Data Science in PwC:

Casi studio di Causal Inference e Anomaly Detection

Alessandro De Bettin - Manager D&A
03/05/2024 - Padova

ALESSANDRO DE BETTIN

Manager

Alessandro è un manager del team di **Data Science & AI** in PwC e si occupa della gestione di progetti di **Machine learning** ed **Intelligenza artificiale** in diversi settori.

Laureato in Statistica, nel corso degli anni ha avuto la possibilità di partecipare a progetti di analisi statistica, Machine Learning e Natural Language Processing.

pwc ITALIA

888 mln €
Ricavi FY 2023

+8.000
Dipendenti

Advisory

3.000+
Dipendenti

27
Uffici operativi e sedi in Italia

Assurance

1.400+ dipendenti

Tax

900+ dipendenti

PwC Data Science & AI

Il team **PwC Data Science & AI** si occupa di sviluppare soluzioni analitiche avanzate per l'estrazione di insights dai dati, e di supportare le aziende nella trasformazione di processi attraverso l'adozione di sistemi di AI, fornendo supporto strategico e consulenza personalizzata ai clienti.

Statistica e Data Science per il business

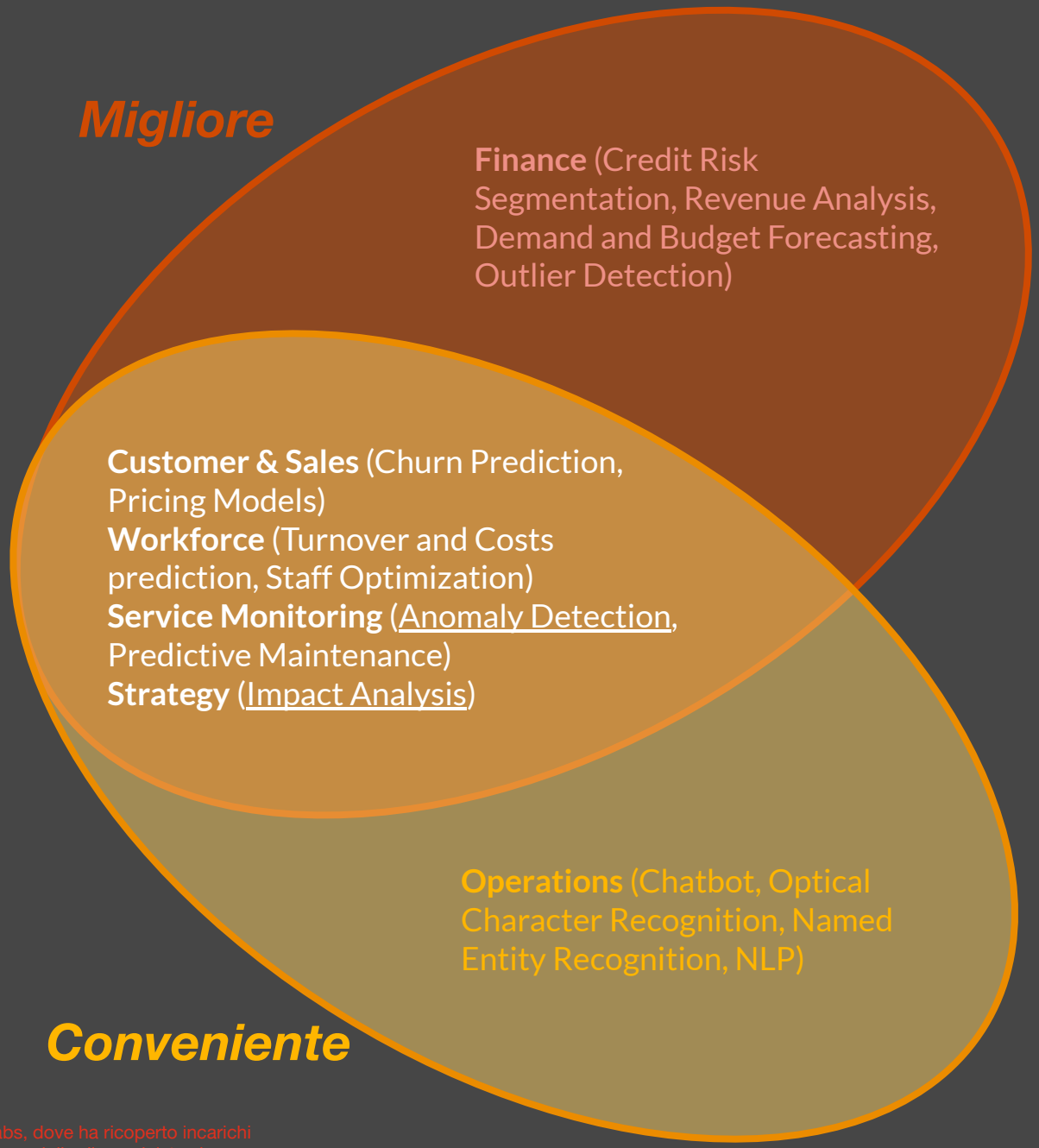
"Better, or cheaper, or both."

