Distributed lags

Lag features

Contents

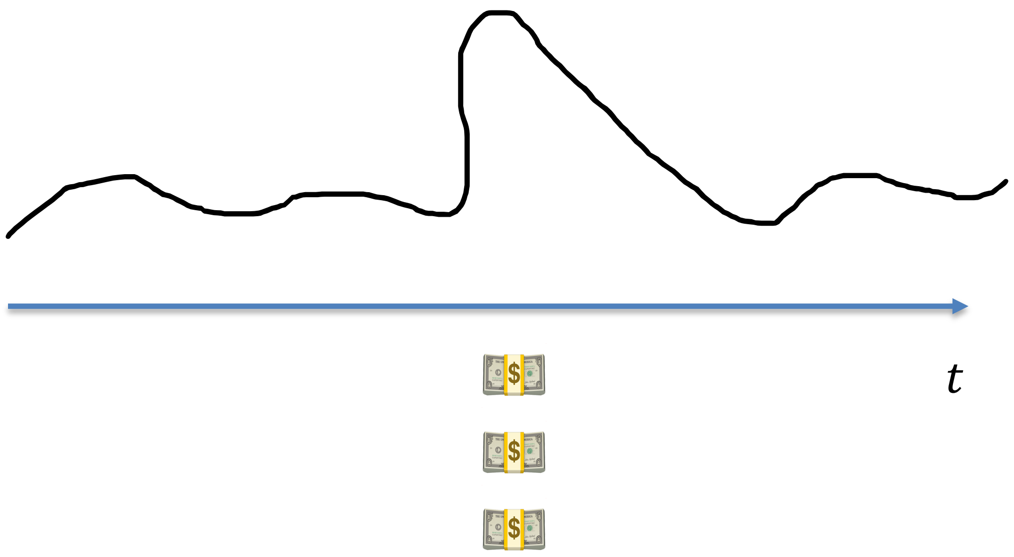
DISTRIBUTED LAGS

WHEN TO USE THEM

Example: advertising spend

- Let's consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t.
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.

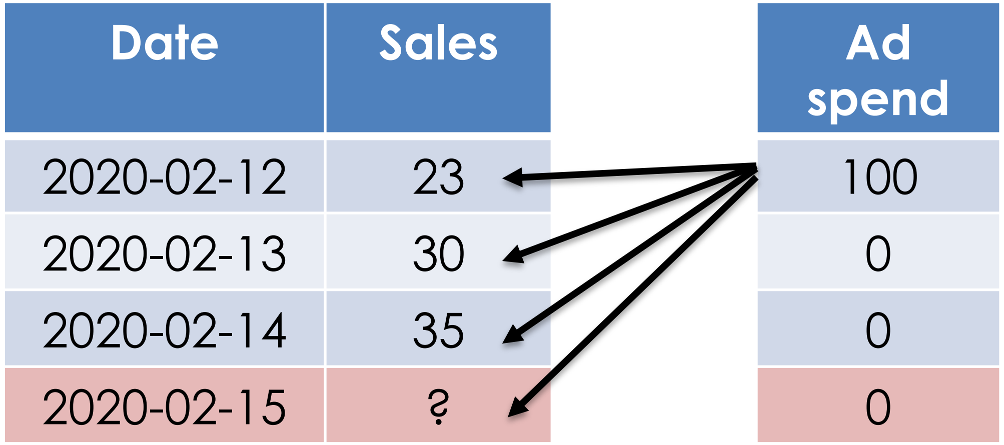
sales



Advertising (ad) spend

Example: advertising spend

- Let’s consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t.
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



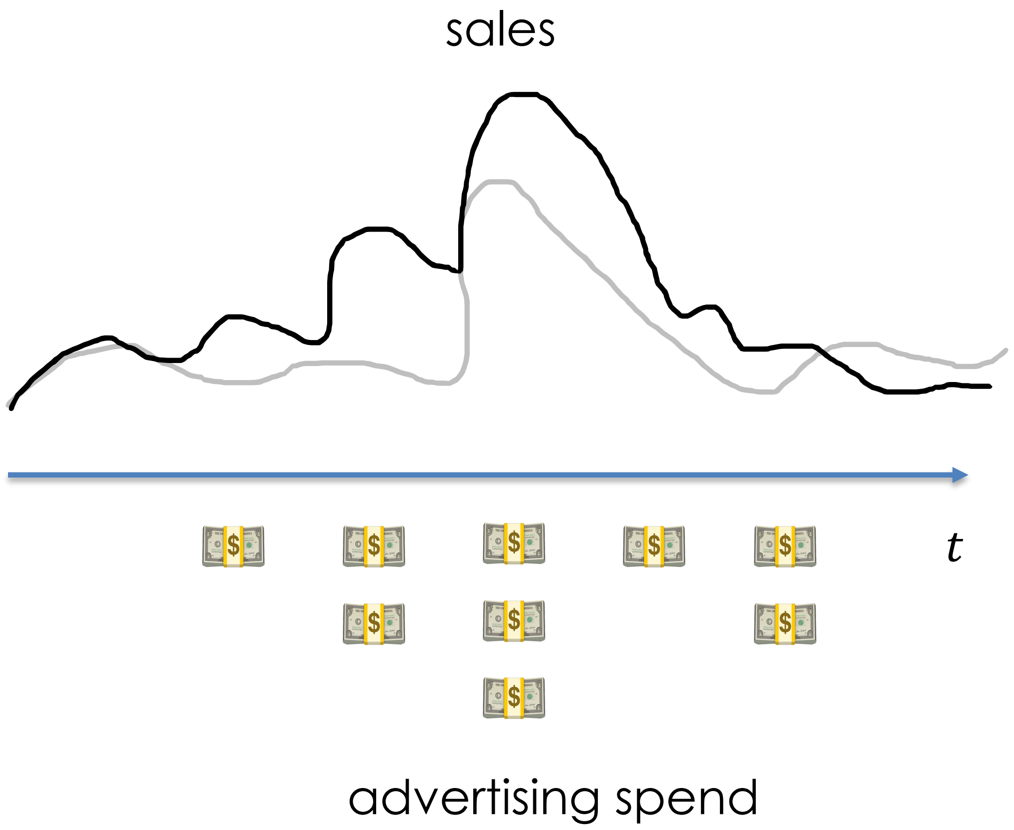
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	0
2020-02-14	35	0
2020-02-15	?	0

Example: advertising spend

- Let's consider sales and advertising spend.

