**Reducing Levels in Categorical Variables**

We often come across scenarios in which a categorical variable has many attribute levels, like the branches of a bank, postal codes, or products listed on an e-commerce website. Handling many attributes or levels might become cumbersome, and looking at the frequency distribution might reveal that a subset of such levels accounts for about 90 percent of the observations. Building a predictive model without treating the levels most likely will lead to a less robust model, and the computational efficiency also gets impacted negatively.

For example, a decision tree or random forest will tend to give more importance to the categorical variable with many levels, even though it may not deserve it. We can treat such categorical variables with predictive-modeling techniques, domain expertise, or even a simple frequency-distribution approach.