

Metodologia per estrarre informazioni e valore dai dati del fatturato di un e-commerce

Merlo Fabrizio 847203
Poterti Daniele 844892
Sanvito Alessio 844785
Sanvito Simone 844794

Data Science Lab
AA 2021/2022

Sinossi

Con l'evoluzione tecnologica e la diffusione sempre maggiore delle piattaforme di e-commerce si ha un mondo sempre più immerso nei dati, anche per lo "shopping online".

L'obiettivo delle nuove aziende data-driven è quello di prendere decisioni motivate, basate su elementi oggettivi e non su istinto o sensazioni del momento poiché, nonostante alcuni eventi non siano prevedibili (come, ad esempio, la pandemia di Covid-19), tendenzialmente tutti gli eventi futuri hanno delle circostanze simili, degli schemi che si ripetono.

Un esempio di ciò è il periodo natalizio: sarebbe interessante capire quali tipi di promozioni fornire ai propri clienti per aumentare le vendite basandosi sui dati raccolti negli anni precedenti; infatti, i dati possono suggerire le risposte a domande come queste.

Quindi il lavoro che è stato fatto mira proprio ad esplorare ed elaborare i dati per poter rispondere a quesiti che possano portare informazioni utili all'interno dell'azienda.

In particolare, in questo elaborato, viene presentata una metodologia da applicare a più settori di vendita di un sito ecommerce per estrarre informazioni e valore dai dati raccolti.

L'analisi parte da una fase esplorativa che risulta essere molto preziosa nel fornire informazioni per effettuare delle previsioni dettagliate ed il più accurate possibili. Le informazioni estratte vengono successivamente testate e validate con l'analisi predittiva.

Sommario

Sinossi	1
Sommario	2
Analisi esplorativa	3
Analisi predittiva	5
Conclusioni e sviluppi futuri	7
Riferimenti	8
Appendice	9
Aspetti metodologici	9
ARIMA(X)	9
PROPHET	10
ADF	12
Preprocessing	12
Calcio	14
Casual	17
Fitness	20
Pesca	21
Running	24
Costruzione modelli di previsione	26
Analisi risultati modelli	28
Calcio	28
Casual	28
Pesca	29
Running	29

1. Analisi esplorativa

Come prima fase del progetto è stata effettuata un'analisi esplorativa-descrittiva sui dati presenti nel dataset iniziale. Per effettuare una precisa analisi esplorativa si è scelto di selezionare le categorie più adatte a questo tipo di lavoro, ovvero quelle con dati relativi a più di 2000 giorni in modo da poter disporre di una numerosità tale da coprire la maggior parte del periodo storico in esame.

L'approccio iniziale è stato quello di visualizzare i dati graficamente per vedere gli andamenti di ogni settore.

Essendo la granularità di questo dataset molto densa, il focus si è concentrato nel creare un grafico il più riassuntivo ed esplicativo possibile; dunque la prima analisi è stata effettuata sulla somma cumulata dei fatturati di ogni mese.

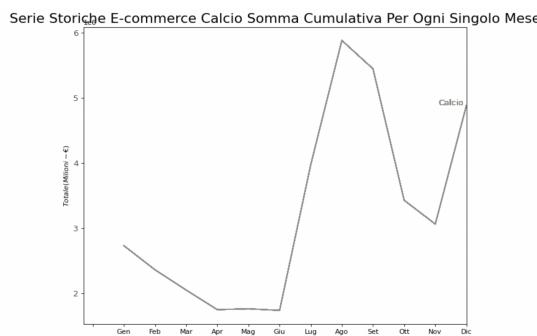


Figura 1.1: Serie storica del settore Calcio sommando i mesi di ogni anno

Per quanto riguarda il settore calcio, è ben visibile nel grafico un picco nel mese di luglio ed agosto. Sembra che ogni anno in quel periodo le vendite aumentino in modo sostanziale.

Anche provando a giustificare questo aumento integrando questa informazione con le conoscenze in nostro possesso in

ambito calcistico, non siamo riusciti a spiegare perché ciò avvenisse.

Rivolgendosi al fornitore del dataset e richiedendo informazioni, si è scoperto che l'azienda responsabile dell'e-commerce predispose maggiormente il budget pubblicitario nel periodo dei saldi (invernali, estivi, Black Friday) e durante le festività natalizie.

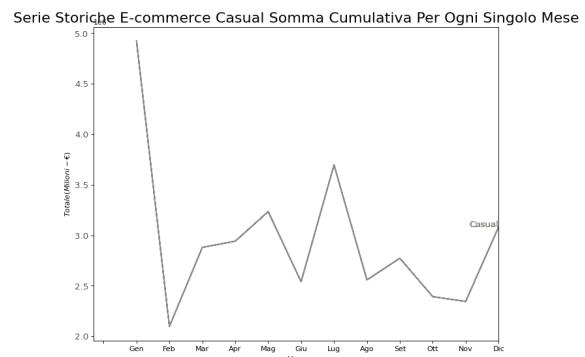


Figura 1.2: Serie storica del settore Casual sommando i mesi di ogni anno

Grazie a queste informazioni si riescono a spiegare maggiormente anche i picchi che caratterizzano il grafico del settore Casual. Si giunge alla conclusione che, nella futura analisi predittiva, può risultare fondamentale prendere in considerazione questi periodi.

Successivamente, dopo aver riflettuto sulla granularità e sul tipo di dato a disposizione, è stata condotta un'analisi volta a confrontare l'andamento del fatturato con alcuni indici statistici (come ad esempio Double Exponential Smoothing o Triple Exponential Smoothing) che evidenziano caratteristiche delle serie storiche quali trend e stagionalità.

Questa analisi statistica è fondamentale per capire quali fattori influenzano maggiormente l'evoluzione delle serie storiche e quindi per effettuare al meglio l'analisi predittiva.

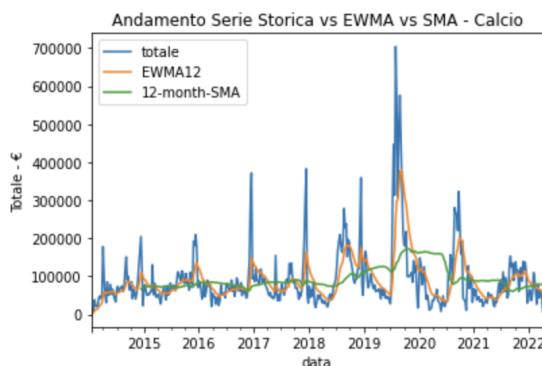


Figura 1.3: Andamento serie storica del settore Calcio con l'EWMA e SMA

Vediamo in figura 1.3 come l'exponential moving average riesce a seguire meglio l'andamento della variabile totale rispetto al simple moving average dando un peso maggiore ai valori più recenti. Tramite questa analisi, si è raggiunta l'evidenza statistica della necessità di adottare, nella fase di predictive analytics, dei modelli rolling che tengono conto degli ultimi giorni. Grazie a questi modelli si è in grado di spiegare meglio l'andamento dei vari settori e si riesce ad avere una visione più chiara sull'evoluzione futura delle vendite.

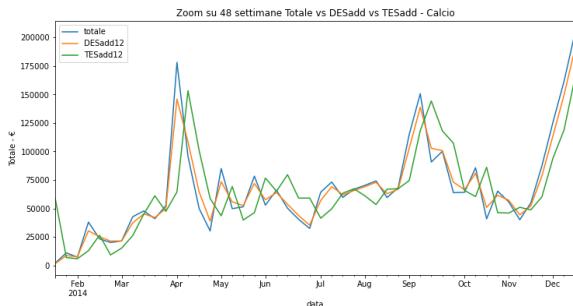


Figura 1.4: Zoom dell'andamento serie storica del settore Calcio con DESadd12 e TESadd12

Il triple exponential smoothing sembra descrivere accuratamente l'andamento del fatturato totale, ma non è molto preciso in quanto appare traslato di due settimane in avanti; si afferma che sono presenti sia la componente di trend (evidenziata dal DES)

che una debole componente di stagionalità (mostrata dal TES).

Le considerazioni effettuate su queste tecniche statistiche sono estendibili e valide per tutti gli altri settori considerati nell'analisi (tranne fitness che presenta un bias molto elevato che porta a dei risultati poco esplicativi).

In ultima analisi, dopo aver studiato l'andamento delle serie storiche relative a tutte le categorie, è stata cercata una possibile correlazione tra il fatturato e la diffusione del Covid; in particolare, questa correlazione è stata molto evidente per il settore fitness durante la prima quarantena .

Per rendere evidente questa asserzione sono stati utilizzati un grafico e una matrice di correlazione.

Dal grafico in figura 1.5 (in cui vengono mostrati rispettivamente l'andamento del fatturato e l'andamento dell'indice Rt) è evidente che, in vari momenti (tra il 2020-03-09 e il 2020-03-31), al diminuire del valore dell'indice cresce la vendita di prodotti della categoria Fitness.

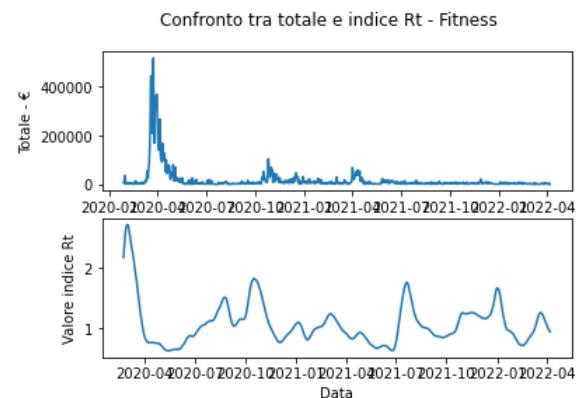


Figura 1.5: Confronto tra il totale del settore Fitness e l'indice Rt del Covid

Questo fenomeno è interpretabile nel seguente modo: una volta iniziato il lockdown, essendo la popolazione costretta a rimanere in casa, il valore dell'indice di contagio è andato in calo (inevitabilmente); di contro il fatturato della categoria fitness è andato in crescita, poiché le persone, non potendo accedere, ad esempio, alle palestre, hanno iniziato ad attrezzarsi per disporre di propri strumenti per tenersi in forma.

Allo stesso modo la matrice di correlazione evidenzia che l'indice Rt e il fatturato (del settore Fitness) sono correlati negativamente, con un coefficiente di correlazione di Pearson che vale -0.73; quindi al crescere del fatturato corrisponde una diminuzione del numero dei casi (e viceversa).

	totale	rt_positivi
totale	1.00000	-0.72963
rt_positivi	-0.72963	1.00000

Figura 1.6: Matrice di correlazione tra il fatturato totale di Fitness e l'andamento dell'indice Rt (periodo dal 2020-03-09 al 2020-03-31)

Concludendo, grazie a queste analisi sono stati rilevati aspetti fondamentali da tenere in considerazione per la fase di predizione: innanzitutto si è visto come i periodi di saldi portino, solitamente, ad un aumento di vendite, e quindi ad un aumento del fatturato; è stata riscontrata, nell'andamento di tutti i settori, la presenza di trend e di una leggera stagionalità. Inoltre, è stata individuata una maggior influenza dei dati recenti: per questo si è deciso di procedere con la costruzione di modelli rolling. Infine, si è notato come l'andamento del fatturato è abbastanza legato all'andamento del covid, in alcuni periodi.

Queste informazioni sono utili non solo per un'analisi predittiva, ma anche per comprendere meglio la natura del business.

Per ulteriori considerazioni, anche relative alle altre categorie, sono presenti nell'[appendice](#) tutti i grafici specifici, i riferimenti statistici e le analisi fatte sopra.

2. Analisi predittiva

Nella fase di analisi predittiva l'obiettivo è stato quello di effettuare delle previsioni sui primi giorni del 2022 usando come dati di training tutti quelli precedenti al 01-01-2022.

Sono stati creati dei modelli usando due librerie che fossero in grado di prevedere i dati futuri considerando anche delle variabili esogene: Prophet e ARIMA(X).

Prophet è molto utile per le previsioni di business che si riferiscono a osservazioni orarie, giornaliere o settimanali con forti stagionalità multiple. Facebook Prophet è progettato anche per apprendere e effettuare previsioni tenendo in considerazione le festività note. Inoltre è sviluppato per operare su serie che subiscono cambiamenti di regime, dovuti a eventi particolari e non ordinari come ad esempio il lancio di un prodotto.

ARIMA(X) viene utilizzato per vedere quali siano gli eventi che influenzano le vendite (e quindi il fatturato) e per prevedere il fatturato giornaliero dei settori considerati.

Come variabili esogene sono utilizzate delle serie che identificano il periodo di saldi, il periodo di lockdown e i pattern di stagionalità con Fourier; questa scelta è stata dettata da quanto emerso nella fase di analisi descrittiva.