Mervin Kelly*

Il ruolo del team Data Science & AI di PwC è quello di portare **innovazione** nelle aziende utilizzando i dati. L'innovazione consiste o nel **migliorare** un determinato **processo**, o nel renderlo **meno costoso**, o in **entrambe** le cose.

Le soluzioni di AI sono spesso capaci di coniugare entrambi questi aspetti, grazie alla scalabilità e alla versatilità dei modelli e alle notevoli potenzialità di sviluppo e avanzamento degli attuali sistemi.

L'intelligenza Artificiale permette sia di supportare il business nei **processi decisionali**, fornendo evidenze ed insights a partire dai dati, sia di rivoluzionare **processi operativi**, consentendo di raggiungere livelli di efficienza maggiori e di facilitare il lavoro umano attraverso l'automatizzazione di operazioni ripetitive.



*Mervin Kelly è stato una figura fondamentale dei Bell Labs, dove ha ricoperto incarichi come direttore della ricerca, presidente e presidente del consiglio di amministrazione

Use-case #1

Anomaly Detection nei Servizi Digitali

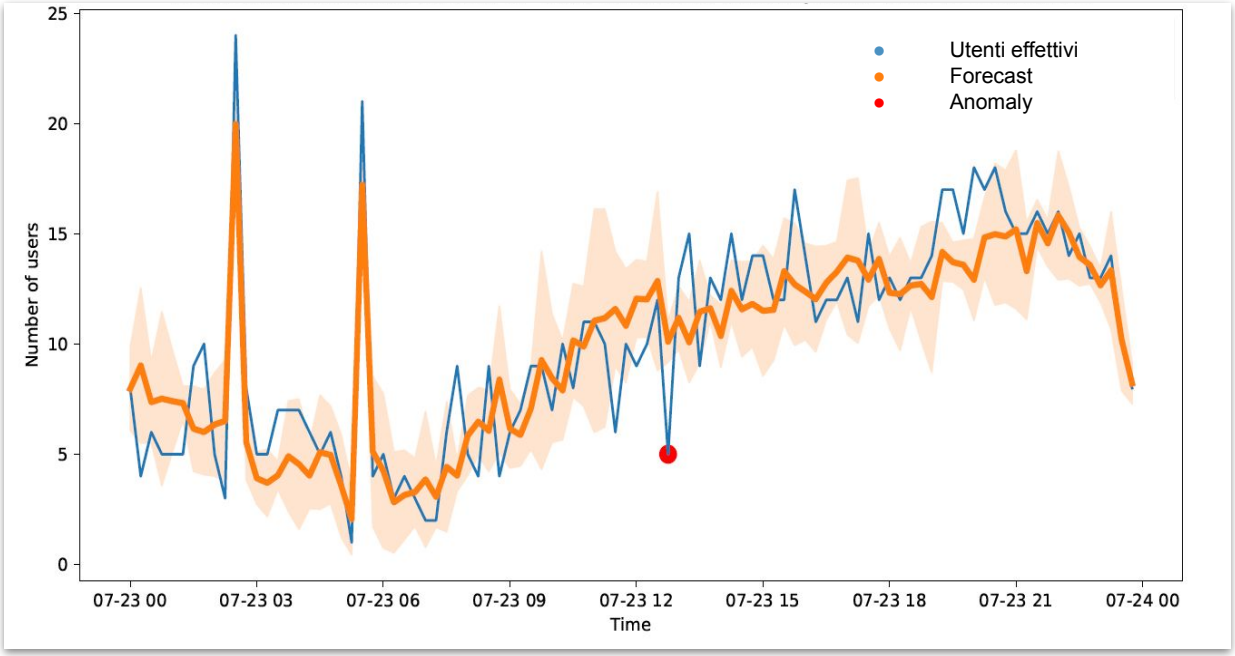
Rilevamento di anomalie nel numero di utenti attivi per un prodotto digitale, al fine di individuare eventuali disservizi di Internet Service Provider.

Attività

- Analisi dei **dati storici** degli utenti connessi e delle **anomalie passate.**
- Sviluppo di uno strumento in grado di **prevedere** il numero di utenti connessi per area geografica in **assenza di anomalie.**
- Confronto **in tempo reale** tra previsioni e dati effettivi.

Risultati

- Uno strumento in grado di **prevedere il numero futuro di utenti** connessi e di **migliorare continuamente** incorporando nuovi dati.
- Identificazione continua di **anomalie a livello di ISP**, indice di possibili disservizi a livello di rete internet.



Use-case #1

Approccio all'Analisi

Le **città** sono state suddivise in **cluster** in base al numero medio di utenti connessi

Abbiamo sviluppato **modelli di forecast** per ogni **città e provider**, utilizzando i dati storici di connessione degli utenti.

Per garantire un riconoscimento accurato delle anomalie, abbiamo definito **soglie specifiche per ciascun cluster**, considerando le variazioni nel comportamento degli internet service provider.

Per valutare l'efficacia dei modelli, abbiamo confrontato le anomalie rilevate con le segnalazioni di effettivi **disservizi di internet service provider**.

Use-case #1

Modellazione Statistica

Modello di Forecast

- Sono stati addestrati modelli ai seguenti livelli di granularità:
  - Città e Internet Service Provider
  - Intervalli temporali di 15 minuti
- Dopo svariati test, è stato identificato **XGBoost** come modello target per:
  - Robustezza agli outlier (è un model ensemble basato su alberi)
  - Possibilità di effettuare training in maniera incrementale
- Applicazione incrementale
  - Per poter confrontare le previsioni con i dati effettivi, il modello viene applicato iterativamente

Anomaly Detection

- Soglie basate sullo z-score modificato dei residui.
- Lo **Z-score modificato** è una misura statistica che quantifica il numero di deviazioni standard di un punto rispetto alla mediana.
- Le soglie sono state ottimizzate per ogni cluster separatamente sulla base delle anomalie effettive registrate.
- Questo metodo identifica la maggior parte delle anomalie registrate.

$$M_i = \frac{0.6745(x_i - \tilde{x}_i)}{MAD}$$

Use-case #1

Metriche di Valutazione dei Risultati

Recall Modificata

La Recall modificata valuta la capacità del modello di catturare tutte le istanze positive, rappresentando la percentuale di istanze positive identificate correttamente su tutte le istanze positive effettive.

$$MR = \frac{\sum_{i=1}^{n} \left(TP_i \cdot w_i\right)}{\sum_{i=1}^{n} \left(TP_i \cdot w_i\right) + \sum_{j=1}^{n} \left(FN_j \cdot w_j\right)}$$

w = Numero di utenti interessati
n = Numero di anomalie reali

Precision Modificata

La Precision modificata misura l'accuratezza delle previsioni positive, indicando la percentuale di istanze positive correttamente identificate tra tutte le istanze previste come positive.

$$MP = \frac{\sum_{i=1}^{n} \left(TP_i \cdot w_i\right)}{\sum_{i=1}^{n} \left(TP_i \cdot w_i\right) + \sum_{k=1}^{m} FP_k}$$

w = Numero di utenti interessati
n = Numero di anomalie reali
m = Numero di anomalie non reali

