Time series analysis sui dati delle vendite di un portale di e-commerce

Boller Mattia, De Rosso Daniel

Sommario

In questo report si presentano il lavoro di analisi e i risultati ottenuti da esso su dati riguardanti le vendite di un e-commerce italiano dal 2014 al 2022. Gli obiettivi perseguiti sono stati quelli di valutare l'impatto che il lockdown dovuto alla pandemia di SARS-Cov-2 ha avuto sulle entrate dell'azienda e di esaminare il trend passato e futuro dell'andamento delle vendite, con il fine di rilevare eventuali criticità che necessitano di un intervento in termini di modifica del piano di business. Per rispondere ai quesiti presentati si è effettuata in principio un'analisi esplorativa del dataset, volta a familiarizzare con i dati in possesso e all'applicazione delle dovute operazioni per renderli adeguati alle seguenti analisi. In secondo luogo si è andato ad identificare il modello adatto al caso per effettuare predizioni, andando a confrontare i modelli SARIMA e FbProphet in termini di MAPE sul medesimo periodo di test. I risultati della suddetta fase hanno portato alla scelta di FbProphet, rivelatosi migliore alla luce delle performance ottenute, modello quindi utilizzato per valutare l'impatto del periodo Covid-19 e per ottenere una stima sulla vendite attese nell'anno 2022. Dai risultati delle analisi è emerso che le vendite dell'e-commerce hanno subito un incremento dal 2014 al 2017, per poi iniziare a calare dal 2018 al 2019, trend decrescente spezzato momentaneamente dal periodo Covid-19 che ha fatto registrare entrate superiori del 83% rispetto alle entrate attese. Le previsioni per l'anno 2022, secondo le stime del modello creato, indicano un possibile calo del 20% dei guadagni rispetto all'anno 2021, con un valore atteso di introiti annui totali più basso rispetto ai 7 anni precedenti. Dai risultati ottenuti dallo studio si evince la necessità di applicazione di provvedimenti alla strategia di business, con il fine di invertire il trend negativo emerso in fase di studio.

Keywords

E-commerce — Time series analysis — SARIMA — FbProphet

Indice

1. Introduzione

L'e-commerce nel mondo sta vivendo negli ultimi anni una crescita esponenziale. Il numero di utenti fruitori di questo metodo di acquisto ha superato i 3,78 miliardi, con una crescita del 10% nel 2021 con 344 milioni di nuovi acquirenti. Si stima che nel 2022, l'e-commerce mondiale supererà la soglia dei 5mila miliardi di dollari di fatturato, crescita destinata ad aumentare negli anni avvenire. Anche l'Italia sta registrando numeri importanti su questo fronte, con una crescita del 78% del fatturato nel primo trimestre del 2021 e con una spesa media annua per persona su portali di e-commerce pari a 1.608€, con però il 67% dell'utenza che ha acquistato da siti esteri. [1]

In questo report si presenta l'analisi effettuata sulle vendite di un negozio di e-commerce italiano dedicato principalmente alla commercializzazione di forniture sportive, con l'obiettivo di comprendere l'andamento degli introiti dell'azienda ed individuare eventuali problematiche che necessitano di un intervento sul piano di business.

1.1 Dataset

Il dataset in analisi riguarda i dati storici raccolti da un e-commerce italiano negli anni 2013-2022. L'e-commerce in questione presenta un'ampia varietà di articoli nel suo catalogo, raggruppati per categorie di riferimento (30 in totale). Nel dettaglio, ogni record è ripartito in tre colonne: data (secondo il formato YYYY-MM-DD), totale (le vendite in euro) e il settore. Quindi ogni entry rappresenta le entrate totali, ottenute per una singola categoria in uno specifico giorno. In totale il dataset presenta 25.261 righe.

1.2 Domanda di ricerca

L'obiettivo dello studio riguarda l'analisi generale della serie storica, evidenziando in primo luogo gli effetti del periodo seguente l'inizio della pandemia sulle vendite, il quale può aver influenzato negativamente o positivamente l'andamento degli introiti dell'e-commerce. In secondo luogo si cerca di fornire una stima sugli andamenti futuri dell'attività, per cercare di anticipare trend di crescita o decrescita delle vendite con il fine di fornire un quadro sul quale costruire un'adeguata strategia di business.

