Feature engineering in time series forecasting

Summary

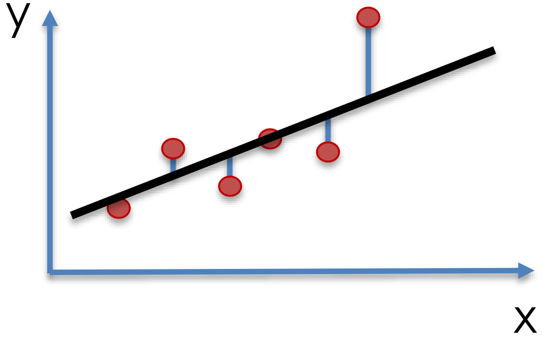
Time Series - definition

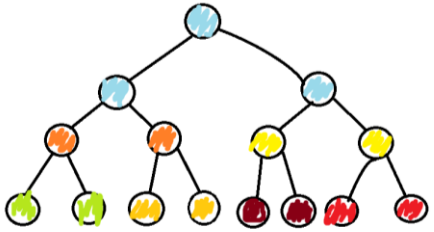
- ✓ Time series are data points **indexed in time order**.
- ✓ We can have 1 or more time series.

Time series forecasting models

Machine learning models

LightGBM





Statistical models

ARIMA

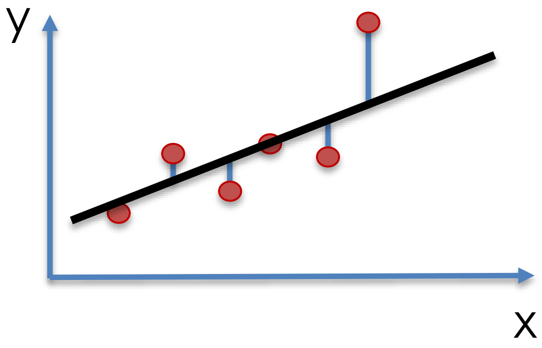
ETS

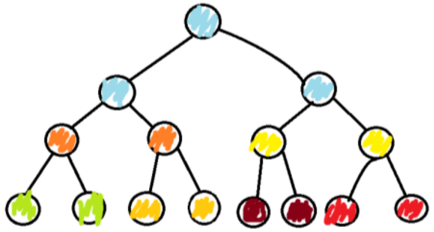
PROPHET

Forecasting with traditional ML models

Machine learning models

LightGBM

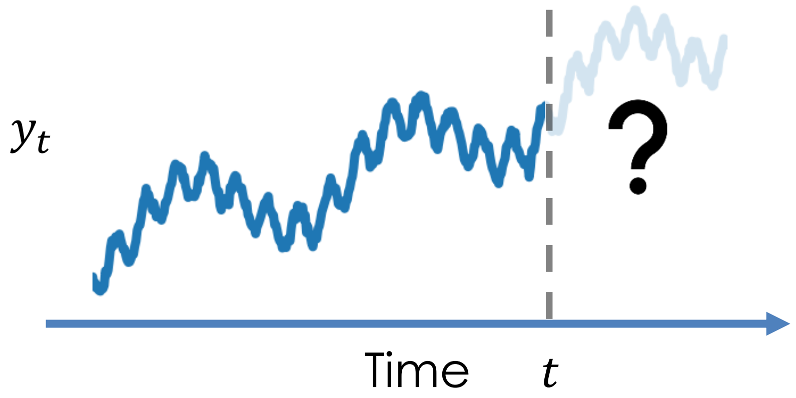




Traditional machine learning models can also leverage the power of including exogenous explanatory variables.

