MODULE 07

DATA SCIENCE BOOTCAMP

Models Evaluation and Hyperparameter Tuning

Vertical
Institute

Reminder

1 Please turn on Zoom camera the whole duration of classes.

2 At the start of all classes, please rename yourselves to: Name + Last 3 digits and letter of your NRIC. Example: John Tan (123A)

Agenda

- Review of Past Lessons
- Model Evaluation Metrics
- Grid Search
- Pipelines
- Cross Validation

📸

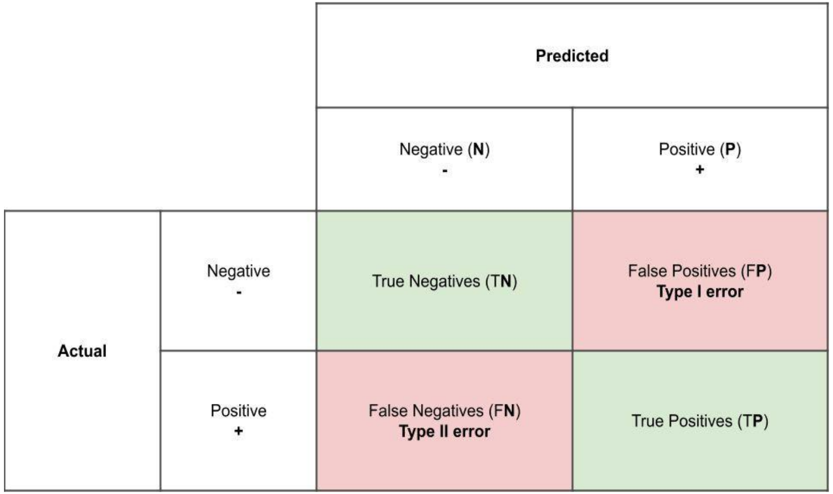
Attendance Photo Taking

MODULE 07: MODELS EVALUATION AND HYPERPARAMETER TUNING

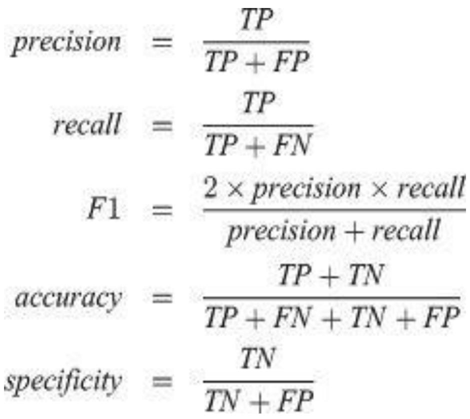
Model Evaluation

Model Evaluation - Confusion Matrix

A confusion matrix is **a summary of prediction results on a classification problem.** The number of correct and incorrect predictions are summarized with count values and broken down by each class.



		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (**TN**)	False Positives (**FP**) **Type I error**
	Positive +	False Negatives (**FN**) **Type II error**	True Positives (**TP**)



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

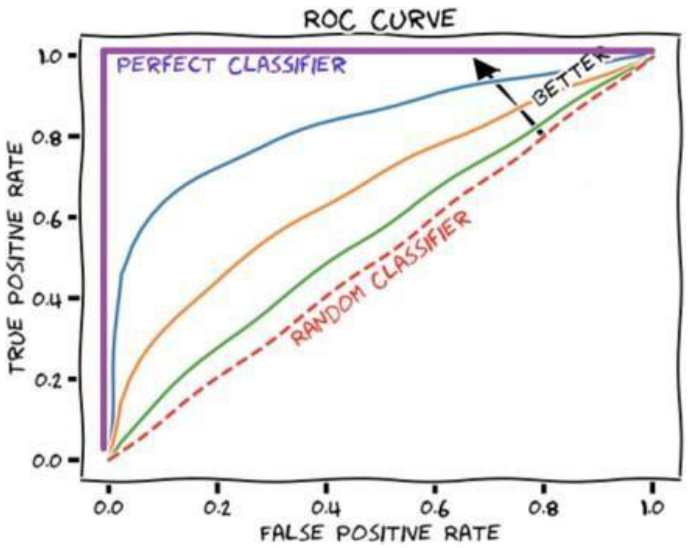
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

Performance Metrics (ROC Curve)

The Receiver Operating characteristic (ROC) curve is explicitly used for binary classification. Illustrating the **true positive rate** against the **false positive rate** of our classifier.