La caratteristica dei modelli creati è che sono rolling: prendono una finestra di un giorno e un numero di iterazioni iter=14. In questo modo si prevede un solo giorno, a partire dal primo giorno t, proseguendo poi con la previsione del singolo giorno al tempo t+1, con la previsione del singolo giorno al tempo t+2, e così via, per finire con la previsione del singolo giorno al tempo t+13. Si prevede dunque una finestra di tempo giornaliera che si trasla di giorno in giorno fino ad arrivare all'ultimo (quattordicesimo) giorno.

Questo è molto utile per avere delle previsioni soggette a un basso errore, ovvero il più fedeli possibile ai valori reali del fatturato.

Sono stati dunque inizialmente usati i due modelli senza variabili esogene, successivamente sono stati previsti i valori del fatturato ed è stato calcolato l'errore percentuale (RMSPE); dal momento che gli errori calcolati erano parecchio alti (tra il 30 e il 50% nella maggior parte dei casi) si è deciso di inserire in ARIMA(X) la variabile esogena delle serie di Fourier per essere in grado di dare importanza alla stagionalità dei dati.

Anche in questo caso i risultati non sono stati molto validi, quindi, valutando ciò che era stato trovato nell'analisi esplorativa, la decisione che è stata presa è stata quella di inserire la variabile esogena relativa ai saldi (infatti questa componente sembrava molto influente nei grafici). Questa scelta si è rivelata corretta in quanto tra Prophet e ARIMA l'aggiunta della componente dei saldi ha portato a trovare i modelli con l'errore minimo nella maggior parte dei casi (su 14 giorni che vengono predetti, il modello con la variabile esogena saldi ha

l'errore minore in circa 6/7 casi per ogni categoria).

Un'altra componente che era stata ritenuta rilevante nella fase di analisi esplorativa è stata quella relativa al covid e al periodo lockdown; è stato notato però che, inserendo questa variabile esogena, l'errore delle previsioni non si è discostato molto da quello trovato dai modelli che utilizzano i saldi come variabili esogene.

Questo modo di procedere assume un valore considerevole nel momento in cui si vanno a confrontare gli errori ottenuti con i vari modelli e si vanno a scegliere giorno per giorno i modelli migliori (con l'errore minimo); in questo modo si potrebbe valutare quale tipo di modello potrebbe essere usato nelle previsioni di un determinato settore, valutando anche l'inserimento delle variabili esogene più influenti e significative.

Questo metodo permette di verificare e capire quanto ipotizzato nell'analisi descrittiva. Ad esempio, per quanto riguarda il settore calcio, si nota che la variabile esogena saldi riesce a spiegare meglio l'andamento delle vendite nel periodo in esame.

I risultati ottenuti sono visibili nelle immagini sottostanti.

	rmseProphet	rmseProphetsSaldi	rmseProphetCovid	rmseARIMAFourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid
0	82.009184	49.625998	39.679220	60.385979	8.857473	8.807679	8.735079
1	31.744204	8.273310	15.525257	247.941564	114.978100	114.879764	114.736390
2	45.230588	35.010448	32.299804	13.864773	35.918204	35.947516	35.990254
3	54.343857	47.252377	44.774845	7.286638	44.844629	44.869859	44.906643
4	29.466332	18.472414	17.249469	102.466610	4.428713	2.399746	2.463485
5	72.448174	68.844986	64.593471	5.280402	58.663735	57.786171	57.813739
6	45.876180	42.653074	41.293601	103.096634	17.318589	15.583273	15.618415
7	70.542228	69.964276	68.614647	218.696127	22.303090	24.897415	24.815649
8	67.478881	67.792131	67.869588	122.934280	18.398132	16.665734	16.720157
9	229.409834	220.082858	232.724049	825.941982	227.498205	234.451973	234.233556
10	0.314139	10.054821	7.475921	172.012275	5.723172	3.721666	3.784561
11	16.304813	30.837864	26.872043	129.210364	20.978744	19.301132	19.353834
12	13.985453	23.786521	25.213218	130.283717	19.764160	18.060763	18.114274
13	31.398705	35.087321	44.845689	103.725415	27.046094	25.497291	25.545946

Tabella 2.1: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Calcio

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAFourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid
0	130.990376	139.649368	127.773376	48.342874	54.171957	41.187940	41.027719
1	38.543938	38.257181	39.454624	62.489434	61.404110	64.054572	64.094682
2	3.622652	5.985434	1.258489	43.899257	42.855853	47.668406	47.727795
3	5.212833	6.824626	3.087039	42.798631	42.316407	47.174383	47.234340
4	17.527881	16.487709	19.768856	52.801916	52.876596	39.794883	39.833854
5	24.873369	26.564753	21.668769	29.842377	30.631580	11.359800	11.431898
6	10.494961	10.731491	6.838645	35.387579	36.714212	19.132303	19.198071
7	27.809324	27.996767	23.625484	24.829227	27.036988	6.766575	6.842400
8	40.427148	38.642117	34.966167	18.935116	21.992185	0.320237	0.401305
9	512.209923	504.773360	486.546124	233.446887	218.269507	306.690584	306.359831
10	15.048958	12.385782	9.849710	34.471878	37.928763	20.684276	20.747874
11	185.285211	176.835456	170.264106	78.242053	67.653561	114.230679	114.056450
12	133.337708	125.716969	120.313801	51.042467	41.155551	80.371054	80.224361
13	14.195713	9.153863	6.549961	19.546915	25.251563	4.488511	4.563191

Tabella 2.2: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Casual

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAFourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid
0	26.111025	32.257752	13.805024	3.798041	10.636075	3.012696	7.583869
1	6.322372	12.030684	0.160988	30.467473	22.405092	27.752524	35.184376
2	21.265906	26.222138	16.650599	40.049864	55.165629	39.633041	45.842785
3	127.863718	141.551702	122.555196	22.458388	28.308592	19.467302	7.178107
4	28.201668	34.149243	23.646980	23.737283	22.535837	7.430492	14.849906
5	80.066021	96.684684	79.662589	23.728761	22.050902	45.796790	34.111227
6	83.419343	101.920483	87.499350	34.033863	28.560423	53.633731	41.320040
7	12.856647	25.252185	15.575118	8.849264	14.853245	1.774997	6.382235
8	31.482259	24.471638	20.501387	51.949394	50.042648	47.471058	51.681235
9	91.504054	107.740192	98.836725	3.612726	7.004298	11.129664	2.222669
10	120.884628	141.483079	133.394506	27.830618	12.941329	34.964644	24.147274
11	48.838415	65.030277	57.860505	4.350312	16.624375	0.246808	8.241992
12	28.324320	42.544794	37.305932	9.260200	21.498903	6.191358	13.710088
13	24.431539	38.781229	33.305161	9.448257	22.097071	6.906168	14.367607

Tabella 2.3: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Pesca

Si noti, ad esempio, nella categoria Pesca, che il 50% dei modelli costruiti funziona meglio se al loro interno viene utilizzata la variabile esogena relativa ai saldi.

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAFourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid
0	64.522606	58.942795	50.926623	80.716164	80.817830	69.887439	72.886761
1	77.482301	78.251887	79.409494	66.638076	67.033030	48.247780	53.371557
2	57.543511	58.493295	60.222942	78.714423	78.851238	66.800242	70.087215
3	66.351360	66.805674	67.573503	61.368590	61.639867	39.780068	45.742199
4	39.429507	39.965271	41.378333	78.020673	78.186864	45.359696	67.750031
5	66.958598	67.209237	67.771097	46.591136	47.025240	13.444971	21.678646
6	10.712026	10.542672	11.459087	19.841656	20.537860	29.832535	17.482031
7	27.876911	28.406125	25.293190	45.213830	45.719936	11.312125	19.748799
8	18.027608	17.981200	19.351327	6.339040	7.255553	51.534420	37.119493
9	40.137090	41.016388	39.308377	60.582676	58.924280	159.665068	134.964061
10	177.643965	179.304201	176.776930	755.734616	746.337793	128.987721	151.429054
11	1287.798682	1269.778758	1262.443395	46.414613	47.025240	13.444971	21.678646
12	21.800549	21.993191	22.862663	110.166452	107.661059	239.295690	207.019707
13	222.801402	222.916767	219.865059	32.527566	30.878819	113.841822	93.499815

Tabella 2.4: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Running

Da tutti i settori si ottiene che i modelli migliori, ovvero quelli che hanno il maggior numero di occorrenze con RMSPE minimo, sono quelli che utilizzano i saldi come variabile esogena.

In questo modo è stato verificato che questa variabile risulta fondamentale per spiegare le vendite nel periodo in esame.

Di conseguenza, in ottica di un'analisi prescrittiva, diventa quindi necessario considerare, soprattutto nei periodi di saldi, una corretta allocazione delle risorse finanziarie e di marketing al fine di massimizzare le vendite.

Anche per questa sezione, per ulteriori analisi, fare riferimento all'apposito paragrafo dell'appendice ([analisi risultati modelli](#)).

3. Conclusioni e sviluppi futuri

Concludendo, in questo lavoro è stata analizzata una serie di dati forniti da un e-commerce.

In primis sono state fatte diverse analisi esplorative e descrittive per capire l'andamento della serie storica; è stato confrontato il trend del fatturato con diversi modelli matematico-statistici e con il trend di alcune variabili esterne al dataset. In particolare l'obiettivo è stato quello di analizzare delle condizioni anomale, in modo da individuare quali fossero le più influenti sull'andamento del fatturato e considerarle come variabili esogene nella fase di predizione.

Successivamente si è passati alla fase di modellazione: per le categorie prese in esame sono stati costruiti diversi modelli (utilizzando ARIMA, ARIMAX e Prophet) che realizzano, con una tecnica rolling, delle previsioni su una finestra temporale di 14 giorni, prevedendo giorno dopo giorno il fatturato totale. I risultati sono visibili nella sezione dedicata ([capitolo 2](#)).

Dai risultati si evidenzia come i modelli che utilizzano i saldi come variabile esogena, soprattutto quelli costruiti con ARIMA(X), sono i migliori, ovvero con RMSPE minore: è importante, quindi, in una logica di business, considerare questa caratteristica dei dati, in quanto i saldi sono particolarmente influenti nell'andamento del fatturato.

L'informazione utile che è stata trovata in questo studio è che nel periodo dei saldi, questi giocano un ruolo fondamentale, hanno un'influenza significativa sull'andamento del fatturato: diventa quindi fondamentale considerare questa variabile. Quindi va studiato il fenomeno tramite una fase di analisi esplorativa-descrittiva, ne va verificata l'influenza tramite un'analisi predittiva e successivamente deve essere considerato all'interno delle strategie di business, in modo da poter massimizzare la produttività, l'efficienza e possibilmente anche i guadagni della piattaforma e-commerce.

Possibili sviluppi futuri per questo lavoro potrebbero consistere nei seguenti punti:

- avendo a disposizione un id per ogni utente a cui è associato un acquisto, si potrebbe costruire un sistema di raccomandazione;
- si potrebbe, se forniti di un maggior numero di dati per le categorie meno numerose, ampliare l'analisi effettuata, esaminando anche le altre categorie presenti;
- se si avessero più informazioni relative agli acquisti (ad esempio un timestamp dell'acquisto) si potrebbe provare a prevedere quale tipologia di prodotto verrà acquistata in un determinato orario di un determinato

giorno (in base a giorno ed orario degli acquisti storici);

- si potrebbe procedere all'aggiunta di ulteriori eventi esogeni (come investimenti/pubblicità oppure condizioni meteo e temperatura) per vedere se questi influiscono o meno nel processo di acquisto (e successivamente nelle previsioni dei modelli);
- si potrebbe, per questo dataset in particolare, provare a fare un forecast per prevedere l'andamento del periodo mancante (metà aprile 2022 - fine 2022).

4. Riferimenti

1. <https://covid19.infn.it/iss/>
2. <https://facebook.github.io/prophet/>
3. Fattore M. (2020). Fundamentals of time series analysis, for the working data scientist (DRAFT).
4. [Analyzing seasonality with Fourier transforms using Python & SciPy](#)
5. <https://robjhyndman.com/hyndsoft/longseasonality/>
6. <https://www.mathworks.com/help/econ/arima-model-including-exogenous-regressors.html>
7. Taylor SJ, Letham B. 2017. Forecasting at scale.
8. <https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/>
9. [http://idosi.org/waj/waj\(ITEMIES\)13/28.pdf](http://idosi.org/waj/waj(ITEMIES)13/28.pdf)
10. <https://www.italiaonline.it/risorse/e-commerce-data-driven-los-e-e-quali-sono-i-vantaggi-3351>

Appendice

Aspetti metodologici

All'interno del progetto sono stati utilizzati, per effettuare le previsioni, due strumenti: ARIMA(X) e Prophet.

