Streaming Data Management and Time Series Analysis

Sanvito Alessio 844785

Università degli Studi di Milano Bicocca. Dipartimento di Informatica, Sistemistica e Comunicazione. CdM Data Science F9101Q. Email: a.sanvito6@campus.unimib.it

February 10, 2023

Abstract

La previsione di serie temporali è un'attività sempre più importante all'interno di un mondo sempre più data-driven, in quanto essa permette di prevedere eventi futuri basati su dati storici e permette quindi di prendere decisioni informate in diversi ambiti aziendali; inoltre essa aiuta a identificare tendenze e pattern nei dati, che possono essere utilizzati per migliorare i processi aziendali e aumentare l'efficienza. In particolare, la previsione del consumo di energia elettrica è un compito importante che trasmette informazioni ai fornitori di elettricità e li aiuta a migliorare le prestazioni dei loro sistemi in termini di produttività ed efficienza. Il compito di questo progetto è quello di analizzare l'andamento del consumo e confrontare diversi metodi di previsione; più precisamente vengono utilizzati diversi algoritmi (ARIMA, UCM, Machine Learning) e sono valutati i loro risultati in termini di differenza tra le previsioni e il valore reale (usando il Mean Absolute Error come metrica di riferimento).

Keywords: Time Series Analysis, ARIMA, UCM, Machine Learning

1 Introduzione

Il focus principale di questo progetto è posto sull'analisi e la stima di modelli di previsione per una serie storica relativa al consumo di energia elettrica. La modellizzazione di serie storiche è diventata un'area di studio importante nell'ambito dell'analisi dei dati, grazie alla sua vasta applicabilità in vari domini e grazie alla sua capacità di rispondere a molte domande all'interno di questi domini. In questo studio l'obiettivo è stata la modellizzazione e previsione della serie temporale del consumo di energia elettrica in Marocco nel 2017. Viene effettuato il confronto delle prestazioni dei metodi ARIMA, UCM e Machine Learning; si cerca qundi di identificare il modello che prevede con la massima accuratezza i valori di consumo di energia ogni 10 minuti nel periodo compreso tra il 1° dicembre 2017 e il 30 dicembre 2017.

2 Analisi esplorativa della serie

Il dataset fornito riguarda una time series univariata regolare osservata ogni 10 minuti, relativa a misurazioni di consumo di elettricità. I dati sono organizzati nelle seguenti 2 colonne:

- date - stringa codificante la data-ora della misurazione, in formato dd/mm/yyyy HH:MM:SS
- power - consumo rilevato

I dati coprono il periodo da 01/01/2017 00:00:00 - 30/11/2017 23:50:00 (per un totale di 48096 osservazioni); in particolare è stato considerato come periodo di training quello che va dal 01/01/2017 00:00:00 al 31/10/2017 23:50:00, mentre come validation set tutto il mese di novembre (quindi dal 01/11/2017 00:00:00 al 30/11/2017 23:50:00). Lo scopo è quello di prevedere il mese di dicembre (test set), quindi le 4320 osservazioni successive ai dati disponibili per l'analisi.

Dopo una iniziale fase di preprocessing (per verificare la presenza di valori duplicati o NA), il primo step è stato quello di realizzare dei grafici in grado di far comprendere l'andamento della serie storica. Mostrando i dati relativi alla prima settimana e al primo mese si può individuare una netta componente stagionale sia giornaliera (come mostrato in Figura 1) che settimanale (come mostrato in Figura 2).

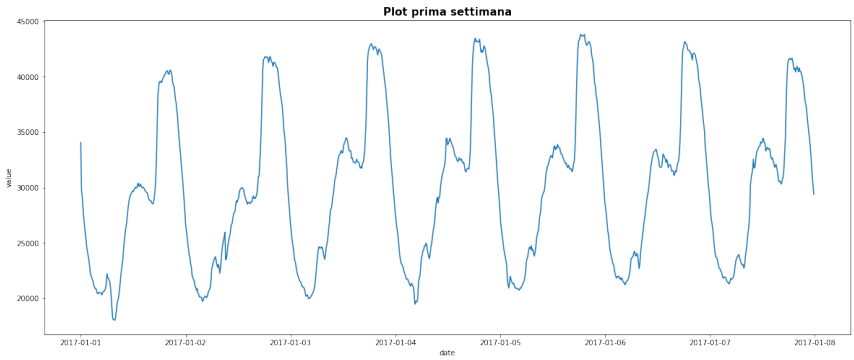

Figure 1. Andamento settimana 1 di Gennaio

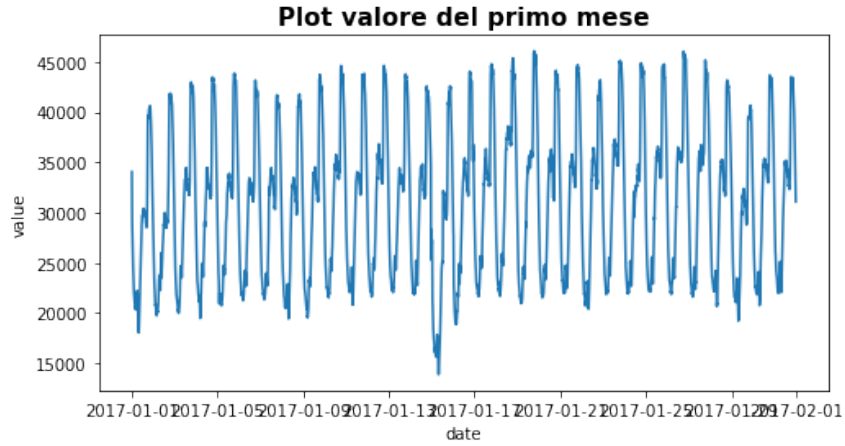

Figure 2. Andamento di Gennaio

Inoltre, per analizzare l'andamento del consumo energetico sono state estratte diverse feature dalla data (trasformata in datetime) come hour, minute, dayofweek, day_name, dayofyear, month, year ecc; queste sono state utilizzate per realizzare ulteriori grafici che permettessero di comprendere ulteriori informazioni sul valore del consumo elettrico. Per esempio, raggruppando i valori della serie in base ai valori della colonna 'day' si è notato come il consumo giornaliero sia massimo

il terzo giorno della settimana (ovvero il mercoledì) e minimo nell'ultimo (ovvero la domenica). Ciò è evidente nelle Figure 3 e 4.