Use-case #2

Analisi dell'Impatto di Eventi Significativi

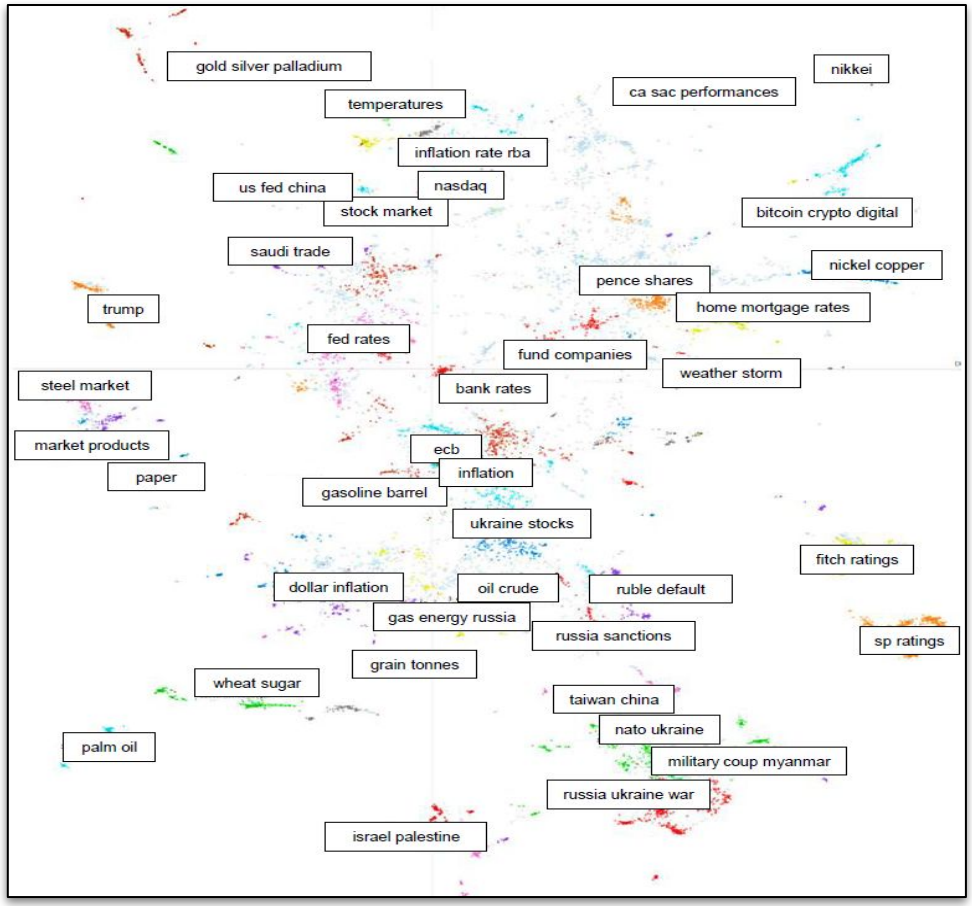
Identificazione di eventi significativi all'interno di dati non strutturati e calcolo del loro impatto sui processi di business usando tecniche di Causal Inference

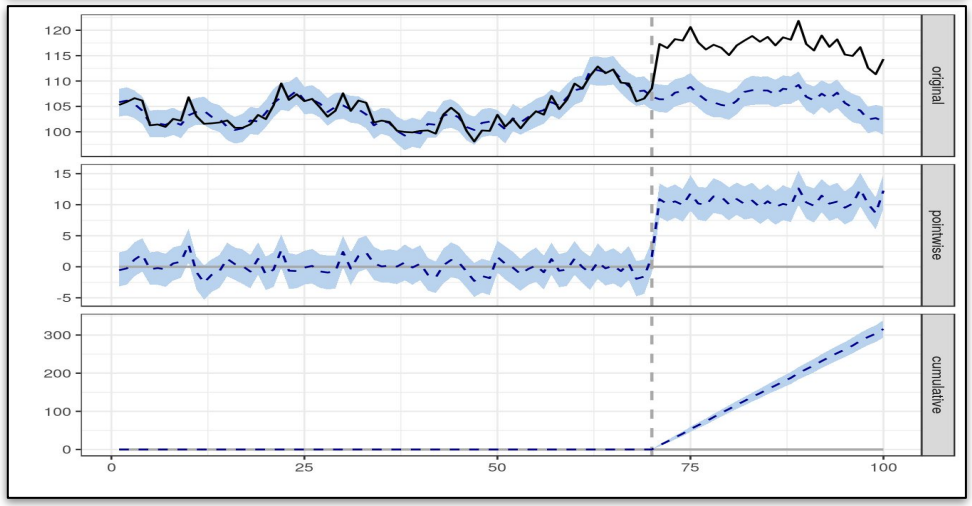
Attività

- **Topic Modelling** - identificazione di topic ricorrenti nei testi non strutturati
- **Sentiment Analysis e Summarization** - a partire dai testi per facilitare l'estrazione di informazioni di interesse
- **Causal Inference** - utilizzo di un modello di Deep Learning per time series forecasting e tecniche specifiche per quantificare l'impatto di ciascun topic

