

Random Forest Appendix

Catalin Starica* and Jan Marton†

As some readers may not be familiar with the RF algorithm, we provide an intuitive explanation of the method. RF starts by building a large number of decision trees, each using a random subset of data and a random subset of attributes (the factors that might influence prices, see next subsection). Once all trees are built, they each propose a shape of the mapping of earnings into prices based on their subset of data and attributes. The random forest algorithm then combines (averages) the inferred relationships of all the trees to propose a final shape of the mapping. This sequence of steps (construction of many trees followed by aggregation of tree-specific inferred shapes) prevents over-fitting, that is, it ensures that the model does not become too tailored to the specifics of the sample.¹

Decision trees divide the space of independent variables into regions by repeatedly splitting the range of each variable (earnings and attributes) to best fit the dependent variable (price). These splits occur one variable at a time and are binary: the values above the chosen split point belong to one region, while the values below belong to another.² Each region identifies a group of peers (firms with attribute values within the region) and a range of earnings for which these peers share a similar relationship of earnings with prices. Within each region, a simple model, such as a constant, is estimated for the range of earnings it covers. This process results in a step function approximation of the complex relationship under study.

The functioning of decision trees also provides another intuitive interpretation of the regression function in equation

$$Mktv_i = g(NI_i, \mathbf{Attributes}_i) + \varepsilon_i, \quad (1)$$

familiar to researchers accustomed to linear regression. Nonlinear regression of firm prices in the cross section t_0 on the corresponding vectors of earnings and attributes, $(NI, \mathbf{Attributes})$, means estimating the relationship between earnings and prices while holding constant the levels

*School of Economics and Business, University of Neuchâtel, Switzerland.

†School of Business, Economics and Law, University of Gothenburg, Sweden.

¹By not fitting the noise in the sample, the algorithm yields consistent estimates of the relation of interest.

²The term "decision" in "decision tree" refers to where to split the range of each regressor variable.

of attributes that determine the shape of the earnings-price relationship. While linear regression performs this estimate globally, RF does it locally.

Decision Trees

Figure 1 shows—in a simplified example with two explanatory (independent) variables (X_1, X_2) ³—how recursive splitting in decision trees works. Let us assume that X_1 is earnings, while X_2 is a proxy for attributes, e.g., P/B. The left panel of Figure 1 shows that the first split occurs at the value t_1 of earnings, e.g., $t_1 = 0$; that is, the earnings–price relation for firms with negative earnings is fundamentally different from that of firms for which earnings are greater than 0. We now have two regions, where we approximate the response Y , the price, by a constant with a different value in each region.

To improve the fit, the decision tree continues to subdivide the earnings–attribute space into smaller regions. For example, in the region defined by earnings smaller than 0 ($X_1 \leq t_1$),⁴ the next split occurs for the attribute $X_2 = P/B$ at value t_2 , which could distinguish between value and growth firms. Consequently, the earnings-to-price mapping for value firms (P/B less than t_2) with earnings smaller than 0 is fundamentally different from that of firms with negative earnings and attribute value P/B greater than t_2 (growth firms). We now have three regions, where we approximate the response Y (the price) by three different constants, one in each region. Next, the positive earnings range is split, for example, into moderate earnings ($X_1 < t_3$) and high earnings ($X_1 > t_3$). Finally, the relation of high earnings to prices varies systematically with the type of firm (split along the P/B dimension at t_4).

The left panel of Figure 1 shows four splits along the two dimensions, i.e., earnings and P/B, in a tree formation (which explains the name of the method). It emphasizes that a tree is a fully nonparametric structure constructed with a logic that departs significantly from that of traditional parametric regressions. A tree grows sequentially, with a new branch added at each step. Each new branch separates the remaining data from the previous step by further splitting the range of one of the independent variables. As a result, the earnings–attributes space is

³Decision trees can be applied to regressions with a very large number of explanatory variables. Limiting the number to two enables us to graphically describe the method in three dimensions.

⁴Note that t_1 , t_2 , t_3 , and t_4 are values taken by the variables X_1 or X_2 . They should not be mistaken for time points, that is, cross sections.

partitioned into rectangles.

In our example, the splits define five regions or rectangles ($R_1, R_2, \dots, R_5$) of the two-dimensional space of earnings and P/B, (X_1, X_2) , represented in the center panel. Within each region, the response variable Y takes a different price, obtained by averaging the prices of all firms with earnings and P/B belonging to the rectangle, as vertically shown in the right panel of Figure 1. For example, Y takes the lowest value in R_2 (for growth firms with negative earnings) and the highest value in R_5 (corresponding to growth firms with high earnings).

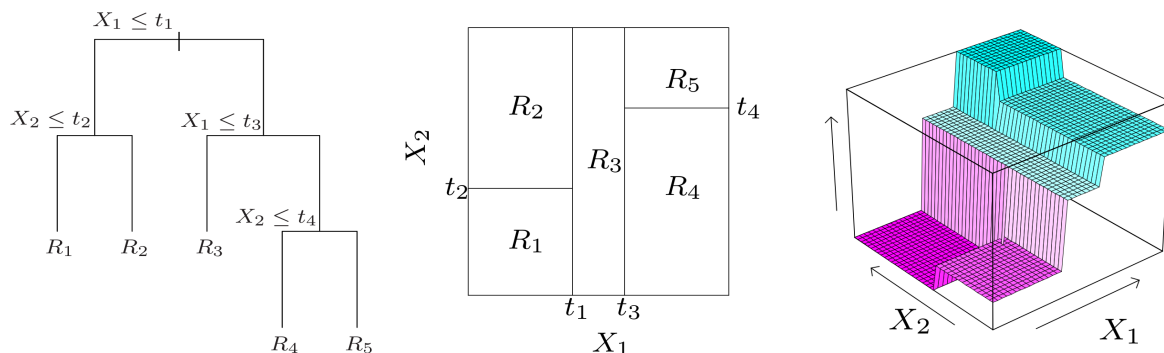


Figure 1: **Tree-based nonparametric estimation.** The left and center panels show the partition of a two-dimensional explanatory variable space by recursive binary splitting, as used in decision trees. The right panel shows a plot of the estimated regression function, which is locally constant on each of the five regions $R_1, R_2, \dots, R_5$ of the explanatory variable space (reproduced from Hastie et al. (2009)).

The right panel indicates that there is a nonlinear positive relationship between earnings and prices that is steeper in the middle range than at the extreme low and high values of earnings. In addition, the relationship is a function of the value of the attribute X_2 , the price-to-book ratio. Value firms have a relatively flatter mapping of earnings into prices (lower growth) than growth firms, which exhibit a relatively steeper mapping (higher growth). Figure 1 provides an intuitive picture of how decision trees allow for a local estimation of the earnings–price relationship for similar firms and model the interactions between different independent variables. In the actual study, the space is multidimensional with several variables that proxy for attributes. Additionally, the large number of observations makes the multidimensional surface (cf. the right panel in Figure 1) approximately continuous.

References

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer.