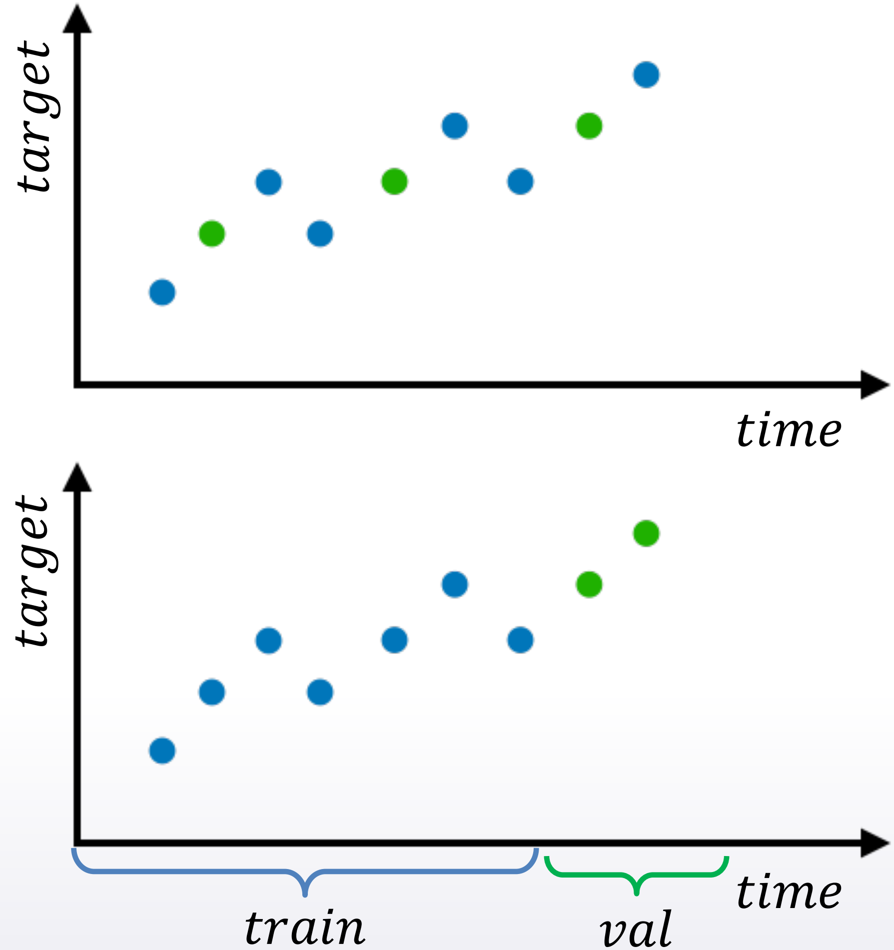
Data splitting strategies

Different approaches to validation

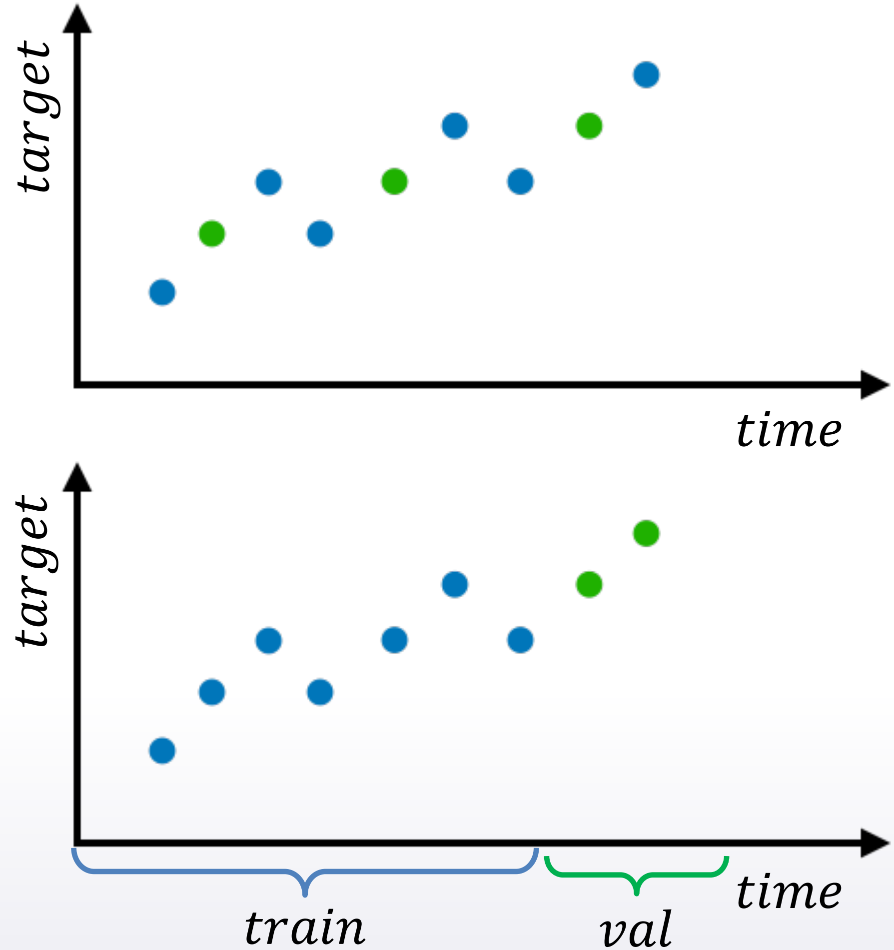
set up validation to replicate train/test split



Different approaches to validation

Important features:

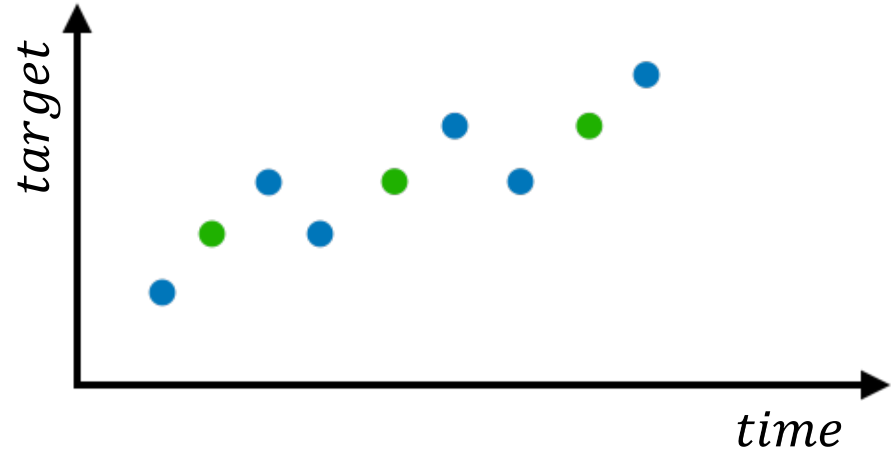
1. Previous and next target values



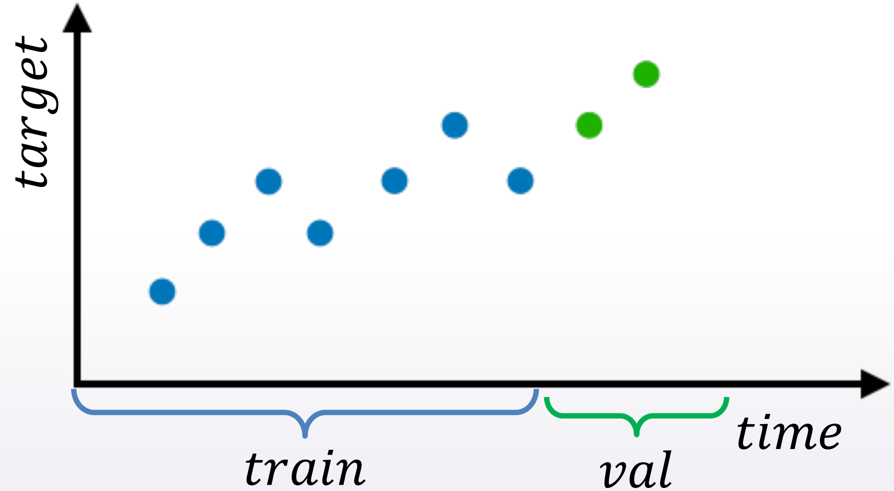
Different approaches to validation

Important features:

1. Previous and next target values



2. Time-based trend

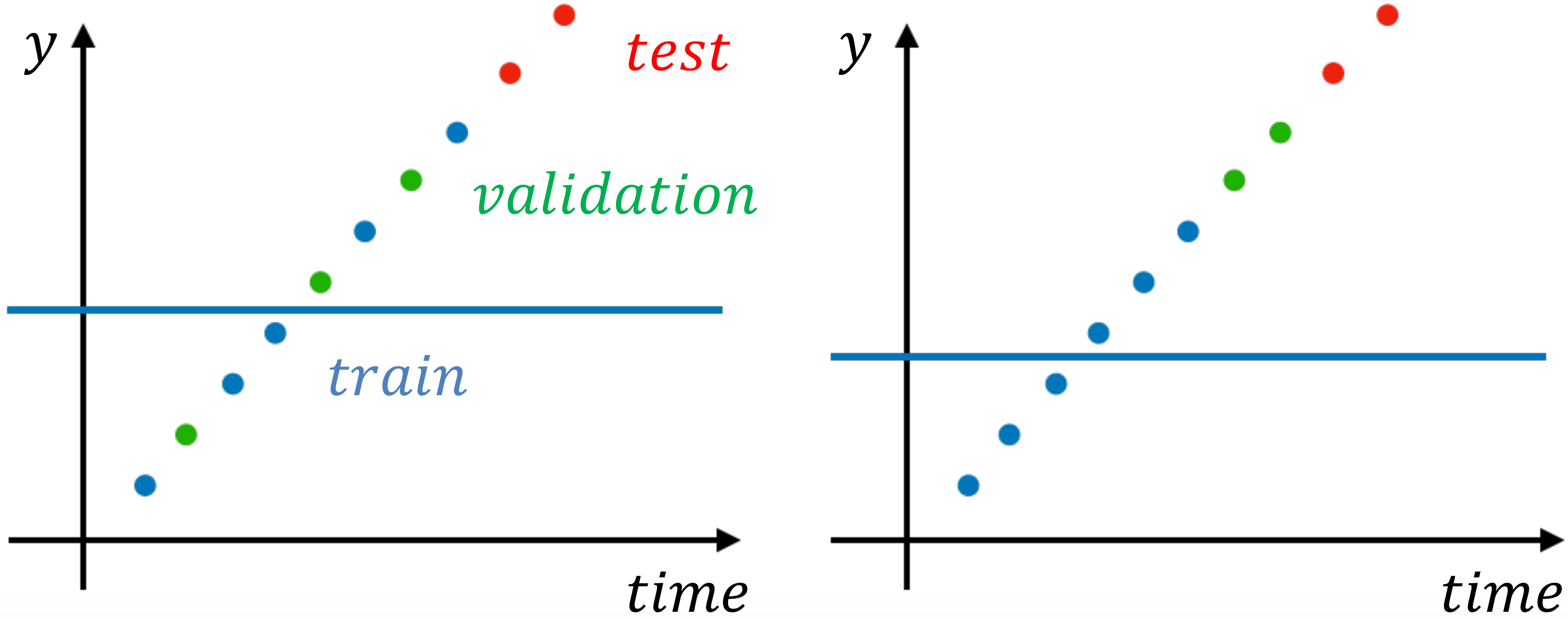


Question screen

If we carefully generate features that are drawing attention to time-based patterns, will we get a reliable validation with a random-based split?

