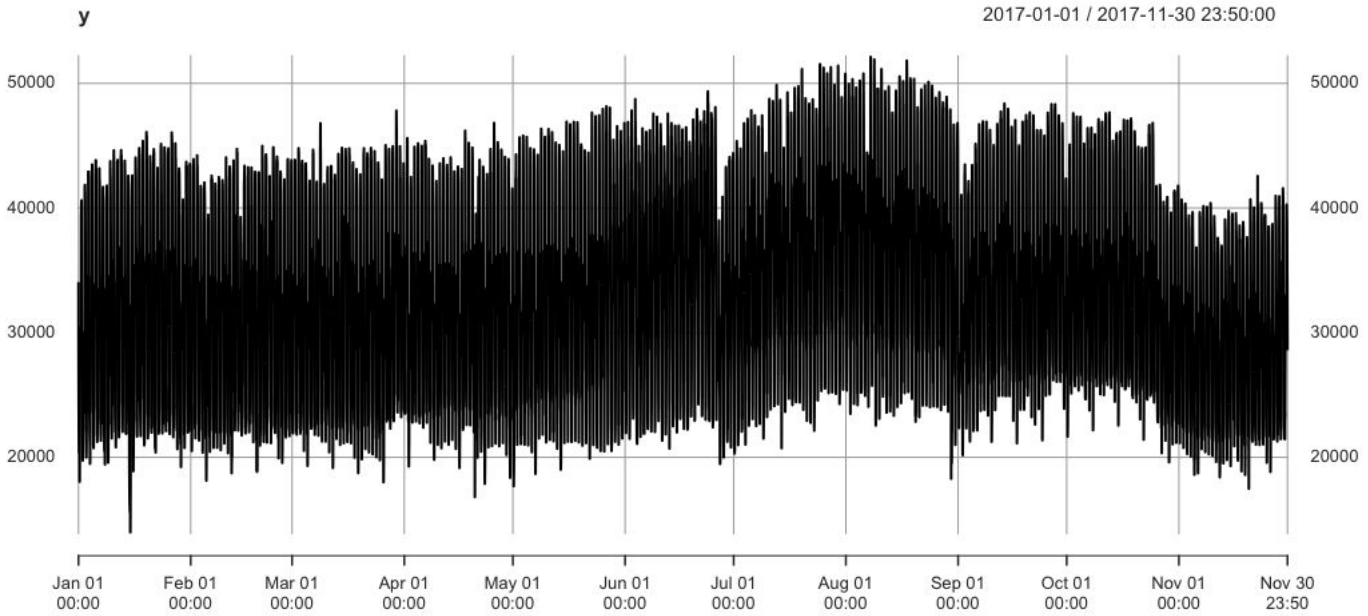
Streaming Data Management and Time Series Analysis

Emanuela Elli [892901]

Corso di Laurea Magistrale in Data Science
Anno accademico 2022-2023

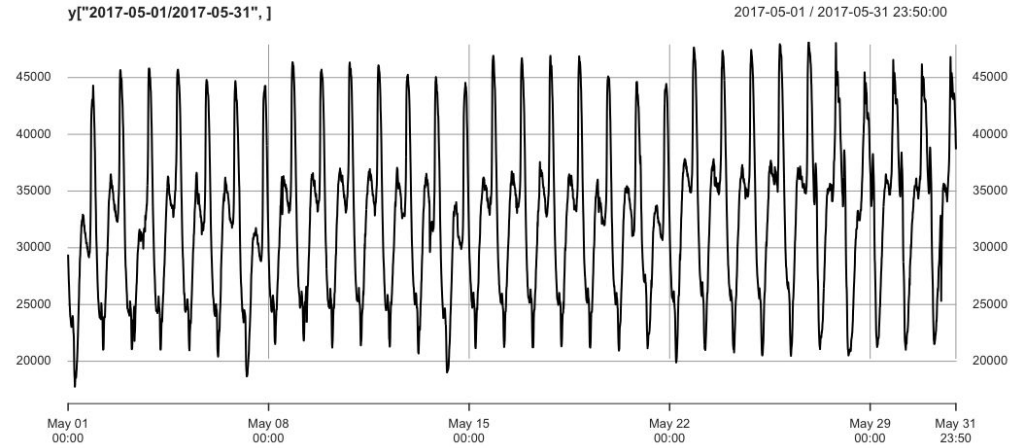
Data Exploration and Preprocessing

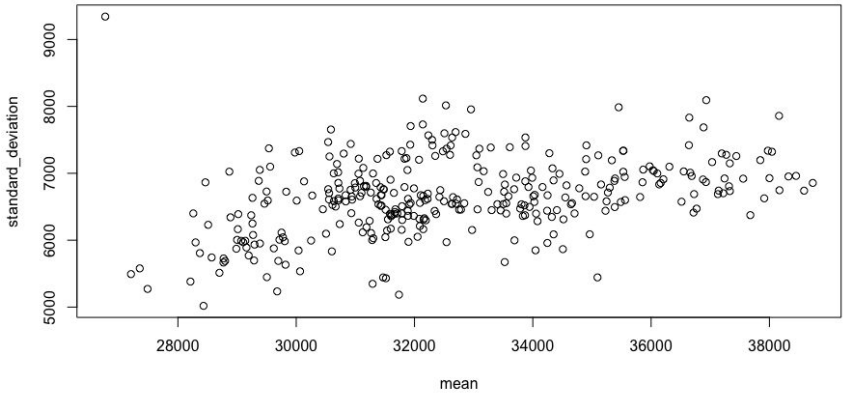
- Verifica valori mancanti e del formato data e ora.
- Osservazione di stagionalità e trend.



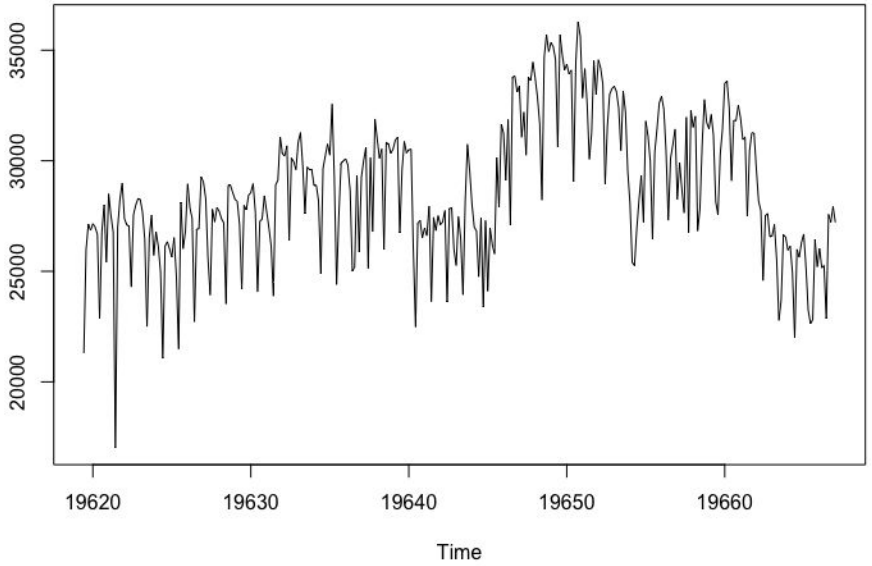
Data Exploration and Preprocessing

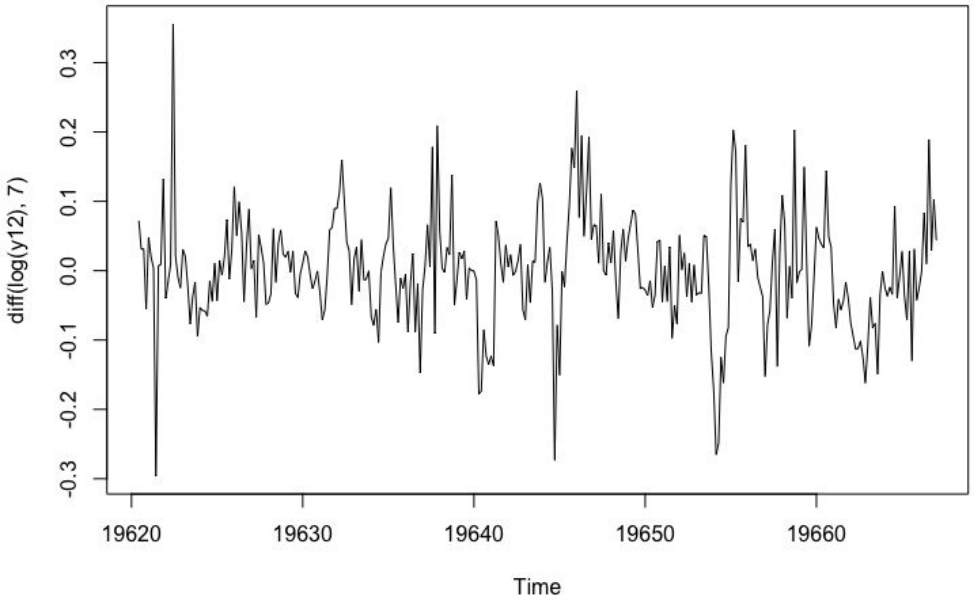
Si osserva la presenza di stagionalità giornaliera, settimanale e mensile oltre ad un trend lineare nel grafico di media e deviazione standard.





Data Exploration and Preprocessing



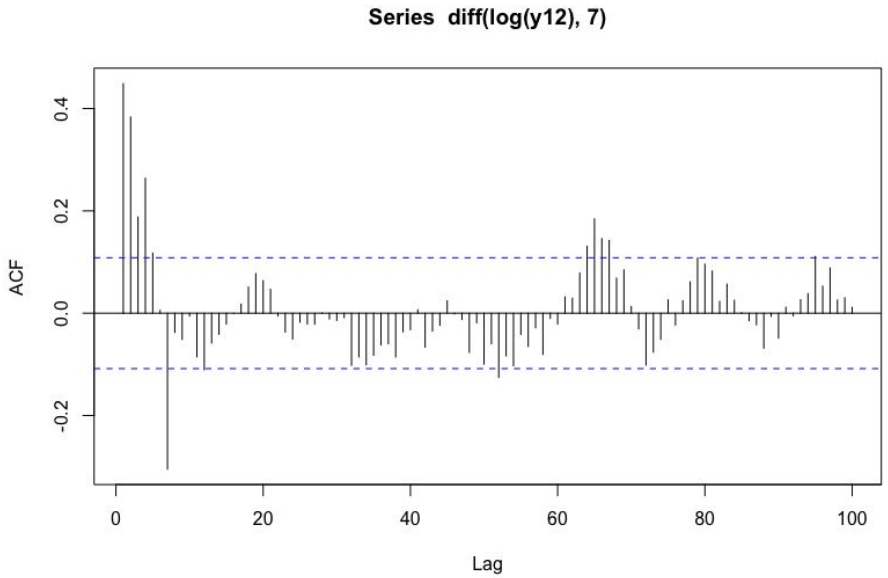


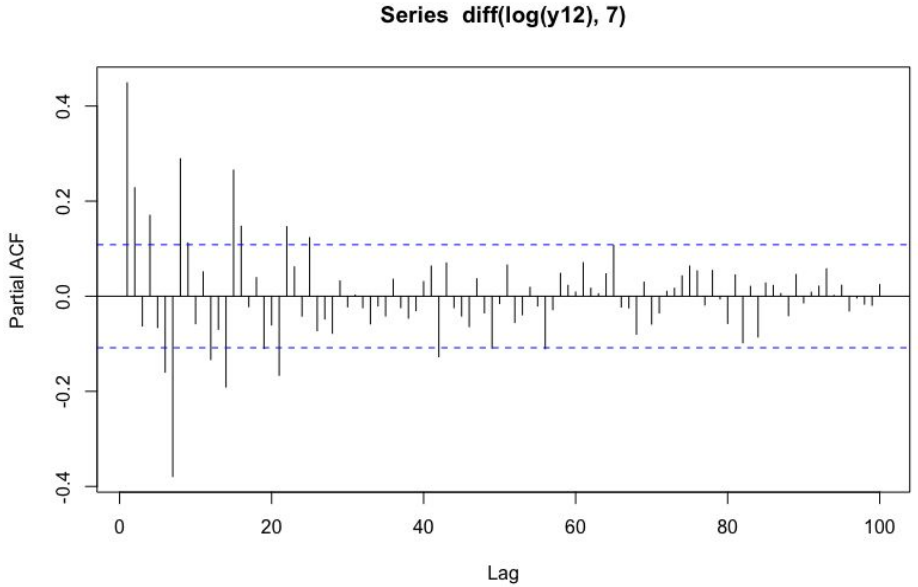
Viene considerato un singolo orario (ore `09:00:00`) e vengono applicati:

- Logaritmo
- Differenza stagionale

Data Exploration and Preprocessing

Vengono osservati i grafici di autocorrelazione e autocorrelazione parziale della serie per verificare la stazionarietà e la presenza di processi AR e MA.





