How to choose the lags

Lag features

How to choose the lags

Date	Sales (target)	Ad spend (feature)
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

Lag feature

	target	lag features **from target**		original features	lag features **from original features**	
Date	**Sales**	**Sales Lag 1**	**Sales Lag 3**	**Ad spend**	**Ad spend Lag 1**	**Ad spend Lag 2**
2020-02-12	23	NaN	NaN	100	NaN	NaN
2020-02-13	30	23	NaN	120	100	NaN
2020-02-14	35	30	NaN	90	120	100
2020-02-15	30	35	23	80	90	120
2020-02-16	?	30	30	100	80	90

We can create multiple lag features with different lags from the target and features.

Problem: Which lags to use? How many lag features to create?

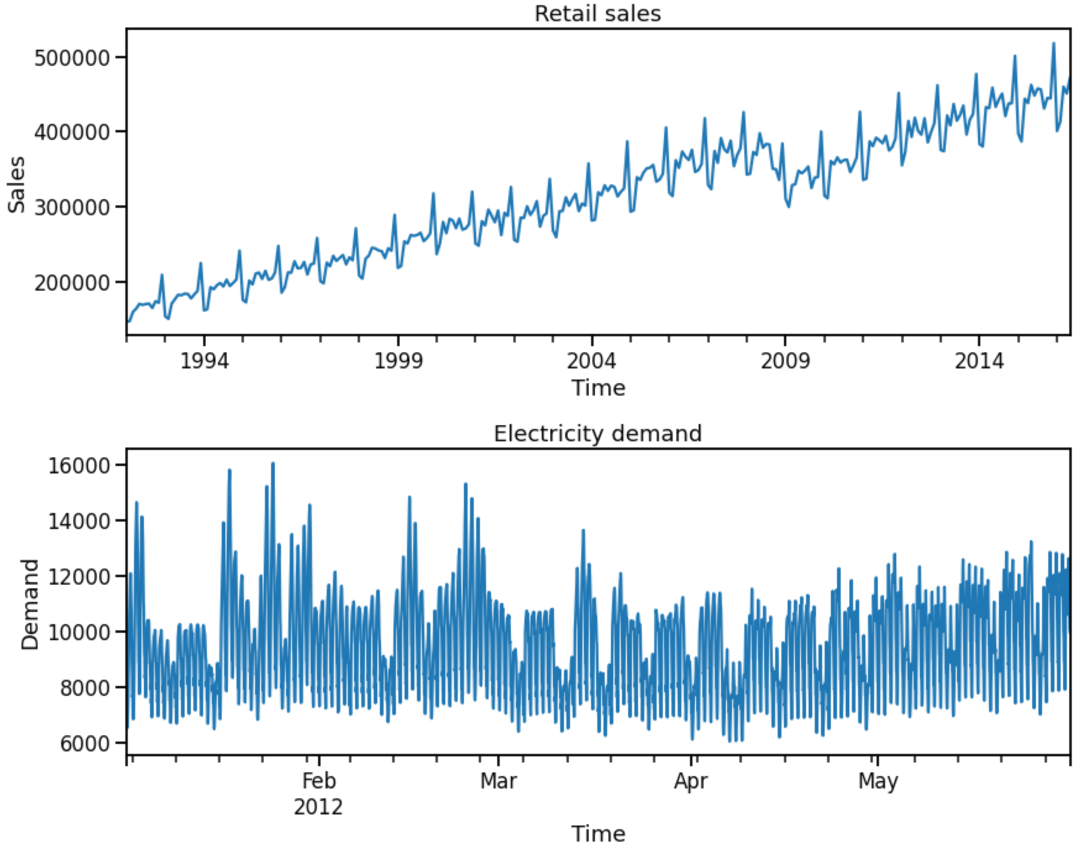
How to choose the lags

- Domain knowledge
- Feature selection and modelling
- Time-series correlation methods

Domain knowledge

When lagging the target

- If seasonality is known use a lag of the same seasonal order (aka seasonal lag).
- Example retail sales: yearly seasonality ➔ use lag of 1 year.
- Example electricity demand: multiple seasonalities such as yearly, weekly, and daily ➔ use lag of 1 year, 1 week, and 1 day.
- Most recent values tend to be predictive ➔ use small lags.



