Feature Engineering Overview

Creating predictor variables

Dataset and Features

Month	Day	Temp	Rain	inflation
3	30	15	50	0.2
3	31	16	10	0.2
4	1	17	0	0.19
4	2	?	?	?

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

- We want to predict a value at time ***t***.
- We can use data at time ***t*** if available and previous data (t-1, t-2, etc.).
- As past events can inform future behaviour, we can create new features from past information taking the temporal aspect of the data into account.

Lagged Features

Month	Day	Temp-1	Rain-1	Infl. -1
3	30	nan	nan	nan
3	31	15	50	0.2
4	1	16	10	0.2
4	2	17	0	0.19

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Lag features

- We can infer the value at ***t*** utilizing the previous value of the feature.
- A lagged feature is any feature that is from a fixed period in the past relative to the target.
- The lag can vary, we can create multiple features with multiple lags.

Sliding Window Features

Month	Day	Temp	Rain	inflation
3	30	15	50	0.2
3	31	16	10	0.2
4	1	17	0	0.19
4	2	?	?	?

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Window features

- We can infer the value at ***t*** utilizing previous values within a certain period ➔ **windows**.
- We can use statistical parameters within those windows, i.e., min, max, mean, std., etc.
- The window can vary, we can create multiple features from multiple window sizes.

Sliding Window Features

Month	Day	Temp	Rain	inflation
3	30	nan	nan	nan
3	31	15	50	0.2
4	1	15.5	30	0.2
4	2	16.5	5	0.195

Window features

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

- We can infer the value at ***t*** utilizing previous values within a certain period ➔ **windows**.
- We can use statistical parameters within those windows, i.e., min, max, mean, std., etc.
- The window can vary, we can create multiple features from multiple window sizes.

Temporal Features

Month	Day
3	30
3	31
4	1
4	2

weekend	Elapsed
0	1
0	1
1	1
1	1

Temp-1	Rain-1	Infl. -1
nan	nan	nan
15	50	0.2
16	10	0.2
17	0	0.19

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Temporal features

- We can create new features from the **timestamp** ➔ day, month, hr, time, is weekend, business hrs, public holidays, etc.
- We can capture elapsed time, for example "time since last transaction".

Target Features – lags and windows

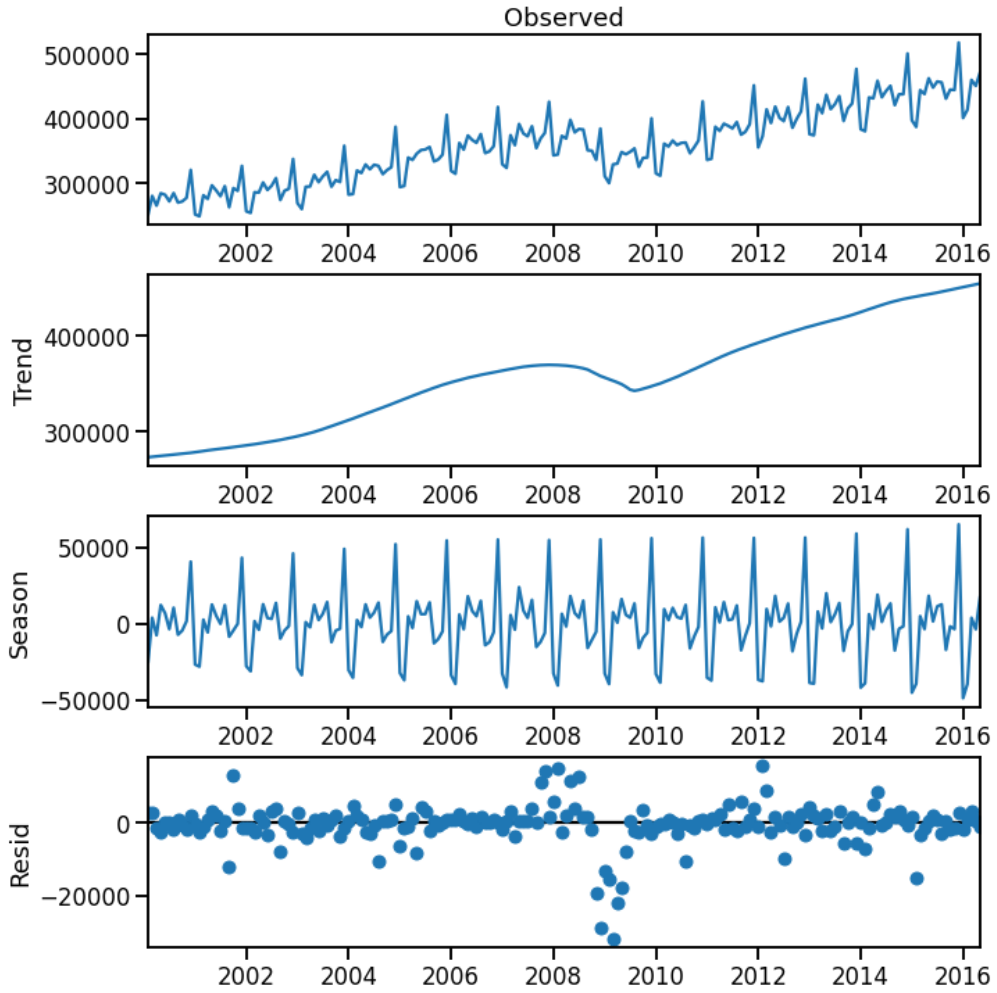
Month	Day	weekend	Temp-1	Rain-1	Infl. -1	Sales-1	Mean
3	30	0	nan	nan	nan	nan	nan
3	31	0	15	50	0.2	200	200
4	1	1	16	10	0.2	220	210
4	2	1	17	0	0.19	230	225