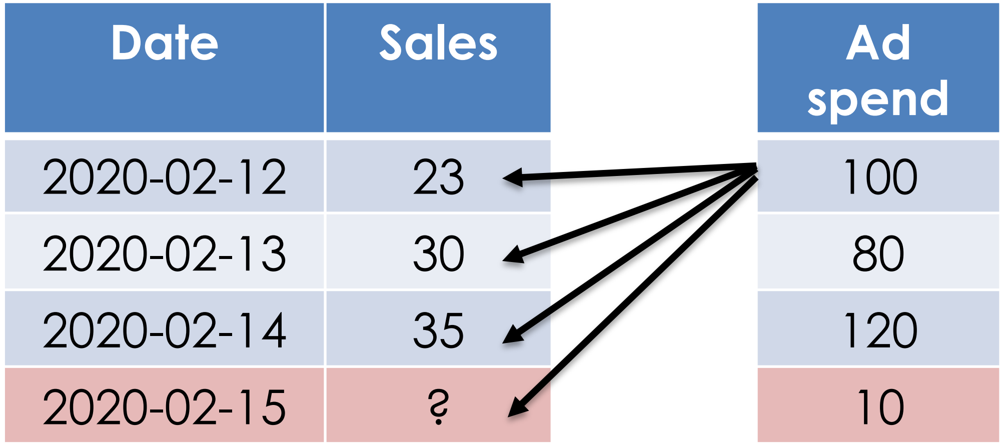
- The impact of advertising on day t will probably last for multiple days into the future after time t.

- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



Example: advertising spend

- Let's consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t.
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



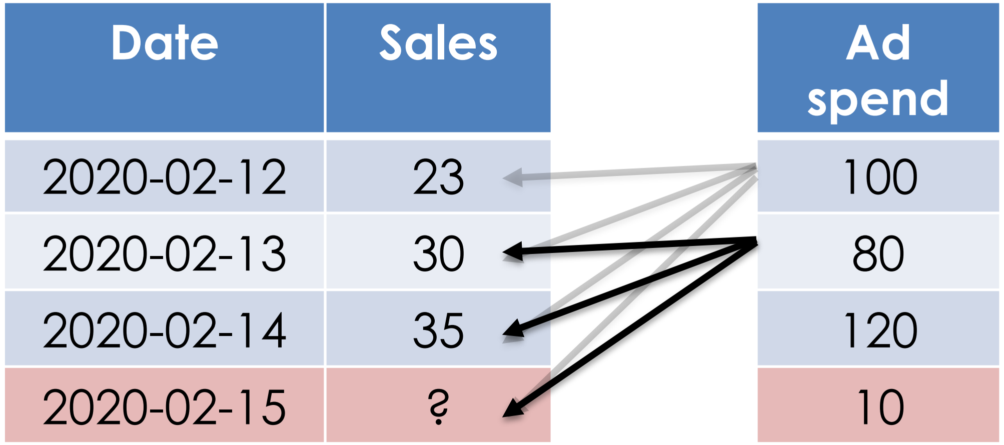
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

Example: advertising spend

- Let’s consider sales and advertising spend.

- The impact of advertising on day t will probably last for multiple days into the future after time t.

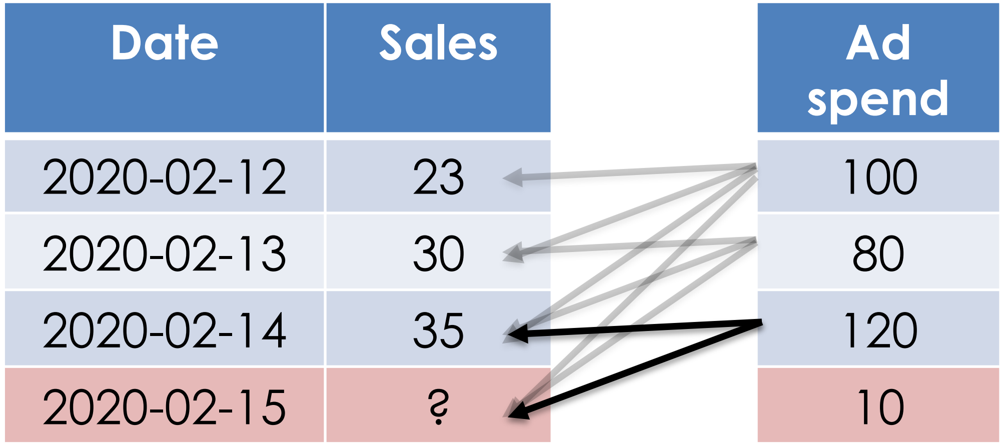
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

Example: advertising spend

- Let’s consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t.
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



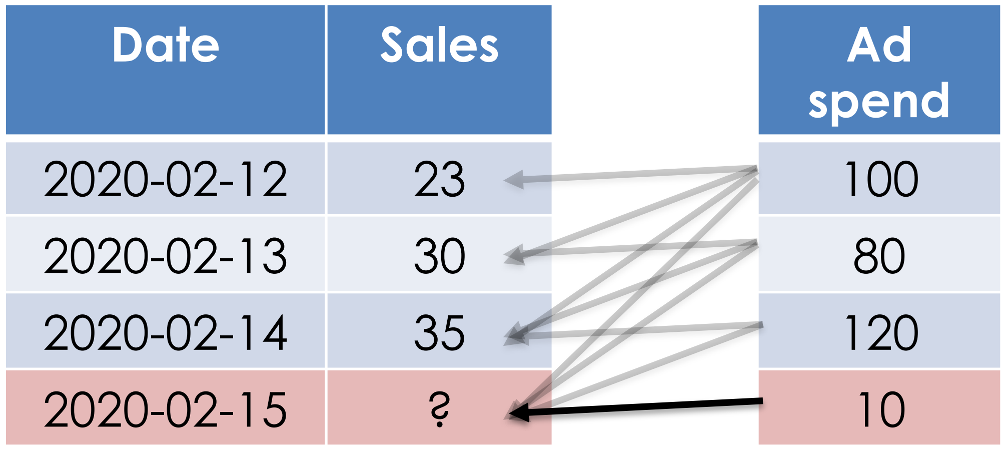
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

Example: advertising spend

- Let’s consider sales and advertising spend.

