2019

Prediction on: Adult data set

SUBMITED BY:

DEBASISH BHOL

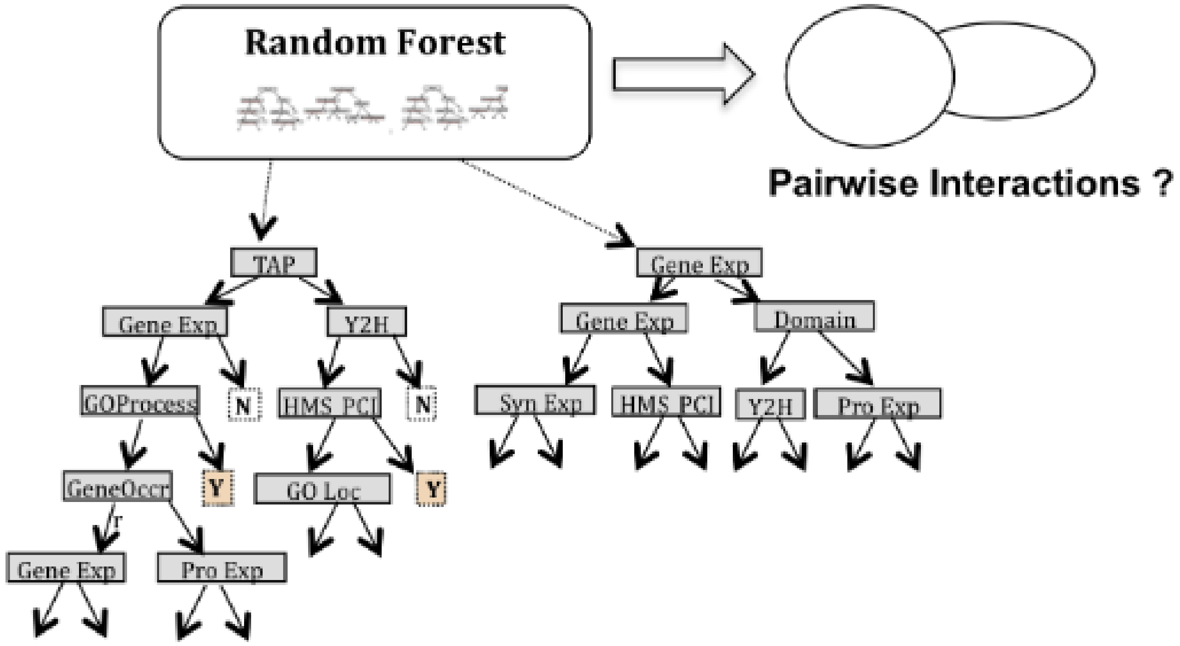
LOVELY PROFESSIONAL UNIVERSITY

Introduction:

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results - in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

For its practical demonstration - specifically in a classification context - I'll be walking through an example using a famous data set from (UCI) Machine Learning Repository. The data set, called the Adult Data Set, deals with binary classification and includes features computed from digitized images. In this data set features present are age, workclass, fnlwgt, education, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss, hours_per_week, country, salary. Here the target attribute is the salary.

In this I have implemented the Random forest classification algorithm to predict the salary of an individual using the other attributes present in the algorithm. I have used three estimators to produce the output. Here, three estimator mean that the random forest will create three decision tree and it will predict the output by voting the individual output of the tree.



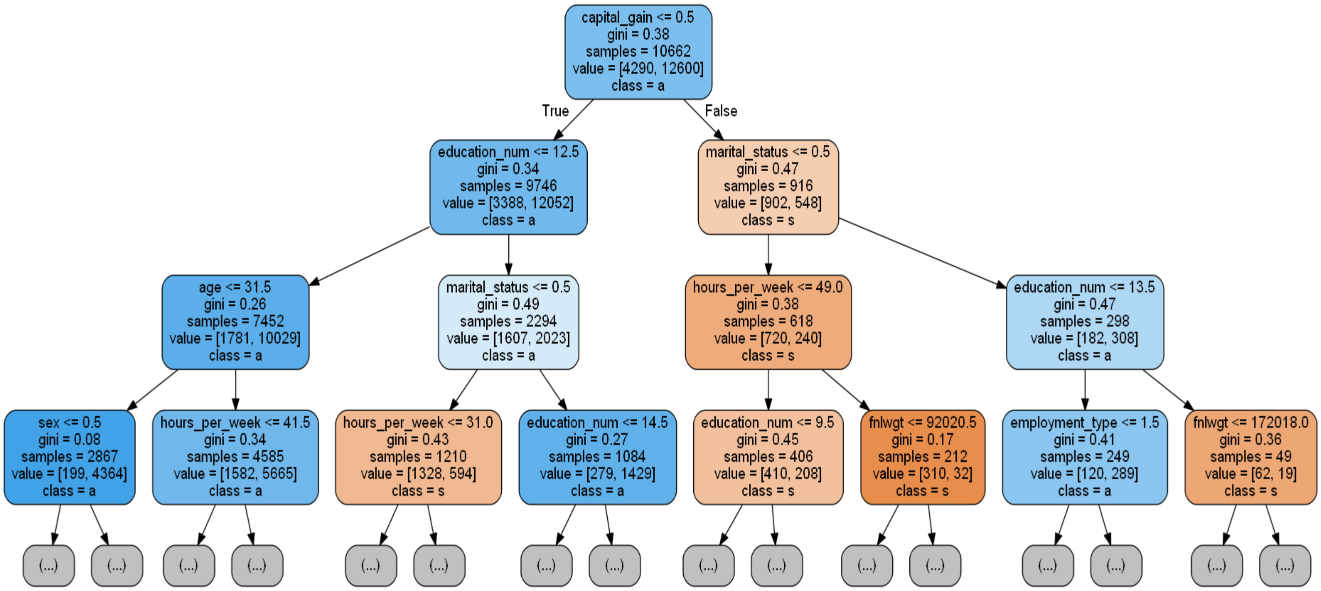
Decision tree:

Decision tree are the powerful and popular tools for classification and prediction. Decision tree represent rules, which are human understandable. A decision tree is a hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree consist of internal decision nodes and terminal leaves. Each decision nodes implements a test function FM(x) with discrete outcomes labelling the branches. Given an input to each node, a test is applied and one of the branches is taken depending on the outcome. This process start at the root node and ends at the leaf node.

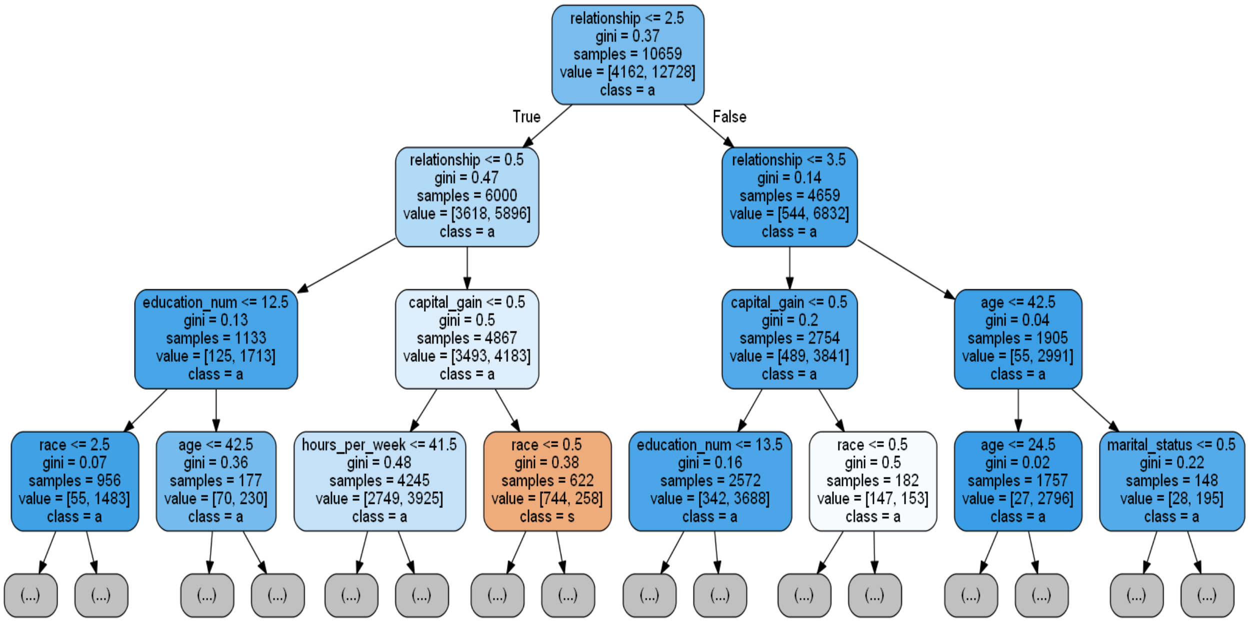
A decision tree is also a nonparametric model in the sense that we do not assume any parametric from the class densities and the tree structure is not fixed a priori but the tree grows, branches and leaves are added, learning depends on the complexity of the problem inherent in the data. Decision tree is classified in the form of a tree structure which consist of:

- **Decision node**: Decision nodes are specific to a test on a single attribute.
- **Leaf node**: It indicates the value of the target attribute.
- **Edge**: Edge splits of one attribute.
- **Path**: A path is a disjunction of test to make the final decision.

