Università degli studi di
Milano-Bicocca

Streaming Data Management and Time Series Analysis

Final Project

Previsione del prezzo giornaliero di energia elettrica

Authors:
Chiaretti Giulia - 800928 - g.chiaretti@campus.unimib.it
Giugno 2020

1 Introduzione

L'obiettivo del progetto è svilupare un sistema in grado di predire i prezzi giornalieri del mercato energetico. Più nello specifico, vengono implementati diverse tipologie di algoritmi: i modelli lineari ARIMA e UCM e i modelli non lineari kNN, LSTM e GRU. Verranno analizzate, discusse e confrontate le performance di tutti i gli algoritmi sviluppati al fine di determinare il modello migliore per il fine predittivo dell'analisi.

2 Dataset

Il dataset utilizzato è costituito dalla serie storica giornaliera riferita al prezzo dell'energia elettrica. I dati sono relativi al periodo che va dal 1 gennaio 2010 al 31 dicembre 2018, per un totale di 3287 osservazioni. L'obiettivo è prevedere i valori giornalieri degli 11 mesi successivi: dal 1 gennaio 2019 al 30 novembre 2019. Per l'analisi il dataset è stato diviso in training e validation set. Si è scelto di considerare come validation set gli ultimi due anni di dati a disposizione in modo da riuscire a testare e visualizzare anche le stagionalità intra-annue che caratterizzano gli andamenti dei prezzi. La suddivisione del dataset viene rappresentata in Figura 1.

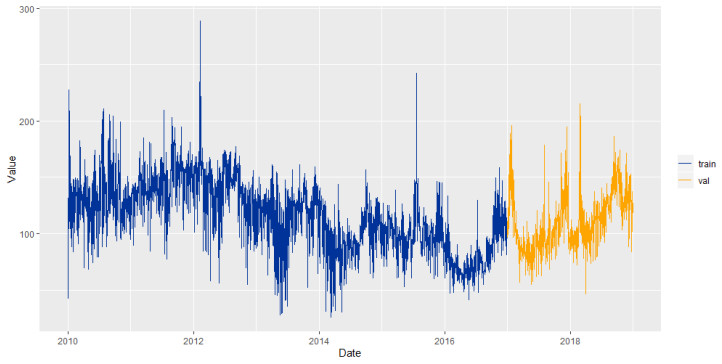

Figure 1: Serie storica divisa per train e test

Emerge la presenza di alcuni picchi che potrebbero essere considerati outlier ma in questo caso si preferisce non trattarli come tali, quindi eliminarli o sostituirli con valori standard, in quanto potrebbero rappresentare degli eventi significativi.

Dal grafico è inoltre evidente la presenza di una stagionalità probabilmente settimanale, evidenza che aumenta riducendo l'intervallo di date visualizzate in ascissa.

La presenza di una stagionalità settimanale è confermata anche da una prima analisi descrittiva del livello medio di prezzo dell'energia elettrica differenziato per giorno della settimana. In Figura 2 è possibile vedere che mediamente il prezzo nei weekend, ma in particolar modo di domenica, risulta essere più basso rispetto ai giorni infrasettimanali.

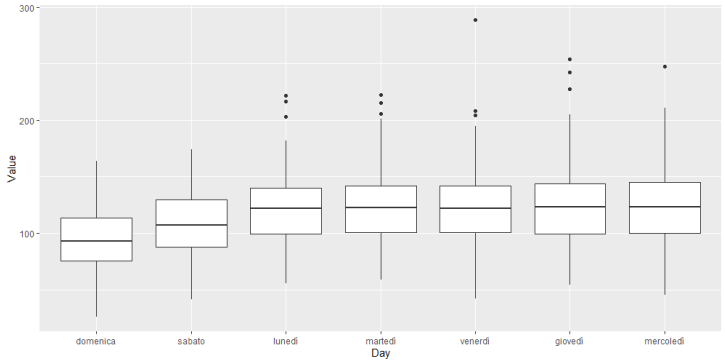

Figure 2: Prezzo per giorno della settimana

3 Approccio Metodologico

Come precedentemente descritto, nel progetto sono stati sviluppati modelli differenti. Verranno presentati prima i modelli statistici lineari, quali ARIMA e UCM, e successivamente gli algoritmi di Machine Learning.