Risultati

- Accesso alle informazioni attraverso una vista unificata per ottenere **insights** significativi
- Estrazione di **entità rilevanti** dal testo, come luoghi, organizzazioni, menzioni di indici di rischio
- Comprensione approfondita dell'**impatto sul processo di interesse** degli eventi identificati





Use-case #2

Approccio all'Analisi

Overview

0 Raccolta dei dati

- Serie storiche
- Dati non strutturati

1 Topic Modeling

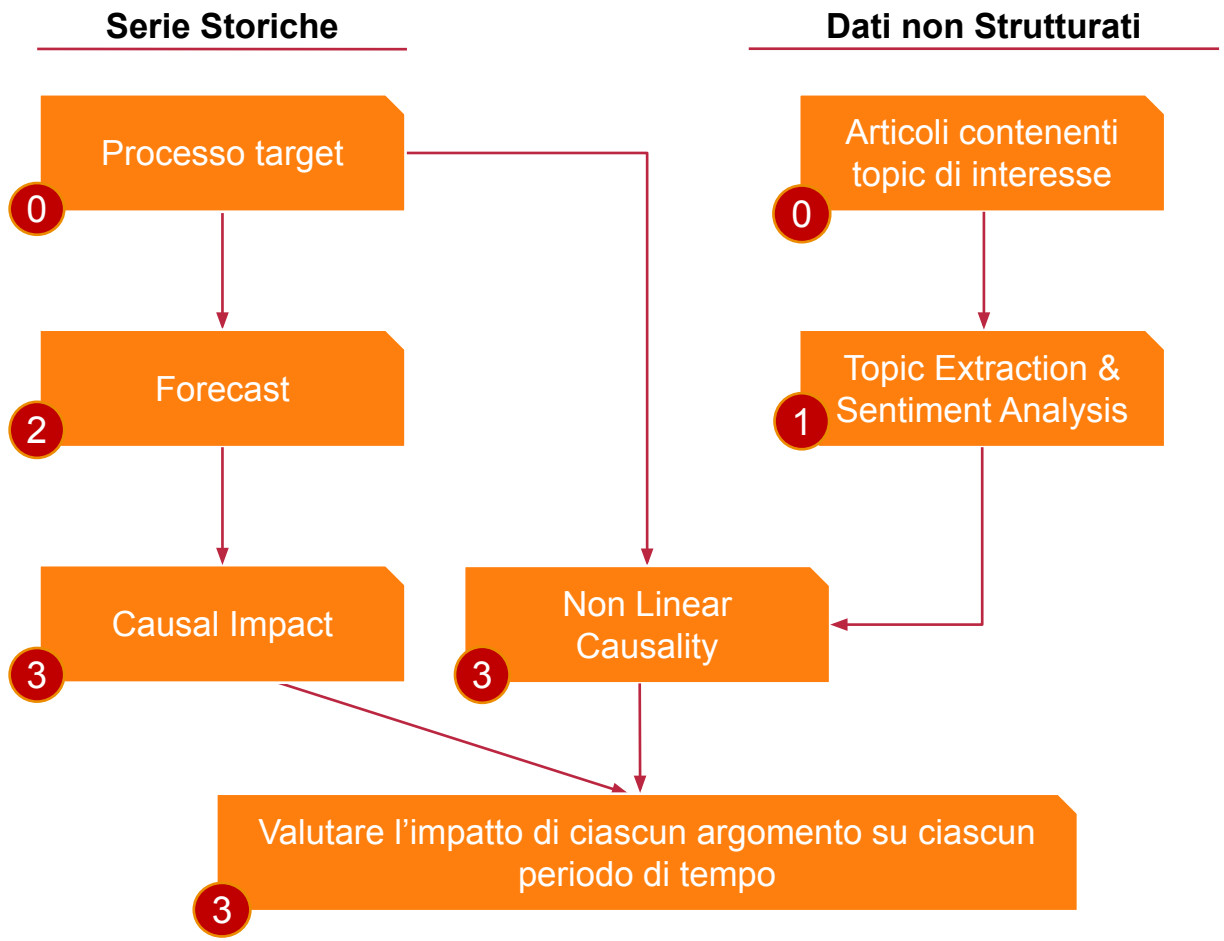
- Analisi dei testi per estrarre i principali argomenti.
- Topic modelling per ottenere una comprensione organizzata dei temi discussi.
- Estrazione del sentiment e delle parole chiave più rappresentative dai testi utilizzati.

