Decision Trees— What Are They?

Introduction

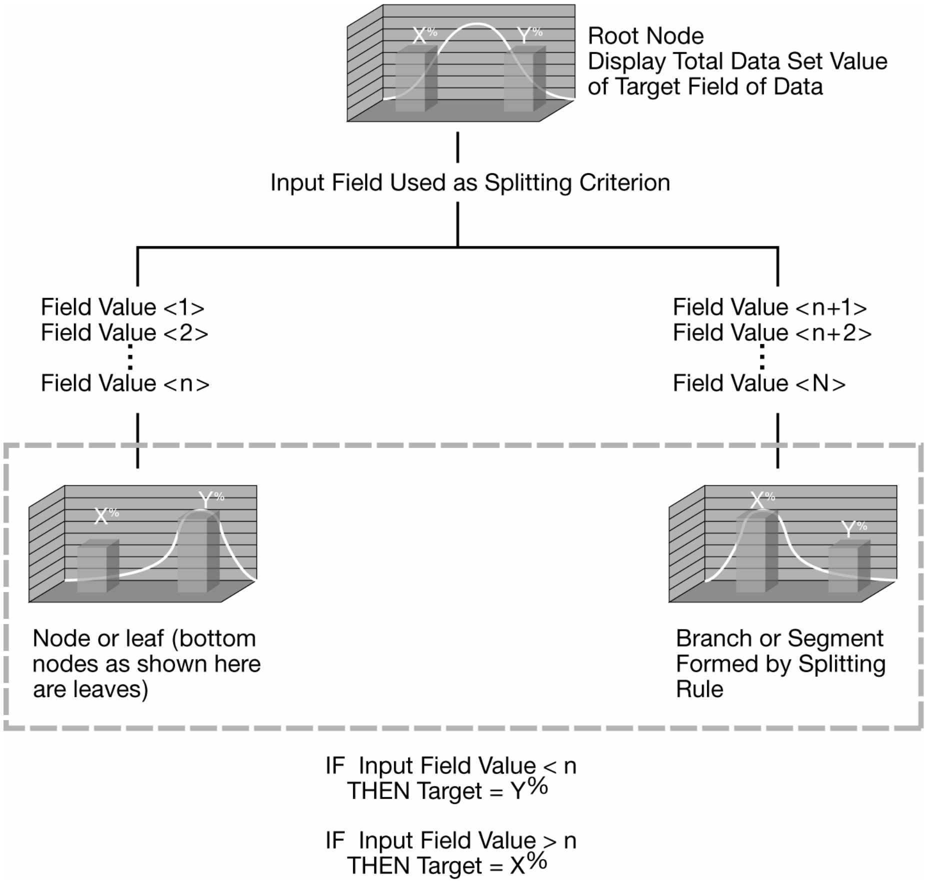
Decision trees are a simple, but powerful form of multiple variable analysis. They provide unique capabilities to supplement, complement, and substitute for

- traditional statistical forms of analysis (such as multiple linear regression)
- a variety of data mining tools and techniques (such as neural networks)
- recently developed multidimensional forms of reporting and analysis found in the field of business intelligence

Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field. A sample decision tree is illustrated in Figure 1.1, which shows that the decision tree can reflect both a continuous and categorical object of analysis. The display of this node reflects all the data set records, fields, and field values that are found in the object of analysis. The discovery of the decision rule to form the branches or segments underneath the root node is based on a method that extracts the relationship between the object of analysis (that serves as the target field in the data) and one or more fields that serve as input fields to create the branches or segments. The values in the input field are used to estimate the likely value in the target field. The target field is also called an outcome, response, or dependent field or variable.

The general form of this modeling approach is illustrated in Figure 1.1. Once the relationship is extracted, then one or more decision rules can be derived that describe the relationships between inputs and targets. Rules can be selected and used to display the decision tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values. Decision rules can predict the values of new or unseen observations that contain values for the inputs, but might not contain values for the targets.

Figure 1.1: Illustration of the Decision Tree

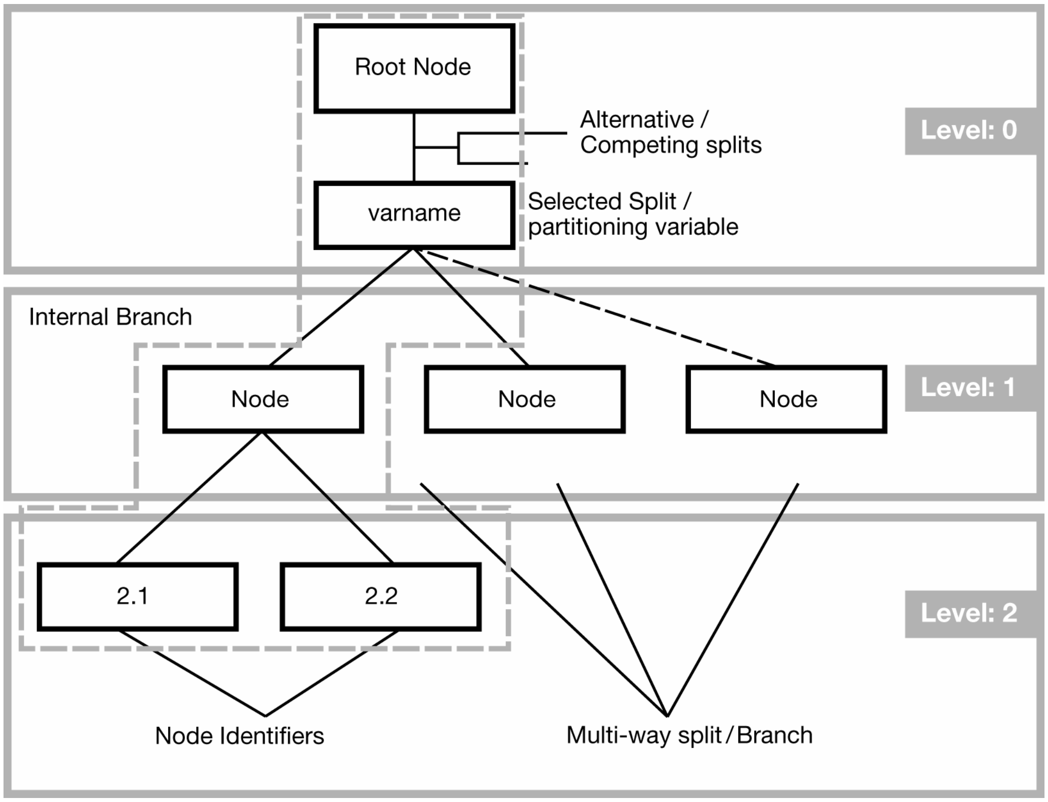


Each rule assigns a record or observation from the data set to a node in a branch or segment based on the value of one of the fields or columns in the data set.[1] Fields or columns that are used to create the rule are called *inputs*. Splitting rules are applied one after another, resulting in a hierarchy of branches within branches that produces the characteristic inverted decision tree form. The nested hierarchy of branches is called a

[1] The SAS Enterprise Miner decision tree contains a variety of algorithms to handle missing values, including a unique algorithm to assign partial records to different segments when the value in the field that is being used to determine the segment is missing.

decision tree, and each segment or branch is called a *node*. A node with all its descendent segments forms an additional segment or a branch of that node. The bottom nodes of the decision tree are called *leaves* (or *terminal nodes*). For each leaf, the decision rule provides a unique path for data to enter the class that is defined as the leaf. All nodes, including the bottom leaf nodes, have mutually exclusive assignment rules; as a result, records or observations from the parent data set can be found in one node only. Once the decision rules have been determined, it is possible to use the rules to predict new node values based on new or unseen data. In predictive modeling, the decision rule yields the predicted value.

Figure 1.2: Illustration of Decision Tree Nomenclature



Although decision trees have been in development and use for over 50 years (one of the earliest uses of decision trees was in the study of television broadcasting by Belson in 1956), many new forms of decision trees are evolving that promise to provide exciting new capabilities in the areas of data mining and machine learning in the years to come. For example, one new form of the decision tree involves the creation of *random forests*. Random forests are multi-tree committees that use randomly drawn samples of data and inputs and reweighting techniques to develop multiple trees that, when combined, provide for stronger prediction and better diagnostics on the structure of the decision tree.

