VALUTAZIONE ED ANALISI DI MODELLI PREVISIONALI PER LE VENDITE NELLA RISTORAZIONE:

UN ESEMPIO DI APPLICAZIONE DELLA DATA SCIENCE IN UN CONTESTO IMPRENDITORIALE.

Luca Ballarati l.ballarati@campus.unimib.it
Mohamed Helmi Ben Hassine m.benhassine@campus.unimib.it
Francesco Oliviero f.oliviero2@campus.unimib.it
Daniele Quattrocchi d.quattrocchi2@campus.unimib.it
Letterino Sauro l.sauro@campus.unimib.it

INDICE

SINOSSI

Tramite strumenti statistici e di data science, è possibile monitorare i risultati ottenuti da un'azienda, al fine di poter correggere eventuali azioni che sono dannose per la stessa e migliorarne i risultati. Questo report ha proprio l'obiettivo di innovare il settore della ristorazione dal punto di vista della scienza dei dati fornendo analisi statistiche al ristoratore che, dopo essere state spiegate, possono far comprendere allo stesso se i potenziali risultati coincidono con quelli ottenuti, date le caratteristiche del ristorante. Per fare questo, sono state considerate solamente le entrate giornaliere di ogni ristorante nel periodo precedente al Covid, al fine di non considerare situazioni economiche estreme. Per raggiungere gli obiettivi è stato eseguito un clustering dei ristoranti. Inoltre, si sono studiati i risultati giornalieri confrontati per settimane al fine di identificare potenziali pattern ripetitivi.

Successivamente, per ciascun ristorante rappresentativo del clustering, si sono applicati due diversi modelli di previsione al fine di confrontarne l'efficacia.

Infine, si sono confrontati i risultati previsti con i valori dello scenario che abbiamo vissuto.

PAROLE CHIAVE

Stagionalità, *pattern ripetitivi*, *entrate giornaliere*, *previsione*, *confronto scenari*.

1. INTRODUZIONE

"Ciò che non può essere monitorato non può essere valutato."

Grazie all' innovazione tecnologica, disponiamo di strumenti che ci consentono di fare delle analisi e previsioni sempre più accurate sul core business di un'azienda. Nell'ambito della ristorazione si fa poco riferimento ai dati generati giornalmente che, come sappiamo, sono utili per la previsione di risultati futuri ma anche al miglioramento dell'attività stessa. Con il dataset a nostra disposizione, riguardante le entrate giornaliere di sei ristoranti, abbiamo scelto di adottare un approccio data-driven per consentire ai ristoratori una migliore allocazione delle risorse e una più intelligente presa delle decisioni.

2. OBIETTIVO/PROBLEMA AFFRONTATO

L'obiettivo di questa ricerca è stato quello di studiare se nel corso di ciascuna settimana del periodo di riferimento dei dati vi fossero dei pattern settimanali che avrebbero potuto dare informazioni importanti sul comportamento del ristorante, indipendentemente dalle sue caratteristiche. Inoltre, si è verificato se il periodo festivo influenzasse in maniera significativa l'andamento settimanale delle entrate giornaliere. Per soddisfare questo obiettivo, si è scelto di escludere il periodo della pandemia, al fine di mettere in risalto le potenzialità delle tecnologie, senza ottenere bias all'interno dei risultati. Tuttavia, per sopperire a questa problematica, una volta applicati i modelli di previsione a 91 giorni (i.e. 12 settimane successive ai dati processati) si sono confrontati i valori previsti con i valori ottenuti durante il Covid. Per tale ragione, non si è considerato il ristorante 3 che ha iniziato la sua attività soltanto un anno prima dell'inizio della pandemia. Dunque, non si avevano dati sufficienti per soddisfare gli obiettivi posti relativi a tale ristorante. Un'altra problematica è stata quella di considerare un periodo temporale di analisi comune per tutti i ristoranti. Per semplicità si è deciso di considerare il periodo di attività che coincide con il giorno di apertura del ristorante 6. Questo ha comportato ad una perdita di valori per i ristoranti rappresentativi del clustering 1 e 2 che, nonostante ciò, non ha modificato i pattern e le caratteristiche della serie storica.

3. ASPETTI METODOLOGICI

Al fine di esplorare le serie storiche, concentrandoci sulle vendite giornaliere per ogni ristorante, abbiamo usato diversi approcci metodologici e strumenti. La motivazione dietro l'uso di certi modelli rispetto ad altri sarà chiarita e spiegata più avanti nella sezione Analisi/Processo di trattamento dei dati.

Il metodo utilizzato per procedere in questa parte dovrà essere ripetuto per il secondo e il sesto ristorante, perché ciascuno di essi ha una serie temporale differente.

Nel descrivere queste serie temporali, abbiamo usato parole come "trend" e "stagionale" che devono essere definite in modo più dettagliato:

- Un trend o tendenza esiste quando c'è un aumento o una diminuzione nei dati. Non deve essere necessariamente lineare. Infatti, esso potrebbe avere un "cambio di direzione" quando passa da una tendenza all'aumento a una tendenza alla diminuzione.

- Un modello stagionale si verifica quando una serie temporale è influenzata da fattori stagionali come il periodo dell'anno o il giorno della settimana. La stagionalità è sempre di una frequenza fissa e nota.

- Un ciclo si verifica quando i dati mostrano valori alti e bassi senza che abbiano una frequenza fissa. Queste fluttuazioni sono solitamente dovute a condizioni economiche e sono spesso legate al "ciclo economico".

Per approfondire la previsione e trovare il miglior modello per adattarsi ai dati, abbiamo considerato i seguenti modelli:

3.1)SARIMA
3.2)PROPHET

3.1 SARIMA

L'abbreviazione sta per Seasonal Autoregressive Integrated Moving Average e questo modello cerca di tenere conto di tutti gli effetti dovuti ai fattori che influenzano le vendite, per prevedere le vendite giornaliere di ogni ristorante.

Per capire il suo funzionamento, dobbiamo introdurre il modello ARIMA (AutoRegressive Integrated Moving Average), una classe di modelli statistici utilizzati per analizzare serie temporali stazionarie. La stazionarietà può essere forte o debole: la prima richiede la shift-invariance nel tempo delle distribuzioni dimensionali finite di un processo stocastico, mentre la seconda richiede solo la shift-invariance nel tempo del primo momento e del momento incrociato (i.e. auto-covarianza).

Questo acronimo racchiude tutte le parti principali del modello:

- AR (AutoRegressivo): utilizza la relazione di dipendenza tra un'osservazione e un'osservazione n-lagged.
- I (Integrated): rende stazionaria una serie attraverso la differenziazione della stessa.
- MA (Moving Average): utilizza la relazione di dipendenza tra un'osservazione e un errore residuo, attraverso un modello di media mobile applicato alle osservazioni ritardate.

Un modello ARMA integrato di ordine *d* è un processo stocastico che diventa stazionario dopo essere stato differenziato *d* 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:

$$\Phi(B)(1-B)^d Y_t = \Psi(B)\epsilon_t = \ AR\ IM\ A(p,d,q)$$

Il limite del modello ARIMA è che esso non supporta i dati stagionali, ossia una serie temporale con un ciclo che si ripete. Le stagioni sono importanti quando si tratta di analizzare serie temporali. Certi eventi accadono ogni anno, alcuni esempi possono essere:

vacanze, settimana del ritorno a scuola, festival e conferenze annuali, grandi eventi sportivi, premi annuali.