Per effettuare il fit dei modelli è stato utilizzato il dataset con granularità giornaliera.

Inizialmente sono stati utilizzati i modelli SARIMA e Prophet.

Studiando e sperimentando con il modello si è capito che SARIMAX non è progettato per supportare grandi frequenze stagionali come quella annuale a granularità giornaliera (365). Infatti per tutte le frequenze maggiori di 200 il programma termina con errore a causa della saturazione della memoria. Anche con un'istanza potente di AWS EC2, in particolare c5.12xlarge con 96 GB di RAM e 48vCPU, non è stato possibile ottenere un risultato. Si è quindi compreso che questo approccio non è l'ideale. Quindi per spiegare la componente stagionale si è fatto ricorso ad un approccio molto più simile a quello utilizzato dal modello di Facebook Prophet, ovvero quello di fare uso delle serie di Fourier.

Tramite le serie di Fourier è infatti possibile modellare il pattern stagionale:

$$y_t = a + \sum_{k=1}^K [\alpha_k \sin(2\pi kt/m) + \beta_k \cos(2\pi kt/m)] + N_t,$$

Figura a.1: Serie di Fourier

dove Nt è ARIMA e K, ovvero il numero di termini, seno e coseno della serie, è scelto nella combinazione che minimizza il valore AIC.

Tuttavia, come vedremo dopo dai risultati, la differenziazione stagionale di ordine elevato non ha molto senso: per i dati giornalieri implicherebbe confrontare ciò che è accaduto oggi con ciò che è accaduto esattamente un anno fa, ma non vi è alcun vincolo che l'andamento stagionale sia regolare. Nonostante ciò non possiamo dire che questo sia valido per tutti i giorni dell'anno: infatti abbiamo festività particolari come Natale, o eventi speciali come il Black Friday, dove sapere ciò che è successo l'anno precedente può rivelarsi utile.

Per questo nella fase dei test del modello prenderemo in considerazioni diverse configurazioni di modelli, proprio per avere una prospettiva completa di quello che potrà succedere nel futuro.

1. ARIMA(X)

Questo modello viene utilizzato per vedere quali siano gli eventi che influenzano le vendite (e quindi il fatturato) e per prevedere il fatturato giornaliero dei settori considerati.

ARIMA: un pilastro della modellazione delle serie storiche.

In ARIMA si hanno 3 componenti:

1. AR: Modellazione AutoRegressiva (modella i valori della serie)
2. I: Integrazione della serie storica (differenza), utile per serie non stazionarie
3. MA: Modellazione Moving Average (modella gli errori)

Un modello ARMA integrato di ordine d è un processo stocastico che diventa stazionario dopo essere stato differenziato due volte. Poiché tutti i processi stazionari possono essere descritti in rappresentazioni

ARMA(p,q), usando i polinomi AR e MA nell'operatore di backward B, si ha che qualsiasi processo integrato obbedisce a un'equazione del tipo:

$$\Phi(B)(I - B)dY_t = \Psi(B)\varepsilon_t$$

Modello ARIMAX

Con ARIMAX l'approccio usato è leggermente diverso: si fa interagire la componente autoregressiva (AR) e coefficienti di regressione, facendo regredire y_t sui suoi valori ritardati e su regressori esogeni; formalmente:

$$\begin{aligned}\Phi(B)y_t &= \beta^T x_t + \Theta(B)\varepsilon_t \Rightarrow \\ y_t &= \beta^T \Phi(B)^{-1}x_t + \Phi(B)^{-1}\Theta(B)\varepsilon_t\end{aligned}$$

Dalla formula precedente si nota che i regressori vengono ritardati dall'inverso del polinomio autoregressivo e dai coefficienti di regressione in entrambi i parametri AR. Come nel caso dell'errore ARMA, i polinomi stagionali possono fare la loro comparsa nelle parti AR e MA e, in caso di non stazionarietà, bisogna differenziare l'input y_t .

Come variabili esogene verranno utilizzate delle serie che identificano il periodo di saldi, il periodo di lockdown e i pattern di stagionalità con Fourier.

2. PROPHET

Prophet è un modello di regressione additiva con quattro componenti principali:

- un trend della curva di crescita logistica lineare a tratti;
- una componente stagionale annuale modellata utilizzando le serie di Fourier;
- una componente stagionale settimanale creata utilizzando variabili fittizie;

- un elenco fornito dall'utente di festività importanti.

Il software è molto utile per le previsioni di business che si riferiscono a osservazioni orarie, giornaliere o settimanali con forti stagionalità multiple. Prophet è progettato anche per affrontare le festività note in anticipo, mentre mancano osservazioni e grandi valori anomali. È progettato per far fronte a serie che subiscono cambiamenti di regime, come il lancio di un prodotto, e che affrontano limiti naturali dovuti alla saturazione del mercato del prodotto.

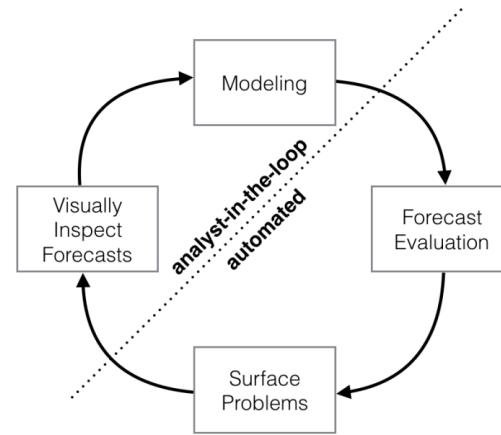


Figura a.2: Forecasting process usato da Prophet

Le tre componenti principali del modello, ovvero trend, stagionalità e festività sono combinate nella seguente equazione:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

dove

1. $g(t)$ rappresenta la funzione di crescita che modella l'andamento generale dei dati (trend);
2. $s(t)$ rappresenta la funzione di stagionalità (che è una serie di Fourier in funzione del tempo);
3. $h(t)$ sono gli effetti delle ferie; questa componente consente a Facebook Prophet di modificare le previsioni

- quando una vacanza o un evento importante può modificare la previsione;
4. ε_t è il termine di errore; tiene conto di eventuali modifiche insolite non soddisfatte dal modello.

La funzione di crescita ha tre sviluppi principali:

- crescita lineare: utilizza un insieme di equazioni lineari a tratti con diverse pendenze tra i punti di cambio; quando viene utilizzata la crescita lineare, il termine di crescita sarà simile alla forma
 $y = mx + b$
 tranne per il fatto che la pendenza(m) e l'offset(b) sono variabili e cambieranno valore ad ogni punto di cambio.
- crescita logistica: usata quando le serie temporali hanno un limite o un livello minimo che i valori non possono superare. Quando viene utilizzata la crescita logistica, il termine di crescita sarà simile a una tipica equazione per una curva logistica, tranne per il fatto che la capacità di carico (C) varierà in funzione del tempo, mentre il tasso di crescita (k) e l'offset (m) sono variabili e cambieranno valore ad ogni punto di cambiamento.

$$g(t) = \frac{C(t)}{1 + x^{-k(t-m)}}$$

Figura a.3: Equazione che esprime la crescita logistica

- flat: assume un valore costante e viene usata quando non c'è crescita nel tempo (ma potrebbe esserci ancora stagionalità).

La funzione di stagionalità

La funzione di stagionalità è semplicemente una serie di Fourier in funzione del tempo.

La trasformata di Fourier permette di modellare una funzione del tempo e del segnale in una funzione della frequenza e della potenza. Questo mostra quali frequenze compongono il tuo segnale e quanto sono forti.

Nel nostro caso, il segnale è il fatturato giornaliero di un determinato settore e potremmo aspettarci una sorta di frequenza settimanale, mensile o annuale.

$$s(t) = \sum_{n=1}^N (a_n \cos(\frac{2\pi nt}{P}) + b_n \sin(\frac{2\pi nt}{P}))$$

Figura a.4: Funzione di stagionalità

Questa somma può approssimare quasi tutte le curve; nel caso di Facebook Prophet approssima la stagionalità, se presente all'interno dei dati.

La funzione vacanza

La funzione vacanze consente a Facebook Prophet di modificare le previsioni quando una vacanza o un evento importante può influenzare il fenomeno. Prende in input un elenco

di date (ci sono le date delle holidays negli Stati Uniti incorporate o possono essere definite delle date personalizzate); quando ogni data è presente nella previsione, viene aggiunto o sottratto il valore della previsione dai termini di crescita e stagionalità in base ai dati storici nelle date individuate.

Usando il tempo come regressore, Prophet cerca di adattare diverse funzioni lineari e non lineari del tempo come componenti; la modellazione della stagionalità come componente additiva è lo stesso approccio adottato dallo smoothing esponenziale nella tecnica Holt-Winters.

3. ADF

Nell'analisi delle serie storiche è stata anche verificata la stazionarietà delle stesse: per fare ciò è stato utilizzato il Augmented Dickey Fuller Test.

Questo test è una versione "aumentata" del test di Dickey-Fuller, un test di radice unitaria che testa l'ipotesi nulla che α sia uguale a 0 nella seguente equazione:

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e_t$$

Figura a.5: Equazione Dickey-Fuller

dove:

- $y(t-1)$ rappresenta il ritardo di ordine 1 della serie temporale;
- $\Delta Y(t-1)$ rappresenta la prima differenza di tempo della serie (t-1).

Il test ADF espande l'equazione del test di Dickey-Fuller per includere il processo regressivo di ordine elevato nel modello.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} \dots + \phi_p \Delta Y_{t-p} + e_t$$

Figura a.6: Equazione augmented Dickey-Fuller

L'ipotesi nulla è la stessa del test di Dickey Fuller.

Dal momento che l'ipotesi nulla presuppone il fatto che $\alpha=1$ (quindi presuppone la presenza di una radice unitaria), il p-value dovrebbe essere inferiore rispetto al livello di significatività scelto (solitamente 0.05) per rigettare l'ipotesi nulla; per non rifiutare l'ipotesi nulla di stazionarietà quindi il valore del p-value deve risultare minore del livello di significatività.

Preprocessing

Fase 1

Importando il dataset sono stati riscontrati alcuni problemi, principalmente di formattazione: sono state effettuate quindi delle operazioni per avere un formato dei dati adatto all'utilizzo.

Sono stati poi creati dei data frame singoli, ognuno riferito ad una determinata categoria/settore di vendita.

Sono state definite poi delle funzioni per la creazione di data frame che consentissero di avere i data frame relativi alla categoria con il totale aggregato per anno o per mese (somma cumulata).

Fase 2

Nella fase 2 sono stati integrati all'interno dei data frame delle categorie i dati relativi al Covid-19: è stata inserita una colonna, chiamata `rt_positivi`, in cui sono contenuti i valori dell'indice Rt a partire dal giorno 2020-02-23 fino all'ultimo giorno disponibile

nel dataset; è stata inserita un'ulteriore colonna, chiamata lockdown, che contiene valori binari: 1, se nel giorno considerato la popolazione si trovava in lockdown e/o zona rossa; 0 altrimenti.

In seguito sono state raccolte e salvate le date dei periodi dei saldi, distinguendo fra saldi invernali, saldi estivi e saldi del periodo Black Friday + Natale. Queste date sono state utilizzate poi per la costruzione dei modelli.