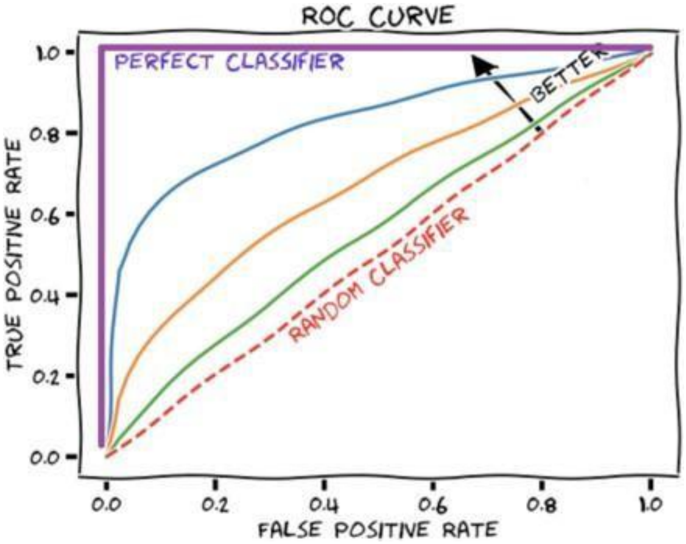
Performance Metrics (ROC Curve)

True positive rate is another name for recall which is ratio of the true positive predictions compared to all values that are actually positive.

TPR = TP/(TP + FN)

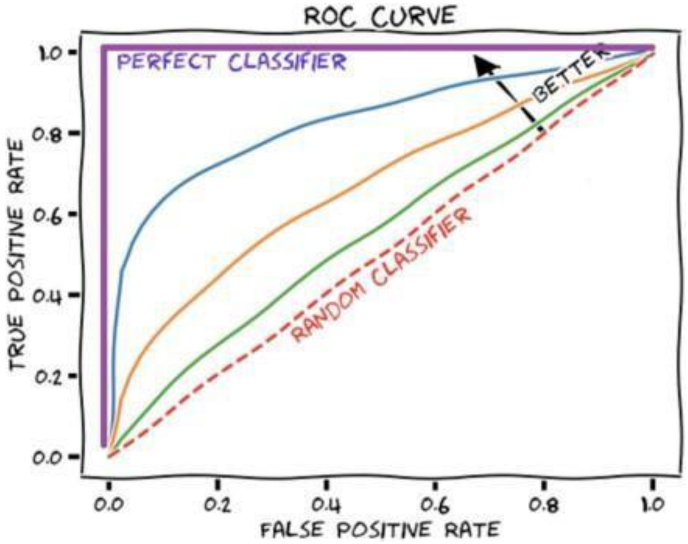
False positive rates is the ratio of false positive predictions compared to all values that are actually negative.

FPR = FP/(FP + TN)



Performance Metrics (ROC Curve)

If TPR is our y-axis and FPR is our x-axis, we want our ROC curve to hug the left side of our chart, meaning higher the TPR the lower the FPR.



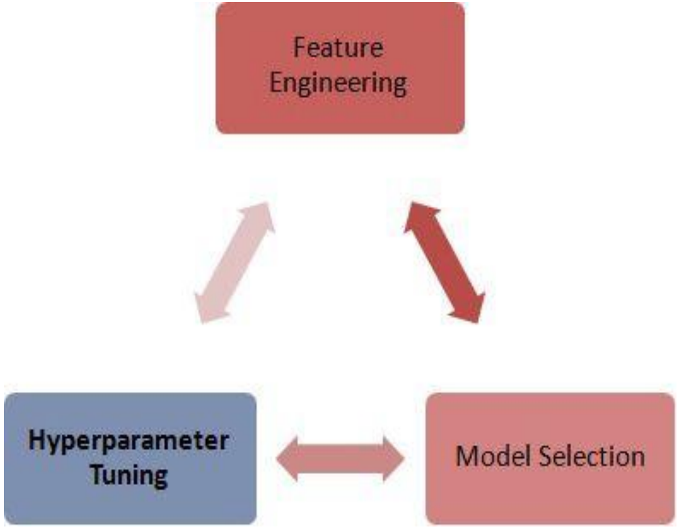
MODULE 7: MODELS EVALUATION AND HYPERPARAMETER TUNING

Grid Search

Problem Statement

For simplicity sake, we will divide the analytical aspects of a data science problem into 3 parts:

1. Gathering the required data and engineering the features.
2. Choosing the right machine learning model.
3. Finding the optimal hyperparameters.



Problem Statement

Choosing the right hyperparameters is an art.