Alcuni eventi avvengono una volta al mese come pagamenti dell'affitto o del mutuo, bollette della carta di credito, le buste paga o anche settimanalmente.

Quindi, se sappiamo che qualcosa è probabile che accada con una cadenza regolare e che abbia un impatto sulla nostra variabile target, in modo simile ogni volta che si verifica, dovremmo tenerne conto quando costruiamo il nostro modello. Nelle serie temporali, questa ricorrenza di eventi d'impatto a frequenza costante è nota come stagionalità.

Il modello SARIMA è così composto:

(p,d,q)×(P,D,Q)

Dove:

- p, d, q sono gli ordini di autoregressione, differenza e media mobile dei polinomi non stagionali.
- P, D, Q sono gli ordini analoghi per la parte stagionale.

La formulazione di tale modello è la seguente:

$$\Phi(B)\Phi_s(B^s)(1-B)^d(1-B^s)^D Y_t = \ \Psi(B)\Psi_s(B^s)\epsilon_t$$

3.2 PROPHET

Prophet è una procedura per la previsione dei dati delle serie temporali, basata su un modello additivo in cui le tendenze non lineari si adattano alla stagionalità annuale, settimanale e giornaliera, oltre agli effetti delle festività. Esso funziona meglio con serie temporali che hanno forti effetti stagionali e diverse stagioni di dati storici. Prophet è resistente ai dati mancanti e ai cambiamenti di tendenza e in genere gestisce bene i valori anomali.
È un software open source rilasciato dal team *Core Data Science* di Facebook, implementato in R e Python.
È veloce e fornisce previsioni completamente automatizzate che possono essere regolate.

Utilizza un modello di serie temporale scomponibile (Harvey & Peters 1990) in tre componenti principali: trend, stagionalità e festività. Sono combinate nella seguente equazione:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t:$$

Dove:

- g(t): funzione di trend che modella i cambiamenti non periodici nel valore della serie storica
- s(t): cambiamenti periodici (stagionalità settimanali, mensili e/o annuali)
- h(t): feste ricorrenti che non ricadono sempre nello stesso giorno o periodo (es. Pasqua);
- ϵ_t: errore che il modello non riesce a spiegare

I trend g(t) possono essere non lineari e costanti (come l'aumento della popolazione) o lineari ma non costanti. Per definire questa funzione, si identificano dei punti specifici (detti punti di cambiamento) dove il tasso di crescita κ subisce una modifica a cui vengono aggiunti degli aggiustamenti. Analogamente, in questo modo viene trattato il parametro compensativo m.
Considerando la portata C(t) anch'essa mutabile nel tempo si giunge alle seguenti formulazioni:

- trend lineari ma non costanti:

$$g(t) = (\kappa + a(t)^T \delta)t + (m + a(t)^T \gamma)$$

  k: tasso di crescita
  δ: aggiustamenti del tasso
  m: parametro offset
  γ_j: rende la funzione continua (= $-s_j \delta_j$)

- trend non lineari ma costanti (crescita logistica):

$$g(t) = \frac{C(t)}{1 + e^{-(k+a(t)^T\delta)(t-(m+a(t)^T\gamma))}}$$

  C(t): capacità attese del sistema in qualsiasi punto nel tempo

L'incertezza nella predizione dei trend futuri è misurata dall'assunzione che il tasso di cambiamenti in un intervallo T del futuro abbia la stessa frequenza e numerosità media di un analogo periodo T nel passato. Per definire le componenti stagionali, si ricorre a funzioni periodiche. Considerando una serie di Fourier standard1 si stimano i 2N parametri:

$$\beta = [a1, b1, \ldots, aN, bN]^T$$

attraverso la costruzione di una matrice *X*(*t*) di vettori stagionali per ogni istante t, per i dati passati e futuri. Assumendo β distribuito normalmente si ottiene:

$$s(t) = X(t)\beta$$

Le feste ricorrenti sono trattate come mutuamente indipendenti. Viene richiesto all'analista di fornire una lista delle date che ritiene festività non ancora considerate e, mediante una funzione indicatrice Z(t), si associa ogni istante 't' al cambiamento nella previsione relativo all'evento in questione. Supponendo κ distribuito normalmente risulta:

$$h(t) = Z(t)\kappa$$

Questo modello fornisce una serie di vantaggi pratici:

- Flessibilità
- Le misurazioni non devono essere spaziate regolarmente e non è necessario interpolare i valori mancanti (ad es. rimozione degli outlier).
- Adattamento molto veloce
- I parametri del modello di previsione sono facilmente interpretabili

4. DESCRIZIONE DEI DATI ORIGINALI

Il dataset è stato fornito dall'Università degli Studi di Milano-Bicocca. Esso contiene i dati relativi alle entrate giornaliere, in euro ed in numero di scontrini registrati, di sei differenti attività di ristorazione, avviate nel Nord Italia e localizzate in Emilia-Romagna. Il periodo temporale di riferimento dei dati inizia dal 1° gennaio 2017. Tuttavia, alcuni esercizi hanno avviato l'attività dopo tale data. Si tratta dei ristoranti registrati nel dataframe come ristorante 3, 4 e 6. In particolare, la data di ultima rilevazione dei dati è la medesima per tutti i ristoranti, ossia il 12 aprile 2021. Inoltre,

sono presenti dati mancanti nel dataset relativi alle entrate giornaliere, ed in particolare, questo fattore è fortemente presente nel periodo di chiusura delle attività a causa del lock-down per il COVID-19 (i.e. mancanza giustificata).

Nonostante tale incompletezza dei dati, non si ritiene tale fattore un punto di debolezza rispetto al raggiungimento degli obbiettivi prefissati. Come già accennato, il periodo covid non è di interesse per lo studio in questione e i restanti dati mancanti sono esigui per il rimanente, e dunque precedente, periodo di analisi. Si ritiene di notevole importanza evidenziare che le caratteristiche dei ristoranti sono ignote, ovvero non sono state fornite informazioni sulla tipologia del ristorante, sulla sua planimetria, sulla sua precisa localizzazione e così via. Questo rappresenta un punto di debolezza per lo studio, in quanto non è possibile spiegare i pattern settimanali delle analisi in questione che generano i dati. Inoltre, non è possibile essere precisi sulle indicazioni da fornire al ristoratore per poter modificare e/o migliorare tali pattern e dunque l'attività stessa.