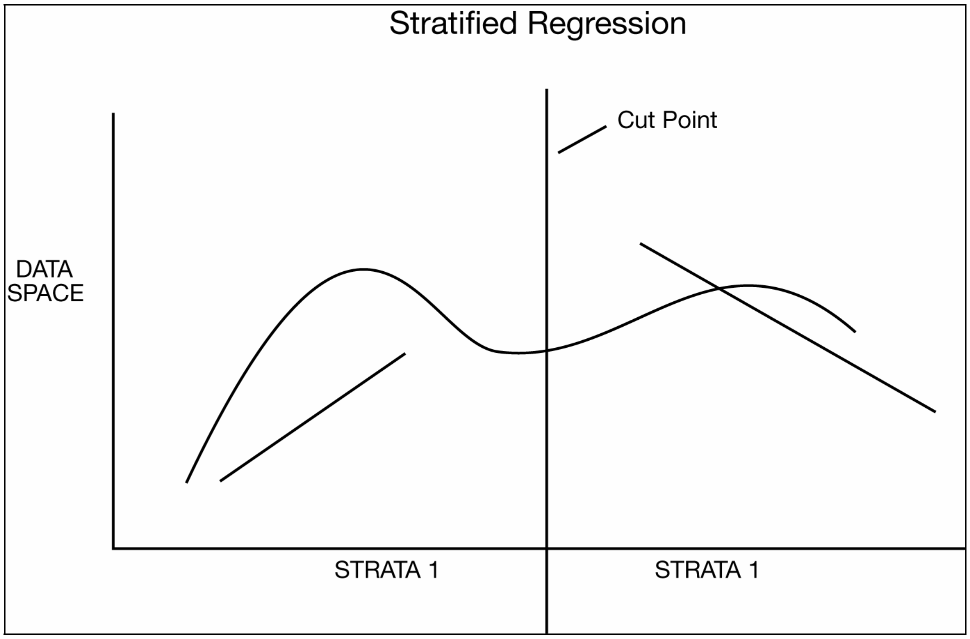
Besides modeling, decision trees can be used to explore and clarify data for dimensional cubes that can be found in business analytics and business intelligence.

Using Decision Trees with Other Modeling Approaches

Decision trees play well with other modeling approaches, such as regression, and can be used to select inputs or to create dummy variables representing interaction effects for regression equations. For example, Neville (1998) explains how to use decision trees to create stratified regression models by selecting different slices of the data population for in-depth regression modeling.

The essential idea in stratified regression is to recognize that the relationships in the data are not readily fitted for a constant, linear regression equation. As illustrated in Figure 1.3, a boundary in the data could suggest a partitioning so that different regression models of different forms can be more readily fitted in the strata that are formed by establishing this boundary. As Neville (1998) states, decision trees are well suited in identifying regression strata.

Figure 1.3: Illustration of the Partitioning of Data Suggesting Stratified Regression Modeling