2 Forecast della serie target

- Modellazione statistica del processo target, attraverso modelli di forecast multivariati.

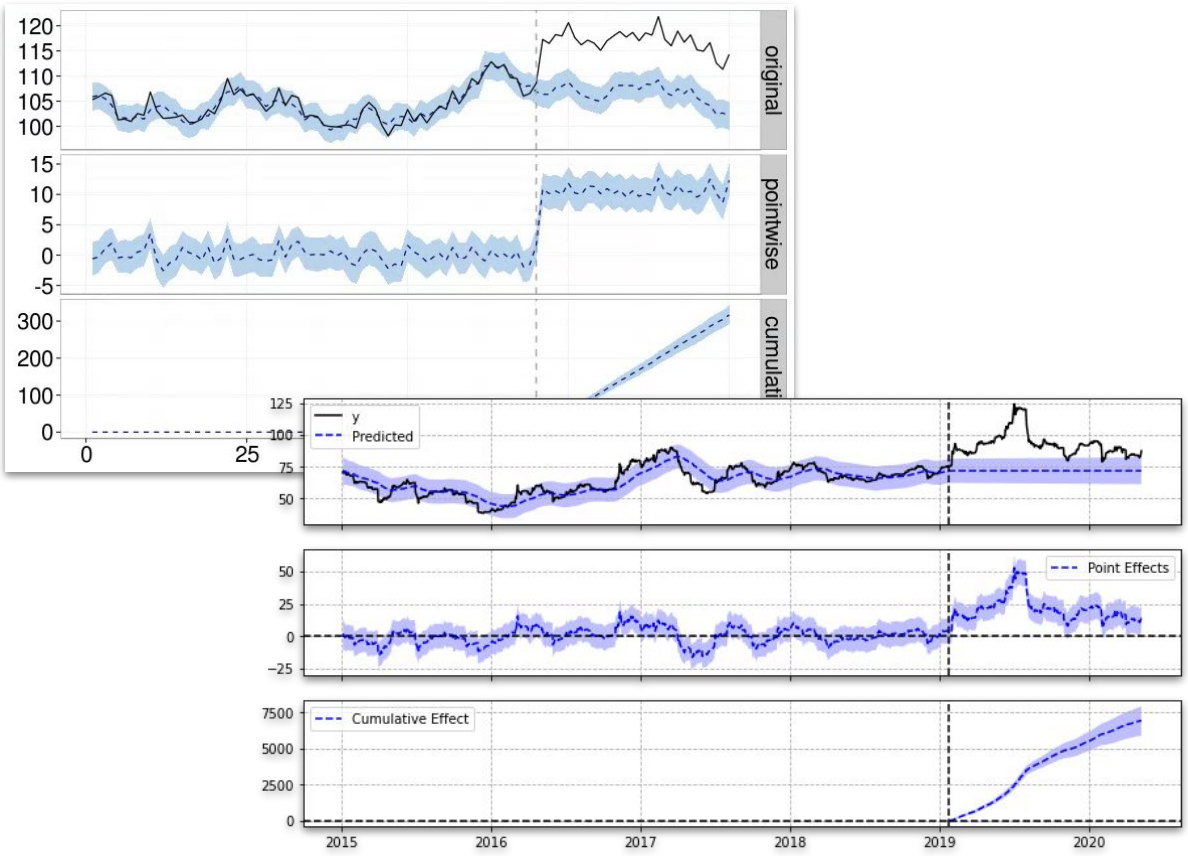
3 Analisi dell'Impatto dei Topic

- Creazione di serie storiche per ogni topic e valutazione del loro impatto temporale sul processo di interesse.
- Confronto tra l'errore dei modelli rispetto alla serie target e il calcolo della causalità per valutare l'efficacia della previsione basata sui topic.
- Calcolo dell'indice di causalità per determinare l'influenza dei topic sull'andamento della serie storica target.