2. Esplorazione dei dati

I dati originali sono stati sottoposti a numerose trasformazioni per permettere diverse analisi. In prima battuta si è verificata l'eventuale presenza di dati mancanti in termini di date non presenti. Si è riscontrata la mancanza di informazioni per alcune giornate comprese tra il 01/02/2013 e il 20/02/2014, periodo di tempo corrispondente al primo anno di

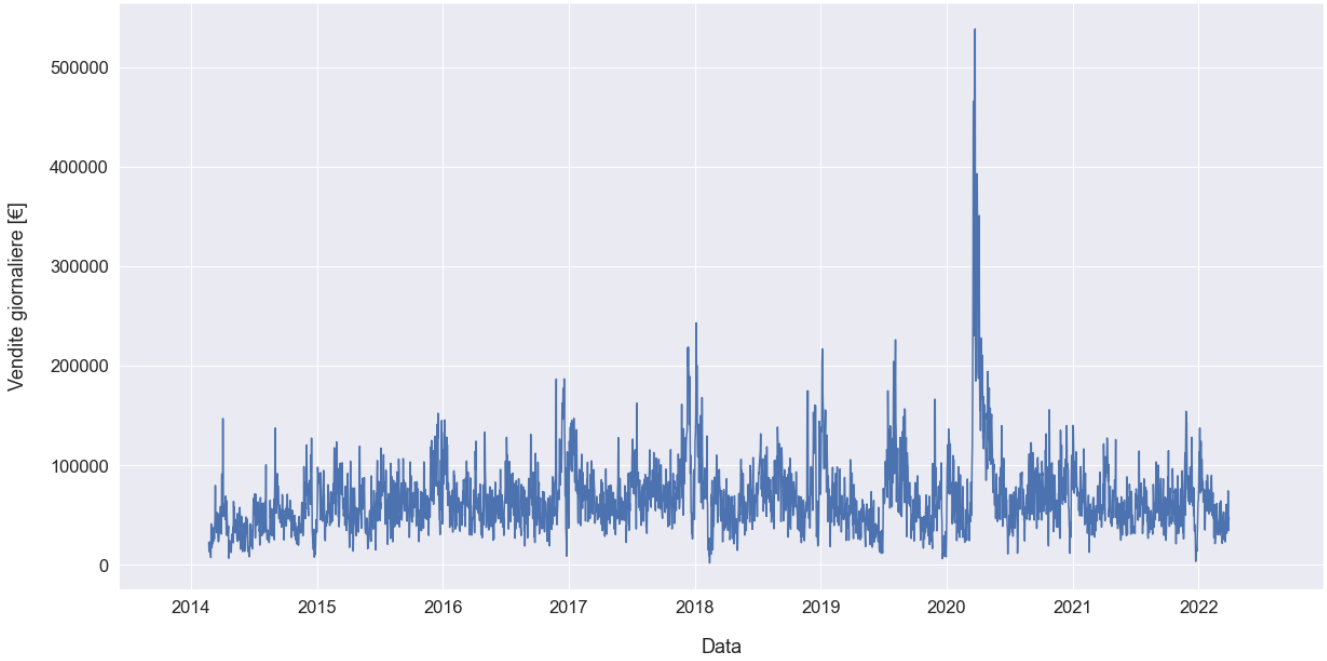

Figura 1. Vendite giornaliere totali

dati sulle vendite dell'azienda di e-commerce. Non essendo in possesso di informazioni riguardati la modalità di estrazione dei dati e del significato della mancanza di alcune giornate, non potendo sapere quindi se si trattasse di giornate prive di vendite o di dati mancanti causati da altri motivi, si è preferito utilizzare solo i dati dal 20/02/2014, scelta dettata anche dall'elevata estensione temporale del periodo privo di dati assenti.

I dati riguardati le vendite giornaliere divise per settori sono stati successivamente raggruppati per totale giornaliero, totale settimanale e totale mensile, in modo da analizzare l'andamento nel tempo delle vendite totali dell'e-commerce sotto diversi punti di vista e senza tenere conto dei settori. In particolare, in figura 1 è presente la visualizzazione riguardate il raggruppamento giornaliero delle vendite totali.

Da questa prima rappresentazione è già possibile notare un picco anomalo di vendite nei primi mesi del 2020, corrispondente con l'inizio del lockdown indetto a causa della pandemia da Covid-19. Analizzando gli introiti divisi per settore durante il periodo pandemico, rappresentati in figura 2, è facile notare come questo notevole aumento nelle vendite sia implicabile principalmente al settore "Fitness", presumibilmente causato dalla totale chiusura di palestre e impianti sportivi in quel determinato periodo e quindi alla conseguente volontà della popolazione di continuare a fare attività fisica in casa.

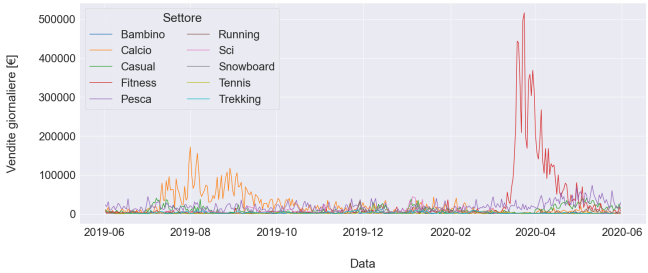

Figura 2. Vendite giornaliere per settore (limitati ai primi 10 settori per vendite totali)

Nonostante questo notevole incremento di vendite nel settore "Fitness", un'analisi sul totale delle

vendite per settore durante l'intero periodo di dati disponibili ha rivelato che le tipologie di prodotti che hanno portato all'e-commerce la quantità maggiore di entrate tra il 2014 e il 2020 sono stati quelli appartenenti alle categorie:

1. Pesca: 55.076.428 €
2. Calcio: 39.063.769 €
3. Casual: 35.470.317 €
4. Fitness: 25.187.121 €

2.1 Confronto con Google Trends

Si è deciso in questa fase di confrontare il dataset originale con dati provenienti da Google Trends [2]. Nella fase di esplorazione del dataset infatti, si è potuto notare un forte picco delle vendite nel settore del fitness a ridosso del periodo Covid-19. L'idea quindi è stata quella di usare i dati relativi alle ricerche degli utenti per tentare di spiegare questa varianza. Appoggiandosi alle API di Google Trends, si è ricavato un dataset che, per il periodo di tempo 2014/02/01-2022/03/31, riporta il valore relativo di interesse riguardante il topic "Fitness".

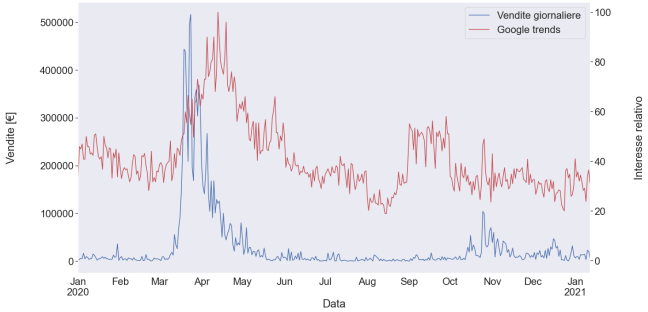

Figura 3. Valore di interesse relativo alla parola "Fitness" estratto da Google Trends (linea rossa) e quantità di vendite giornaliere nella categoria fitness (linea blu).

Osservando la figura 3 si può apprezzare la sovrapposizione dei picchi delle vendite del settore fitness con il relativo interesse nelle ricerche Google. Correlazione che però non viene confermata dall'indice di Pearson il quale, per l'arco temporale dell'anno 2020, ritorna un valore di R2=0.47325729. Difatti, sebbene da un punto di vista grafico si evince una coincidenza tra i due picchi, con un'analisi più attenta risalta un ritardo di qualche giorno tra le ricerche di Google e le vendite effettive, rendendo cosi inefficace lo studio della correlazione numerica.

Ai fini dello sviluppo del modello futuro quindi, non è stata ritenuta funzionante l'integrazione con i dati di Google Trends e il dataset di partenza, per i motivi precedentemente riportati. Ad ogni modo, questi dati sono tornati utili per ricondurre le cause dell'aumento delle vendite del settore "Fitness" nel periodo di pandemia, alla crescita nell'interesse delle persone in questo campo.

3. Analisi delle serie storiche

In questo capitolo verranno presentate le varie tecniche di analisi di serie storiche adottate per il trattamento dei dati sopradescritti, le scelte effettuate in termini di modelli di predizione e relativi parametri e i risultati ottenuti dalla loro applicazione.

3.1 Decomposizione

La prima operazione eseguita è stata quella di decomposizione in trend e seasonality, applicata in particolare alla serie storica con granularutà mensile. Il risultato della decomposizione è visulizzato in figura 4.

