linea orizzontale

Modello Predittivo Prophet

# Introduzione

In questo documento c’è la spiegazione delle complete funzionalità del Modello Gideon, creato appositamente per l’analisi dei movimenti finanziari di un’azienda. In allegato c’è il Jupyter Notebook che contiene tutto il codice nel quale troverete lo svolgimento pratico di ciò che segue.

Il notebook è diviso in due parti: la prima contiene un’analisi descrittiva delle entrate e uscite dell’azienda con misurazione del cash flow giornaliero, la seconda contiene il modello predittivo con misurazione delle performance e proiezione ai 6 mesi successivi all'ultima data osservata. Per comodità e maggiore chiarezza nella sua descrizione, il codice è stato suddiviso in blocchi e la spiegazione seguirà la precedente divisione.

(Come campione è stata presa l’azienda con codice identificativo 57 dalle 4000 e passa disponibili, in quanto essa è la prima in ordine numerico a far parte del cluster preso in analisi.)

# 

# Prima Parte: Analisi Entrate/Uscite e Cash Flow

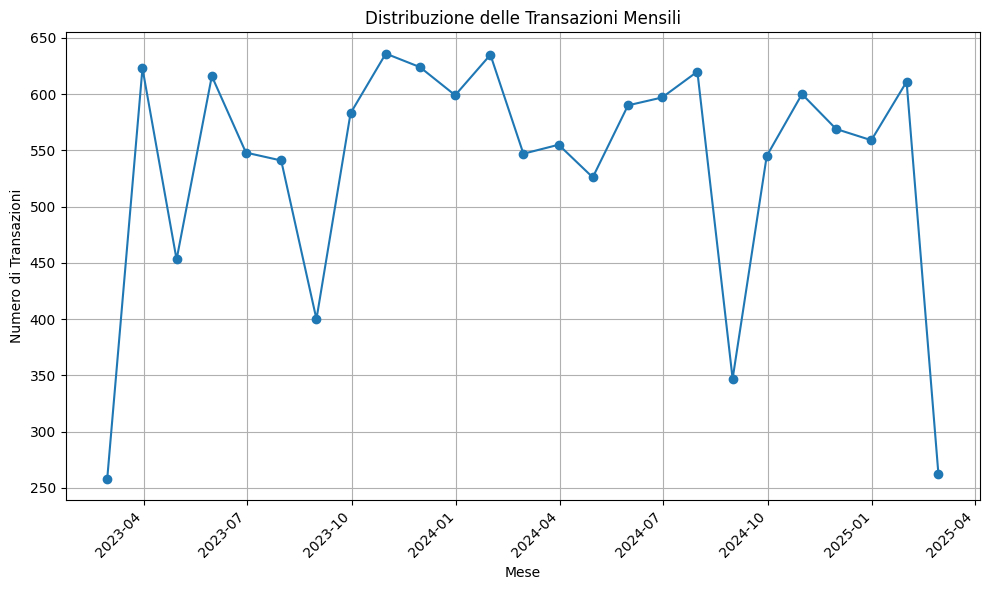
## Primo Blocco

Nella prima parte di codice, dopo l’installazione delle librerie e il caricamento dei vari dataset, l’algoritmo dà la possibilità all'operatore di scegliere una azienda tra le 4000 disponibili, in modo tale da poter visualizzare i comportamenti caratteristici di entrare ed uscite della singola azienda lungo i due anni. Le prime informazioni che vengono restituite sono:

- il numero di spostamenti registrati e il numero di giorni in cui sono stati effettuati;

- il numero di conti dell’azienda e il saldo iniziale presente sui conti;

- il primo e ultimo giorno registrato degli spostamenti ed un grafico che mostra la frequenza degli spostamenti sui due anni osservati.