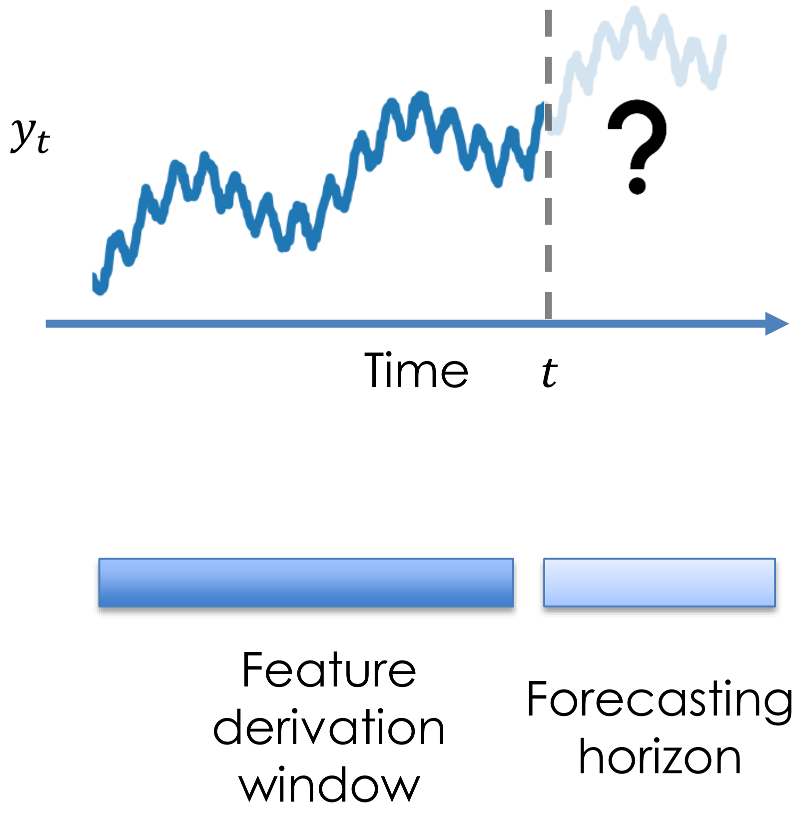
Forecasting with machine learning

Time	Sales	
		⋮
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t



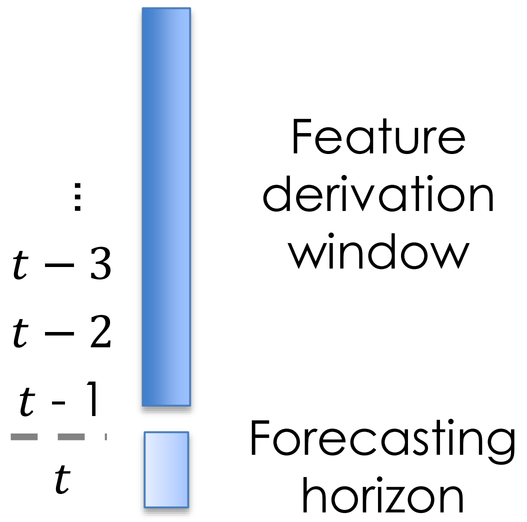
Forecasting with machine learning

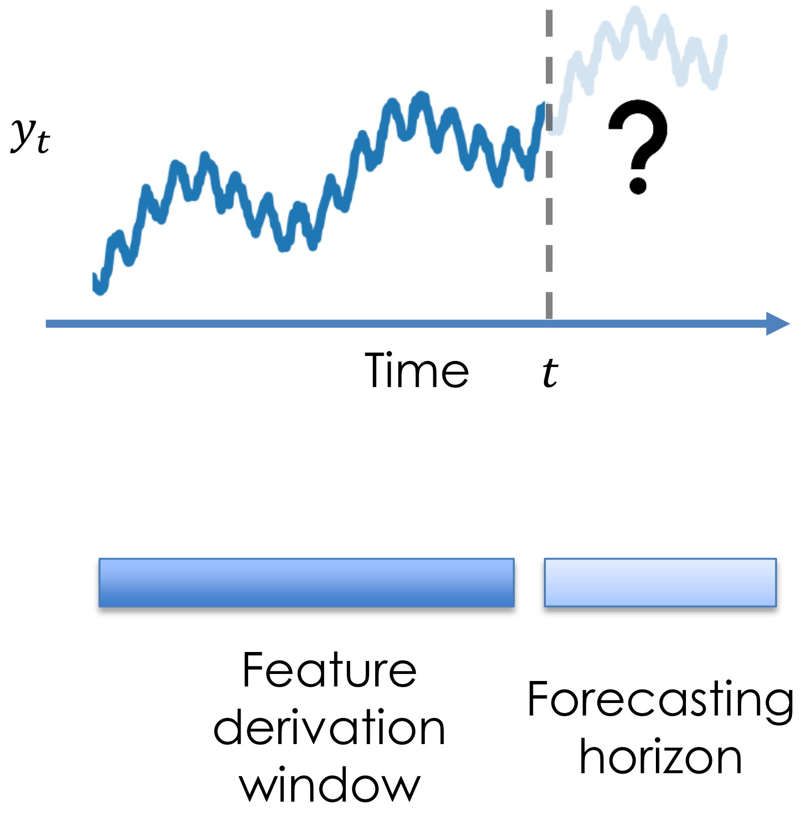
Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	⋮
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t



Forecasting with machine learning

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?