Dal grafico rappresentante il trend, è possibile notare un andamento crescente delle vendite iniziato nel 2014 e durato fino a circa il 2018, punto in cui ha avuto inizio un calo degli introiti, interotto solo temporaneamente dal periodo pandemico.

Si è effettuato uno studio sulla componente stagionale estratta, andando a verificare quali fossero i

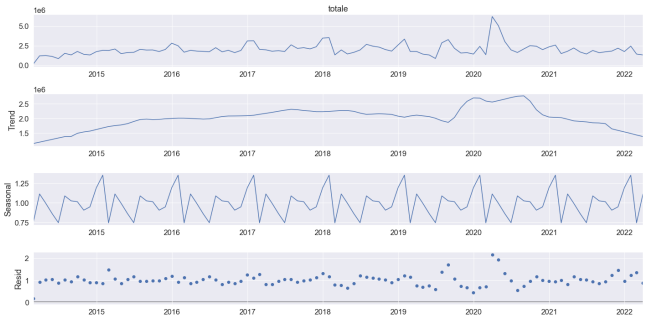

Figura 4. Decomposizione in trend, seasonality e residuals delle vendite totali mensili

mesi con una quantità di vendite che si discostasse maggiormente dalla quantità media. Ne è risultato che i mesi di gennaio e dicembre sono i mesi che registrano il maggior aumento di vendite (rispettivamente 36% e 19% superiore alla media), mentre febbraio e giugno presentano un calo degli introiti (rispettivamente 26% e 25% volte inferiore alla media). In figura 5 sono presenti i dati dei mesi citati e dei restanti.

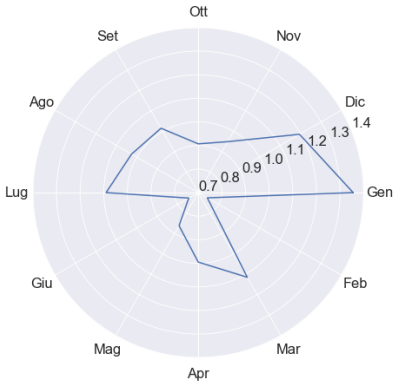

Figura 5. Analisi mensile della componente stagionale. Considerare che un valore di 1 corrisponde alla media delle vendite mensili

3.2 Model selection

Si è deciso in questo progetto di effettuare un'analisi comparativa delle performance tra due modelli per lo studio di serie storiche, SARIMA e FbProphet, con il fine di identificare il più adatto nel caso trattato e quindi di adottarlo per le fasi seguenti. Si specifica che da questo punto in poi, i dati utilizzati per la model selection e per le successive analisi, sono stati presi con granularità settimanale.

3.2.1 SARIMA

SARIMA (Seasonal Autoregressive Integrated Moving Average) è un estensione del modello ARIMA, con l'aggiunta della componente di modellazione della stagionalità. [3] Nella fase di decomposizione è infatti emersa una ripetitività periodica nella serie storica, da includere nel modello, corrispondente a picchi di vendite nei mesi di dicembre e gennaio, e altri movimenti minori ma pur sempre ripetuti.

Il modello SARIMA riceve in input un totale di 7 parametri (p,d,q,P,D,Q,m), gli ultimi 4 sono un'aggiunta al modello ARIMA:

- P: ordine autoregressivo stagionale.
- D: ordine della differenza stagionale.
- Q: ordine della media mobile stagionale.
- m: il numero di fasi temporali per un singolo periodo stagionale.

I parametri del modello sono stati estratti con una funzione di tuning degli iperparametri per ottenere la configurazione ottimale. L'output di tale ricerca è stato un modello SARIMA(4,1,1)(0,1,1)[52] con un valore AIC=4436.532.

La figura 6 permette di diagnosticare il modello. Attraverso tali grafici si può osservare:

- un'assenza di pattern evidenti nei residui, con media zero e varianza uniforme
- la curva KDE è paragonabile ad una normale
- la maggior parte dei punti della normal Q-Q si estendono su una linea retta