Analisi esplorativa

Il dataset è composto da 30 settori, ognuno contenente un numero diverso di istanze. I dati sono stati resi disponibili con granularità giornaliera.

settore	totale								
	count	mean	std	min	25%	50%	75%	max	
Arceria	7.0	519.330000	479.763142	71.23	143.6800	569.820	628.9000	1449.10	
Arti marziali	139.0	760.717194	711.336248	122.81	243.1500	670.520	1104.0150	6432.53	
Bambino	1250.0	2083.043672	2172.618008	81.87	762.2100	1361.770	2676.4675	25858.09	
Baseball	60.0	463.274500	264.859403	161.01	365.9600	379.330	399.7975	1844.80	
Basket	922.0	1793.860390	1409.098156	120.08	787.3150	1471.200	2307.7075	12255.94	
Buoni / contatti	7.0	2729.000000	1929.694406	1364.50	1364.5000	2729.000	2729.0000	6822.50	
Calci	2956.0	13215.077498	13910.280138	106.43	5171.7975	9407.970	15947.5250	171365.37	
Casual	2901.0	12226.183633	9169.902835	-2159.18	5642.2100	10252.030	16419.3000	98255.46	
Ciclismo	759.0	978.939789	1416.752589	54.58	270.7200	517.960	1078.7700	13372.10	
Danza	53.0	512.007547	556.998460	109.16	316.8400	352.040	486.3100	3520.41	
Fitness	2834.0	8887.481138	27940.749429	45.44	2341.6200	4328.330	7727.8275	516104.93	
Freccette	10.0	1633.632000	924.861542	734.10	1225.5900	2144.7900	3676.78		
Golf	31.0	2401.829677	2099.746604	212.86	733.1450	2033.110	3905.2000	9349.55	
Intime	4.0	1487.305000	620.909974	955.15	955.1500	1432.725	1964.8800	2128.62	
Mare	1009.0	1799.722131	1705.351755	135.09	685.8000	1220.410	2330.8400	13740.79	
Nuoto	949.0	777.835985	608.102485	57.04	368.4200	633.670	979.9800	5185.10	
Padel	191.0	5190.539948	4013.113861	136.45	2999.1700	4584.720	6860.7050	33070.02	
Pattini	154.0	1286.765130	1058.878659	242.88	408.7350	1225.5950	1654.7925	5887.27	
Pesca	2978.0	18494.435480	11543.885015	36.84	10070.4900	16851.300	25013.0575	84700.25	
Ping-pong	105.0	718.353429	810.160836	54.31	188.3000	485.220	971.5200	5703.61	
Rugby	73.0	1201.385783	710.748701	169.47	757.3000	1226.690	1500.9500	3745.55	
Running	2271.0	5040.842554	4261.376770	0.00	2213.9000	3891.550	6787.4300	55733.00	
Sci	1209.0	7842.052672	8391.974503	2.73	1760.2100	5387.050	10902.3600	58780.48	
Skateboard	94.0	1275.693085	993.344902	98.24	611.5700	917.490	1606.9025	5654.49	
Snowboard	1372.0	4368.104964	5006.850072	95.52	1134.1700	2469.060	5681.0950	50980.72	
Soft air	1.0	136.450000	NaN	136.45	136.4500	136.450	136.4500	136.45	
Subacquea	162.0	1204.627901	1833.318167	147.37	488.7600	712.270	1225.5900	17192.70	
Tennis	972.0	2039.352294	2054.405058	36.84	614.0300	1397.930	2863.4025	15268.76	
Trekking	1514.0	3764.138137	3661.227758	121.44	1335.3000	2524.870	4947.2025	28574.54	
Volley	274.0	1142.115292	887.329796	215.59	486.3100	1075.225	1526.0600	5865.17	

Figura a.7: Descrizione del dataset

Nell' immagine a.7 si riscontra che i settori con il maggior numero di occorrenze (più di 2000) sono:

- **Calcio:** 2956 occorrenze, con valori del totale compresi tra [106.34, 171365.37];
- **Casual:** 2901 occorrenze, con valori del totale compresi tra [-2159.18, 98255.46]. Il valore è negativo in quanto in quel giorno è stato fatto un reso; questo valore è stato rimosso nella fase di pre-processing;
- **Fitness:** 2834 occorrenze, con valori del totale compresi tra [45.44 , 516104.93];
- **Pesca:** 2978 occorrenze , con valori del totale compresi tra [36.84 , 84700.25];
- **Running:** 2271 occorrenze , con valori del totale compresi tra [0.00 , 55733.00]

Essendo queste categorie le più numerose sono state scelte sia per l'analisi esplorativa sia per la previsione, in quanto hanno un numero di occorrenze minimo adeguato per poter eseguire degli studi.

Sono stati utilizzati alcuni modelli per confrontare il loro andamento con quello della serie storica (con i dati, per questa sezione, *raggruppati settimanalmente*):

- Simple Moving Average: rappresenta l'andamento della media non ponderata dei valori dei precedenti K dati, dove K si riferisce al periodo di tempo di rilevazione di questi dati; SMA presenta alcuni punti deboli:
 - finestre di tempo piccole portano a più rumore, quindi a performance peggiori;
 - difficilmente raggiunge i punti di picco o di minimo dei dati a causa dell'uso della media;

- non informa sui possibili comportamenti futuri della serie di dati, tutto ciò che fa è descriverne la tendenza.
- **Exponentially Weighted Moving Average:** EWMA consente di ridurre l'effetto lag di SMA e dà più peso ai valori che si sono verificati più di recente. La quantità di peso applicata ai valori più recenti dipenderà dai parametri effettivi utilizzati nell'EWMA.
- **Double Exponential Smoothing:** Double Exponential Smoothing aggiunge un secondo fattore di smoothing beta (rispetto al simple exponential smoothing) che affronta le tendenze dei dati; il vantaggio è che il modello può anticipare incrementi o diminuzioni futuri laddove il Simple Exponential Smoothing funzionerebbe solo in base a calcoli recenti; il Simple Exponential Smoothing non va bene quando c'è una tendenza nei dati. Ci possono essere diversi tipi di cambiamento (crescita e/o decadimento) nel trend. Se una serie temporale visualizza un andamento inclinato in linea retta, si utilizzerà un aggiustamento *additivo*. Se la serie storica mostra un andamento esponenziale (curvo), si utilizzerà un aggiustamento *moltiplicativo*.
- **Triple Exponential Smoothing:** il Triple Exponential Smoothing aggiunge supporto nei dati sia per i trend che per la stagionalità. Il Triple Exponential Smoothing applica il livellamento esponenziale tre volte; viene comunemente utilizzato quando ci sono tre segnali ad alta frequenza da rimuovere da una serie

temporale in studio. Esistono diversi tipi di stagionalità: principalmente vengono usati i Triple Exponential Smoothing di natura moltiplicativa o additiva.

1. Calcio

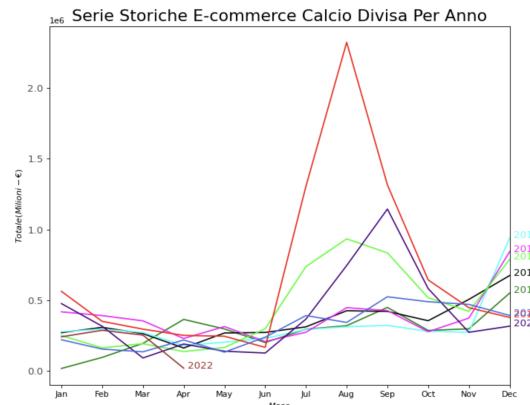


Figura a.8: Serie storica del settore Calcio divisa per anno

In primis si è studiato l'andamento mensile dei fatturati nei vari anni di interesse. Come si può vedere nell'immagine a.8 il settore del calcio ha un grosso incremento di vendite tra il mese di Agosto e Settembre, specialmente nell'anno del 2019.

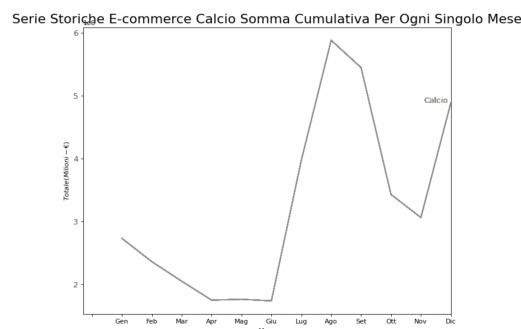


Figura a.9: Serie storica del settore Calcio sommando i mesi di ogni anno

Successivamente, si è fatta una somma cumulativa di ogni mese dei singoli anni per capire qual è il mese che storicamente ha un fatturato maggiore per la categoria in questione. Come si vede nella foto a.9 i

mesi con più vendite sono agosto e settembre.

In seguito, si è posta l'attenzione sullo studio dell'autocorrelazione tra i valori della serie storica. I risultati sono i seguenti:

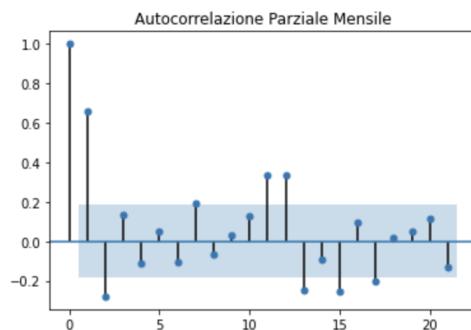


Figura a.10: Autocorrelazione parziale mensile del settore Calcio

Dalla figura a.10, si evince che la serie storica è caratterizzata da una componente stagionale molto forte (come si vede ai lag 1, 2, 11, 12, 13, 16). Sembra che possa esistere una stagionalità annuale all'interno della serie storica dei dati.

Si è proceduto poi con un'analisi dell'andamento del fatturato confrontato con l'andamento del valore dell'indice Rt (indice relativo al contagio del CoronaVirus). Questo poiché si pensava che ci potesse essere una correlazione tra questi due fattori.

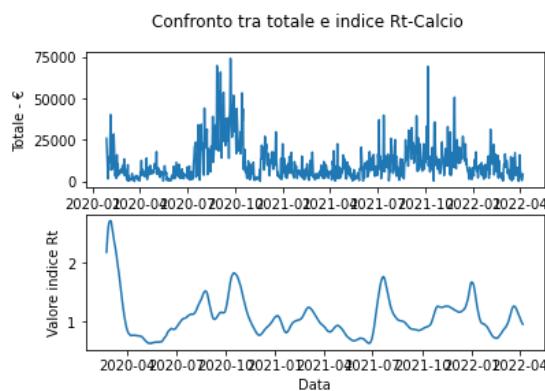


Figura a.11: Confronto tra il totale del settore Calcio e l'indice Rt del Covid

Il periodo considerato va da gennaio 2020 fino a fine aprile/inizio maggio 2022; gli unici due picchi in cui indice Rt e fatturato totale sembrano muoversi nella stessa direzione sono ad inizio 2020 e intorno a settembre-novembre 2020. Nel resto dei periodi sembrano avere un andamento indipendente l'uno dall'altro; la lettura del grafico non risulta facilissima, in quanto i dati sono stati raggruppati giornalmente e, soprattutto quelli del fatturato, soffrono di forti oscillazioni tra un giorno ed un altro.

È stata poi costruita la matrice di correlazione tra il fatturato totale e l'indice Rt per il periodo riguardante il primo mese di lockdown circa (dal 9/3/2020 (inizio lockdown) al 31/3/2020).

	totale	rt_positivi
totale	1.000000	0.215082
rt_positivi	0.215082	1.000000

Figura a.12: Matrice di correlazione tra il fatturato totale di Calcio e l'andamento dell'indice Rt (periodo dal 2020-03-09 al 2020-03-31)

Si può vedere nella matrice di correlazione che l'andamento dei contagi è correlato positivamente con l'andamento del fatturato, ma non influenza particolarmente sui guadagni del settore calcio (il coefficiente di correlazione è pari a 0.21).

Si è passati dunque ad analizzare l'andamento del fatturato totale confrontandolo con degli approcci matematico-statistici per capire di più su stagionalità e trend della serie storica in questione.

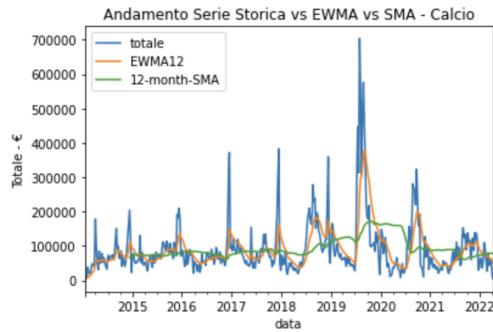


Figura a.13: Andamento serie storica del settore Calcio con l'EWMA e SMA

L'exponential moving average riesce a seguire meglio l'andamento della variabile totale rispetto al simple moving average poiché riesce a dare un peso maggiore ai valori più recenti e riesce a ridurre l'effetto lag dei simple moving average.

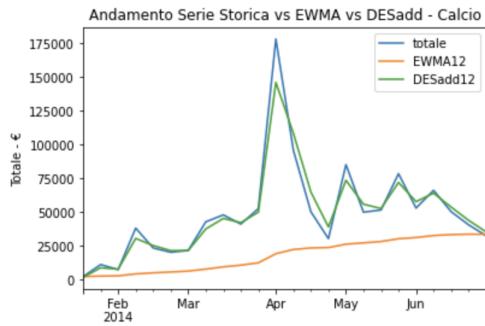


Figura a.14: Andamento serie storica del settore Calcio con l'EWMA12 e DESadd12

In questo grafico si può notare quanto più efficiente sia il double exponential smoothing rispetto all'exponential moving average; questo è dovuto al fatto che il DES può anticipare incrementi o diminuzioni futuri laddove il EWMA funzionerebbe solo in base a calcoli recenti. Si può supporre che ci sia un trend nella serie storica del fatturato.

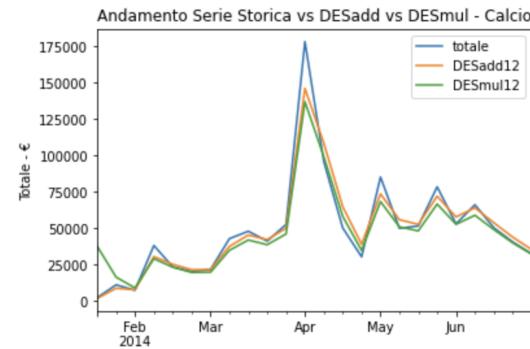


Figura a.15: Andamento serie storica del settore Calcio con DESadd12 e DESmul12

Questo grafico evidenzia che, per il Double Exponential Smoothing, sarebbe preferibile un aggiustamento additivo rispetto a uno moltiplicativo.