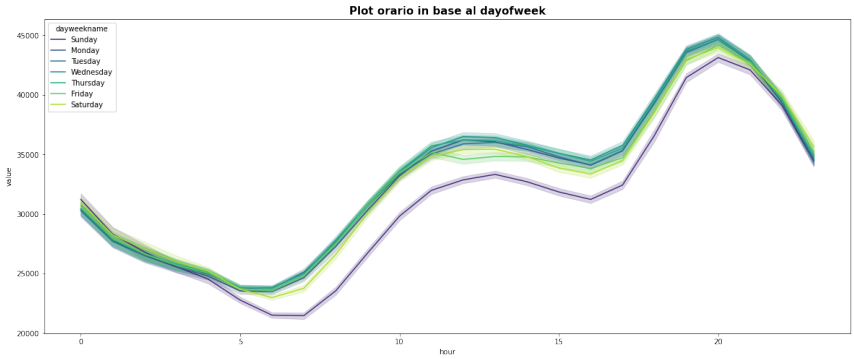

Figure 3. Andamento relativo ai giorni divisi per ore

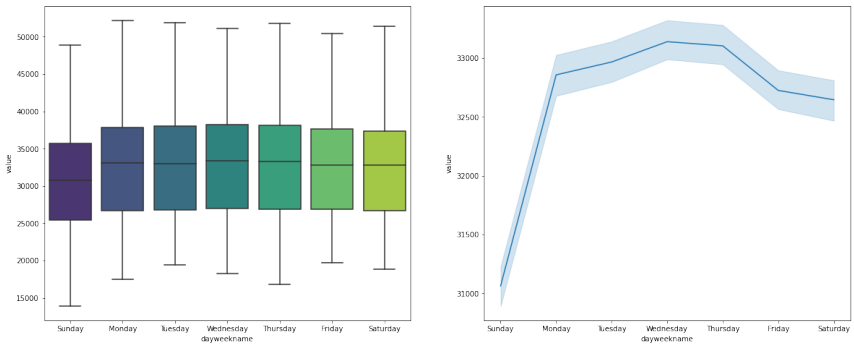

Figure 4. Andamento giornaliero

Si è condotta anche un'analisi riguardante il consumo relativo ai mesi dell'anno che ha evidenziato Agosto come mese con il consumo maggiore; si verifica un aumento nei mesi primaverili e estivi e una diminuzione nei mesi autunnali e invernali (Figure 5 e 6). Analizzando invece l'andamento della serie in base alle diverse ore del giorno si è notato che questo è minimo nelle ore notturne (dalle 21 alle 6), raggiunge un massimo locale intorno alle 12, torna a scendere (di poco) fino alle 17 e successivamente cresce fino a raggiungere il massimo intorno alle ore 20 (vedi Figura 7).

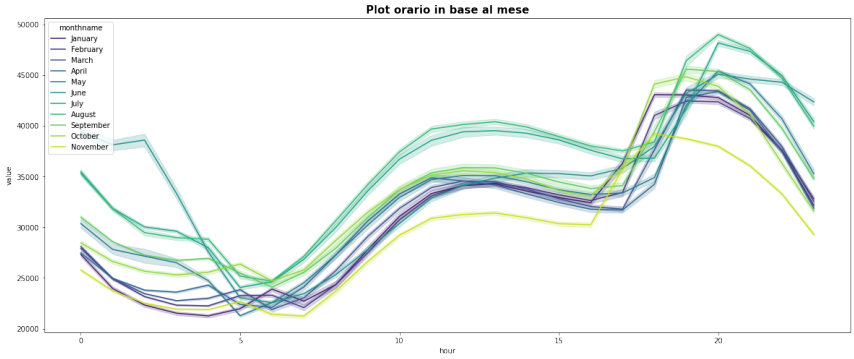

Figure 5. Andamento relativo ai mesi divisi per ore

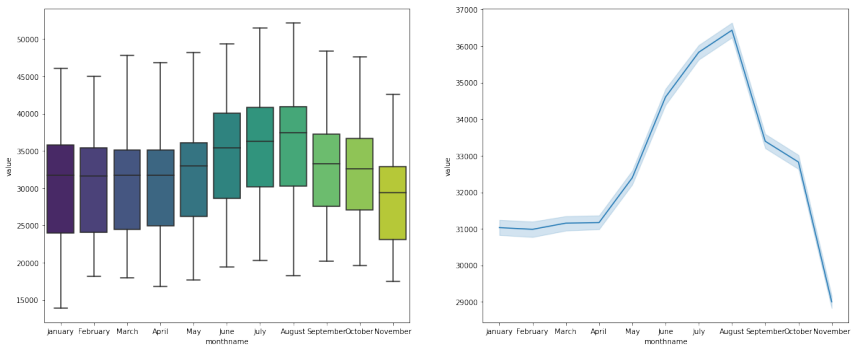

Figure 6. Andamento mensile

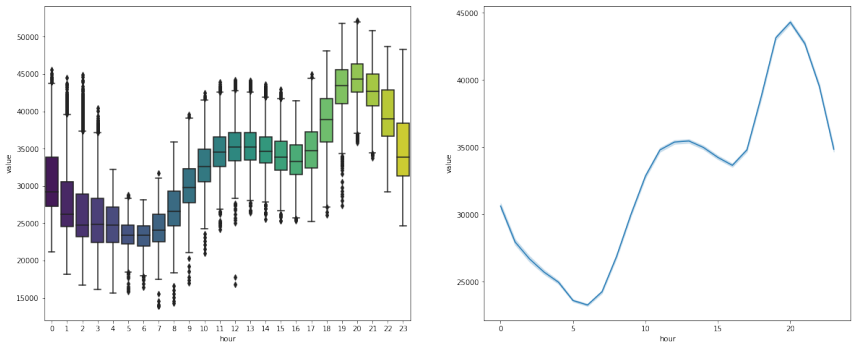

Figure 7. Andamento relativo alle ore

Come ultima analisi di tipo grafico sono stati plottati i correlogrammi riguardanti la serie originale e la serie differenziata di un giorno (ovvero di 144 osservazioni); questi grafici, mostrati in Figura 8, evidenziano degli spike sia per valori di lag pari a 144 che 1008 (e multipli), sottolineando le due stagionalità descritte in precedenza.

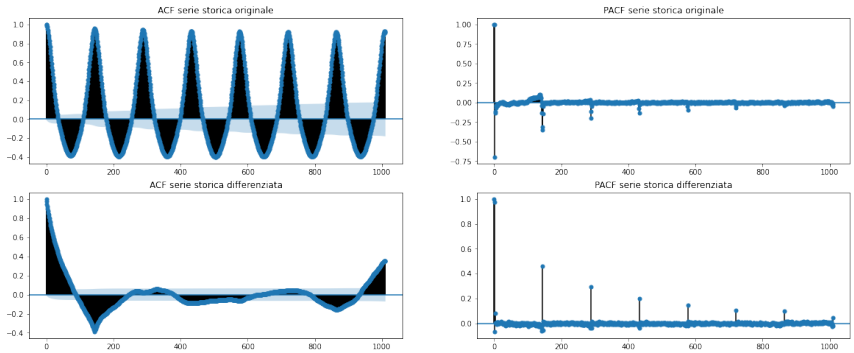

Figure 8. Acf e Pacf della serie storica originale e differenziata (144)

In seguito sono stati condotti dei test statistici per verificare la stazionarietà della serie, tra cui i test ADF e KPSS. Il test ADF viene utilizzato per determinare la presenza di una radice unitaria nella serie e quindi aiuta a capire se la serie è stazionaria o meno; l'ipotesi nulla è che la serie ha una radice unitaria, l'ipotesi alternativa è che la serie non ha una radice unitaria. Se l'ipotesi nulla non viene respinta, questo test può fornire evidenze che la serie non è stazionaria. Il KPSS è un altro test per verificare la stazionarietà di una serie temporale; le ipotesi nulla e alternativa per il test KPSS sono opposte a quelle del test ADF.

I risultati di questi due test sono discordanti:

- I risultati del test di Dickey-Fuller mostrano che il valore del test statistico è -33.54, che è molto più basso rispetto ai valori critici del 1% (-3.43), 5% (-2.86) e 10% (-2.57). Ciò significa che possiamo rigettare l'ipotesi nulla che afferma che la serie ha una radice unitaria; quindi la serie è stazionaria. Il p-value è anche zero (0.00), il che fornisce una forte evidenza contro l'ipotesi nulla.
- I risultati del test KPSS mostrano che il valore del p-value è di 0.01, il che indica che c'è una forte evidenza a favore dell'ipotesi alternativa, ovvero che la serie ha una radice unitaria; quindi la serie non è stazionaria.

In sintesi, dal momento che KPSS indica non stazionarietà e ADF indica stazionarietà, la serie non è stazionaria; un'idea potrebbe quindi essere quella di applicare una differenziazione per rendere le serie stazionaria.

3 Approccio metodologico

Sono stati inizialmente tentati diversi approcci che non hanno portato a dei miglioramenti sostanziali:

- la serie è stata differenziata di 144 osservazioni: i test ADF e KPSS non sono più discordanti (quindi la serie è stazionaria), ma resta stagionalità sia giornaliera che settimanale;
- è stata applicata una trasformazione di Box-Cox (selezionando il lambda ottimale); i test ADF e KPSS sono discordanti e restano stagionalità giornaliera e settimanale;
- trasformazione logaritmica: si ottengono risultati uguali a quelli ottenuti tramite la trasformazione di Box-Cox.

Sono stati analizzati i correlogrammi della serie differenziata (mostrati in figura 9) che evidenziano la necessità di introdurre una componente autoregressiva di ordine 3 (AR(3)) con una differenziazione di 144 periodi.

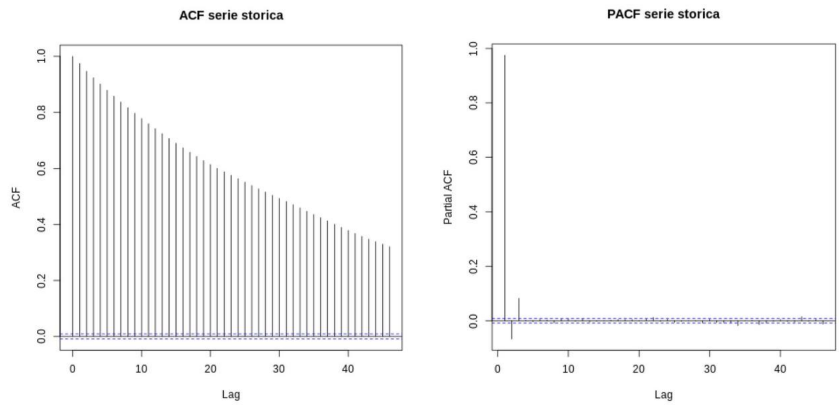

Figure 9. Acf e Pacf della serie differenziata

4 Modelli ARIMA

Il primo modello utilizzato (sulla serie originale) è quindi stato il seguente:

$$ARIMA(3,0,0)(0,1,0)[144] \tag{1}$$

Dalla Figura 10 si può notare che restano sia stagionalità giornaliera (ogni 144 lags) sia stagionalità settimanale, quindi si può constatare che il modello non riesce a catturare globalmente molta memoria; inoltre questo modello ottiene un MAE di 1771.99.