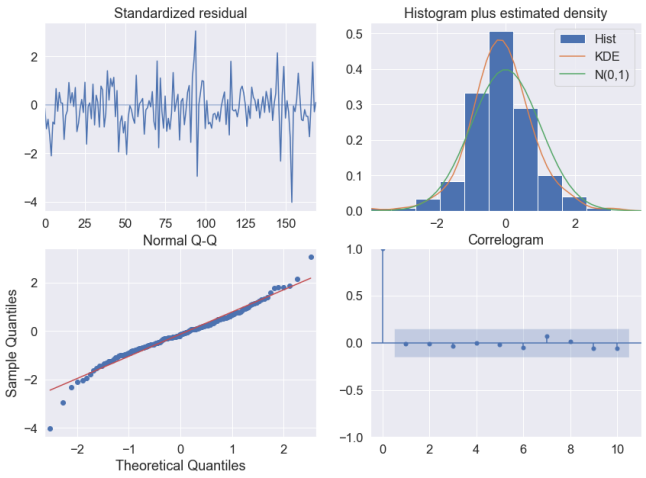

Figura 6. Diagnostica modello SARIMA

- nessun punto del correlogramma cade al di fuori della regione di confidenza e quindi non c'é significatività

Per permettere il confronto con altri modelli, in particolare con FbProphet di cui si parla nella successiva sezione, si è allenato il modello su una parte dei dati, che vanno dal 20/04/2014 al 01/06/2018, e si è utilizzato per prevedere le vendite nel periodo dal 01/06/2018 al 01/03/2020, confrontando i valori restituiti dal modello con quelli reali in termini di MAPE (Mean Absolute Percentage Error), definito come:

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \qquad (1)$$

dove A_t è il valore attuale e F_t è il valore predetto. Si tratta di una misura ampiamente usata nella valutazione di predizioni in serie temporali data la sua facile interpretabilità. [4]

Le performance in termini di MAPE del modello SARIMA ottenuto sono pari al 42,18%.

3.2.2 FbProphet

Prophet è un processo di previsione dei dati delle serie temporali basato su un modello additivo, in cui le tendenze non lineari vengono adattate alla stagionalità annuale, settimanale e giornaliera (con la possibilità di aggiungere giorni festivi). Funziona in maniera ottimale con le serie temporali che hanno forti componenti stagionali e diverse periodicità di dati storici. Prophet è robusto in caso di valori mancanti e alle variazioni del trend, è inoltre in grado di gestire gli outliers. [5]

FbProphet viene formalmente formulato come:

$$y(t) = g(t) + s(t) + h(t) + e_t \qquad (2)$$

dove:

- $g(t)$ è la funzione di trend (tasso di crescita) che modella i cambiamenti non periodici della serie temporale (il trend di una funzione a tratti);
- $s(t)$ rappresenta i cambiamenti periodici (stagionalita giornaliere, settimanali, mensili e annuali);
- $h(t)$ rappresenta i periodi di vacanza che avvengono irregolamente durante uno o piu giorni a seconda dell'anno;
- e_t rappresenta ogni cambiamento che il modello non e in grado di catturare, ovvero l'errore (ipotizzato normalmente distribuito).

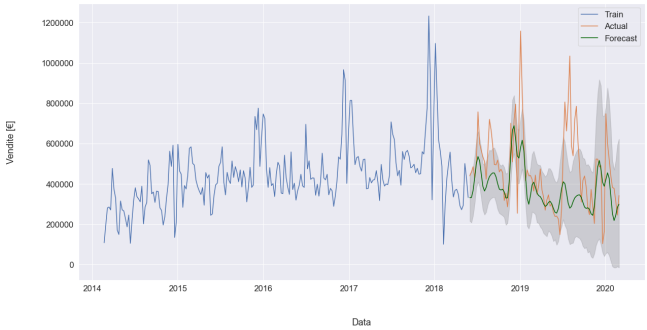

Figura 7. Comparazione valori predetti (linea verde) e attuali (linea arancione) modello FbProphet

Per identificare i migliori parametri del modello si è effettuata un'operazione di hyperparameters tuning con performance misurate tramite cross validation, funzione messa a disposizione dalla libreria fbprophet. Questa tecnica si basa sul selezionare diversi punti di cutoff nella serie storica e ad ogni passaggio allenare il modello solo sui dati precedenti al determinato punto di cutoff. Ad ogni iterazione si confrontano i valori predetti fino ad un determinato periodo, chiamato punto di horizon, con i valori reali della serie storica, in modo da ottenere misure di performance affidabili per ogni modello creato con diversi parametri. [5]

Come risultato della fase di hyperparameter tuning si sono ottenuti i seguenti valori:

- changepoint_prior_scale=1,0. Parametro che determina la flessibilità del trend, in particolare quanto il trend cambia nei changepoints. Da intendersi come un regolatore di underfitting (valori bassi) e overfitting (valori alti).
- seasonality_prior_scale=1,0. Parametro che determina la flessibilità nella componente di stagionalità.
- seasonality_mode='multiplicative'. Parametro che specifica la serie storica additiva o moltiplicativa.