No

Different approaches to validation



Time-based splits

- “Rossman Store Sales”
- “Grupo Bimbo Inventory Demand”

Important outcome

Different splitting strategies can differ significantly

1. in generated features
2. in a way the model will rely on that features
3. in some kind of target leak

Splitting data into train and validation

1. Random, rowwise
2. Timewise
3. By id

Splitting data into train and validation

1. Random, rowwise
2. Timewise
3. By id

the most common way of making a train/test split is to split data randomly by rows. This usually means that the rows are independent of each other.

Splitting data into train and validation

1. Random, rowwise
2. Timewise
3. By id

ROSSMANN

GRUPO BIMBO

If we are to predict a number of customers for the shop for each day in the next week, we can came up with something like the number of customers for the same day in the previous week, or the average number of customers for the past month.

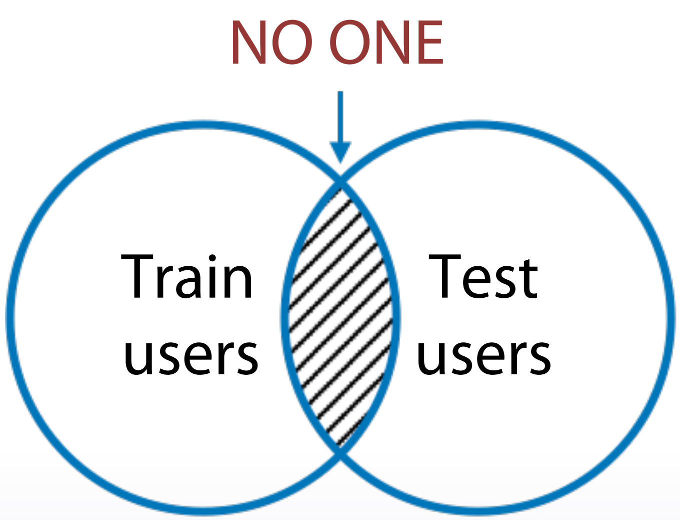
Moving window

Moving window validation

<table>
<tr><th>week1</th><th>week2</th><th>week3</th><th>week4</th><th>week5</th><th>week6</th></tr>
<tr><td colspan="3">train</td><td>validation</td><td></td><td></td></tr>
<tr><td colspan="4">train</td><td>validation</td><td></td></tr>
<tr><td colspan="5">train</td><td>validation</td></tr>
</table>

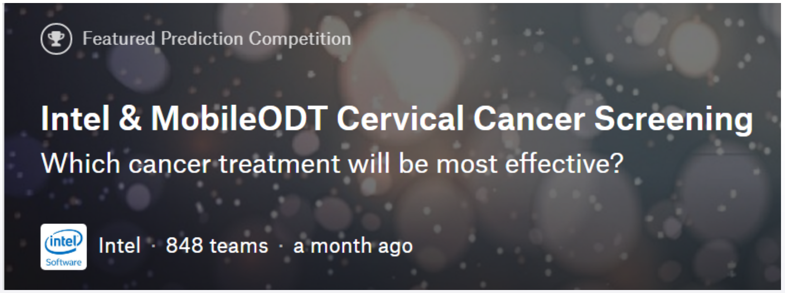
Splitting data into train and validation