## 

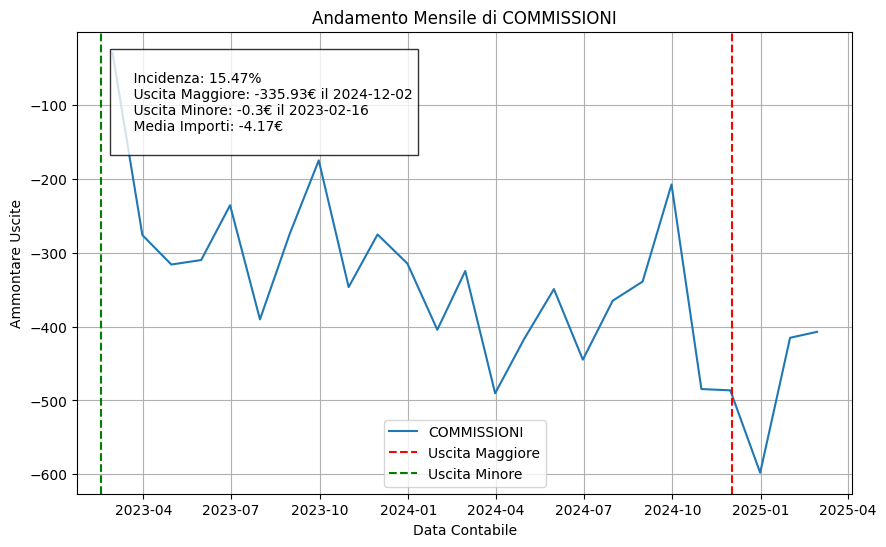
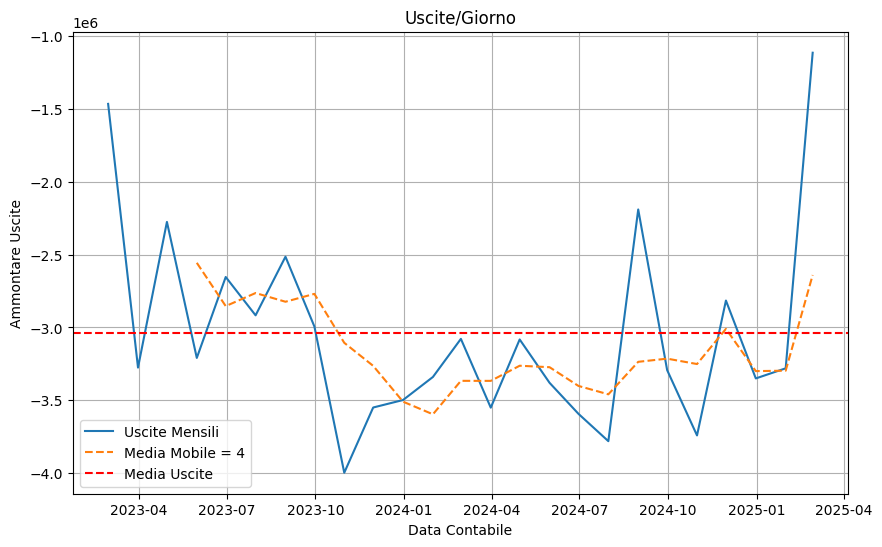
## Secondo e Terzo Blocco

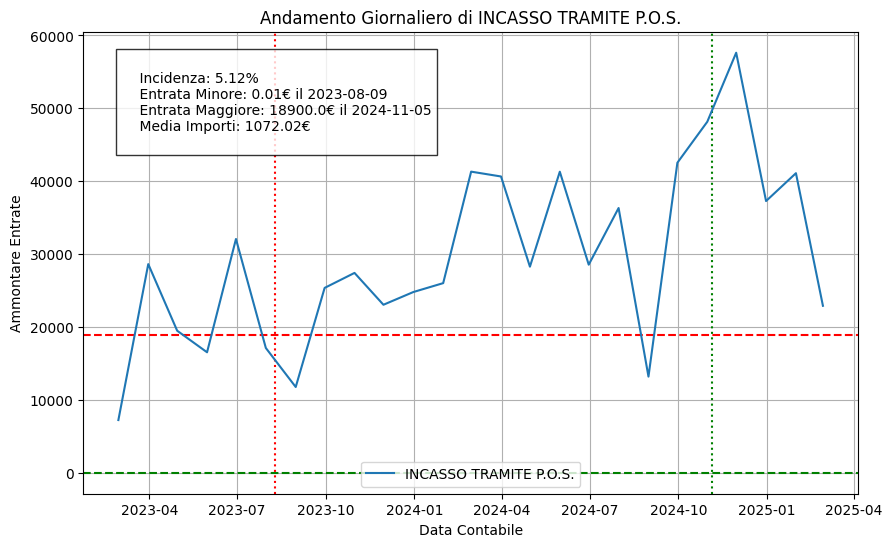
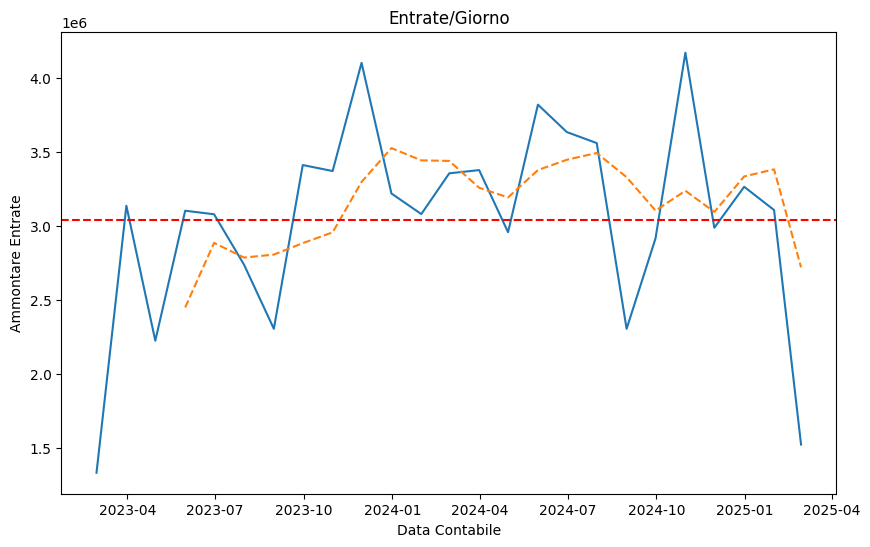
I movimenti vengono categorizzati in uscite ed entrate e analizzati separatamente in blocchi diversi. Le caratteristiche dei movimenti vengono catturate da grafici, indicando anche:

- il numero di spostamenti in entrata e in uscita con i rispettivi giorni interessati, in quanto diverse operazioni potrebbero essere effettuate nello stesso giorno;

- un grafico con aggregazione mensile delle totali uscite ed entrate;

- grafici per i cinque movimenti più frequenti, cinque uscite e cinque entrate, con aggregazione mensile con focus sul valore massimo e minimo specificando la data, valore medio e incidenza dello specifico movimento.



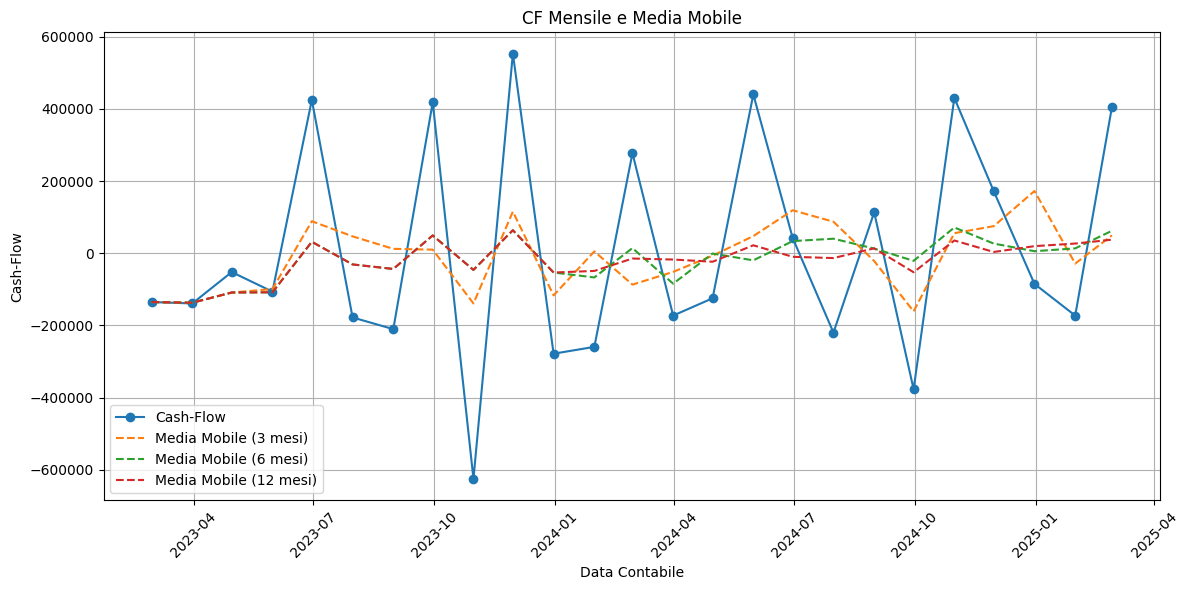


## Quarto Blocco

Il Cash Flow è calcolato come la differenza tra le entrate e le uscite in un giorno. Esso viene calcolato sui due anni osservati, aggregato mensilmente per una migliore rappresentazione grafica e accompagnato con medie mobili su vari periodi (3 mesi, 6 mesi e 12 mesi). Infine viene generato un commento d’analisi che indica:

- il valore massimo e minimo con data;

- il numero di giorni consecutivi in cui il Cash Flow ha differenza positiva e negativa, cioè i periodi in cui le uscite sono maggiori delle entrate e viceversa (di seguito al grafico c’è il commento generato per l’azienda 57).



Il periodo più lungo con uscite maggiori delle entrate è stato dal 2023-02-28 al 2023-06-30 (122 giorni).

Il valore di uscita giornaliera più basso è stato di -625454.36€ il giorno 2023-10-31.

Il periodo più lungo con entrate maggiori delle uscite è stato dal 2024-05-31 al 2024-07-31 (61 giorni).

Il valore di entrata giornaliera più alto è stato di 553450.78€ il giorno 2023-11-30.

## Quinto Blocco

Questo è l’ultimo blocco della prima parte, rappresenta infatti il punto di partenza per l’analisi predittiva. Prima però di continuare con la spiegazione, trovo doveroso spiegare il perché ho preferito un analisi predittiva diversa, infatti i modelli non prevedono direttamente il Cash Flow ma la **Liquidità**. Perché?

La serie temporale del Cash Flow presenta una **non stazionarietà**. Questo significa che le sue proprietà statistiche come media e varianza **non sono costanti nel tempo** rendendo difficile estrarre informazioni utili per la previsione, come trend e stagionalità.

