Validation strategies

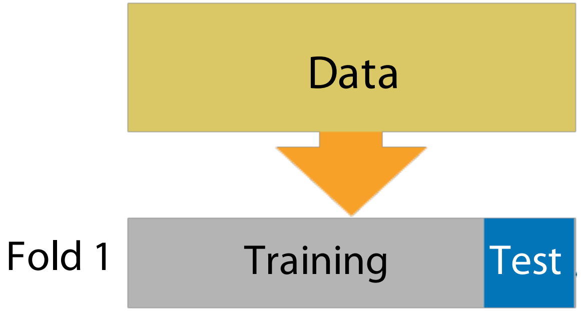
Validation types

- Holdout
- K-fold
- Leave-one-out

Validation types

- Holdout: ngroups = 1

  `sklearn.model_selection.ShuffleSplit`



Validation types

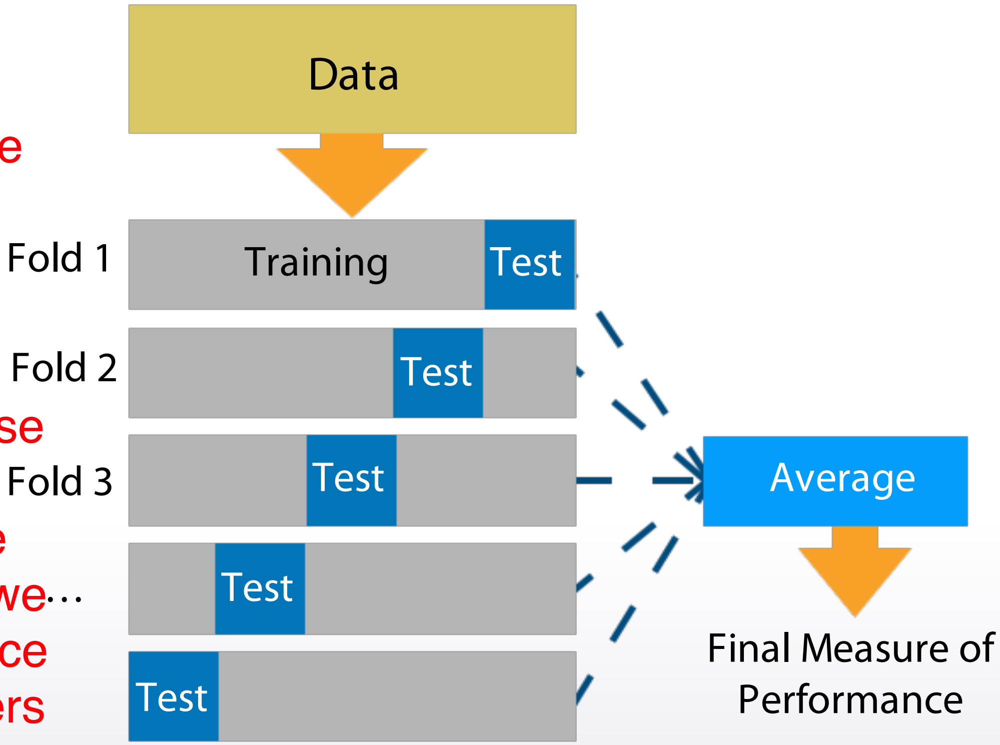
- Holdout: ngroups = 1

  ```
  sklearn.model_selection.ShuffleSplit
  ```

- K-fold: ngroups = k

  ```
  sklearn.model_selection.Kfold
  ```

. Here it is important to understand the difference between K-fold and usual holdout or bits of K-times. While it is possible to average scores they receive after K different holdouts. In this case, some samples may never get invalidation, while others can be there multiple times. On the other side, the core idea of K-fold is that we want to use every sample for validation only once. This method is a good choice when we have a minimum amount of data, and we… can get either a sufficiently big difference in quality, or different optimal parameters between folds.

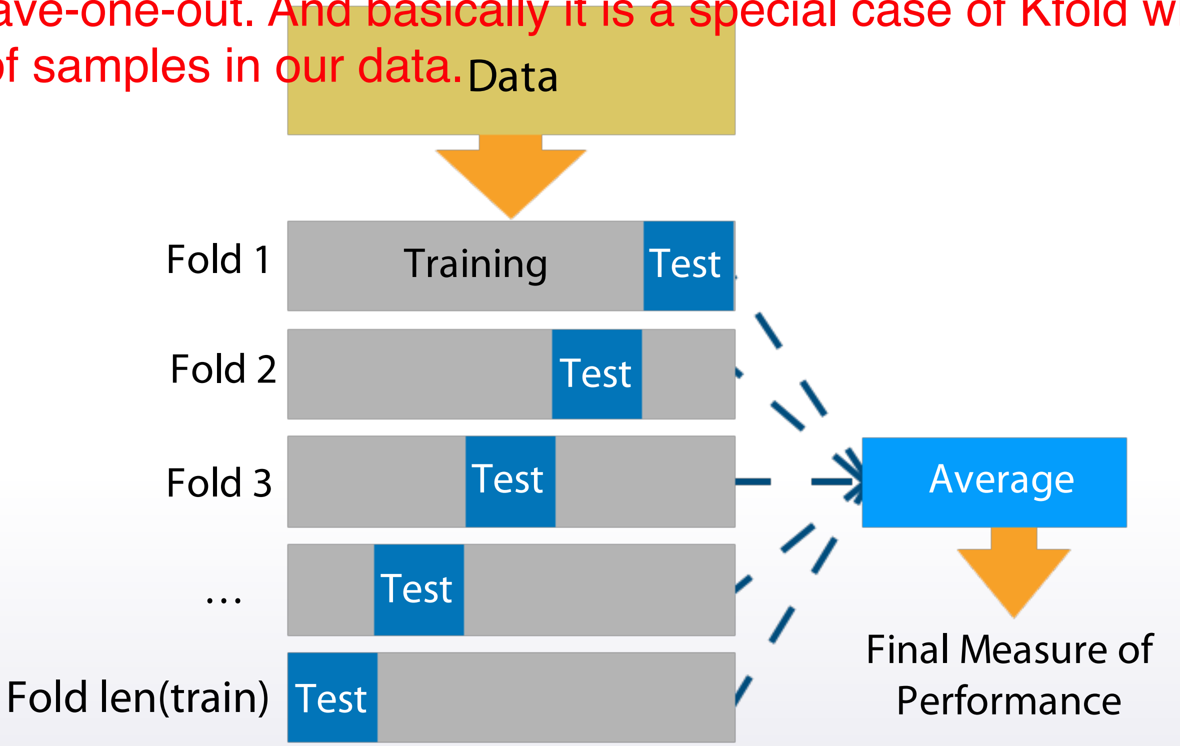


samples may

Validation types

- Holdout: ngroups = 1
  ```
  sklearn.model_selection.ShuffleSplit
  ```
- K-fold: ngroups = k
  ```
  sklearn.model_selection.Kfold
  ```
- Leave-one-out: ngroups = len(train)
  ```
  sklearn.model_selection.LeaveOneOut
  ```

It is called leave-one-out. And basically it is a special case of Kfold when K is equal to the number of samples in our data.



Stratification

Samples and their target values

0	1	0	0	1	1	1	0

Stratification

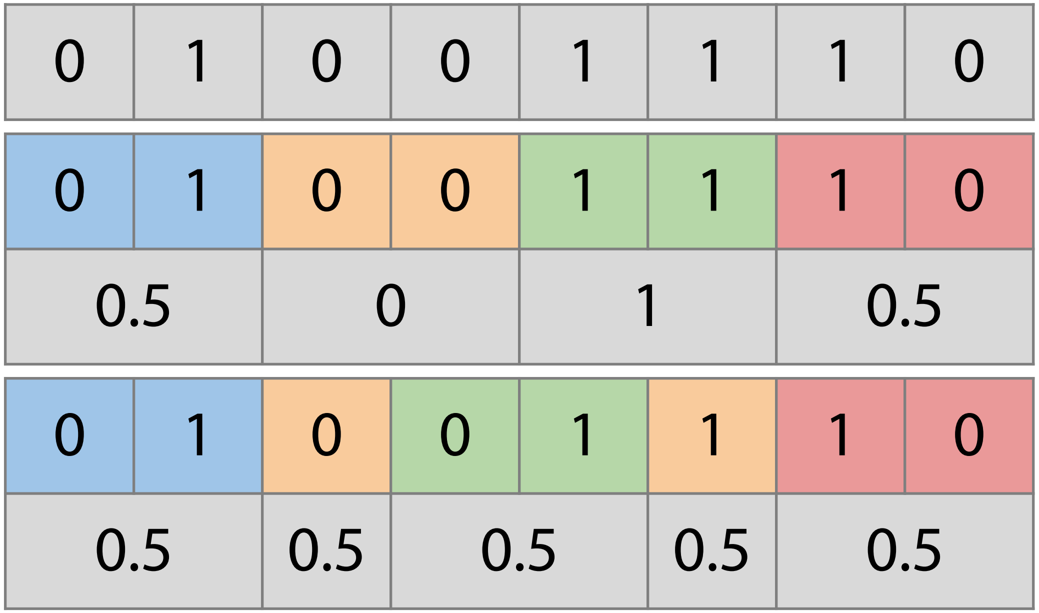
Samples and their target values

0	1	0	0	1	1	1	0

0	1	0	0	1	1	1	0
0.5		0		1		0.5	

Stratification

Samples and their target values



0 1 0 0 1 1 1 0

0 1 0 0 1 1 1 0

0.5 0 1 0.5

0 1 0 0 1 1 1 0

0.5 0.5 0.5 0.5 0.5

Stratification is useful for:

- Small datasets
- Unbalanced datasets
- Multiclass classification

Conclusion

There are three main validation strategies:

1. Holdout
2. KFold
3. LOO

Stratification preserve the same target distribution over different folds

If we have enough data, and we're likely to get similar scores and optimal model's parameters for different splits, we can go with Holdout.

If on the contrary, scores and optimal parameters differ for different splits, we can choose KFold approach. And event, if we too little data, we can apply leave-one-out.

The second big takeaway from this video for you should be stratification. It helps make validation more stable, and especially useful for small and unbalanced datasets.