Università degli studi di
Milano-Bicocca

Streaming Data Management and Time Series Analysis

Final Project

Previsione dei prezzi del mercato energetico

Authors:
Federica Fiorentini - 807124 - f.fiorentini1@campus.unimib.it
July 6, 2020

1 Introduzione

Il progetto ha come obiettivo la definizione, lo sviluppo e l'implementazione di diversi algoritmi con lo scopo di effettuare una previsione dei prezzi del mercato energetico. In particolare, vengono sviluppati diversi modelli appartenenti a 3 categorie:

- Modelli **ARIMA** *(AutoRegressive Integrated Moving Average);*
- Modelli **UCM** *(modelli a componenti non osservabili);*
- Modelli **non-lineari** *(ML).*

Per realizzare la previsione viene utilizzato un Dataset messo a disposizione che include il valore dei prezzi dell'energia elettrica aggregati a livello giornaliero. I dati si riferiscono ad un periodo di 8 anni, dal 1° Gennaio 2010 al 31 Dicembre 2018. Il file messo a disposizione è costituito semplicemente da due colonne:

- Data [yyyy-mm-dd]
- Prezzo [euro]

Per l'analisi, il dataset viene suddiviso in training e validation set considerando rispettivamente il periodo dal 1° Gennaio 2010 al 30° Giugno 2017 e l'anno e mezzo restante come validation. Il criterio tramite il quale è stato suddiviso il dataset nelle due porzioni corrispondenti è basato sul fatto che una serie storica riferita al prezzo può contenere stagionalità intra annue e, valutando le performance del modello, è necessario verificare la bontà in tutti i periodi dell'anno per assicurarsi che sia stata modellata correttamente la componente stagionale. Di seguito *(figura 1)* viene rappresentato la suddivisione dei dati in train e validation set.

L'obiettivo dell'analisi consiste nel prevedere l'andamento della serie temporale, ovvero l'andamento del prezzo giornaliero dell'energia elettrica, per quanto riguarda parte dell'anno successivo, dal 1° Gennaio 2019 al 30 Novembre 2019.

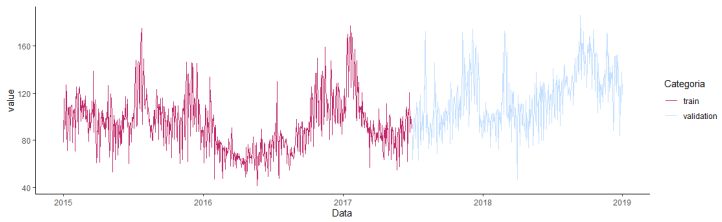

Figure 1: Suddivisione della serie storica in Train e Validation set.

2 Approccio metodologico

Come anticipato nella sezione precedente, sono stati sviluppati alcuni modelli per realizzare la previsione. Di seguito la descrizione dei singoli approcci, la metodologia di sviluppo affrontata e le scelte considerate per procedere.

Inizialmente è stata effettuata un'analisi qualitativa dei dati, al fine di indagare riguardo la presenza di un andamento crescente/decrescente nella serie, di eventuali picchi negativi o positivi e di comportamenti anomali durante alcuni giorni.

In *figura 2* viene rappresentato l'andamento della serie storica.

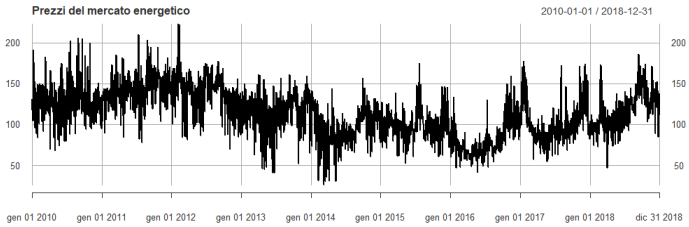

Figure 2: Andamento dell'intera serie storica

Da una prima visualizzazione della serie storica emerge una possibile stagionalità settimanale, abbastanza prevedibile se si pensa al mercato di riferimento dei dati. Il processo, inoltre, sembra essere stazionario in varianza poiché, escludendo alcuni picchi, non si presentano particolari trend crescenti o decrescenti. I picchi potrebbero risultare in corrispondenza di festività o del weekend, come detto in precedenza, oppure in presenza di outlier.

In *figura 3*, invece, viene rappresentato l'andamento della media dei valori aggregati per anno.