- The impact of advertising on day t will probably last for multiple days into the future after time t.

- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



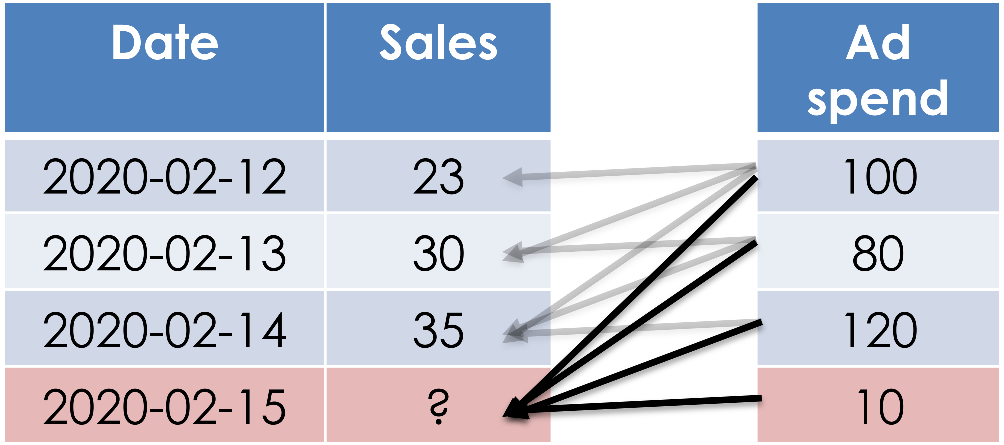
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

Example: advertising spend

- Let’s consider sales and advertising spend.

- The impact of advertising on day t will probably last for multiple days into the future after time t.

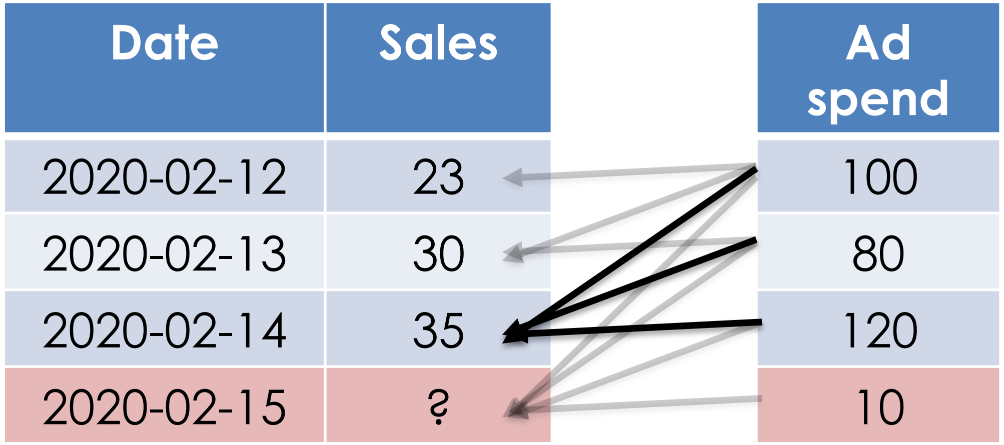
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

Example: advertising spend

- Let’s consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t.
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

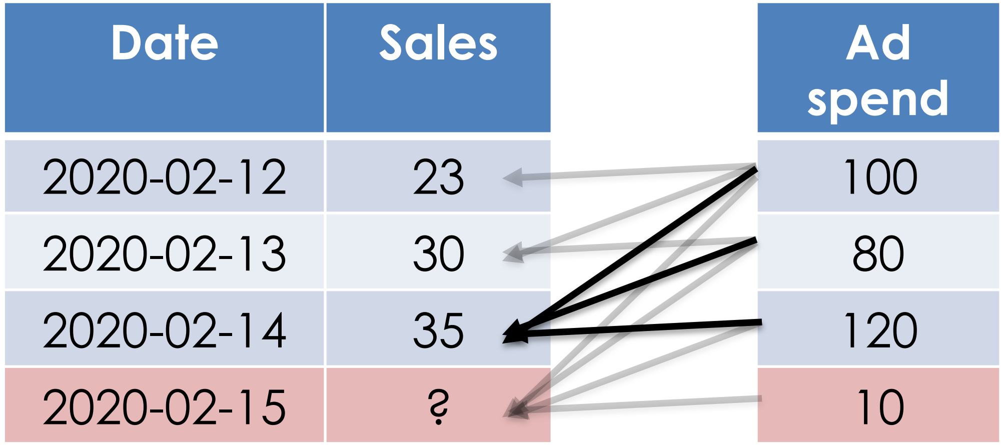
Example: advertising spend

- Let’s consider sales and advertising spend.

- The impact of advertising on day t will probably last for multiple days into the future after time t.

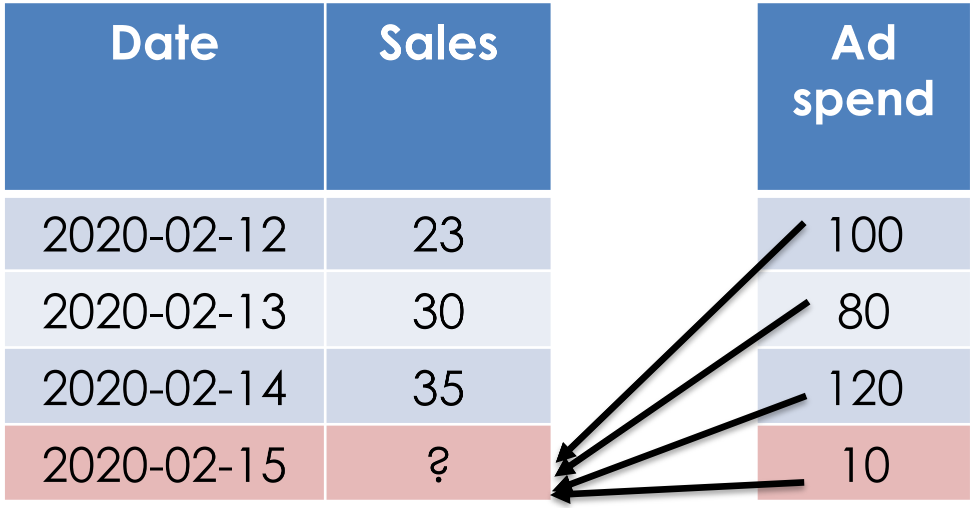
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.

- We can capture this effect using multiple lag features called distributed lags.



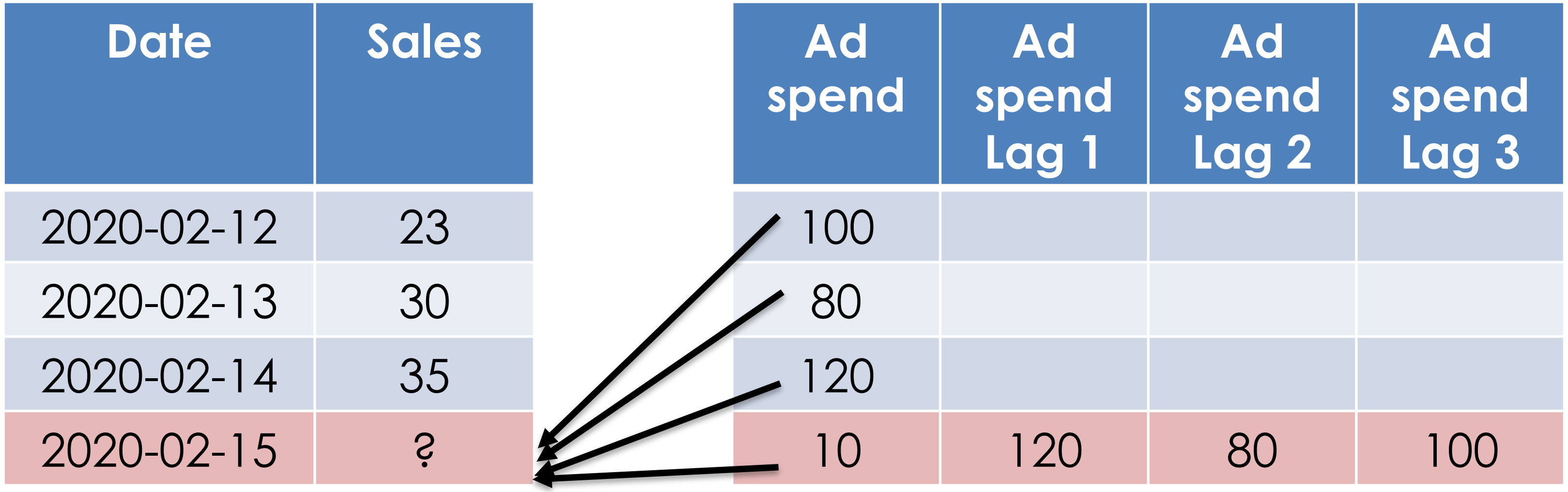
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

Distributed lags



Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

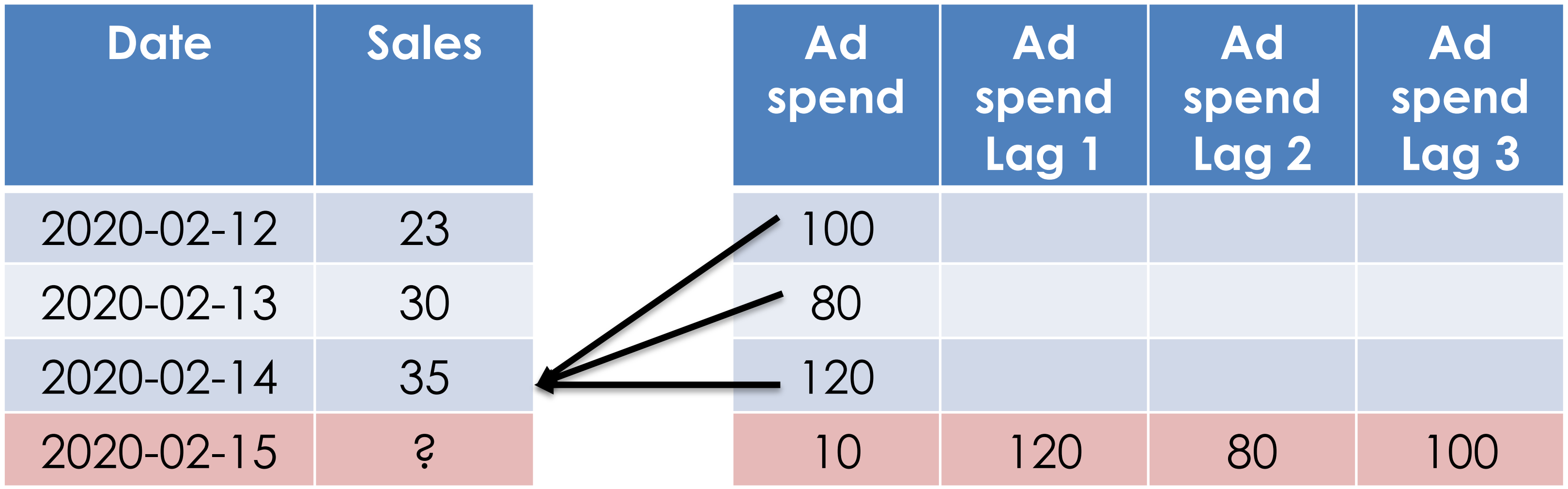
Distributed lags



Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100			
80			
120			
10	120	80	100

