Bevétel előrejelzés és fogyasztás predikció

Idősor elemzés

2024.11.18.

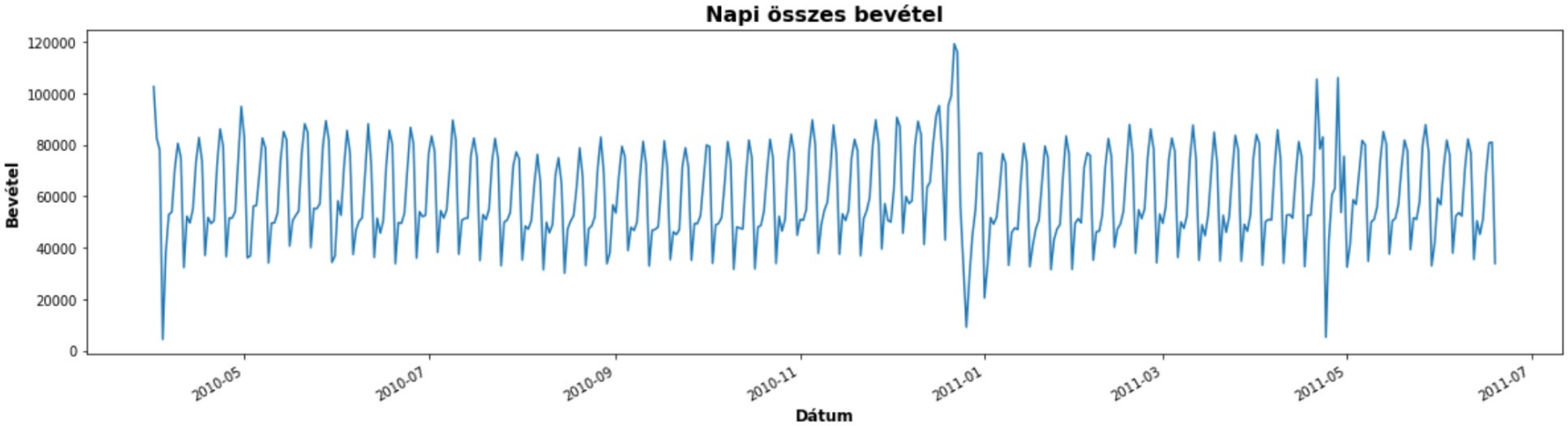
Jónás Dániel
data scientist

Agenda

- Mi az idősor?
- Pandas datetime függvények és attribútumok
- Idősor kiértékelés és a walk forward optimization

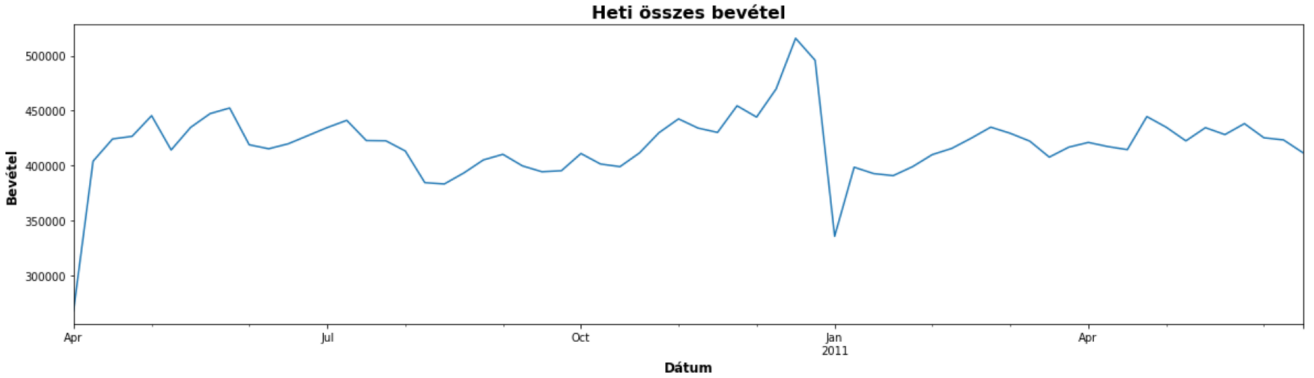
Mi az az idősor?

- Egymást követő állandó intervallumokkal regisztrált adatok
- Ez az időbeliség fontos tulajdonság



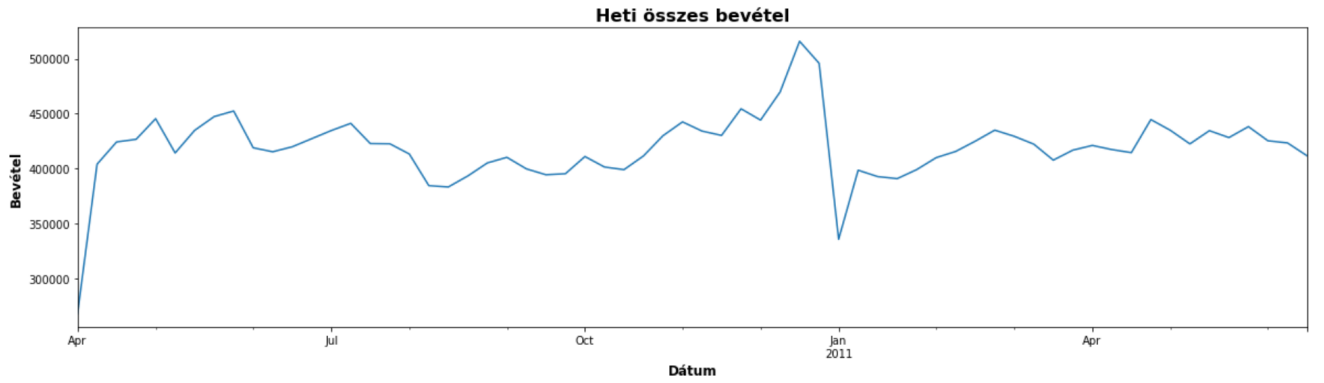
Mi az az idősor?

- Az idő során nőtt vagy csökkent jellemzően a mérték? (trend)
- Vannak e különböző ismétlődő időszakok, amiknek saját jellegzetességük van? (szezonalitás pl.: ünnepek)



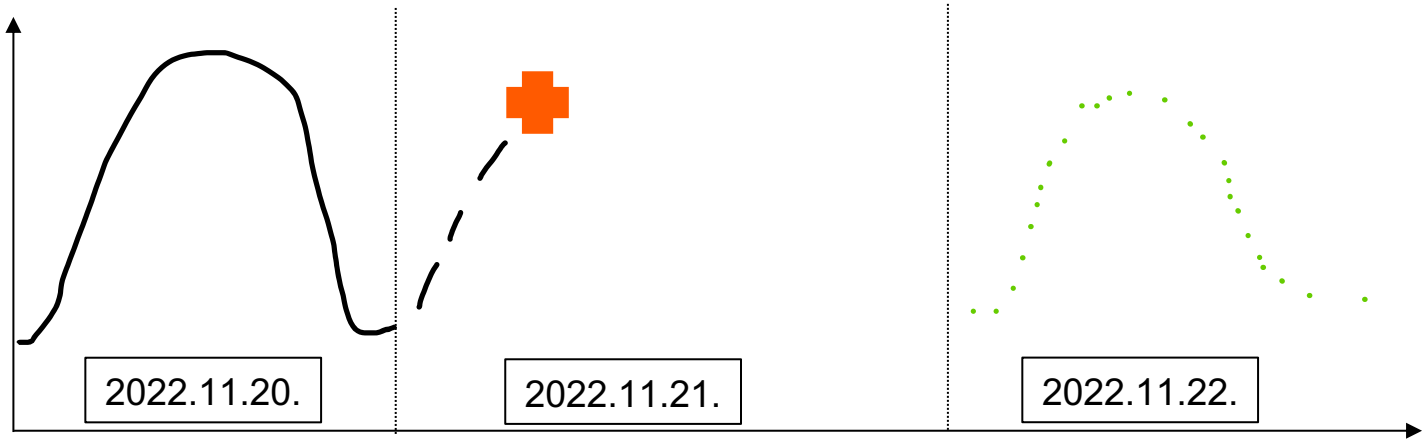
Idősor - kihívások

- Sok évnyi adatra szükség van az egyedi időszakok megértéséhez
- Az egyedi időszakok változhatnak évről évre (emiatt a sok évnyi adat se segítene)



Idősor - kihívások

- Modell folyamatos, akár napi, frissítése
- Mikor kapjuk meg az adatot?
- Milyen adatot használhatunk fel a modellezéshez



Kihívások

- Kevés bemenő változó lehetőség
- Mi lesz az X-ben, ha ránézésre csak egy célváltozónk van?

visit_date	visit_spend
2010-04-01	102625.86
2010-04-02	82660.90
2010-04-03	78081.71
2010-04-04	4411.05
2010-04-05	38144.77
2010-04-06	52928.27
2010-04-07	53908.71
2010-04-08	70863.75
2010-04-09	80590.71
2010-04-10	75106.49

Kihívások

- Kevés bemenő változó lehetőség
- Mi lesz az X-ben, ha ránézésre csak egy célváltozónk van?
- Dátumból nyerünk ki hasznos bemenő változókat

visit_date	visit_spend	dayofweek	month
2010-04-01	102625.86	3	4
2010-04-02	82660.90	4	4
2010-04-03	78081.71	5	4
2010-04-04	4411.05	6	4
2010-04-05	38144.77	0	4
2010-04-06	52928.27	1	4
2010-04-07	53908.71	2	4
2010-04-08	70863.75	3	4
2010-04-09	80590.71	4	4
2010-04-10	75106.49	5	4

Idősor - pandas

Pandas datetime parancsok

- pd.to_datetime(df['oszlop'])
- pd.date_range(start, end, freq)
- .dt

  df['oszlop'].dt.day

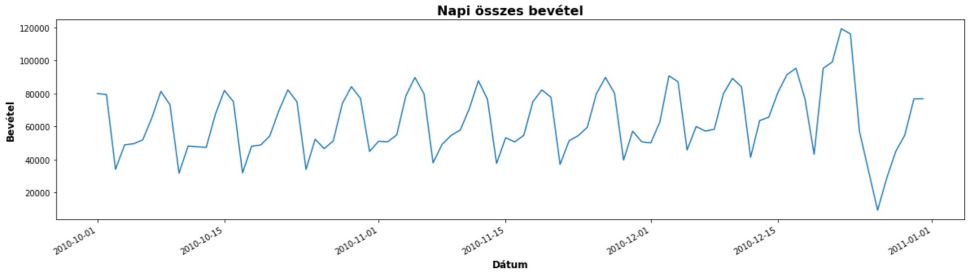
  df['oszlop'].dt.dayofweek

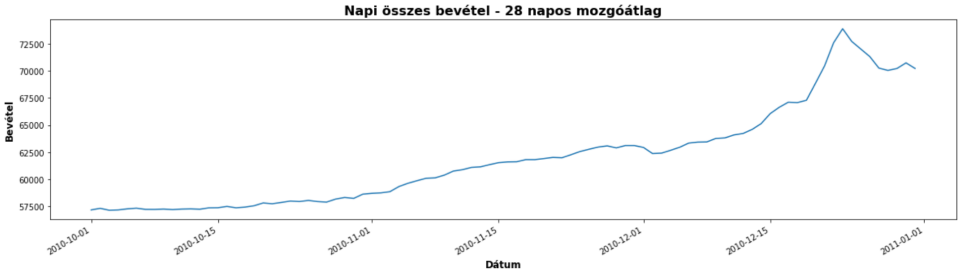
  df['oszlop'].dt.month
- .rolling(window).agg()
- .resample(rule).agg()
- pd.to_timedelta(1, unit='d')

Idősor - pandas

Mozgó aggregáció

- index típus == datetime
- .rolling(window).agg()
- df['visit_spend'].rolling('28d').mean()



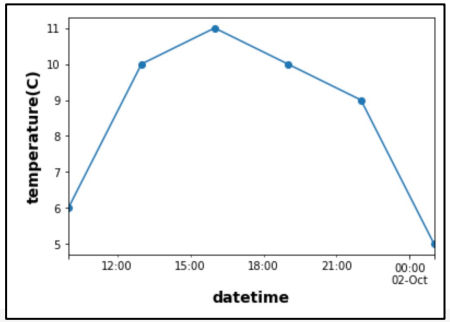


Idősor - pandas

Resampling - downsampling

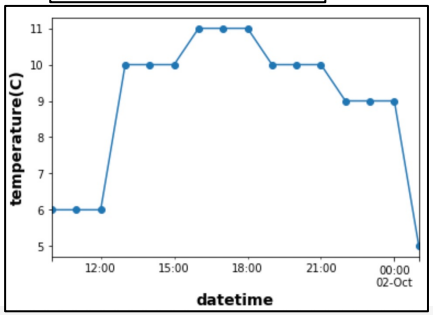
- index típus == datetime
- .resample(rule).agg()

datetime	temp
2022-10-01 10:00:00	6
2022-10-01 13:00:00	10
2022-10-01 16:00:00	11



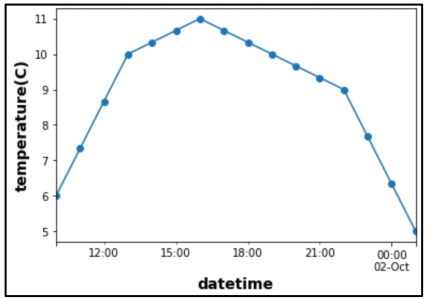
df[‘oszlop’].resample(‘1H’).ffill()

datetime	temp
2022-10-01 10:00:00	6
2022-10-01 11:00:00	6
2022-10-01 12:00:00	6
2022-10-01 13:00:00	10
2022-10-01 14:00:00	10



df[‘oszlop’].resample(‘1H’).interpolate()

datetime	temp
2022-10-01 10:00:00	6.000000
2022-10-01 11:00:00	7.333333
2022-10-01 12:00:00	8.666667
2022-10-01 13:00:00	10.000000
2022-10-01 14:00:00	10.333333

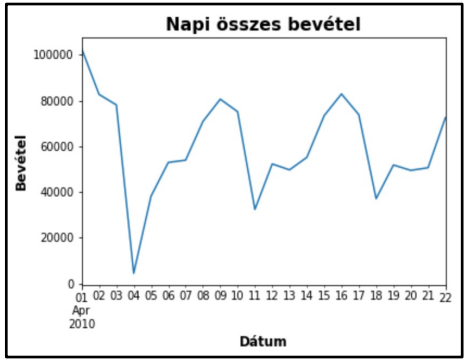


Idősor - pandas

Resampling - upsampling

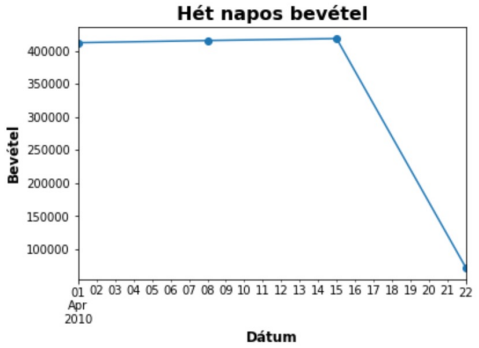
- index típus == datetime
- .resample(rule).agg()

```
visit_date
2010-04-01    102625.86
2010-04-02     82660.90
2010-04-03     78081.71
```



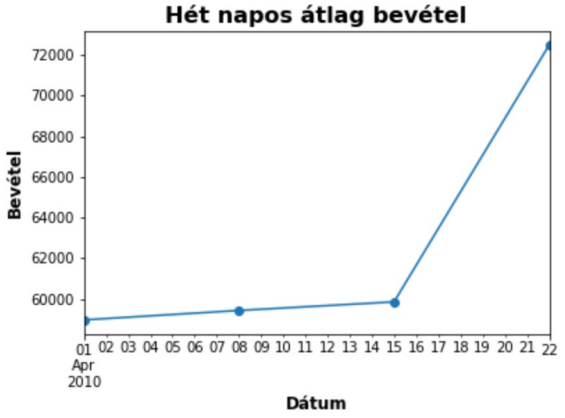
df['oszlop'].resample('7d').sum()

```
visit_date
2010-04-01    412761.27
2010-04-08    416009.54
2010-04-15    418964.38
2010-04-22     72482.41
```



df['oszlop'].resample('7d').mean()

```
visit_date
2010-04-01    58965.895714
2010-04-08    59429.934286
2010-04-15    59852.054286
2010-04-22    72482.410000
```



Kihívások

- Kevés bemenő változó lehetőség
- Mi lesz az X-ben, ha ránézésre csak egy célváltozónk van?
- Dátumból nyerünk ki hasznos bemenő változókat

visit_date	visit_spend	dayofweek	month
2010-04-01	102625.86	3	4
2010-04-02	82660.90	4	4
2010-04-03	78081.71	5	4
2010-04-04	4411.05	6	4
2010-04-05	38144.77	0	4
2010-04-06	52928.27	1	4
2010-04-07	53908.71	2	4
2010-04-08	70863.75	3	4
2010-04-09	80590.71	4	4
2010-04-10	75106.49	5	4

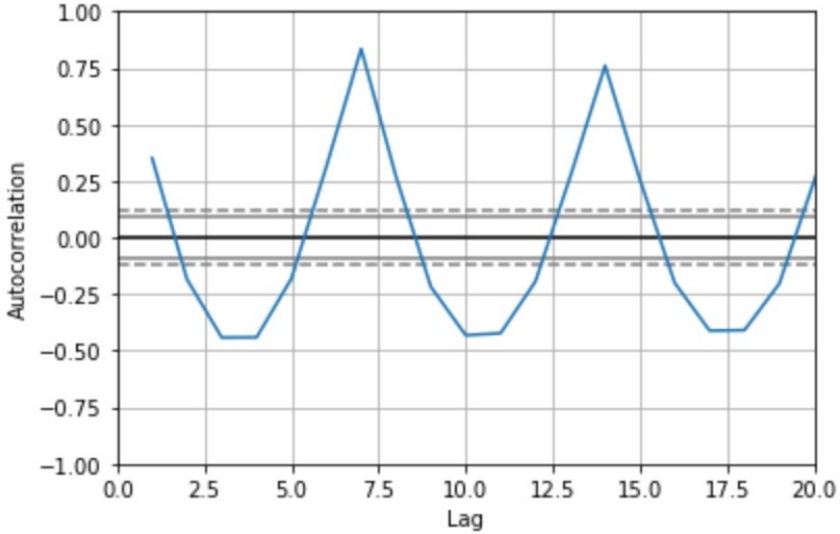
Kihívások

- Kevés bemenő változó lehetőség
- Mi lesz az X-ben, ha ránézésre csak egy célváltozónk van?
- Dátumból nyerünk ki hasznos bemenő változókat
- Célváltozóból nyerünk ki hasznos bemenő változókat

visit_date	visit_spend	dayofweek	month	visit_spend-1d	visit_spend-7d
2010-04-01	102625.86	3	4	NaN	NaN
2010-04-02	82660.90	4	4	102625.86	NaN
2010-04-03	78081.71	5	4	82660.90	NaN
2010-04-04	4411.05	6	4	78081.71	NaN
2010-04-05	38144.77	0	4	4411.05	NaN
2010-04-06	52928.27	1	4	38144.77	NaN
2010-04-07	53908.71	2	4	52928.27	NaN
2010-04-08	70863.75	3	4	53908.71	102625.86
2010-04-09	80590.71	4	4	70863.75	82660.90
2010-04-10	75106.49	5	4	80590.71	78081.71

Idősor - autokorreláció

- A célváltozó önmaga korábbi értékeivel korrelál
- pd.plotting.autocorrelation_plot(df['oszlop'])
- df['oszlop'].autocorr()



Kiértékelés

- Kfold keresztvalidáció nem megfelelő
- Prediktált értékhez képest jövőbeli értékeket is beletehet a tanítóhalmazba
- Megoldás: walk forward optimization