Anomaly Detection by statistical analysis

Fabio Livorno

Marzo 2021

1 Introduzione

Conosciuta anche come Outlier Detection, la Anomaly Detection è il processo di data mining usato per determinare la presenza di valori anomali all'interno di un data set, il tipo di anomalia che si presenta e per determinare ulteriori dettagli sulle loro occorrenze. L'idea è quella di studiare, dati un insieme di file CSV contenenti serie temporali, le anomalie presenti grazie a dei tool presenti in linguaggio Python.

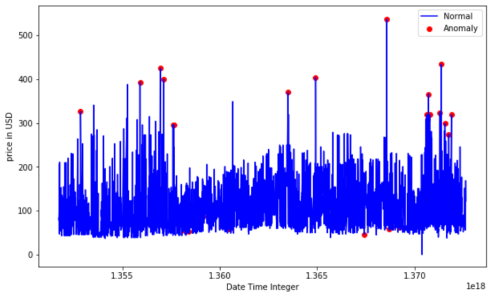

Figure 1: Esempio di raffigurazione di dati con anomalie

2 Primi passi

Dopo un primo momento in cui mi sono dovuto informare su quel che è la Anomaly Detection, ho iniziato a lavorare con i file e le serie temporali in Python, utilizzando diverse librerie per la lettura ed elaborazione dei dati. Tra queste vi è **pandas**, libreria open source flessibile, veloce e facile da utilizzare che permette l'analisi dei dati e la loro rielaborazione.

Il primo passo è stato utilizzare pandas per la lettura dei dati presenti nei file csv e la loro memorizzazione. In seguito per constatarne la corretta estrazione ho utilizzato le funzioni concesse dalla libreria **matplotlib**, che permette la creazione e la visualizzazione di grafici. In particolar modo ho utilizzato una

collezione di funzioni contenute in matplotlib.pyplot, collezione che permette di avere un framework molto simile a quello presente su MATLAB. Di seguito si mostrano un paio di plot d'esempio realizzati utilizzando i dati presenti nel file *ALBIG_elaborato.csv*.

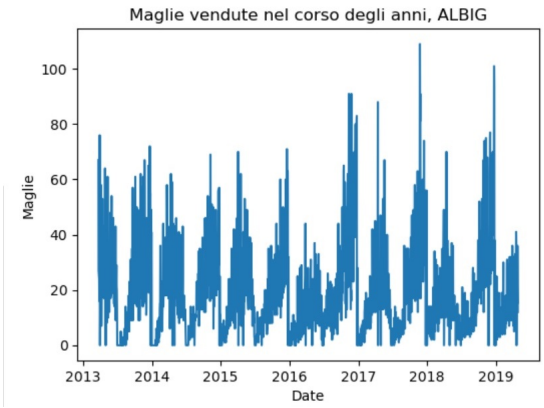

Figure 2: Maglie vendute tra il 2013 ed il 2019 - ALBIG

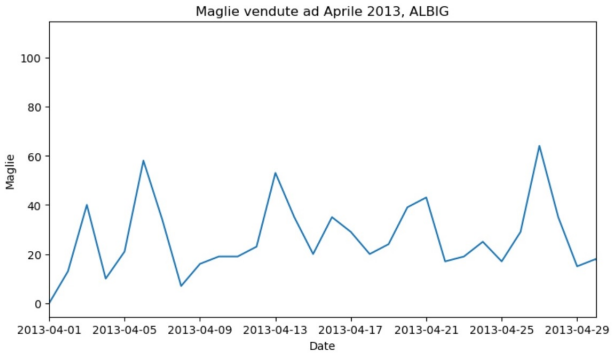

Figure 3: Maglie vendute ad Aprile 2013 - ALBIG

Sebbene ci sia da ridire a livello visivo, ciò che importava maggiormente era che tutto funzionasse e che tutti i valori venissero mostrati correttamente a livello grafico.

3 Tool osservati per il rilevamento

Il prossimo step è stato quello di osservare lo stato dell'arte dell'Anomaly Detection, in particolar modo dei metodi già presenti ed in uso in linguaggio Python.

