Data changes over time

You are not testing right

What is time series data?

When the target variable or one of its features has some behaviour in which time has some influence

$$x = g(t), y = f(x) \rightarrow y = h(t), h = f \circ g$$

$$\mathbb{E}[X_{t+k}|X_t] \neq \mathbb{E}[X_{t+k}]$$

Type	fabric	color	size	price	date	**Units sold**

You work for a company that sells clothing and they want to hire you to plan production for next *period*. You have historic records covering 10 years of sales.

Demand forecasting problem

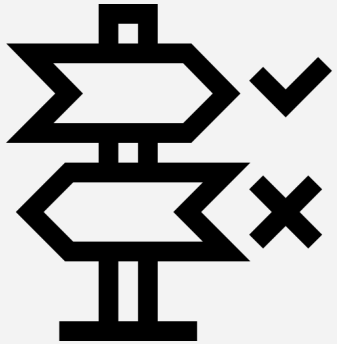
Type fabric color size price date

Is it all? Do you know more ?

- How many units did we sell last month?
- How many did we sell last year?
- How many did we sell from that color last month?
- What about moving average or some other aggregate statistic?
- ...

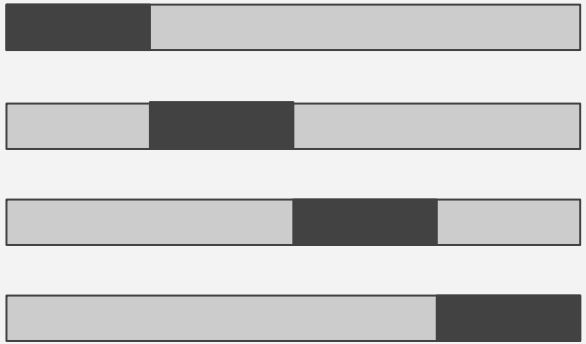
Demand forecasting problem: features

Shuffle and split as usual (`train_test_split`)

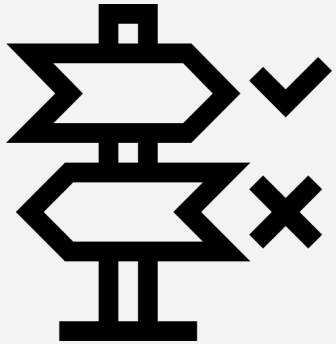


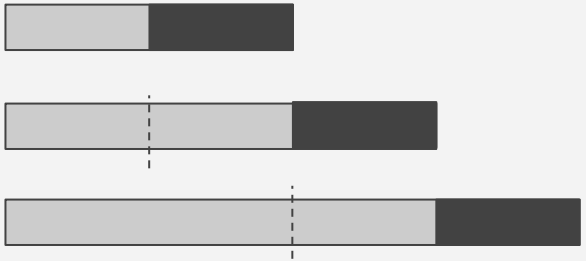
Latest available period, for test, al the rest for train: you are going to use your model to predict the future, test it in the same way.

Demand forecasting problem: train and test sets



As usual (`KFold`)





Validation always ahead of training (`TimeSeriesSplit`)

Demand forecasting problem: making cross validation