Distributed lags



Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

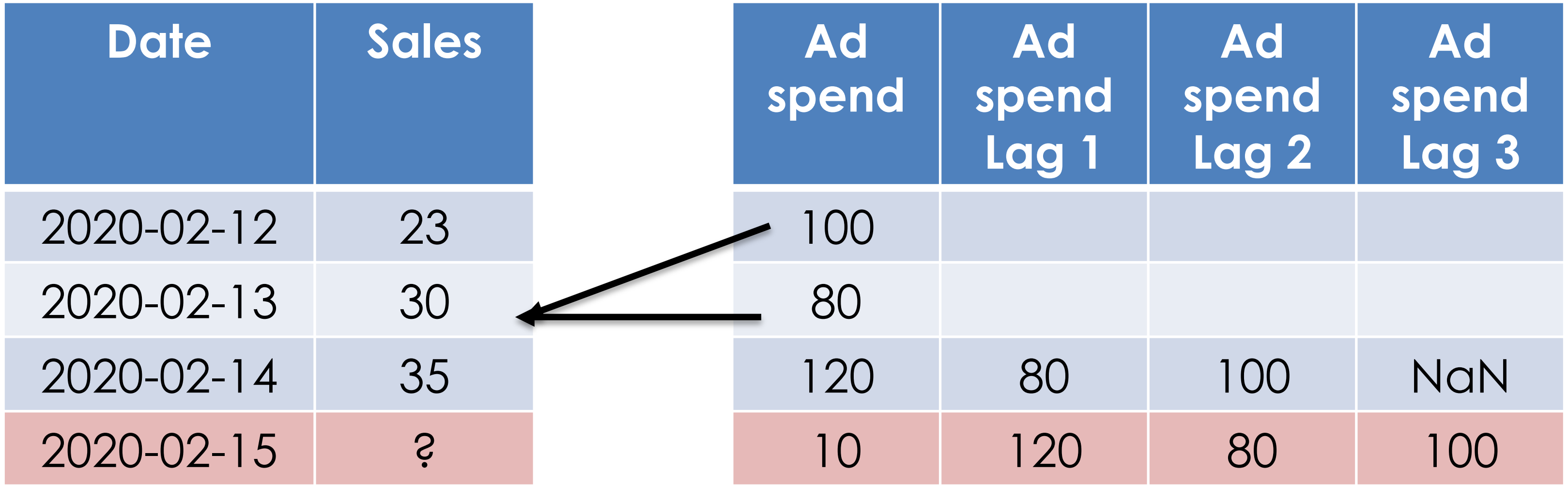
Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100			
80			
120			
10	120	80	100

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100			
80			
120	80	100	NaN
10	120	80	100

Distributed lags



Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100			
80			
120	80	100	NaN
10	120	80	100

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100			
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100			
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

What is the maximum lag to use for the distributed lag? It would be the amount of time that we expect the effect of the feature to influence the target variable.

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Practically speaking the most recent lags will carry most of the predictive information.

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Alternatively a large maximum lag can be set. Then a model & feature selection methods (e.g., LASSO) can decide which lag features to keep.

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Downside: If you apply a distributed lag to many of your original features you will create **a lot** of additional features.

Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

Which features should we pick to lag? Any which we believe can have an impact on future values of the target variable. This can be selected either by domain knowledge or the CCF.

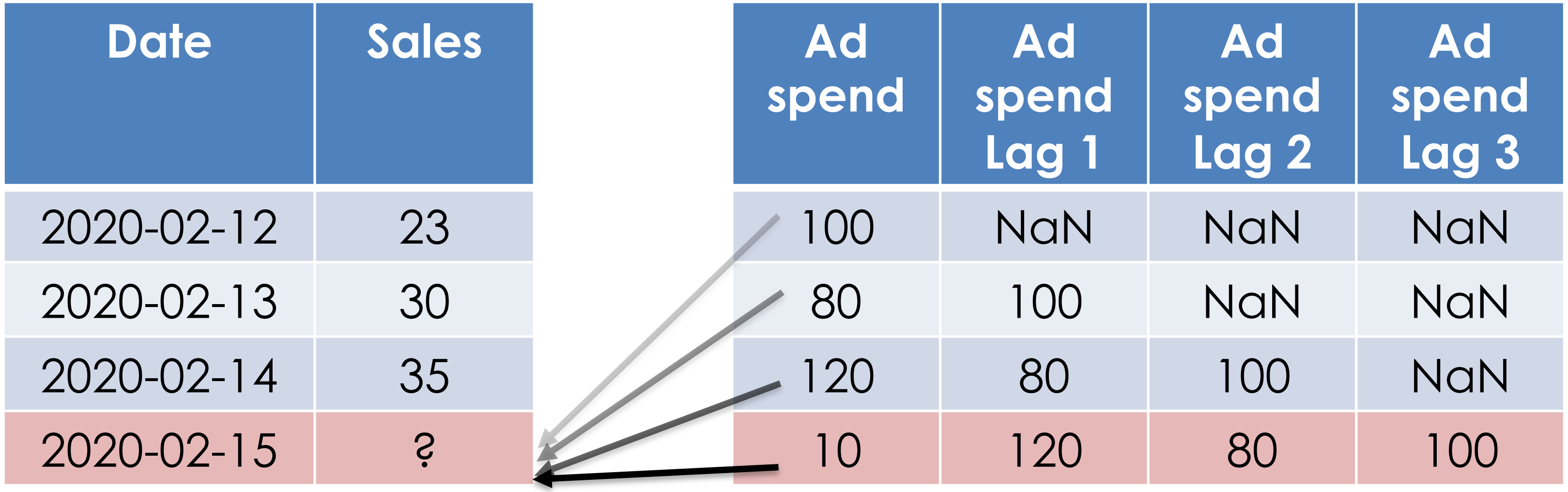
Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

It may be desirable that larger lags have less impact on the target than smaller lags in a model.

