

Distributed lags

Lag features

Contents



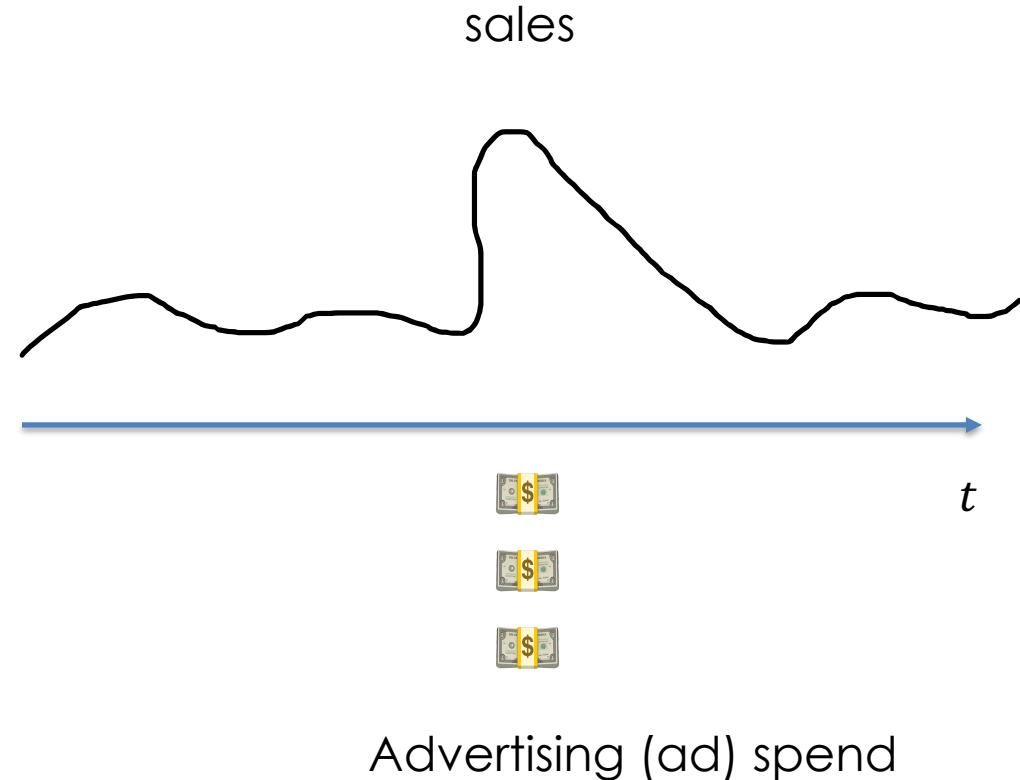
DISTRIBUTED LAGS



WHEN TO USE THEM

Example: advertising spend

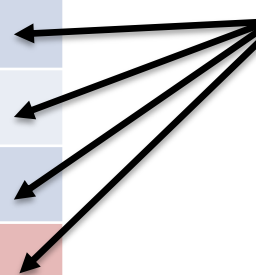
- Let's consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t .
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



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.

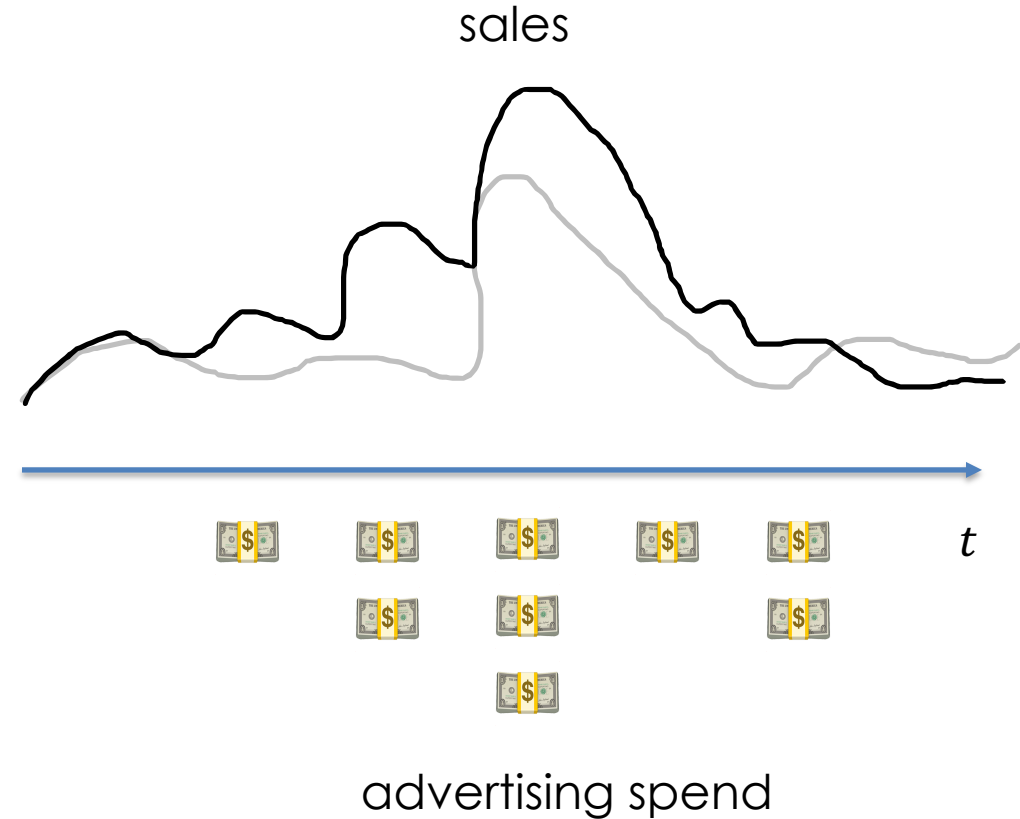
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	0
2020-02-14	35	0
2020-02-15	?	0



The diagram illustrates the lagged effect of advertising spend on sales. Four arrows originate from the 'Ad spend' column and point to the 'Sales' column for the following days: from 100 on Feb 12 to 23 on Feb 12, from 0 on Feb 13 to 30 on Feb 13, from 0 on Feb 14 to 35 on Feb 14, and from 0 on Feb 15 to ? on Feb 15. This shows that the advertising spend on a given day influences the sales on that day and the subsequent days.

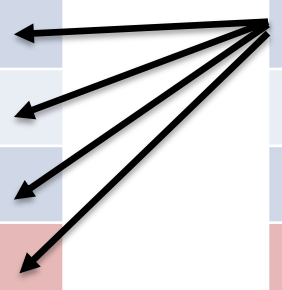
Example: advertising spend

- 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

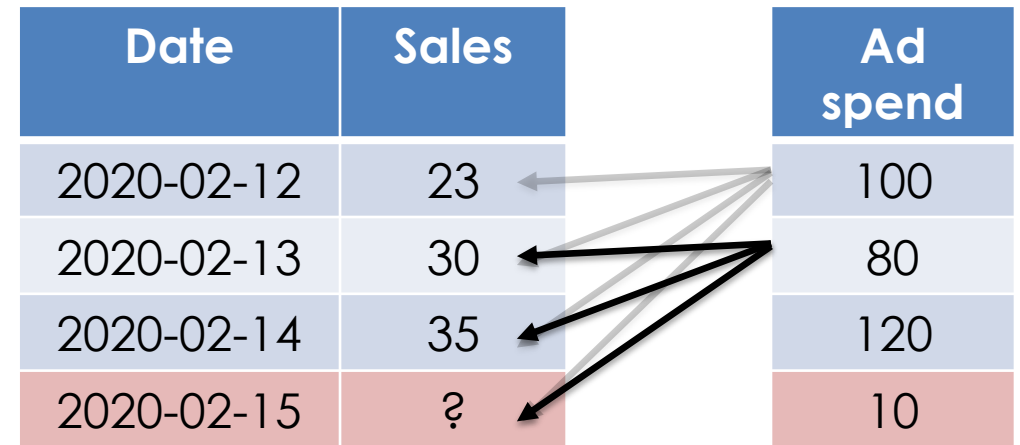
- Let's consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t .
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.



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.



The diagram illustrates the relationship between advertising spend and sales over time. It consists of two tables. The first table, 'Sales', has columns 'Date' and 'Sales'. The second table, 'Ad spend', has columns 'Date' and 'Ad spend'. Arrows point from the 'Ad spend' values to the 'Sales' values for the same day and the following three days, demonstrating that advertising spend on a given day affects sales for multiple subsequent days.

