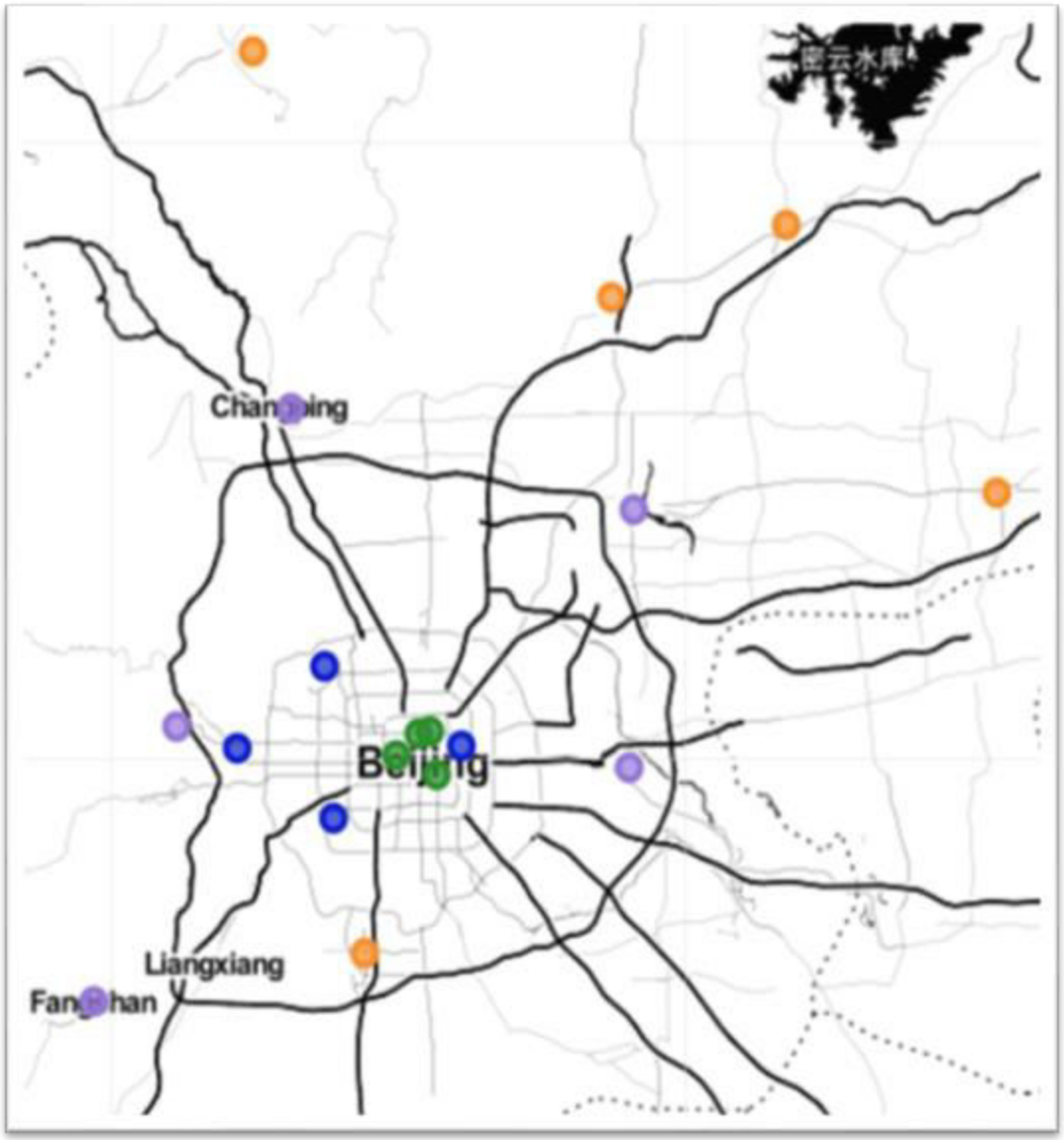
Progetto di Data Science Lab

Analisi e previsione di serie storiche

Presentazione di Giorgio Bini, Lorenzo Famiglini, Pranav Kasela
Università di Milano-Bicocca

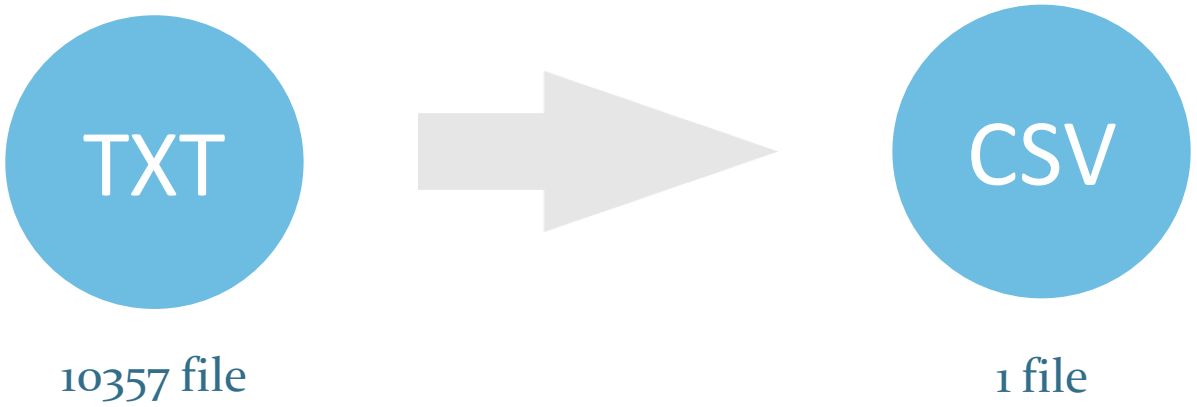
Introduzione al problema

- Dati relativi alle richieste di taxi in 18 distretti di Pechino.
- Dati a disposizione: la posizione GPS del taxi (latitudine, longitudine) e l'orario della richiesta.
- Periodo di riferimento: 2-8 Febbraio 2008
- Obiettivo: costruire dei modelli per la previsione del numero di richieste.



