Types of cross-validation

Types of cross-validation

1. K-fold cross-validation

2. Hold-out cross-validation

3. Stratified k-fold cross-validation

4. Leave-p-out cross-validation

5. Leave-one-out cross-validation

6. Monte Carlo (shuffle-split)

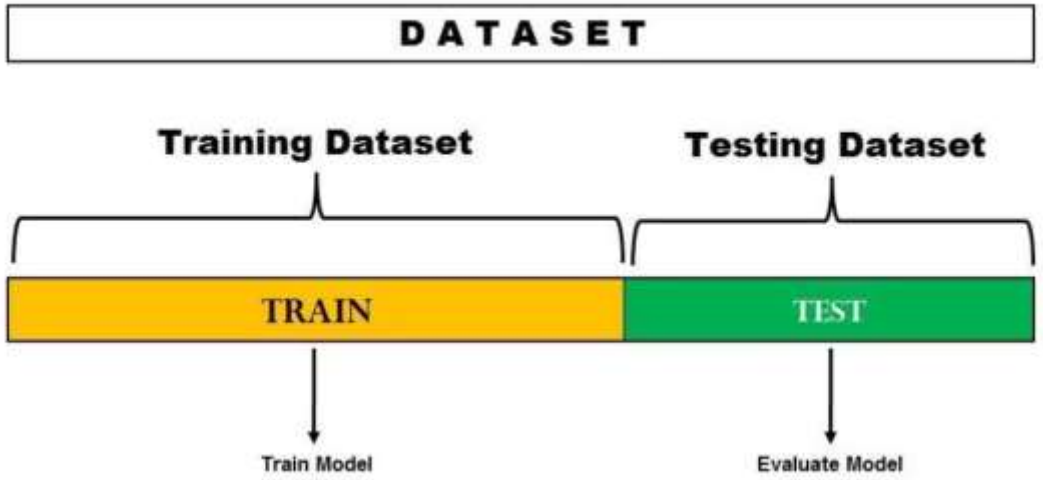
7. Time series (rolling cross-validation)

K-fold cross-validation

K Fold CV, K=5

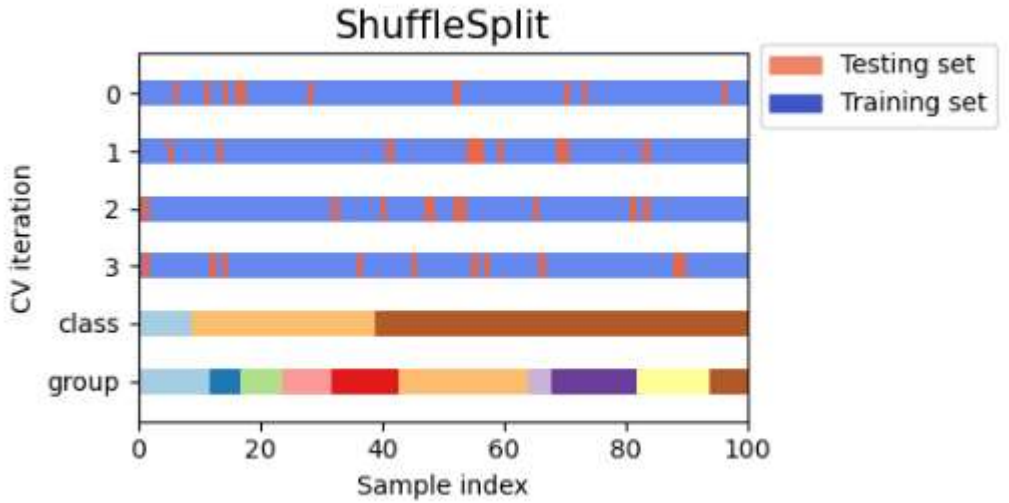
- The dataset is divided into training and test data in different ratios, and a model is created in each iteration.
- If most of the trained model receives a good score, it is a good model.
- If most of the trained models perform poorly, it indicates a bad model.

Hold-out cross-validation



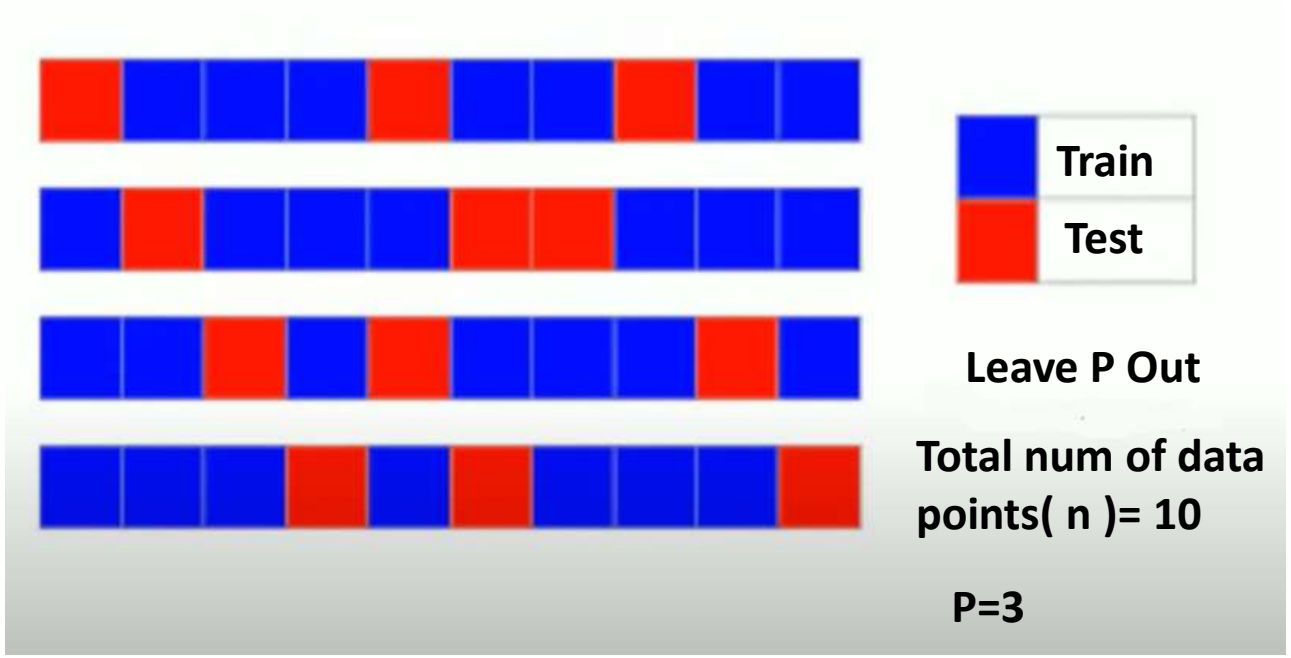
- **Holdout cross-validation** is also called a **train-test split**
- The **dataset** is randomly divided into two parts: a **training set** and a **test set.**
- **70%** of the **dataset** will be used for **training** and **30%** for **validation.**
- The **dataset** is divided into two sets, and the **model** is **built** only **once** on the **training set.**

Stratified k-fold cross-validation



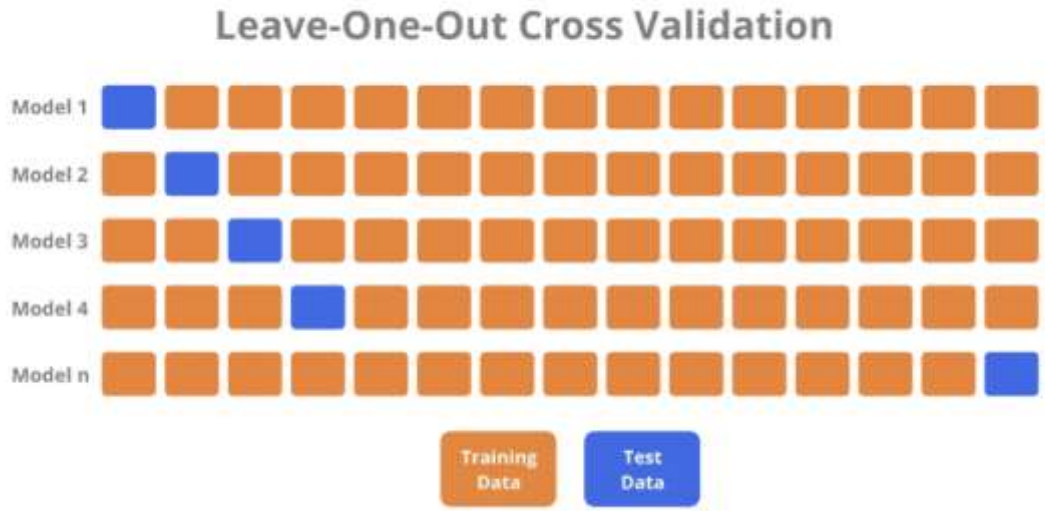
- Stratified k-fold cross-validation is also known as **shuffle split.**
- The Dataset is split into k equal folds, each fold has the same ratio of instances of target variables that are in the complete dataset.

Leave-p-out cross-validation



- A fixed number of **data points**, denoted by **p**, are systematically excluded from the total number of data samples represented by **n**.
- In every iteration, the model is trained on **n-p** data points and later tested on **p** data points.

Leave-one-out cross-validation



- Only **1 sample point** is used as a **validation set,** and the remaining **n-1 samples** are used in the **training set.**

- **Leave-one-out cross-validation** method is **computationally expensive** to perform and shouldn't be used with very large datasets

Monte Carlo (shuffle-split)

- **Splitting** can be done in the percentage of **70-30%** or **60-40%** - or anything. The only condition for each iteration is to keep the **train-test split percentage different.**

- To evaluate the model's performance, we take the average of all the test errors.

Time series (rolling cross-validation)

- Time series is a type of data collected at **different points in time.**
- The **dataset** is split into **train** and **test sets** according to **time.**
- **Time series cross-validation** is also called as **forward chaining** method or **rolling cross-validation.**