Analisi e risultati per business

Colloquio per BitBang - dicembre 2017

Sommario

- Analisi esplorativa, con particolare attenzione alle variabili target
- Previsione con ARIMA stagionali
- Previsione con Recurrent Neural Network
- Risultati
- Conclusione e possibili estensioni

Tabella KPIs

- Dati giornalieri in un periodo compreso tra marzo 2015 e ottobre 2017 (932 record totali)
- 11 variabili numeriche, con una prima ovvia distinzione tra variabili intere (variabili K1-K8) e variabili con decimali (variabili K9-K11)
- Senza conoscenza del contesto, potremmo essere portati a pensare che il primo gruppo rappresenti variabili di conteggio, mentre il secondo delle metriche ricavate da altre variabili
- Le variabili target per questa analisi sono **K5** e **K11**

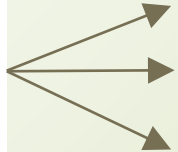
Tabella Spending

- Dati aggregati mensili in un periodo compreso tra gennaio 2015 e dicembre 2018 (48 record totali)
- 3 variabili numeriche, di cui la terza è la somma delle precedenti 2 (E3 = E1 + E2)
- La presenza di dati del futuro potrebbe far pensare ad obiettivi di spesa da raggiungere

Calcolo della spesa media per giorno

- Poiché le variabili target fanno riferimento alla tabella KPIs, è possibile usare i dati di Spending associando la media delle spese sul numero di giorni di quel mese
- Il processo è stato eseguito solo per E1 e per E2 per evitare ridondanze

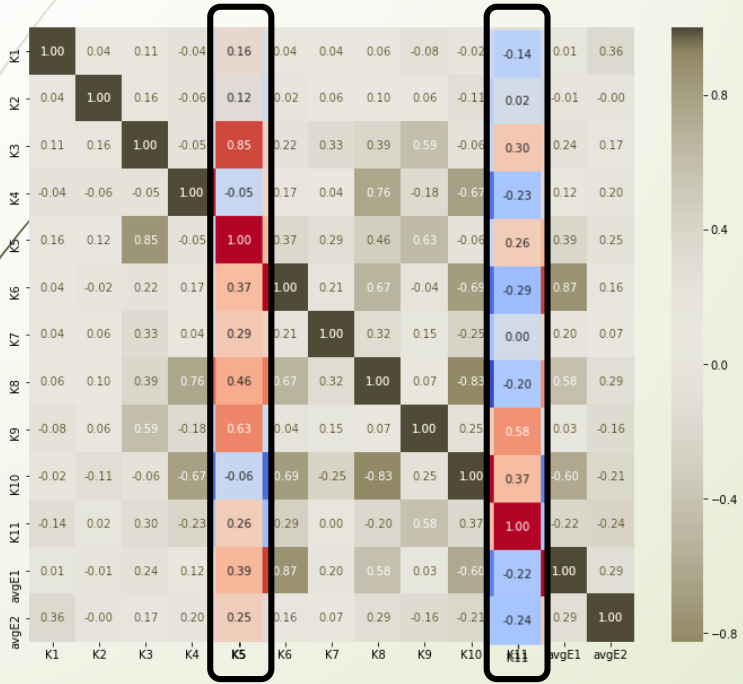
Date	E1	E2	days_in_month
2015-01-01	14830	0	31
2015-02-01	12262	137900	28
2015-03-01	16434	136505	31
2015-04-01	17130	0	30
2015-05-01	18732	1147	31



	Date
0	2015-03-16
1	2015-03-17
2	2015-03-18
3	2015-03-19
4	2015-03-20
5	2015-03-21
6	2015-03-22
7	2015-03-23
8	2015-03-24
9	2015-03-25
10	2015-03-26
11	2015-03-27
12	2015-03-28

Date	avgE1	avgE2
2015-03-16	530.129032	4403.387097
2015-03-17	530.129032	4403.387097
2015-03-18	530.129032	4403.387097
2015-03-19	530.129032	4403.387097
2015-03-20	530.129032	4403.387097
2015-03-21	530.129032	4403.387097
2015-03-22	530.129032	4403.387097
2015-03-23	530.129032	4403.387097
2015-03-24	530.129032	4403.387097
2015-03-25	530.129032	4403.387097
2015-03-26	530.129032	4403.387097
2015-03-27	530.129032	4403.387097
2015-03-28	530.129032	4403.387097