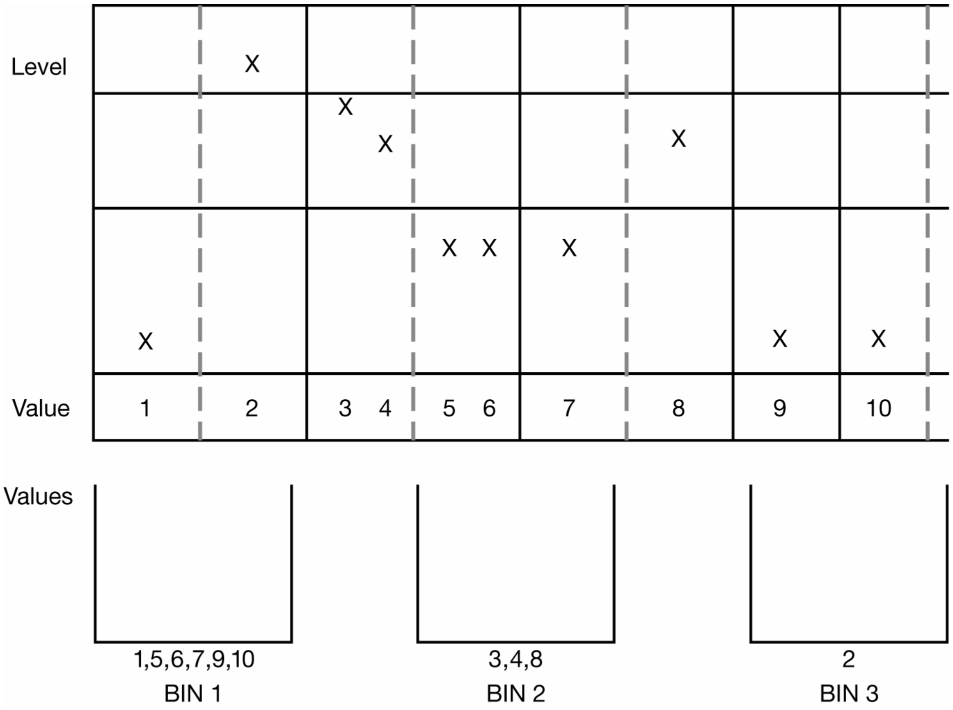


Decision trees are also useful for collapsing a set of categorical values into ranges that are aligned with the values of a selected target variable or value. This is sometimes called *optimal collapsing of values*. A typical way of collapsing categorical values together would be to join adjacent categories together. In this way 10 separate categories can be reduced to 5. In some cases, as illustrated in Figure 1.4, this results in a significant reduction in information. Here categories 1 and 2 are associated with extremely low and extremely high levels of the target value. In this example, the collapsed categories 3 and 4, 5 and 6, 7 and 8, and 9 and 10 work better in this type of deterministic collapsing framework; however, the anomalous outcome produced by collapsing categories 1 and 2 together should serve as a strong caution against adopting any such scheme on a regular basis.

Decision trees produce superior results. The dotted lines show how collapsing the categories with respect to the levels of the target yields different and better results. If we impose a monotonic restriction on the collapsing of categories—as we do when we request tree growth on the basis of ordinal predictors—then we see that category 1 becomes a group of its own. Categories 2, 3, and 4 join together and point to a relatively high level in the target. Categories 5, 6, and 7 join together to predict the lowest level of the target. And categories 8, 9, and 10 form the final group.

If a completely unordered grouping of the categorical codes is requested—as would be the case if the input was defined as "nominal"—then the 3 bins as shown in the bottom of Figure 1.4 might be produced. Here the categories 1, 5, 6, 7, 9, and 10 group together as associated with the highest level of the target. The medium target levels produce a grouping of categories 3, 4, and 8. The lone high target level that is associated with category 2 falls out as a category of its own.

Figure 1.4: Illustration of Forming Nodes by Binning Input-Target Relationships



Since a decision tree allows you to combine categories that have similar values with respect to the level of some target value there is less information loss in collapsing categories together. This leads to improved prediction and classification results. As shown in the figure, it is possible to intuitively appreciate that these collapsed categories can be used as branches in a tree. So, knowing the branch—for example, branch 3 (labeled BIN 3), we are better able to guess or predict the level of the target. In the case of branch 2 we can see that the target level lies in the mid-range, whereas in the last branch—here collapsed categories 1, 5, 6, 7, 9, 10—the target is relatively low.

Why Are Decision Trees So Useful?

Decision trees are a form of multiple variable (or multiple effect) analyses. All forms of multiple variable analyses allow us to predict, explain, describe, or classify an outcome (or target). An example of a multiple variable analysis is a probability of sale or the likelihood to respond to a marketing campaign as a result of the combined effects of multiple input variables, factors, or dimensions. This multiple variable analysis capability of decision trees enables you to go beyond simple one-cause, one-effect relationships and to discover and describe things in the context of multiple influences. Multiple variable analysis is particularly important in current problem-solving because almost all critical outcomes that determine success are based on multiple factors. Further, it is becoming increasingly clear that while it is easy to set up one-cause, one-effect relationships in the form of tables or graphs, this approach can lead to costly and misleading outcomes.

According to research in cognitive psychology (Miller 1956; Kahneman, Slovic, and Tversky 1982) the ability to conceptually grasp and manipulate multiple chunks of knowledge is limited by the physical and cognitive processing limitations of the short-term memory portion of the brain. This places a premium on the utilization of dimensional manipulation and presentation techniques that are capable of preserving and reflecting high-dimensionality relationships in a readily comprehensible form so that the relationships can be more easily consumed and applied by humans.

There are many multiple variable techniques available. The appeal of decision trees lies in their relative power, ease of use, robustness with a variety of data and levels of measurement, and ease of interpretability. Decision trees are developed and presented incrementally; thus, the combined set of multiple influences (which are necessary to fully explain the relationship of interest) is a collection of one-cause, one-effect relationships presented in the recursive form of a decision tree. This means that decision trees deal with human short-term memory limitations quite effectively and are easier to understand than more complex, multiple variable techniques. Decision trees turn raw data into an increased knowledge and awareness of business, engineering, and scientific issues, and they enable you to deploy that knowledge in a simple, but powerful set of human-readable rules.

Decision trees attempt to find a strong relationship between input values and target values in a group of observations that form a data set. When a set of input values is identified as having a strong relationship to a target value, then all of these values are grouped in a bin that becomes a branch on the decision tree. These groupings are determined by the observed form of the relationship between the bin values and the target. For example, suppose that the target average value differs sharply in the three bins that are formed by the input. As shown in Figure 1.4, binning involves taking each input, determining how the values in the input are related to the target, and, based on the input-target relationship, depositing inputs with similar values into bins that are formed by the relationship.

To visualize this process using the data in Figure 1.4, you see that BIN 1 contains values 1, 5, 6, 7, 9, and 10; BIN 2 contains values 3, 4, and 8; and BIN 3 contains value 2. The sort-selection mechanism can combine values in bins whether or not they are adjacent to one another (e.g., 3, 4, and 8 are in BIN 2, whereas 7 is in BIN 1). When only adjacent values are allowed to combine to form the branches of a decision tree, then the underlying form of measurement is assumed to monotonically increase as the numeric code of the input increases. When non-adjacent values are allowed to combine, then the underlying form of measurement is non-monotonic. A wide variety of different forms of measurement, including linear, nonlinear, and cyclic, can be modeled using decision trees.

A strong input-target relationship is formed when knowledge of the value of an input improves the ability to predict the value of the target. A strong relationship helps you understand the characteristics of the target. It is normal for this type of relationship to be useful in predicting the values of targets. For example, in most animal populations, knowing the height or weight improves the ability to predict the gender. In the following display, there are 28 observations in the data set. There are 20 males and 8 females.

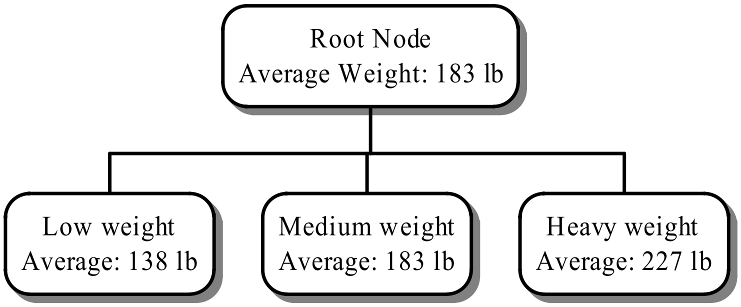
Gender	Weight	Height	Ht_Cent.	BMIndex	BodyType
Female	179	4’10	147	162	slim
Female	160	5’ 4	163	161	slim
Male	191	5’ 8	173	182	average
Male	132	5’1	155	143	slim
Female	167	5’1	180	174	average
Female	128	5’2	157	142	slim
Female	150	5’2	157	154	slim
Male	150	5’2	157	154	slim
Female	215	5’2	157	184	heavy
Female	89	5’3	160	119	slim
Female	167	5’3	160	163	slim
Male	180	5’4	163	171	average
Male	206	5’4	163	183	average
Male	239	5’5	165	199	heavy
Male	161	5’6	168	164	average
Male	188	5’6	168	178	average
Male	284	5’6	168	218	heavy
Female	117	5’7	170	141	slim
Male	163	5’7	170	167	average
Male	194	5’7	170	182	average
Male	201	5’7	170	185	heavy
Male	254	5’8	173	209	heavy
Male	201	5’9	175	188	heavy
Male	206	5’9	175	190	heavy
Male	216	5’9	175	195	heavy
Male	206	6’0	183	194	heavy
Male	220	6’1	185	202	heavy
Female	182	6’2	188	185	heavy

In this display, the overall average height is 5’6 and the overall average weight is 183. Among males, the average height is 5’7, while among females, the average height is 5’3 (males weigh 200 on average, versus 155 for females).

Knowing the gender puts us in a better position to predict the height and weight of the individuals, and knowing the relationship between gender and height and weight puts us in a better position to understand the characteristics of the target. Based on the relationship between height and weight and gender, you can infer that females are both smaller and lighter than males. As a result, you can see how this sort of knowledge that is based on gender can be used to determine the height and weight of unseen humans.

From the display, you can construct a branch with three leaves to illustrate how decision trees are formed by grouping input values based on their relationship to the target.

