Progetto Streaming Data and Time Series Analysis.

Corso di Laurea Magistrale in Data Science.

Gabriele Strano, mat.866563

Introduzione

Lo scopo di questa analisi è la previsione dei valori orari di CO per il periodo intercorrente tra il 01-03-2005 ed il 31-03-2005.

Questo task verrà affrontato con approcci appartenenti alla famiglia dei modelli lineari quali i modelli ARIMA e i modelli UCM ed approcci appartenenti alla famiglia dei modelli non lineari come le reti neurali ricorrenti LSTM e l'algoritmo KNN.

L'ambiente di lavoro principalmente utilizzato è stato R ad eccezione delle reti neurali ricorrenti LSTM che sono state implementate con Google Colab in Python.

I risultati ottenuti dai vari modelli sono stati confrontati principalmente tramite il Mean Absolute Percentage Error (MAPE), utilizzato sia per confrontare i risultati dei diversi metodi utilizzati che per confrontare la bontà dei parametri scelti all'interno dei modelli.

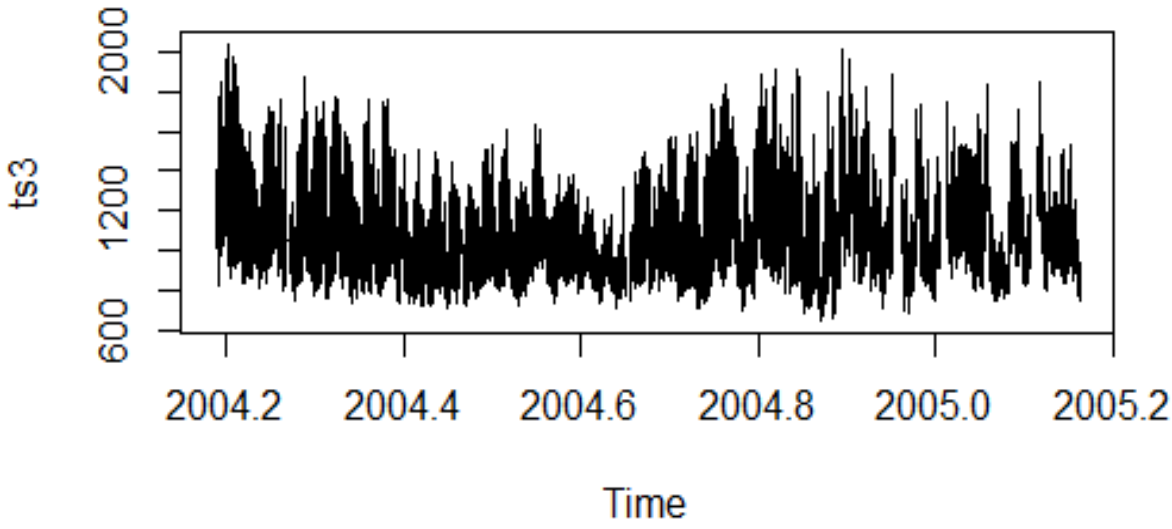
Analisi esplorativa

Il dataset di riferimento per questa analisi è composto da 3 colonne: la prima *'Date'* contenente la data in formato *yyyy-mm-dd,* la seconda *'Hour'* contenente soltanto l'orario della misurazione in valori interi da 0 a 23 e l'ultima '*CO',* contenente i valori effettivi della serie storica da analizzare.

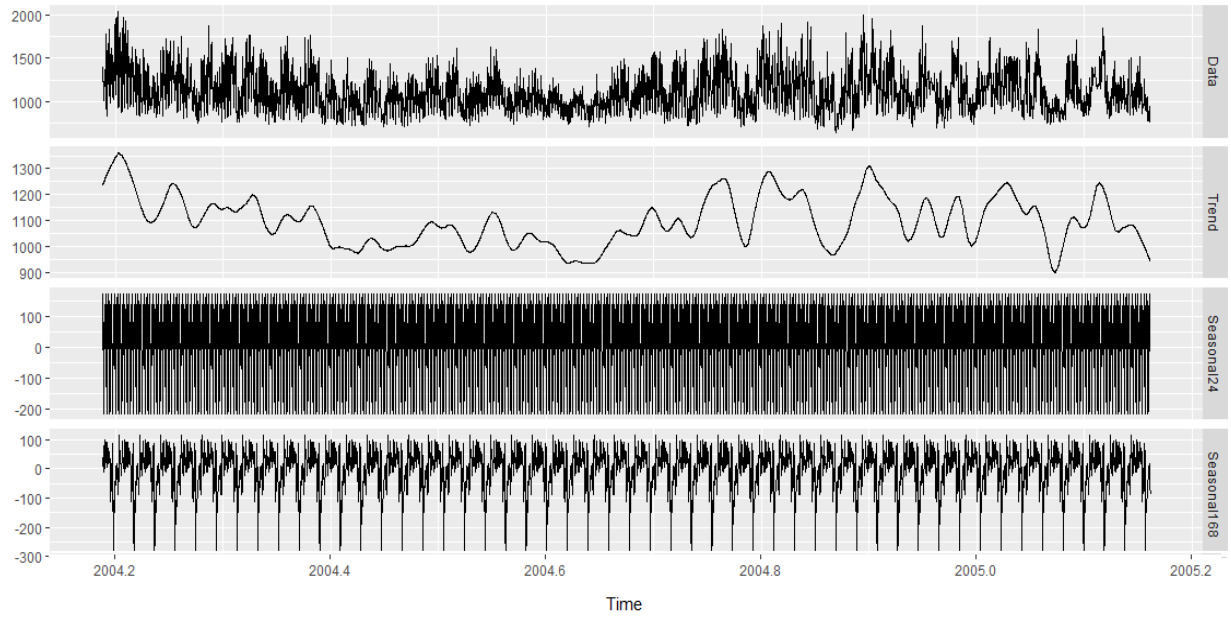
Il periodo di rilevazione della serie storica copriva quasi un anno, dalle ore 18 del 10-03-2004 alle ore 23 del 28-02-2005.

La prima operazione svolta è stata la creazione di una nuova colonna che fungesse da timestamp andando quindi ad unire le informazioni contenute nelle colonne data ed ora.

La serie presentava alcuni intervalli di rilevazioni mancanti come è possibile osservare dal grafico qui in basso.



Sempre dallo stesso grafico è possibile notare come la serie storica sembri non stazionaria.
È anche presente una forte componente stagionale come è possibile notare qui in basso.
Nel secondo e nel terzo riquadro sono rappresentate le componenti stagionali con frequenza giornaliera e settimanale.

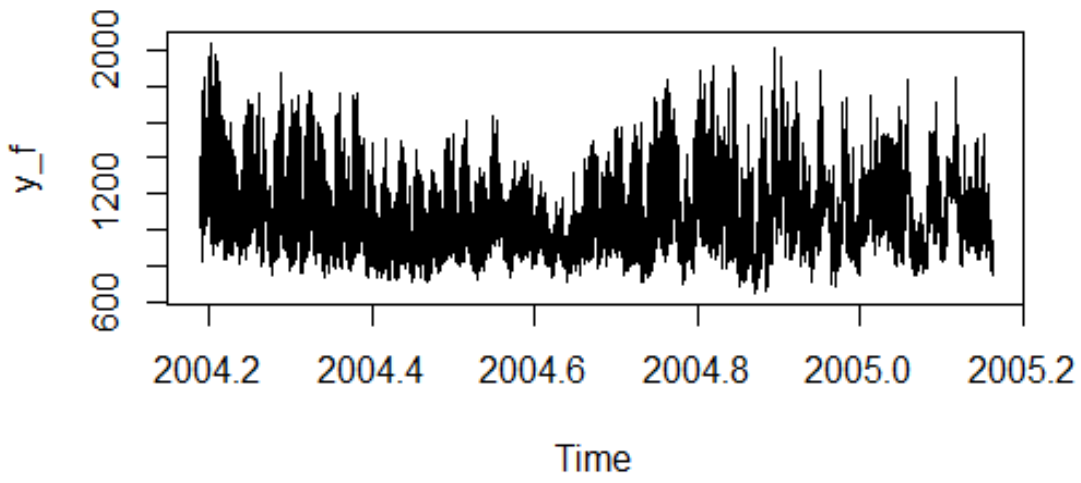


Null Value

È stato scelto di andare a riempire i valori mancanti tramite delle previsioni effettuate con i modelli ARIMA che meglio si adattavano ai dati.
Per evitare di riportare di volta in volta le parti di errore delle stime precedenti anche sulle stime successive, non è stato aggiornato il dataset né la serie storica tra una previsione e l'altra.
Solo alla fine dell'intero processo di previsione delle sequenze di null values queste sono state inserite nel dataset e nella serie storica finali.

Le previsioni sono state effettuate utilizzando come train set l'insieme di dati precedenti al periodo mancante e come validation set le 24 osservazioni successive a questi ultimi.
Una volta verificato che le distribuzioni dei residui dei modelli fossero approssimabili a distribuzioni normali, e che correlogrammi ACF/PACF non risultassero eccessive componenti non spiegate dai modelli si è proceduto alla selezione dei migliori modelli candidati.
Per selezionare quindi i migliori parametri per il modello finale di volta in volta sono stati tenuti in considerazione il valore dell'AiCc (che tiene conto della numerosità dei parametri oltre che della verosimiglianza) ed il Mean Absolute Percentage Error (MAPE) calcolato rispetto alle 24 rilevazioni dei validation set.
Il modello ARIMA maggiormente utilizzato per questa fase di previsione dei null value è stato un $ARIMA(2,1,1)(0,1,1)_{24}$.
Per catturare la doppia stagionalità sono stati provati dei regressori esterni che andavano a modellare anche la componente settimanale della stagionalità.
Sono state provate sia variabili dummies per ogni giorno della settimana che sinusoidi con frequenza settimanale

In basso, il grafico della serie storica senza ulteriori valori mancanti che verrà utilizzata per le previsioni con i modelli descritti nell'introduzione.

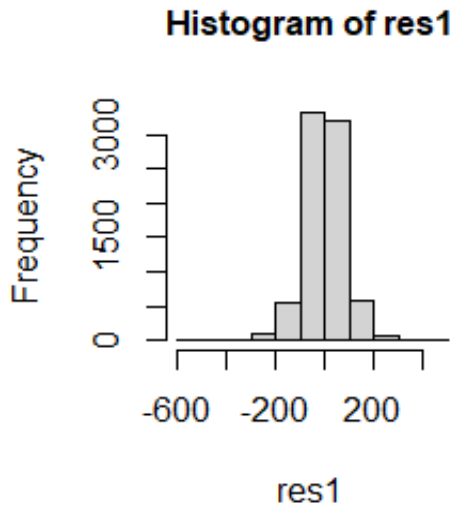


Modelli ARIMA

Sono stati testati modelli ARIMA con svariate combinazioni di parametri e svariate tipologie di regressori.
I tipi di regressori testati sono stati: variabili dummies sia giornaliere (24 variabili) che settimanali (7 variabili) e sinusoidi con diverse frequenze.

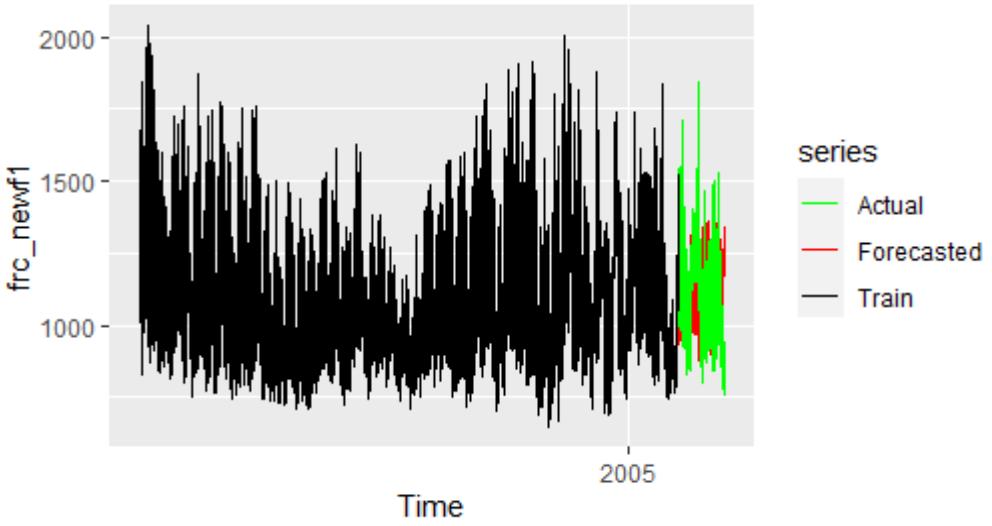
Nel modello finale sono state create queste sinusoidi con la funzione fourier del pacchetto *forecast*, essendo risultate significative sono state utilizzate 4 sinusoidi per modellare la stagionalità giornaliera e 2 sinusoidi per quella settimanale.
Il modello ARIMA scelto dopo i vari test è stato ARIMA(3,1,1)$(1,0,1)_{24}$.

In basso a sinistra, è possibile notare come la distribuzione dei residui del modello si possa considerare simile ad una distribuzione normale il che significa che il nostro modello riesce a rappresentare abbastanza bene i dati.
A destra invece una tabella contenente gli score dei 4 migliori modelli considerati.



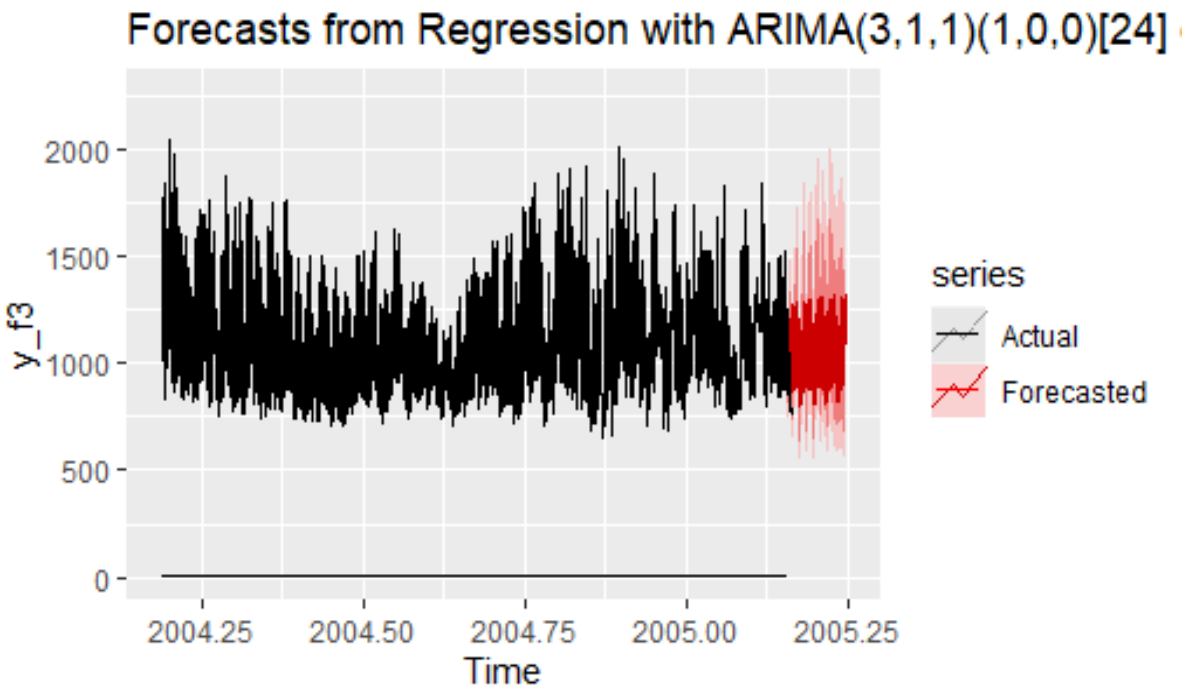
Modello	AiCc	MAPE
$(3,1,1)(1,0,1)_{24}$	90714.34	13.59
$(1,1,2)(1,0,1)_{24}$	90715.23	14.05
$(1,1,2)(1,0,0)_{24}$	91290.26	12.54
$(3,1,1)(1,0,0)_{24}$	91269.45	12.28

Di seguito viene mostrato il grafico delle previsioni riguardanti il mese di febbraio utilizzato come ultimo tra i validation set.
In nero la serie storica usata come train set, in verde i veri valori di febbraio ed in rosso i valori predetti tramite il modello ARIMA finale.



Per le effettuare le previsioni riguardo al mese di Marzo è stato quindi selezionato il modello ARIMA(3,1,1)$(1,0,1)_{24,}$ con l'ausilio delle sinusoidi come regressori per modellare entrambe le componenti stagionali. È stata anche applicata una trasformazione logaritmica per rendere stazionaria in varianza la serie storica prima di calcolare il modello ARIMA.

In basso vengono quindi mostrate le previsioni finali orarie del livello di CO, riguardanti l'intero mese di marzo.
Le previsioni in rosso hanno anche gli intervalli di confidenza al 80% ed al 95% rappresentati entrambi con tonalità di rosso più leggere.

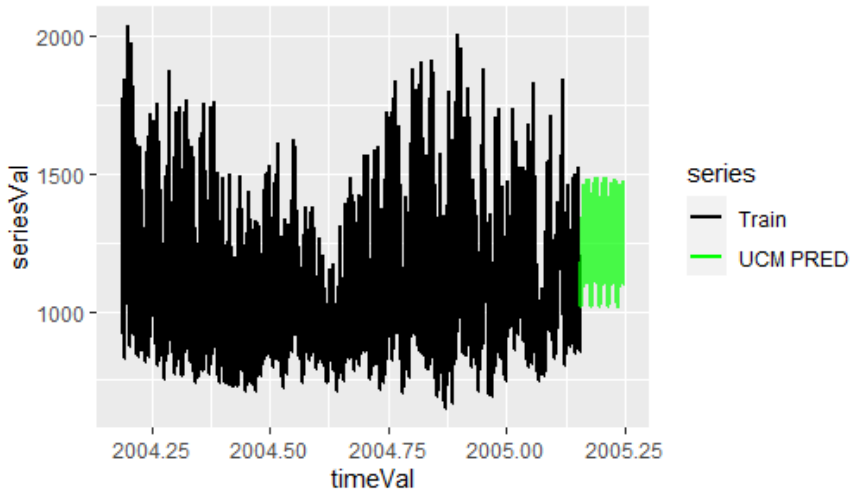


Modelli UCM

Come fatto in precedenza anche per questa tipologia di modelli sono stati testati e confrontati diversi parametri.
Partendo dalla base di conoscenza sviluppata durante la fase di studio relativa ai modelli ARIMA si è deciso di utilizzare due componenti stagionali, rispettivamente ogni 24 e 168 osservazioni, entrambe modellate con componenti trigonometriche.
Sono stati provati modelli con: local linear trend (LLT), random walk (RW), integrated random walk (IRW).
Tra i vari modelli candidati i migliori in termini di MAPE e di adattamento ai dati sono risultati i seguenti 4, si è scelto di proseguire l'analisi andando a migliorare l'ultimo dei quattro modelli rappresentati in basso, nonché quello con il valore del MAPE più basso.

Trend	Stagionalità 24h	Stagionalità 168h	MAPE
RW	4 armoniche	2 armoniche	18.16
IRW	4 armoniche	2 armoniche	18.12
IRW	8 armoniche	2 armoniche	17.54
RW	8 armoniche	2 armoniche	17.42

Per migliorare ulteriormente il modello si è deciso anche di inserire ulteriori regressori esterni calcolati con il modello ARIMA(7,0,0) sul train set considerato.
Il MAPE di questo nuovo modello calcolato sullo stesso validation set utilizzato in precedenza, ovvero l'intero mese di fabbraio, è sceso al 12.83%.
Il modello finale scelto per la tipologia UCM è stato quindi un modello con RW, 8 armoniche per la stagionalità giornaliera, 2 armoniche per la stagionalità settimanale (168h) e regressori ARIMA(7,0,0).
In basso viene mostrato il grafico delle previsioni relative al mese di marzo.



Modelli di Machine Learning

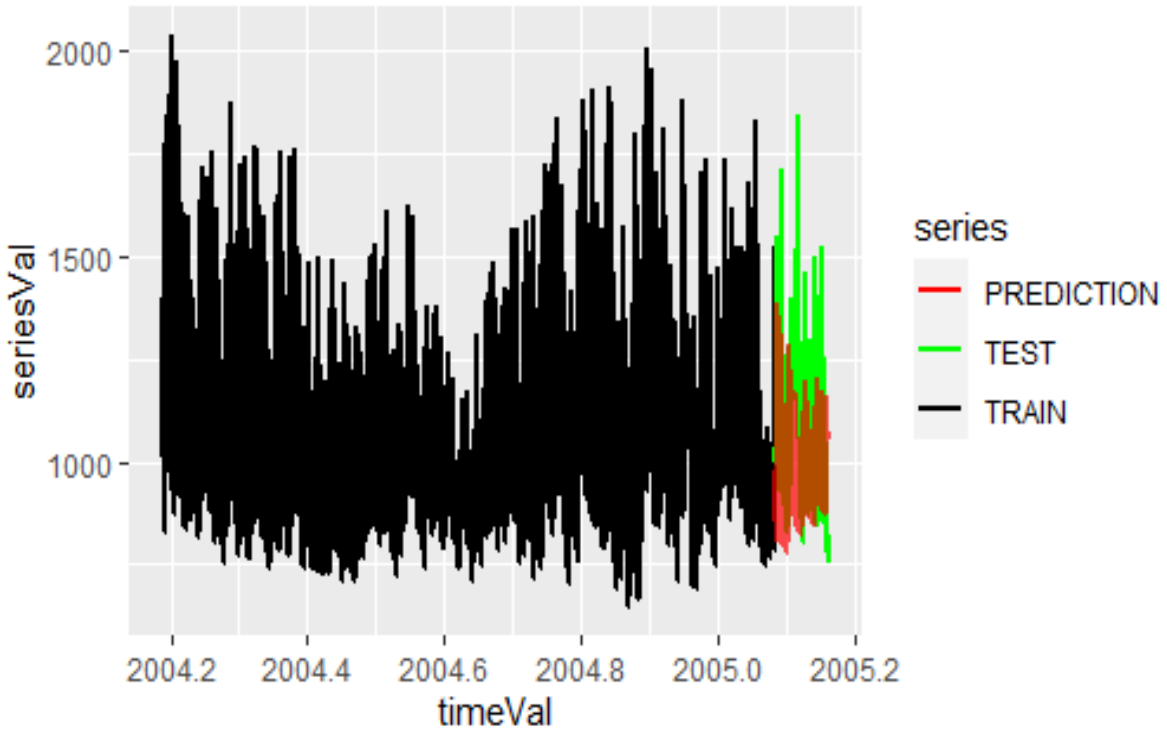
KNN

Il primo dei modelli non lineari utilizzato per il nostro task è stato il modello KNN per il quale è stata utilizzata la libreria *tsfknn.*

I parametri utilizzati per eseguire questa previsione ricorsiva sono stati i seguenti:

- p = 168 (settimanale)
- k = 7 (numero di sequenze da considerare)
- h = 672 (lunghezza del validation set cioè numero di osservazioni di febbraio)

Il MAPE ottenuto per questo tipo di modello è stato del 11.51%. In basso, viene mostrato il grafico delle previsioni riguardante il mese di fabbraio.



Nonostante i risultati sembrino complessivamente buoni si è scelto di puntare su un modello più strutturato per quanto riguarda i modelli non lineari.

LSTM

Per questo ultimo task si è scelto per comodità di lavorare in un altro ambiente, è stato infatti implementato in Python tramite Google Colab. La libreria utilizzata per questo task è stata *Keras*.

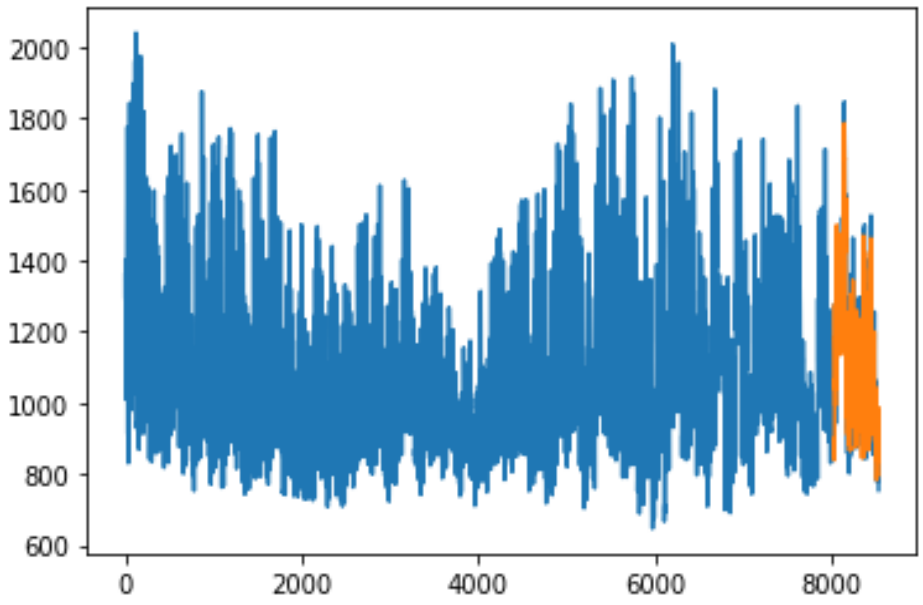
È stata utilizzata una rete neurale ricorrente di tipo LSTM per modellare al meglio la serie anche con modelli del tipo non lineare.

Come prima fase di preparazione di dati è stata utilizzata la funzione *MinMaxScaler (0,1)* per normalizzare i dati, in quanto in questo modo viene facilitato il lavoro della rete neurale. Dopo la divisione in train set e validation set, entrambi sono stati rimodellati con una finestra di lookback pari a 168 osservazioni, che era la stagionalità settimanale che a nostro avviso risultava dominante ed inoltre inglobava almeno in parte anche quella giornaliera.

Per essere passato alla rete ricorrente il train set ha bisogno di avere una nuova forma infatti è stato trasformato in un array 3d con le seguenti dimensioni: *(sample, time steps, features).*

Il modello finale selezionato è stata quindi una rete neurale ricorrente con layer LSTM da 100 neuroni un layer di output di tipo Dense con un solo neurone, funzioni di attivazione 'sigmoid', addestrata su 50 epoche e con batch size pari a 45.

In basso, sono rappresentate con il colore arancione le previsioni effettuate sul mese di fabbraio, mentre in blu i valori reali della serie storica. Il modello sembra adeguarsi bene ai dati.



Conclusioni

Tenendo in considerazione i risultati ottenuti sul validation set, i modelli non lineari utilizzati sembrano essere più performanti rispetto ai modelli lineari.
Tuttavia, con entrambe le tipologie di modelli si ottengono buoni risultati.
In particolare, la RNN con layer LSTM è quella che presenta le maggiori possibilità di essere migliorata, aggiungendo ad esempio un nuovo strato LSTM ed altri layer successivi.

Va inoltre aggiunto che essendo una serie storica di dati reali rimane comunque qualche componente stagionale che non si riesce a catturare con i vari modelli e questo viene anche accentuato dalla scarsità dei dati a disposizione.
È assai probabile che avendo più di un anno di dati a disposizione tutti i modelli considerati possano migliorare in termini di precisione delle previsioni.