4.1 PREPROCESSING

FASE 1

Al fine di confrontare quale sarebbe stato il possibile andamento dei dati per ciascun esercizio in uno scenario no covid con i risultati del mondo reale, si è ritenuto di fondamentale importanza processare i dati in maniera tale da ricostruire la serie storica dello scenario odierno. Per questa ragione, il ristorante 3 non è stato considerato nelle analisi poiché privo di dati per il raggiungimento di tale obbiettivo. Inoltre, tale ristorante non presenta un numero sufficiente di dati per studiare i pattern settimanali precedentemente al periodo di sviluppo del COVID-19 in Italia. In particolare, nella prima fase di preprocessing, i dati mancanti successivi alla data 24 ottobre 2020 (in cui è stato introdotto il decreto che identificava l'emergenza covid di una regione in base al colore) sono stati corretti introducendo il dato zero. La ragione di questa azione è data dal fatto che l'Emilia-Romagna, come le altre regioni, è stata principalmente classificata con il colore rosso, successivamente a tale data. Dopodiché, i dati mancanti relativi al lock-down per tutti gli esercizi sono stati corretti con il dato zero. Successivamente, si sono ricercati eventuali valori estremamente anomali, quali un'entrata giornaliera in euro inferiore a 100. L'identificazione di tali valori ha portato ad una sostituzione del dato con uno zero. Infine, per ciascun ristorante si sono eliminate le osservazioni mancanti precedenti all'inizio dell'attività, e dunque alla rilevazione dei dati poiché privi di significato. In conclusione, il dataset è stato suddiviso in 5 dataframe, ognuno relativo al singolo esercizio. Gli attributi di tale dataframe sono la data, le entrate in euro ed in numero di scontrini registrati e il giorno della settimana.

FASE 2

Al fine di correggere i restanti valori mancanti per i ristoranti in analisi, ad eccezione del ristorante 2 i cui dati sono stati tutti gestiti nella fase 1, si è deciso di utilizzare il software KNIME per interpolare i valori mancanti. A tal fine, ciascun dataset è stato diviso in Training set, in cui non vi erano valori mancanti e Test set, in cui vi erano i valori mancanti in oggetto. Per l'interpolazione, è stato applicato il modello Random Forest, che è risultato molto più efficace rispetto al modello di regressione lineare.

FASE 3

In questa fase è stato realizzato il dataframe che ha permesso di effettuare gli studi di pattern settimanali e successivamente la previsione dei dati, escludendo dal dataset tutto il periodo relativo al COVID-19.

Utilizzando il software KNIME, si è proceduto a raggruppare i ristoranti mediante opportune tecniche di clustering, utilizzando due modelli per confrontare i possibili diversi risultati: Hierarchical Clustering, che calcola la Manhattan Distance per effettuare la clusterizzazione, e K-means, che richiede in input il numero di cluster da creare. A tal fine, per tutti i ristoranti, si è proceduto a riunire i dataframe considerando come base di unione la data di esercizio comune a tutte le attività, e considerando come attributi solamente la data di riferimento e l'entrata giornaliera in euro dell'esercizio. Il numero di scontrini registrati è

stato escluso in quanto esso non è risultato utile per lo studio in analisi, in quanto non permette di conoscere con una precisione sufficiente quanti clienti sono compresi per ciascun scontrino (i.e. singolo o gruppo). Inoltre, si è proceduto ad escludere da tale dataframe tutte le osservazioni successive al 4 gennaio del 2020 al fine di soddisfare l'obbiettivo prefissato. Dunque, il periodo di riferimento del nuovo dataframe parte dal 24 settembre 2017 e termina il 4 gennaio 2020.

Come prima tecnica di clustering elementare, si è studiata la correlazione tra le entrate giornaliere dei ristoranti in questione. Si notano correlazioni molto elevate: le più elevate (>84%) sono quelle tra il ristorante 2 ed 1; 5 ed 1; 5 e 2. È stato poi applicato il modello Hierarchical clustering che non necessita da parte dell'utente di inserire il numero di cluster da realizzare. Questo è molto importante per non avere un risultato artificiale.

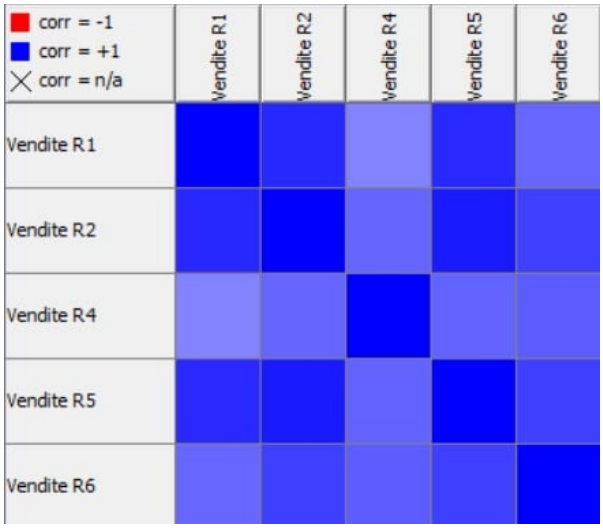

Figura 1: Correlazione

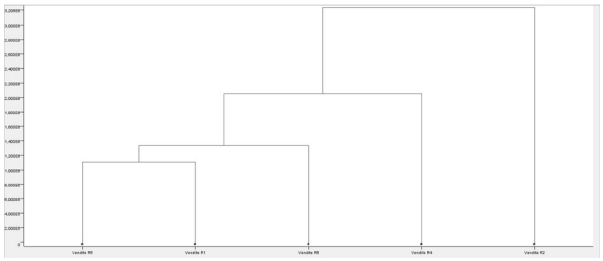

Figura 2: Dendogramma Hierarchical clustering

Successivamente è stato applicato il K-means due volte: la prima per creare 2 cluster, la seconda per creare 3 cluster. Il fine è stato quello di identificare se il dendogramma permette di raggruppare i risultati in 2 o tre cluster e se questi cluster coincidano esattamente con i gruppi realizzati dal K-means.

Ristorante	Cluster
Vendite R1	0
Vendite R2	1
Vendite R4	0
Vendite R5	0
Vendite R6	0

Clusterizzazione K-means: 2 gruppi

Ristorante	Cluster
Vendite R1	0
Vendite R2	1
Vendite R4	2
Vendite R5	0
Vendite R6	2

Clusterizzazione K-means: 3 gruppi

In tutti i casi, i risultati di entrambi i modelli coincidono, e si è proceduto a creare tre cluster.

5. ANALISI/PROCESSO DI TRATTAMENTO DEI DATI

Per ciascun cluster si è considerato un'attività rappresentativa del gruppo stesso. Dunque, sono state eseguite le analisi sul ristorante 1 (i.e. clu.0), ristorante 2 (i.e. clu. 1) e ristorante 6 (i.e. clu. 2).

5.1 VALUTAZIONE DELLE CARATTERISTICHE DELLE SERIE STORICHE

Attraverso l'utilizzo del linguaggio di programmazione R, abbiamo creato la serie storica dove i cicli sono ciascuna settimana. In R, le settimane iniziano da domenica, ed è per questo che il periodo di riferimento dei dati è stato manualmente impostato dal 24 settembre 2017

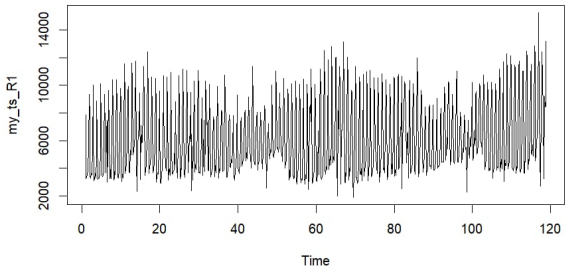

Figura 3: Serie storica R1

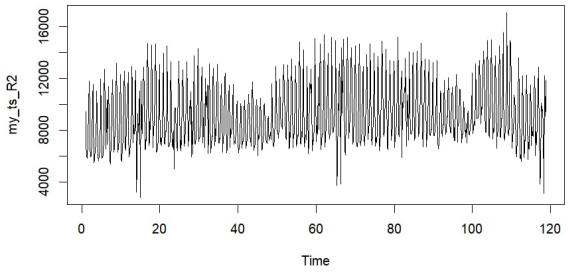

Figura 4: Serie storica R2

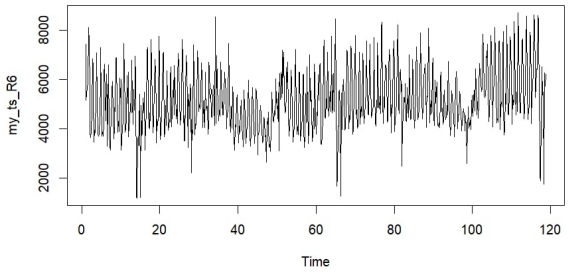

Figura 5: Serie storica R6

Al fine di studiare le serie storiche in oggetto, si è effettuata una decomposizione di ciascuna serie storica, per valutare in modo elementare le componenti della stessa. In altre parole, è stata studiata la stazionarietà della serie storica (funzione 'decompose' del pacchetto 'forecast'). Dunque, emerge dalla decomposizione e dalla visualizzazione delle serie storiche che le serie storiche dei ristoranti 1, 2 e 6 sono stazionarie (i.e. media, varianza, distanza tra i picchi costante e nessuna presenza di trend). Inoltre, è stato eseguito il test Augmented Dickey–Fuller per valutare la stazionarietà e il risultato del test conferma quanto affermato precedentemente.

5.2 VALUTAZIONE DEI CICLI SETTIMANALI

Per la valutazione dei cicli settimanali delle serie storiche, si è deciso di utilizzare il software KNIME. Con questo software, si è deciso di considerare come inizio della settimana il lunedì, in quanto i valori si abbassano significativamente passando dalla domenica al lunedì (si tratta di un puro fine di visualizzazione). Inoltre, il mese di dicembre per tutti i ristoranti è il mese con maggiore variabilità. Questo potrebbe essere dato dall'incidenza delle festività. I cicli di questo mese per gli anni di riferimento sono stati studiati per rilevare l'incidenza delle festività sull'andamento dei valori della settimana.

5.2.1 Ristorante 1

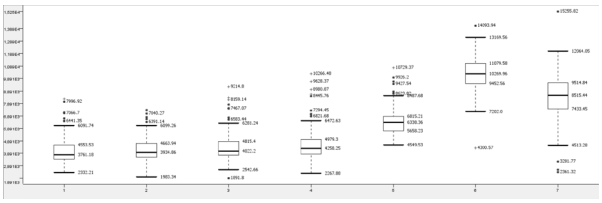

Figura 6: Boxplot vendite giornaliere su frequenza settimanale R1

Dalla figura 6, emerge che i valori si distribuiscono secondo una curva normale asimmetrica negativamente. La variabilità delle entrate giornaliere è la medesima nei giorni che vanno dal lunedì al giovedì, per poi modificarsi in aumento il venerdì e significativamente il sabato e diminuire la domenica. Tale fattore mette in evidenza il fatto che il fenomeno delle entrate giornaliere, nei giorni compresi tra lunedì e giovedì, si comporta allo stesso modo. Dunque, non si verificheranno particolari differenze di entrate in questi giorni, in cui le entrate stesse, per il 75% dei valori, sono comprese tra 2000 e 5000 euro.

Successivamente, il venerdì si presenta un aumento dei valori delle entrate. Si noti che le entrate sono più solide in questo giorno, in quanto vi è una maggiore concentrazione delle stesse, che per il 75% dei valori, sono comprese tra 4500 e 6800 euro. Pertanto, il venerdì si attendono maggiori entrate (e più significative) rispetto ai giorni precedenti.

Il sabato, vi è una variabilità che non permette di fatto di conoscere in quale range di valori le

entrate si concentrino. Tuttavia, possiamo affermare che, dato che il box-plot è spostato significativamente verso l'alto, le entrate in tale giorno sono: 1. Per il 25% dei valori, uguali o simili a quelle del venerdì; 2. Per il restante 75% dei valori, significativamente più elevate del venerdì.

La domenica rappresenta il giorno con maggiore variabilità dei valori. Tuttavia, confrontando questo giorno con il sabato, è possibile affermare che nel giorno di domenica il 75% delle entrate sono simili o uguali a quelle del sabato.

Analisi valori anomali

Essi sono esigui, e questo significa che tuttavia, le festività potrebbero non avere una rilevante incidenza sui valori ottenuti dal ristorante in ogni ciclo. In particolare, le settimane riferite agli anni 2017, 2018, 2019 del mese di dicembre si comportano allo stesso modo delle settimane analizzate nei box-plot (tutte le settimane). Pertanto, si ritiene che le festività non modifichino significativamente l'andamento dei valori delle entrate nei giorni della settimana.

5.2.2 Ristorante 2

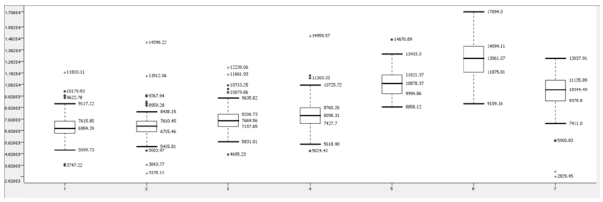

Figura 7: Boxplot vendite giornaliere su frequenza settimanale R2

Dalla figura 7, emerge che le entrate in euro del ristorante 2 si comportano allo stesso modo dei valori del ristorante 1 (i.e. curva normale asimmetrica negativamente). Tuttavia, si ritiene necessario evidenziare quanto segue: Rispetto al ristorante 1, le entrate dal lunedì al venerdì sono in progressivo aumento. Il 75% dei valori si concentra, in aumento, tra 5000 e 8760 euro. Questo significa che tali giorni si differenziano tra loro per quanto riguarda il comportamento delle entrate. Si ritiene questo risultato ottimo da un punto di vista economico, in quanto questo fenomeno potrebbe significare una solida salute finanziaria, che non è sostenuta solamente dagli ultimi giorni della settimana (venerdì, sabato, domenica) ma da tutti i giorni della settimana stessa. In generale, possiamo affermare che il ristorante 2 ottiene valori più elevati rispetto al ristorante 1.

Analisi valori anomali

Come per il ristorante 1 i valori anomali ottenuti dal ristorante 2, identificati dalle croci, sono stati analizzati e sono riferiti ai giorni di festività. Essi sono esigui e questo significa che tuttavia, le festività non hanno una rilevante incidenza sui valori che in ogni ciclo ottiene il ristorante.

5.2.3 Ristorante 6

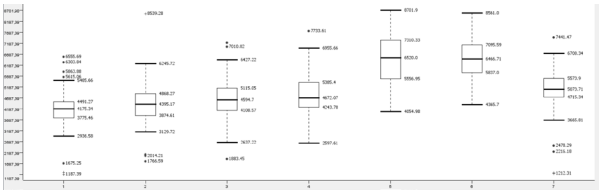

Figura 8: Boxplot vendite giornaliere su frequenza settimanale R6

Dalla figura 8, emerge che le entrate in euro del ristorante 6 si comportano allo stesso modo dei valori del ristorante 1 (i.e. curva normale asimmetrica negativamente). Tuttavia, si ritiene necessario evidenziare quanto segue:

i valori delle entrate tra il lunedì e il giovedì sono in leggero aumento, e per tutti i giorni vi è una elevata variabilità dei valori, ad eccezione del lunedì, martedì e della domenica.

Analisi valori anomali

Si veda quanto affermato per il ristorante 1 e 2.

5.3 Costruzione modelli SARIMA

In questo sottoparagrafo si mostrerà come sono stati costruiti i modelli ARIMA e la previsione dei valori a 91 giorni (i.e. 12 settimane successive ai dati processati).

5.3.1 Ristorante 1

Innanzitutto, è stata utilizzata la funzione *tsclean()* del pacchetto forecast per gestire i valori outliers della serie storica. In particolare, questa funzione sostituisce i valori anomali correnti con i corrispondenti valori interpolati tramite un modello di regressione lineare.

Successivamente, si è studiata l'autocorrelazione tra i valori della serie storica.

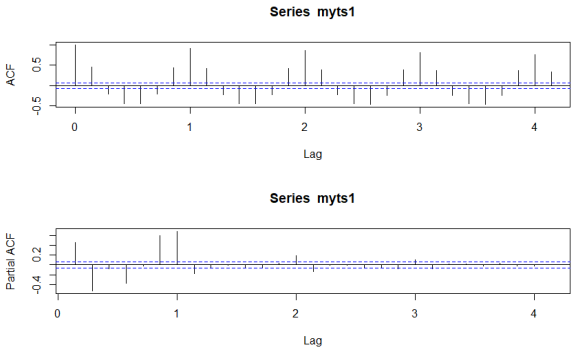

Figura 9: test autocorrelazione ACF vs PACF per R1

Dalla figura 9, si evince che la serie storica è caratterizzata da una componente stagionale molto forte ed in particolare si può notare che l'autocorrelazione diminuisce all'aumentare del numero di lag interi (i.e. 1 lag = una settimana).

L'applicazione del test di autocorrelazione Durbin-Watson conferma quanto affermato.

Fatte tali premesse è stato costruito il modello ARIMA, applicando la funzione del software R "auto.arima()" contenuta nel pacchetto "forecast". Date le componenti di tale modello si evince che il modello costruito sia un modello SARIMA. La sua formulazione è ARIMA (0,0,2) (0,1,1).

Il modello stimato può essere esplicitato formalmente come segue:

$$\Delta y_t = a_1 \Delta y_t - 1 + \epsilon_t + b_1 \epsilon_t + b_2 \epsilon_t$$

Dove:

$$\Delta y_t = (y_t - y_t - 1)$$

con coefficienti:

ma1	ma2	sma1
0.3612	0.0559	-0.3956

É utile sottolineare come i parametri siano tutti significativi in quanto il valore assoluto del rapporto tra coefficiente e standard error risulta maggiore di due.

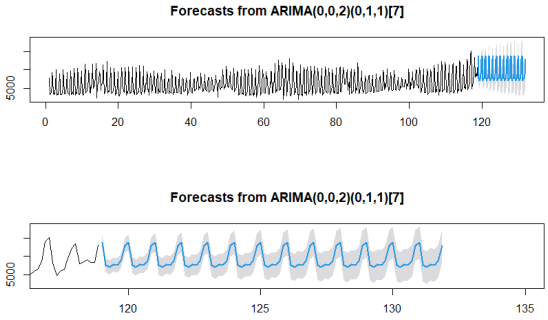

Figura 10: Previsioni Vendite Ristorante 1

Utilizzando la funzione *forecast()* del pacchetto forecast sono state previste le 12 settimane successive. L' intervallo di confidenza dei valori previsti è stato impostato al 95% e si nota una forte componente stagionale nella previsione.

Successivamente si è studiata l'autocorrelazione tra i residui e dal test Durbin-Watson si evince che i residui non sono autocorrelati. Tuttavia, valutando la normalità dei residui, considerato il test statistico di Jarque-Bera, emerge che i residui non seguono la distribuzione normale.

Questo risultato è principalmente imputabile ad una numerosa presenza di valori anomali che sono individuabili sulle code della distribuzione.

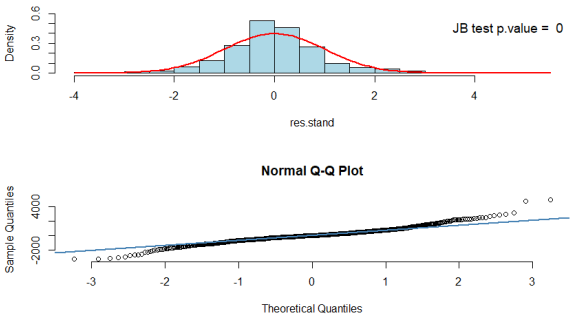

Figura 11: test statico di Jarque-Bera

Si è valutata la bontà di adattamento utilizzando l'indice MAPE (errore percentuale medio assoluto). Il valore di tale indice è pari a 15,84. Pertanto, in media, la differenza tra i valori previsti e i valori reali è del 15,84%.

5.3.2 Ristorante 2

Come per il ristorante 1, è stata utilizzata la funzione *tsclean()* per gestire i valori anomali al fine di ottenere risultati più corretti. Successivamente si è studiata l'autocorrelazione tra i valori della serie storica, confermata dal Test Durbin-Watson. I risultati sono i seguenti:

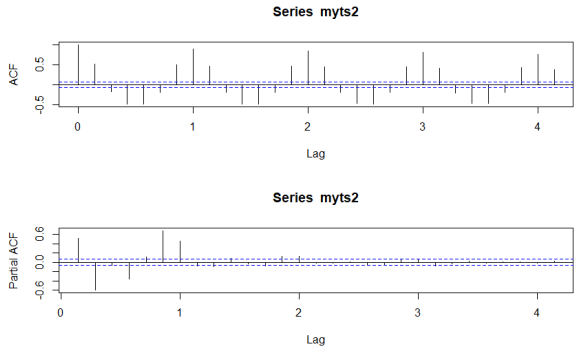

Figura 12: test autocorrelazione ACF vs PACF per R2

La situazione è la medesima del ristorante 1.

Si è costruito il modello ARIMA con la funzione auto.arima(). Il modello stimato è questo ARIMA(4,0,3)(0,1,1) i coefficienti sono i seguenti:

ar1	ar2	ar3	ar4	ma1	ma2	ma3	sma1
0.8746	0.2169	-0.7787	0.3361	-0.4969	-0.4693	0.5383	-0.5209

I parametri sono risultati significativi.

