RANDOM FOREST

Individual decision tree algorithms can be prone to problems, such as bias and over-fitting. A more practical approach is to construct multiple decision trees and combine the results into a single output. This approach is termed **ensemble methods**. The most well-known ensemble method is **bagging** (also known as **bootstrap aggregation**, **bagging=bootstrap+aggregation**). This method was introduced in 1996 by Leo Breiman. In this method, data points in the training set are sampled with replacement (producing a **bootstrap sample**), one-third of which, known as the **out-of-bag (OOB) sample**, is set aside for cross-validation, and the remaining two-thirds of the sample is used to build a decision tree. Decision trees are generated for each bootstrap sample independently from others. The results are then aggregated depending on the type of trees used. If regression trees are trained, the average of predicted values are computed. If classification trees are fitted, the majority of the predictions define the final predicted class. The ensemble method reduces variance and, as the result, yields more accurate predictions than individual decision trees.

The **random forest algorithm** is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. **Feature randomness** (also known as **feature bagging** or the **random subspace method**) generates a random subset of variables (also called **features**), which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible variable splits, random forests only select a subset of those variables.

Random forest algorithms have three main hyper-parameters that need to be set before training. These are node size, the number of trees, and the number of variables sampled.

Variable Importance

Random forest makes it easy to evaluate the **variable importance** (or the **contribution** of each splitting variable to the model). It is sometimes termed the **feature importance**. There are two commonly used ways to evaluate variable importance that are characteristically different from each other. One method is termed the **loss reduction** or **Gini increase** or **Gini importance** or **impurity reduction** or **mean decrease in impurity (MDI)**. It is used to measure how much the model's accuracy decreases when a given variable is excluded. For a random forest with regression trees, the loss functions are the **mean squared error** $MSE = RSS/n$, and the **absolute error** $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. The second feature importance method is the **permutation importance** (or the **mean decrease accuracy (MDA)**), which identifies the average decrease in accuracy by randomly permuting the variable values in OOB samples.