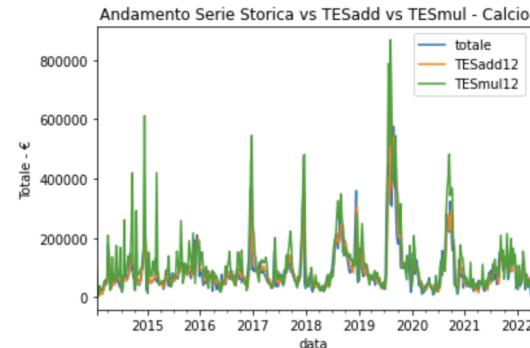


Figura a.16: Andamento serie storica del settore Calcio con TESadd12 e TESmul12

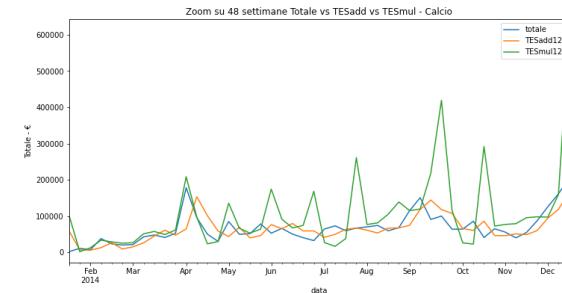


Figura a.17: Zoom dell'andamento serie storica del settore Calcio con TESadd12 e TESmul12

Il triple exponential smoothing sembra descrivere accuratamente l'andamento;

anche in questo caso è preferibile l'utilizzo di un modello additivo.

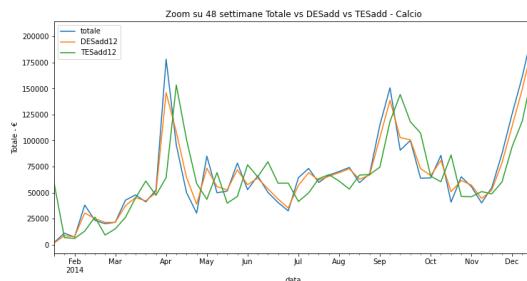


Figura a.18: Zoom dell'andamento serie storica del settore Calcio con DESadd12 e TESadd12

Il triple exponential smoothing sembra descrivere accuratamente l'andamento, ma è come se lo descrivesse con due settimane di lag; si afferma che è più influente la componente di trend piuttosto che quella di stagionalità, ma anche questa sembra essere presente nei dati.

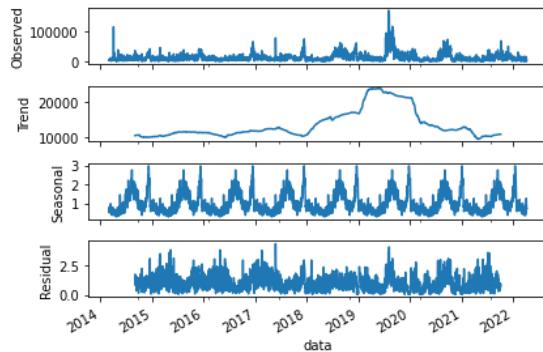


Figura a.19: Andamento della Stagionalità, Trend, Osservazioni e Residui

Il trend è sempre positivo fino al picco del 2019; dopo questo picco il trend decresce sempre più fino a riprendersi nei primi mesi del 2021.

Viene sottolineata un'importante stagionalità nei mesi di agosto e settembre e nel mese di dicembre, mesi dove si verificano sempre dei picchi positivi. Sembra che la

stagionalità influisca sul fatturato, ma a causa della granularità giornaliera del grafico non è del tutto chiaro.

2. Casual

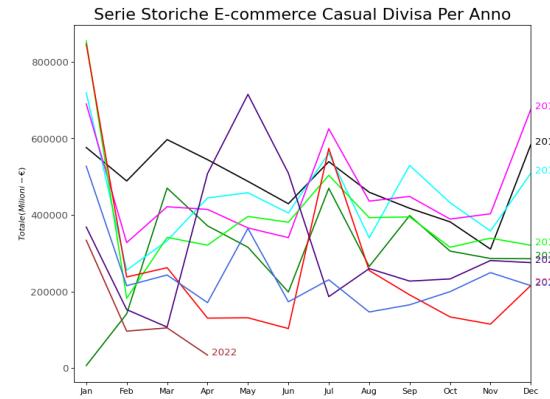


Figura a.20: Serie storica del settore Calcio divisa per anno

Come si può vedere nella figura a.20 nel mese di Gennaio del 2016, 2017 e 2019 vi è un incremento esponenziale, ma durante l'anno tutti i valori decrescono senza mai raggiungere i valori di picco iniziali.

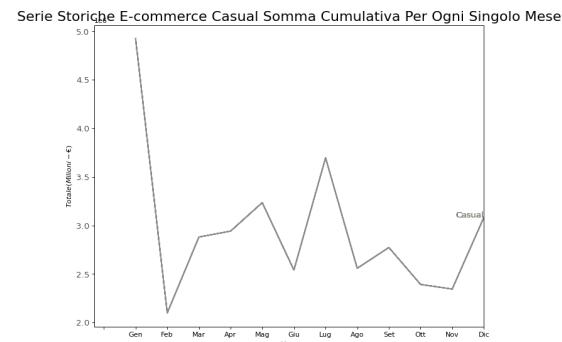


Figura a.21: Serie storica del settore Casual sommando i mesi di ogni anno

Collegandosi alla figura precedente si può affermare, guardando l'immagine a.21, che il mese in cui nel settore Casual si fattura di più è ampiamente quello di Gennaio.

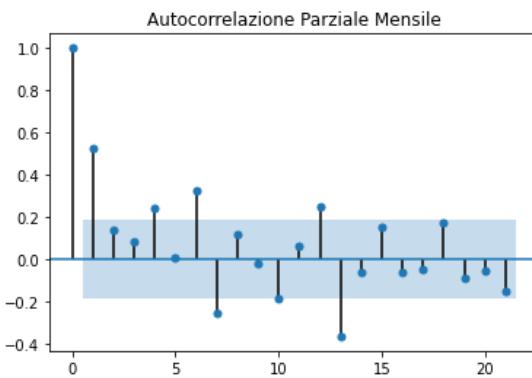


Figura a.22: Autocorrelazione parziale mensile del settore Casual

Esiste una forte correlazione positiva al lag 1; negativa ai lag 7 e 13.

Procedendo con l'analisi si è osservato l'andamento del fatturato totale della categoria Casual rispetto all'andamento storico dell'indice Rt.

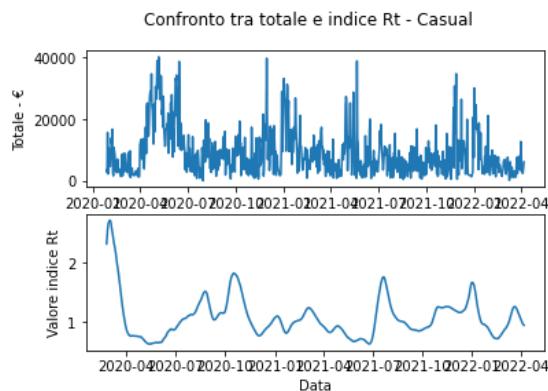


Figura a.23: Confronto tra il totale del settore Casual e l'indice Rt del Covid

Dal grafico si capisce come nel periodo iniziale del primo lockdown le persone non hanno comprato nel settore Casual; nel periodo compreso tra metà maggio e giugno, quando il lockdown era terminato, si vede un netto aumento delle vendite a fronte di un periodo in cui l'indice Rt assumeva valori decisamente bassi (i minori dell'anno 2020).

Con l'avanzare della pandemia non si può dire granché sulla correlazione tra i due andamenti (che sembrano spesso simili).

	totale	rt_positivi
totale	1.00000	0.51494
rt_positivi	0.51494	1.00000

Figura a.24: Matrice di correlazione tra il fatturato totale di Casual e l'andamento dell'indice Rt (periodo dal 2020-03-09 al 2020-03-31)

Nella matrice di correlazione si nota che il valore dell'Rt e il fatturato del settore Casual sono correlati positivamente, con indice pari a 0.5, quindi vi è una discreta correlazione positiva.

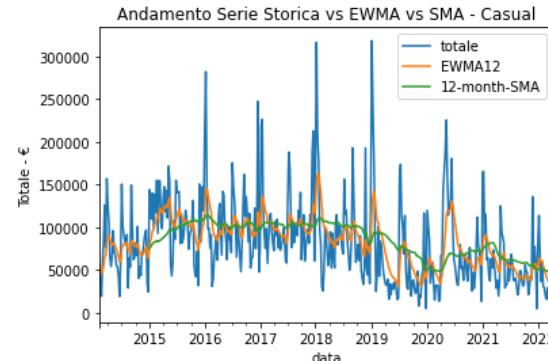


Figura a.25: Andamento serie storica del settore Casual con l'EWMA e SMA

Sia l'exponential moving average che il simple moving average faticano a raggiungere i picchi, ma tra i due approcci il primo è preferibile.

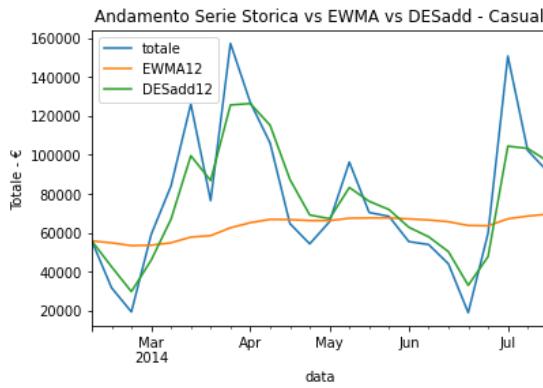


Figura a.26: Andamento serie storica del settore Casual con l'EWMA12 e DESadd12

Anche in questo caso (come nel settore calcio) il DES additivo è preferibile rispetto all'EWMA. Si può affermare che è probabile che ci sia un trend nella serie storica considerata.

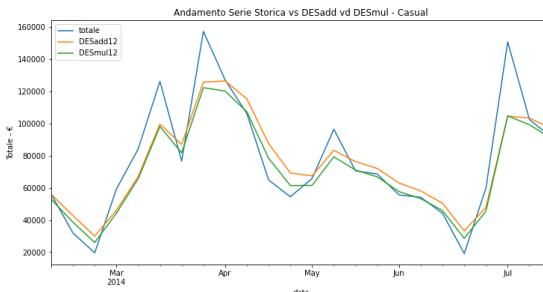


Figura a.27: Andamento serie storica del settore Casual con DESadd12 e DESmul12

L'andamento del DES additivo è molto simile rispetto a quello del DES moltiplicativo; entrambi spiegano bene l'andamento della variabile totale.

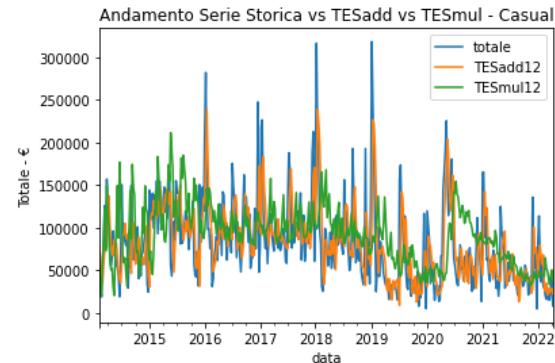


Figura a.28: Andamento serie storica del settore Casual con TESadd12 e TESmul12

Il triple exponential smoothing sembra descrivere accuratamente l'andamento nei punti centrali; è più difficoltoso, per questo modello, descrivere i picchi in tutta la loro ampiezza.

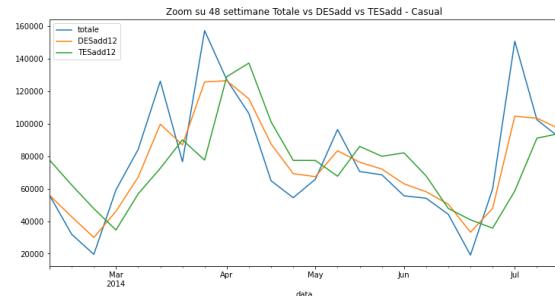


Figura a.29: Zoom dell'andamento serie storica del settore Casual con DESadd12 e TESadd12

Il triple exponential smoothing descrive abbastanza accuratamente l'andamento del fatturato, ma sembra traslato di circa due settimane in avanti. Dunque la componente di trend è più influente della componente stagionale che però è una componente integrante dei dati.

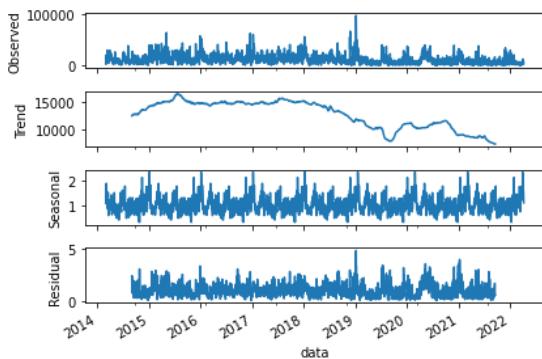


Figura a.30: Andamento della Stagionalità, Trend, Osservazioni e Residui

Il grafico della stagionalità è poco esplicativo e non viene evidenziata una particolare stagionalità; il trend è pressoché stabile fino al 2018, dopo tale anno inizia a calare fino a raggiungere il punto di minimo a fine 2019 senza più raggiungere i livelli dei primi anni.

3. Fitness

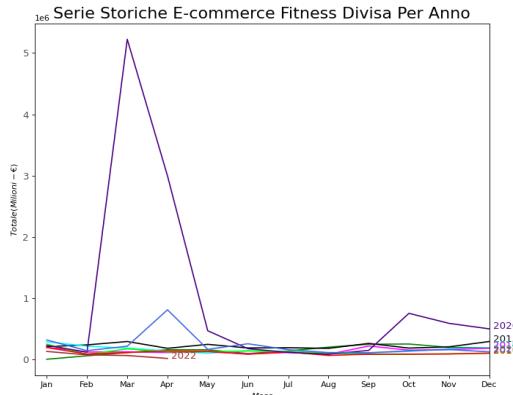


Figura a.31: Serie storica del settore Fitness divisa per anno

Per quanto riguarda il settore Fitness, si può vedere nella foto a.31, che l'andamento negli anni, rapportato alla scala assunta in questo grafico, è quasi sempre omogeneo; a Marzo 2020 si registra un picco positivo che va praticamente a quintuplicare il valore massimo che era stato registrato storicamente. Questo picco non permette

un'analisi specifica di questo settore, in quanto introduce un bias elevatissimo rispetto all'andamento storico medio.

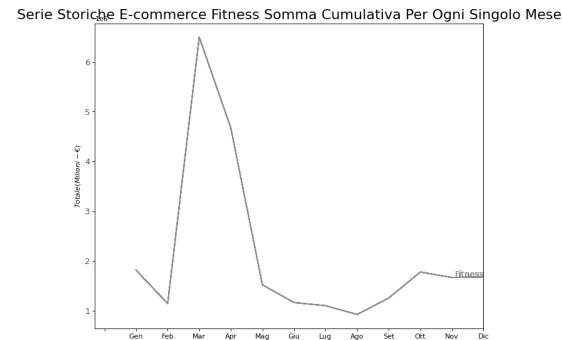


Figura a.32: Serie storica del settore Fitness sommando i mesi di ogni anno

Anche nel grafico a.32 si evidenzia che il mese che cumulativamente e storicamente registra il maggiore fatturato è Marzo.

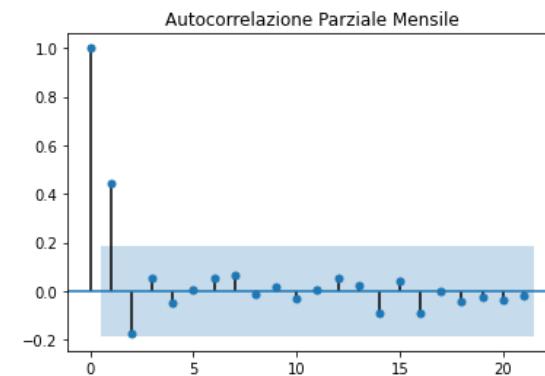


Figura a.33: Autocorrelazione parziale mensile del settore Fitness

C'è una forte componente di autocorrelazione parziale al lag 1.

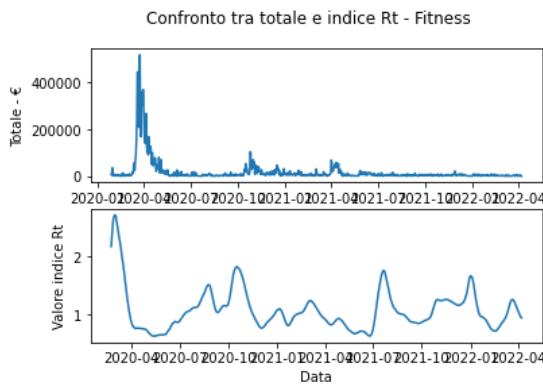


Figura a.34: Confronto tra il totale del settore Fitness e l'indice Rt del Covid

Dal grafico emerge che in vari periodi al diminuire dei contagi cresce la vendita di oggetti della categoria Fitness. Questo fenomeno è interpretabile in questo modo: una volta iniziato il lockdown, essendo la popolazione costretta a rimanere in casa, il numero di contagi è andato in calo (inevitabilmente); di contro il fatturato della categoria fitness è andato in crescita, poiché le persone, non potendo accedere, ad esempio, alle palestre, hanno iniziato ad attrezzarsi per disporre di propri attrezzi personali per tenersi in forma.

	totale	rt_positivi
totale	1.00000	-0.72963
rt_positivi	-0.72963	1.00000

Figura a.35: Matrice di correlazione tra il fatturato totale di Fitness e l'andamento dell'indice Rt (periodo dal 2020-03-09 al 2020-03-31)

La matrice di correlazione evidenzia che l'indice rt e il fatturato del settore Fitness sono correlati negativamente e quindi in generale al crescere corrisponde una diminuzione dell'altro, e viceversa. Questo conferma quanto evidenziato nel grafico precedente: nel periodo in esame (marzo 2020) il coefficiente di correlazione

di Pearson vale -0.73, ribadendo che nel lockdown, mentre il numero dei casi diminuiva, il fatturato della categoria Fitness aumentava.

Non sono riportate le analisi sui modelli statistici (EWMA, DES, TES) in quanto il bias da cui i dati di questo settore sono affetti è molto elevato e porta a dei risultati poco esplicativi.

4. Pesca

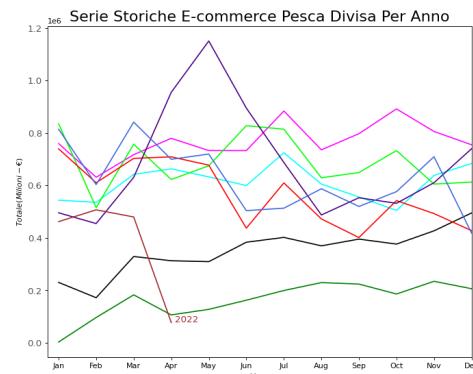


Figura a.36: Serie storica del settore Pesca divisa per anno

In questo grafico viene evidenziato come nel 2020 ci sia stato il picco massimo di vendite per il settore pesca; tendenzialmente le vendite durante gli anni restano più o meno le stesse, tranne per il 2014 e 2015 che sono in netta salita e per il 2022 che stava registrando un picco negativo di fatturato.

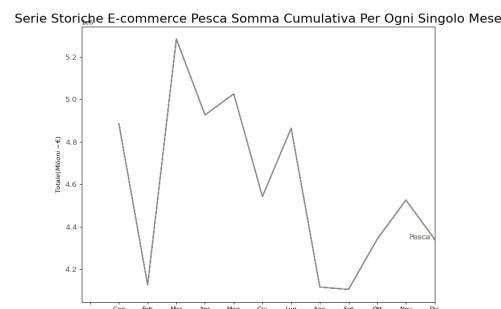


Figura a.37: Serie storica del settore Pesca sommando i mesi di ogni anno

Da questo grafico si nota come i mesi in cui la somma cumulativa dei fatturati è minore sono Febbraio, Agosto e Settembre: nel periodo 2013-2022 sono questi i mesi in cui si è venduto di meno. Il periodo in cui, in questa finestra temporale, si è registrato un fatturato maggiore, invece, è Marzo.

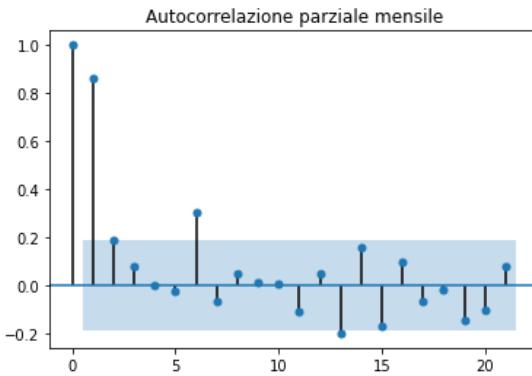


Figura a.38: Autocorrelazione parziale mensile del settore Pesca

Esiste una forte correlazione al lag 1 e al lag 6.

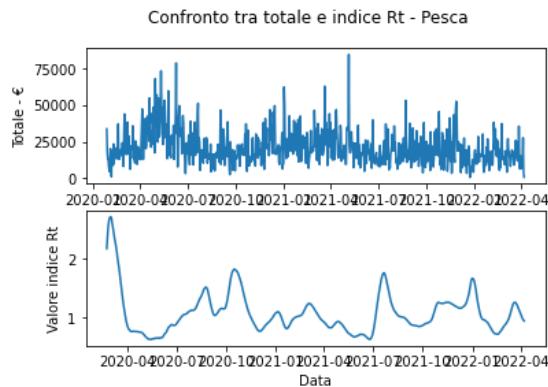


Figura a.39: Confronto tra il totale del settore Pesca e l'indice Rt del Covid

Da questo grafico si nota come nel primo trimestre del lockdown ci sia un andamento opposto tra fatturato e indice Rt: il primo cresce abbastanza, mentre il secondo è in decrescita (a causa, appunto, del lockdown).

	totale	rt_positivi
totale	1.00000	0.15589
rt_positivi	0.15589	1.00000

Figura a.40: Matrice di correlazione tra il fatturato totale di Pesca e l'andamento dell'indice Rt periodo dal 2020-03-09 al 2020-03-31)

Si può vedere nella matrice di correlazione che l'andamento dei contagi è correlato positivamente con l'andamento del fatturato, ma non influisce particolarmente sui guadagni del settore pesca (il coefficiente di correlazione è pari a 0.155).

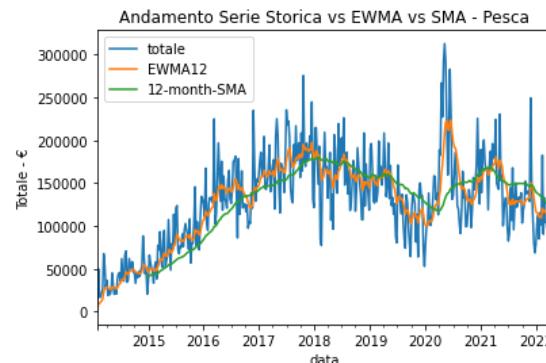


Figura a.41: Andamento serie storica del settore Pesca con l'EWMA e SMA

Anche in questo settore, così come per gli altri, EWMA riesce a seguire meglio l'andamento del totale, nonostante sia molto meno preciso nei punti di picco.

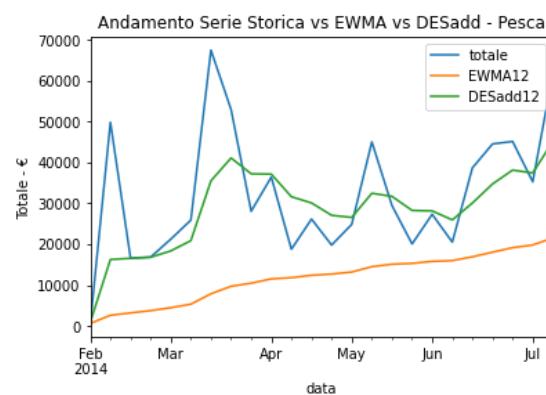


Figura a.42: Andamento serie storica del settore Pesca con l'EWMA12 e DESadd12

Confrontandolo con un altro tipo di modello statistico però l'EWMA risulta decisamente meno preciso; a supporto di ciò si può vedere come il double exponential smoothing riesca a descrivere in maniera più accurata l'andamento dei dati (visibile nella figura a.41).