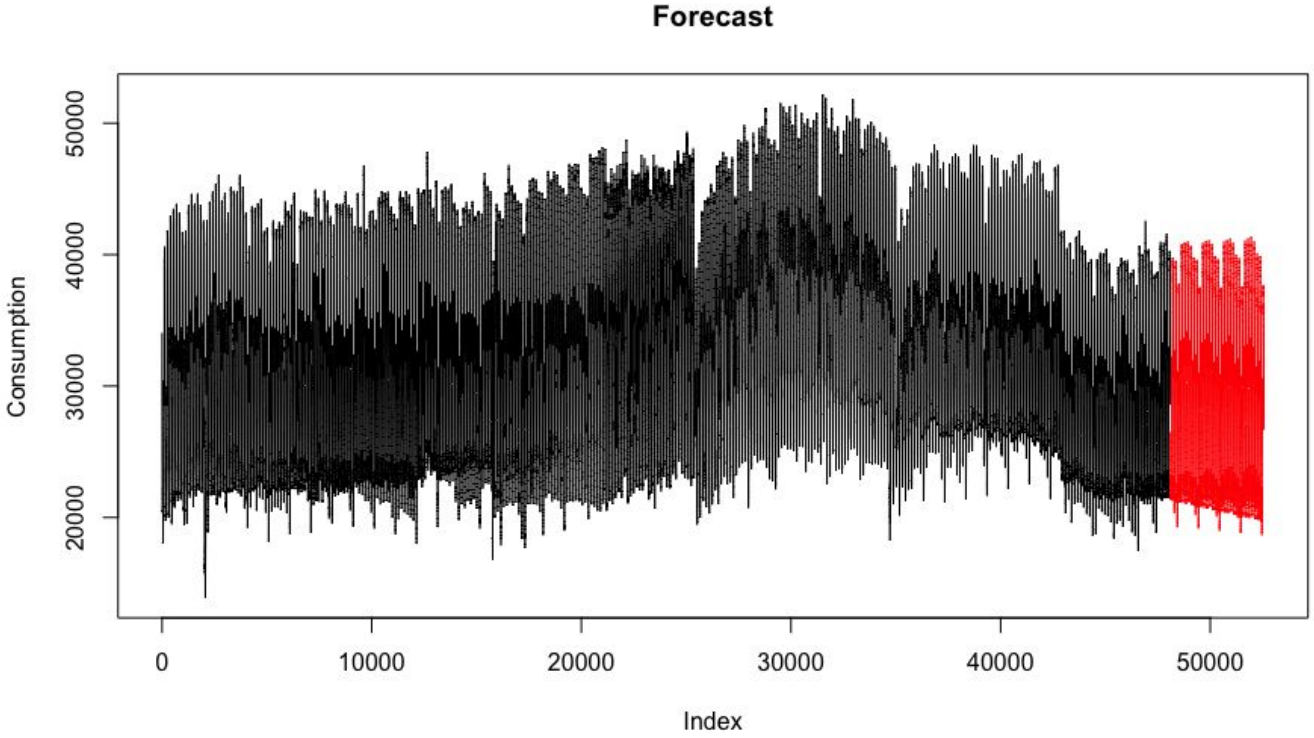
ARIMA

- Si è cercato il modello più adatto per i dati per un singolo orario e si è replicato il modello per tutti gli orari del giorno (144 modelli totali).
- Sono state create variabili dummies per rappresentare i giorni festivi e particolari outliers (26-27 giugno).
- Sono stati testati diversi modelli, tra cui i migliori:

Modello	MAE	MAPE	REG
$\text{ARIMA}(1, 0, 1)(2, 1, 0)_7$	1363.44	4.91	dummy
$\text{ARIMA}(1, 0, 1)(2, 1, 0)_7 + \text{drift}$	1321.74	4.79	dummy
$\text{ARIMA}(2, 1, 0)(0, 1, 1)_7$	1031.17	3.65	dummy

ARIMA

Si è scelto il terzo modello, ovvero ARIMA(2, 1, 0)$(0, 1, 1)_7$, che ha prodotto le seguenti predizioni mostrate in rosso:



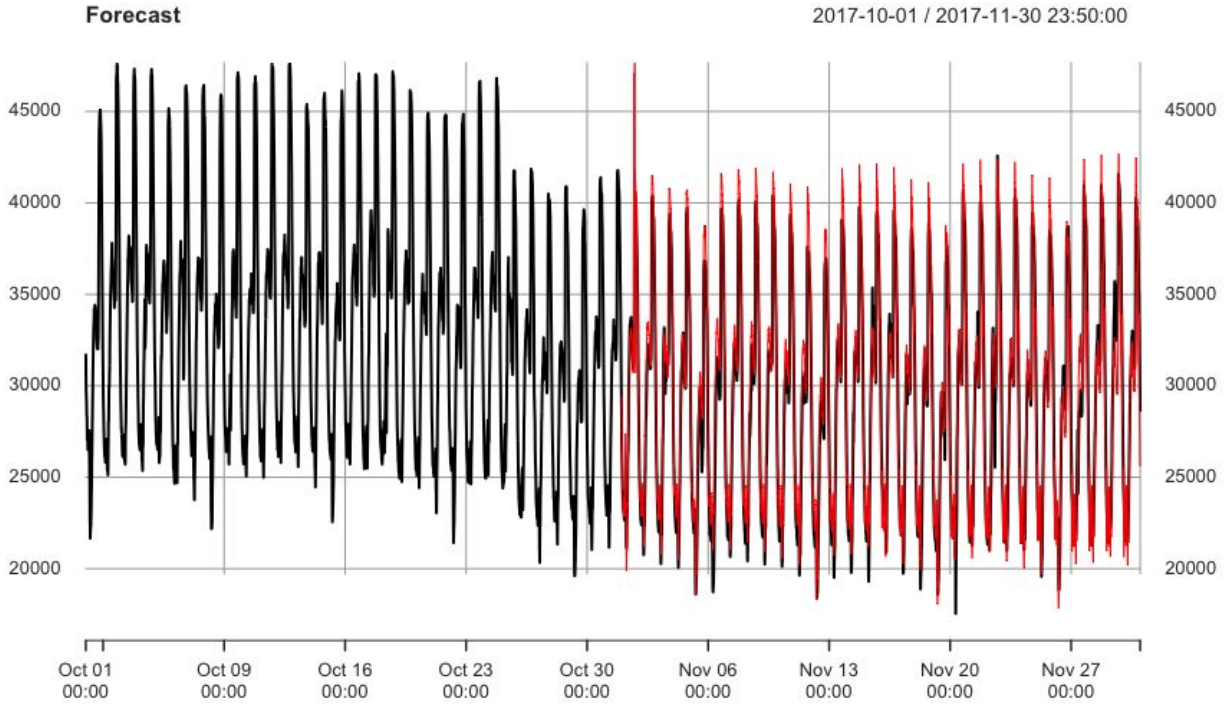
UCM

- Stesso approccio utilizzato per i modelli ARIMA.
- Utilizzo di variabili dummies per rappresentare i giorni festivi e particolari outliers.
- Di seguito alcuni dei modelli testati:

Modello	**MAE**	**MAPE**
Dummies + LLT + seasonal 7 dummy	1253.39	4.49
Dummies + LLT + seasonal 7 dummy + seasonal 365 trigonometric 8 harmonics	1198.71	4.28
Dummies + LLT + seasonal 7 dummy + seasonal 365 trigonometric 2 harmonics	1080.84	3.83

UCM

Di seguito vengono mostrate in rosso le previsioni per il mese di Novembre.



ML

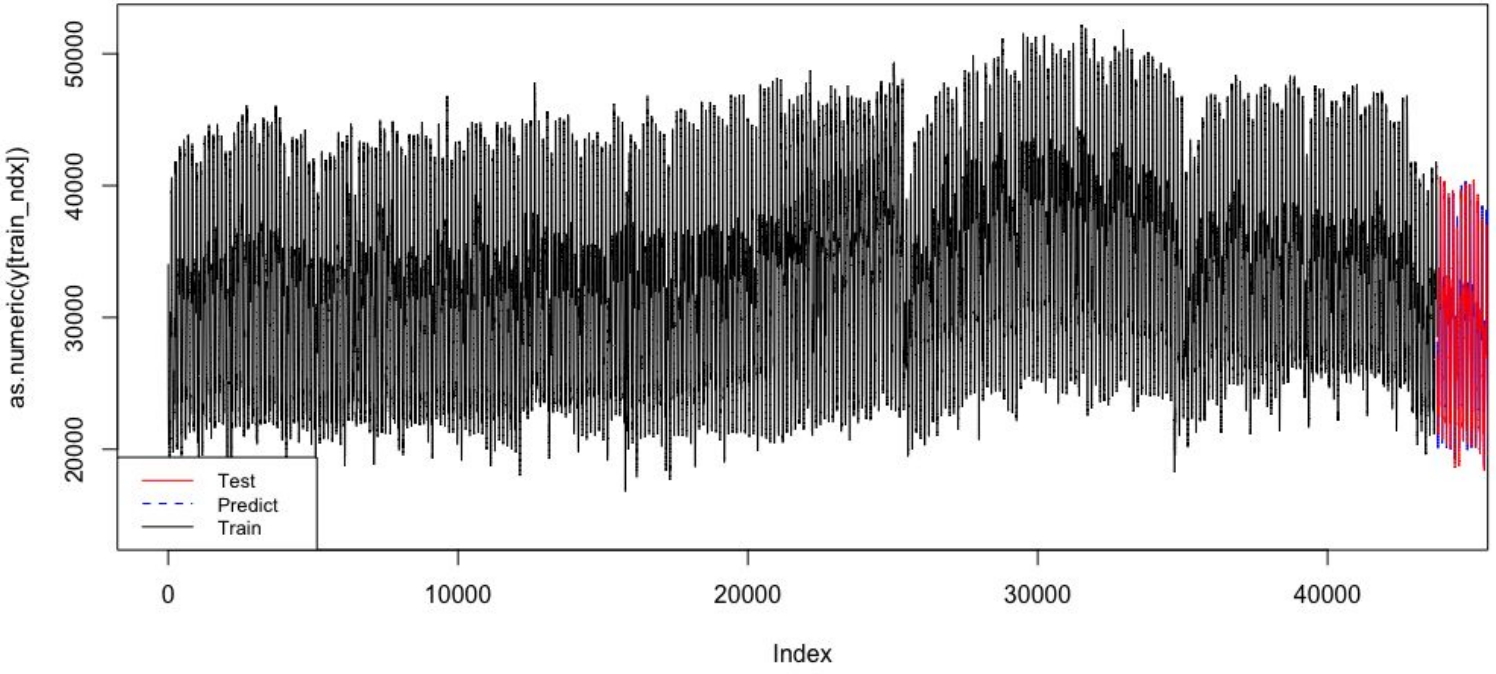
Si è deciso di testare i seguenti approcci:

- **KNN** - K-Nearest Neighbors;
  - Utilizzato sia per task di classificazione che per quelli di regressione.
  - Parametri utilizzati:
    - `K`, numero di Nearest Neighbors;
    - `H`, numero di valori da prevedere (orizzonte previsivo, impostato a 30 giorni);
    - `Lags`: numero di lags da considerare per effettuare le previsioni (stagionalità settimanale);
    - `msas`: assume valori "`recursive`" o "`MIMO`" (Multi Input Multi Output);
    - `cf`: combinazione utilizzata per aggregare i target vicini ("median", "weighted" o "mean");
    - `Transform`: trasformazioni sui campioni ("additive" "multiplicative" oppure "none").

- **LSTM** - Long Short-Term Memory.
  - Architettura basata su Recurrent Neural Network (RNN).
  - Preprocessing:
    - Normalizzazione dei valori nel range (0, 1);
    - Rimodellazione con una finestra di lags pari a 6*24*7;
    - Reshape del train set in un array 3d.

KNN

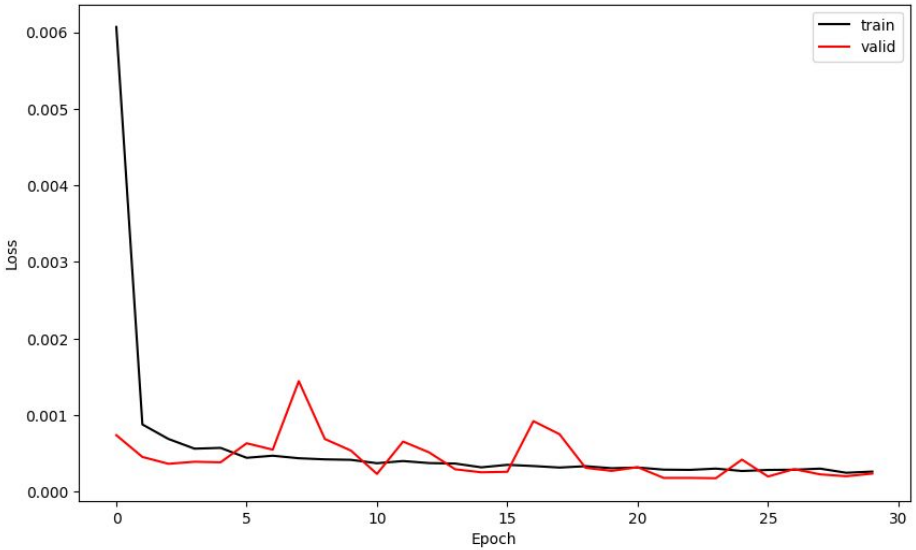
Modello	MAE	MAPE
k = 7, msas = recursive, cf = median, transform = additive	1805.27	6.30
k = 7, msas = MIMO, cf = weighted, transform = multiplicative	1535.39	5.14
k = 12, msas = MIMO, cf = mean, transform = multiplicative	1354.37	4.57