1. Random, rowwise
2. Timewise
3. By id



Splitting data into train and validation

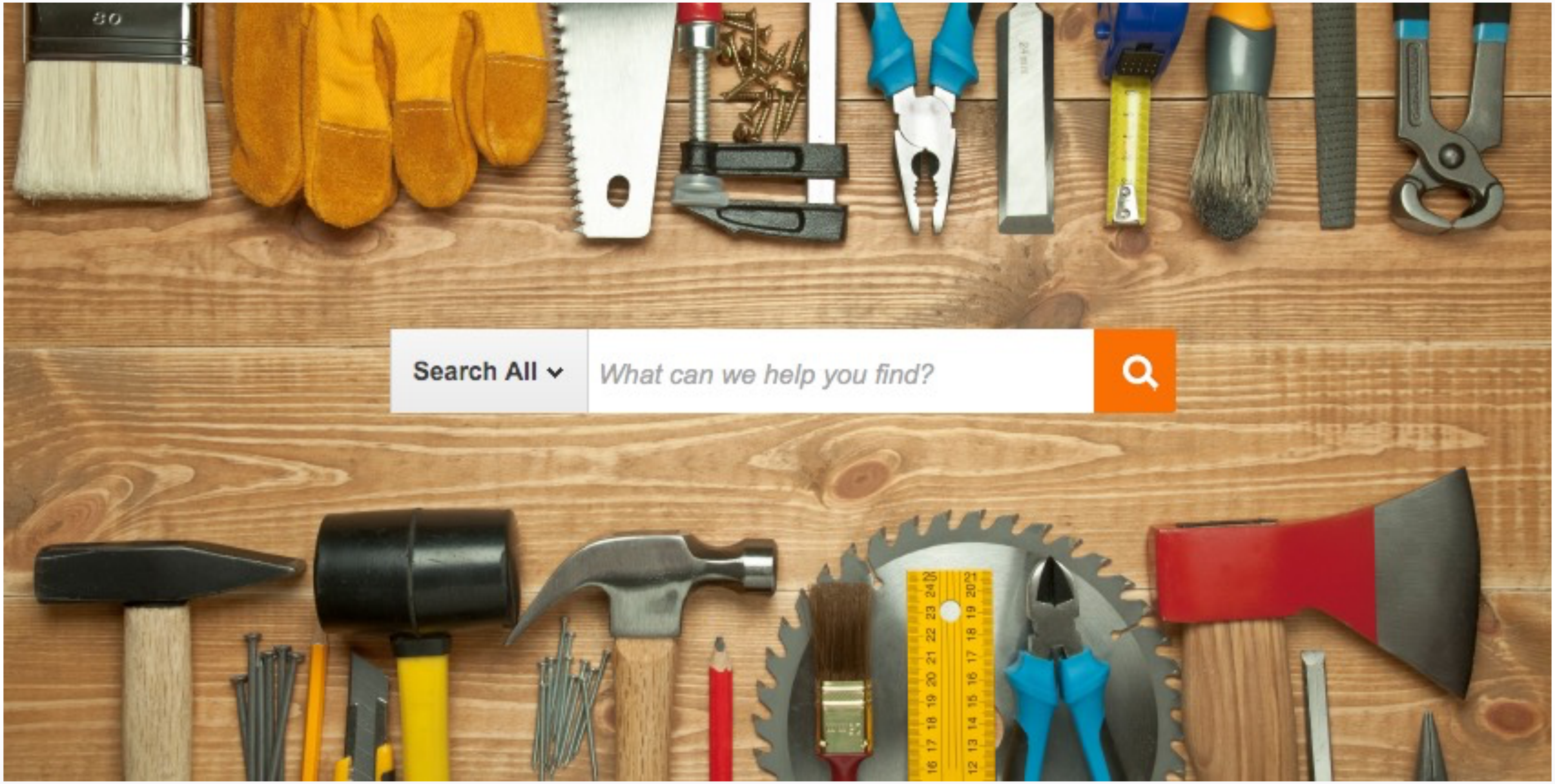
1. Random, rowwise
2. Timewise
3. By id



Splitting data into train and validation

1. Random, rowwise
2. Timewise
3. By id
4. Combined

Home Depot Product Search Relevance



Conclusion

1. In most cases data is split by
    a. Row number
    b. Time
    c. Id
2. Logic of feature generation depends on the data splitting strategy
3. Set up your validation to mimic the train/test split of the competition