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

- We can create new features from the target.
  - Lag and window features, multiple statistical parameters **at previous time points**

Target Features - Seasonality and trend

- Seasonality
- Trend



Feature combination

- We can combine features into new features
  - Total sales / total population
  - Energy demand / area
  - Temperature and rain ➔ humidity
- Transform features mathematically ➔ log
- We can combine and transform time series themselves or the already extracted features

Feature engineering – aim

Raw data

Timestamp
30/03/2021
30/04/2021
30/05/2021
30/06/2021

Temp	Rain	inflation
15	50	0.2
16	10	0.2
17	0	0.19
?	?	?

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Timestamp	Month	Day	weekend	Temp-1	Rain-1	Infl. -1	Sales-1	Mean
30/03/2021	3	30	0	nan	nan	nan	nan	nan
30/04/2021	3	31	0	15	50	0.2	200	200
30/05/2021	4	1	1	16	10	0.2	220	210
30/06/2021	4	2	1	17	0	0.19	230	225

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Pre-processed or engineered dataset

Feature engineering – aim

Raw data

Timestamp
30/03/2021
30/04/2021
30/05/2021
30/06/2021

Temp	Rain	inflation
15	50	0.2
16	10	0.2
17	0	0.19
?	?	?

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

- Some forecasting models are able to take the raw data as input.
- For example the models supported by Facebook's Prophet.
- ARIMA and ETS.
- Neuronal networks

Feature engineering – aim

- For forecasting or classification with off-the-shelf algorithms like linear models, decision tree based algorithms, SVMs, etc, we need to pre-process the datasets.
- For some Deep Learning Models as well.

Timestamp	Month	Day	weekend	Temp-1	Rain-1	Infl. -1	Sales-1	Mean	Sales	
30/03/2021	3	30	0	nan	nan	nan	nan	nan	200	
30/04/2021	3	31	0	15	50	0.2	200	200	220	*t-2*
30/05/2021	4	1	1	16	10	0.2	220	210	230	*t-1*
30/06/2021	4	2	1	17	0	0.19	230	225	?	*t*

Pre-processed or engineered dataset

Feature engineering – aim

Timestamp
30/03/2021
30/04/2021
30/05/2021
30/06/2021

Temp	Rain	inflation
15	50	0.2
16	10	0.2
17	0	0.19
?	?	?

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Timestamp	Month	Day	weekend	Temp-1	Rain-1	Infl. -1	Sales-1	Mean
30/03/2021	3	30	0	nan	nan	nan	nan	nan
30/04/2021	3	31	0	15	50	0.2	200	200
30/05/2021	4	1	1	16	10	0.2	220	210
30/06/2021	4	2	1	17	0	0.19	230	225

Sales	
200	
220	*t-2*
230	*t-1*
?	*t*

Throughout the course, we will discuss how to obtain a fully pre-processed dataset, ready to use with regression models.

Summary

We can create predictive features from past behavior.

There are multiple approaches to create features, i.e., lags, windows, trends and seasonality, etc.

We can extract / derive a lot of features from a time series.