Utilizzando la funzione forecast per prevedere le 12 settimane successive dei valori delle vendite in euro si è ottenuto quanto segue:

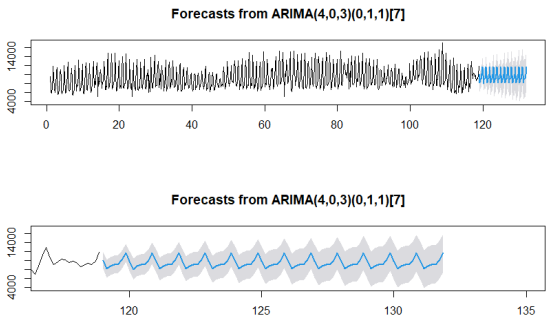

Figura 13: Previsioni Vendite Ristorante 2

Come per il ristorante 1 la componente stagionale è molto forte e si è impostato un intervallo di confidenza dei valori al 95%

Per quanto riguarda i residui la situazione è simile a quella del ristorante 1. Non vi è autocorrelazione tra i residui, ma dal test di normalità si evince che i residui non sono normali a causa di un'elevata concentrazione di valori anomali sulle code.

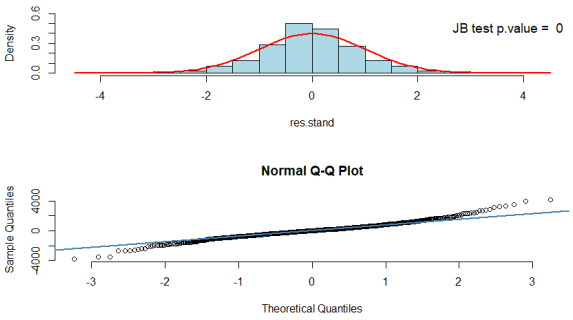

Figura 14: test statico di Jarque-Bera

Si è valutata la bontà di adattamento utilizzando l'indice MAPE (errore percentuale medio assoluto) il valore di tale indice è pari a 11,16%. Pertanto, in media, la differenza tra i valori previsti e i valori reali è del 11,16%.

5.3.3 Ristorante 6

Le stesse considerazioni vengono fatte anche per questo ristorante. In particolare, i valori della serie storica sono autocorrelati, il modello ARIMA è ARIMA (2,0,3) (0,1,2) e i coefficienti sono:

ar1	ar2	ma1	ma2	ma3	sma1	sma2
-0.0436	0.8628	0.3565	-0.7291	-0.2119	-0.5668	-0.1433

I residui del modello non sono autocorrelati e non seguono una distribuzione normale.

In particolare, i grafici riferiti a quanto affermato sono i seguenti:

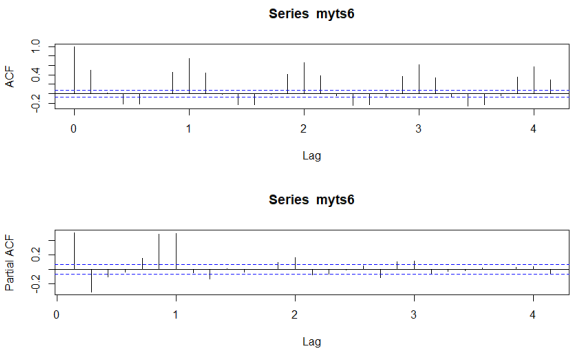

Figura 15: test autocorrelazione ACF vs PACF per R6

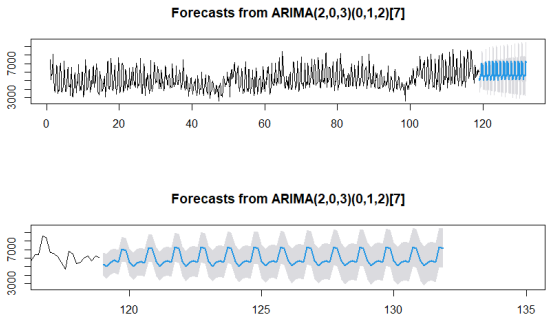

Figura 16: Previsioni Vendite Ristorante 6

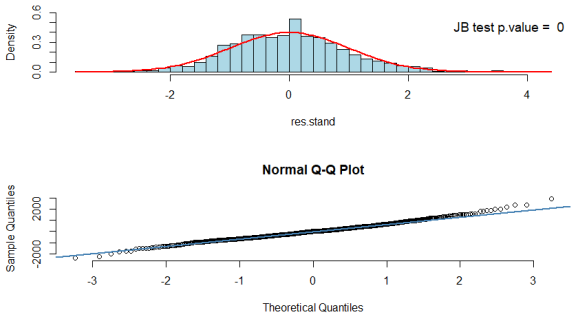

Figura 17: test statico di Jarque-Bera

Si è valutata la bontà di adattamento utilizzando l'indice MAPE (errore percentuale medio assoluto) il valore di tale indice è pari a 13,83%. Pertanto, in media, la differenza tra i valori previsti e i valori reali è del 13,83%.

5.4 COSTRUZIONE MODELLI PROPHET

Il modello Prophet, per essere creato tramite la funzione omonima, richiede in input un data frame storico costituito da due colonne:

- ds: le date, in formato datetime
- y: la variabile di interesse, nel nostro caso le vendite.

Prima di eseguire il fitting del modello, per ogni ristorante analizzato, sono stati rimossi gli outlier ed è stata aggiunta una collezione built-in delle festività per il paese Italia.
Dopodiché, è stato possibile chiamare la funzione di previsione per le vendite nei 3 mesi successivi la data precedente lo scoppio della pandemia Covid. Si è quindi voluto ottenere una prospettiva su un orizzonte temporale di 91 giorni a partire dal giorno 05/01/2020.

Per impostazione predefinita si includono anche le date storiche in modo da poter valutare l'adattamento del campione.
L'oggetto restituito è un data frame con una colonna contente la previsione, e colonne aggiuntive riguardanti gli intervalli di incertezza e le componenti stagionali.
Diamo un'occhiata più da vicino all'analisi delle serie temporali delle vendite dei ristoranti 1, 2 e 6.

5.4.1 Ristorante 1

Qui di seguito, nella figura 18 viene riportata la previsione effettuata attraverso la libreria Prophet. Come possiamo notare, la linea blu riporta le vendite storiche del ristorante sino al giorno precedente indicato come inizio dello scoppio della pandemia, mentre la linea gialla rappresenta la previsione del modello.
Si può vedere che il trend della previsione mantiene costante l'andamento positivo verificatosi negli ultimi mesi.
Inoltre, le linee verde e rossa presenti nel secondo grafico, rappresentano l'intervallo di confidenza della previsione, con un alpha pari a 0,5. La linea verde è il limite superiore, mentre quella rossa rappresenta il limite inferiore.