LSTM

La loss function (impostata con il Mean Squared Error) prodotta nella fase di train mostra una stabilizzazione rapida.

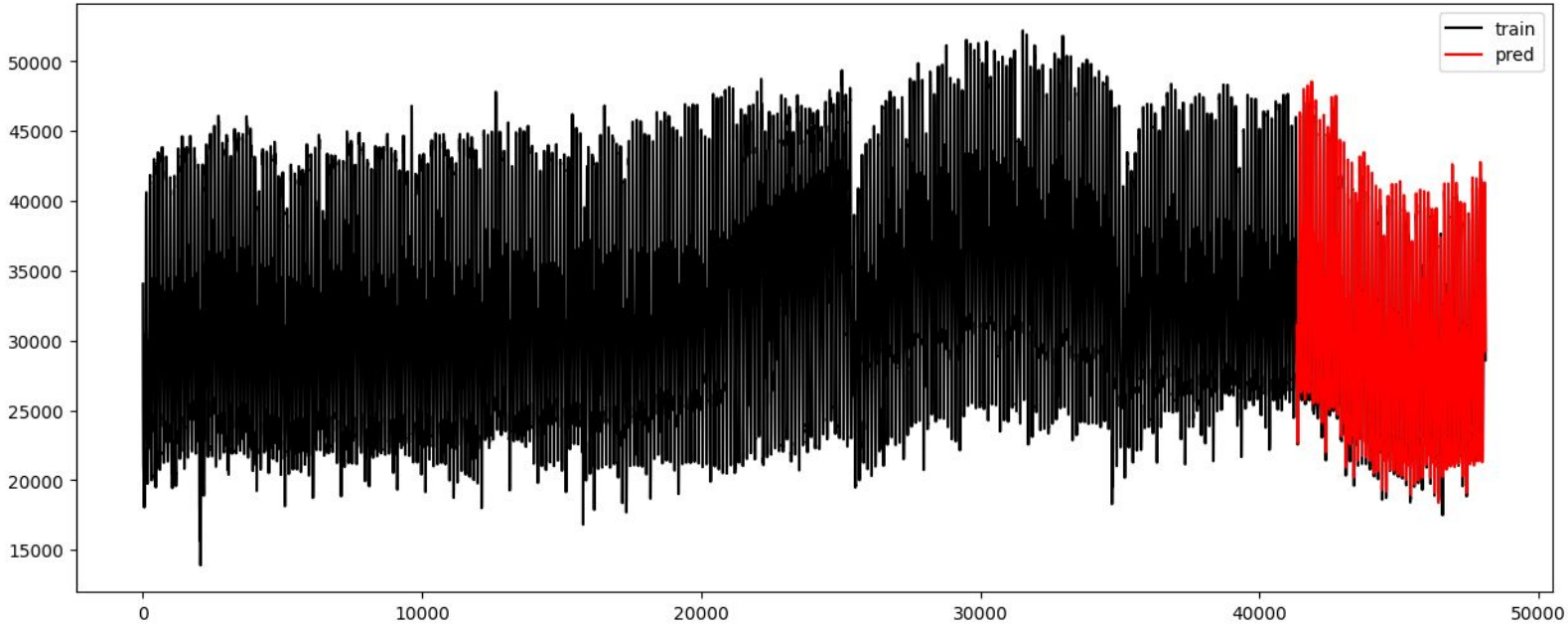
Potrebbe essere sintomo di **overfitting**, avvalorato dai valori di performance ottenuti.



Modello	MAE	MAPE
LSTM con 60 neuroni, un layer Dense con un neurone, 30 epoche, batch size = 64	465.01	0.02

LSTM

Di seguito vengono mostrate in rosso le previsioni per il mese di Novembre.



Conclusioni

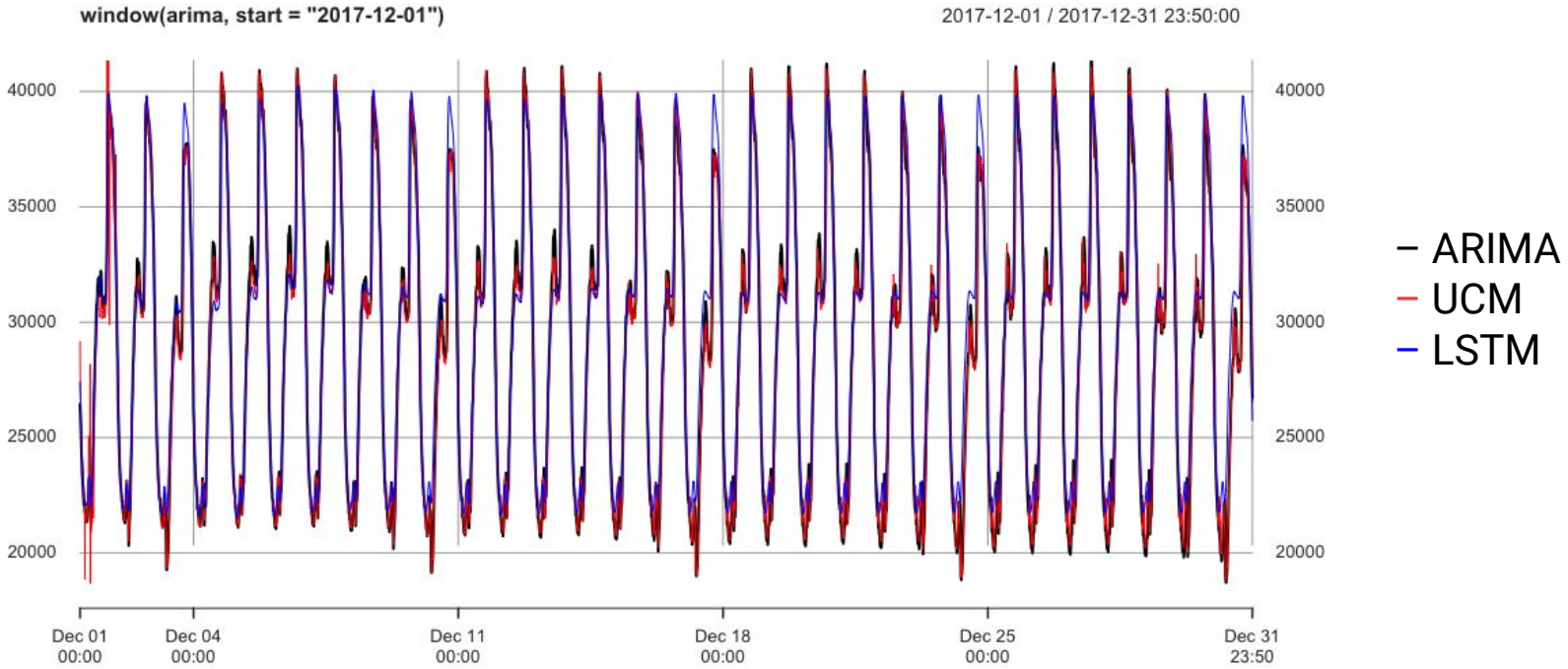
Dai risultati ottenuti sembrerebbe che il modello LSTM ottenga i risultati migliori ma rimane la perplessità che questo sia dovuto ad un overfitting sui dati di train.

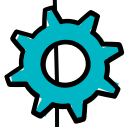
Modello	**MAE**	**MAPE**
ARIMA$(2, 1, 0)(0, 1, 1)_7$ con dummy	1031.17	3.65
Dummies + LLT + seasonal 7 dummy + seasonal 365 trigonometric 2 harmonics	1080.84	3.83
LSTM con 60 neuroni, un layer Dense con un neurone, 30 epoche, batch size = 64	465.01	0.02

Probabilmente effettuando il train su più dati e modificando la configurazione della rete utilizzando un ulteriore numero di layer o di neuroni o inserendo delle regolarizzazioni si potrebbero ottenere dei risultati più affidabili per le previsioni.

Conclusioni

Nel seguente grafico vengono mostrate le predizione dei tre modelli per il mese di Dicembre.





Grazie per l’attenzione!