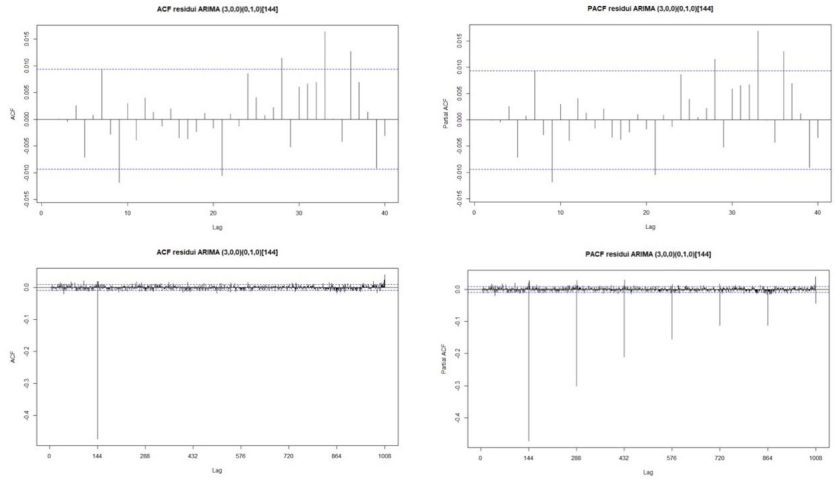

Figure 10. ACF e PACF residui ARIMA (3,0,0)(0,1,0)[144]

Dunque, dopo aver sperimentato altri modelli e aver notato che questo problema persisteva, si è tentato un approccio differente che permettesse di modellare diversamente la stagionalità e eliminare la stagionalità giornaliera, modellando efficacemente la stagionalità settimanale. Sono state quindi utilizzate (invece di una sola serie storica) 144 sotto-serie, ovvero una per ogni 10 minuti di ogni ora; in questo modo si ha una serie contenente i valori corrispondenti a ora 0 - minuto 0, una corrispondente a ora 0 - minuto 10, e così via fino ad ottenere la serie dell'ora 23 - minuto 50.

Queste 144 serie sono state passate in input ai modelli Arima, sono stati studiati gli autocorrelogrammi e sono state effettuate le predizioni di ognuna di queste serie, concatenandole poi all'interno di un unico oggetto di tipo dataframe; questo è stato utilizzato per calcolare il MAE (ovvero la media delle differenze tra il valore reale e il valore predetto).

Per la scelta dei modelli ARIMA il punto di partenza è stato quello di inserire una componente in grado di modellare la stagionalità settimanale, quindi il valore D (considerando Arima(p,d,q)(P,D,Q)) è stato fissato a 1 ed è stato fissato a 7 il numero di periodi. Nel corso dell'implementazione dei modelli, si è adottato un approccio progressivo, iniziando con modelli più semplici e proseguendo con modelli più sofisticati in grado di rappresentare aspetti più complessi. Nel tentativo di ottenere il miglior modello ARIMA che riuscisse a catturare molta memoria e che riuscisse a minimizzare il MAE delle previsioni, sono state eseguite diverse variazioni nei suoi parametri, ovvero:

- p: ordine di ritardo, numero di osservazioni precedenti che influenzano la previsione corrente
- i: grado di integrazione, numero di volte che la serie storica viene differenziata per renderla stazionaria
- q: ordine di media mobile, dimensione della finestra della media mobile utilizzata per eliminare la tendenza non stazionaria dalla serie storica
- P: l'ordine di autoregressione della componente stagionale
- I: il grado di integrazione della componente stagionale (impostato a 1)

- Q: l'ordine di media mobile della componente stagionale

Il primo modello utilizzato è stato un modello che includesse le informazioni sulla stagionalità viste precedentemente; dunque è stato creato un modello con una sola componente di differenziazione stagionale con periodo 7, ovvero

$$ARIMA(0,0,0)(0,1,0)[7] \tag{2}$$

Questo modello però è molto semplice, infatti osservando gli autocorrelogrammi (in Figura 11 sono mostrati quelli di una delle 144 serie) si può notare che resta un'evidente stagionalità ogni 7 osservazioni, ovvero ogni settimana.

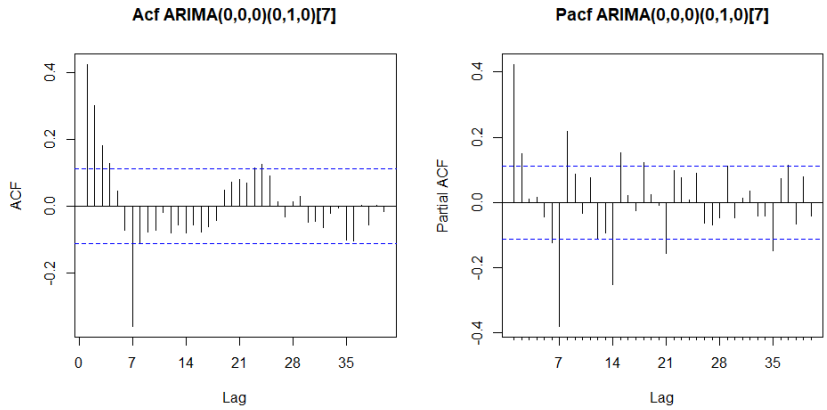

Figure 11. ACF e PACF residui ARIMA(0,0,0)(0,1,0)[7]

Il validation MAE di questo modello nonostante tutto diminuisce rispetto al modello precedente, infatti risulta pari a 1566.

Data la struttura regolare e persistente della stagionalità potrebbe essere adeguato inserire una componente autoregressiva; nel caso di un modello

$$ARIMA(1,0,0)(0,1,0)[7] \tag{3}$$

il termine di auto-regressione (AR(1)) potrebbe catturare eventuali tendenze presenti nella serie storica, consentendo una migliore descrizione della stagionalità. All'interno dei correlogrammi non si notano però delle differenze effettive, e nemmeno per quanto riguarda il validation MAE sono stati ottenuti dei netti miglioramenti (in questo caso ha un valore di 1317). Aumentando l'ordine della componente autoregressiva il Mean Absolute Error si abbassa fino a un valore di 1142 con il modello

$$ARIMA(4,0,0)(0,1,0)[7] \tag{4}$$

Tuttavia resta evidente la componente di stagionalità settimanale nel grafico dell'autocorrelazione parziale (Figura 12).

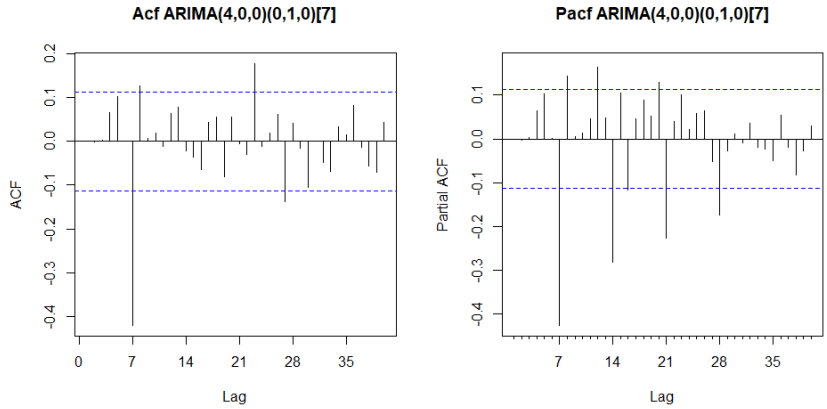

Figure 12. ACF e PACF residui ARIMA(4,0,0)(0,1,0)[7]

Si è passati a un modello più complesso a causa dei risultati insufficienti ottenuti con il modello precedente. Il modello

$$ARIMA(1,1,0)(0,1,1)[7] \tag{5}$$

è stato scelto perché utilizza sia la differenziazione delle osservazioni grezze che una finestra di media mobile per modellare la serie temporale. La differenziazione delle osservazioni rende la serie più stabile e facilmente prevedibile, mentre l'utilizzo della finestra di media mobile consente di modellare la memoria residua nella serie. Osservando i residui (di una delle serie) si può notare una migliore memoria lineare del modello, infatti è notevolmente diminuito il numero di lag al di fuori delle bande di confidenza; inoltre la stagionalità settimanale sembra essere colta dal modello (in Figura 13 non ci sono più i picchi significativi ai lag con valore multiplo di 7).

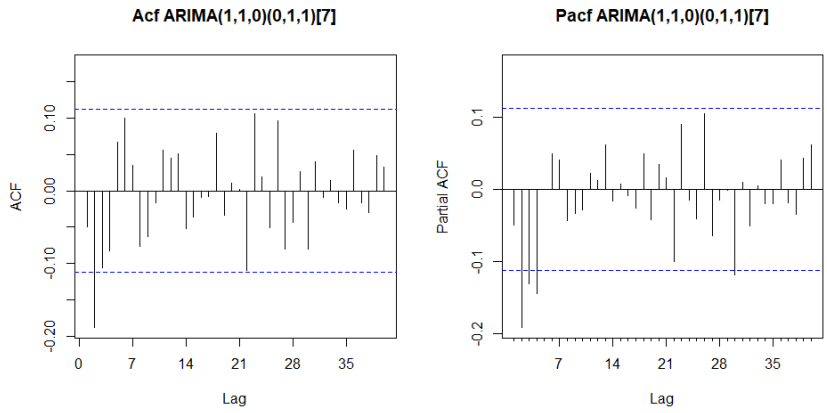

Figure 13. ACF e PACF residui ARIMA(1,1,0)(0,1,1)[7]

Il MAE ottenuto sul validation set da questo modello diminuisce ulteriormente, raggiungendo un valore di 1014.

Come ultimo tentativo per cercare di ottenere dei residui ancora più puliti e migliorare la capacità di previsione del modello (e di conseguenza avere anche un errore minore) è stato aumentato l'ordine auto-regressivo del modello (da 1 a 2) ed è stata inserita una media mobile di ordine maggiore ripetto al modello precedente (anche in questo caso ordine 2); questo può migliorare la capacità del modello di previsione, in quanto può catturare tendenze più lunghe e una maggiore memoria storica dei dati.

Effettivamente con il modello

$$ARIMA(2,1,0)(0,1,2)[7] \tag{6}$$

ottiene dei residui puliti nella maggior parte delle serie (come mostrato in Figura 14) ed è il modello che ottiene le predizioni sul validation che portano all'errore minore, infatti il validation MAE con questo modello risulta 960.

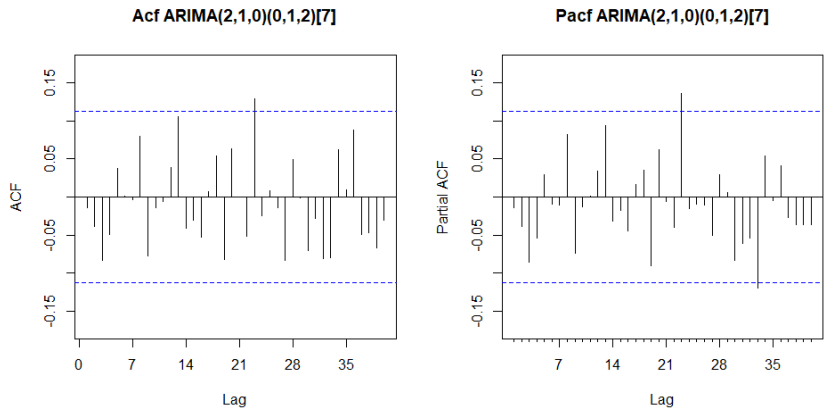

Figure 14. ACF e PACF residui ARIMA(2,1,0)(0,1,2)[7]

5 Modelli UCM

Gli Unobserved Components Models (UCM) sono una classe di modelli statistici utilizzati per l'analisi delle serie temporali. Questi modelli si basano su una decomposizione della serie temporale in componenti osservabili e non osservabili. La componente osservabile corrisponde ai dati stessi, mentre le componenti non osservabili includono fattori come tendenza, ciclo, stagionalità e così via.