Per confrontare le performance predittive dei diversi modelli sviluppati è stato utilizzato come criterio il Mean Absolute Percentage Error (MAPE), dato dalla media aritmetica dei rapporti tra il valore assoluto degli errori di previsione e il valore reale. Proprio per come è costruito questo indice ha il vantaggio di essere facilmente interpretabile in quanto fornisce una valutazione immediata dell'impatto dell'accuratezza previsionale, che in termini assoluti non sarebbe di così facile interpretazione. Si specifica anche che è stato possibile utilizzare il MAPE come criterio di confronto in quanto la serie storica non presenta nessun valore prossimo allo zero. In caso di serie

caratterizzate da valori molto bassi, infatti, il MAPE per costruzione assume valori molto elevati anche per i modelli che in realtà risultano essere buoni.

3.1 ARIMA

Al fine di identificare i processi autoregressivi e a media mobile per i modelli ARIMA, è stata seguita la procedura di Box-Jenkins. Non avendo rilevato una particolare non stazionarietà in varianza si è effettuata una prima analisi dei correlogrammi della serie, riportati in Figura 3.

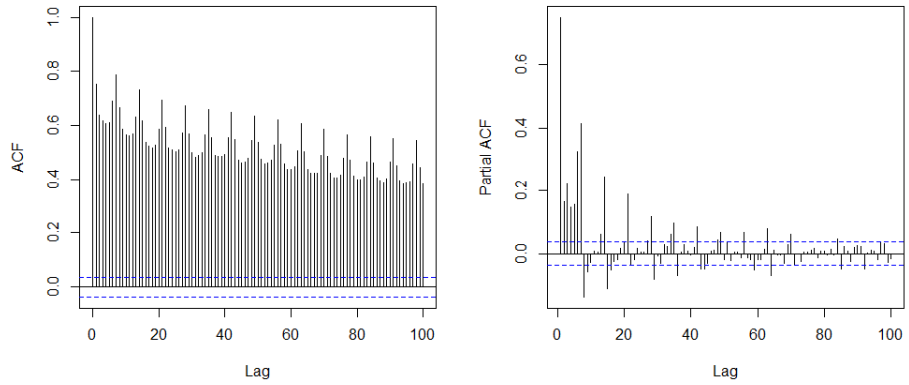

Figure 3: Correlogrammi della serie originale

Ci si concentra inizialmente sulla componente stagionale e, da una prima analisi sono visibili picchi ai lag stagionali (multipli di 7) sia sulla ACF che PACF quindi si conferma una non stazionarietà in media stagionale e si procede con la differenziazione stagionale. Inoltre, si può vedere che L'ACF, sia sui lag stagionali che non, sembra andare a zero anche se molto lentamente. Un andamento simile si ha per la PACF, anche se la discesa verso lo zero è in questo caso più rapida. Non si riescono ad individuare con precisione gli ordini dei processi quindi si procede inizialmente con una differenzazione stagionale di periodo 7 e la successiva stima del modello con una componente stagionale AR e MA di ordine 1.

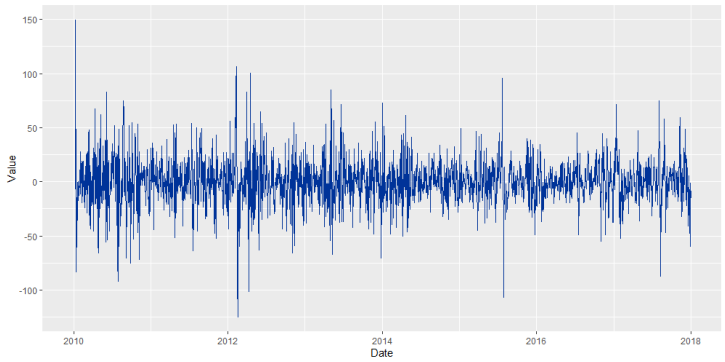

Figure 4: Serie differenziata stagionalmente

E' possibile vedere in Figura 4 che, a seguito della differenzazione stagionale, la serie risulta stazionaria in media. Non si notano segni particolari di persistenza, infatti, la serie oscilla intorno al valore zero. Dopo aver stimato il modello $SARIMA(0,0,0)(1,1,1)_7$, si analizzano i correlogrammi dei residui del modello riportati di seguito in Figura 5 al fine di identificare e modellare la componente non stagionale del modello.

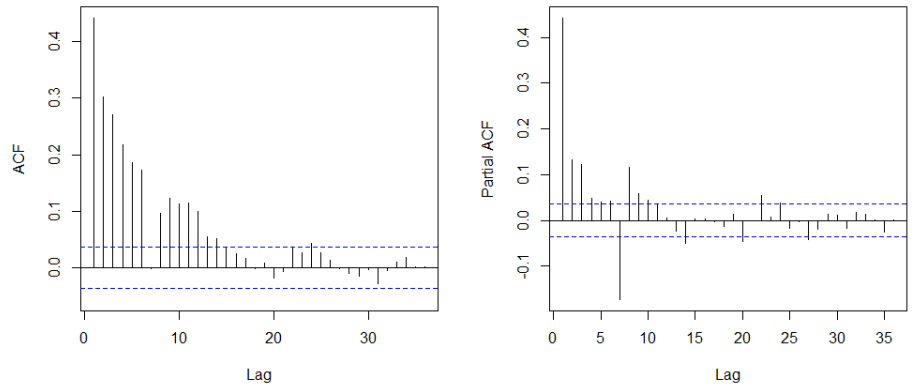

Figure 5: Correlogrammi dei residui del modello $SARIMA(0,0,0)(1,1,1)_7$

Si può osservare dai correlogrammi dei residui che l'ACF tende esponenzialmente a zero mentre la PACF rimane sicuramente significativa nei premi 3 ritardi.Anche i lag 4, 5 e 6, tuttavia, hanno valori che risultano significativi anche se vicini alla soglia limite. Anche il lag 7 della PACF risulta ancora significativo ma per ora si preferisce ignorare la cosa in quanto potrebbero

essere delle rimanenze della componente stagionale. Quindi, si procede prima stimando i modelli con una componente autoregressiva con ordine da 2 a 6 andando ad indagare come varia il criterio di Akaike nei diversi modelli.

Il modello con il valore dell'AIC minore risulta essere $SARIMA(6,0,0)(1,1,1)_7$. Si procede, quindi, con l'analisi dei correlogrammi dei residui di questo modello mostrati in Figura 6.

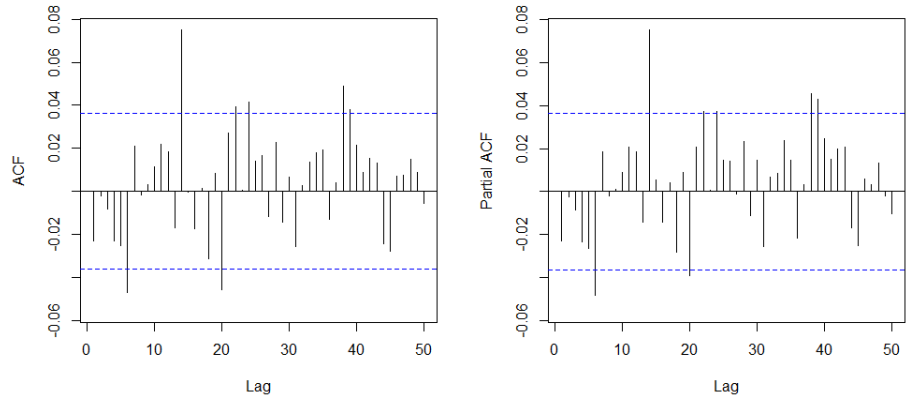

Figure 6: Correlogrammi dei residui del modello $SARIMA(6,0,0)(1,1,1)_7$

Dall'analisi di questi correlogrammi se ne deduce che gran parte della variabilità dei dati è stata colta del modello, infatti, le autocorrelazioni e le autocorrelazioni parziali risultano, a meno di qualche eccezione, non significative.

Analizzando il modulo delle radici della componente autoregressiva si nota che una risulta vicina ad uno, per questo motivo si è valutata anche la possibilità di effettuare un'integrazione di primo ordine. Tuttavia, l'AIC del modello $SARIMA(6,1,0)(1,1,1)_7$ è risultato maggiore rispetto a quello del modello non integrato e non ci sono stati grandi miglioramenti nei correlogrammi dei residui che risultavano avere ancora qualche ritardo significativo. Si è preferito, quindi, procedere con il modello $SARIMA(6,0,0)(1,1,1)_7$ senza l'integrazione di ordine 1.

Proprio la presenza di queste autocorrelazioni tra i residui ancora presenti potrebbe essere sintomo della presenza di una stagionalità intra-annua; come è possibile vedere in Figura 7, infatti, le previsioni del modello $SARIMA(6,0,0)(1,1,1)_7$ sul validation set non sono accurate.