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

Date	Ad spend
2020-02-12	100
2020-02-13	80
2020-02-14	120
2020-02-15	10

Example: advertising spend

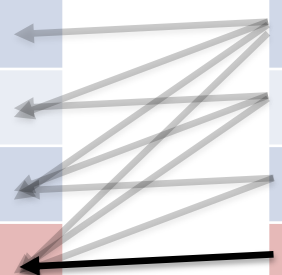
- Let's consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t .
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.

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



The diagram illustrates the relationship between advertising spend and sales over time. Arrows point from the 'Ad spend' column to the 'Sales' column, showing that advertising on a given day affects sales on that day and subsequent days. Specifically, arrows point from the 'Ad spend' row for 2020-02-12 to the 'Sales' rows for 2020-02-12, 2020-02-13, and 2020-02-14. Similarly, arrows point from the 'Ad spend' row for 2020-02-13 to the 'Sales' rows for 2020-02-13, 2020-02-14, and 2020-02-15. An arrow points from the 'Ad spend' row for 2020-02-14 to the 'Sales' row for 2020-02-14. Finally, a thick black arrow points from the 'Ad spend' row for 2020-02-15 to the 'Sales' row for 2020-02-15, indicating the immediate effect of advertising on 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.

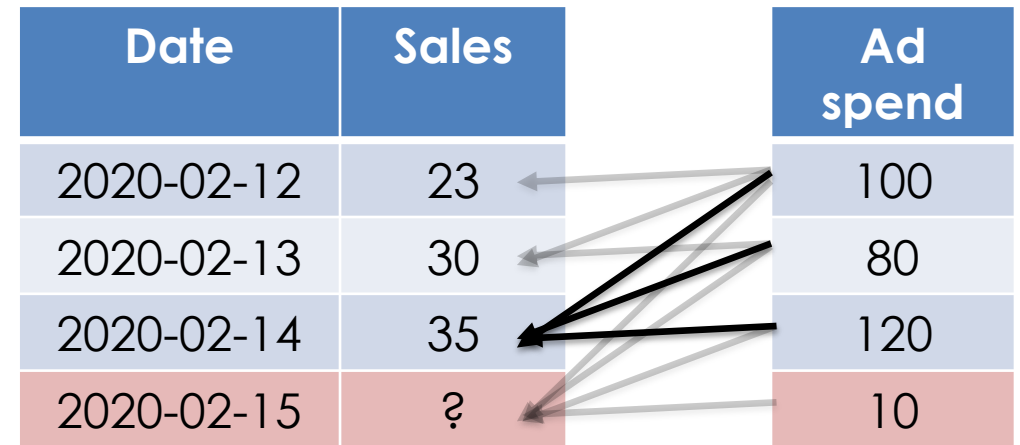
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10

The diagram illustrates the lagged effect of advertising spend on sales. Arrows point from the 'Ad spend' column to the 'Sales' column for the same day and the following three days. Specifically, arrows connect (2020-02-12, 100) to (2020-02-12, 23), (2020-02-13, 30), (2020-02-14, 35), and (2020-02-15, ?). Similarly, arrows connect (2020-02-13, 80) to (2020-02-13, 30), (2020-02-14, 35), and (2020-02-15, ?). Finally, arrows connect (2020-02-14, 120) to (2020-02-14, 35) and (2020-02-15, ?). The row for 2020-02-15 is highlighted in red, indicating the current day of interest.

Example: advertising spend

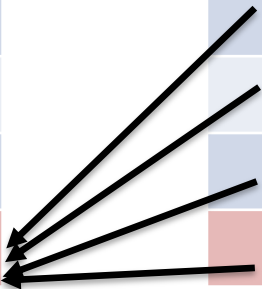
- Let's consider sales and advertising spend.
- The impact of advertising on day t will probably last for multiple days into the future after time t .
- Therefore, the sales on a given day is influenced by ad spend on previous days as well as the same day.
- We can capture this effect using multiple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

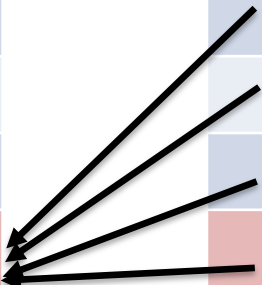
Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	80
2020-02-14	35	120
2020-02-15	?	10



The diagram illustrates the concept of distributed lags. Four black arrows originate from the 'Ad spend' column and point to the 'Sales' column for the date 2020-02-15. The arrows originate from the rows for 2020-02-12, 2020-02-13, 2020-02-14, and 2020-02-15, indicating that the sales on 2020-02-15 are influenced by ad spend from all four days.