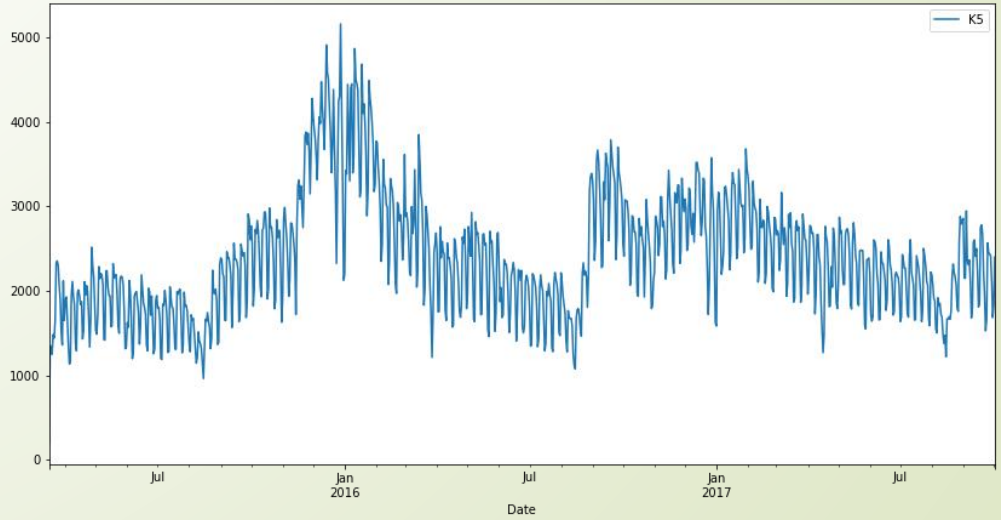
K11 sembra essere poco correlata con le altre variabili



- Dalla tabella delle correlazioni si evince che K5 è correlata con diverse variabili, in particolar modo con le variabili (in ordine decrescente di correlazione) K3, K9, K8 e K6.
- Per quanto riguarda la variabile K11, sembrano esserci meno variabili correlate con essa. La correlazione è evidente per K9 e K10.

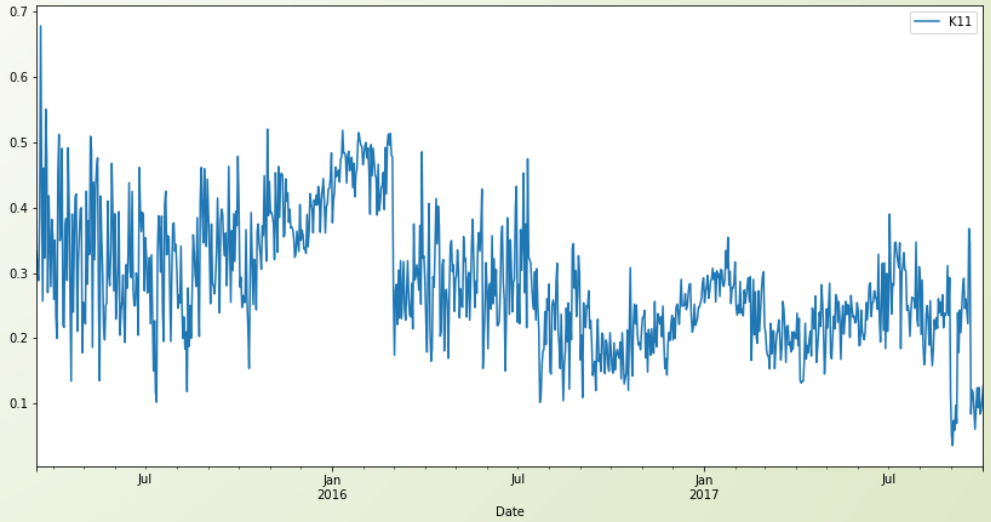
K5 ha un forte trend positivo nella seconda metà del 2015

- In seguito al picco, il trend diminuisce nella prima metà del 2016 e non si ripete nell'anno successivo
- Sembrano esserci degli outliers che si discostano di molto dalla linea di trend. Questi potrebbero ricondotto alla stagionalità della serie storica (con un'ipotesi di trend settimanale)



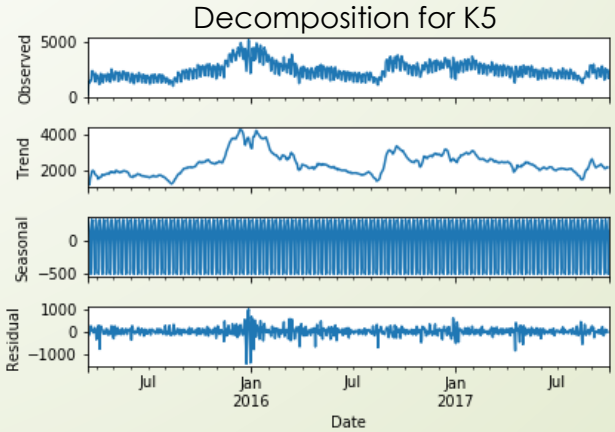
K11 si comporta in modo instabile

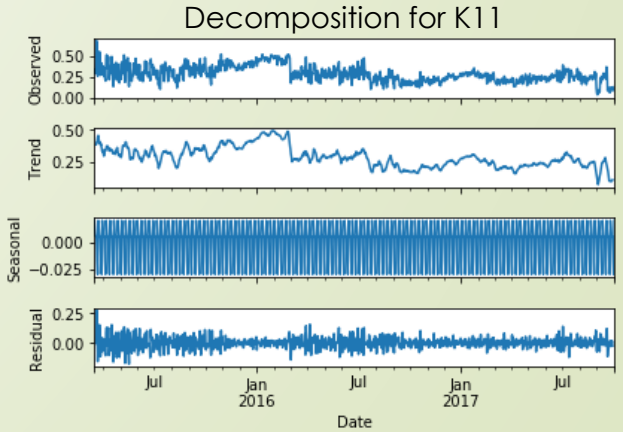
- L'ampiezza della serie storica varia in modo non prevedibile, all'inizio e nella parte centrale sembra essere di diversa intensità rispetto al resto della serie
- E' possibile notare un punto di discontinuità in prossimità di marzo 2016



K5 si mantiene stazionaria nel tempo, K11 no

- Una condizione importante per fare inferenza sulle serie storiche è che queste siano **stazionarie**
- Ad un primo sguardo la variabile K5 sembra stazionaria nonostante l'andamento irregolare, mentre lo stesso non si può dire per K11 che mostra un trend decrescente
- Tali ipotesi vengono confermate per entrambe dall'**augmented Dicker-Fuller test** per testare la stazionarità, con p-value rispettivamente di 0.0412 per K5 e 0.4413 per K11 contro l'ipotesi nulla





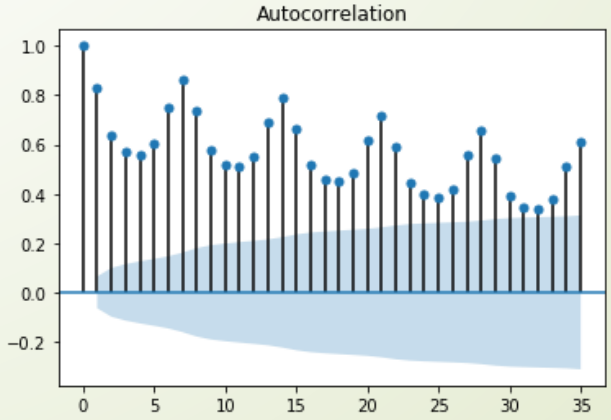
Idea: previsione di valori futuri per K5 e K11

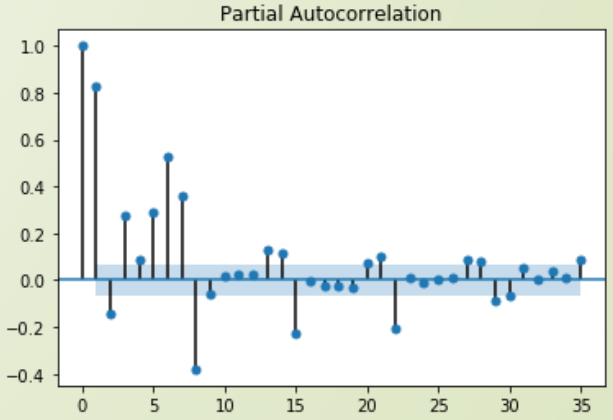
- Viste le caratteristiche numeriche di entrambe le serie, potrebbe essere interessante capire se si può stimare un modello che permetta la previsione di valori futuri in modo efficace
- Un modello di previsione di questo tipo può essere utile soprattutto in contesti di pianificazione, ad esempio per programmare attività di marketing che permettano di superare il target previsto dall'andamento naturale della serie
- Verranno tentati due approcci:
  - Approccio parametrico: Modello ARIMA stagionale
  - Approccio non parametrico: Recurrent Neural Network