Proprio per questo si procede valutando la possibilità di aggiungere regressori dummy e sinusoidali per migliorare l'adattabilità del modello ai dati. Solamente dopo aver trovato il modello ARIMA migliore si procederà con un'analisi sui residui di esso, per confermare che siano white noise.

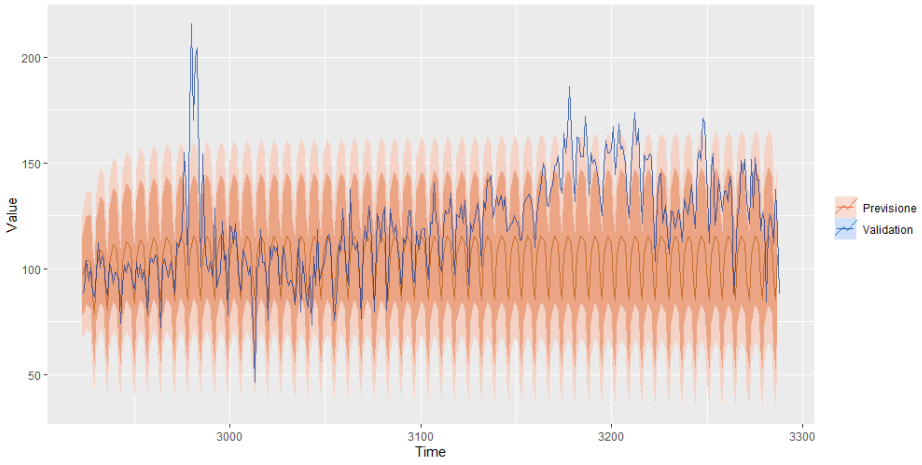

Figure 7: Previsione del modello $SARIMA(6,0,0)(1,1,1)_7$

Per catturare anche la componente stagionale intra-annua all'interno dei dati, sono stati costruiti dei regressori sinusoidali. In particolare, sono state considerate le prime 16 serie di seni e coseni con frequenza $\frac{2\pi}{365.25}$ in modo da modellare stagionalità piuttosto liscie.

Inoltre, per migliorare l'adattamento dei valori stimati alla serie originale sono stati introdotti dei regressori dummy rappresentanti i principali giorni di festività. E' stato assunto che i dati riguardassero il mercato elettrico italiano e, per questo motivo, sono stati considerate le seguenti festività: capodanno (31 dicembre-1 gennaio), epifania, Pasqua, festa della liberazione, festa dei lavoratori, festa della Repubblica, ferragosto, tutti i Santi e Natale (24-26 dicembre).

Sono stati creati 3 differenti modelli a partire dall'$SARIMA(6,0,0)(1,1,1)_7$: nel primo sono stati aggiunti solamente i regressori dummy, nel secondo solamente i regressori sinusoidali e nell'ultimo sono stati aggiunti entrambi. Valutando le performance di questi 3 modelli basandosi sia sul criterio AIC, sia sul valore del MAPE calcolato sul training set, sia sull'analisi visiva della previsione sul validation set, il terzo modello è risultato essere il migliore tra

gli ARIMA. E' possibile vedere in Figura 8 che, con l'aggiunta dei regressori esterni, sia quelli riferiti alle festività sia quelli che modellano la stagionalità intra-annua, la serie prevista si avvicina molto ai dati originali del validation set.

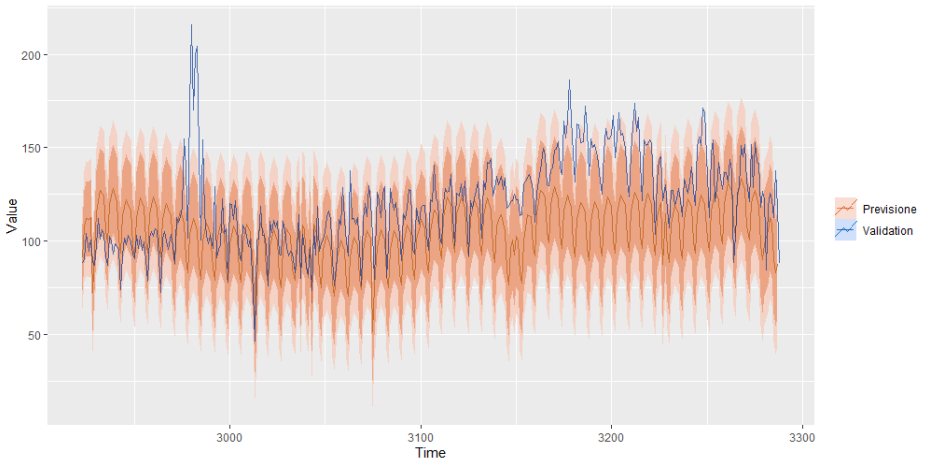

Figure 8: Previsione del modello $SARIMA(6,0,0)(1,1,1)_7$ con l'aggiunta di regressori sinusoidali e dummy

Avendo individuato il modello ARIMA migliore si procede con l'analisi dei residui. L'obiettivo è quello di verificare che siano generati da un processo white noise: a media nulla varianza costante e incorrelati con il proprio passato.

Dal plot del residui presente in Figura 9 è evidente che la media sia nulla, cosa che viene confermata anche dal test t per cui la media risulta essere non significativamente diversa da zero. Tuttavia, il test di Ljung-Box non permette di accettare l'ipotesi di assenza di autocorrelazione globale, infatti è possibile vedere, sempre in Figura 10, che l'autocorrelazione ad alcuni ritardi rimane comunque significativa. Questo perchè trattandosi di dati reali, è difficile arrivare ad avere residui effettivamente white noise. Una cosa che invece sembra essere verificata dal grafico è la normalità dei residui: è evidente che l'istogramma empirico dei residui assomigli molto ad una normale con media nulla.

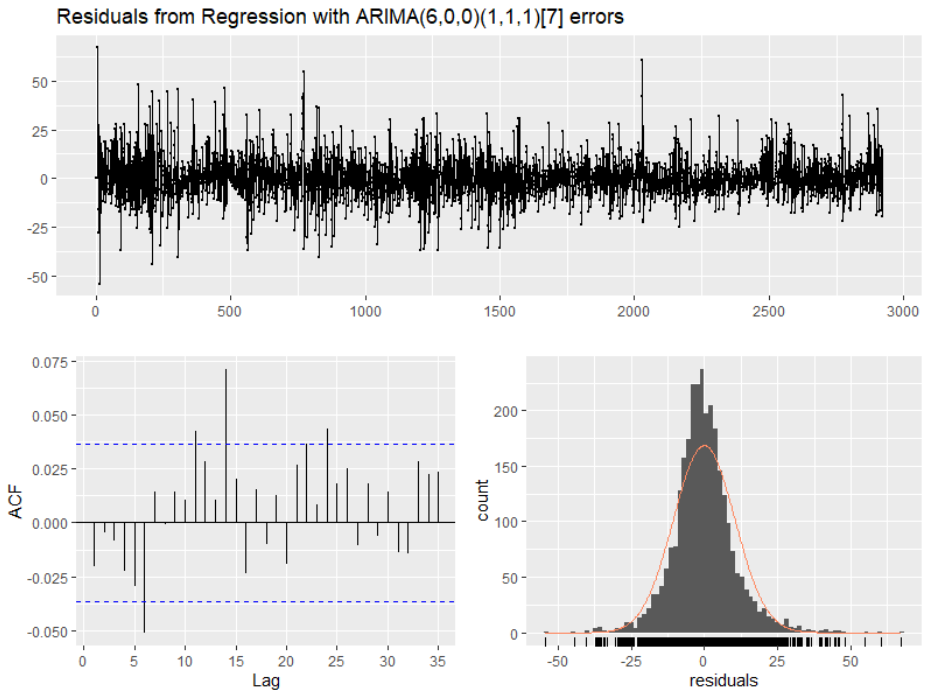

Figure 9: Analisi residui modello $SARIMA(6,0,0)(1,1,1)_7$ con l'aggiunta di regressori sinusoidali e dummy

3.2 UCM

I modelli ucm sono modelli a componenti non osservabili e permettono di combinare diverse componenti come trend, ciclo e stagionalità. In questo caso, benchè si tratti di una serie economica, non viene inserita la componente ciclica in quanto difficilmente questa influenza le previsioni one-step-ahead di un modello.

Per stimare la componente stagionale sono state utilizzate delle dummy stocastiche per modellare la stagionalità settimanale e delle sinusoidi stocastiche per modellare quella intra-annua. Per la stima del trend, invece sono state testati modelli con componenti differenti: il *local linear trend*, il *random walk* e l'*integrated random walk*. Dal momento che nei modelli ARIMA i regressori dummy che identificano le vacanze hanno portato un miglioramento nelle previsioni, sono stati aggiunti anche in acluni modelli ucm stimati. Non sono stati utilizzati come regressori esterni, invece, le serie di seni e coseni in

quanto la stagionalità intra-annua viene già modellata tramite le componenti non osservabili.

Tra i modelli UCM stimati, quello che ottiene performance migliori in termini di MAPE è il modello costituito da:

- *random walk* per modellare il trend;
- *sinusoidi stocastiche* per modellare la stagionalità intra-annua;
- *dummy stocastiche* per modellare la stagionalità settimanale;
- *regressori dummy* utilizzati come regressori esterni per identificare i giorni di vacanza.

L'adattamento ai dati del validation set di tale modello è riportato in Figura 10.

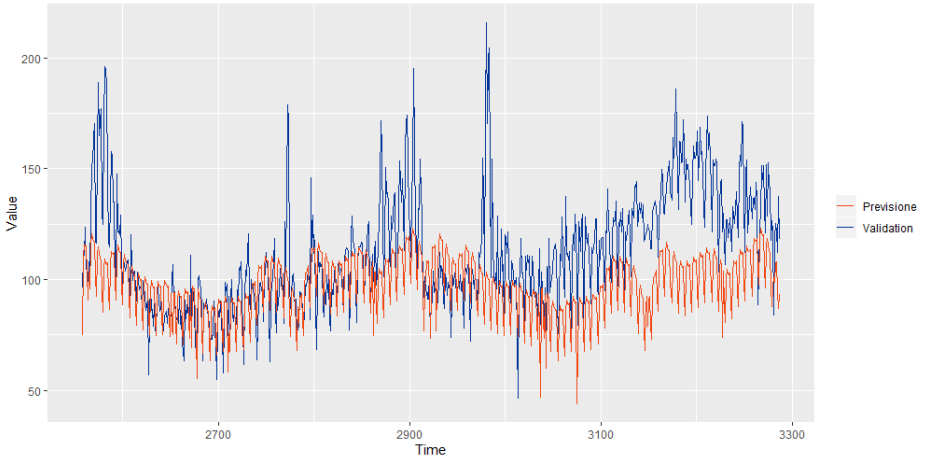

Figure 10: Previsione UCM sul validation set

Tramite lo sviluppo dei modelli UCM, oltre che stimare le componenti non osservabili, è possibile anche stimare gli shock che fanno evolvere queste componenti e, infine, l'errore di osservazione. In Figura 11 vengono riportati i grafici del disturbance smoother e dell'errore di osservazione del modello UCM suddetto.

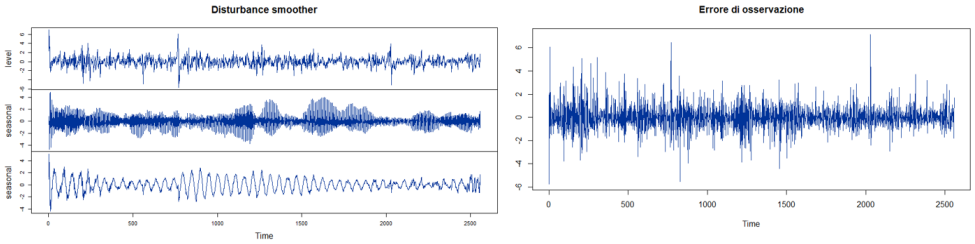

Figure 11: Disturbance Smoother

Emerge la presenza di alcuni cambiamenti repentini nell'evoluzione delle componenti non osservabili. Infatti, è visibile qualche picco che però non viene ritenuto così significativo da introdurre regressori esterni nel modello per modellare questi shock.

3.3 Machine Learning

Per quanto riguarda le tecniche di Machine Learning utilizzate per effettuare la previsione, sono stati implementate due tecniche differenti: k Nearest Neighbors e Recurrent Neural Network.

3.3.1 kNN

Il primo metodo utilizzato è stato quello della previsione tramite kNN: vengono previsti i valori futuri basandosi sui k Nearest Neighbors, ovvero le k serie più simili all'ultimo lag temporale che precede i valori da prevedere. 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. Nell'effettuare questa media si è scelto di pesare maggiormente le serie più recenti tra le k identificate. Si è scelto di utilizzare il metodo ricursivo, anzichè la metodologia MIMO-Multi Input Multi Output, in modo da avere una previsione one-step-ahead. Tramite il metodo ricursivo, in ogni iterazione vengono considerati, non solo tutti i dati della serie, ma anche i dati di previsione generati fino a quell'iterazione. In questo modo si aumenta la base campionaria del training set.

Per definire k, ovvero il numero di Nearest Neighbors da considerare per poi calcolare il valor medio, è stato seguito inizialmente un approccio euristico per cui si pone k pari alla radice quadrata della numerosità del training set.

Successivamente, però, è stata messa in pratica un'altra strategia che prevede di utilizzare più modelli kNN con k differenti per generare le previsioni e, tramite la media di questi valori, ottenere la previsione finale. Sono stati combinati, quindi, 9 modelli kNN con k pari a 30, 35, 40, 45, 50, 55, 60, 65 e 70 e, grazie a questa tecnica, è stato notato un lieve miglioramento in termini di MAPE. Si riportano in Figura 12 le previsioni ottenute sul validation set utilizzando questa tecnica.

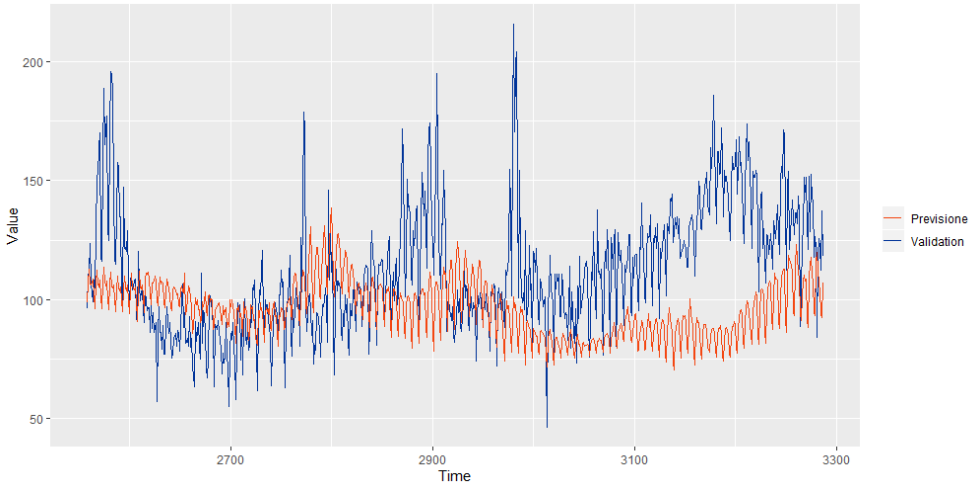

Figure 12: Previsioni kNN sul validation set

3.3.2 RNN

Successivamente sono state implementate due diverse architetture di reti neurali ricorrenti: l'LSTM (Long Short-Term Memory) e la GRU (Gated Recurrent Unit). Il vantaggio di queste tipologie di algoritmi è che permettono di conservare le informazioni sul passato analizzando sequenzialmente i dati.

La divisione tra training e validation set è rimasta la stessa per permettere un confronto con i modelli lineari.

I dati sono stati sottoposti ad una prima fase di pre-processing: per facilitare l'apprendimento delle reti sono stati inizialmente scalati e centrati. Successivamente, sono stati posti sotto forma di array avente 3 dimensioni: la numerosità del campione, il numero di lag (timesteps) e il numero di features. In questo caso il timestep è pari a 1, così come il numero di features.

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 rete 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. Come si può vedere in Figura 13, le previsioni ottenute grazie a questo modello si adattano molto bene ai dati del validation set.

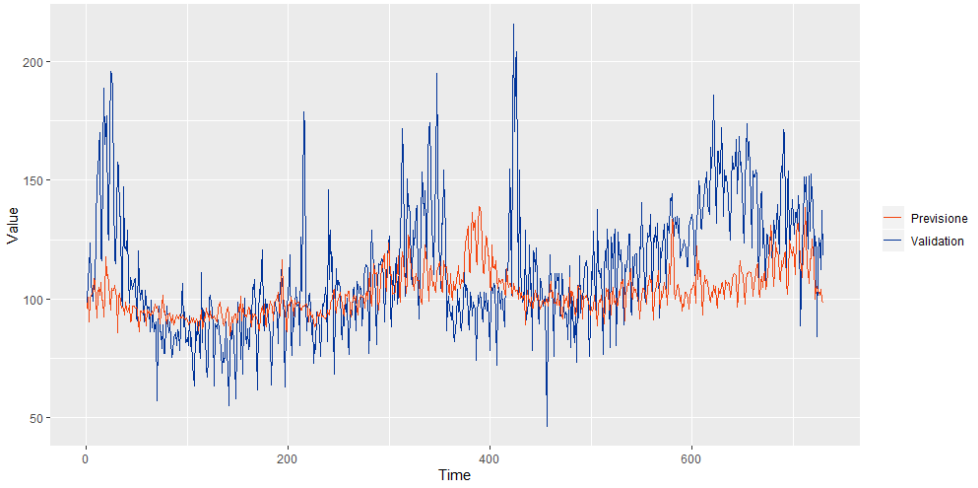

Figure 13: Previsioni GRU sul validation set

4 Conclusioni

Come anticipato nel paragrafo 3, per valutare le performance dei modelli è stata utilizzata la metrica MAPE calcolato sugli ultimi due anni di dati, ovvero sul validaiton set.

L'unica eccezione è stata fatta per i diversi modelli ARIMA, che, per essere confrontati tra di loro in fase di definizione del modello, sono stati confrontati tramite il criterio AIC. Per questo motivo, si riportano la performance in termini di MAPE solamente del miglior modello ARIMA stimato: $SARIMA(6,0,0)(1,1,1)_7$ a cui sono stati aggiunti i regressori dummy e sinusoidali.

	ARIMA
Training	9.28
Validation	14.15

Di seguito si riportano i risultati ottenuti dai modelli a componenti non osservabili.

	LLT (no regr)	LLT	RW	IRW
Training	7.91	7.32	7.34	7.32
Validation	18.78	18.43	18.27	22.28

Si ricorda che in tutti i modelli UCM la stagionalità è stata stimata sia tramite dummy stocastiche (stagionalità settimanale) sia tramite sinusoidi stocastiche (stagionalità intra-annua). 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 il random walk per stimare il trend stocastico. Tuttavia, è possibile vedere che, tra tutti i modelli lineari, il migliore rimane il modello ARIMA.

I modelli di machine learning, invece, hanno riportato i seguenti risultati.

	kNN (k=50)	kNN (multiple k)	LSTM	GRU
Validation	17.64	17.51	20.17	16.20

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. Questo è stato fatto affinchè i dati

degli ultimi due anni, molto importanti per effettuare le previsioni future, fossero inseriti nel processo di stima dei modelli. 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.

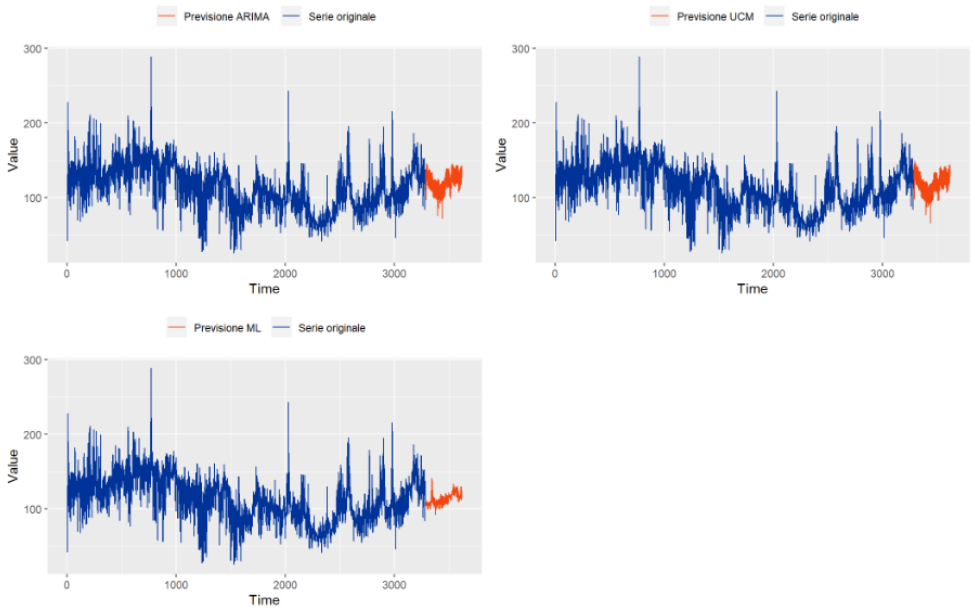

Figure 14: Previsioni ARIMA, UCM, GRU sul test set