Split types of Machine Learning

Fast.ai

Aishwarya Deshmane

Why to split data?

If we train a model with all our data, and then evaluate the model using that same data, we would not be able to tell how well our model can perform on data it hasn’t seen.

Dataset can be split into three sets:

- training set - which our model sees in training
- validation set - which is used only for cross validation, model selection and
- test set - which is used for final evaluation

This is done to avoid **‘overfitting’**.

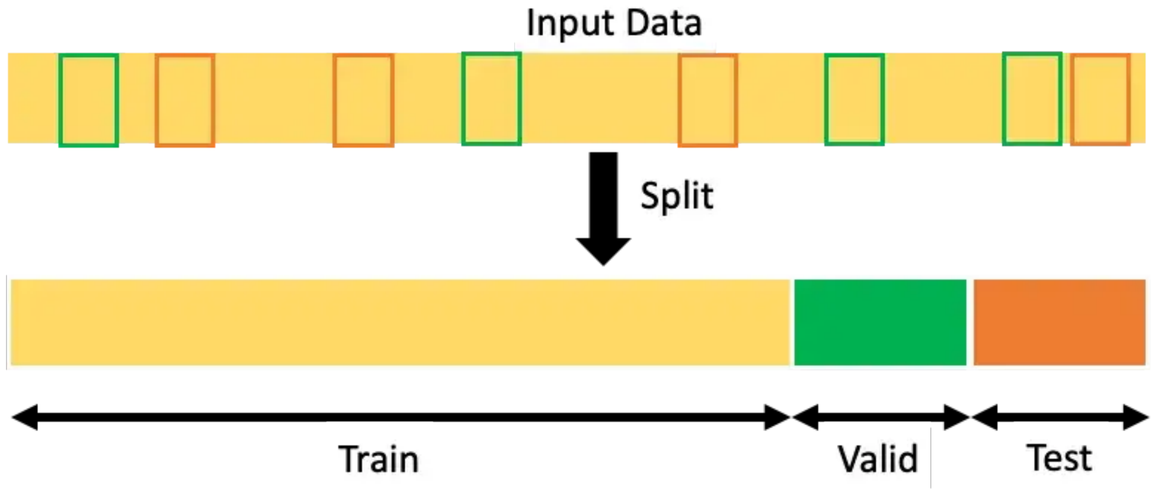
Types of data splitting

1. Splitting randomly

Randomly splits items/rows of dataset into user defined percentages.

Most commonly used method for that unbiased evaluation

It may result in uneven distribution of data.



Functions

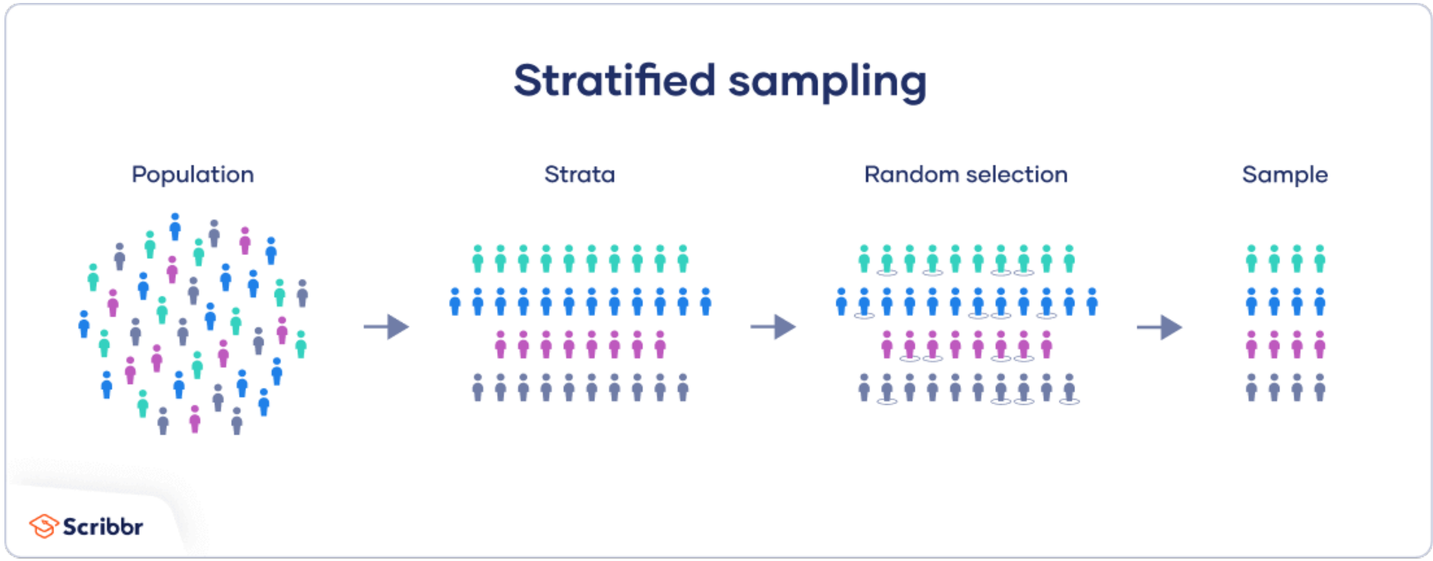
RandomSplitter (valid_pct=0.2, seed=None)

train_valid_test_split(df, target = 'SalePrice', train_size=0.8, valid_size=0.1, test_size=0.1)

2. Stratified random sampling

Selects data samples at random within specific parameters.

Ensures the data is correctly distributed in training and test sets.



https://cdn.scribbr.com/wp-content/uploads/2020/09/stratified-sample-7.png

Functions

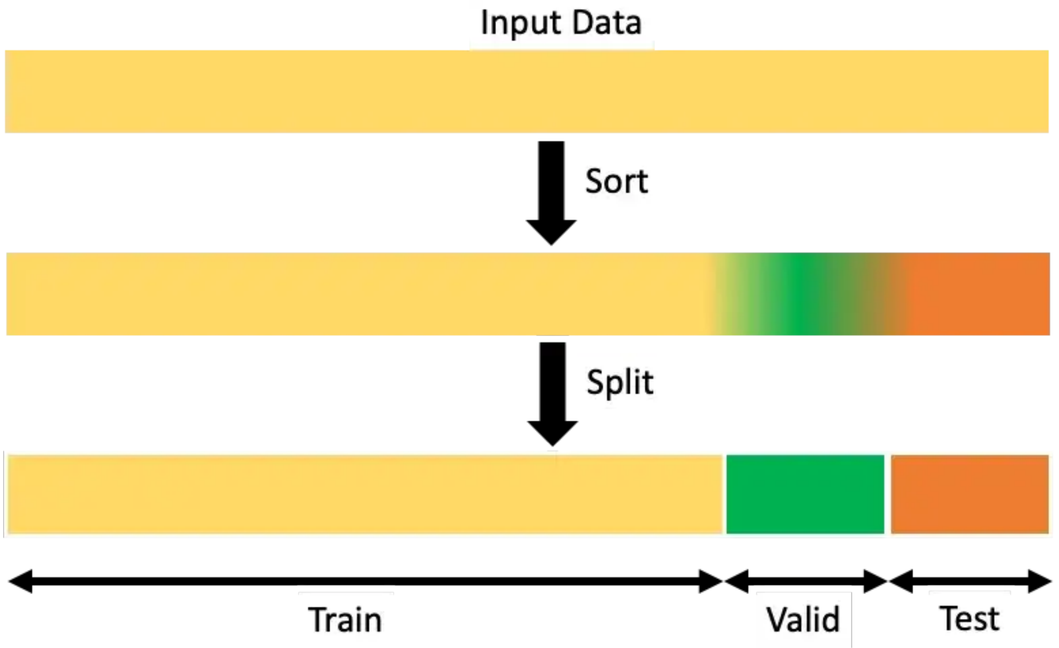
TrainTestSplitter (test_size=0.2, random_state=None, train_size=None, shuffle=True, stratify=labels) - this is sklearn train test utility

3. Nonrandom sampling / Splitting using the temporal component or order

Data modelers want the most recent data as the test set.

Using that temporal variable is a more reliable way of splitting datasets whenever the dataset includes the date variable.

Data need to be ordered before splitting.



Functions

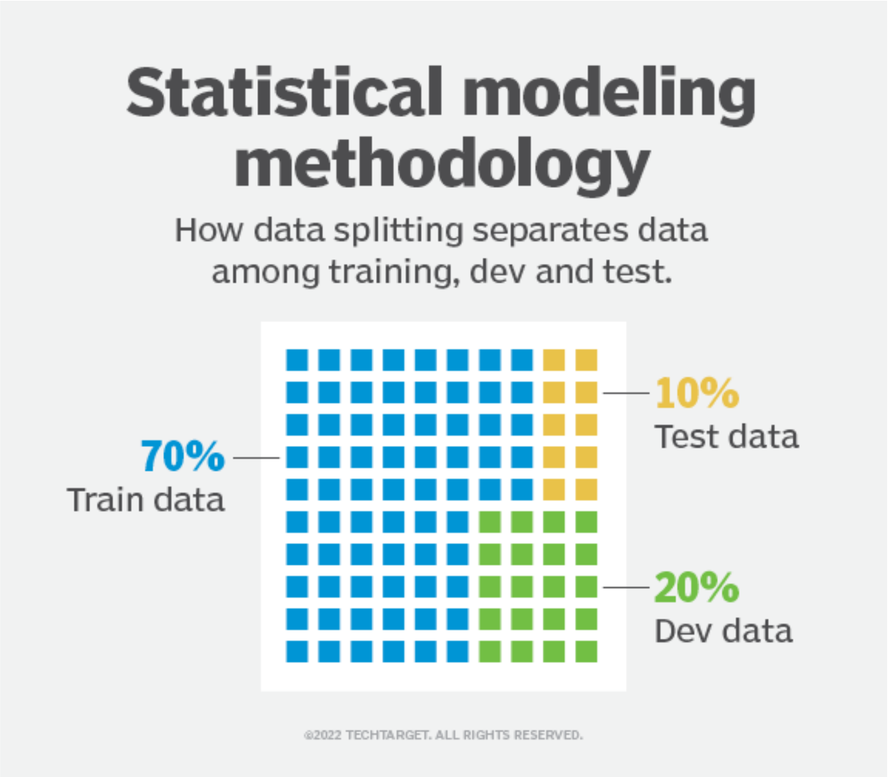
IndexSplitter (valid_idx)

EndSplitter (valid_pct=0.2, valid_last=True)

What proportion to split your dataset?

Depends on total number of samples in your data, and on the actual model you are training.

- The exact ratio depends on the data, but a 70-20-10 ratio for training, dev and test splits is optimal for small data sets.
- Models with very few hyper-parameters will be easy to validate and tune, so it can use small validation set
- Model with many hyper-parameters needs significant validation set
- Model with no hyper-parameters might not need a validation set at all



https://www.techtarget.com/searchenterpriseai/definition/data-splitting

References:

https://docs.fast.ai/data.transforms.html

https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c

https://www.techtarget.com/searchenterpriseai/definition/data-splitting

https://machinelearningmastery.com/difference-test-validation-datasets/

https://www.scribbr.com/methodology/stratified-sampling/