Al contrario, sommando il saldo iniziale al cash flow, si ottiene una serie temporale che rappresenta la **Liquidità**, ovvero il saldo del conto corrente nel tempo. Questa serie risulta essere **stazionaria** (e con una auto-correlazione molto più alta della serie storica precedente), il che implica che le sue proprietà statistiche **rimangono costanti nel tempo**, o meglio, visibili anche nella brevità del periodo osservato.

Inoltre, una serie storica che misura la Liquidità ci dà più informazioni. Per comprendere meglio la differenza, si consideri il seguente esempio: se il cash flow in un giorno è negativo, questo potrebbe significare che l'azienda sta affrontando difficoltà finanziarie. Tuttavia, se il saldo del conto corrente è ancora positivo, l'azienda potrebbe avere ancora una buona situazione finanziaria.

**Ma così facendo non perdiamo il focus sugli obiettivi posti dal Datathon?**

Assolutamente no. La previsione della Liquidità è una “trasformazione naturale” con la quale rendiamo la serie storica del Cash Flow stazionaria e di conseguenza più prevedibile, ma una volta prevista la Liquidità il passaggio al Cash Flow è matematico:

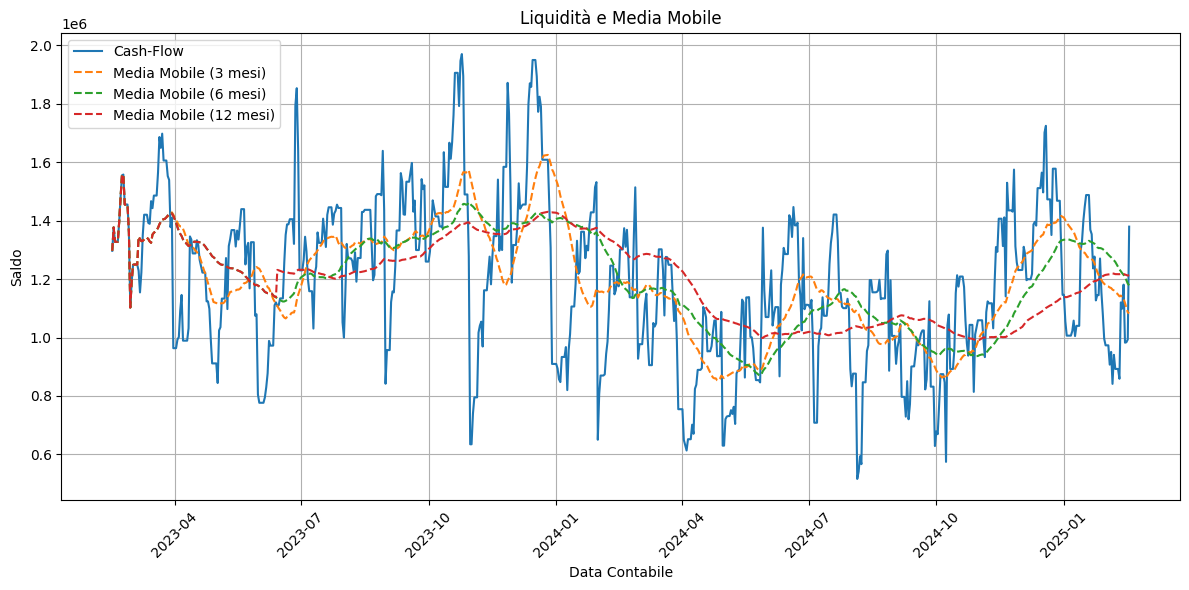
Questa è la formula della Liquidità nel giorno “t”. Essa è ottenuta dalla Liquidità del giorno precedente più la differenza tra le entrate e le uscite del giorno stesso. Rendiamo così la Liquidità una somma cumulata del Cash Flow.

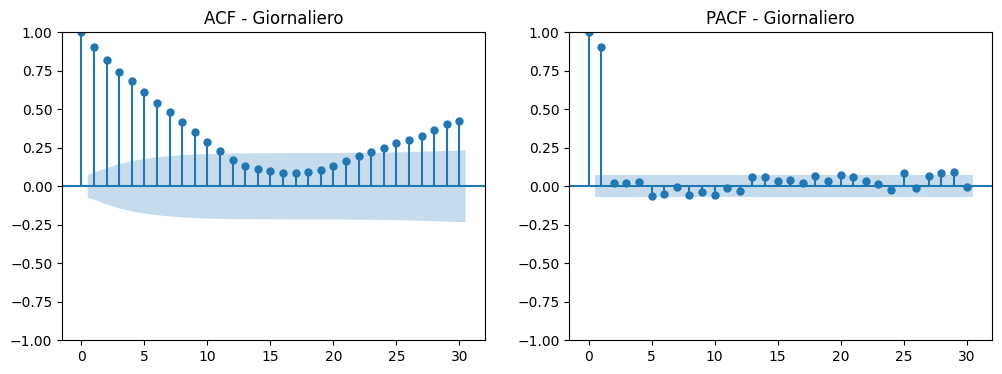
Una volta che siamo nella zona dei valori della Liquidità predetta basta la formula inversa per calcolare il Cash Flow predetto:

Posizionandosi nel primo giorno delle previsioni (“tp” sta per “time predicted”), il Cash Flow verrà calcolato come la differenza tra la Liquidità predetta e la Liquidità del giorno precedente, cioè dell’ultimo giorno reale. Così otteniamo una predizione per costruzione del CF nei 6 mesi successivi.