Mi sono quindi soffermato nella lettura di vari paper ed articoli che elencavano e mostravano l'utilizzo di diversi tool presenti nel mondo dell'Anomaly Detection.

L'idea era quella di utilizzare uno dei tool più recenti open source che ho avuto modo di trovare su GitHub. Considerato che uno dei tool più documentati e mantenuti è **PyOD**, inizialmente ho pensato di utilizzare quello.

Ma ciò di cui questo toolkit non si occupa è rilevamento di anomalie quando si tratta di lavorare con serie temporali, ossia il nostro caso. Motivo per cui è risultato necessario trovare delle alternative valide: con il contributo dello stesso sviluppatore di PyOD ho trovato una repository che trattava **TODS**, un sistema per il rilevamento di anomalie con le serie temporali. Nonostante potesse essere ideale per il lavoro di cui mi sto occupando, trattandosi di un sistema relativamente nuovo non lo reputo abbastanza ben documentato per il suo utilizzo.

Ho deciso quindi di utilizzare **ADTK**, un toolkit contenente gli strumenti che facevano al caso mio.

4 ADTK: Quick start

Di seguito ho dovuto modificare il codice per permettere l'utilizzo di alcuni tool forniti da ADTK. Ho iniziato con la lettura del file CSV *ALBIG_elaborato.csv* e la validazione della serie temporale per il training.

```
import pandas as pd

df = pd.read_csv('ALBIG_elaborato.csv', parse_dates=True,
                 squeeze=True)
df.rename(columns={'Unnamed: 0': 'DATE'}, inplace=True)
datetime_series = pd.to_datetime(df['DATE'])

# create datetime index passing the datetime series
datetime_index = pd.DatetimeIndex(datetime_series.values)

df2 = df.set_index(datetime_index)

# dropping the column that is a duplicate of the
# new index
df2.drop('DATE', axis=1, inplace=True)

# validating the series with the adtk.data method
# validate_series()
from adtk.data import validate_series
s_train = validate_series(df2)

# printing the result
print(s_train)
```

Il passo successivo è stato quello di rilevare le presunte anomalie e stamparne il risultato.

```
from adtk.detector import SeasonalAD
seasonal_ad = SeasonalAD()
anomalies = seasonal_ad.fit_detect(s_train)
```

```
4
5 # plotting the anomalies found
6 plot(s_train, anomaly=anomalies, anomaly_color="red",
7     anomaly_tag="marker")
8 plt.show()
```

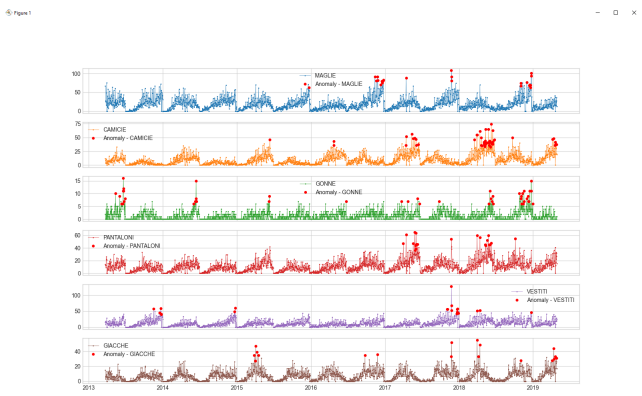

Figure 4: Anomaly Detection - ALBIG

L'analisi basata sui singoli capi rende il rilevamento di anomalie più complicato del necessario. Proviamo ad analizzare la somma dei capi venduti giornalmente e riscontrare eventuali anomalie presenti in quel determinato negozio o su più città contemporaneamente.

5 Modifiche al Dataframe

Per sommare tutti i capi venduti giornalmente mi è bastato utilizzare la funzione **sum** che permette, all'interno di un Dataframe, di sommare i valori presenti nelle varie colonne.

```
1 # somma dei capi d'abbigliamento venduti giorno per
2 # giorno
3 somma_1 = df_1.sum(axis=1, numeric_only=True)
```

A questo punto, avendo la somma dei capi venduti, non avrei avuto più bisogno delle altre colonne presenti nel Dataframe.

```
1 # rimuovo i dati in eccesso dal dataframe
2 df_1.drop(['DATE', 'MAGLIE', 'CAMICIE', 'GONNE',
3 'PANTALONI', 'VESTITI', 'GIACCHE'], axis=1, inplace=True)
4
5 # inserisco la somma dei capi d'abbigliamento venduti
6 # precedentemente calcolata
7 df_1['Albig'] = somma_1
```

Il risultato ottenuto è un segnale unico con la somma dei capi, come mostrato in Figure 5.

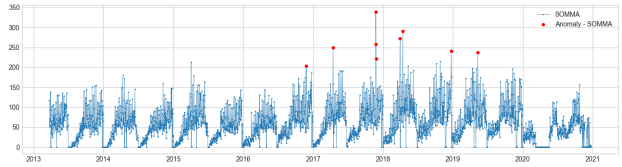

Figure 5: Anomaly Detection - Somma dei capi di ALBIG

Una volta ottenuto questo risultato per una sola città, ho ripetuto il passaggio ed ho unito le somme ottenute in un unico dataframe.

```
# LETTURA SECONDO CSV

# [...]

# somma dei capi d'abbigliamento venduti giorno per giorno
somma_2 = df_2.sum(axis=1, numeric_only=True)

# inserisco la somma dei capi d'abbigliamento venduti
# precedentemente calcolata
df_1['Alghe'] = somma_2

# [... e cos  via anche per il terzo e il quarto ...]
```

Grafici risultanti:

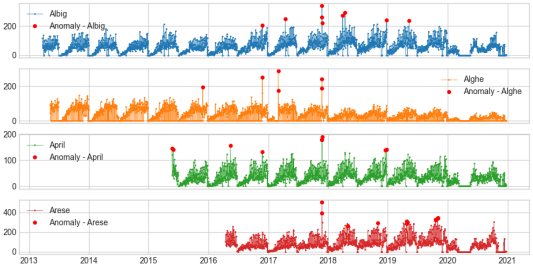

Figure 6: Anomaly Detection - Somma dei capi divisi per città

6 Derivata e resample del segnale

A seguire ho effettuato diversi test per capire quali fossero le anomalie rilevate e perché. Si nota che le anomalie rilevate sono i solo picchi in positivo, motivo per cui il passo successivo è stato quello di fare la derivata del segnale e vedere se vengono evidenziati periodi differenti come anomalie. Per calcolarne la derivata ho utilizzato la funzione *diff()* resa disponibile dalla libreria **pandas**.

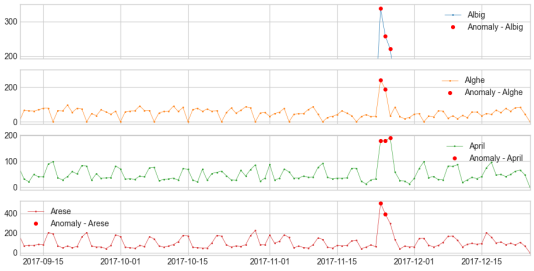

Figure 7: Anomaly Detection - Somma dei capi divisi per città (Zoom)

```
# calcolo la derivata del segnale
diff = s_train.diff()
```

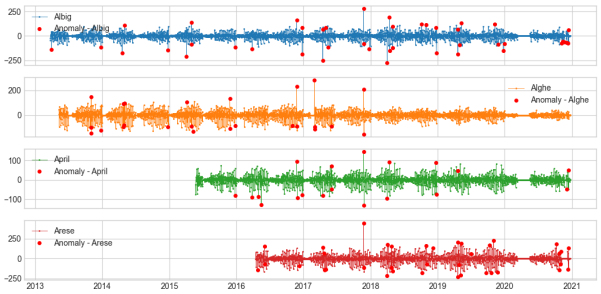

Figure 8: Anomaly Detection - Somma dei capi divisi per città (Derivata)

Ne ho analizzato il segnale sommandone i valori giornalieri e controllandone l'andamento settimanale.

```
# analizzo l'andamento settimanale del segnale
diff = diff.resample('W').sum()
```

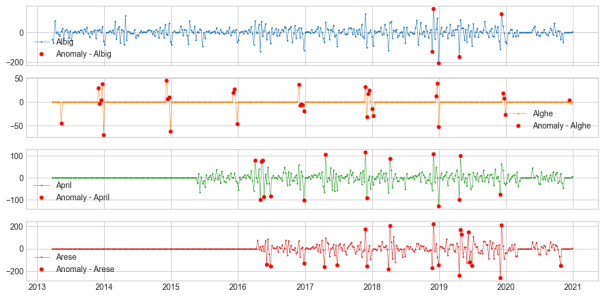

Figure 9: Anomaly Detection - Somma dei capi divisi per città (Derivata -¿ Resample)

7 Ulteriori esempi

Di seguito ulteriori grafici con prove effettuate modificando l'ordine delle operazioni per analizzarne i risultati.

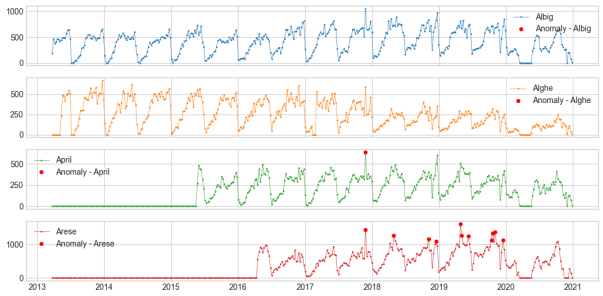

Figure 10: Anomaly Detection - Somma dei capi divisi per città (Resample)

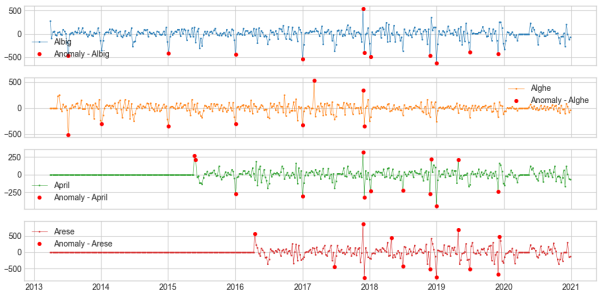

Figure 11: Anomaly Detection - Somma dei capi divisi per città (Resample -¿ Derivata)

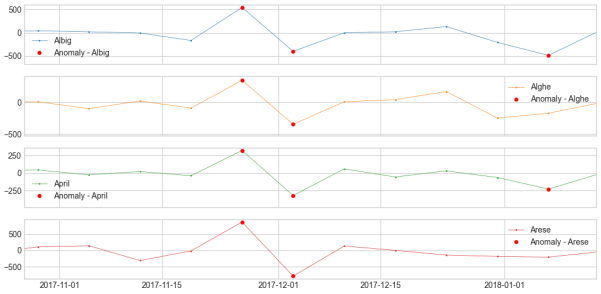

Figure 12: Anomaly Detection - Somma dei capi divisi per città (Resample -¿ Derivata) (Zoom)

Non abbiamo ottenuto degli ottimi risultati ma alcuni potrebbero considerarsi dei miglioramenti in termini di precisione nel rilevamento di anomalie. Cerchiamo di valutare meglio le anomalie riscontrate, comprendendo se sono giustificate o meno.

Anomalie Albig

Dal	Al	Capi venduti	Note
2018/11/19	2018/11/25	255	Black Friday 2018
2018/11/26	2018/12/02	605	Settimana post Black Friday
2018/12/24	2018/12/30	353	Periodo natalizio
2019/04/15	2019/04/21	696	Pasqua 2019
2019/12/01	2019/12/08	466	Settimana post Black Friday

Apparentemente tutte queste anomalie riscontrate sono giustificate da ricorrenze festive o meno che comportano sconti o comunque un aumento di acquisti

all'interno del negozio. Osserviamo altre anomalie rilevate per il negozio Alghe. Per evitare che queste tabelle risultino di più facile lettura e comprensione, eviteremo di mostrare riscontri di anomalie motivate che si presentano annualmente.