Time series to a table of features and a target

Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	$\vdots$
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

?

x1	x2	x3	y

Time series to a table of features and a target

Time	Sales	
		⋮
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

?

x1	x2	x3	y_t
			35
			30
			23
			21
			40
			31
			?

Time series to a table of features and a target

Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	$\vdots$
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

Can only use data up to $t-1$ to predict t.

This is to avoid look-ahead bias.

x1	x2	x3	y_t
			35
			30
			23
			21
			40
			31
			?

Time series to a table of features and a target

Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	⋮
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

x1	x2	x3	y_t
			35
			30
			23
			21
			40
			31
			?

Time series to a table of features and a target

Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	⋮
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

x1	x2	x3	y_t
			35
			30
			23
			21
			40
			31
21	40	31	?

Time series to a table of features and a target

Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	⋮
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

x1	x2	x3	y_t
			35
			30
			23
			21
			40
23	21	40	31
21	40	31	?

Time series to a table of features and a target

Time	Sales	
2020-02-12	35	
2020-02-13	30	
2020-02-14	23	$\vdots$
2020-02-15	21	$t-3$
2020-02-16	40	$t-2$
2020-02-17	31	t - 1
2020-02-18	?	t

x1	x2	x3	y_t
			35
			30
			23
			21
30	23	21	40
23	21	40	31
21	40	31	?

Time series to a table of features and a target

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?

x1	x2	x3	y_t
NaN	NaN	NaN	35
NaN	NaN	35	30
NaN	35	30	23
35	30	23	21
30	23	21	40
23	21	40	31
21	40	31	?

Time series to a table of features and a target

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?

y_{t-3}	y_{t-2}	y_{t-1}	y_t
NaN	NaN	NaN	35
NaN	NaN	35	30
NaN	35	30	23
35	30	23	21
30	23	21	40
23	21	40	31
21	40	31	?

Features derived from **past values** (e.g., lag features).

Time series to a table of features and a target

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?

Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
100	NaN	NaN	NaN	35
120	NaN	NaN	35	30
116	NaN	35	30	23
120	35	30	23	21
101	30	23	21	40
90	23	21	40	31
190	21	40	31	?

Features with **known values in the future.**

Features derived from past values (e.g., lag features).

Time series to a table of features and a target

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?

Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
100	100	NaN	NaN	NaN	35
120	120	NaN	NaN	35	30
116	116	NaN	35	30	23
120	120	35	30	23	21
101	101	30	23	21	40
90	90	23	21	40	31
?	190	21	40	31	?

Features with **unknown values in the future**.

Features with known values in the future.

Features derived from past values (e.g., lag features)

Time series to a table of features and a target

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?

Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
100	100	NaN	NaN	NaN	35
120	120	NaN	NaN	35	30
116	116	NaN	35	30	23
120	120	35	30	23	21
101	101	30	23	21	40
90	90	23	21	40	31
$\hat{x}_{t+1}$	190	21	40	31	?

Features with **unknown values in the future.**

Features with known values in the future.

Features derived from past values (e.g., lag features)

Time series to a table of features and a target

Time	Sales	Country	Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
2020-02-12	35	UK	100	100	NaN	NaN	NaN	35
2020-02-13	30	UK	120	120	NaN	NaN	35	30
2020-02-14	23	UK	116	116	NaN	35	30	23
2020-02-15	21	UK	120	120	35	30	23	21
2020-02-16	40	UK	101	101	30	23	21	40
2020-02-17	31	UK	90	90	23	21	40	31
2020-02-18	?	UK	$\hat{x}_{t+1}$	190	21	40	31	?

Static features.

Features with unknown values in the future.

Features with known values in the future.

Features derived from past values (e.g., lag features)

Time series to a table of features and a target

Time	Sales	Country	Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
2020-02-12	35	UK	100	100	NaN	NaN	NaN	35
2020-02-13	30	UK	120	120	NaN	NaN	35	30
2020-02-14	23	UK	116	116	NaN	35	30	23
2020-02-15	21	UK	120	120	35	30	23	21
2020-02-16	40	UK	101	101	30	23	21	40
2020-02-17	31	UK	90	90	23	21	40	31
2020-02-18	?	UK	$\hat{x}_{t+1}$	190	21	40	31	?