Il preprocessing

CONVERSIONE | AGGREGAZIONE | FEATURE ENGINEERING



Riduzione dei tempi

✖ Da 30 minuti...

✓ ...a 11 minuti

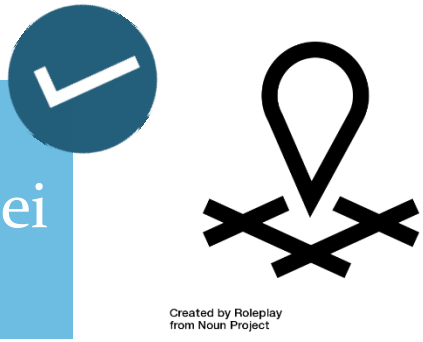
multiprocessing

Il preprocessing

CONVERSIONE | **AGGREGAZIONE** | FEATURE ENGINEERING

Latitudine, Longitudine → Nome del distretto

Calcolo della distanza minima tra la latitudine e la longitudine centrale dei distretti.



Il preprocessing

CONVERSIONE | AGGREGAZIONE | **FEATURE ENGINEERING**

È ora di mangiare? (Bool)

Numero di richieste

Periodo del Giorno (Mattina, pomeriggio, sera, notte)

È un giorno lavorativo? (Bool)

Giorno della settimana

Frequenze del seno (Le prime 24, aggregazione ogni 10 minuti)

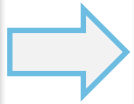
I primi 6 ritardi (più il 12-esimo e il 24-esimo)

C'è il sole? (Bool)

Il preprocessing

In sintesi...

```
1,2008-02-02 15:36:08,116.51172,39.92123
1,2008-02-02 15:46:08,116.51135,39.93883
1,2008-02-02 15:46:08,116.51135,39.93883
1,2008-02-02 15:56:08,116.51627,39.91034
1,2008-02-02 16:06:08,116.47186,39.91248
1,2008-02-02 16:16:08,116.47217,39.92498
1,2008-02-02 16:26:08,116.47179,39.90718
1,2008-02-02 16:36:08,116.45617,39.90531
1,2008-02-02 17:00:24,116.47191,39.90577
```



	taxi_count	working_or_not	is_sun_up	time_to_eat	sin_frequency_1	sin_frequency_2
0	25228	0	1	0	-1.470814e-15	1.000000e+00
1	26383	0	0	0	-1.470814e-15	1.000000e+00
2	24781	0	0	0	-1.470814e-15	1.000000e+00
3	25488	0	0	1	-2.204364e-15	1.102182e-15
4	24913	0	0	1	-2.204364e-15	1.102182e-15
5	24911	0	0	1	-2.204364e-15	1.102182e-15
6	25687	0	0	1	-2.204364e-15	1.102182e-15
7	27271	0	0	1	-2.204364e-15	1.102182e-15
8	24935	0	0	1	-2.204364e-15	1.102182e-15
9	28155	0	0	1	5.879543e-15	-1.000000e+00
10	28155	0	0	1	5.879543e-15	-1.000000e+00