Distributed lags

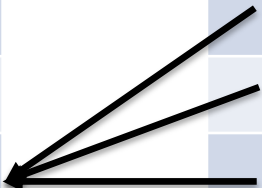
Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100			
2020-02-13	30	80			
2020-02-14	35	120			
2020-02-15	?	10	120	80	100



The diagram illustrates the concept of distributed lags. Four arrows originate from the 'Ad spend' values of the four preceding dates (100, 80, 120, and 10) and point to the 'Sales' cell for 2020-02-15, which contains a question mark. This indicates that the sales on a given day are influenced by ad spend from that day and the three preceding days.

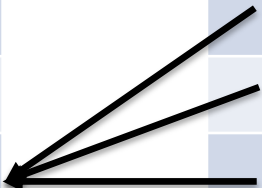
Distributed lags

Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100			
2020-02-13	30	80			
2020-02-14	35	120			
2020-02-15	?	10	120	80	100



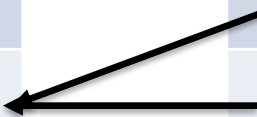
Distributed lags

Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100			
2020-02-13	30	80			
2020-02-14	35	120	80	100	NaN
2020-02-15	?	10	120	80	100



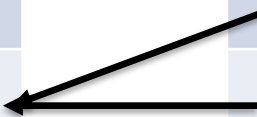
Distributed lags

Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100			
2020-02-13	30	80			
2020-02-14	35	120	80	100	NaN
2020-02-15	?	10	120	80	100



Distributed lags


Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100			
2020-02-13	30	80	100	NaN	NaN
2020-02-14	35	120	80	100	NaN
2020-02-15	?	10	120	80	100



The diagram illustrates the concept of distributed lags. Two arrows originate from the 'Ad spend' column and point to the 'Sales' column. One arrow points from the 'Ad spend' value of 100 on 2020-02-12 to the 'Sales' value of 23 on 2020-02-12. The other arrow points from the 'Ad spend' value of 80 on 2020-02-13 to the 'Sales' value of 30 on 2020-02-13. This shows that the impact of advertising spend is distributed over time, affecting sales in the current period and subsequent periods.

Distributed lags

Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100			
2020-02-13	30	80	100	NaN	NaN
2020-02-14	35	120	80	100	NaN
2020-02-15	?	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

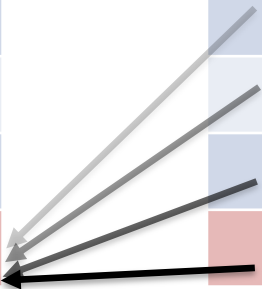
Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	35
2020-02-15	?

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	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100	NaN	NaN	NaN
2020-02-13	30	80	100	NaN	NaN
2020-02-14	35	120	80	100	NaN
2020-02-15	?	10	120	80	100



The diagram illustrates the concept of distributed lags. It shows four arrows pointing from the 'Ad spend' column to the 'Sales' cell for the date 2020-02-15. The arrows originate from the values 100, 80, 120, and 10 in the 'Ad spend' column, which correspond to the dates 2020-02-12, 2020-02-13, 2020-02-14, and 2020-02-15 respectively. This indicates that the sales on 2020-02-15 are influenced by ad spend from all four days, with the impact being distributed across the lags.

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

Distributed lags

Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100	NaN	NaN	NaN
2020-02-13	30	80	100	NaN	NaN
2020-02-14	35	120	80	100	NaN
2020-02-15	?	10	120	80	100

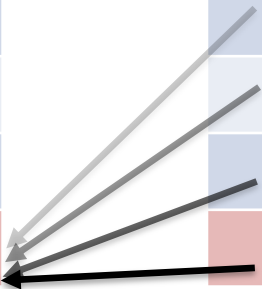
$w_0 > w_1 > w_2 > w_3$

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

Distributed lags

Date	Sales	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100	NaN	NaN	NaN
2020-02-13	30	80	100	NaN	NaN
2020-02-14	35	120	80	100	NaN
2020-02-15	?	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	Ad spend	Ad spend Lag 1	Ad spend Lag 2	Ad spend Lag 3
2020-02-12	23	100	NaN	NaN	NaN
2020-02-13	30	80	100	NaN	NaN
2020-02-14	35	120	80	100	NaN
2020-02-15	?	10	120	80	100

$w_0 > w_1 > w_2 > w_3$

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

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

```
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.