# Cross Validation Techniques

Tushar B. Kute,
http://tusharkute.com

# Cross Validation

- Cross-validation is a statistical method used to estimate the performance (or accuracy) of predictive statistical models.
- It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.
- In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

# Cross Validation

- When dealing with a predictive modeling task, you have to properly identify the problem so that you can pick the most suitable algorithm which can give you the best score. But how do we compare the models?
- Say, you have trained the model with the dataset available and now you want to know how well the model can perform.
- One approach can be that you are going to test the model on the dataset you have trained it on, but this may not be a good practice.

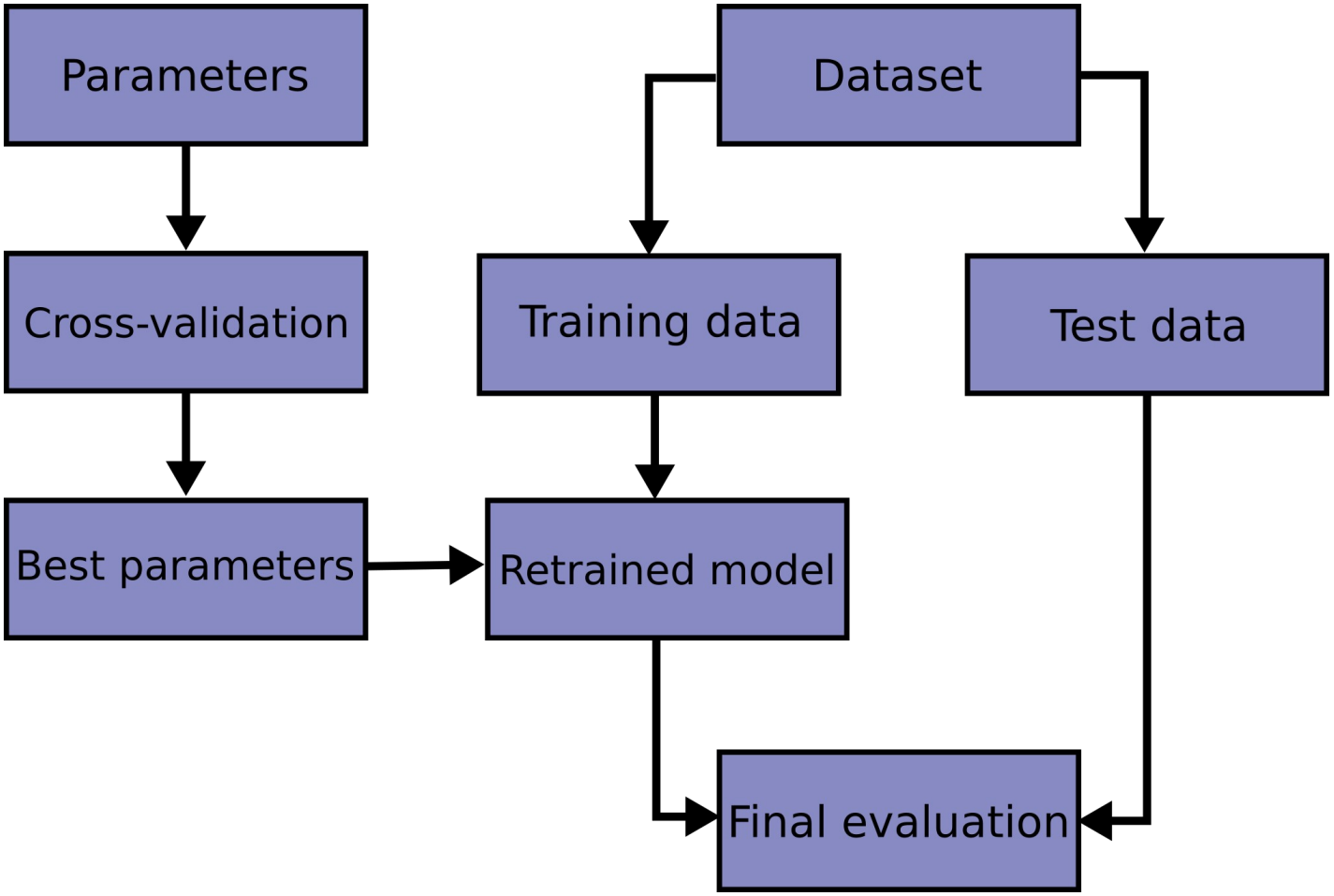
# Cross Validation

- So what is wrong with testing the model on the training dataset?
- If we do so, we assume that the training data represents all the possible scenarios of real-world and this will surely never be the case.
- Our main objective is that the model should be able to work well on the real-world data, although the training dataset is also real-world data, it represents a small set of all the possible data points(examples) out there.

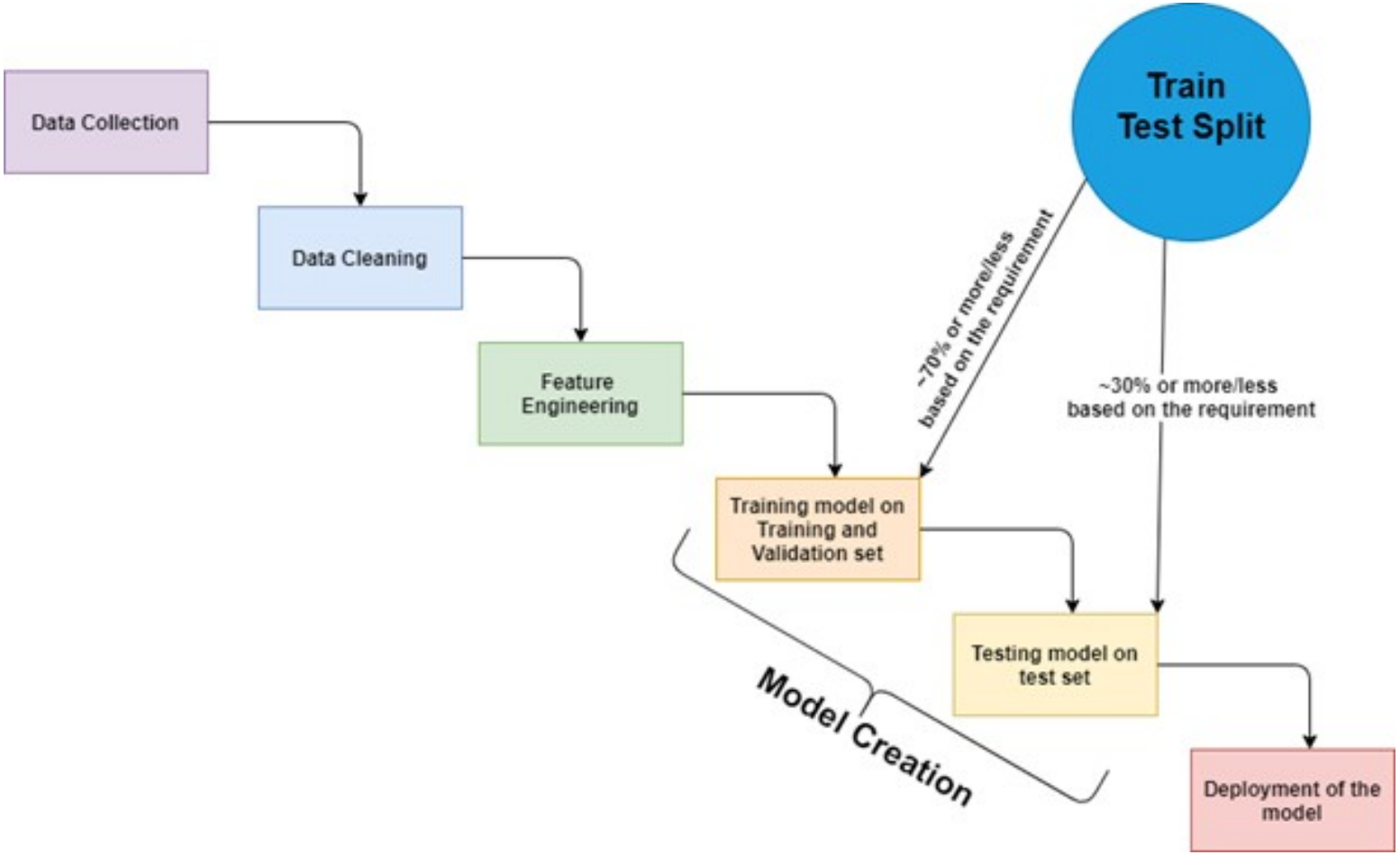
# Cross Validation

- To know the real score of the model, it should be tested on the data that it has never seen before and this set of data is usually called testing set.
- But if we split our data into training data and testing data, aren’t we going to lose some important information that the test dataset may hold?

# Cross Validation



# Cross Validation – Where?



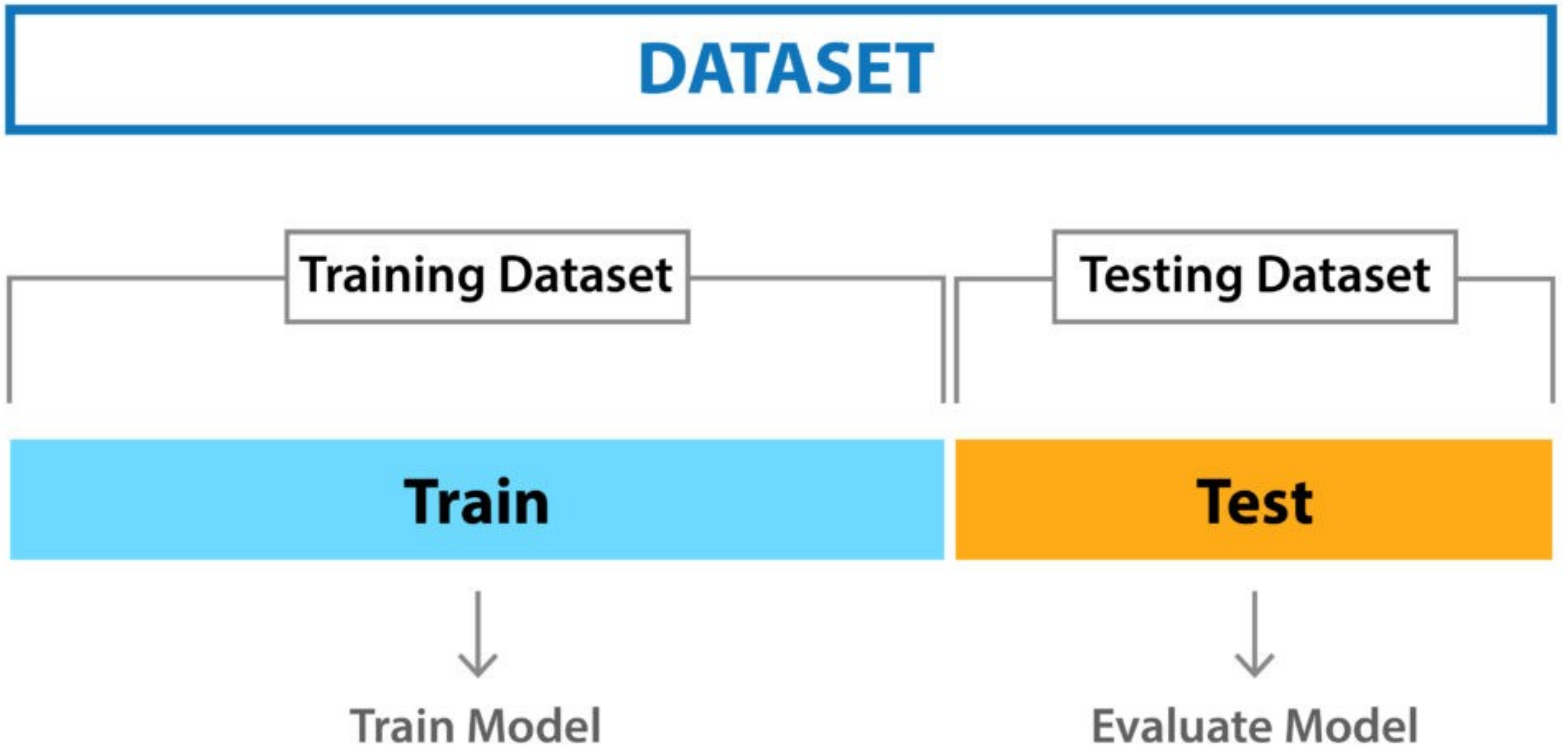
# Cross Validation – Types

- There are different types of cross validation methods, and they could be classified into two broad categories –
  - Non-exhaustive
    - Holdout
    - K-Fold
    - Stratified K-fold
  - Exhaustive Methods.
    - Leave-P-Out cross validation
    - Leave-1-Out cross validation

# Holdout Method

- This is a quite basic and simple approach in which we divide our entire dataset into two parts viz- training data and testing data.
- As the name, we train the model on training data and then evaluate on the testing set.
- Usually, the size of training data is set more than twice that of testing data, so the data is split in the ratio of 70:30, 75:25 or 80:20.

# Holdout Method



# Holdout Method

- In this approach, the data is first shuffled randomly before splitting.
- As the model is trained on a different combination of data points, the model can give different results every time we train it, and this can be a cause of instability.
- Also, we can never assure that the train set we picked is representative of the whole dataset.

# Holdout Method

- Also when our dataset is not too large, there is a high possibility that the testing data may contain some important information that we lose as we do not train the model on the testing set.
- The hold-out method is good to use when you have a very large dataset, you're on a time crunch, or you are starting to build an initial model in your data science project.

# Holdout Method

| | |
|---|---|
| 16 | A |
| 13 | A |
| 13 | A |
| 12 | A |
| 15 | B |
| 12 | B |
| 11 | B |
| 10 | B |
| 10 | B |
| 10 | B |
| 8 | B |
| 10 | C |
| 9 | C |
| 8 | C |
| 6 | C |
| 7 | D |
| 6 | D |
| 6 | D |
| 6 | D |
| 5 | D |

| | |
|---|---|
| 16 | A |
| 13 | A |
| 12 | A |
| 15 | B |
| 11 | B |
| 10 | B |
| 10 | B |
| 10 | B |
| 8 | B |
| 10 | C |
| 8 | C |
| 6 | C |
| 7 | D |
| 6 | D |
| 6 | D |
| 5 | D |

**Train Set**

| | |
|---|---|
| 13 | A |
| 12 | B |
| 9 | C |
| 6 | D |

**Test Set**

# Thank you



@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

**Web Resources**
https://mitu.co.in
http://tusharkute.com

@mituskillologies

**contact@mitu.co.in**

**tushar@tusharkute.com**