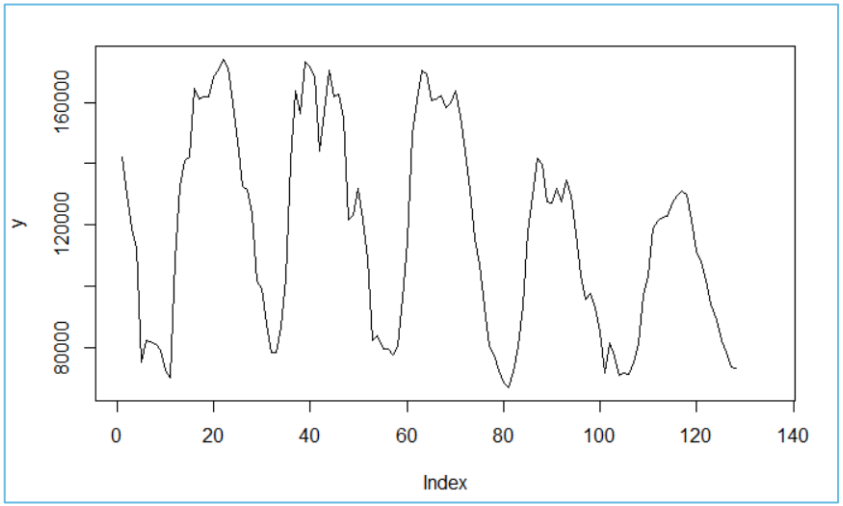
✖ Da questo... ✓ ...a questo

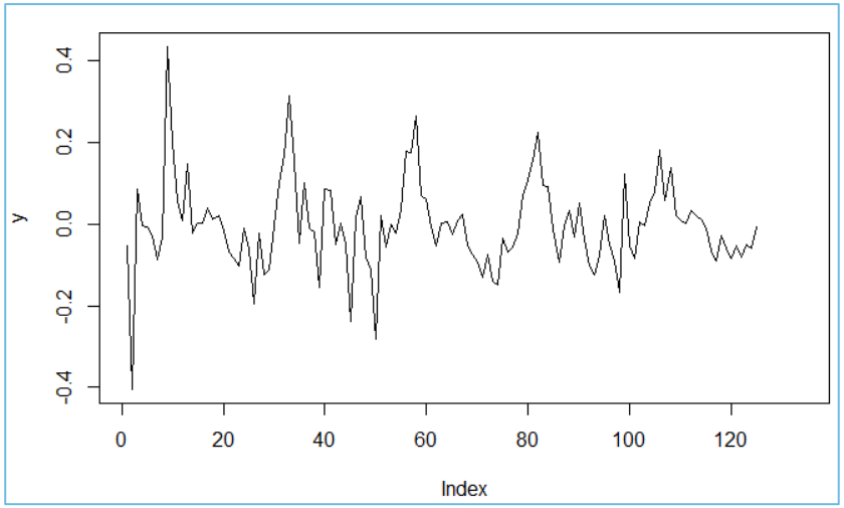
Numero di file	10357, uno per ogni taxi (0.7 GB)	1
Schema	TaxiID, DateTime, Latitudine, Longitudine	DateTime, Numero di Richieste, E' ora di mangiare?, Periodo del giorno, Giorno della settimana, Frequenze del Seno, Ritardi, C'è il sole? E' un giorno lavorativo?

Il modello ARIMA

Aggregazione:
un'osservazione ogni ora.

Previsione:
le ultime 10 ore.



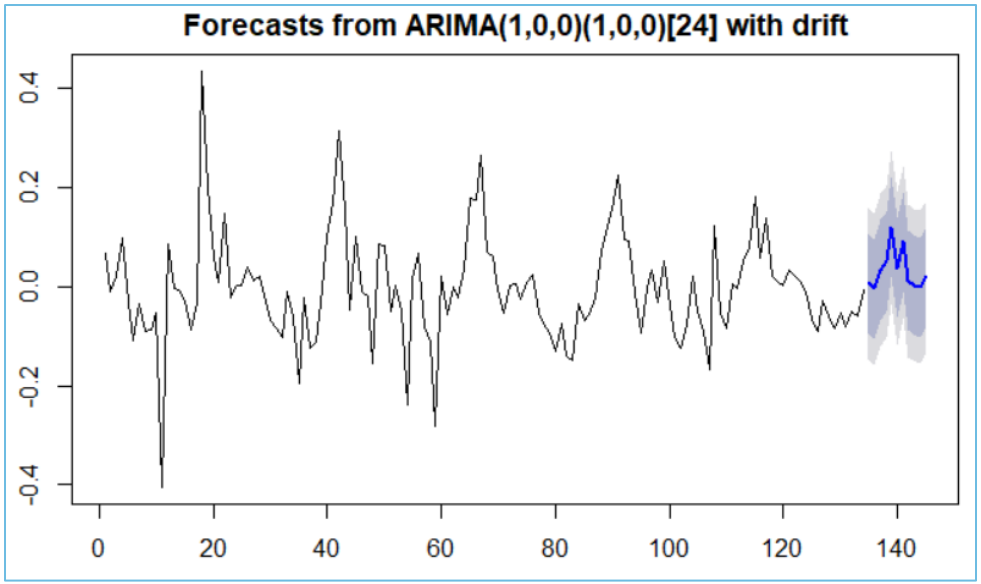


Alla variabile y è stata applicata una trasformazione logaritmica e un differenziale orario per ottenere stazionarità.

Il modello ARIMA

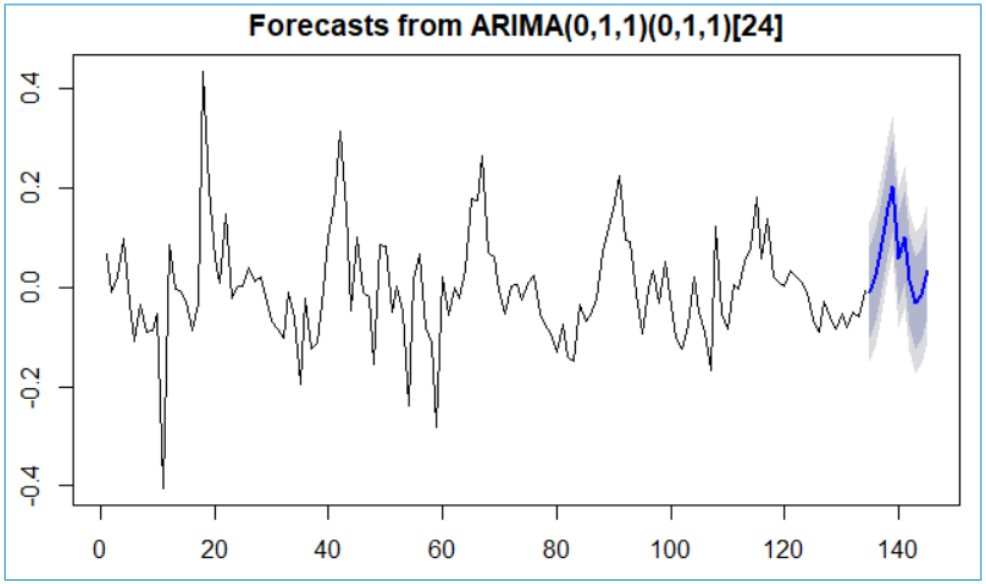
AR(1) SAR(1) (1,0,0) (1,0,0)