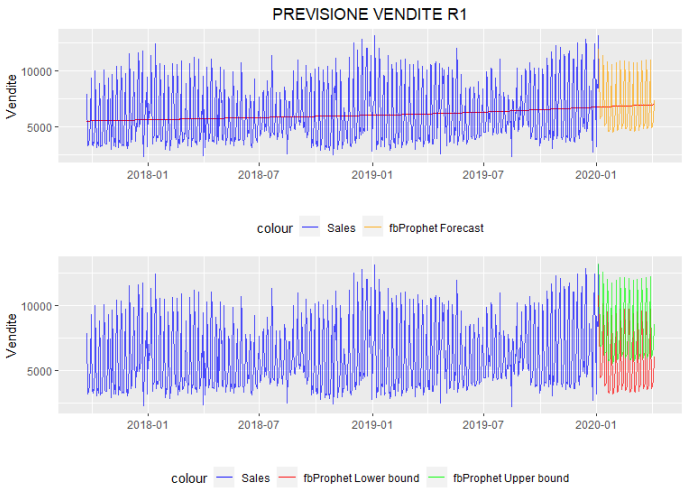

Figura 18: Previsione vendite Ristorante 1 con Prophet

Nella figura 19 sono riportati gli effetti, quali trend, festività, stagionalità settimanale e stagionalità annuale, delle componenti della serie storica. Vale la pena notare che l'effetto delle festività ha un grande impatto sulla vendita giornaliera del ristorante, denotando perdite su alcune festività e guadagni su altre. Ad esempio, a Natale e a Pasqua di ogni anno, le perdite raggiungono un valore rispettivamente di circa -3000 euro e -6000 euro, suggerendo che il ristorante dovrebbe essere

chiuso piuttosto che aperto. Considerando invece festività come la 'Festa della Liberazione' e 'Tutti i Santi', si può notare che si registra un guadagno ogni anno, di circa 4000 euro e 3000 euro rispettivamente, aumentando la vendita giornaliera del ristorante. Tuttavia, l'effetto che impatta maggiormente sulle vendite è la stagionalità settimanale. Il grafico mostra che quando ci sono i giorni del fine settimana, le vendite aumentano molto di più rispetto ai restanti giorni settimanali. Anche la stagionalità annuale ha un impatto abbastanza rilevante, mostrando che a gennaio e a settembre le vendite sono maggiori rispetto agli altri mesi.

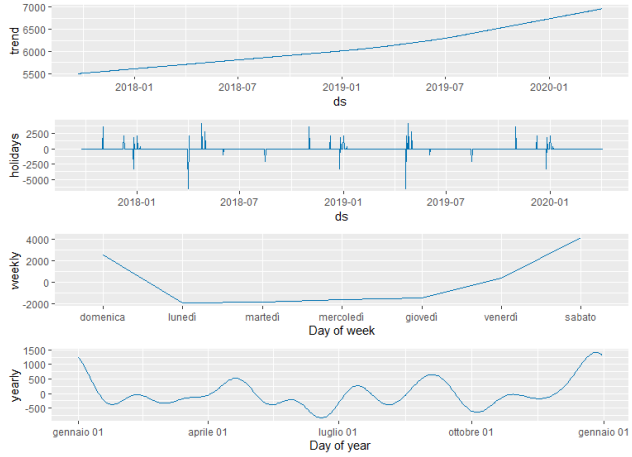

Figura 19: Plot delle rispettive stagionalità Ristorante 1

5.4.2 Ristorante 2

Nel secondo ristorante la situazione è un po' diversa: il trend che nel lungo periodo è in leggera crescita vede diminuirsi nell'ultimo anno. La previsione del modello prophet segue questa tendenza negativa. Anche in questo caso abbiamo lo storico delle vendite rappresentato dalla linea blu e la prospettiva delle vendite in caso di assenza della pandemia rappresentata dalla linea gialla. Possiamo anche individuare limite superiore (linea verde) e limite inferiore (linea rossa) della previsione.

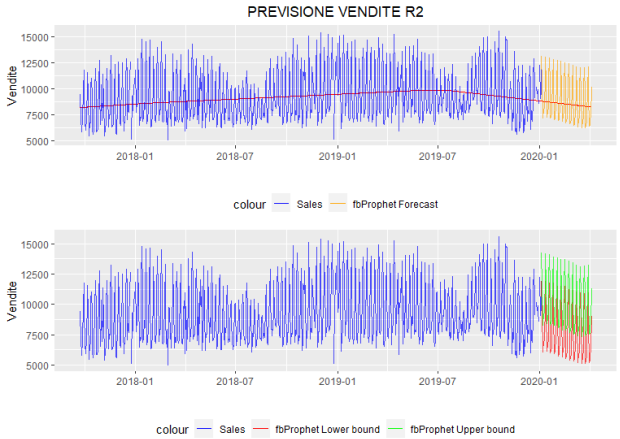

Figura 20: Previsione vendite Ristorante 2 con Prophet

Andando ad analizzare i vari componenti del modello si può percepire che la stagionalità settimanale rimane simile a quella del primo ristorante: dei picchi di vendite nel weekend, percependo già un aumento il venerdì. Nei giorni della settimana la situazione rimane pressoché stabile.

Per quanto riguarda la stagionalità annuale, il mese di agosto è il mese in cui si registra il fatturato più basso. Con l'inizio di settembre le vendite effettuate tornano ad essere subito importanti, registrando il picco più alto.

Le festività hanno sempre un impatto importante sulla vendita giornaliera del ristorante. 'Capodanno', 'Festa dei Lavoratori' e 'Tutti i Santi' sono le ricorrenze che più influenzano positivamente il fatturato. A Natale e Pasqua di ogni anno, come nel caso del ristorante 1, si registrano invece perdite importanti.

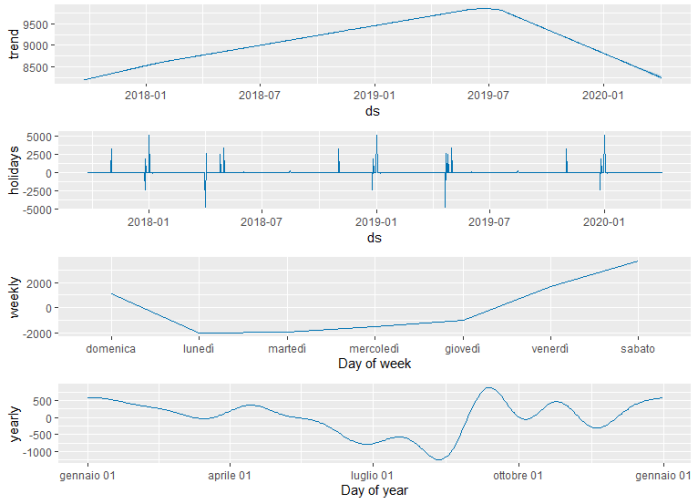

Figura 21: Plot delle rispettive stagionalità Ristorante 2

5.4.3 Ristorante 6

La serie temporale del Ristorante 6 è un po' più irregolare, come si può vedere dalla figura 22.
Si individuano diversi punti di cambiamento del trend, che segue quasi un andamento convesso: prima decresce per poi successivamente crescere. Tale incremento continua anche nella previsione effettuata dal modello per l'orizzonte temporale di 91 giorni. Come nei casi precedenti, la linea gialla rappresenta la prospettiva di vendita, le linee verde e rossa sono rispettivamente il limite superiore e il limite inferiore.