Domain knowledge

When lagging the features

- Use the subset of features you think are most important in affecting the target.
- Only use the value of the feature that you know at predict time to avoid data leakage.
- Most recent values tend to be predictive
  ➔ use small lags.

	target	features	
Date	**Sales**	**Ad spend**	**weekend**
2020-02-12	23	100	0
2020-02-13	30	120	0
2020-02-14	35	90	0
2020-02-15	30	80	1
2020-02-16	?	100	1

Domain knowledge

Pros

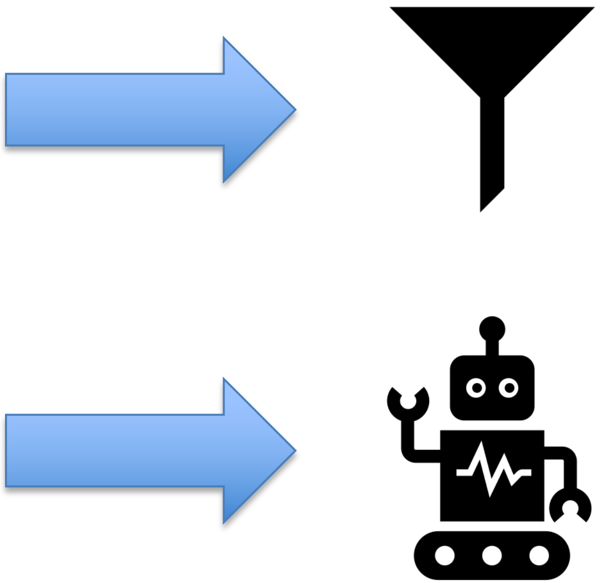
- Likely to result in fewer additional features as we will pick a small number of lags and features known to be important.

Cons

- We may not know all the seasonal patterns or which ones are most important.
- We may not know which features are important to lag.
- Not scalable.

Feature selection and modelling

Sales Lag 1	Sales Lag 3	Ad spend Lag 1	Ad spend Lag 2
NaN	NaN	NaN	NaN
23	NaN	100	NaN
30	NaN	120	100
35	23	90	120
30	30	80	90



- Create a bunch of different lags which are reasonable given the feature and use case (e.g., ad spend more than 1 year ago unlikely to be help for sales forecasting).

- Use feature selection and/or modelling (e.g., LASSO) to best utilize the features and determine a subset which minimizes forecast error.

Feature selection and modelling

Pros

- Automatic. Less hands-on decision making.
- May find useful features which you may not have been used otherwise.

Cons

- Will create very large number of features.
- The lags of the same features will be highly correlated to each other.
- More complex model than necessary.
- Computationally expensive.

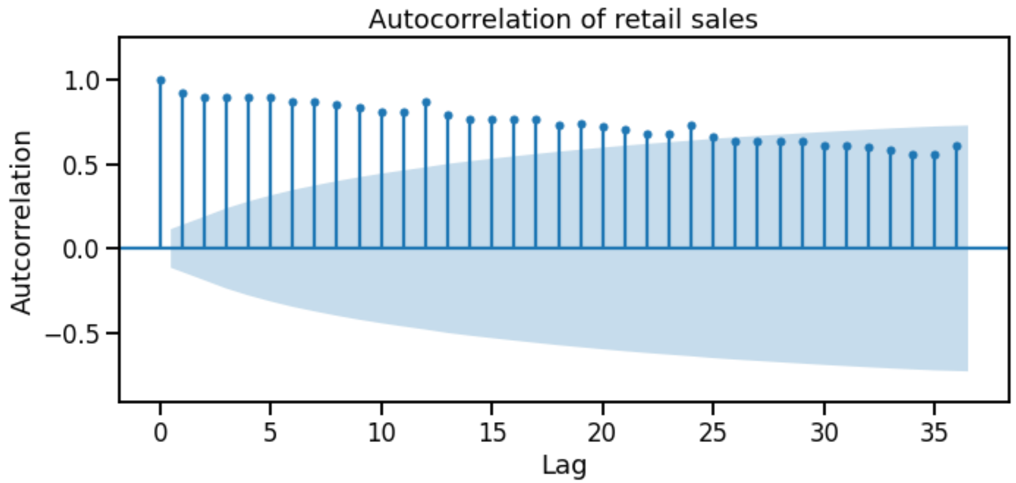
Time-series correlation methods

The main idea

- Measure how correlated the lag features are with the target.
- If the lag feature is highly correlated to the target then it might be helpful.

Three main methods

- Autocorrelation function (ACF)
- Partial autocorrelation function (PACF)
- Cross-correlation function (CCF)



Time-series correlation methods

Pros

- More robust way to find relevant lags.
- Can indicate whether there is any predictive information in the historic time series at all.
- Can help identify important seasonalities.

Cons

- Can be difficult to interpret correlation plots.
- Time consuming to interpret and read correlation plots ➔ not scalable to large number of features.
- Even if one feature is not highly correlated with the target it could still be predictive in the presence of other features ➔ not captured in these methods.
- These methods only measure linear relationship between variables.

Summary

Lags of the target and other features can create predictive features for forecasting.

User must decide which lag features to create.

Domain knowledge, feature selection & modelling, and time-series correlation methods can help inform this decision.