AIC = -282
MAE = 0.07



Airline Model

AIC = -247
MAE = 0.06

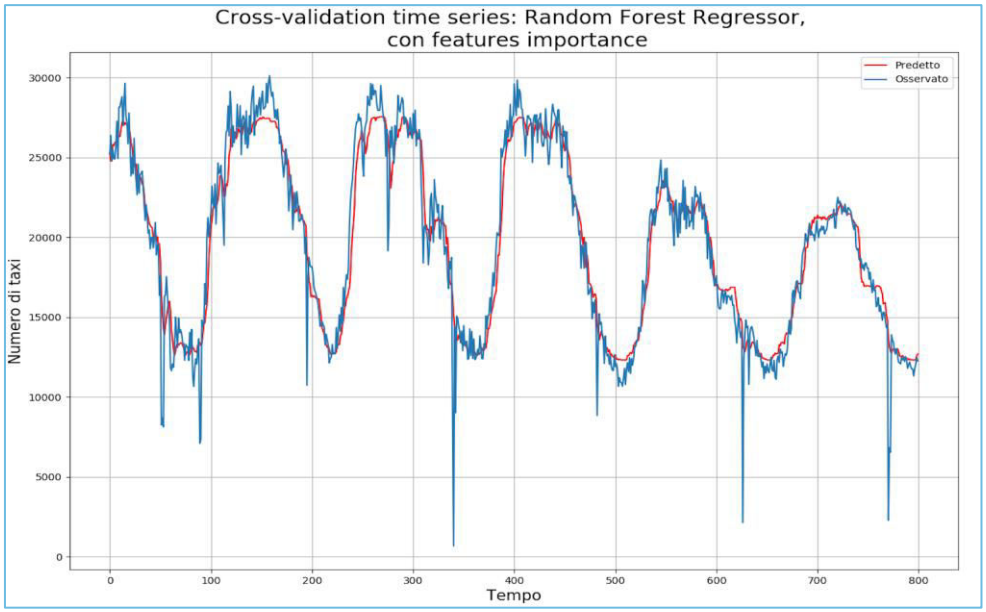


Random Forest Regressor

Holdout

- Train Size: 800 (8000 minuti)
- Test Size: [illegible] minuti
- Tra[illegible] = 1
- Train MAE = 0
- Test R^2 = 0.83
- Test MAE = 0.17

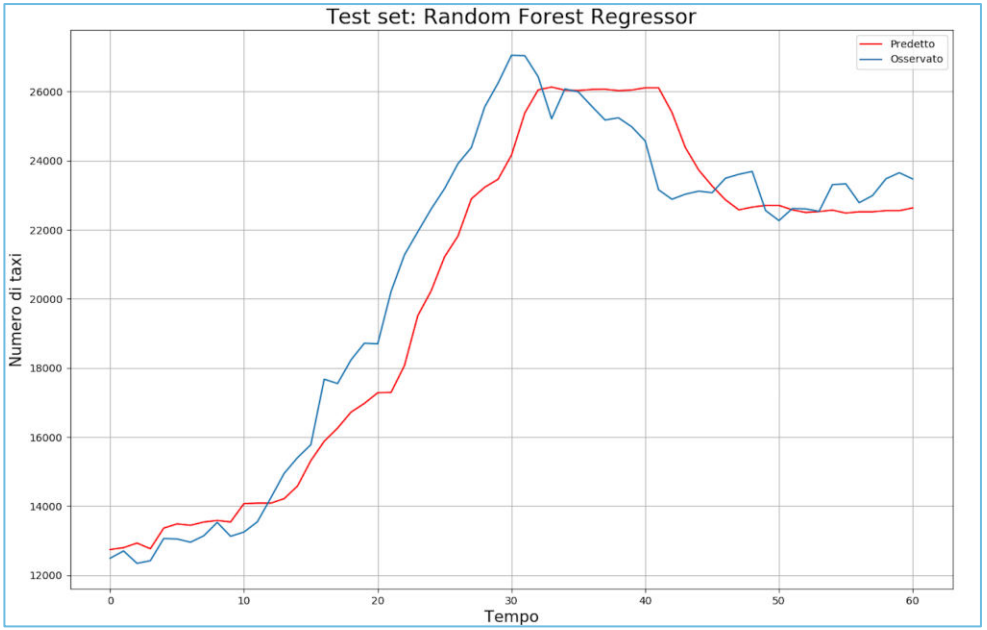
Overfitting



TS-CV, Test set

Tuning dei parametri

- Features Importance
- CV size: 800
- Test size: 61
- CV R^2 = 0.91
- CV MAE = 0.1
- Test R^2 = 0.92
- Test MAE = 0.07