Come nel modello SARIMA, anche in questo caso si è utilizzata come metrica di performance il valore MAPE (maggiori dettagli nel paragrafo precedente) sui dati compresi nel periodo 01/06/2018 - 01/03/2020. Nel dettaglio si è ottenuta una misura pari a 28,72%, valore nettamente inferiore rispetto al 42,18% ottenuto dal modello SARIMA, motivo per cui si è preferito FbProphet.

3.3 Analisi impatto covid

Uno degli obiettivi primari fissati in fase iniziale, è stato quello di comprendere in che modo il Covid-19 e le conseguenti restrinzioni avessero impattato sulle vendite del sito di e-commerce. Dalla semplice analisi presentata in fase di esplorazione dei dati, nel capitolo 2, è stato possibile intuire che il periodo di lockdown abbia portato ad un aumento degli introiti.

Per quantificare questo incremento si è sfruttato un modello FbProphet, con i parametri ottenuti in fase di model selection e hyperparameters tuning presentate nel capitolo 3.2.2, utilizzando come dati di training le vendite totali settimanali fino al 01/01/2020, escludendo quindi il periodo Covid. Successivamente si è utilizzato il modello per prevedere le vendite dei successivi 2 anni, fino al 31/12/2022, di cui erano disponibili i dati reali, permettendo quindi un confronto. In figura 8 sono visualizzati i valori di vendite settimanali predetti dal modello allenato con dati pre-Covid e i valori reali.

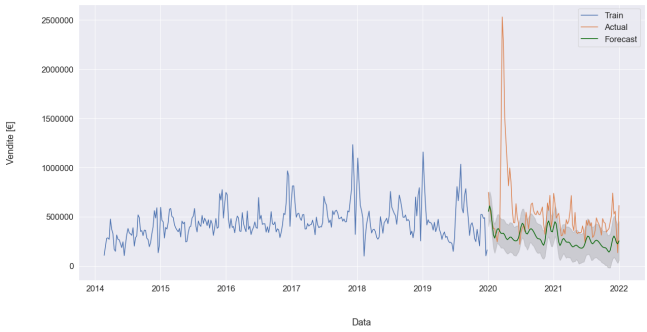

Figura 8. Predizione vendite durante il periodo Covid-19 (linea verde) a confronto con i valori reali (linea arancione)

Confrontando i dati predetti con quelli reali relativi all'intero anno 2020, secondo il modello FbProphet creato, la pandemia ha portato ad una quantità di vendite totali annuali superiori del 83% rispetto a quelle attese.

3.4 Analisi vendite future

L'analisi predittiva delle vendite è stata svolta con una proiezione di un anno degli introiti. Il modello FbProphet è stato allenato su tutti i dati a disposizione, con l'obiettivo di prevedere le vendite future per ottenere una stima dei guadagni nell'anno 2022. Nella figura 9 viene mostrato l'output del modello, con relativa stima puntuale e barre di confidenza al 95%.

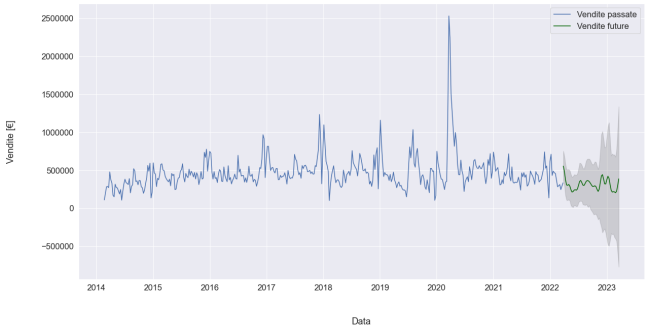

Figura 9. Predizione vendite future per l'anno 2022

Esaminando la figura è ipotizzabile un ribasso della vendite future. Il trend della serie storica, in prossimità della fine del lockdown, mostra infatti una tendenza decrescente, che viene intercettata dal modello predittivo.

I dati ottenuti dal modello sono stati confrontati con i dati passati in termini di introiti mensili medi nell'arco di un anno e di vendite totali annuali. I valori per ogni anno sono presentati in tabella 1.

