CONSUMO GIORNALIERO DI ENERGIA ELETTRICA IN MAROCCO

Progetto del corso: Streaming Data Management and Time series Analysis

Matteo Lanzillotti – 843283

Università degli studi di Milano-Bicocca

14/02/2023

Introduzione

La serie storica in Analisi contiene le osservazioni, rilevate ogni 10 minuti, del consumo di energia elettrica in Marocco, nel periodo dal 1 Gennaio 2017 al 30 Novembre 2017, per un totale di **52560 osservazioni**. Si richiede per questo task la previsione dei consumi nel mese di Dicembre.

Il principale problema di questo task è la gestione di dati con una frequenza così elevata, che aumenta drasticamente i costi computazionali nonchè le tempistiche per l'addestramento dei modelli. Inoltre dati con una frequenza così elevata hanno una o più stagionalità, che va di conseguenza inclusa nei modelli per essere stimata. Il lavoro svolto prevede l'utilizzo di tre famiglie di modelli per effettuare le previsioni:

1. Arima
2. UCM
3. Machine Learning

Per valutare i modelli verrà utilizzato il **MAE** (*mean absolute error*)

Analisi effettuate

Per prima cosa è stata visualizzata la serie storica per dare un primo giudizio sulla stazionarietà e sull'eventuale presenza di stagionalità all'interno della serie.

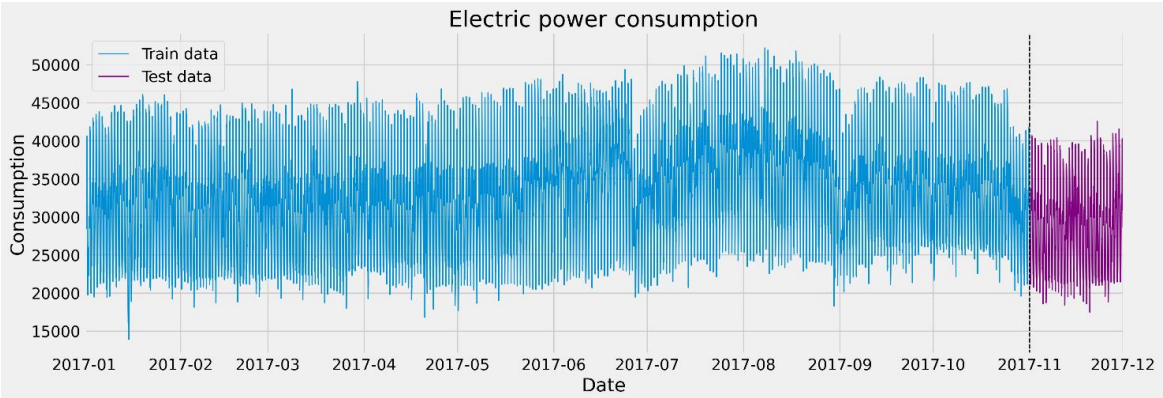

Figura 1 - Serie storica

Come vediamo la serie storica ha per i primi mesi un andamento abbastanza regolare, con un calo drastico in corrispondenza di fine Giugno, per poi andare a risalire durante il mese di Luglio, che è seguito da un nuovo calo che fa registrare valori più bassi fino a Novembre, il mese in cui sono stati registrati in media i valori più bassi.

Questa differenza tra Novembre e gli altri mesi si traduce poi in un problema nella stima dei modelli, in quanto la divisione della serie in training e test set, che ricopre proprio quest'ultimo mese, renderà difficile produrre stime accurate in quanto queste non terranno automaticamente in conto di questo calo.

Come possiamo vedere anche dai boxplot seguenti, la serie presenta diverse stagionalità al suo interno:

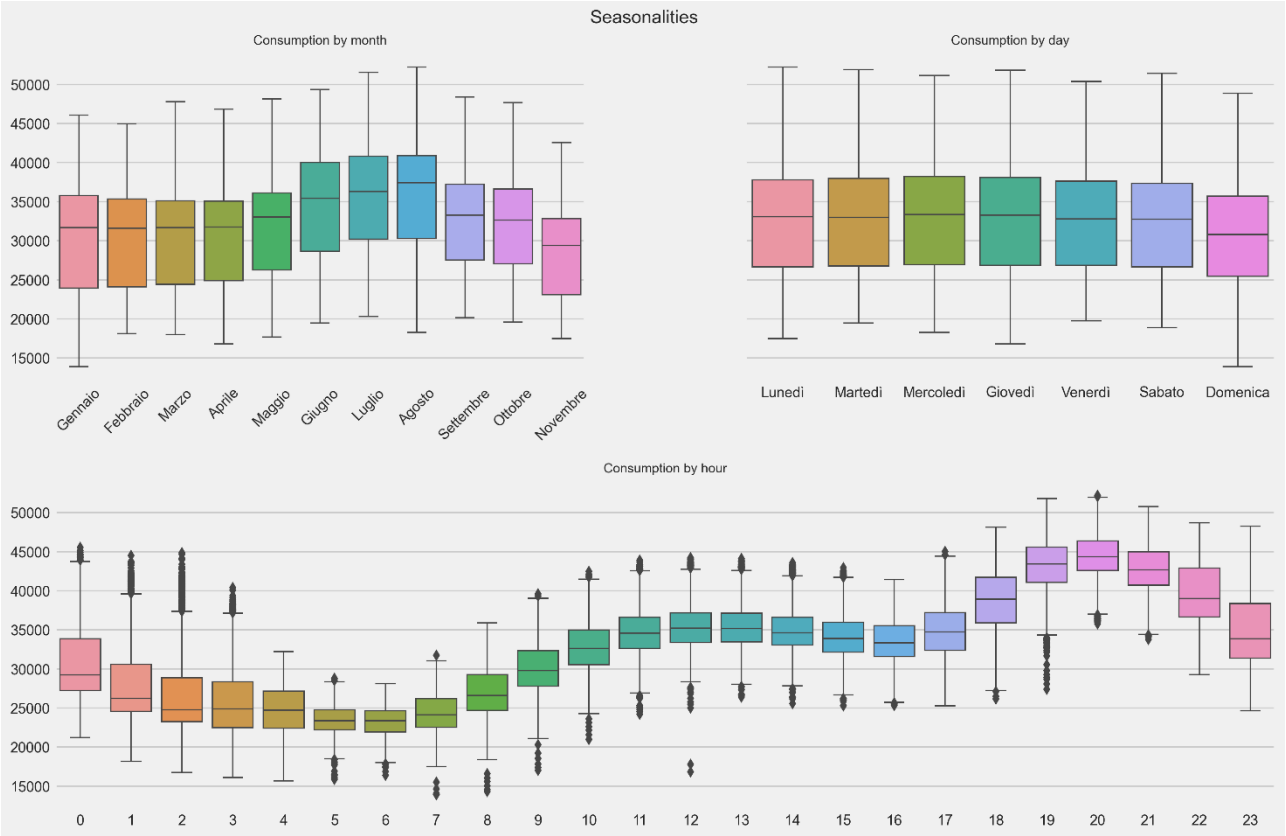

Figura 2 - Distribuzioni delle osservazioni per varie frequenze

La Domenica, com'è possibile vedere, si hanno sempre valori più bassi, probabilmente questo calo è dovuto alla chiusura delle aziende. Inoltre vediamo come durante la notte si registrano ovviamente valori più bassi, mentre durante il giorno i consumi si alzano fino a raggiungere il loro picco alle 20.

La serie così com'è tuttavia, come già anticipato nell'introduzione, necessitava di un preprocessing importante ai fini di ridurre i costi computazionali, soprattutto nella stima dei modelli ARIMA e UCM. Al fine di risolvere questo problema è stato necessario campionare il dataset, selezionando un osservazione per ogni ora, in corrispondenza del minuto :00. Il modello migliore è stato scelto sulla base dei risultati ottenuti sui dati campionati.

Per effettuare le stime sul mese di Dicembre si è poi stimato lo stesso modello su sei serie differenti, estrapolate dalla serie originale, divise per minuto di osservazione.

Come anticipato, per la stima dei modelli i dati sono stati divisi in training (da Gennaio a Ottobre) e test set (Novembre).

Il dataset fornito è completo, non presenta valori nulli. Lo step successivo è la valutazione dei grafici dell'ACF e PACF della serie storica.

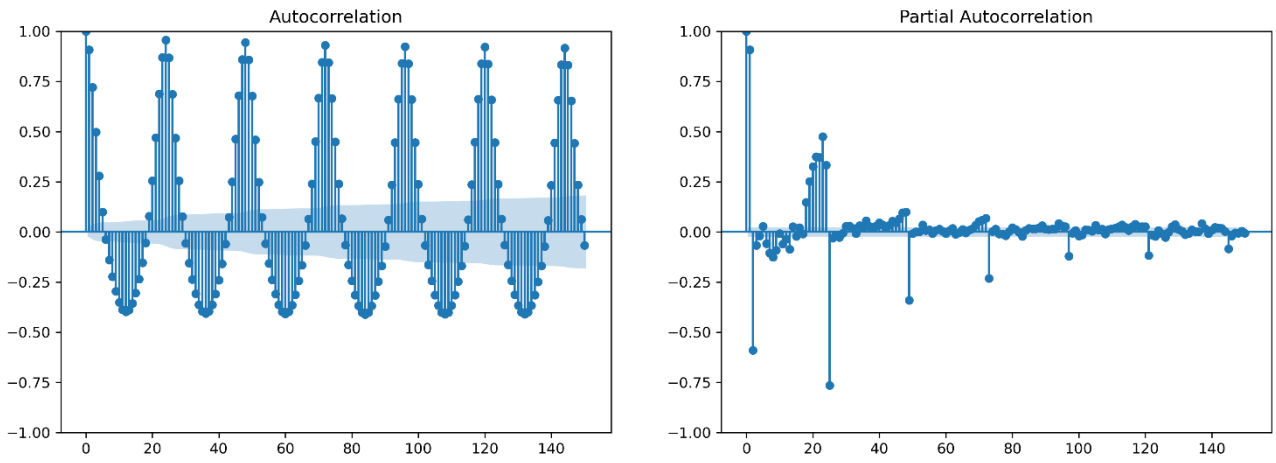

Figura 3 - Correlogrammi della serie storica

Dando un rapido sguardo ai due grafici, è evidente la già anticipata stagionalità di ordine 24 nella serie storica campionata, che verrà tenuta in considerazione nei modelli.

Dummy inserite

Per modellare alcune peculiarità della serie storica, sono stati aggiunti i seguenti regressori:

Giorno della settimana	Categorica → Dummies	Una colonna per giorno della settimana.
Stagione	Categorica → Dummies	Una colonna per stagione.
Festività[1]	Dummy	Se il giorno è festivo.
Ramadan[2]	Dummy	Periodo di Ramadan (dal 2017-5-26 al 2017-6-24).
Ore di luce[3]	Numerica (float)	Ore medie di luce, un valore per ogni mese.
Temperatura media diurna[4]	Numerica (int)	Temperatura media di giorno, un valore per ogni mese.
Temperatura media notturna[5]	Numerica (int)	Temperatura media di note, un valore per ogni mese.

[1] Festività islamiche 2017 - Feiertagskalender.ch
[2] Islam, Quando inizia il Ramadan 2017 · Immezcla
[3] Times for sunrise and sunset in Morocco (worlddata.info)
[4,3] Marrakesh, Morocco - Average Annual Weather - Holiday Weather (holiday-weather.com)

ARIMA

A partire dai grafici dell'ACF e PACF, si è cercato di trovare, attraverso successive differenze, il modello più adatto a stimare la serie storica. Dopo una differenziazione stagionale di ordine 24 e una nuova differenziazione di ordine 1 è stato possibile ottenere i seguenti correlogrammi:

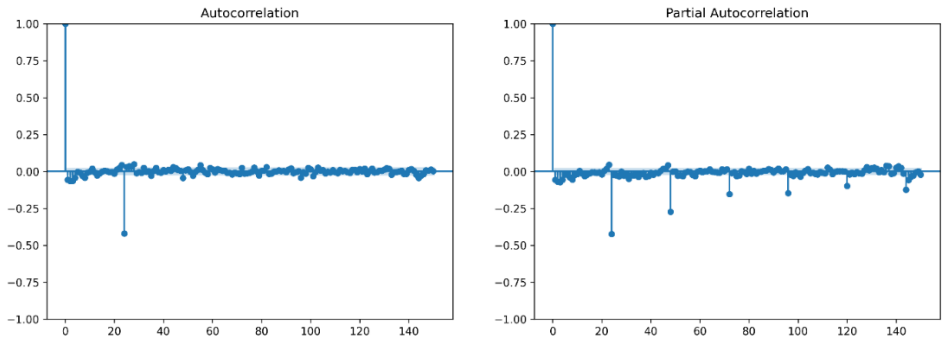

Figura 4 - Correlogrammi dopo due differenziazioni (24,1)

Che sono facilmente riconducibili a un processo MA(1).

Si è quindi stimato il modello ARIMA(1,1,1), con una parte stagionale di ordine (0,1,1,24), con e senza le varaibili esogene:

SENZA VARIABILI ESOGENE		**CON VARIABILI ESOGENE**	
AIC	119436.71	**AIC**	119450.702
MAE sul training set	587.41	**MAE sul training set**	586.95
MAE sul test set	1395.88	**MAE sul test set**	1232.88

Com'è possibile notare, l'introduzione delle variabili esogene non ha migliorato particolarmente le capacità di stima del modello, tuttavia è leggermente migliorato l'errore sul test set (**-13%**) , dimostrando quindi una migliore capacità di generalizzazione su dati nuovi.

Il modello è stato quindi stimato sulle 6 serie per intero, con le variabili esogene, e sono state ottenute le seguenti previsioni per il mese di Dicembre:

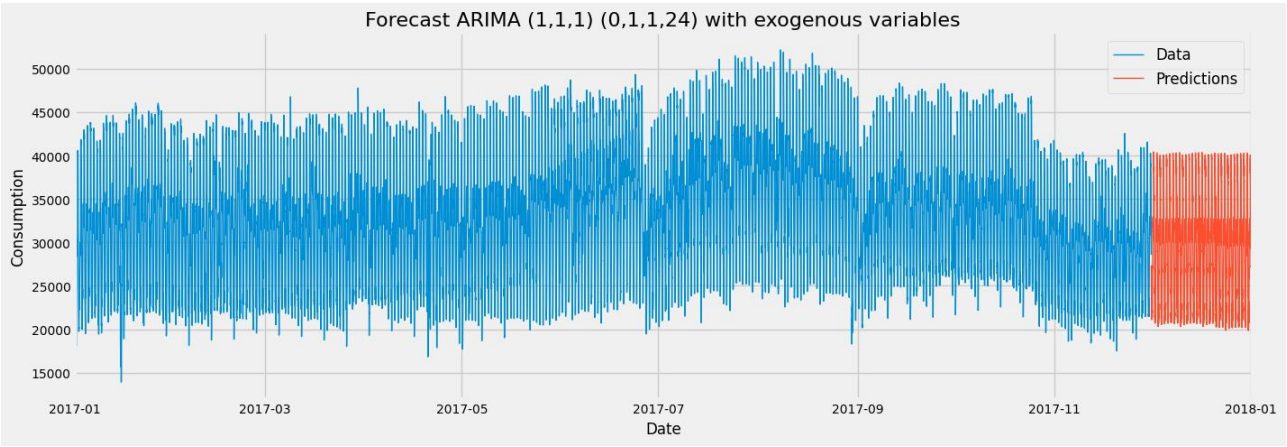

Figura 5 - Previsioni Modello ARIMA

Modello UCM

Il modello UCM stimato è composto dalle seguenti componenti:

- Trend Deterministico
- Stagionalità stocastica di ordine 24
- Componente autoregressiva di ordine 1
- Variabili esogene

I risultati ottenuti sono i seguenti:

UCM	
AIC	122389.07
MAE sul training set	797.19
MAE sul test set	1200.65

Il modello è soggetto ad overfitting. A facilitare questo fenomeno è anche il brusco calo della media nel test set, mai verificato nei dati di training, e quindi imprevedibile.

Il modello è stato ristimato sulle 6 serie storiche e sono state ottenute le seguenti previsioni:

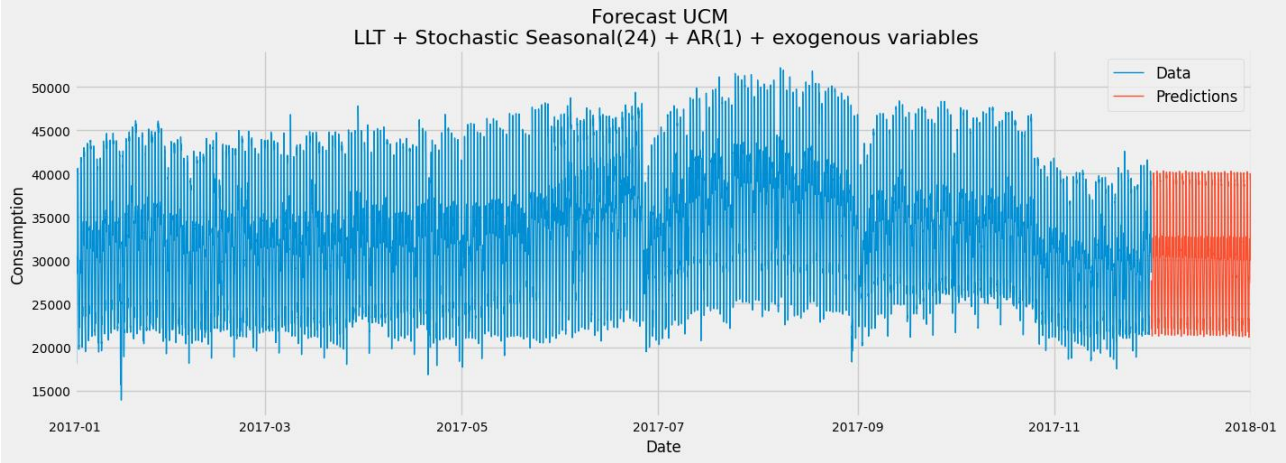

Figura 6 - Previsioni modello UCM

Machine Learning

XGBoost

Per svolgere il task è stato utilizzato l'algoritmo XGBoost, dopo aver arricchito il dataset con nuove variabili dummies che esprimono le caratteristiche del momento di osservazione. In particolare, le variabili aggiunte sono le seguenti:

FEATURE	DESCRIZIONE	FEATURE	DESCRIZIONE
Lags	Una colonna per ciascuna delle 6 osservazioni precedenti	**Settimana dell'anno**	In quale settimana dell'anno è stata fatta l'osservazione
Giorno del mese	Indice del giorno del mese	**Ora**	Ora del giorno dell'osservazione
Trimestre	In quale trimestre è stata fatta l'osservazione	**Mese**	Mese dell'osservazione

Ovviamente non tutte queste feature, insieme a quelle già aggiunte, saranno ugualmente importanti per il modello. Il modello è stato addestrato sui dati di training con 130 stimatori, un **learning rate** di 0.05, usando alberi con profondità 50.Il modello addestrato sul training set è stato poi utilizzato per predire il mese di Novembre.

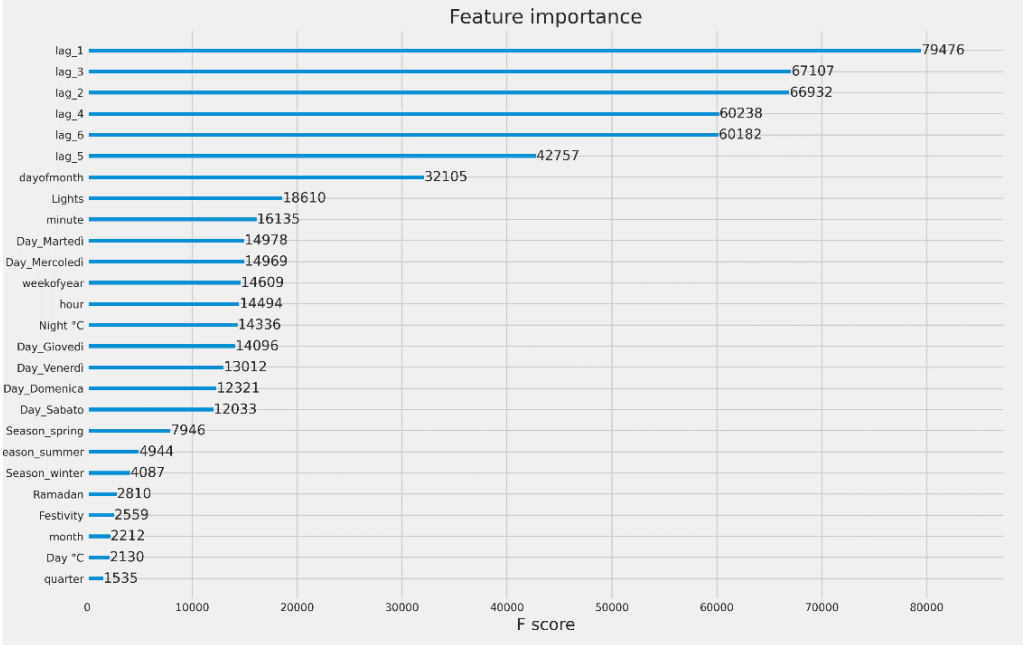

Figura 7 - Ranking dell'importanza delle feature

L'algoritmo implementato predice un valore, che poi viene inserito nella riga successiva come lag, traslando anche gli altri valori. In questo modo è stato possibile simulare il comportamento in presenza di dati mai visti, come accade nel mese di Dicembre. Il MAE qui ottenuto è molto alto, 2961.12, tuttavia si è ritenuto accettabile data la semplicità del modello. In prove precedenti infatti, anche includendo un numero molto più elevato di lag all'interno del modello, si è sperimentato che le previsioni collasserebbero molto più in fretta verso un valore molto piccolo rispetto al resto della serie.

Si è poi proceduto alla stima del mese di Dicembre, con la stessa metodologia applicata sopra, ottenendo i seguenti risultati:

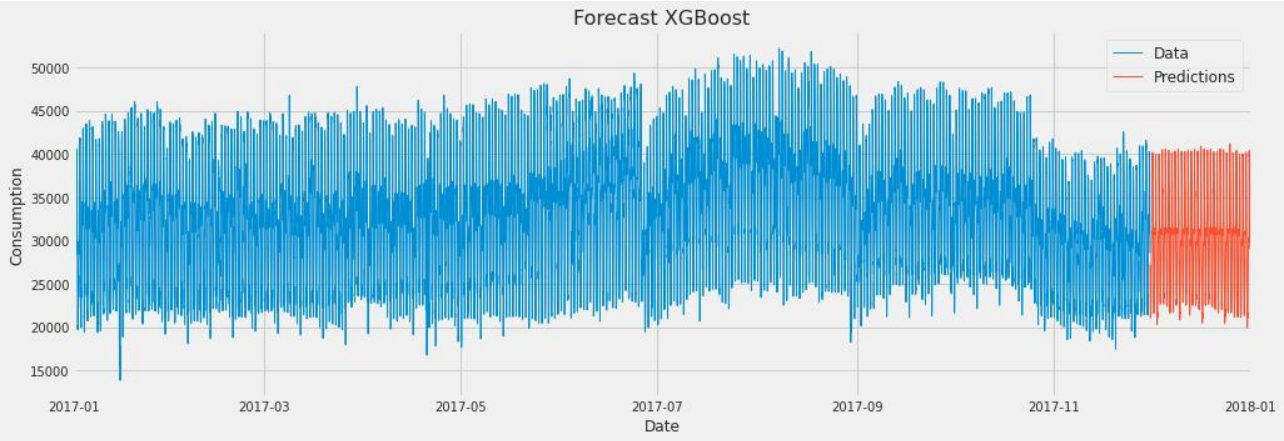

Figura 8 - Previsioni algoritmo XGBoost

Per tentare un altro approccio con tecniche di deep learning, sono state testate le performance di previsione con una rete neurale, in particolare una TCN (Temporal Convolutional Network).

TCN

[6]La Temporal Convolutional Network (TCN) è una tipologia di rete neurale artificiale che si concentra sull'elaborazione di dati temporali. La logica alla base della TCN consiste nell'utilizzo di convoluzioni per catturare relazioni a lungo termine tra i dati, permettendo alla rete di imparare a prevedere eventi futuri. La struttura di base di una TCN consiste in una serie di strati di convoluzione seguiti da uno o più strati di pooling. I layer convoluzionali utilizzano filtri che scorrono sui dati di input, mentre i pooling layer riducono la dimensionalità dei dati elaborati. Ciò consente alla rete di catturare relazioni a lungo termine tra i dati, poiché gli strati di convoluzione e pooling vengono ripetuti più volte. La TCN utilizza una struttura di dilatazione dei layer convoluzionali, che aumenta la capacità della rete di catturare relazioni a lungo termine tra i dati. Ciò è ottenuto incrementando la distanza tra i kernel di convoluzione, permettendo alla rete di catturare relazioni tra i dati a distanze maggiori.

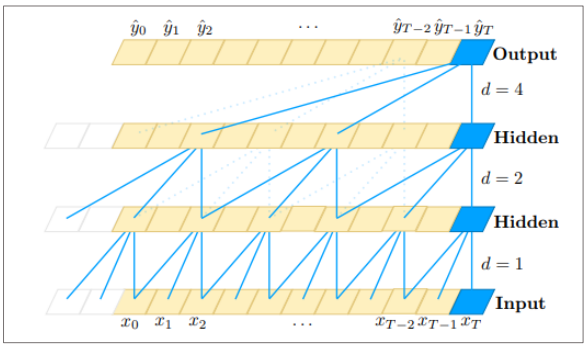

Figura 9 - Funzionamento di base di una TCN

[6] TEMPORAL CONVOLUTIONAL NETWORKS. Learning sequences efficiently and… | by Raushan Roy | Medium

Per l'addestramento della TCN, i dati sono stati normalizzati al fine di evitare problemi di convergenza durante l'addestramento della rete neurale. Per questo modello non sono state usate variabili esogene, ma solo la serie storica originale, andando a considerare 288 osservazioni per prevedere la successiva. La rete è stata addestrata sui seguenti parametri:

PARAMETER			
Lags	4	**Use skip connections** **Use batch normalization** **Use weight normalization** **Use layer normalization**	No No No No
Dilations	1,2,3,23,24,25,47,48,49, 142,143,144	**Optimizer**	Adam con learning rate = 0.001
Dropout rate	0.2	**Loss**	Mean squared error
Activation	ReLU	**Epochs**	56

La rete tuttavia, ha riportato dei risultati inverosimili per il mese di Dicembre, e pertanto si è preferito il modello XGBoost per questo task.

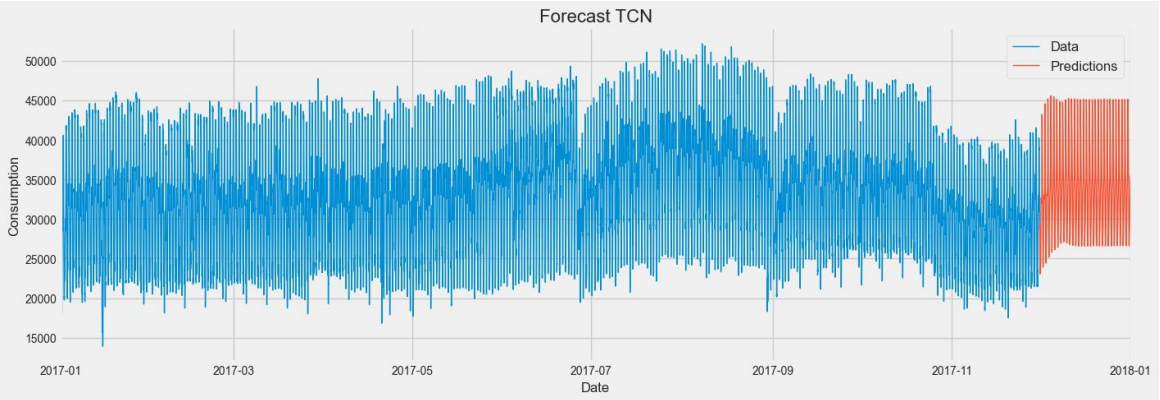

Figura 10 - Previsioni Temporal Convolutional Network

Conclusioni

In conclusione, le analisi svolte sul consumo di energia elettrica in Marocco, basate sui dati raccolti da gennaio a novembre 2017, hanno portato allo sviluppo di tre modelli: ARIMA, UCM e un algoritmo di Machine Learning. Il modello con il MAE più basso è risultato essere quello UCM, anche se la differenza con il modello ARIMA non è stata significativa. Questi risultati suggeriscono che entrambi i modelli possono essere utilizzati efficacemente per prevedere il consumo di energia elettrica in Marocco. Inoltre, per ulteriori sviluppi, sarebbe stato utile avere informazioni più significative sul luogo di rilevazione dei dati, per comprendere se rappresentano una regione specifica o l'intero paese. Inoltre, informazioni relative a eventuali scioperi o periodi prolungati di ferie, che potrebbero dipendere più da fattori culturali che stagionali, sarebbero state utili per modellare i dati, in particolare per spiegare il drastico calo avvenuto nel mese di novembre. Questi fattori potrebbero avere un impatto significativo sul consumo di energia elettrica e sarebbe opportuno tenerli in considerazione nei futuri modelli. La raccolta di queste informazioni avrebbe potuto migliorare la precisione dei modelli e fornire una comprensione più dettagliata del consumo di energia elettrica in Marocco.