Figure 1.5: Illustration of Decision Tree Partitioning of Physical Measurements



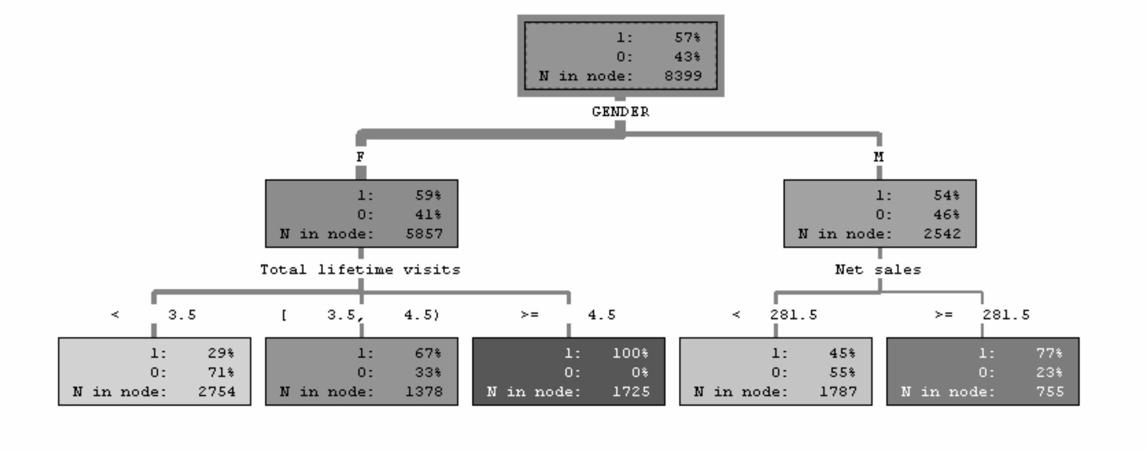
Level of Measurement

The example as shown here illustrates an important characteristic of decision trees: both quantitative and qualitative data can be accommodated in decision tree construction. Quantitative data, like height and weight, refers to quantities that can be manipulated with arithmetic operations such as addition, subtraction, and multiplication. Qualitative data, such as gender, cannot be used in arithmetic operations, but can be presented in tables or decision trees. In the previous example, the target field is weight and is presented as an average. Height, BMIndex, or BodyType could have been used as inputs to form the decision tree.

Some data, such as shoe size, behaves like both qualitative and quantitative data. For example, you might not be able to do meaningful arithmetic with shoe size, even though the sequence of numbers in shoe sizes is in an observable order. For example, with shoe size, size 10 is larger than size 9, but it is not twice as large as size 5.

Figure 1.6 displays a decision tree developed with a categorical target variable. This figure shows the general, tree-like characteristics of a decision tree and illustrates how decision trees display multiple relationships—one branch at a time. In subsequent figures, decision trees are shown with continuous or numeric fields as targets. This shows how decision trees are easily developed using targets and inputs that are both qualitative (categorical data) and quantitative (continuous, numeric data).

Figure 1.6: Illustration of a Decision Tree with a Categorical Target



The decision tree in Figure 1.6 displays the results of a mail-in customer survey conducted by HomeStuff, a national home goods retailer. In the survey, customers had the option to enter a cash drawing. Those who entered the drawing were classified as a HomeStuff *best* customer. Best customers are coded with **1** in the decision tree.

The top-level node of the decision tree shows that, of the 8399 respondents to the survey, 57% were classified as best customers, while 43% were classified as *other* (coded with **0**).

Figure 1.6 shows the general characteristics of a decision tree, such as partitioning the results of a 1–0 (categorical) target across various input fields in the customer survey data set. Under the top-level node, the field **GENDER** further characterizes the best – other (1–0) response. Females (coded with **F**) are more likely to be best customers than males (coded with **M**). Fifty-nine percent of females are best customers versus fifty-four percent of males. A wide variety of splitting techniques has been developed over time to gauge whether this difference is statistically significant and whether the results are accurate and reproducible. In Figure 1.6, the difference between males and females is statistically significant. Whether a difference of 5% is significant from a business point of view is a question that is best answered by the business analyst.

The splitting techniques that are used to split the 1–0 responses in the data set are used to identify alternative inputs (for example, income or purchase history) for gender. These techniques are based on numerical and statistical techniques that show an improvement over a simple, uninformed guess at the value of a target (in this example, best–other), as well as the reproducibility of this improvement with a new set of data.

Knowing the gender enables us to guess that females are 5% more likely to be a best customer than males. You could set up a separate, independent *hold out* or *validation* data set and (having determined that the gender effect is useful or interesting) you might see whether the strength and direction of the effect is reflected in the hold out or validation data set. The separate, independent data set will show the results if the decision tree is applied to a new data set, which indicates the generality of the results. Another way to assess the generality of the results is to look at data distributions that have been studied and developed by statisticians who know the properties of the data and who have developed guidelines based on the properties of the data and data distributions. The results could be compared to these data distributions and, based on the comparisons, you could determine the strength and reproducibility of the results. These approaches are discussed at greater length in Chapter 3, “The Mechanics of Decision Tree Construction.”

Under the female node in the decision tree in Figure 1.6, female customers can be further categorized into best–other categories based on the total lifetime visits that they have made to HomeStuff stores: those who have made fewer than 3.5 visits are less likely to be best customers compared to those who have made more than 4.5 visits: 29% versus 100%. (In the survey, a shopping visit of less than 20 minutes was characterized as a half visit.)

On the right side of the figure, the decision tree is asymmetric; a new field—**Net sales**—has entered the analysis. This suggests that **Net sales** is a stronger or more relevant predictor of customer status than **Total lifetime visits**, which was used to analyze females. It was this kind of asymmetry that spurred the initial development of decision trees in the statistical community: these kinds of results demonstrate the importance of the combined (or interactive) effect of two indicators in displaying the drivers of an outcome. In the case of males, when **Net sales** exceed $281.50, then the likelihood of being a best customer increases from 45% to 77%.

As shown in the asymmetry of the decision tree, female behavior and male behavior have different nuances. To explain or predict female behavior, you have to look at the interaction of gender (in this case, female) with **Total lifetime visits**. For males, **Net sales** is an important characteristic to look at.

In Figure 1.6, of all the k-way or n-way branches that could have been formed in this decision tree, the 2-way branch is identified as best. This indicates that a 2-way branch produces the strongest effect. The strength of the effect is measured through a criterion that is based on strength of separation, statistical significance, or reproducibility, with respect to a validation process. These measures, as applied to the determination of branch formation and splitting criterion identification, are discussed further in Chapter 3.

Decision trees can accommodate categorical (gender), ordinal (number of visits), and continuous (net sales) types of fields as inputs or classifiers for the purpose of forming the decision tree. Input classifiers can be created by binning quantitative data types (ordinal and continuous) into categories that might be used in the creation of branches—or splits—in the decision tree. The bins that form total lifetime visits have been placed into three branches:

- < 3.5 … less than 3.5
- [3.5 – 4.5) … between 3.5 to strictly less than 4.5
- >= 4.5 … greater than or equal to 4.5

Various nomenclatures are used to indicate which values fall in a given range. Meyers (2000) proposes an alternative, which is shown below:

- < 3.5 … less than 3.5
- [3.5 – 4.5[… between 3.5 to strictly less than 4.5
- >= 4.5 … greater than or equal to 4.5

The key difference from the convention used in the SAS decision tree is in the second range of values, where the designator “[” is used to indicate the interval that includes the lower number and includes up to any number that is strictly less than the upper number in the range.

A variety of techniques exist to cast bins into branches: 2-way (binary branches), n-way (where **n** equals the number of bins or categories), or k-way (where **k** represents an attempt to create an optimal number of branches and is some number greater than or equal to 2 and less than or equal to n).

Figure 1.7: Illustration of a Decision Tree—Continuous (Numeric) Target

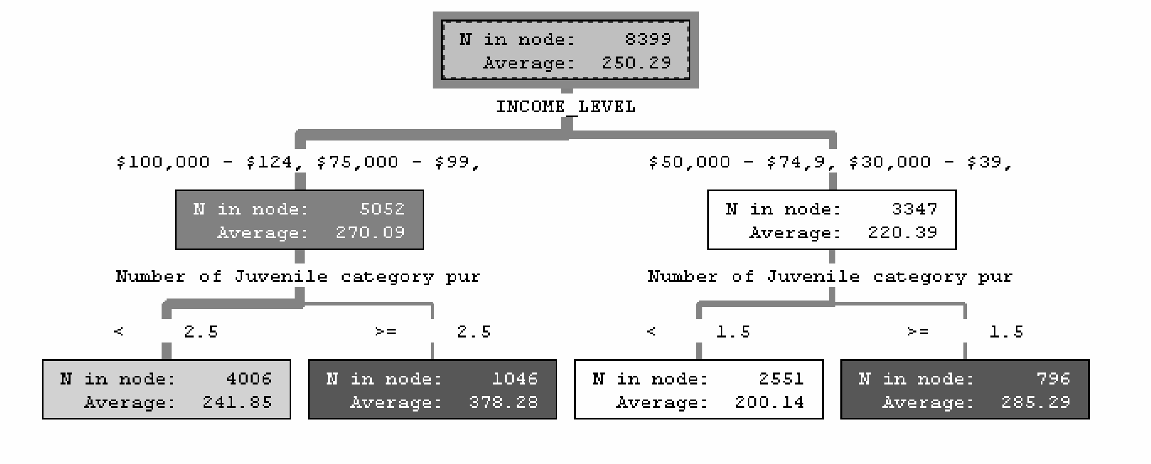


Figure 1.7 shows a decision tree that is created with a continuous response variable as the target. In this case, the target field is **Net sales**. This is the same field that was used as a classifier (for males) in the categorical response decision tree shown in Figure 1.6.