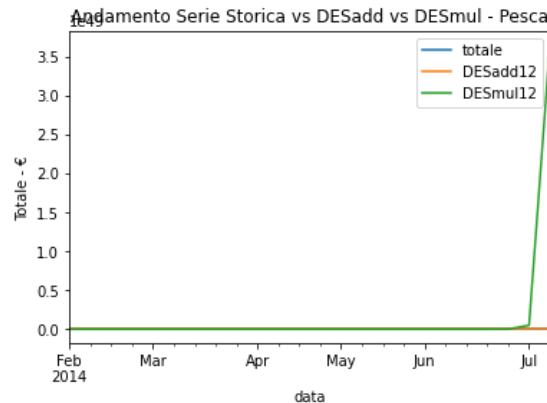


Figura a.43: Andamento serie storica del settore Pesca con DESadd12 e DESmul12

Il DES moltiplicativo è esponenzialmente più grande del DES additivo, quindi è più adatto scegliere un modello di tipo additivo.

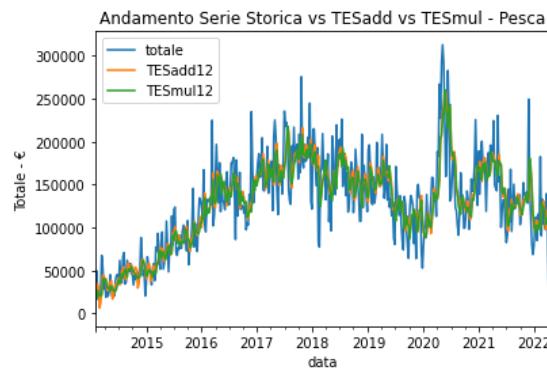


Figura a.44: Andamento serie storica del settore Pesca con TESadd12 e TESmul12

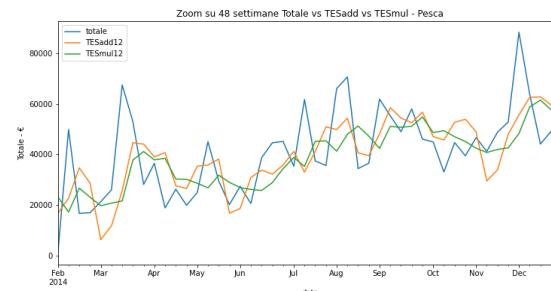


Figura a.45: Zoom dell'andamento serie storica del settore Pesca con TESadd12 e TESmul12

Il modello triple exponential smoothing additivo accentua in maniera più marcata i picchi rispetto al TES moltiplicativo che segue più accuratamente l'andamento del fatturato totale; è preferibile utilizzare un modello TES moltiplicativo.

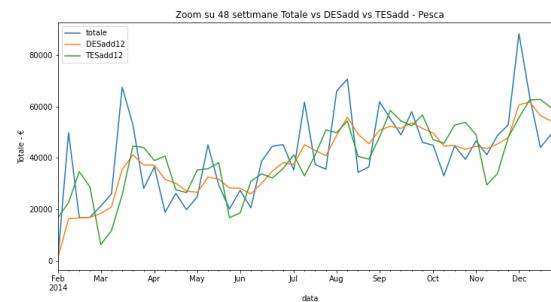


Figura a.46: Zoom dell'andamento serie storica del settore Pesca con DESadd12 e TESadd12

Il triple exponential smoothing sembra descrivere abbastanza accuratamente l'andamento. Sembra essere presente nei dati una componente di stagionalità.

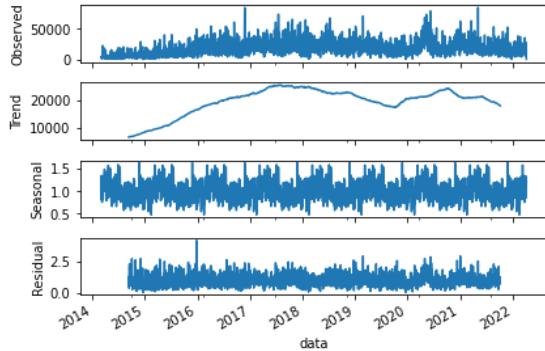


Figura a.47: Andamento della Stagionalità, Trend, Osservazioni e Residui

Il trend del settore pesca è pressoché sempre crescente (ha un piccolo calo nel periodo del primo lockdown).

Non si può dire molto sulla stagionalità dei dati a causa della granularità giornaliera che risulta di difficile interpretazione a causa dell'alta volatilità dei dati.

5. Running



Figura a.48: Serie storica del settore Running divisa per anno

Il fatturato nei diversi anni è molto variabile, ma è pressoché sempre compreso tra i 50000 e i 250000; il 2022 nel complesso sembra essere l'anno con il fatturato minore in assoluto.

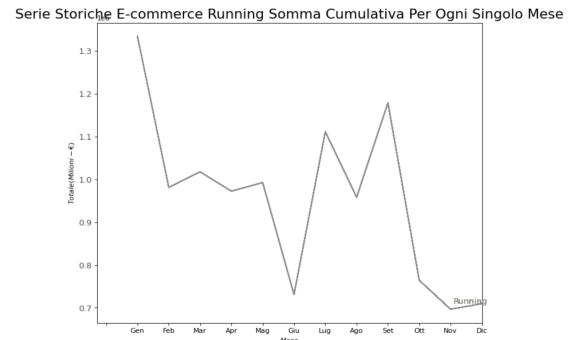


Figura a.49: Serie storica del settore Running sommando i mesi di ogni anno

Nella foto a.49 si può notare come i mesi in cui i prodotti di Running vengono venduti di più sono Gennaio e Settembre, ma si può notare che anche a Luglio è presente un picco positivo influenzato dall'anno 2015 e 2017.

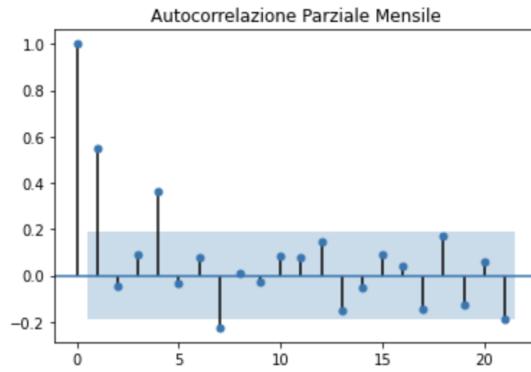


Figura a.50: Autocorrelazione parziale mensile del settore Running

Esiste una forte autocorrelazione positiva al lag 1, al lag 4 e negativa al lag 7.

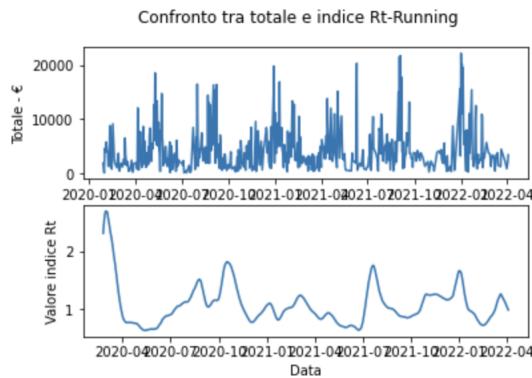


Figura a.51: Confronto tra il totale del settore Running e l'indice Rt del Covid

Il fatturato e l'indice Rt sono correlati positivamente nel periodo di ottobre 2020 (secondo picco nel grafico dell'Rt); nel complesso i due andamenti sono simili.

	totale	rt_positivi
totale	1.000000	0.559365
rt_positivi	0.559365	1.000000

Figura a.52: Matrice di correlazione tra il fatturato totale di Running e l'andamento dell'indice Rt

Si può vedere nella matrice di correlazione che la percentuale di rt e il fatturato del settore di Running sono correlati positivamente, con indice pari a 0.5, quindi vi è una discreta correlazione.

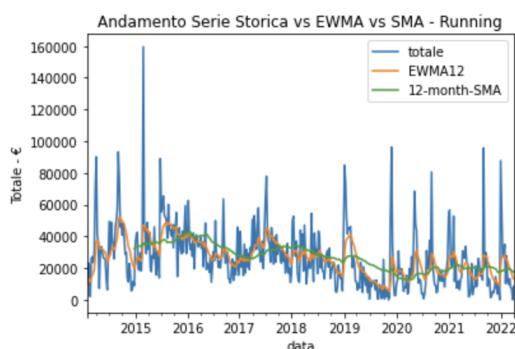


Figura a.53: Andamento serie storica del settore Running con l'EWMA e SMA

EWMA è il modello che spiega meglio l'andamento della variabile totale, nonostante non riesca a spiegare tutta l'ampiezza dei picchi.

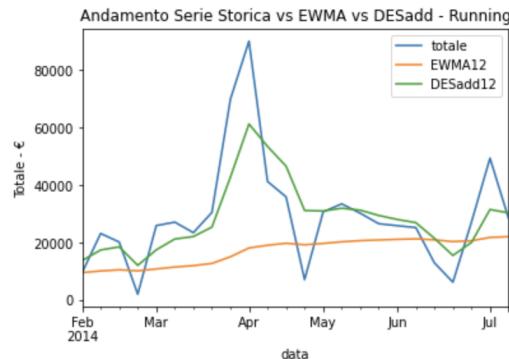


Figura a.54: Andamento serie storica del settore Running con l'EWMA12 e DESadd12

Il double exponential smoothing continua a essere un modello più accurato rispetto all'EWMA così come per gli altri settori. Sembra, dunque, che ci sia trend nei dati.

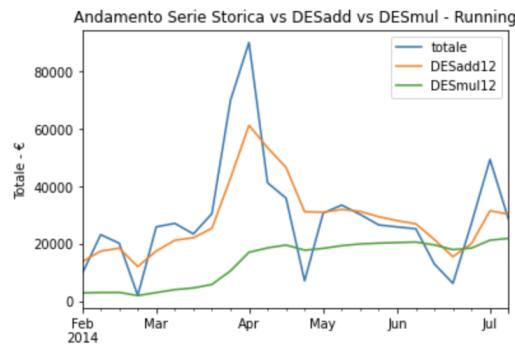


Figura a.55: Andamento serie storica del settore Running con DESadd12 e DESmul12

Per il settore Running il double exponential smoothing di tipo additivo segue meglio l'andamento della serie storica del totale; è quindi preferibile rispetto al DES moltiplicativo.

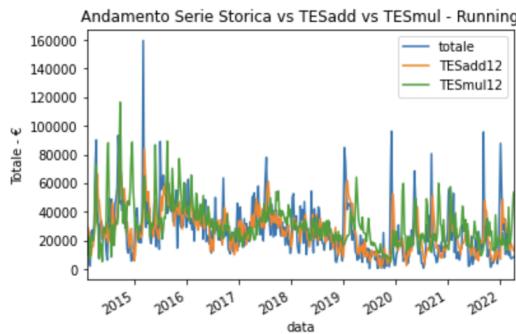


Figura a.56: Andamento serie storica del settore Running con TESadd12 e TESmul12

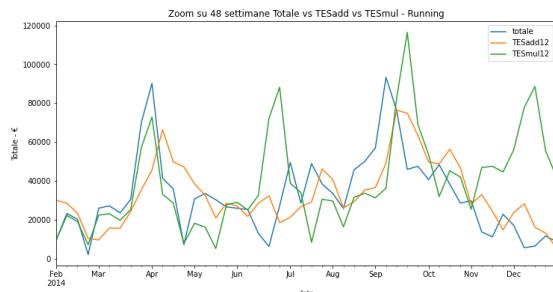


Figura a.57: Zoom dell'andamento serie storica del settore Running con TESadd12 e TESmul12

Così come per il DES anche per il triple exponential smoothing è preferibile un modello di tipo additivo che riesce a spiegare più precisamente l'andamento del fatturato.

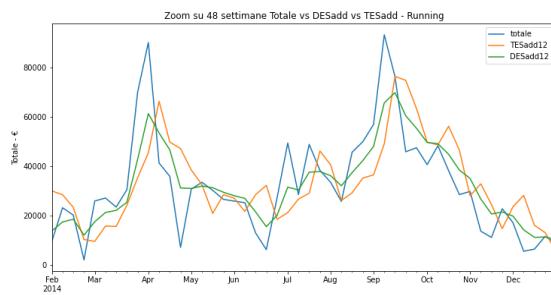


Figura a.58: Zoom dell'andamento serie storica del settore Running con DESadd12 e TESadd12

Il DES e il TES hanno andamenti abbastanza simili: sicuramente è presente un trend nei dati, probabilmente vi è anche

una componente di stagionalità che non ha un valore eccessivamente elevato, in quanto il TES non si distacca eccessivamente dal DES.

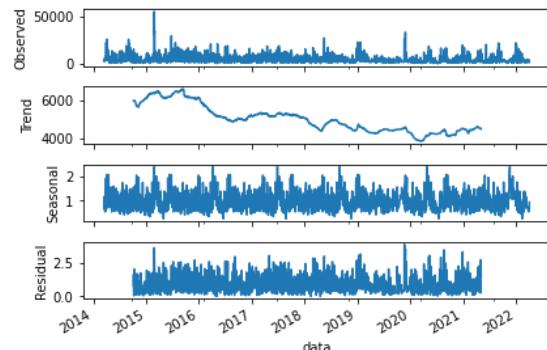


Figura a.59: Andamento della Stagionalità, Trend, Osservazioni e Residui

La granularità dei dati non consente di trarre conclusioni sulla stagionalità dei dati (che sembra esserci nella parte iniziale di ogni anno).