Modello Arima per K5

K5 presenta una stagionalità di frequenza settimanale

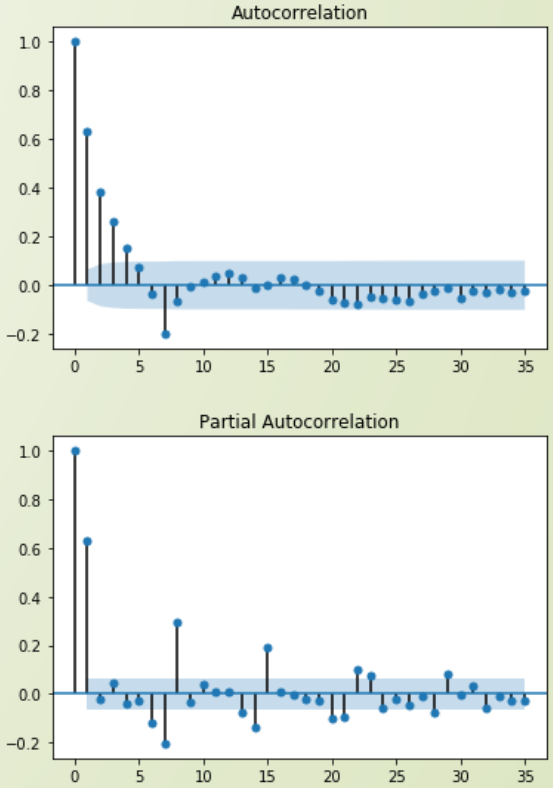
- Prendendo in considerazione la sola serie per K5, si possono calcolare sia le autocorrelazioni (**acf**) che le autocorrelazioni parziali (**pacf**), che permettono di capire quanto il valore ad una certa data dipenda dai giorni precedenti (**ritardo** o **lag**)
- Dal grafico a sinistra si deduce che la stagionalità per questa serie è effettivamente settimanale, visto che i picchi sono in corrispondenza dei ritardi multipli di 7





Il modello usato è un ARIMA$(1, 0, 0)(0, 0, 1)_7$

- Prima di appurare quale modello potrebbe essere adatto, si può destagionalizzare la serie, sottraendo i valori della serie stessa a ritardo 7
- La presenza di una forte **pacf** a lag 1 e un decadimento graduale dell'**acf** è indice di una componente autoregressiva nella serie
- La situazione opposta per i ritardi settimanali (multipli di 7), suggerisce invece la presenza di una componente a media mobile stagionale
- Il modello risultante sarà quindi un **modello ARIMA$(1, 0, 0)(0, 0, 1)_7$**



Riassunto per il modello ARIMA per K5

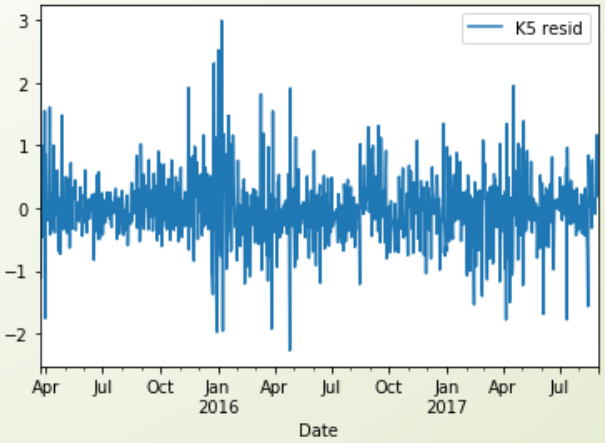
	coef	std err	z	P>|z|	[0.025	0.975]
K3	0.2936	0.016	18.430	0.000	0.262	0.325
K6	0.1935	0.033	5.931	0.000	0.130	0.257
K8	0.2550	0.032	7.897	0.000	0.192	0.318
K9	0.0994	0.022	4.545	0.000	0.057	0.142
ar.L1	0.8382	0.015	56.146	0.000	0.809	0.867
ma.S.L7	-0.8447	0.013	-63.614	0.000	-0.871	-0.819
sigma2	0.3058	0.010	31.913	0.000	0.287	0.325

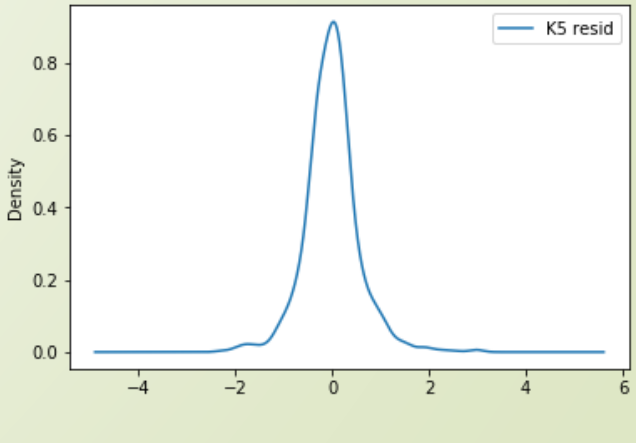
Ljung-Box (Q):	85.57	**Jarque-Bera (JB):**	398.19
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.87	**Skew:**	0.34
Prob(H) (two-sided):	0.23	**Kurtosis:**	6.20

- I valori dei coefficienti e dei relativi test di nullità degli stessi (colonna **z**) rassicurano sul fatto che tutti i parametri stimati nel modello sono utili al modello stesso
- La bonta' del modello viene misurata dal criterio **AIC**, che, con un valore di 1495, risulta essere il miglior risultato tra i modelli ARIMA simili a questo

La distribuzione dei resuidui viola l'ipotesi di normalità

- Nei dati considerati per stimare il modello, l'analisi dei residui dopo la stima del modello mostra criticità come si può osservare dai grafici sottostanti
- In particolare, applicando il **test Shapiro-Wilk** per la normalità, si ottiene un p-value pari a 0, confermando la distribuzione non normale per i residui





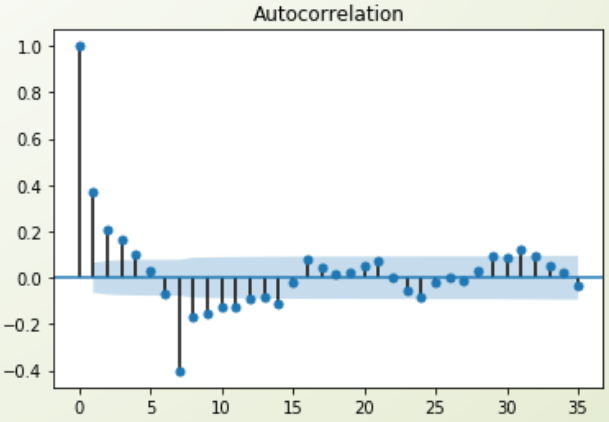
La previsione attraverso il modello ARIMA segue l'andamento con qualche incertezza

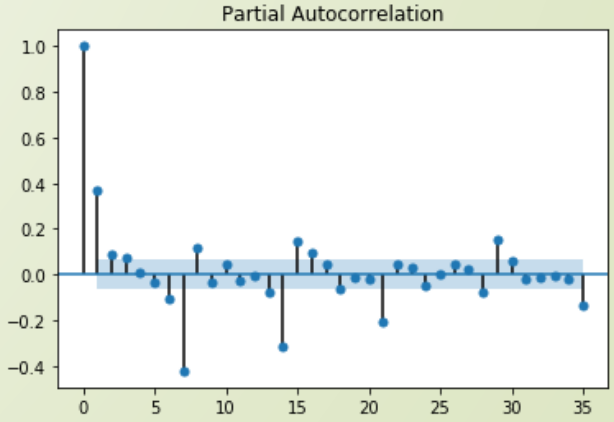
- Sebbene il modello colga l'andamento della serie, ci sono situazioni in cui la variabilità dei dati previsti è più ampia di quanto la serie non faccia originariamente

Modello Arima per K11

Anche K11 può essere stimata da un ARIMA(1, 0, 0)(0, 0, 1)$_7$

- A causa della mancanza di stazionarietà per K11, si forza una differenziazione a ritardo 1 e, fortunatamente, il p-value per testare l'ipotesi di stazionarietà è pressoché zero.
- I grafici di **acf** e di **pacf** prospettano una situazione simile a quanto visto per K5, nonostante l'orientamento opposto nel grafico delle autocorrelazioni parziali
- Possiamo quindi provare a stimare un **modello ARIMA(1, 0, 0)(0, 0, 1)$_7$**





Riassunto per il modello ARIMA per K11

	coef	std err	z	P>|z|	[0.025	0.975]
K9	0.2256	0.027	8.244	0.000	0.172	0.279
K10	0.0336	0.030	1.103	0.270	-0.026	0.093
ar.L1	0.4219	0.027	15.548	0.000	0.369	0.475
ma.S.L7	-0.8809	0.018	-49.642	0.000	-0.916	-0.846
sigma2	0.5430	0.019	28.169	0.000	0.505	0.581

Ljung-Box (Q):	135.99	**Jarque-Bera (JB):**	236.30
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.34	**Skew:**	-0.47
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.34

- La matrice di correlazione sulle serie differenziate a ritardo 1 mostra come nessuna variabile sia correlata con K11, ad eccezione di K9
- Questo viene confermato dalla ridondanza dal parametro stimato per la variabile K10, che mostra p-value alto contro l'ipotesi nulla di non nullità del parametro

Sommario per il modello ARIMA per K11 (senza K10)