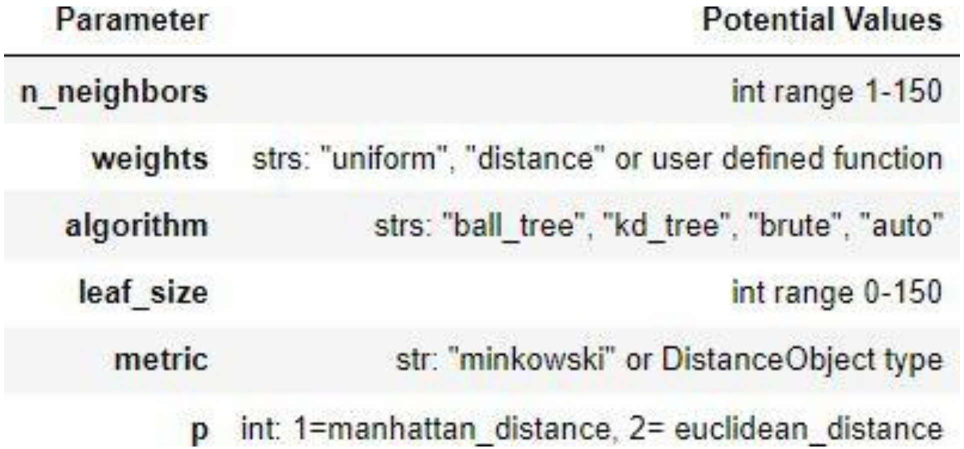
It is a tedious task and requires a lot of time and effort!

That’s where GridSearch will save us time, effort and resources to find the optimal hyperparameter.

What is a Hyperparameter?

A machine learning model has multiple parameters that are not trained by the training set.

These parameters control the accuracy of the model. Therefore, the hyperparameters are particularly important in a data science project.

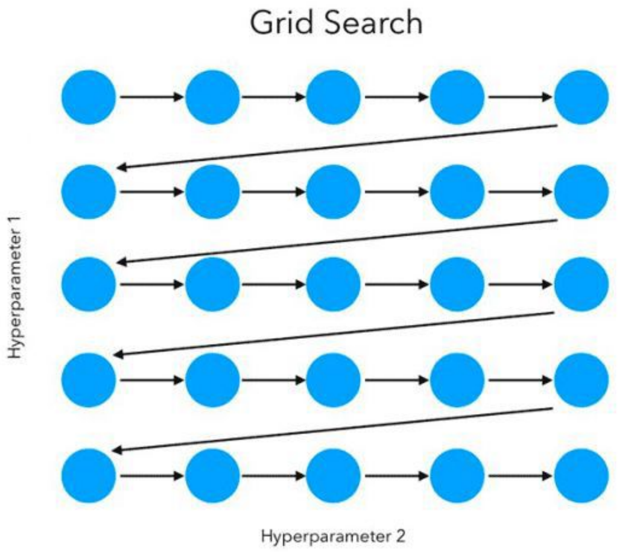


Parameter	Potential Values
n_neighbors	int range 1-150
weights	strs: "uniform", "distance" or user defined function
algorithm	strs: "ball_tree", "kd_tree", "brute", "auto"
leaf_size	int range 0-150
metric	str: "minkowski" or DistanceObject type
p	int: 1=manhattan_distance, 2= euclidean_distance

GridSearch

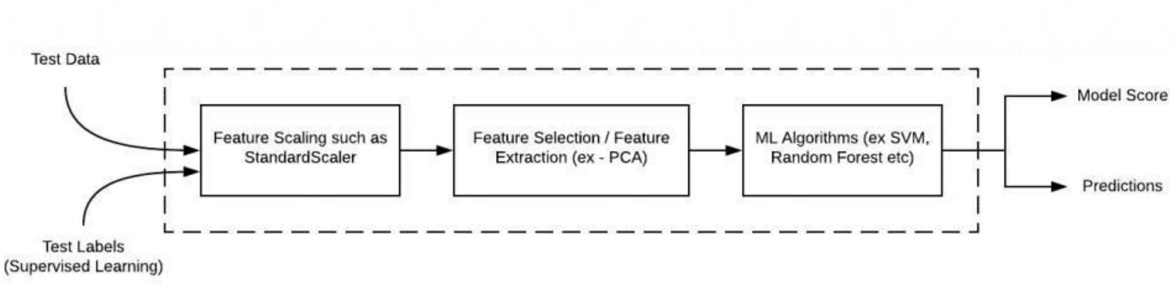
GridSearch is a tuning technique that attempts to compute the optimum values of hyperparameters. It is an exhaustive search that is performed on a specific parameter values of a model.