Distributed lags

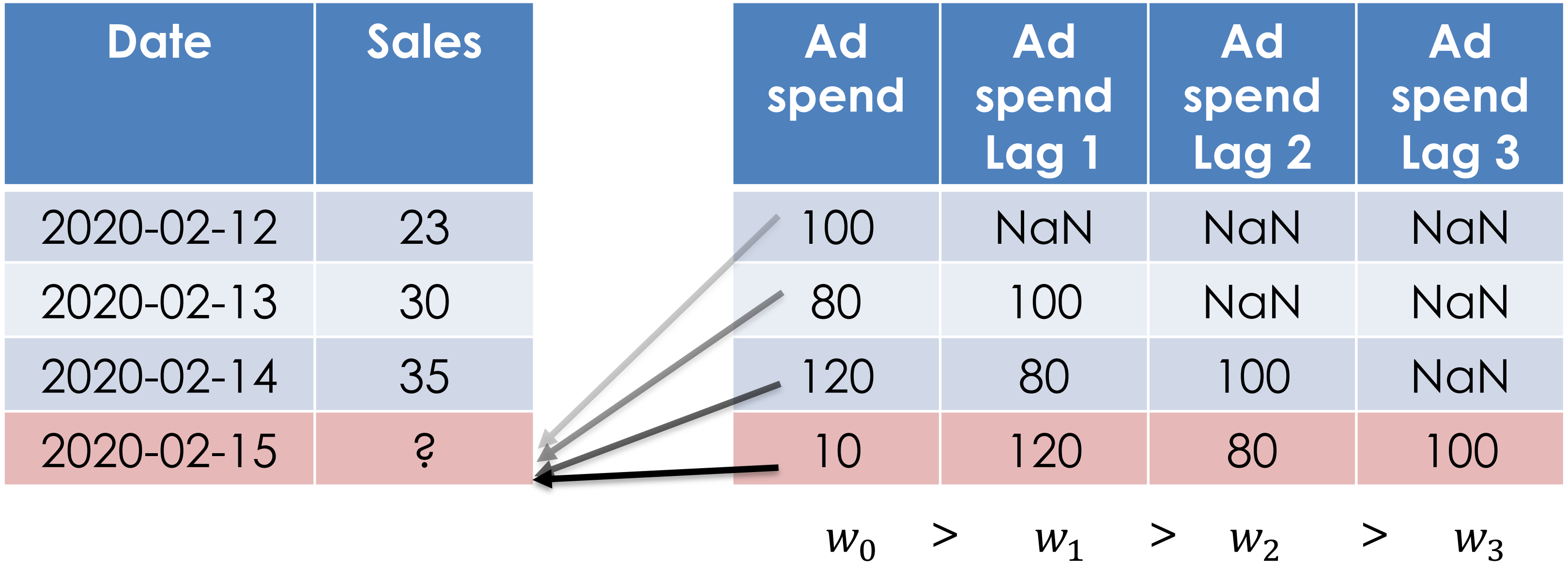


Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

It may be desirable that larger lags have less impact on the target than smaller lags in a model.

Distributed lags



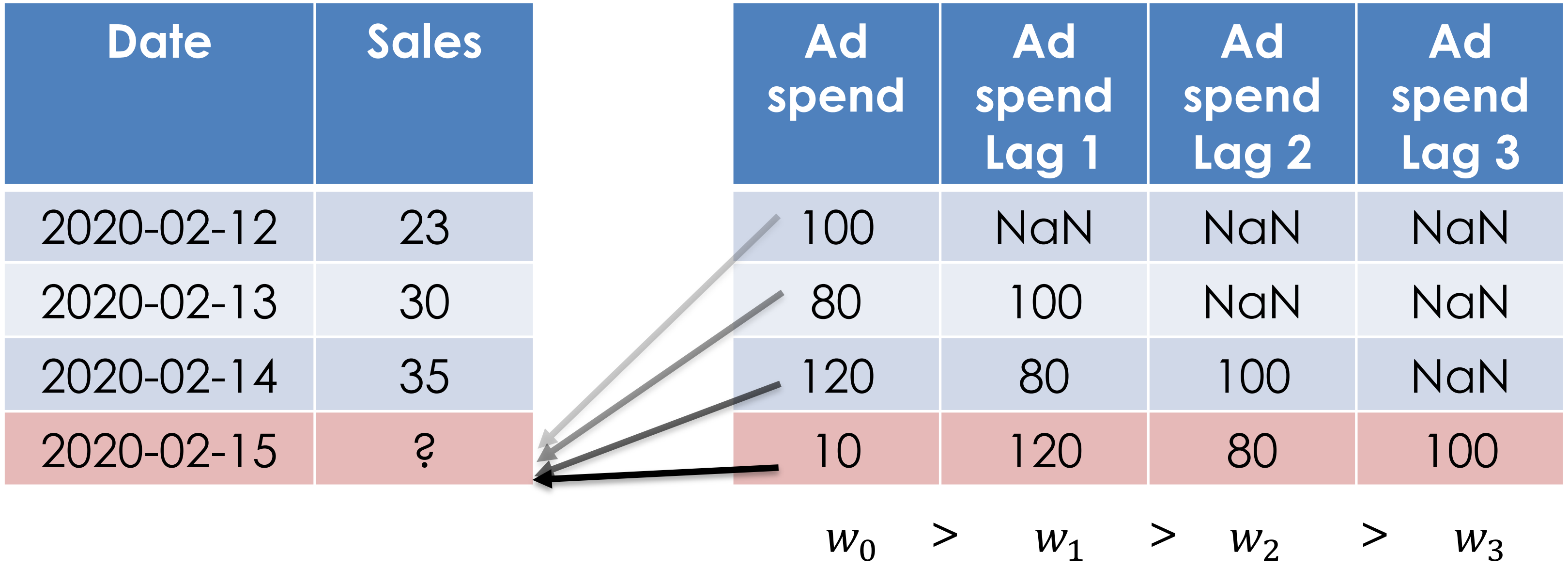
Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

$w_0 > w_1 > w_2 > w_3$

It may be desirable that larger lags have less impact on the target than smaller lags in a model.