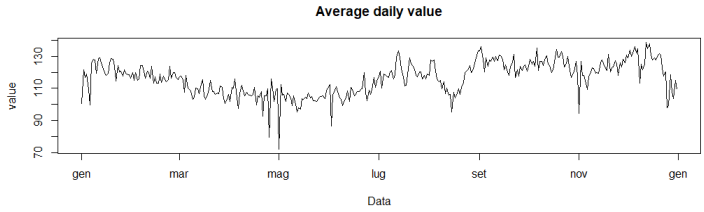

Figure 3: Serie storica aggregata

Da questo grafico si evince che sono presenti alcuni picchi negativi, come detto in precedenza, forse in corrispondenza delle festività.

Infine, in *figura 4* vengono rappresentati i boxplot che mostrano l'aggregazione della serie per giorno della settimana.

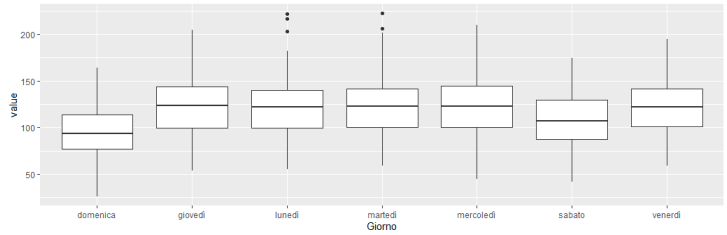

Figure 4: Boxplot dei valori della serie aggregati per giorno della settimana

Anche questo grafico conferma la differenza del consumo di energia elettrica tra i giorni della settimana. Mediamente, infatti, la domenica si rileva un consumo più basso così come il sabato, a differenza degli altri giorni in cui all'incirca mantiene lo stesso livello. Come evidenziano i boxplot, inoltre, sono presenti diversi outlier, precisamente 26 osservazioni pari allo 0.6% dell'intero dataset. I valori, quindi, sono stati sostituiti e calcolati tramite l'interpolazione lineare della serie.

2.1 Modelli ARIMA

Per stimare i coefficienti del modello ARIMA è stata seguita la procedura di Box e Jenkins che prevede, inizialmente, l'analisi dei correlogrammi della serie (rappresentati in *figura 5 e figura 6*) per stimare i coefficienti stagionali e non sia della parte autoregressiva che quella a media mobile.

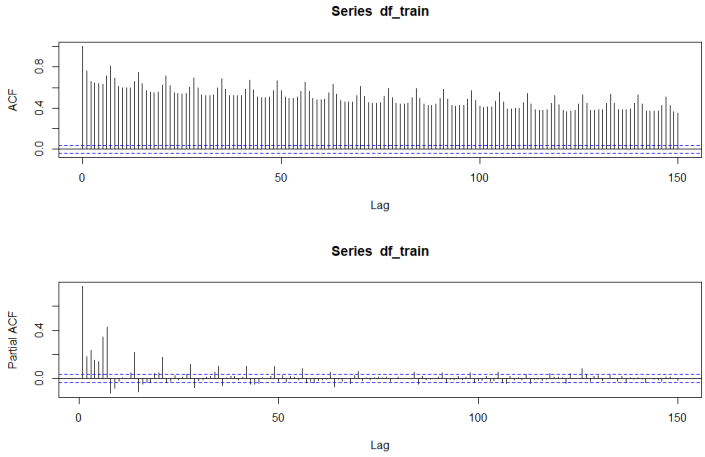

Figure 5: ACF e PACF della serie

Osservando i correlogrammi della serie, si osserva che l'ACF tende a zero molto lentamente. Si evince, inoltre, l'evidenza di picchi con cadenza settimanale e questo va a confermare la presenza di una leggera *non stazionarietà in media*. Si procede, quindi, con una differenziazione settimanale e, successivamente, con l'applicazione di un primo modello $ARIMA(0;0;0)(1,1,1)_7$.

Dai correlogrammi dei residui del modello sviluppato rappresentati in *figura 7 e 8*, si nota che la PACF rimane significativa nei primi 6 ritardi. Si cerca il coefficiente della parte autoregressiva non stagionale testando 7 modelli di ordine da 0 a 6. Il modello migliore viene selezionato tramite il criterio dell'AIC, ovvero il modello che minimizza questo valore. Si preferisce non aumentare ancor di più il valore del coefficiente dell'AR (anche se in realtà al ritardo 7 rimane ancora significativo) perché, a fronte di un piccolo miglioramento del modello, si preferisce non appesantirlo con troppi coefficienti. Inoltre, la significatività del ritardo 7 potrebbe essere dovuta a rimanenze della componente stagionale.

In base al criterio dell'AIC, come rappresentato nella tabella seguente, il modello migliore risulta essere un $ARIMA(6,0,0)(1,0,1)_7$

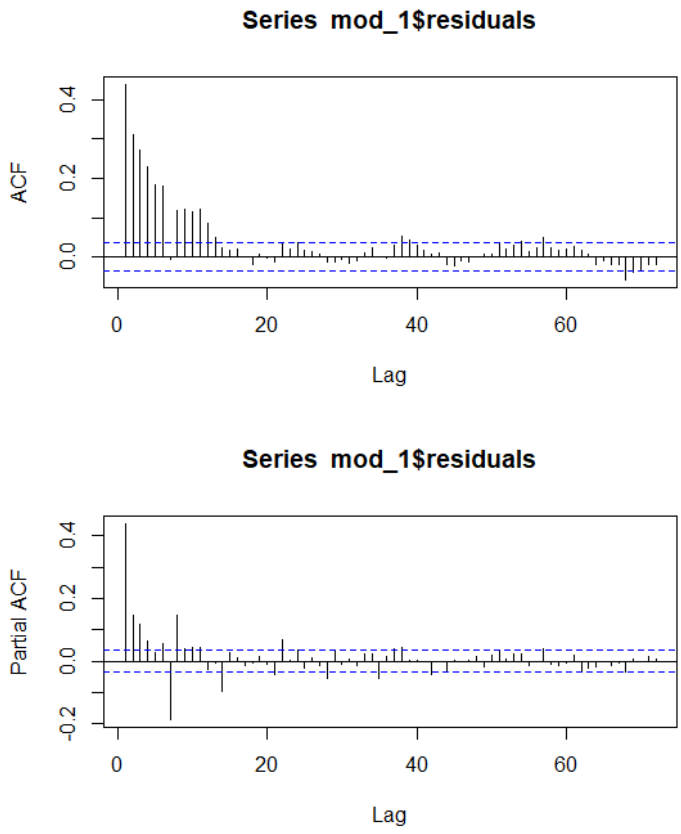

Figure 6: ACF e PACF dei residui

Modello	AIC
ARIMA(1,0,0)(1,1,1)	21998
ARIMA(2,0,0)(1,1,1)	21907
ARIMA(3,0,0)(1,1,1)	21832
ARIMA(4,0,0)(1,1,1)	21801
ARIMA(5,0,0)(1,1,1)	21792
ARIMA(6,0,0)(1,1,1)	21747

Dall'analisi del grafico delle previsioni confrontate con il validation set, emerge che il modello stimato non riesce a modellare le stagionalità intra-annue. A tal proposito, vengono aggiunti al modello ARIMA precedente inizialmente dei regressori sinusoidali, in particolare 20 serie di seni e coseni

con frequenza pari a $2\pi//365.25$, e successivamente dei regressori dummy in corrispondenza delle festività italiane (si suppone che i dati si riferiscano ad un mercato energetico italiano). Sono state considerate le seguenti festività: *Capodanno, Pasqua, 25 Aprile, 1 Maggio, 2 Giugno, Ferragosto, Tutti i Santi, Immacolata, Vigilia e Natale, Ultimo dell'anno.*

Sia da una analisi qualitativa dei grafici in *figura 9*, sia dal confronto dell'AIC dei 3 modelli ARIMA sviluppati, emerge che il modello migliore è il modello $ARIMA(6,0,0)(1,1,1)_7$ sia con i regressori sinusoidali che le dummy relative alle festività.

Modello	AIC
ARIMA(6,0,0)(1,1,1) semplice	22627
ARIMA(6,0,0)(1,1,1) con regressori sinusoidali	23556
ARIMA(6,0,0)(1,1,1) con regressori sinusoidali e dummy stocastiche	22357