Use-case #2

Causal Impact



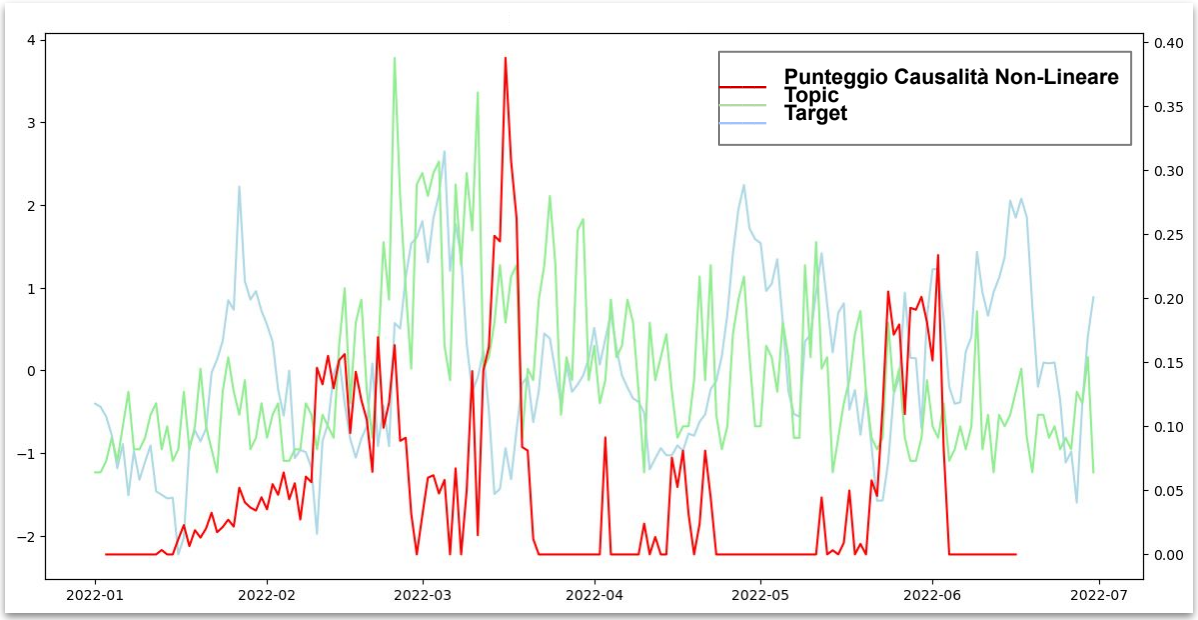
Funzionamento dell'algoritmo

- Addestramento di un **modello di forecast** sullo storico dati, usando i dati fino all'istante dell'evento di interesse
- Utilizzo il modello per **prevedere il comportamento atteso** della serie target (un "controfattuale")
- Confronto fra i dati effettivi e la previsione effettuate
- **Calcolo dell'impatto** dell'evento, puntuale e cumulato

Il paper di riferimento (Brodersen, K.H., et al. (2015). Inferring Causal Impact Using Bayesian Structural Time-Series Models) utilizza modelli bayesiani, ma sono utilizzabili anche altri modelli.

Use-case #2

Non Linear Causality



Funzionamento dell'algoritmo

La causalità non lineare (Rosol, M., et al. (2022). Granger Causality Test with Nonlinear Neural-Network-Based Methods: Python Package and Simulation Study) si basa sul concetto di **causalità di Granger**, superando l'utilizzo dei modelli lineari e l'assunzione di **stazionarietà delle variabili.**

Il punteggio di causalità **non lineare** in una finestra di osservazione viene calcolato come segue:

$$Causality(topic, time) = \frac{2}{1 + e^{1 - \frac{RMSE\ standard\ features}{RMSE\ standard\ features + Topic\ series}}} - 1$$

In questo modo siamo in grado di quantificare la causalità attraverso il **miglioramento che ogni topic apporta** al modello di forecast del processo target.