L’aggiunta di questo dato esterno ci permette di misurare il saldo per ogni giorno della serie osservata, andando anche ad aumentare il numero di dati disponibili (che non guasta mai).

Di seguito c’è il grafico della Liquidità con medie mobili per vari periodi temporali, e la misurazione dell’autocorrelazione con il test di Dickkey-Fuller per diverse aggregazioni temporali.





# Seconda Parte: Analisi Predittiva

Il modello utilizzato è Prophet, un modello statistico sviluppato da Meta per l’analisi delle serie storiche, scelto appositamente per la sua semplicità e per l’attenzione alle varie stagionalità settimanali, mensili ed annuali. Esso utilizza delle funzioni lineari con cambi di pendenza per modellare il **trend** della serie temporale, combinando **funzioni di Fourier** per catturare la **stagionalità** e **funzioni indicatrici** per rappresentare **eventi speciali**.

## Sesto e Settimo Blocco

Per avere un modello che non si adatti troppo alla funzione di train bisogna dividere i due anni in:

- train set, la porzione di dati che il modello utilizza per studiare e imparare i pattern e le correlazioni tra i dati;

- validation set, la porzione di dati mai vista del modello, che viene utilizzata per valutare le performance durante l'addestramento e ottimizzare gli iperparametri, evitando overfitting;

- test set, costituito da dati completamente separati, su cui il modello viene testato alla fine per verificare la sua capacità di previsione su nuovi dati reali.

Il modello viene addestrato su un cluster di 8 aziende, tutte appartenenti allo stesso settore produttivo, così facendo il modello è capace di cogliere stagionalità comuni aumentando la sua precisione.

Le dimensioni dei vari set (contando le 8 aziende combinate) è:

- Train set: 2989 giorni (il primo anno);

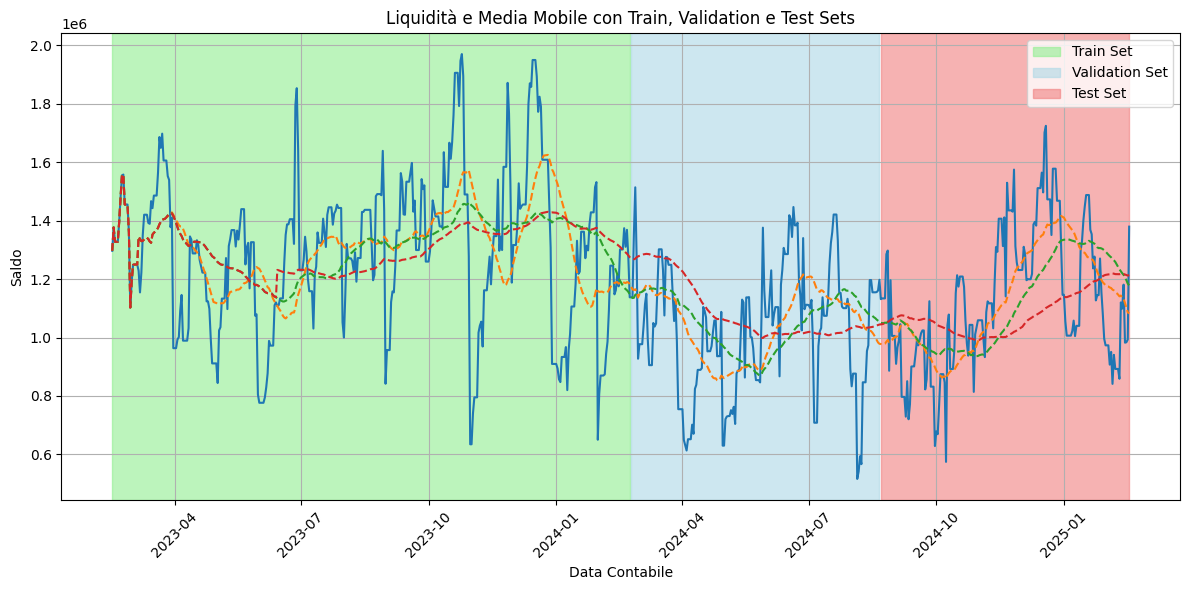
- Validation Set: 1440 giorni (primi 6 mesi del secondo anno);

- Test Set: 1440 giorni (ultimi 6 mesi del secondo anno).

**Perché proprio 8?**

Questo numero rappresenta un ottimo trade-off tra **generalizzazione** e **personalizzazione**. Se il modello si addestra su tante aziende, esso apprende pattern più generali del settore, rendendo le previsioni più robuste e meno sensibili alle anomalie di una singola azienda. Tuttavia, potrebbe perdere specificità rispetto alle peculiarità di ciascuna azienda. Se invece si addestra su poche aziende, o su una sola, il modello si adatta meglio alle dinamiche specifiche di quella realtà, ma rischia di sovradattarsi ai dati e di essere meno generalizzabile.

Il grafico successivo rappresenta graficamente la differenza per l’azienda 57.



## Ottavo Blocco

Il modello Prophet si compone di due parti. La prima parte, che interessa l’ottavo blocco, ha il compito di creare un modello generale, con l’obiettivo di trovare i migliori iperparametri per la previsione delle sedi storiche e selezionare il modello con l’errore più basso prevedendo il validation set. Il primo modello infatti studia il train set (il primo anno delle 8 aziende) e prevede il validation set (i primi 6 mesi del secondo anno delle 8 aziende), iterando su tutte le combinazioni di parametri per selezionare il modello con l’errore più basso.

Strutturalmente, i parametri del modello sono:

- gli iperparametri che controllano come Prophet modella i cambiamenti di trend e la stagionalità, come “changepoint\_prior\_scale”, “seasonality\_mode” e “seasonality\_prior\_scale”;

- componenti stagionali, con "model.add seasonality", che catturano pattern ricorrenti nel tempo con un indice di complessità, il “Fourier order”, che varia a seconda delle stagionalità;

- regressori, con "model.add regressor", per introdurre variabili esterne che potrebbero influenzare le previsioni, come il giorno, il mese o direttamente i Cash Flow (nel codice riportato come “Importo”).

Alla fine il modello risultante ha i seguenti errori:

Best Validation RMSE: 616036.59 €

Best Validation MAPE: 3.81%

**Perché sono stati scelti questi due indicatori di performance?**

Per diversi motivi:

- l’RMSE (**Radice dell'errore quadratico medio**) eleva al quadrato le differenze tra i valori previsti e quelli reali, dando maggiore peso agli errori più grandi. Questo è importante perché in molti casi, gli errori più grandi sono più costosi o problematici di quelli piccoli. In più, è espresso nelle stesse unità della variabile dipendente (euro in questo caso), il che lo rende facile da interpretare nel contesto del problema.

- il MAPE (**Errore percentuale assoluto medio**) è molto facile da visualizzare essendo un percentuale, è una misura relativa e quindi non dipende dalle unità di misura della variabile dipendente, il che lo rende utile per confrontare le prestazioni di modelli su dataset diversi, ma può essere problematico quando i valori reali sono vicini a zero o negativi, poiché può portare a valori di errore indefiniti o molto grandi.

Combinando queste due metriche, si può ottenere una comprensione più approfondita di quanto bene il modello stia prevedendo i valori e se ci sono particolari aree in cui il modello sta commettendo errori significativi.

## Nono e Decimo Blocco

Una volta che abbiamo il modello validato, prediciamo i valori di tutta la serie temporale (train, validation e test set) per calcolare i “Residui”, cioè la differenza tra i valori reali e i valori predetti, che poi verranno aggiunti come regressori nel secondo modello.

Questi due blocchi servono ad aggregare il train set e il validation set in un unico “Train set”, su cui il secondo modello si addestra per la previsione e il forecast dei sei mesi successivi.

## Undicesimo Blocco

In questo blocco è presente il secondo modello che utilizza i migliori parametri del primo, con l’aggiunta dei residui come nuovo regressore, per prevedere il test set delle 8 aziende.

I parametri di performance sono:

Final RMSE sul Test Set delle 8 Aziende: 1662048.29€

Final MAPE sul Test Set delle 8 Aziende: 14.45%

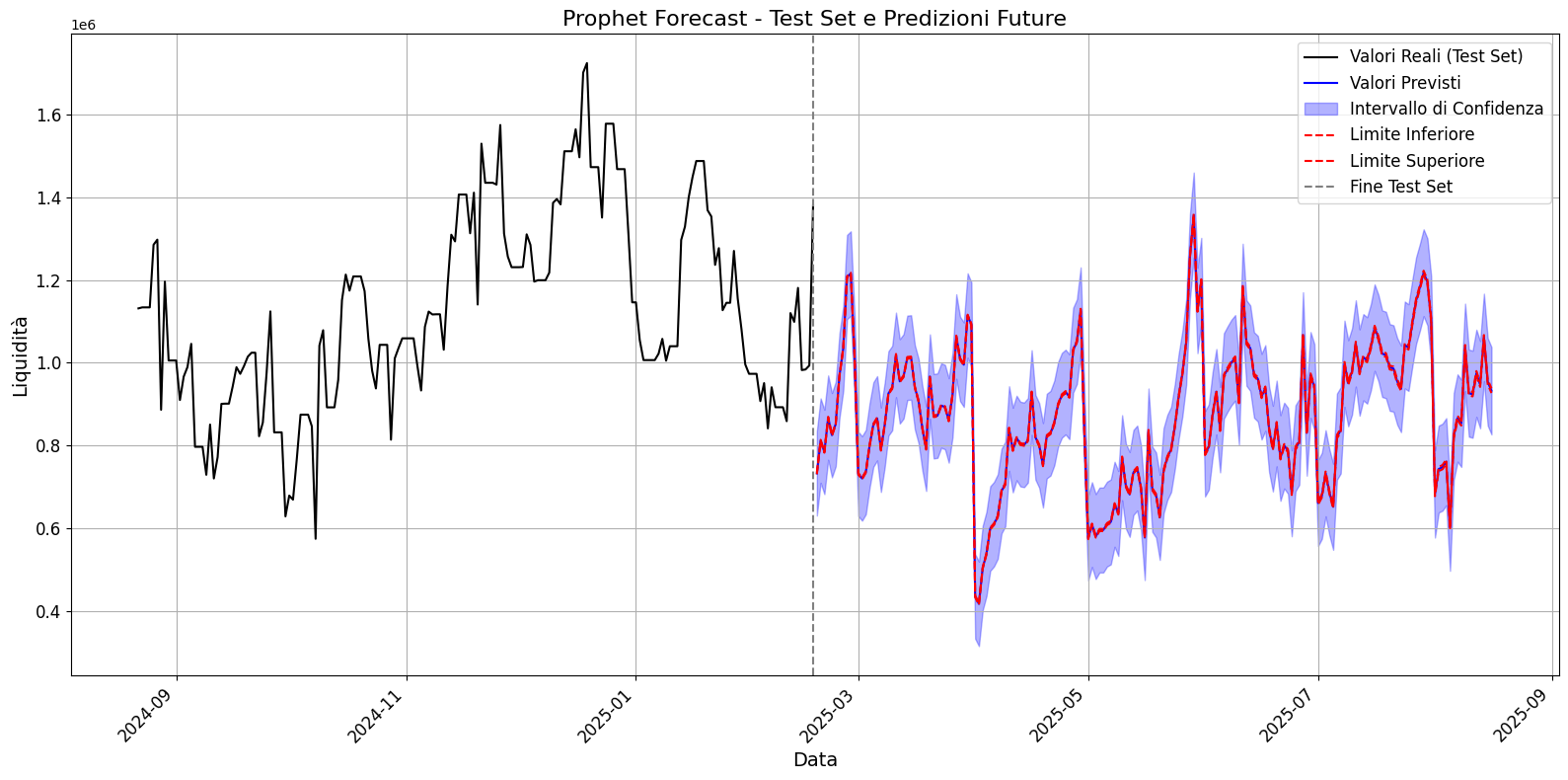
## Dodicesimo Blocco

In questo blocco il modello crea la previsione della Liquidità per i 6 mesi successivi. Per farlo crea un nuovo dataset, utilizzando la data del giorno successivo all’ultima data reale fino ai 6 mesi dopo, utilizzando i regressori (giorno, mese, anno, Importo e Residui) che poi verrà “compilato” dal modello con i valori predetti.

(I valori dei Residui sono anche predetti da una mini regressione lineare per aiutare le previsioni.)

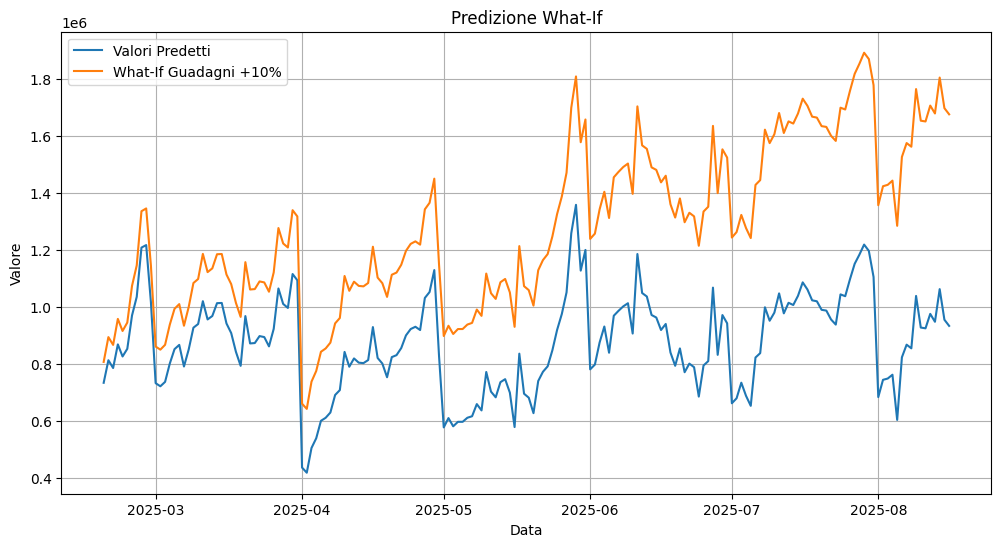
(Nel codice è presente un valore di soglia che aggiusta la previsione nel caso essa sia troppo sfasata dai valori reali. Questo succede perché, diversamente dai Residui, gli importi utilizzati sono quelli degli anni precedenti per questione di alta probabilità di stagionalità. Questi ultimi, contenendo anche molti zero nella serie, abbassano le previsioni, cioè fanno partire il primo valore predetto vicino allo zero.)

Questo è il grafico della previsione dell’azienda 57.



## Tredicesimo e Quattordicesimo Blocco

Vengono sviluppati degli scenari “What-If”, cioè delle situazioni alternative dove le entrate o le uscite aumentano del 10%. In entrambi i casi viene riportato il grafico e la variazione degli importi, cioè delle 5 principali entrate e uscite. Di seguito i grafici e i commenti che vengono generati.



Guadagni di Liquidità in 6 mesi: 7425648.57€

Guadagni di Liquidità What-If: 8168213.42€

La possibile aumenti dei guadagni di: 742564.86, è data da:

## Incidenza delle Top 5 Entrate e Potenziale Variazione nel What-If ##

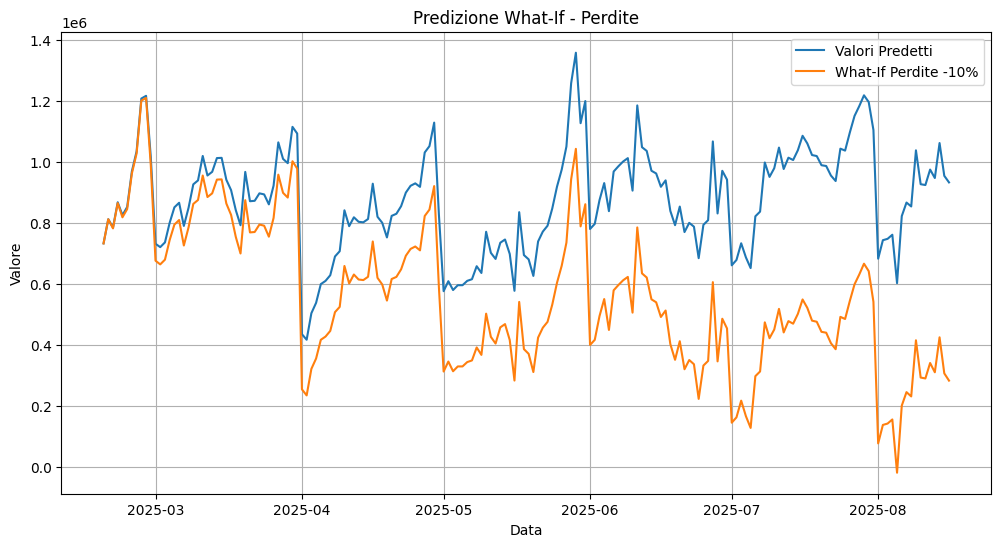
L'entrata 'BONIFICO A VS FAVORE' ha una percentuale di incidenza del 67.68%, che porta ad un'entrata possibile di 502574.64€ nel what-if.

L'entrata 'INCASSO TRAMITE P.O.S.' ha una percentuale di incidenza del 10.86%, che porta ad un'entrata possibile di 80660.13€ nel what-if.

L'entrata 'VERSAMENTO CONTANTI/VAL.ASSIM.' ha una percentuale di incidenza del 7.91%, che porta ad un'entrata possibile di 58768.34€ nel what-if.

L'entrata 'VERSAMENTO DI ASSEGNI BANCARI' ha una percentuale di incidenza del 7.17%, che porta ad un'entrata possibile di 53266.12€ nel what-if.

L'entrata 'VERSAMENTO DI ASSEGNI CIRCOLARI EMESSI DA ALTRE BANCHE' ha una percentuale di incidenza del 2.14%, che porta ad un'entrata possibile di 15921.30€ nel what-if.



## Incidenza delle Top 5 Uscite e Potenziale Variazione nel What-If ##

L'uscita 'PAGAMENTI DIVERSI' ha una percentuale di incidenza del 33.56%, che porta ad una variazione di 249194.77€ nel what-if.

L'uscita 'COMMISSIONI' ha una percentuale di incidenza del 29.29%, che porta ad una variazione di 217509.49€ nel what-if.

L'uscita 'DISPOSIZIONI DIVERSE' ha una percentuale di incidenza del 18.70%, che porta ad una variazione di 138871.44€ nel what-if.

L'uscita 'EFFETTI RITIRATI' ha una percentuale di incidenza del 9.01%, che porta ad una variazione di 66926.00€ nel what-if.

L'uscita 'IMPOSTE E TASSE' ha una percentuale di incidenza del 3.18%, che porta ad una variazione di 23633.24€ nel what-if.

# Conclusione e Ringraziamenti

Ci terrei a ringraziare tutto il team di Flash Flow per la loro disponibilità e gentilezza nonché per la grande professionalità che hanno dimostrato. Grazie al Datathon ho avuto la fortuna di conoscere persone molto preparate e molto simpatiche.

Grazie a Credem per aver organizzato questa fantastica iniziativa e per la macchinetta del caffè gratis.