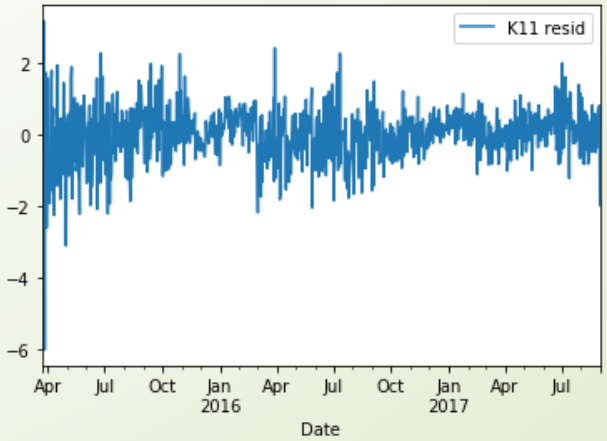
	coef	std err	z	P>|z|	[0.025	0.975]
K9	0.2273	0.027	8.271	0.000	0.173	0.281
ar.L1	0.4281	0.027	15.813	0.000	0.375	0.481
ma.S.L7	-0.8841	0.018	-50.209	0.000	-0.919	-0.850
sigma2	0.5436	0.019	28.052	0.000	0.506	0.582

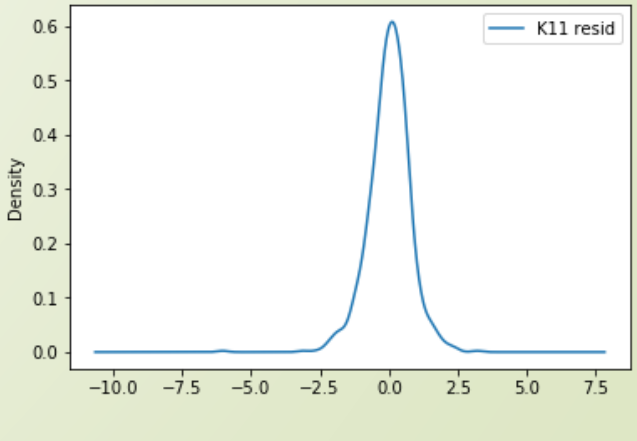
Ljung-Box (Q):	135.09	**Jarque-Bera (JB):**	224.91
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.34	**Skew:**	-0.44
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.29

- Togliendo K10 dalle variabili per la stima del modello, la situazione si presenta più solida
- Tuttavia il miglioramento sul criterio AIC per la bontà del modello è appena percettibile: 2008 per questo modello, contro 2009 per il modello precedente

Anche in questo caso la distribuzione dei resuidui viola l'ipotesi di normalità

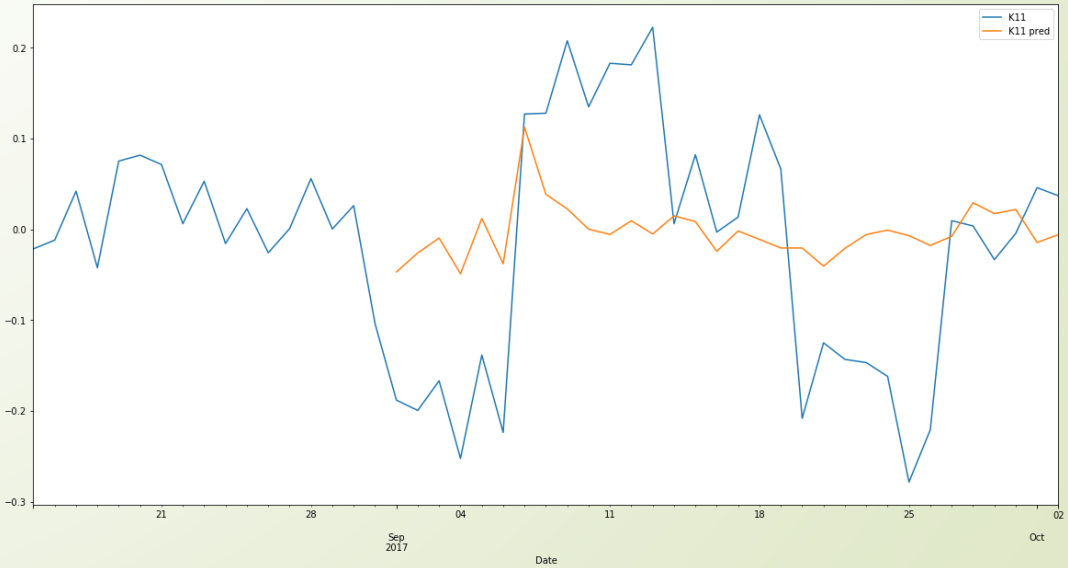
- Come per il modello per K5, la distribuzione dei residui non sembra essere assimilabile ad una distribuzione normale, come da assunzione del modello
- Il **test Shapiro-Wilk** per la normalità calcola un p-value pari a 0, confermando quanto ipotizzato