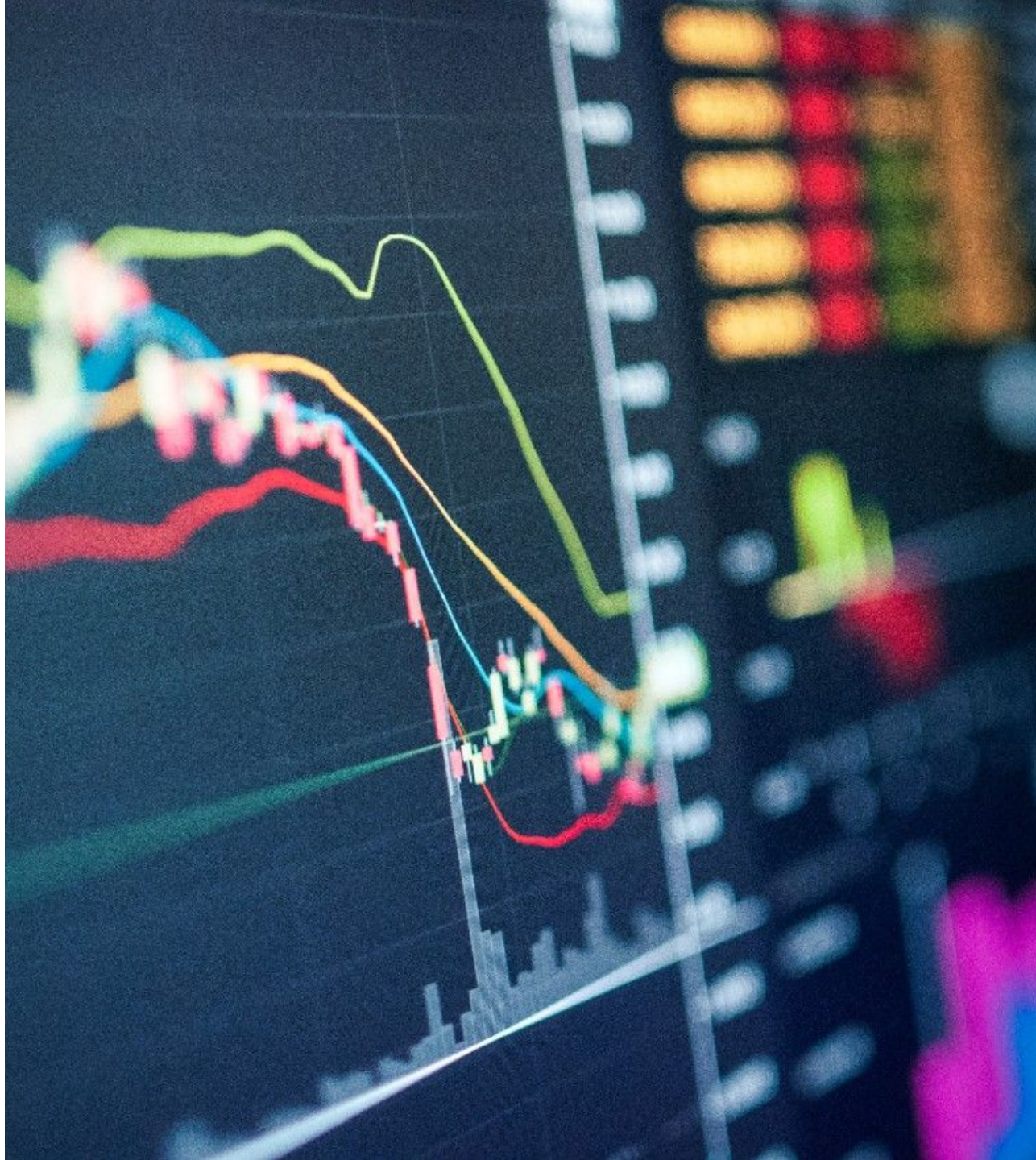
Conclusioni

Analisi evoluta delle serie storiche

- La combinazione di modelli di forecasting e altri modelli statistici abilita analisi evolute di serie storiche
- Combinando forecasting e test statistici è possibile individuare eventi anomali
- Combinando forecasting e Causal Inference è possibile stabilire l'impatto degli eventi sul processo di business target

Alcuni ambiti di applicazione

- Marketing
- Application monitoring
- Finanza
- ...



Grazie a tutti

Alessandro De Bettin - Contatti