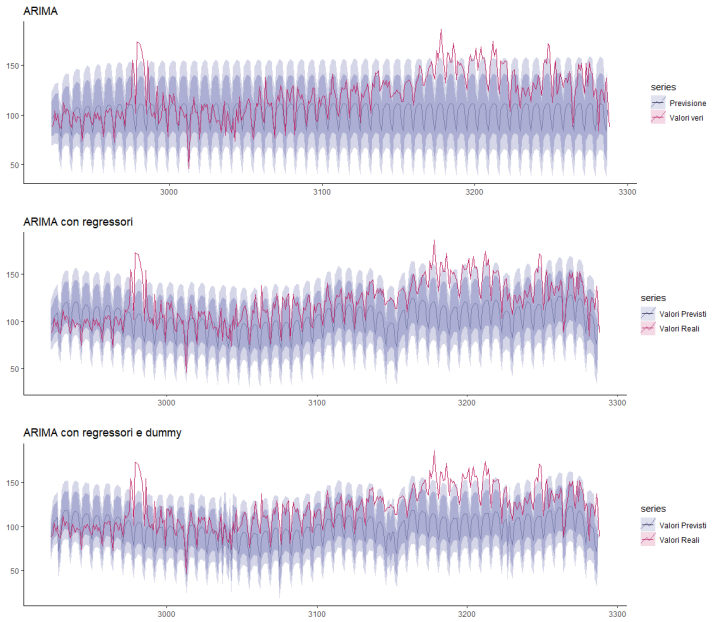

Figure 7: Previsione dei modelli sul validation set

Per concludere la procedura di Box-Jenkins è stata effettuata l'analisi dei residui del modello. Si testano le ipotesi che i residui abbiano media nulla, siano distribuiti normalmente e siano incorrelati. Dall'analisi effettuata, risulta che i residui rispettino le condizioni di normalità e nullità della

media. Tuttavia, il test di Ljung-Box non permette di accettare l'ipotesi di assenza di autocorrelazione globale. Questo perché, trattandosi di dati reali, è difficile avere dei residui che rispettino tutte le condizioni e siano White-Noise.

2.2 Modelli UCM

I modelli UCM a componenti non osservabili vengono utilizzati per combinare diverse componenti quali ciclo, trend e stagionalità.

In questo progetto, in particolare, in tutti i modelli state utilizzate delle dummy stocastiche utili a modellare la stagionalità settimanale, mentre per quanto riguarda la stagionalità intra-annua sono state utilizzate delle sinusoidi stocastiche. Per la stima del trend, invece, vengono testati tre diversi modelli aventi componenti differenti:

- il **local linear trend**;
- il **random walk**;
- l'**integrated random walk**.

In particolare, sono stati sviluppati i diversi modelli, come detto in precedenza, utilizzando le tre componenti per quanto riguarda il trend e, in più, sono state aggiunte le dummy stocastiche in corrispondenza delle vacanze solamente nel modello con il local linear trend, dato che nei modelli ARIMA avevano portato a miglioramenti nelle previsioni.

I modelli ottengono risultati piuttosto soddisfacenti in termini di MAPE, ma il migliore risulta essere quello con l'Integrated Random Walk.

2.3 Modelli non-lineari

Per quanto riguarda i modelli non-lineari, sono state utilizzate due diverse tecniche per effettuare la previsione, in particolare il **K-Nearest Neighbors** e **Recurrent Neural Network** con due diverse architetture.

2.3.1 K-Nearest Neighbors

Il primo algoritmo utilizzato è stato il kNN, ovvero le k serie più simili all'ultimo lag temporale che precede i valori da prevedere. In particolare, in questo caso è stato utilizzato come lag temporale un anno di dati. Una

volta individuate le k serie più simili basandosi sulla distanza euclidea, i 334 valori futuri sono previsti tramite una media dei 334 valori che succedono le k serie identificate e nell'effettuare questa media si è scelto di pesare maggiormente le serie più recenti tra le k identificate. Inoltre, è stato utilizzato il metodo ricursivo in modo da avere una previsione one-step-ahead, poiché in ogni iterazione vengono considerati, non solo tutti i dati della serie, ma anche i dati di previsione generati

fino a quell'iterazione.

Per definire il valore di k sono stati implementati diversi modelli con k differenti appartenenti ad un range da 5 a 100, con un salto di 5. In particolare, il numero di Neighbors k migliore in termini di MAPE del modello risulta essere 35.

2.3.2 Recurrent Neural Network

Sempre per quanto concerne i modelli non lineari, sono state sviluppate due Reti Neurali Ricorrenti aventi due diverse achitetture, l'LSTM (*Long Short-Term Memory*) e la GRU (*Gated Recurrent Unit*). Entrambi gli algoritmi sono stati scelti poiché permettono di conservare le informazioni sul passato analizzando i dati in maniera sequenziale.

Prima di sviluppare il modello, è stato necessario scalare e centrare i dati e, successivamente, trasformarli sottoforma di array con 3 dimensioni: la numerosità del campione, il numero di lag e il numero di features. In questo caso rispettivamente le ultime due misure sono state poste pari a 1.

La prima rete ricorrente è stata costruita con 2 layers LSTM rispettivamente di 100 e 90 neuroni con funzione di attivazione tangente iperbolica. E'stato inserito del dropout (0.3) per evitare l'overfitting e si è terminato con uno strato denso con funzione di attivazione lineare. La seconda architettura ha una struttura simile alla precedente con la differenza che i due layer LSTM sono stati sostituiti da un layer GRU di 90 neuroni e funzione di attivazione tangente iperbolica. I modelli sono stati trainati per 300 epoche con una batch size pari a 365. Come ottimizzatore è stato unitilizzato Adam con un learning rate pari a 0.001 e come funzione di perdita il mean absolute error.

I risultati di entrambe le reti sono stati soddisfacenti, ma l'architettura GRU ha raggiunto performance migliori in termini di MAPE, come vedremo nel capitolo successivo.

3 Conclusioni

Per valutare da un punto di vista quantitativo le performance dei diversi modelli sviluppati, è stata utilizzata la metrica MAPE calcolata sul validation set. Tale indice è stato scelto principalmente perché la serie storica non presenta alcun valore prossimo allo zero; in questo caso, infatti, non sarebbe possibile utilizzare questa metrica poiché potrebbe assumere valore molti elevati anche se il modello ha buone performance.

Dal momento che gli algoritmi di ML sviluppati sono allenati per prevedere 334 periodi in avanti, per poter rendere confrontabili i vari modelli, si è deciso di calcolare con lo stesso criterio il MAPE anche per i modelli ARIMA e UCM. Sostanzialmente, le 31 serie storiche sul quale si basa la valutazione dei modelli sono il risultato di una operazione di sliding window in cui viene fatta scorrere una finestra temporale di 334 elementi sequenzialmente sul 2018.

Nella tabella seguente, vengono mostrati i valori del MAPE raggiunti dal modello ARIMA sviluppato sia in fase di training che validation:

	ARIMA con regressori e dummy
Train	18.78
Validation	8.8

Si ottiene che il modello $ARIMA(6,0,0)(1,1,1)_7$ con regressori sinusoidali e dummy in corrispondenza delle festività, ottine delle performance piuttosto soddisfacenti, anche in fase di validation. Successivamente verrà confrontato con le altre tipologie di modelli.

Per quanto riguarda i modelli UCM, di seguito le performance raggiunte dai tre modelli.

	LLT (reg)	LLT	RW	IRW
Train	7.11	7.49	7.6	7.49
Validation	19.06	18.68	17.27	16.74

La differenza tra i modelli UCM stimati consiste, quindi, nella stima della componente trend e nell'utilizzo di regressori dummy esterni rappresentanti le festività. Tra questi 4 modelli ottiene la performance migliore quello in cui è stato utilizzato l'integrated random walk per stimare il trend stocastico.

I modelli di Machine Learning, infine, hanno raggiunto i seguenti risultati:

	kNN	LSTM	GRU
Validation	16.95	18.04	16.19

La rete neurale ricorrente GRU con un solo layer ottiene buoni risultati rispetto agli altri modelli di Machine Learning.

Quindi, dopo aver selezionato il modello migliore nelle tre categorie ARIMA,

UCM e ML, questi sono stati allenati nuovamente considerando non più solamente il training set ma la serie intera. Si riportano, quindi, le tre serie di previsioni ottenute sul test set, ovvero per il periodo che va dal 2019-01-01 al 2019-11-30.

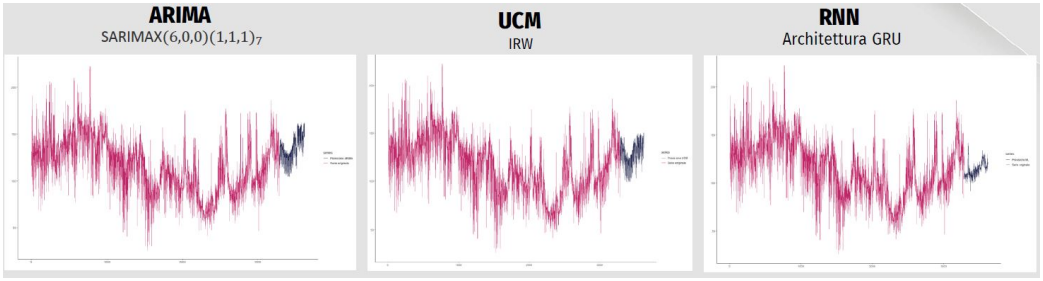

Figure 8: Previsioni