La previsione non riesce a cogliere i repentini cambi di trend

- Per quanto i cambi di trend in questa porzione di serie potrebbero essere dovuti al caso, la previsione mantiene semplicemente la linea di trend



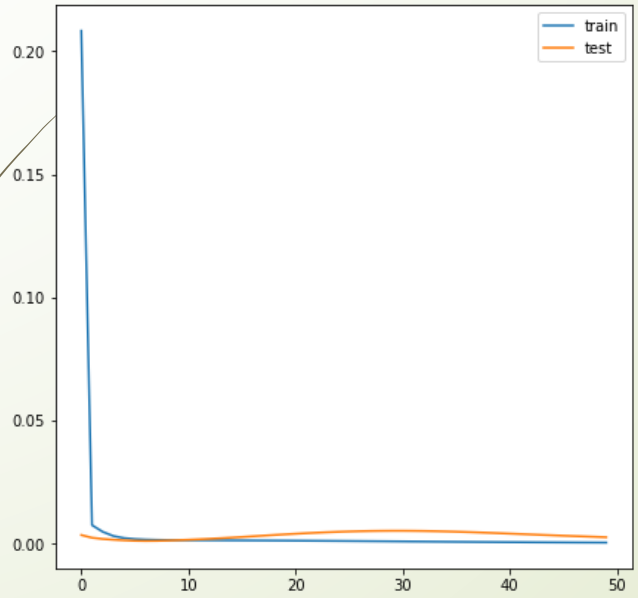
RNN per K5

RNN e Long Term Short Memory

- L'utilizzo di una Recurrent Neural Network e, soprattutto, di un'**architettura LSTM** (o cellula LSTM) si è dimostrata particolarmente efficace per l'apprendimento automatico in presenza di dati organizzati in serie storiche: riconoscimento vocale e scritto, apprendimento del ritmo musicale, ecc…

- Vista la natura di K5 e K11, sembra più che giustificata la creazione di un tal tipo di rete neurale, in cui possiamo comunque considerare le altre variabili a disposizione ed eventuali lag nei dati

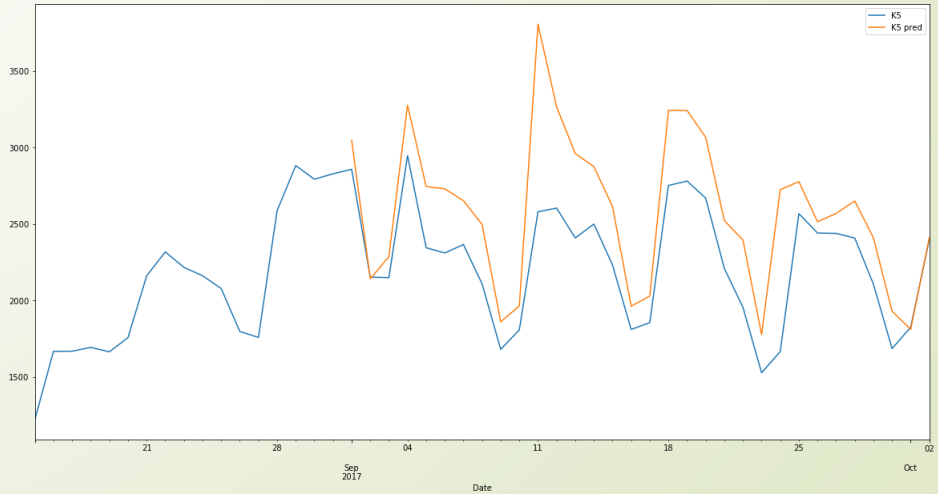
La previsione per K5 mostra spunti interessanti



- Per la costruzione del modello, i dati sono stati nuovamente divisi nei gruppi di train e test, con un mese di dati per la validazione
- Dopo aver testato alcuni modelli, si ottiene un risultato particolarmente soddisfacente per un semplice modello con **14 cellule LSTM** su tutte le variabili a disposizione, comprese delle variabili dummy per identificare il giorno della settimana e il mese dell'anno

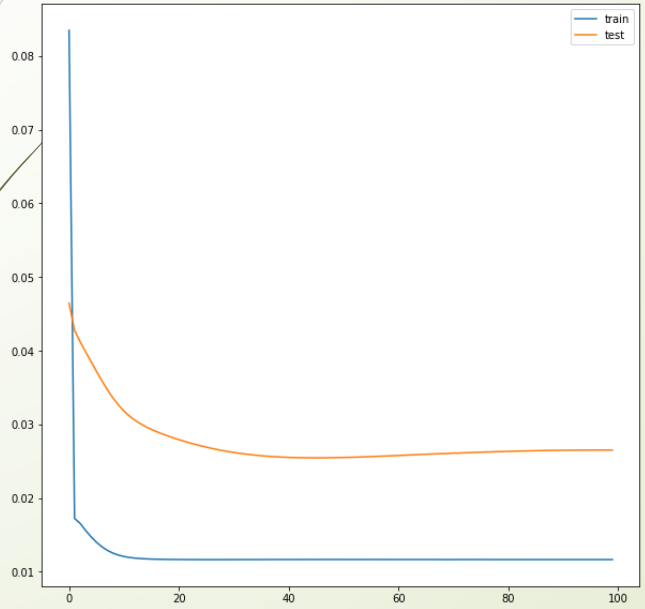
La rete neurale segue l'andamento settimanale, ma sovrastima i valori della serie

