Model Performance

* Feature Selection
* Cross Validation
* Dimensionality Reduction
* Feature Selection
  + All all features important?
  + What:
    - Selecting the best feature to predict the dependent variable.
  + Why?
    - Improves the model accuracy
    - Reduce the training time
    - Reduce computational cost
    - Solve the problem of overfitting / underfitting
  + How?
    - Univariate Selection
      * Analyze each feature and get top 10
      * Sklearn.feature\_selection 🡪 selectKbest
    - Feature Importance
      * We analyze the importance of feature
      * We use ExtratreeClassifer
      * We plot the top 10 features
    - Correlation Matrix (Heatmap) (-1 to +1)
      * Positive
      * Negative
* Cross Validation
  + We splitted data into train and test (70 & 30)
  + We make sure that the model the properly trained
  + 2 Class – 0 and 1
  + How?
    - Holdout validation Approach – train\_test\_split
    - Basic K-Fold Cross Validation
      * We have 100 samples (eg – K=5)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| FOLD1 | FOLD2 | FOLD3 | FOLD4 | FOLD5 |
| FOLD1 | FOLD2 | FOLD3 | FOLD4 | FOLD5 |
| FOLD1 | FOLD2 | FOLD3 | FOLD4 | FOLD5 |
| FOLD1 | FOLD2 | FOLD3 | FOLD4 | FOLD5 |
| FOLD1 | FOLD2 | FOLD3 | FOLD4 | FOLD5 |

* Cons – In case of classification (Imbalanced Dataset)
* Stratified Cross Validation
  + It makes sure that it takes all class in a good proportion
* Leave One Out Cross Validation (LOOCV)
  + It takes a longer time
* Repeated Random Test-Train Split – ShuffleSplit
  + It a hybrid of train\_test\_split and K-Fold
* Dimensionality Reduction – A technique for unsupervised learning
  + What?
    - Dimensions – No. of features/inputs/independent variables
    - Reducing the no. of features
  + Why?
    - Training time decreases
    - Model Accuracy will improve
    - Get better insights(visualizations)
  + How?
    - PCA – Principal Component Analysis
      * We convert multiple features in small chunks (they will linearly uncorrelated features) by maintaining the same variation
      * These chunks are called components