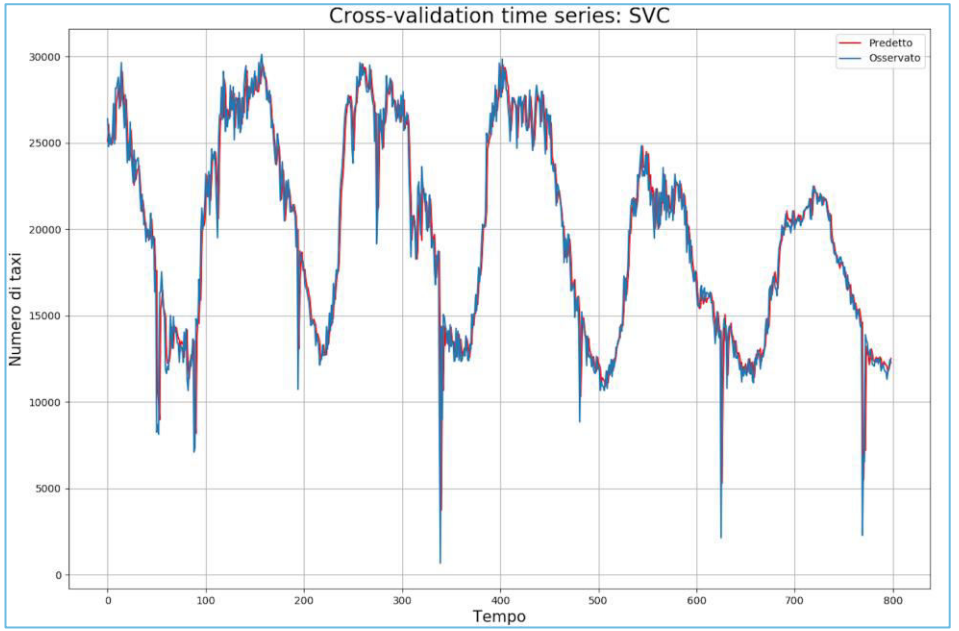


Elastic Net

Holdout

Lambda = 0.1, l1_ratio = 0.5

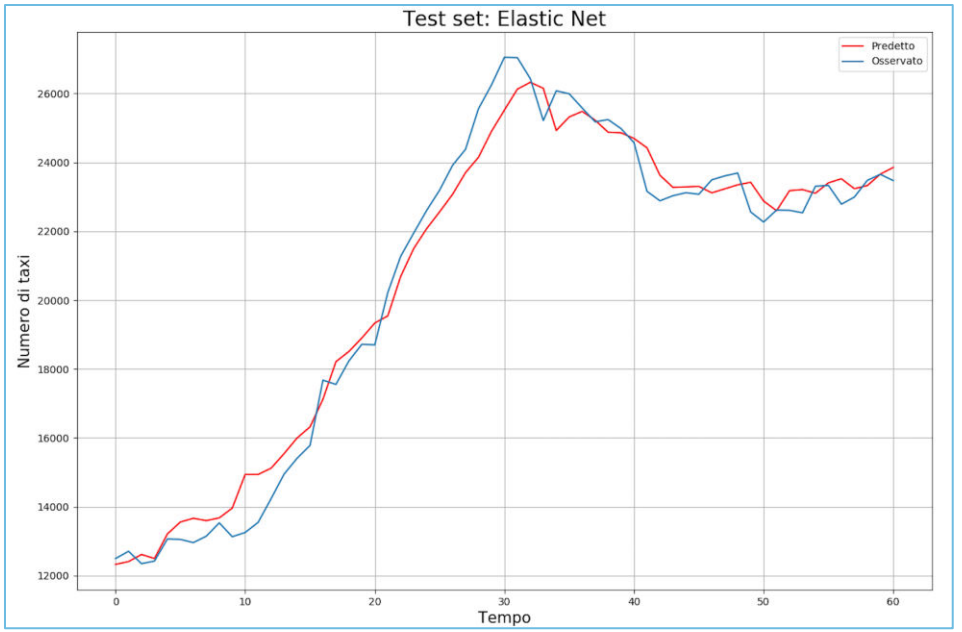
- Train Size: 800 (8000 minuti)
- Test Size: 61 (610 minuti)
- Train R^2 = 0.93
- Train MAE = 0.08
- Test R^2 = 0.91
- Test MAE = 0.09



TS-CV, Test set

Lambda = 0.01, l1_ratio = 1

- CV size: 800
- Test size: 61
- CV R^2 = 0.93
- CV MAE = 0.08
- Test R^2 = 0.97
- Test MAE = 0.04