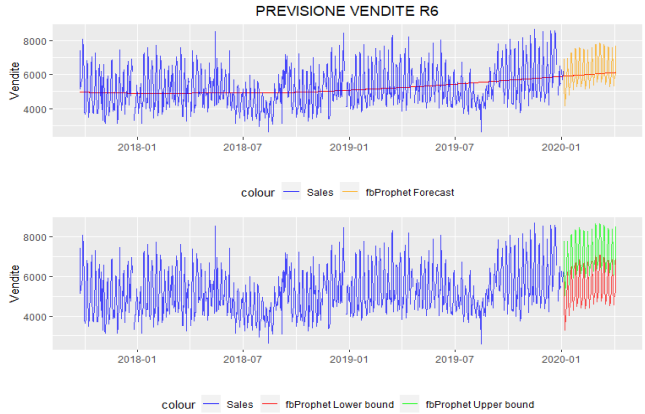

Figura 22: Previsione vendite Ristorante 6 con Prophet

La stagionalità settimanale di quest'ultimo ristorante è leggermente differente dai precedenti. Il venerdì e il sabato sono i giorni in cui si registrano le vendite più alte. La domenica si registra un decremento, che prosegue il lunedì, come anche nei ristoranti 1 e 2. Durante i giorni della settimana le vendite sono basse ma si individua un leggero incremento man mano che ci si avvicina al fine settimana. L'andamento della stagionalità annuale è molto simile alla realtà del ristorante 2.
Per quanto riguarda le festività, durante Capodanno il ristorante dovrebbe rimanere aperto perché si registra un valore positivo delle vendite, magari sfruttando tale ricorrenza organizzando eventi particolari che porterebbero ad un ulteriore aumento del fatturato giornaliero. La festività dell'Immacolata Concezione invece non rappresenta una buona possibilità di guadagno, si registrano ogni anno perdite di circa – 1000 euro.

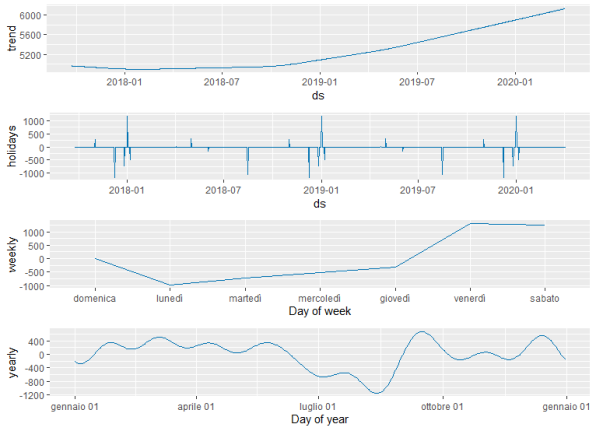

Figura 23: Plot delle rispettive stagionalità Ristorante 6

5.4.5 Valutazione bontà di adattamento Prophet

Per la valutazione del modello è stato usato come unità di misura il MAPE (Mean Absolute Percentage Error) che misura quanto sia accurato un modello di previsione calcolando l'errore medio assoluto per ciascuno periodo di tempo.
Abbiamo suddiviso il database storico costituito da 826 osservazioni in train e test set.
Allenando il modello sul train set, che costituisce l'80% delle osservazioni dell'intero dataframe, abbiamo confrontato i valori previsti con i valori reali presenti nel test set, ottenendo così i seguenti risultati:

MAPE		
R1	**R2**	**R6**
15,18	10,59	12,96

6. RISULTATI

Il modello migliore per prevedere i valori delle entrate dei ristoranti in questione è il modello Prophet che raggiunge un valore inferiore al valore raggiunto dal modello SARIMA. In particolare, di seguito si mostreranno le serie storiche comprendenti anche il periodo della pandemia e post pandemia al fine di confrontare la differenza tra lo scenario che non si è verificato e quello che abbiamo vissuto.

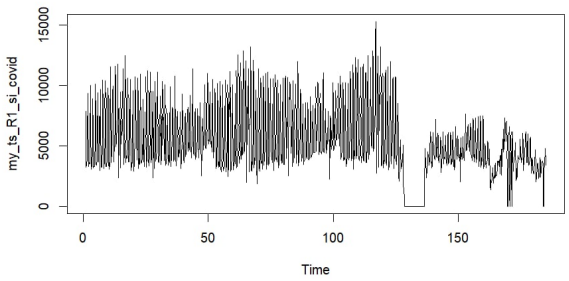

Figura 24: Serie storica delle vendite del Ristorante 1 considerando la pandemia

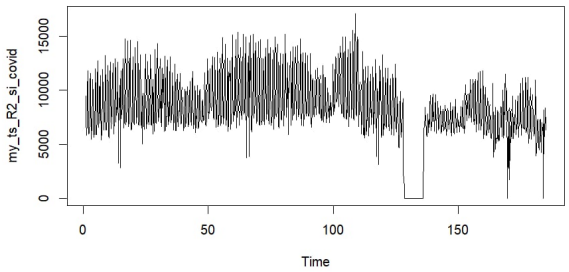

Figura 25: Serie storica delle vendite del Ristorante 2 considerando la pandemia

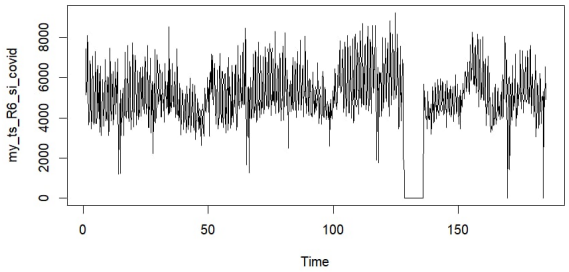

Figura 26: Serie storica delle vendite del Ristorante 6 considerando la pandemia

Nonostante il periodo di Lockdown, ad eccezione del ristorante 1, assistiamo comunque al fatto che, successivamente a tale evento, il ristorante 2 e 6 raggiungono una soglia di valori molto simili al periodo pre-Covid. Pertanto, per il ristorante 1 questo potrebbe significare la presenza di difficoltà finanziarie.

7. CONCLUSIONI E POSSIBILI SVILUPPI

Come si è potuto notare dalle analisi, la Data Science permette di monitorare i risultati che ogni giorno il ristorante raggiunge. Il fine è quello di poterli migliorare suggerendo l'implementazione di azioni, considerati i pattern che continuano a ripetersi nel tempo. Questo aspetto è molto importante in quanto attuando sempre le stesse azioni si ottengono sempre gli stessi risultati. Le nostre analisi non ci permettono di fare delle considerazioni economiche precise in merito alla soglia delle entrate raggiunte dal ristorante in quanto non conosciamo le sue caratteristiche. Tuttavia, possiamo affermare con certezza che il fatturato di un ristorante non debba essere generato/sostenuto solamente dagli ultimi giorni della settimana. Ad esempio, come si può notare, il ristorante 2 presenta delle entrate che si distribuiscono in aumento in tutta la settimana, diminuendo poi la domenica. Inoltre, dalle analisi emerge che le festività portano a delle entrate che non vanno a modificare l'andamento settimanale delle stesse. Questo significa che le festività potrebbero accentuare il valore delle entrate e solo in rari casi ridurlo drasticamente. Dai modelli implementati si evince che la componente stagionale è molto forte all'interno delle serie storiche e che i modelli fanno previsione utilizzando principalmente questa componente.

8. RIFERIMENTI BIBLIOGRAFICI E SITOGRAFICI

1.Fattore M. (2020). Fundamentals of time series analysis, for the working data scientist (DRAFT).

2.K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks | by Imad Dabbura | Towards Data Science

3. https://facebook.github.io/prophet/

4.https://robjhyndman.com/papers/ComplexSeasonality.pdf

5.Forecasting: Principles and Practice (2nd ed) (otexts.com)