Distributed lags



Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

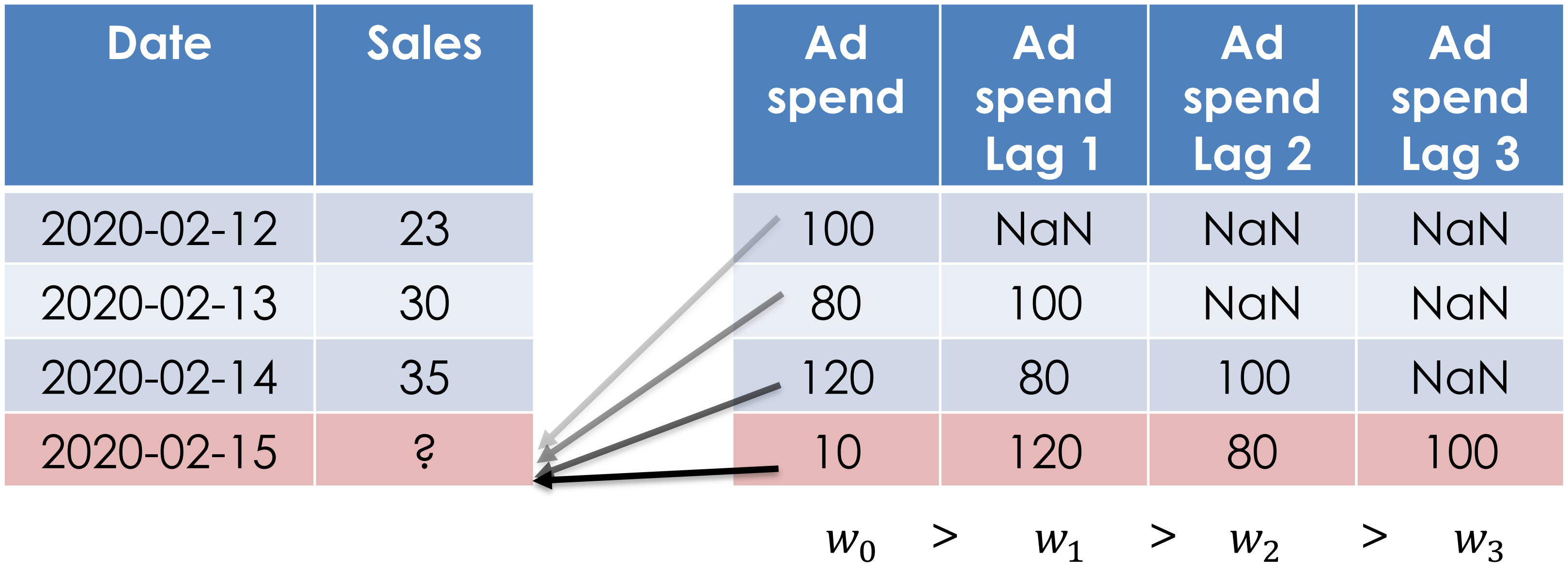
Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

$w_0 > w_1 > w_2 > w_3$

It may be desirable that larger lags have less impact on the target than smaller lags in a model.

The ability to enforce this behaviour depends on the type of model (e.g., linear model vs tree-based models).

Distributed lags



Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
100	NaN	NaN	NaN
80	100	NaN	NaN
120	80	100	NaN
10	120	80	100

$w_0 \quad > \quad w_1 \quad > \quad w_2 \quad > \quad w_3$

We will see in later sections that **window features** allows us to capture this intuition in a feature that is usable by any regression model. It also produces fewer new features relative to distributed lags to capture the same idea.