Ridge

Holdout

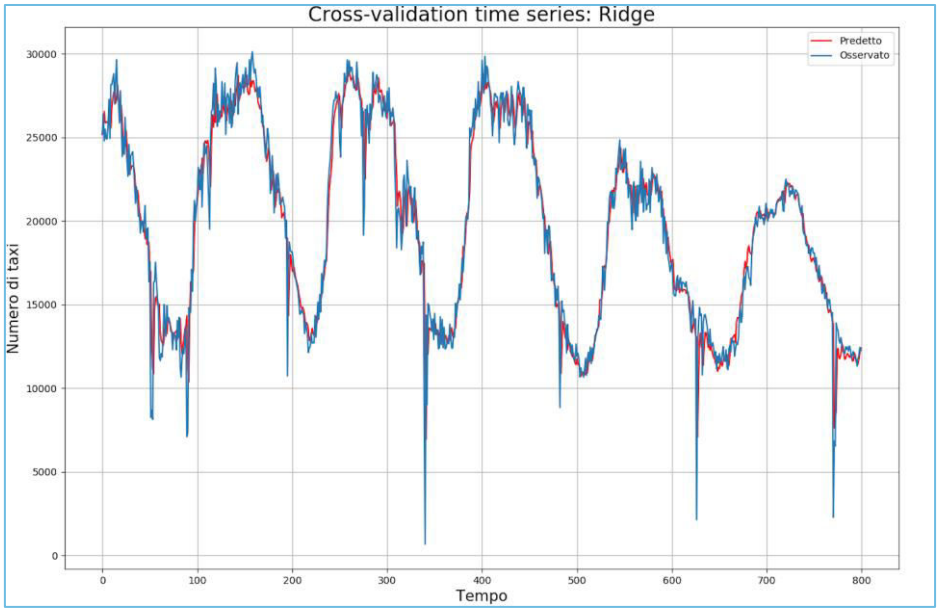
Alpha = 0.1

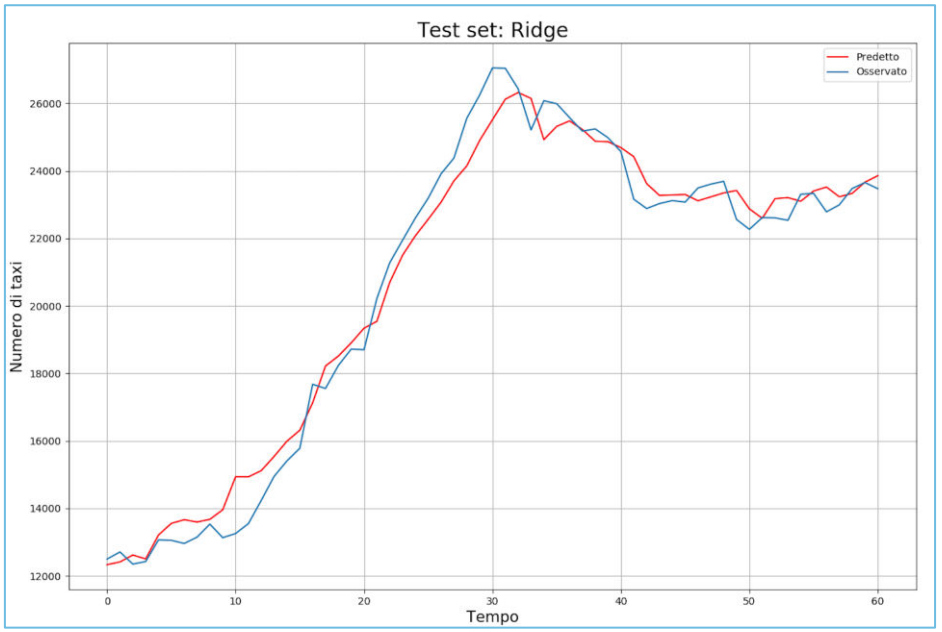
- Train Size: 800 (8000 minuti)
- Test Size: 61 (610 minuti)
- Train R^2 = 0.93
- Train MAE = 0.09
- Test R^2 = 0.98
- Test MAE = 0.03

TS-CV, Test set

Alpha = 0.1

- CV size: 800
- Test size: 61
- CV R^2 = 0.93
- CV MAE = 0.08
- Test R^2 = 0.98
- Test MAE = 0.02