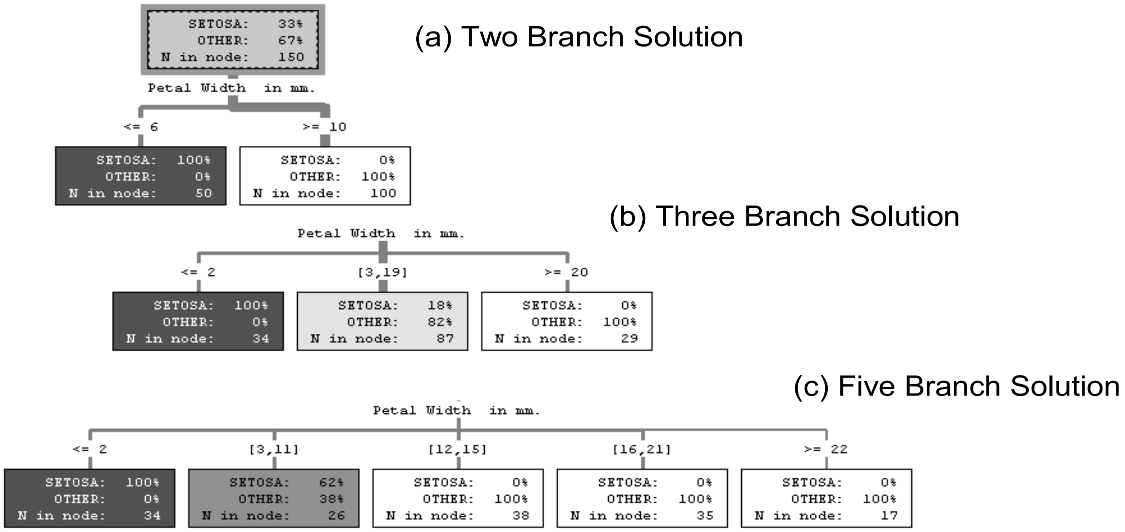
Overall, as shown in Figure 1.7, the average net sale amount is approximately $250. Figure 1.7 shows how this amount can be characterized by performing successive splits of net sales according to the income level of the survey responders and, within their income level, according to the field **Number of Juvenile category purchases**. In addition to characterizing net sales spending groups, this decision tree can be used as a predictive tool. For example, in Figure 1.7, high income, high juvenile category purchases typically outspend the average purchaser by an average of $378, versus the norm of $250. If someone were to ask what a relatively low income purchaser who buys a relatively low number of juvenile category items would spend, then the best guess would be about $200. This result is based on the decision rule, taken from the decision tree, as follows:

```
IF Number of Juvenile category purchases     <      1.5
AND INCOME_LEVEL     $50,000 - $74,9,
                     $40,000 - $49,9,
                     $30,000 - $39,9,
                     UNDER $30,000
                     THEN Average Net Sales = $200.14
```

Decision trees can contain both categorical and numeric (continuous) information in the nodes of the tree. Similarly, the characteristics that define the branches of the decision tree can be both categorical or numeric (in this latter case, the numeric values are collapsed into bins—sometimes called buckets or collapsed groupings of categories—to enable them to form the branches of the decision tree).

Figure 1.8 shows how the Fisher-Anderson iris data can yield three different types of branches when classifying the target SETOSA versus OTHER (Fisher 1936); in this case, 2-, 3-, and 5-leaf branches. There are 50 SETOSA records in the data set. With the binary partition, these records are classified perfectly by the rule **petal width <= 6 mm**. The 3-way and 5-way branch partitions are not as effective as the 2-way partition and are shown only for illustration. More examples are provided in Chapter 2, “Descriptive, Predictive, and Explanatory Analyses,” including examples that show how 3-way and n-way partitions are better than 2-way partitions.

Figure 1.8: Illustration of Fisher-Anderson Iris Data and Decision Tree Options



(a) Two Branch Solution

(b) Three Branch Solution

(c) Five Branch Solution