Si può notare come, secondo il modello creato, nel 2022 il sito di e-commerce potrebbe registrare dei valori di vendite mensili medie e di vendite totali inferiori rispetto ai precedenti 7 anni, superiori solamente ai dati del 2014, primo anno di dati a disposizione e incompleto del primo mese di vendite, essendo i dati riferiti al mese di gennaio non disponibili. In particolare si nota un possibile calo delle vendite totali attese nel 2022 del 20% rispetto all'anno precedente (2021) e del 46% rispetto all'anno che ha registrato gli introiti maggiori (2020).

Tabella 1. Tabella con i valori di vendite mensili medie e di vendite totali divise per anno

	Vendite mensili medie [€]	**Vendite annuali totali [€]**
2014	1.249.578	13.745.360
2015	1.938.444	23.261.330
2016	1.997.196	23.966.350
2017	2.298.046	27.576.550
2018	2.147.691	25.772.290
2019	1.957.926	23.495.120
2020	2.758.411	33.100.930
2021	1.855.226	22.262.710
2022	1.491.914	17.902.970

In figura 10 si mostra il grafico delle vendite mensili suddivise per anno, con l'aggiunta della predizione per il periodo 01/01/2022-31/12/2022. Anche da questa situazione è evidente il calo delle vendite del 2022 rispetto agli anni precedenti. Si specifica che dal grafico è stato omesso l'anno 2014 essendo privo di dati per quanto riguarda il mese di gennaio.

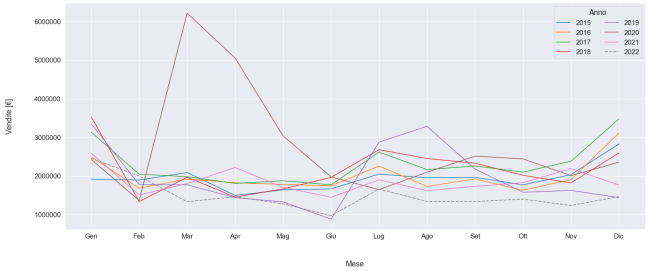

Figura 10. Guadagni totali per mese, con predizione dell'anno 2022 (linea tratteggiata)

4. Risultati e conclusioni

In questo lavoro si è mostrato come, per quanto riguarda il dataset a disposizione, il modello Fb-Prophet riporti dei valori di MAPE migliori rispetto al modello SARIMA, risultati che hanno fatto preferire il primo per le analisi di business dell'e-commerce. Rispondendo alle domande di ricerca precedentemente poste:

- il periodo della pandemia SARS-Cov-2 ha influenzato positivimanete le entrate dell'e-commerce, registrando un aumento del 83% rispetto alle vendite totali attese del modello predittivo;
- si prevede un trend a ribasso per l'anno 2022, con dei rialzi stagionali estivi e picchi positivi nei mesi di dicembre e gennaio. Complessivamente, le vendite totali previste nell'anno 2022, predette a 17.902.970€, risultano più basse del 20% rispetto al 2021 e inferiori del 46% rispetto al 2020, anno con che ha registrato gli introiti maggiori.

I dati ottenuti da queste analisi possono rivelarsi utili ai fini di gestione dell'attività, per effettuare ribilanciamenti del magazzino o investire in campagne di marketing per risollevare le vendite. Inoltre lo studio della serie storica in periodo di Covid-19 ha evidenziato come il picco di vendite del settore "Fitness" sia stato un evento anomalo e non protratto nel tempo.

In un momento di così grande crescita per l'e-commerce in Italia e nel mondo, è importante identificare le possibili cause del trend decrescente previsto, con il fine di adottare le corrette strategie di business, operazione possibile solo se in possesso di una maggiore quantità di informazioni riguardanti il caso in questione.

Riferimenti bibliografici

[1] C. Associati, "E-commerce in italia 2022," report sull'andamento dell'e-commerce in Italia.

[2] Google, "Google trends," sito di ricerca trends per parola chiave. [Online]. Available: https://trends.google.com/trends

[3] "Sarima," introduction to SARIMA for time series forecasting. [Online]. Available: https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

[4] "Mean absolute percentage error." [Online]. Available: https://docs.oracle.com/en/cloud/saas/planning-budgeting-cloud/pfusu/insights_metrics_MAPE.html

[5] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018. [Online]. Available: https://doi.org/10.1080/00031305.2017.1380080