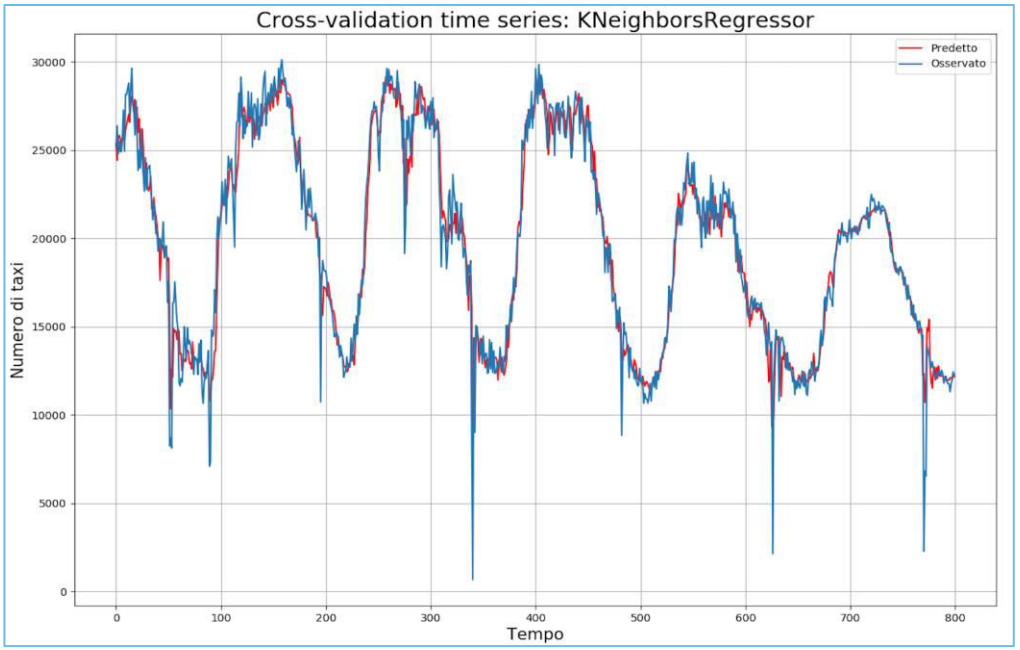
Kneighbors-Regressor

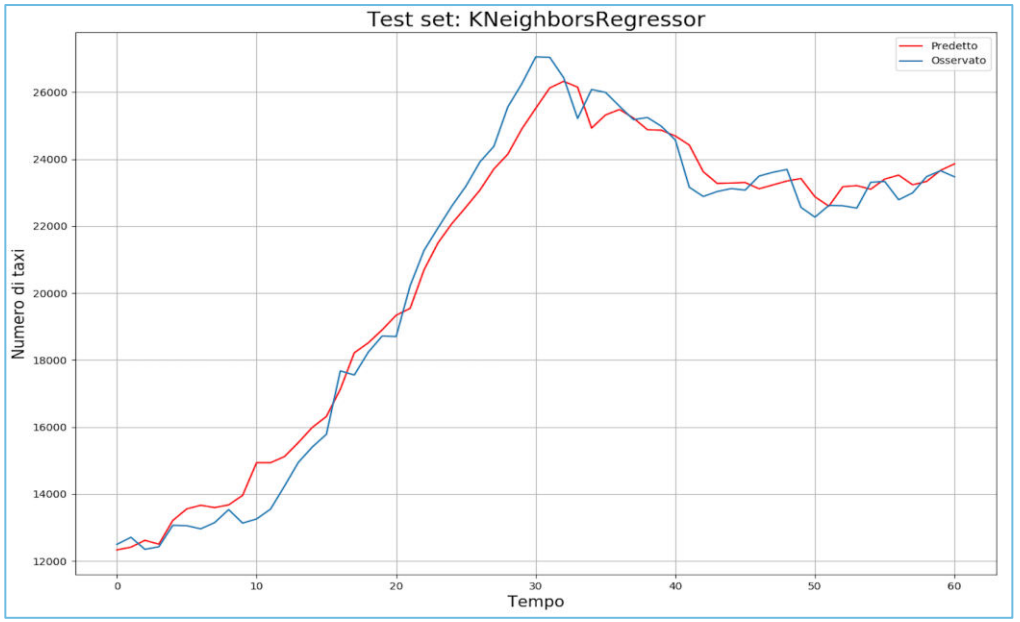
Holdout
N = 5

- Train Size: 800 (8000 minuti)
- Test Size: 61 (610 minuti)
- Train R^2 = 0.94
- Train MAE = 0.07
- Test R^2 = 0.94
- Test MAE = 0.04

TS-CV, Test set
N = 5, metrica = Euclidean

- CV size: 800
- Test size: 61
- CV R^2 = 0.95
- CV MAE = 0.04
- Test R^2 = 0.94
- Test MAE = 0.03

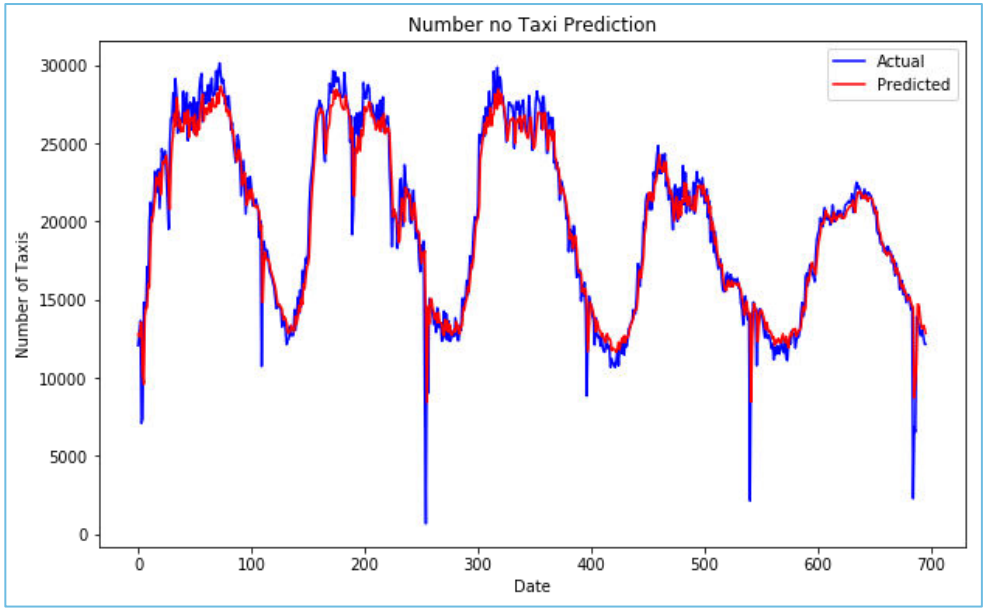




LSTM - Univariate feature set

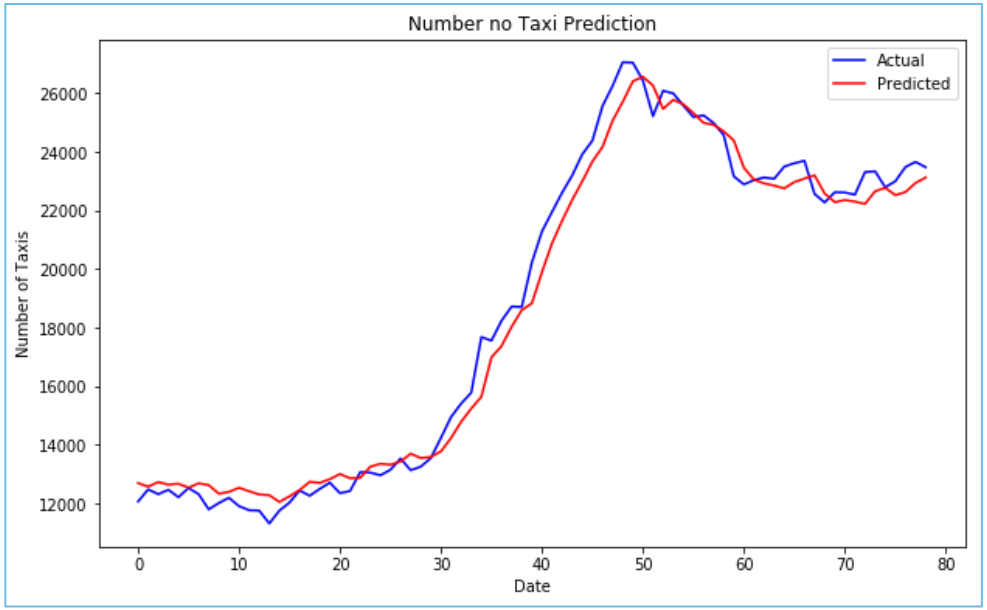
Training Set

- Train Size: 800 * 10 minuti (80 rimossi a causa dei ritardi)
- Epochs: 150
- Tempo di allenamento: 15 minuti
- R^2 = 0.92
- MAE = 0.09



