosa serve: fornisce una valutazione generale della bontà del modello.

Vantaggi: semplice e globale.

Limite: può risultare fuorviante in presenza di dati sbilanciati (es. pochi segnali Short rispetto ai Long).

Precision

Cos'è: indica quante delle previsioni positive (es. segnali Long o Short) si sono rivelate corrette.

$$Precision = \frac{TP}{TP+FP}$$ |

A cosa serve: misura la qualità dei segnali emessi, minimizzando i falsi allarmi.

Importanza: utile quando è fondamentale evitare operazioni sbagliate.

Recall

Cos'è: indica quanti degli eventi realmente accaduti (es. rialzi o ribassi) sono stati intercettati dal modello.

$$Recall = \frac{TP}{TP+FN}$$

A cosa serve: misura la capacità del modello di non perdere segnali rilevanti.

Importanza: cruciale quando è importante non trascurare opportunità di trading.

F1-Score
Cos'è: media armonica tra Precision e Recall.

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

A cosa serve: fornisce un indicatore bilanciato tra accuratezza dei segnali emessi e capacità di catturare le opportunità reali.

Importanza: ideale in contesti in cui è necessario bilanciare rischio di falsi positivi e falsi negativi.

5. Valutazione Separata per Long e Short

Le metriche sono calcolate indipendentemente per le due direzioni previsionali (Long e Short) per garantire una valutazione più precisa e granulare del comportamento del modello. Questa distinzione èfondamentale per diverse ragioni operative e analitiche:

Fasi di mercato differenti: i mercati non si comportano simmetricamente nelle fasi rialziste e ribassiste. Analizzare separatamente le due direzioni consente di verificare l'efficacia del modello in ciascuna fase e adattare di conseguenza l'esposizione.

Bias strutturale: in molti mercati (come l'azionario), esiste una tendenza storica al rialzo. Calcolare metriche aggregate senza distinzione potrebbe mascherare scarse performance nella gestione dei ribassi.

Ottimizzazione selettiva: permette di utilizzare modelli diversi o strategie differenti per Long e Short, scegliendo il migliore per ciascuna direzione.

Controllo del rischio: individuare un modello che funziona bene solo in una direzione pu√≤ essere utile per strutturare strategie di hedging più efficaci o per limitarne l'uso in contesti sfavorevoli.

6. Perchè una soglia a 0,49?

L'utilizzo della soglia 0,49 al posto di 0,50 ha una motivazione operativa:

Leggero bias verso Long: in alcuni mercati (es. azionario), la probabilità di trend rialzisti pu√≤ essere leggermente superiore nel medio periodo.

Stabilità: evita oscillazioni eccessive tra neutralità e segnali inversi.

7. Periodicità e Intervalli Temporali

Le metriche di performance vengono calcolate settimanalmente, su base rolling, e riferite a diversi intervalli temporali del passato:

- 20 giorni di trading (circa 1 mese)
- 60 giorni di trading (circa 3 mesi)
- 120 giorni di trading (circa 6 mesi)
- 320 giorni di trading (circa 1 anno e 1/2)

Questi intervalli sono scelti per garantire una visione dinamica e multilivello delle prestazioni, considerando sia la robustezza storica del modello sia la sua reattività recente ai cambiamenti di mercato.

8. Selezione dei Modelli

Le metriche vengono calcolate per tutti i modelli AI attivi in Penelope, su ciascuna asset class e timeframe.

Insieme agli indicatori di bontà finanziaria (es. rendimento netto, drawdown, Sharpe Ratio, max consecutive loss, ecc.), queste metriche costituiscono il sistema di selezione dei modelli utilizzato da Penelope.

L'obiettivo finale èscegliere i modelli più performanti e stabili nel tempo, ottimizzando il rapporto tra accuratezza predittiva e redditività reale.

Appendice A
esempio di indicatori misurati alla data 19.07.2025

Di seguito si riportano, a titolo esemplificativo, i valori medi e massimi di precisione relativi alla classe Long, rilevati nel periodo compreso tra il 17 maggio 2025 e il 19 luglio 2025.

I dati sono suddivisi per mese e calcolati per i modelli AI attivi su Nasdaq 100 e S&P 500. Le misurazioni sono effettuate in ambiente di produzione.

Nasdaq (NDQ)		1 day		3 days		5 days	
		Average of precision	Max. of precision	Average of precision	Max. of precision	Average of precision	Max. of precision
Ai v1.0	May	63%	77%	56%	60%	55%	59%
	Jun	61%	73%	61%	77%	60%	72%
	Jul	64%	74%	60%	67%	58%	63%
Ai v 1.5o	May	57%	77%	52%	59%	53%	56%
	Jun	58%	73%	58%	77%	58%	67%
	Jul	60%	74%	58%	67%	57%	63%

S&P500 (SPX)		1 day		3 days		5 days	
		Average of precision	Max. of precision	Average of precision	Max. of precision	Average of precision	Max. of precision
Ai v1.0	May	55%	64%	52%	58%	53%	58%
	Jun	57%	70%	56%	61%	59%	66%
	Jul	62%	70%	62%	70%	58%	64%
Ai v 1.5o	May	61%	78%	44%	54%	54%	60%
	Jun	57%	64%	53%	63%	58%	66%
	Jul	64%	74%	64%	76%	58%	63%

Data measured on July 19th 2025 on all timeframes

Visual Classification of data	<40%	>40%, <50%	>50%, <70%	>70%

⚠️ Disclaimer ⚠️

Le statistiche riportate nel presente documento si riferiscono al periodo 17 maggio 2025 – 19 luglio 2025 e sono suddivise per mese. Esse rappresentano i valori medi e massimi di precisione relativi alla classe di previsione Long, calcolati su due strumenti finanziari e due modelli AI attivi durante il periodo considerato.

Questi dati sono forniti a scopo puramente informativo e illustrativo e non devono in alcun modo essere interpretati come una promessa di rendimento, una previsione finanziaria o una garanzia di performance futura.

Tutte le metriche presentate sono generate in modo automatizzato, senza intervento umano, e si basano su dati effettivamente osservati in ambiente di produzione, raccolti nei database interni di BXT e Penelope.

Non si tratta di backtest, simulazioni o analisi ipotetiche, ma di risultati calcolati su dati reali disponibili al momento della misurazione.

Le analisi vengono effettuate su diversi orizzonti temporali storici (20, 60, 120 e 320 giorni di trading) per garantire valutazioni robuste e multilivello, utili a comprendere la stabilità e l'affidabilità dei modelli in differenti contesti di mercato.

I valori riportati non includono commissioni, slippage o altri costi operativi e non riflettono il rendimento netto di una strategia di investimento reale.

Il presente documento è destinato esclusivamente a finalità esplicative, anche in relazione alle metriche comunicate sul sito pubblico e su eventuali materiali informativi o divulgativi diffusi tramite canali digitali o social.