The model is also known as an estimator.

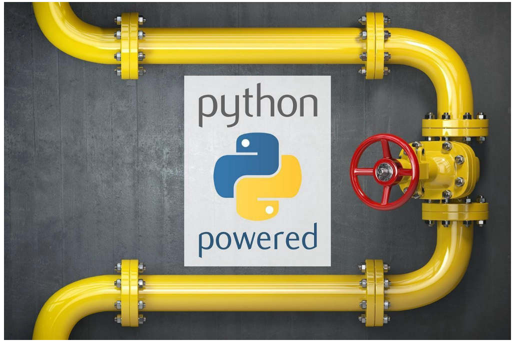


Automating ML Workflows with Pipelines

Machine Learning (ML) pipeline, theoretically, represents different steps including data transformation and prediction through which data passes. The outcome of the pipeline is the trained model which can be used for making the predictions.


Test Data
Feature Scaling such as StandardScaler
Feature Selection / Feature Extraction (ex - PCA)
ML Algorithms (ex SVM, Random Forest etc)
Model Score
Predictions
Test Labels (Supervised Learning)

Machine Learning Pipeline



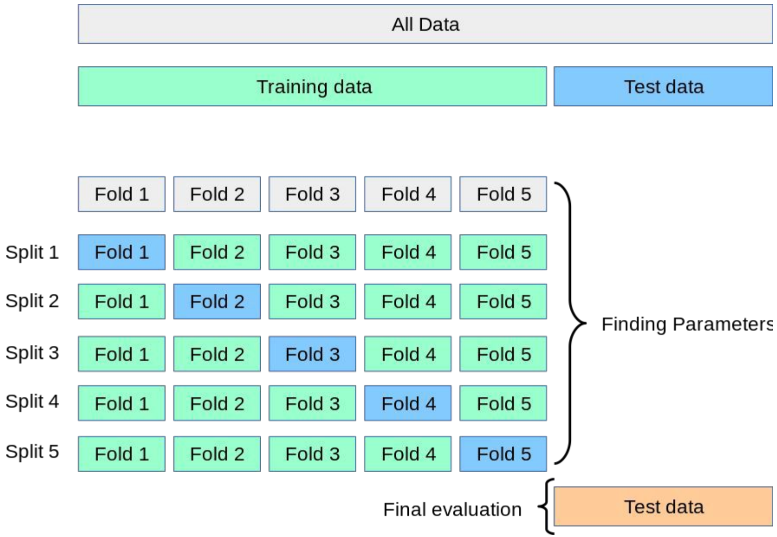
MODULE 7: MODELS EVALUATION AND HYPERPARAMETER TUNING

Cross Validation

Cross-Validation

Cross-validation is a statistical method used to estimate the skill of machine learning models.

It is a resampling procedure used to evaluate machine learning models on a limited data sample.



All Data

Training data | Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

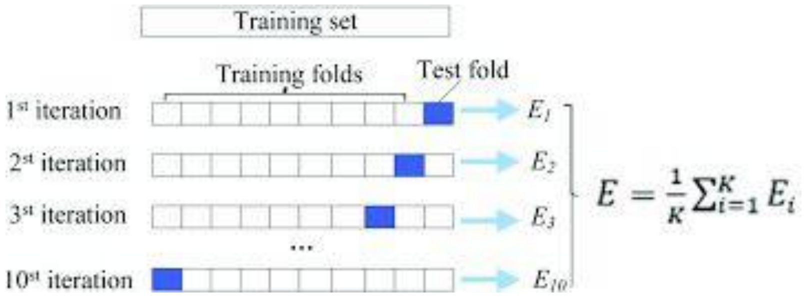
Finding Parameters

Final evaluation | Test data

Cross-Validation

The general procedure is as follows:

- Shuffle the dataset randomly.
- Split the dataset into k groups.
- For each unique group:
  1. Take the group as a hold out or test dataset.
  2. Take the remaining groups as a training dataset.
  3. Fit a model on the training set and evaluate it on the test set.
  4. Retain the evaluation score and discard the model.
- Summarize the skill of the model using the sample of model evaluation scores.



Some things to take note…

Link and resource could be accessed in the Learning Portal.

https://elearn.verticalinstitute.com/users/sign_in

MODULE 07

DATA SCIENCE BOOTCAMP

Thank you!

Vertical Institute