Test Set

- Test Size: 800*10 minuti
- R^2 = 0.98
- MAE = 0.03

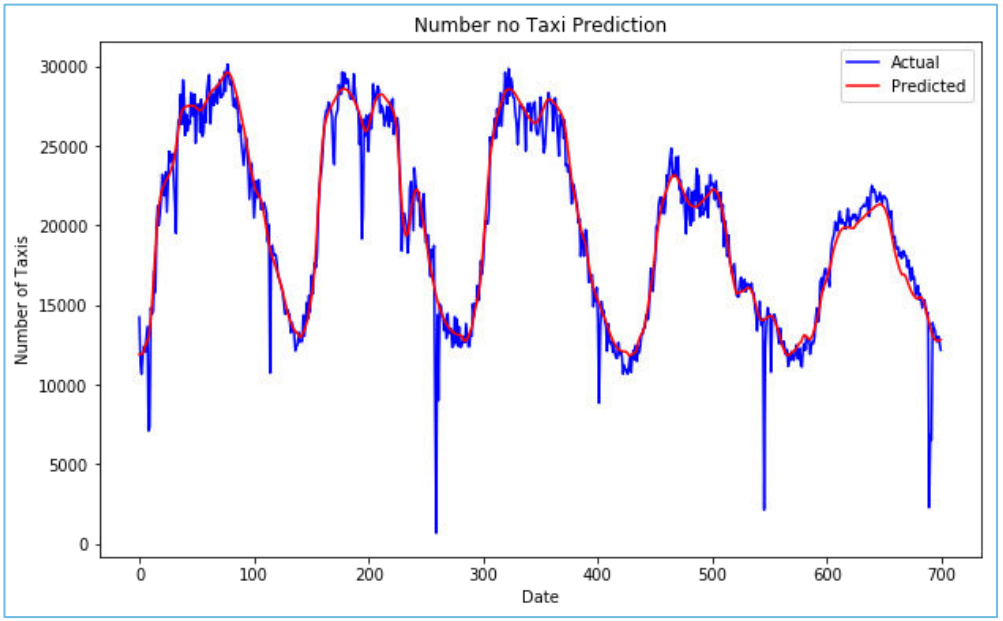


Non c'è underfitting a causa dei picchi nel training set.

LSTM - Univariate feature set

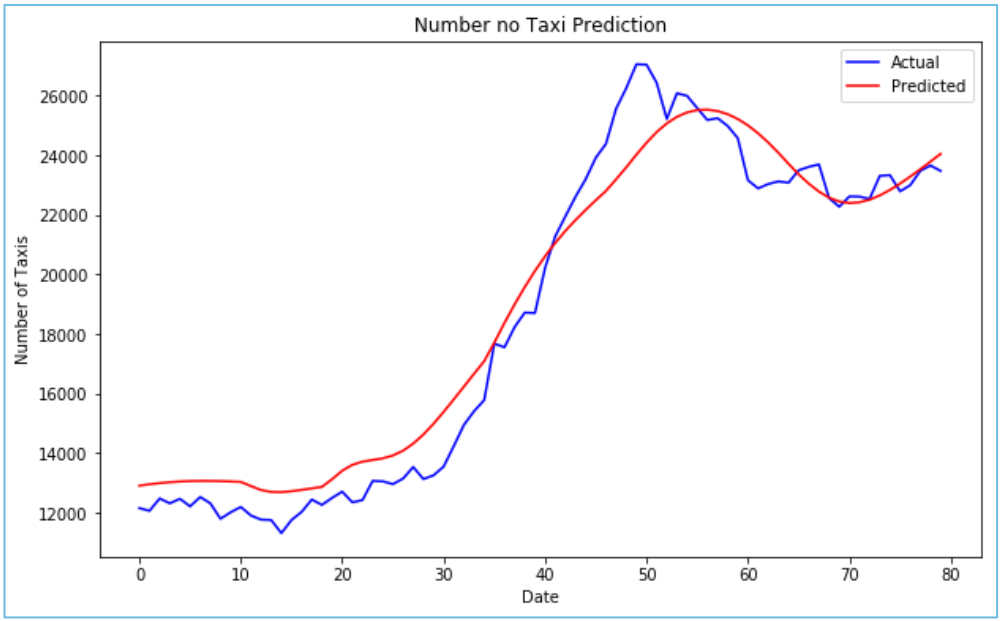
Training Set

- Train Size: 800 * 10 minuti (80 rimossi a causa dei ritardi)
- Epochs: 50
- Tempo di allenamento: 22 minuti
- R^2 = 0.89
- MAE = 0.11



Test Set

- Test Size: 800*10 minuti
- R^2 = 0.96
- MAE = 0.05



Non c'è underfitting per lo stesso motivo di prima.

Grazie per l’attenzione!