- Nel dettaglio qui esposto, si vede come la previsione della rete neurale riesca a seguire l'andamento settimanale con una buona precisione.
- Tuttavia **la serie prevista risulta in generale sovrastimata**. Nella fattispecie, la serie originale inizia una leggera fase di diminuzione del trend, che la rete neurale non riesce a comprendere.



RNN per K11

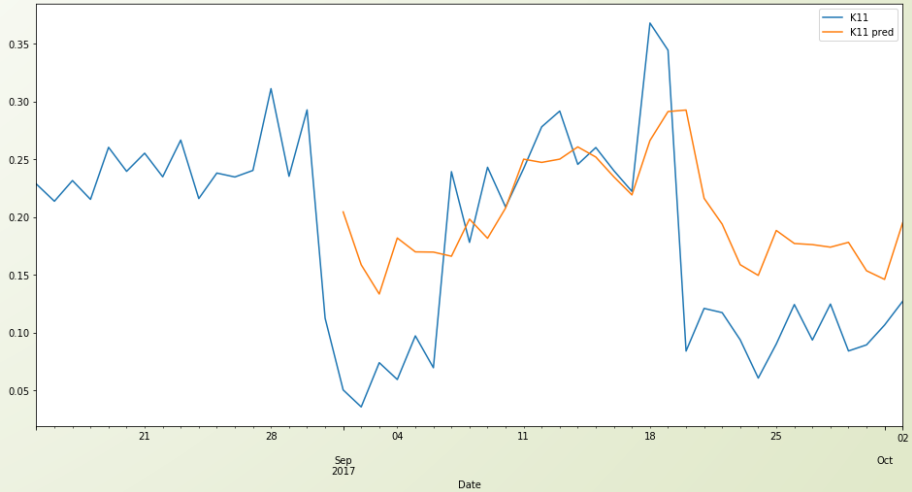
La previsione su K11 funziona su dati destagionalizzati



- Come per K5, i dati sono stati in train e test, con un mese per la validazione
- Tuttavia, a causa dei risultati carenti sulla serie originale, sono state effettuate delle trasformazioni più complesse
- I dati sono stati destagionalizzati e la rete neurale è stata applicata solo alla serie composta dagli elementi di trend e rumore.
- La rete stessa è più complessa: prevede **tre livelli di celle LSTM** (28, 14, 7) e **due livelli di rete neurale classica** (di 5 neuroni e un neurone rispettivamente)

In questo caso, la rete neurale sembra leggermente in ritardo

- Rispetto al modello ARIMA per la stessa serie, la rete neurale riesce ad intuire meglio gli improvvisi cambi di trend, anche se sembra sia leggermente in ritardo.
- Un altro commento può essere fatto sulla variabilità della della previsione, con una variabilità insufficiente, contrariamente alla rete per K5



Risultati delle previsioni

Precisione sulla previsione

	K5	K11
ARIMA	68.7%	167.2%
RNN	16.7%	45.8%

- Per ottenere una misura di precisione sulla previsione, verrà utilizzato la metrica **SMAPE** (errore percentuale assoluto medio simmetrico), che fornisce miglior precisione tanto più il valore è minore
- Per entrambe le serie K5 e K11, il modello basato sulla rete neurale fornisce risultati drasticamente migliori rispetto ai modelli ARIMA

Conclusioni

- I risultati delle previsioni per le variabili K5 e K11 ottenuti attraverso le reti neurali, per quanto non completamente soddisfacenti, dimostrano come l'approccio possa essere applicato con facilità a questo tipo di problemi
- Inoltre, questi modelli sono stati applicati in modo diretto con pochi accorgimenti. Possibili estensioni possono riguardare:
  - Test su strutture di reti neurali più complesse, con un opportuno algoritmo di ricerca degli iperparametri ottimali
  - Più operazioni sulle variabili presenti, combinando le variabili a disposizione in modo diverso
  - Utilizzo di fonti di date esterne, come ad esempio delle informazioni sulle festività che possono influire in determinati periodi dell'anno