Gli UCM sono uno strumento molto utile per la comprensione della struttura sottostante delle serie temporali e per la previsione dei valori futuri. Inoltre, gli UCM sono molto flessibili e possono essere adattati a molteplici situazioni e problemi specifici, rendendoli uno strumento molto potente per molte applicazioni. In generale, gli UCM sono uno strumento essenziale per chi lavora con le serie temporali e desidera comprenderne la struttura sottostante e prevedere i valori futuri.

Come per i modelli ARIMA, anche nei modelli UCM sono state utilizzate 144 serie usando come periodo di training quello dall'1 Gennaio fino al 31 Ottobre. Sono stati testati diversi modelli con il fine di ottenere il modello migliore nell'adattarsi ai dati; è stata inizialmente considerata la componente di level che modella il trend della serie, dunque diversi valori del parametro level sono stati testati e hanno portato a diversi risultati, alcuni più soddisfacenti di altri. I diversi valori del parametro level che sono stati considerati sono: 'ntrend' (No trend), 'dconstant' (Deterministic constant), 'llevel' (Local level), 'rwalk' (Random walk), 'dtrend' (Deterministic trend), 'lldtrend' (Local linear deterministic trend), 'rwdrift' (Random walk with drift), 'lltrend' (Local linear trend), 'strend' (Smooth trend), 'rtrend' (Random trend). Sono state calcolate le previsioni per ognuna delle serie utilizzando una componente seasonal = 7 che indica che il modello sta considerando una stagionalità di periodo 7 e utilizzando le diverse tipologie di level citate sopra; il tipo di livello che ha permesso di ottenere il Mean Absolute Error minore è stato il Random Walk (come mostrato nella Tabella 1) che è stato scelto anche per la sua notevole capacità di adattamento ai dati (infatti è una delle componenti con AIC minore). La scelta del modello di Random Walk potrebbe essere motivata dalla sua capacità di rappresentare la tendenza di una serie temporale come basata sul valore più recente osservato.

Table 1. Componenti trend modelli UCM

Level	AIC	MAE
ntrend	7065.98	29003.28
dconstant	6640.01	3629.52
llevel	5117.63	1432.67
rwalk	5259.08	1143.69
dtrend	6426.63	9061.89
lldtrend	5123.63	1660.64
rwdrift	5187.76	1341.19
lltrend	5119.58	2312.34
strend	5137.54	3470.13
rtrend	5310.98	5241.69

Una volta individuata la componente di level che meglio descrive l'andamento delle serie, a questa sono state aggiunte delle armoniche per modellare la componente stagionale settimanale (poichè anche in questo caso la componente giornaliera è stata eliminata usando 144 serie). La scelta è ricaduta sulle dummy sinusoidali in quanto consentono di catturare più facilmente la forma ciclica della stagionalità, quindi nei casi in cui la stagionalità è regolare è preferibile usare questo tipo di dummy; è stata perciò inserita una componente 'freq_seasonal' all'interno del modello con period = 7 (stagionalità settimanale appunto) e numero di armoniche variabile. Come per la

Table 2. Componenti trend modelli UCM

Level	Numero sinusoidi	AIC	MAE
rwalk	1	5166.51	1115.49
rwalk	2	5122.29	1127.44
rwalk	3	5086.49	1136.93
rwalk	4	5036.82	1123.23
rwalk	5	5004.50	1144.87
rwalk	6	4967.93	1129.62
rwalk	7	4917.37	1105.00
rwalk	8	4872.68	1101.44
rwalk	9	4838.01	1096.33
rwalk	10	4805.05	1098.63
rwalk	11	4769.04	1093.78
rwalk	12	4734.95	1096.41
rwalk	13	4700.05	1098.94
rwalk	14	4659.68	1116.77
rwalk	15	4624.71	1132.91
rwalk	16	4591.41	1134.69

scelta della componente level, anche il numero di armoniche è stato scelto combinando due fattori; infatti il numero ottimale è quello che minimizza il criterio informativo (che permette di misurare la qualità di adattamento del modello ai dati considerando sia la precisione della previsione che la complessità del modello stesso) e minimizza il Mean Absolute Error delle previsioni. É stato effettuato un ciclo che permette di selezionare il numero ottimale di armoniche considerando un range da 1 a 16, come mostrato nella Tabella 2.

Il numero di sinuoidi ottimale risulta quindi essere 11 (in quanto minimizza il MAE e ottiene uno tra i valori più bassi in assoluto dell'AIC). In conclusione, il modello UCM scelto è quello con level pari a Random Walk e con la componente stagionale modellata tramite 11 sinusoidi di periodo 7. Il MAE ottenuto dalle previsioni effettuate tramite questo modello sul validation set di Novembre è di 1093.

6 Modelli Machine Learning

L'ultima tecnica utilizzata nella previsione della serie è stata quella dei modelli di Machine Learning. Sono stati sperimentati diversi approcci, tra cui tecniche di Deep Learning e modelli di Machine Learning, sia utilizzando la serie originale che le 144 serie descritte in precedenza. Questi modelli sono utili in quanto non necessitano di una modellizzazione preventiva della stagionalità presente nei dati, poiché questa può essere acquisita dal modello durante la fase di addestramento; quindi non è necessario definire in anticipo la stagionalità nei dati (ad esempio, la presenza di una stagionalità settimanale), poiché il modello è in grado di rilevarla durante il processo di training.

Relativamente al pre-processing, ovvero la scelta di quale serie utilizzare, i risultati non sono molto differenti tra le due opzioni. Pertanto, per semplicità, si è optato per l'utilizzo della serie originale.

Per quanto riguarda la scelta tra le tecniche di Machine Learning e quelle di Deep Learning, i risultati ottenuti sono stati confrontati e i modelli di Machine Learning sono risultati decisamente più performanti; questo potrebbe essere dovuto a diversi fattori quali:

- i modelli di Machine Learning sono più semplici e più veloci da addestrare rispetto a quelli di Deep Learning, che richiedono molte più risorse computazionali e molto più tempo per l'addestramento
- i modelli di Machine Learning sono meno propensi a overfitting, ovvero a essere troppo adattati ai dati di addestramento e meno generalizzabili a nuovi dati; questo è un problema comune nelle reti neurali, che sono molto flessibili e possono facilmente memorizzare i dati di addestramento
- l'errore nei modelli di Machine Learning si propaga meno rispetto a quello nei modelli di Deep Learning; nelle reti neurali, l'errore accumulato a ogni passo della previsione può diventare molto grande, rendendo le previsioni meno affidabili

Perciò vengono riportati solo i risultati dei modelli di Machine Learning applicati all'intera serie originale.

Inizialmente è stato utilizzato il dataset sottoposto a pre-processing, quindi il dataset che comprende le colonne con le informazioni estratte dalla data (come ad esempio hour, minute, dayofweek, day_name, dayofyear, month, year, ecc.) che sono state usate come regressori. Siccome i risultati ottenuti non sono stati considerati del tutto soddisfacenti, si è deciso di inserire ulteriori regressori, in particolare i più influenti come la notte, il ramadan, le festività marocchine e le stagioni.

Per selezionare i modelli di Machine Learning più adatti alla previsione di una serie storica ad alta frequenza come quella a disposizione ed implementarli all'interno del progetto, è stata effettuata una ricerca; uno dei risultati ottenuti è stato un notebook su Kaggle che effettua previsioni su una serie riguardante consumi energetici e individua come migliori modelli quelli mostrati nella Figura 15.

Come primo modello si è dunque deciso di implementare un Extra Trees Regressor.
Extra Trees Regressor è una tecnica di apprendimento di insiemi basata su alberi di decisione che viene utilizzata per risolvere problemi di regressione. In questo metodo, vengono addestrati e combinati più alberi di decisione per ottenere una previsione finale. Il processo di formazione di ciascun albero è basato su un sottoinsieme casuale delle caratteristiche, il che rende la tecnica meno suscettibile all'overfitting rispetto a un albero di decisione convenzionale. La previsione finale viene ottenuta mediante la media delle previsioni di ogni albero individuale.

Come prima operazione, viene effettuato un fine-tuning sui parametri da passare al modello in modo da regolare queste variabili per far sì che il modello ottenga le prestazioni migliori.

Model	MAE	MSE
Extra Trees Regressor	1402.2131	5033280.8305
Random Forest Regressor	1682.1848	6622596.4779
CatBoost Regressor	1951.5235	7231201.3203
Extreme Gradient Boosting	2019.3625	8021253.8500
Light Gradient Boosting Machine	2382.0319	10534556.9617
Decision Tree Regressor	2139.7482	12583693.4651
Gradient Boosting Regressor	4354.3410	31812612.1472
Ridge Regression	4394.6535	35585153.6000
Bayesian Ridge	4394.7324	35584881.6094
Lasso Regression	4393.8217	35585706.4000
Linear Regression	4395.6196	35587004.8000
Lasso Least Angle Regression	4480.4321	37195026.0016
Huber Regressor	6360.5166	67443876.0526
K Neighbors Regressor	6449.6787	90638036.8000
Orthogonal Matching Pursuit	8677.1907	119895490.5348
AdaBoost Regressor	10744.2761	163256804.6956
Elastic Net	11268.2127	190573809.6000
Passive Aggressive Regressor	9376.0177	199127538.8709
Dummy Regressor	13940.2672	295698883.2000
Least Angle Regression	293746963.5823	1137856085495628672.0000

Figure 15. Modelli ML performanti

Addestrando questo modello sul periodo Gennaio-Ottobre ed effettuando le previsioni su Novembre però si può notare che, contrariamente a quanto evidenziato nella Figura 15, il MAE è parecchio alto, infatti ha un valore di circa 3031. Per comprendere l'importanza dei regressori inseriti si analizza la feature importance di ognuno di essi e si rappresenta graficamente all'interno di un grafico (quello in Figura 16).

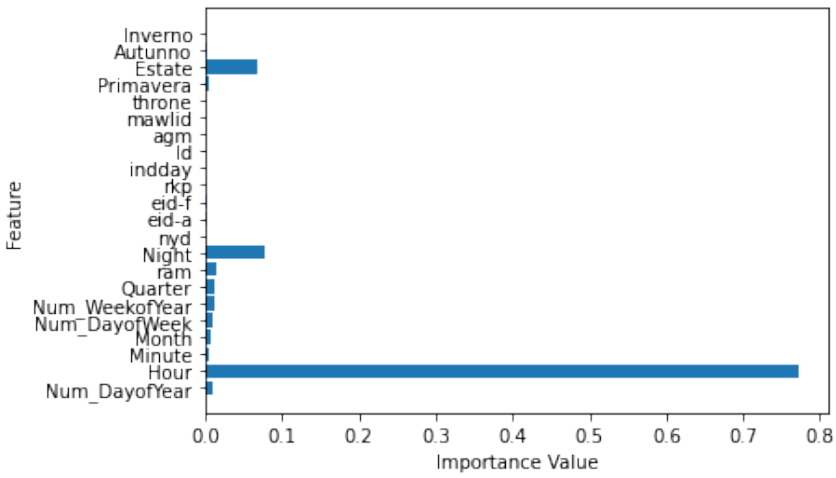

Figure 16. Feature importance

Si decide quindi di mantenere solo i regressori più importanti, ovvero quelli estratti dalla data (hour, Num_dayofweek, quarter, ecc.), notte (periodo compreso tra le ore 21 e le 6), ramadan (periodo che va dal 26 Maggio al 24 Giugno), stagioni (primavera e estate) e una delle festività (Eid al-Fitr, dal 25 al 27 di Giugno). Il MAE di questo nuovo modello rimane pressochè lo stesso.

Visti i risultati non molto buoni del primo algoritmo si è deciso di implementare un ulteriore tecnica di Machine Learning, ovvero Gradient Boosting.
Gradient Boosting è una tecnica di apprendimento di insiemi basata su alberi di decisione utilizzata per risolvere problemi di regressione e classificazione. Questa tecnica prevede l'addestramento sequenziale di più alberi di decisione, in cui ogni passo mira a correggere l'errore di previsione del passo precedente mediante l'aggiunta di un nuovo albero che si adatta al gradiente negativo della funzione di perdita.

Anche su questo modello viene effettuato un fine-tuning e successivamente vengono effettuate le previsioni sul modello con i parametri ottimali; il MAE ottenuto sulle predizioni del mese di Novembre è molto più basso rispetto a quello della tecnica precedente, infatti risulta 1272.

Per provare a migliorare ulteriormente le previsioni è stato implementato un ulteriore modello (tra quelli più performanti della Figura 15): CatBoost.
Questo è un algoritmo di machine learning che usa gradient boosting su decision tree progettato specificamente per le caratteristiche categoriche; queste caratteristiche sono quelle che assumono un numero limitato di valori, ad esempio genere o tipo di prodotto. CatBoost gestisce in modo nativo questo tipo di caratteristiche e fornisce risultati superiori rispetto ad altri algoritmi di gradient boosting. Inoltre, CatBoost include anche meccanismi di regolarizzazione integrati per prevenire l'overfitting, rendendolo una scelta popolare per molte applicazioni reali.

La procedura seguita per implementare questo algoritmo è la stessa dei precedenti: fine tuning sul modello e previsioni su Novembre che ottengono un MAE di 1393.

Dunque il modello migliore risulta essere Gradient Boosting; questo è il modello che viene usato per effettuare le previsioni su Dicembre.

7 Conclusioni

In generale, l'analisi delle serie storiche sui consumi energetici ad alta frequenza rappresenta un'opportunità per comprendere meglio i trend e le dinamiche dei consumi energetici in un determinato contesto. La modellizzazione e la previsione di tali serie storiche possono aiutare a identificare i fattori che influiscono sui consumi energetici. Essendo la serie a disposizione una serie ad alta frequenza però è molto difficile effettuare delle previsioni accurate. Sono stati utilizzati tre diversi metodi per analizzare e prevedere i consumi energetici ad alta frequenza: ARIMA, UCM e ML; i risultati delle previsioni su un periodo di validation (ovvero quello che va dall'1 Novembre al 30 Novembre) è mostrato nella Figura 17 e il MAE relativo a ogni modello è riassunto nella Tabella 3.

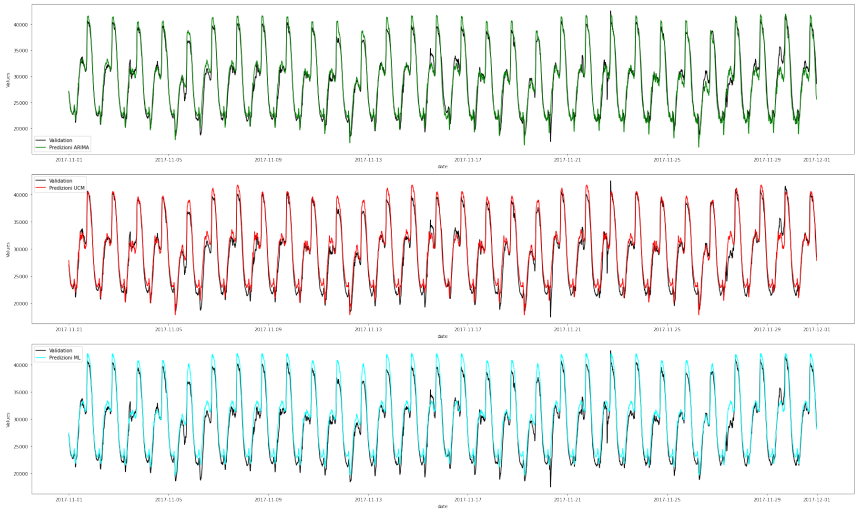

Figure 17. Previsioni sul validation set

Table 3. MAE sul validation set

Modello	Modello migliore	MAE
ARIMA	ARIMA(2,1,0)(0,1,2)[7]	960.37
UCM	Random Walk con 11 sinusoidi di periodo 7	1093.78
ML	Gradient Boosting	1272.49

Le previsioni di dicembre sono invece rappresentate graficamente e mostrate nella Figura 18.

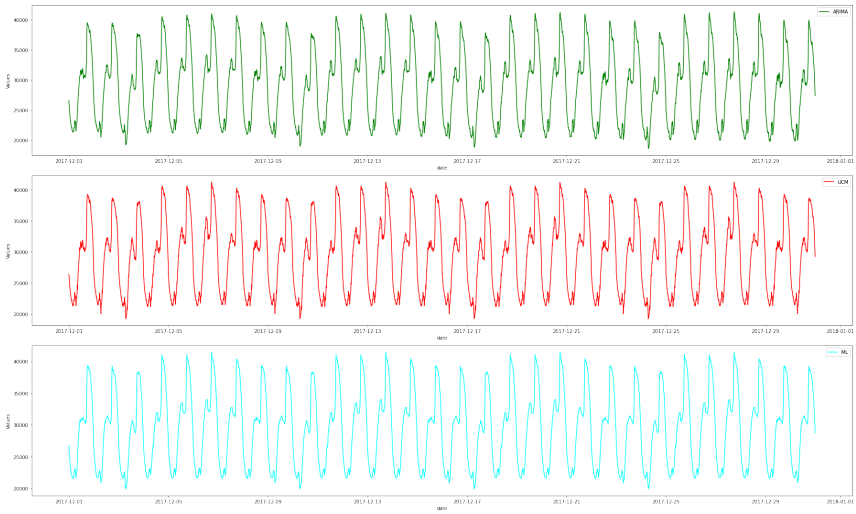

Figure 18. Previsioni sul test set

I risultati ottenuti in questo progetto dimostrano l'importanza di sperimentare diversi metodi per analizzare e prevedere serie storiche complesse, in quanto ognuno può fornire informazioni uniche e valide che possono essere utilizzate per prendere decisioni informate.