Data used to do time series forecasting

Tom Rochette <tom.rochette@coreteks.org>

March 22, 2020 — 42955d63

1 Question

What data do I need to do time series forecasting?

2 Answer

There are three values that you must know for each data point of your time series:

- its entity, which represents a unique value identifying the time series (e.g., a product SKU). Without this information, it is not possible to construct a sequence of points since there's no logical grouping between the points.
- its timestamp, which represents the moment in time the data point was recorded. Without this information, it is not possible to construct a sequence of points since there's no sequential ordering between the points.
- its target, which represents the measurement of the data point itself that we want to predict. Without this information, we have effectively nothing to base ourselves on.

Such information would look as follow when organized in a table:

Entity	Timestamp	Target
A	1	5
A	2	6
A	3	7
B	1	13
B	2	27
B	3	55

Additionally, you may also have recorded additional values at the same time, which can be a useful source of information when trying to predict a time series.

Entity	Timestamp	Target	Value 1
A	1	5	3
A	2	6	2
A	3	7	1
B	1	13	47
B	2	27	33
B	3	55	5

Let see what happened if we removed each of these columns to illustrate their necessity.

2.1 Removing the entity column

Timestamp	Target
1	5
2	6
3	7
1	13
2	27
3	55

Removing the entity effectively leaves us with two values for the same timestamp. If the data was in this format and we were told that each time the timestamp goes below its previous value a new entity was defined, we would be able to reconstruct the initial table with its entity column.

2.2 Removing the timestamp column

Entity	Target
A	5
A	6
A	7
B	13
B	27
B	55

Removing the timestamp gives us the values the entity may take, but we don't know when. Again, if we're told that the rows have been kept in some order, we could reconstruct the timestamp column.

2.3 Removing the target column

Entity	Timestamp
A	1
A	2
A	3
B	1
B	2
B	3

Removing the target column makes this problem impossible to solve. We're left with only the entities that were measured and the time of measurement, but no measurement, which makes the two other values useless.