Static features.

Features with unknown values in the future.

Features with known values in the future.

Features derived from past values (e.g., lag features)

Exogenous variables.

Time series to a table of features and a target

Time	Sales
2020-02-12	35
2020-02-13	30
2020-02-14	23
2020-02-15	21
2020-02-16	40
2020-02-17	31
2020-02-18	?

Country	Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
UK	100	100	NaN	NaN	NaN	35
UK	120	120	NaN	NaN	35	30
UK	116	116	NaN	35	30	23
UK	120	120	35	30	23	21
UK	101	101	30	23	21	40
UK	90	90	23	21	40	31
UK	$\hat{x}_{t+1}$	190	21	40	31	?

X
(features)

y
(target)

Time series to a table of features and a target

Country	Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
UK	100	100	NaN	NaN	NaN	35
UK	120	120	NaN	NaN	35	30
UK	116	116	NaN	35	30	23
UK	120	120	35	30	23	21
UK	101	101	30	23	21	40
UK	90	90	23	21	40	31
UK	$\hat{x}_{t+1}$	190	21	40	31	?

Training data

X
(features)

y
(target)

Time series to a table of features and a target

model.fit

Country	Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
UK	100	100	NaN	NaN	NaN	35
UK	120	120	NaN	NaN	35	30
UK	116	116	NaN	35	30	23
UK	120	120	35	30	23	21
UK	101	101	30	23	21	40
UK	90	90	23	21	40	31
UK	$\hat{x}_{t+1}$	190	21	40	31	?

X
(features)

y
(target)

Time series to a table of features and a target

Country	Rainfall	Ad spend	y_{t-3}	y_{t-2}	y_{t-1}	y_t
UK	100	100	NaN	NaN	NaN	35
UK	120	120	NaN	NaN	35	30
UK	116	116	NaN	35	30	23
UK	120	120	35	30	23	21
UK	101	101	30	23	21	40
UK	90	90	23	21	40	31
UK	$\hat{x}_{t+1}$	190	21	40	31	?

model.fit

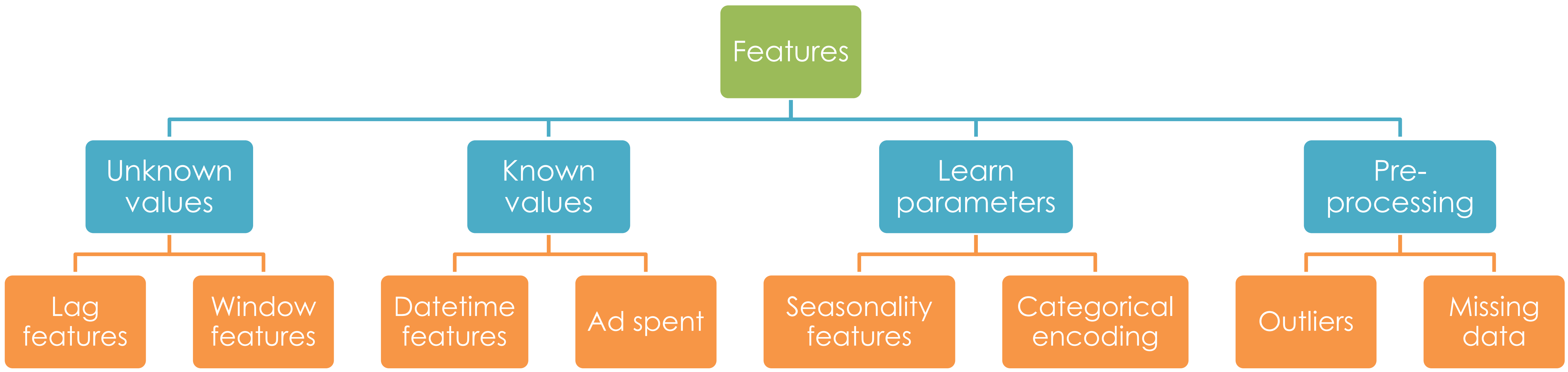
model.predict

$\hat{y}_{t+1}$

X
(features)

y
(target)

Feature engineering



Features: Domain knowledge vs automation

Traditionally, feature creation in forecasting has been mostly done using **domain knowledge.**

Only recently, there is a trend to try and automate this process.