Anomalie Alghe

Dal	Al	Capi venduti	Note
2013/05/06	2013/05/12	233	Motivazione sconosciuta
2013/11/25	2013/12/01	485	Black Friday 2013
2013/12/02	2013/12/08	515	Settimana post Black Friday
2013/12/09	2013/12/15	505	Periodo natalizio
2013/12/16	2013/12/22	673	Periodo natalizio
2013/12/23	2013/12/29	437	Periodo natalizio
2014/12/01	2014/12/07	517	Settimana post Black Friday
2018/01/01	2018/01/07	44	Periodo natalizio
2018/12/24	2018/12/30	98	Periodo natalizio

Per quanto riguarda i dati di Alghe, si riscontrano delle anomalie particolari. Nella settimana tra il 6 ed il 12 Maggio 2013, ad esempio, non riusciamo a dare una motivazione all'anomalia riscontrata. Degli altri valori anomali sono i cali di vendite riscontrati come anomalie nelle ultime due righe della tabella mostrata. In quest'ultimo caso, non si spiega il motivo per il quale il negozio abbia avuto un calo così drastico di vendite rispettivamente post Dicembre 2017 e durante le festività di Natale di Dicembre 2018.

Passiamo ad April:

Anomalie April

Dal	Al	Capi venduti	Note
2016/03/28	2016/04/03	315	Pasqua 2016
2016/04/25	2016/05/01	294	Motivazione sconosciuta
2016/05/02	2016/05/08	371	Motivazione sconosciuta
2016/05/09	2016/05/15	495	Motivazione sconosciuta
2016/05/16	2016/05/22	384	Motivazione sconosciuta
2016/06/20	2016/06/26	149	Motivazione sconosciuta
2016/12/19	2016/12/25	440	Periodo natalizio
2017/04/17	2017/04/23	404	Pasqua 2017
2017/11/20	2017/11/26	643	Black Friday 2017
2019/11/25	2019/12/01	76	Black Friday 2019

In questa tabella si presentano una serie di anomalie sconosciute che si presentano particolarmente nel periodo che va da fine Aprile a fine Maggio 2016. Va capito se queste vendite, che si esprimerebbero come picchi positivi, sono dovuti ad una campagna sconti effettuata dal negozio. Ciò spiegherebbe questo boom di vendite. Per quanto concerne, invece, i picchi negativi di Giugno 2016 e fine Novembre 2019, non se ne comprende il motivo di questo calo sconsiderato di vendite.

Passando, infine, ai dati di Arese:

Anomalie Arese

Dal	Al	Capi venduti	Note
2016/05/29	2016/06/05	965	Motivazione sconosciuta
2016/06/20	2016/06/26	293	Motivazione sconosciuta
2016/12/19	2016/12/25	391	Periodo natalizio
2017/04/10	2017/04/16	505	Pasqua 2017
2017/06/19	2017/06/25	232	Motivazione sconosciuta
2017/11/20	2017/11/26	1438	Black Friday 2017
2018/03/25	2018/04/01	733	Pasqua 2018
2018/12/24	2018/12/30	175	Periodo natalizio
2020/10/19	2020/10/25	506	Motivazione sconosciuta

Anche qui si presentano verso Giugno dei picchi di anomalie. Si potrebbe pensare che possano essere dovuti a saldi estivi. Si noti inoltre l'ultima riga: considerata la chiusura dei negozi dovuti al Lockdown, è possibile ipotizzare che il picco di Ottobre 2020 sia dovuto a degli sconti del negozio per poter rientrare delle mancate vendite nel periodo di chiusura del negozio.

Queste sono solo supposizioni, non vi sono abbastanza dati per esserne certi delle motivazioni date o meno per ogni anomalia rilevata. In seguito si potrebbe pensare di trainare il rilevatore di anomalie con degli strumenti forniti sempre da **adtk**.