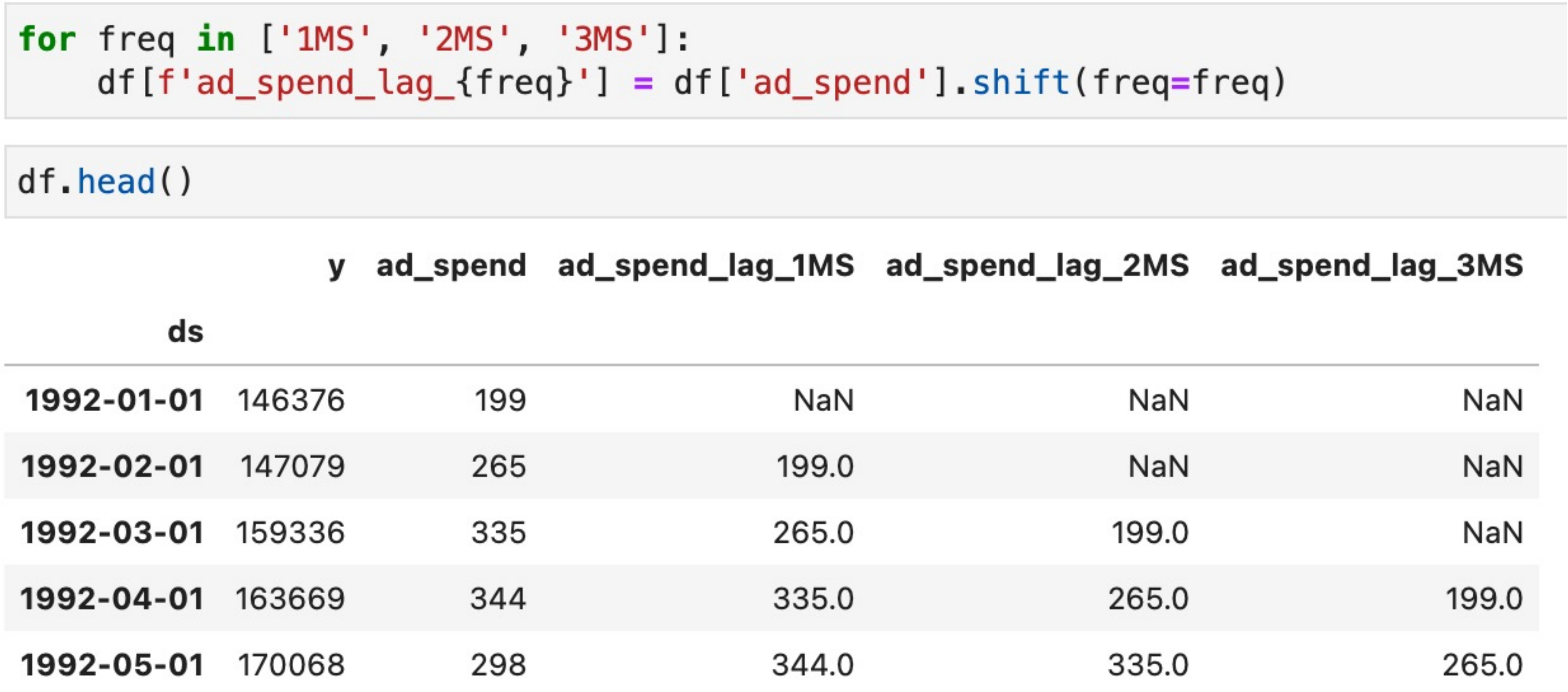
Distributed lags

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

Ad spend	Ad Spend window
100	
80	
120	
10	65

We will see in later sections that **window features** allows us to capture this intuition in a feature that is usable by any regression model. It also produces fewer new features relative to distributed lags to capture the same idea.

Implementation in Pandas

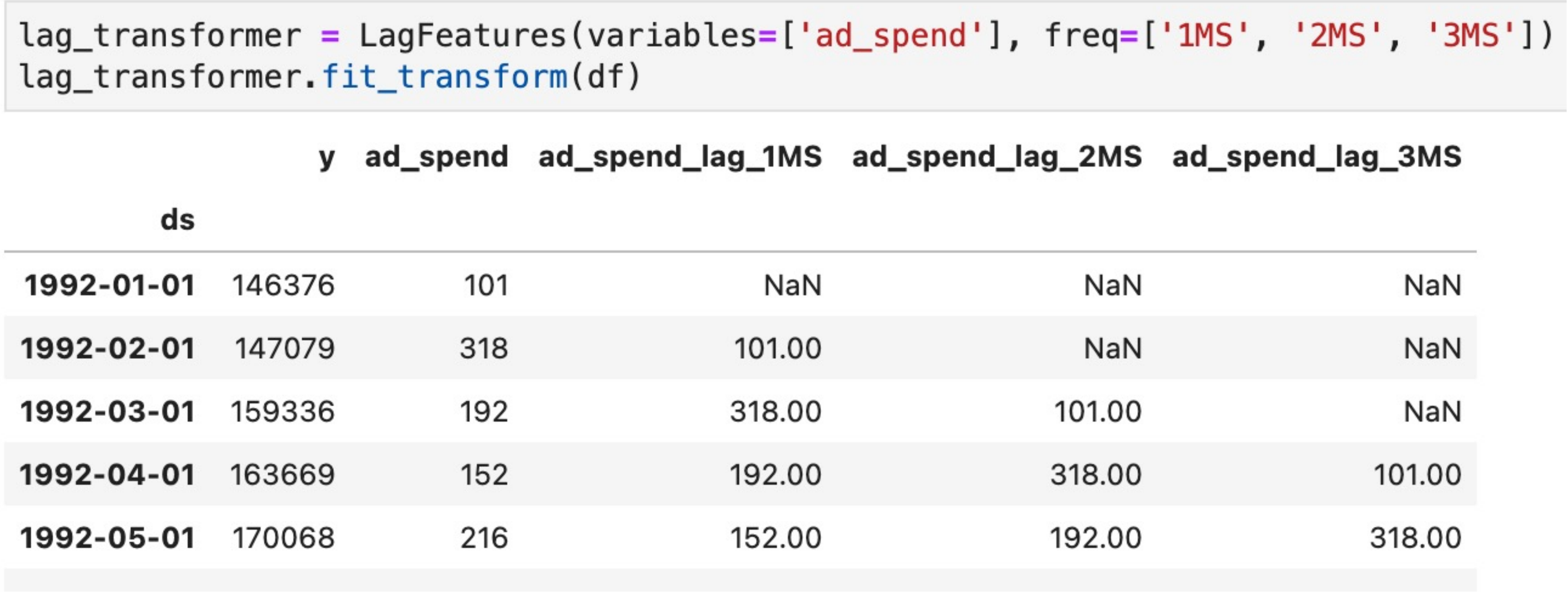


```python
for freq in ['1MS', '2MS', '3MS']:
    df[f'ad_spend_lag_{freq}'] = df['ad_spend'].shift(freq=freq)
```

```python
df.head()
```

	y	ad_spend	ad_spend_lag_1MS	ad_spend_lag_2MS	ad_spend_lag_3MS
ds					
1992-01-01	146376	199	NaN	NaN	NaN
1992-02-01	147079	265	199.0	NaN	NaN
1992-03-01	159336	335	265.0	199.0	NaN
1992-04-01	163669	344	335.0	265.0	199.0
1992-05-01	170068	298	344.0	335.0	265.0

Implementation in Feature-engine



```
lag_transformer = LagFeatures(variables=['ad_spend'], freq=['1MS', '2MS', '3MS'])
lag_transformer.fit_transform(df)
```

	y	ad_spend	ad_spend_lag_1MS	ad_spend_lag_2MS	ad_spend_lag_3MS
ds					
1992-01-01	146376	101	NaN	NaN	NaN
1992-02-01	147079	318	101.00	NaN	NaN
1992-03-01	159336	192	318.00	101.00	NaN
1992-04-01	163669	152	192.00	318.00	101.00
1992-05-01	170068	216	152.00	192.00	318.00

Summary

Distributed lags are multiple lags of a variable that has an impact distributed over time.

The maximum lag to use for a distributed lag depends on how much impact that variable has on future values of the target.

Distributed lags increase the number of features by the max lag. Doing this for many variables can result in a lot of features.