Il trend del totale del settore Running raggiunge un picco tra il 2015 e il 2016 ma negli anni successivi è in costante calo.

Costruzione modelli di previsione

Per ognuna delle 4 categorie considerate (Fitness è stata esclusa per i motivi descritti nell'analisi esplorativa) sono stati costruiti diversi modelli tramite ARIMA (ARIMAX) e Prophet: per trovare il modello migliore sono stati costruiti 7 modelli, 4 utilizzanti ARIMA o ARIMAX e tre utilizzanti Prophet.

All'interno di alcuni modelli sono state inserite delle variabili esogene relative, in particolare, ai saldi e ai periodi di lockdown che hanno caratterizzato il 2020 e il 2021.

La scelta di implementare questo tipo di approccio è stata dettata dal fatto che alcune categorie (calcio su tutte)

presentano un picco anormale (rispetto all'andamento del fatturato nel periodo precedente) tra giugno e settembre 2019 (con il picco massimo raggiunto nel mese di agosto).

Questo picco è stato spiegato dall'azienda in possesso dei dati sottolineando una forte campagna pubblicitaria nel periodo dei saldi; perciò è stato considerato rilevante l'inserimento della variabile esogena (che potrebbe aiutare il modello a adattarsi e prevedere meglio il fatturato in quei periodi).

Sono stati utilizzati dei modelli che hanno tutti lo stesso train set (fino al 31-12-2021), utilizzano dati di granularità giornaliera e prevedono i primi 14 giorni del 2022. La caratteristica di questi modelli è che sono rolling: prendono una finestra di un giorno e un numero di iterazioni iter=14. In questo modo si prevede un solo giorno, a partire dal primo giorno t, proseguendo poi con la previsione del singolo giorno al tempo t+1, con la previsione del singolo giorno al tempo t+2, e così via, per finire con la previsione del singolo giorno al tempo t+13. Si prevede dunque una finestra di tempo giornaliera che si trasla di giorno in giorno fino ad arrivare all'ultimo (quattordicesimo) giorno.

Per quelli che utilizzano ARIMA sono stati scelti come iperparametri quelli forniti dalla funzione `auto_arima`: questa funzione cerca di identificare i parametri **ottimali** per il modello ARIMA effettuando una grid search per minimizzare il valore AIC. Questo processo si basa sulla funzione R comunemente usata, `forecast:auto.arima`. Auto-ARIMA funziona conducendo test di differenziazione (ad esempio, Kwiatkowski–Phillips–Schmidt–Shin, Augmented Dickey–Fuller o Phillips–Perron) per determinare l'ordine di differenziazione,

d, e quindi adattare i modelli all'interno di intervalli definiti dai parametri `start_p`, `max_p`, `start_q`, `max_q`.

I quattro modelli ARIMA si differenziano, per lo stesso settore, in quanto:

- il primo è stato costruito con `auto_arima` con l'aggiunta dei termini di Fourier come variabile esogena. Come descritto precedentemente negli aspetti metodologici vengono inclusi nel modello i termini della serie di Fourier che mirano a spiegare i pattern stagionali.
- il secondo costruito con ARIMA;
- il terzo è costruito con ARIMAX, inserendo come variabile esogena la lista delle date dei saldi (invernali, estivi, del Black Friday e Natale).
- il quarto è costruito con ARIMAX, inserendo come variabile esogena la lista dei saldi e la lista riguardante il lockdown.

Allo stesso modo sono stati costruiti i tre modelli con Prophet, in modo da confrontare i risultati ottenuti e vedere quale sia il modello migliore:

- il primo è costruito con Prophet, senza variabili esogene;
- il secondo è costruito con Prophet, inserendo come variabile esogena la lista delle date dei saldi (invernali, estivi, del Black Friday e Natale);
- il terzo è costruito con Prophet inserendo come variabile esogena la lista dei saldi e la lista riguardante il lockdown.

Analisi risultati modelli

I risultati ottenuti dai vari modelli sono stati raccolti in una tabella: questi risultati sono stati ottenuti mettendo a confronto il valore di previsione del modello con il valore reale della serie storica in un dato giorno e calcolando l'RMS(P)E delle predizioni.

$$RMSPE = \sqrt{\frac{mean}{i=1,n} (100 \cdot |p_i|)^2},$$

RMSPE:

$$p_t = \frac{|e_t|}{y_t}$$

dove

Per calcolare l'RMSPE è stata utilizzata la seguente formula, composta da 6 “step”:

Step 1: Differenza tra il valore reale e il valore predetto.

(test_data - predictions)

Step 2: Divisione del valore ottenuto dalla precedente sottrazione per il valore reale del fatturato (test_data) in quel giorno (a cui è stato sommato un EPSILON piccolo a piacere).

Step 3: Elevato al quadrato il risultato di questo rapporto.

Step 4 : Presa la media dei risultati ottenuti nello step 3 (ad esempio media di 7 valori se la window size è di 7, nel caso in questione media di 1 valore siccome la dimensione della window è 1).

Step 5 : Calcolata la radice quadrata del risultato dello step 4.

Step 6 : Moltiplicato il risultato dello step 5 per 100.

L'RMSPE è il risultato che viene riportato in tabella. I risultati esprimono quindi un valore **percentuale**.

Sono stati evidenziati, per ogni categoria, i modelli che forniscono il risultato migliore di

un giorno specifico nella finestra temporale considerata; in questo modo è stato possibile vedere quale modello sia il migliore (selezionato quello con l'errore minimo) per prevedere un determinato giorno.

1. Calcio

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAFourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid
0	82.009184	49.625999	39.679220	60.38579	8.857473	8.807679	8.735079
1	31.744204	8.273310	15.525257	247.041564	114.978100	114.879764	114.736390
2	45.23058	35.010448	32.299804	13.866773	35.918204	35.947516	35.990254
3	54.343857	47.252377	44.774845	7.286638	44.844629	44.869859	44.906643
4	29.466332	18.472414	17.249469	102.486810	4.428713	2.399746	2.464985
5	72.446174	68.844995	64.593471	5.280402	58.663735	57.786171	57.813739
6	45.876180	42.693074	41.293601	103.096634	17.318589	15.563273	15.618415
7	70.542226	69.964276	68.614647	218.696127	22.300980	24.897415	24.818949
8	67.478881	67.792131	67.86958	122.934280	18.389132	16.665734	16.720157
9	229.409834	220.092858	232.724049	825.941982	227.499205	234.451973	234.233556
10	0.311439	10.054821	7.475920	172.012275	5.723172	3.721686	3.784561
11	16.304813	30.873784	28.872043	129.210366	20.978744	19.353834	19.353834
12	13.985463	23.786521	25.213218	130.283717	19.764160	18.060763	18.114274
13	31.398705	35.087321	44.84568	103.725415	27.046094	25.497291	25.545946

Tabella a.60: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Calcio

Nella tabella a.60 si evince che non esiste un singolo modello che effettua le previsioni in maniera più precisa di tutti gli altri; nei giorni presi in considerazione, il modello ARIMASaldi, ovvero il modello Arima contenente le variabili esogene Saldi, ha il numero maggiore di previsioni più accurate (4), invece i peggiori che contano solo un'occorrenza sono:

- ProphetSaldi
- ARIMASaldiCovid

2. Casual

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAFourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid
0	130.990376	139.649368	127.773376	48.342874	54.171967	41.187940	41.027719
1	38.543936	36.251781	39.454824	62.489434	61.404110	64.654572	64.094682
2	3.626252	5.989543	1.258489	43.899257	42.855853	47.668409	47.727795
3	5.212833	6.824626	3.087039	42.796631	42.316407	47.174393	47.234340
4	17.527881	16.487709	19.768686	52.801916	52.876596	39.784683	39.833854
5	24.873369	26.564753	21.668769	28.842377	30.631580	11.359803	11.431888
6	10.494961	10.731491	8.838645	35.387579	36.714212	19.132303	19.198071
7	27.809324	27.09667	23.625484	24.829227	27.03698	6.766575	6.842400
8	40.427148	36.642117	34.966167	18.935116	21.992185	0.320237	0.401305
9	512.209923	504.173360	486.546124	233.468687	216.239657	306.690584	306.359831
10	15.048995	12.385782	9.549710	34.471878	37.928763	20.684278	20.748784
11	185.285211	176.835456	170.264106	78.242053	67.653561	114.230679	114.056450
12	133.037708	125.716969	120.313801	51.042467	41.155551	80.371054	80.224361
13	14.195713	9.153863	6.549861	19.546915	25.251863	4.485511	4.563191

Tabella a.61: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Casual

In questa tabella si può notare che i modelli ARIMASaldi e ProphetSaldi hanno il numero maggiore di predizioni più accurate (4 ciascuno); il peggiore è Prophet.

3. Pesca

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAfourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid	
0	26.111025	32.257752	13.805024	3.798041	10.636075	3.012696	7.583869	
1	6.322372	12.063064	0.160580	30.467473	22.405002	27.752524	35.184576	
2	21.265604	26.222138	16.650590	40.049564	35.165629	39.630041	45.842785	
3	127.863718	141.151702	122.555196	22.458388	28.308392	19.467302	7.178107	
4	25.201668	34.149243	23.646966	23.737283	22.535637	7.430492	14.849906	
5	80.060021	96.684884	79.662580	23.728761	22.000502	45.796790	34.111227	
6	83.419343	101.920483	87.499350	34.033863	28.564023	53.633731	41.320040	
7	12.636647	25.252185	15.571518	8.949265	14.832485	1.774697	6.388235	
8	31.482259	24.471830	28.501387	51.949394	56.042648	47.471058	51.681235	
9	91.504054	107.740192	98.830725	3.612726	7.004298	11.129664	2.222059	
10	120.884628	141.453079	133.394506	27.839016	12.941529	34.964644	24.147274	
11	48.383415	65.030277	57.860505	4.350312	16.524375	0.246808	8.241992	
12	28.324320	42.544794	37.305932	9.250200	21.498903	6.181356	13.710088	
13	24.431539	38.781229	33.305161	9.448257	22.087071	6.906188	14.367607	

Tabella a.62: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Pesca

Per quanto riguarda il settore Pesca le previsioni più accurate vengono effettuate dal modello ARIMASaldi con 5 occorrenze; il modello meno preciso (come per il settore Casual) rimane Prophet con 0 occorrenze.

4. Running

	rmseProphet	rmseProphetSaldi	rmseProphetCovid	rmseARIMAfourier	rmseARIMA	rmseARIMASaldi	rmseARIMASaldiCovid	
0	64.522606	58.942795	50.926023	80.716164	80.817830	69.887439	72.868761	
1	77.482301	78.251887	79.409494	66.838976	67.033031	48.247780	53.371557	
2	57.543511	58.493295	60.222942	78.714423	78.851238	66.800242	70.087215	
3	66.351360	66.806754	67.573503	61.368590	61.638867	39.780068	45.742199	
4	39.425907	39.965271	41.373833	78.020573	78.186864	64.359696	67.750031	
5	66.958596	67.209237	67.771097	46.591136	47.025240	13.444977	21.678646	
6	10.712026	10.542672	11.459087	19.841656	20.537860	29.832535	17.482031	
7	27.876911	28.406125	28.233190	45.213830	45.719836	11.312251	19.748799	
8	18.027608	17.981206	19.351327	6.339040	7.255532	51.534202	37.119493	
9	40.137090	41.016388	39.308377	60.582076	58.524280	159.665009	134.964061	
10	177.643965	179.304201	176.769930	755.734616	746.437793	1282.987721	1151.429054	
11	1267.789682	1269.778756	1262.443395	46.144613	47.025240	13.444977	21.678646	
12	21.800540	21.993191	22.862863	110.166452	107.661059	239.295690	207.019707	
13	222.801402	222.916767	219.865059	32.527566	30.878819	113.841822	93.499815	

Tabella a.63: Confronto fra gli RMSPE ottenuti dai vari modelli costruiti per la categoria Running

I modelli migliori per il settore Running sono quelli di Prophet e ARIMASaldi che, in 5 casi a testa, effettuano le previsioni caratterizzate dall'errore percentuale minore; al contrario i modelli con l'errore più alto sono ProphetSaldi e ARIMASaldiCovid.

In questo settore anche aggiungendo la componente di Fourier non sembrano

esserci grossi cambiamenti a livello di errore rispetto al modello ARIMA classico; si può dunque ipotizzare che la serie storica del settore Running non presenti una componente stagionale.