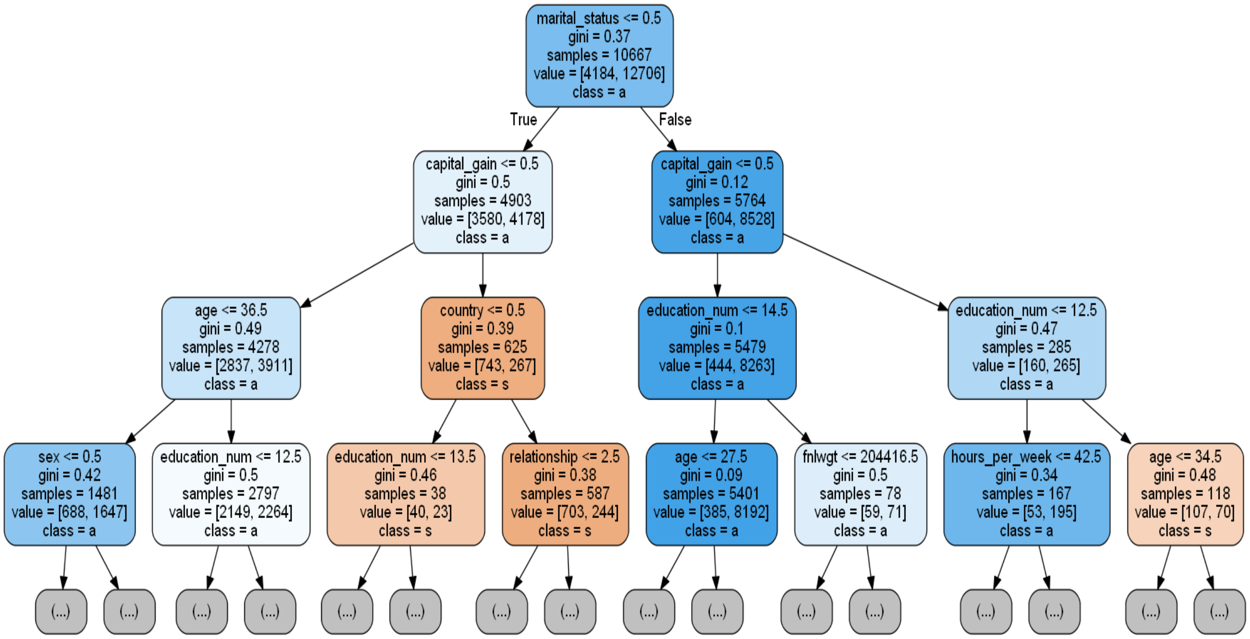
Estimator 1:



Estimator 2:

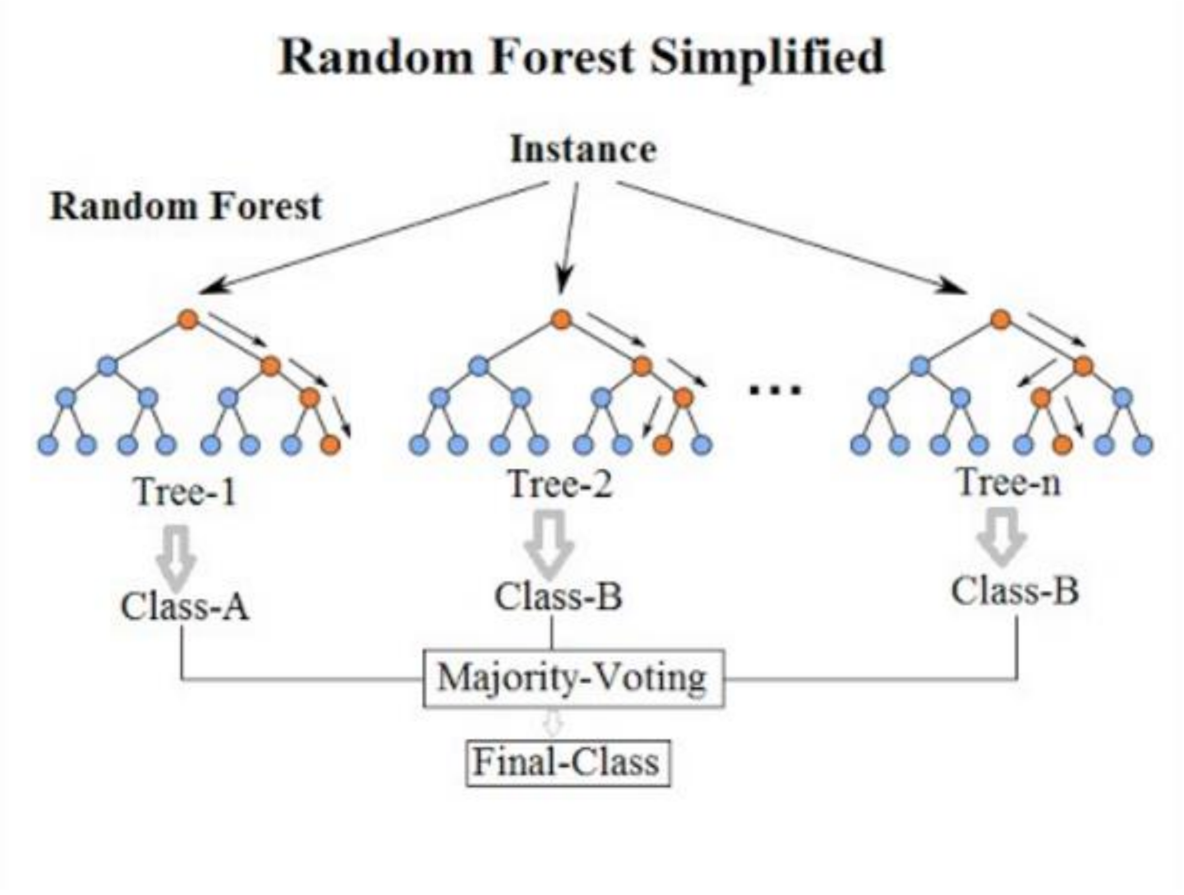


Estimator 3:



Random Forest Classification:

Random forest classification is one of the best algorithm for regression and classification. The random forest is assemble learning model, composed of multiple decision tree. By averaging the output of several decision tree random forest tend to improve its prediction. Random forest overcome the demerits of decision tree, i.e. in decision tree data may be over fitted, but this will not happen in random forest, because in random forest we generate many decision tree and vote for the output. The random forest choose the classification having the highest vote and in regression it take the average of the all output produce by different tree.



The methodology includes construction of decision trees of the given training data and matching the test data with these. Random forests are used to rank the importance of variables in a classification problem. To measure the importance of a variable in a data setDn = {(Xi, Yi)} = 1 to n we fit a random forest to the data. During the fitting process the error for each data point is calculated and averaged over the forest. To measure the importance of the i-th feature after training, the values of the i-th feature are permuted among the training data and the error is again computed on this data set. The importance score for the i-th feature is computed by averaging the difference in error before and after the permutation for all the trees. Normalization of the score is done by the standard deviation of these differences. Features which produce large values for this score are more important than features which produce small values. Random forests provide information about the importance of a variable and also the proximity of the data points with one another.

Training:

In order to discover the optimal split we must iterate over the predictors and their possible values. For each possible split we will calculate the target variable average of the resulting sub-groups. Using those means as predictions for instances of each sub-sample, the next step is to calculate the Mean Squared Error (MSE). Comparing MSE is not enough to determine the best split because it could happen that a sub-group contains just a single sample (and the other sub-group all the remaining instances), something not useful at all. To avoid that situation we take the **weighted average** (MSE * number of samples in the sub-group) as the evaluation metric. This process is repeated until we reach some limit: minimum number of instances in each node, tree max depth or nodes unable to be separated (containing a single training sample).

Advantages of random forest:

1. It provides accurate predictions for many types of applications.
2. It can measure the importance of each feature with respect to the training data set.
3. Pairwise proximity between samples can be measured by the training data set.

Applications of random forest:

1. Is used for image classification for pixel analysis.
2. Is used in the field of Bioinformatics for complex data analysis.
3. Is used for video segmentation (high dimensional data).

Random forest classification up to 5th level: