Prediction modelling with decision tree

By

Dr Shaik A Qadeer

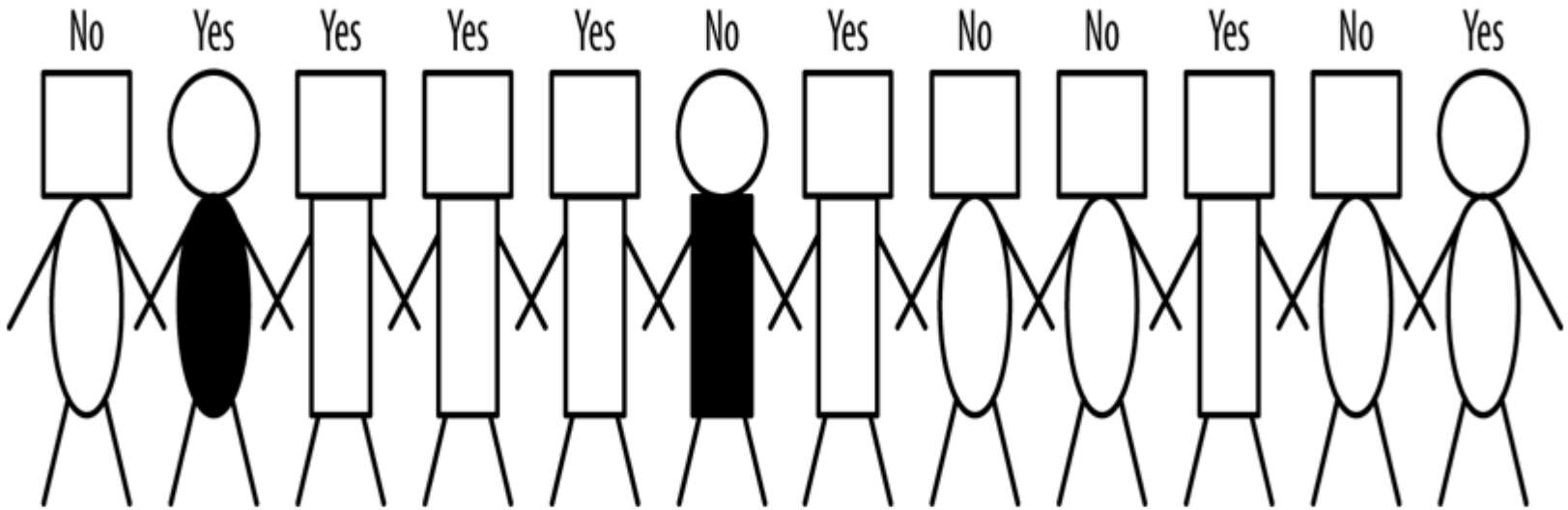
Professor

Classification using decision tree

- Task1:**Selecting Informative Attributes**
- **Task2: Visualising the segmentation**
- **Task3: Trees as set of rules**

Classification using decision tree:Task1

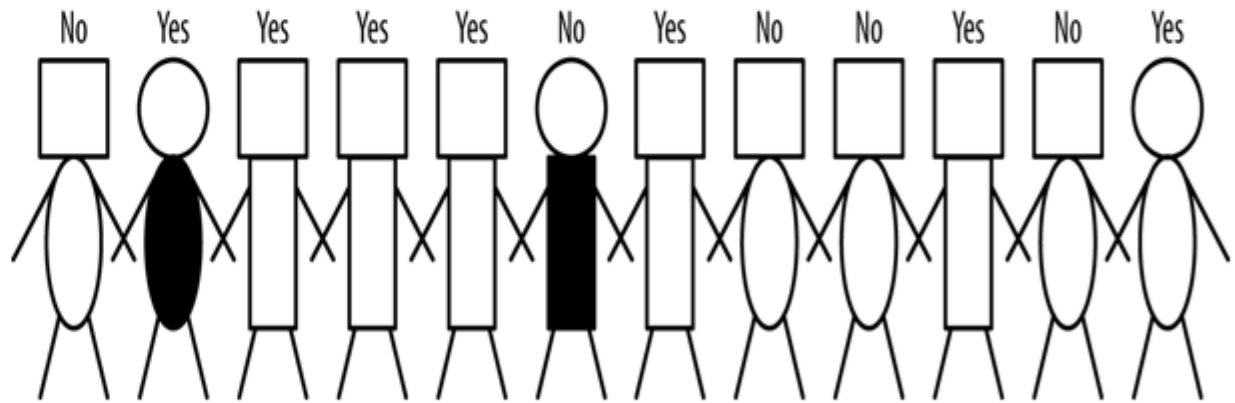
Selecting Informative Attributes



- The label over each head represents the value of the target variable (write-off or not).
- Colors and shapes represent different predictor attributes.

Selecting Informative Attributes

- Attributes:
  - •head-shape: square, circular
  - •body-shape: rectangular, oval
  - •body-color: gray, white
- Target variable:
  - •write-off: Yes, No



Selecting Informative Attributes

- It can be done with Information gain calculation

Example

- A researcher is trying to identify the root node to design a decision tree classifier to classify the students based on scores greater than or equal to 50% which is a pass and less than 50% which is a fail. The data is given in the below table.

Student ID	No. of Assignments Completed	Number of Hours Studied	Attendance in %	Marks out of 100
1	Less than 5	Less than 3 hours	Less than 70%	30
2	Less than 5	Less than 3 hours	Less than 70%	45
3	Less than 5	More than 3 hours	Less than 70%	50
4	More than 5	More than 3 hours	Less than 70%	70
5	More than 5	More than 3 hours	More than 70%	75
6	More than 5	More than 3 hours	More than 70%	85

Q1:Find the entropy of the target column.

- Ans:0.92

Q2Calculate the information gain for the parameter 'Attendance %'.

Q3:Which of the following can be accepted as root node

•

Visualization of segmentation

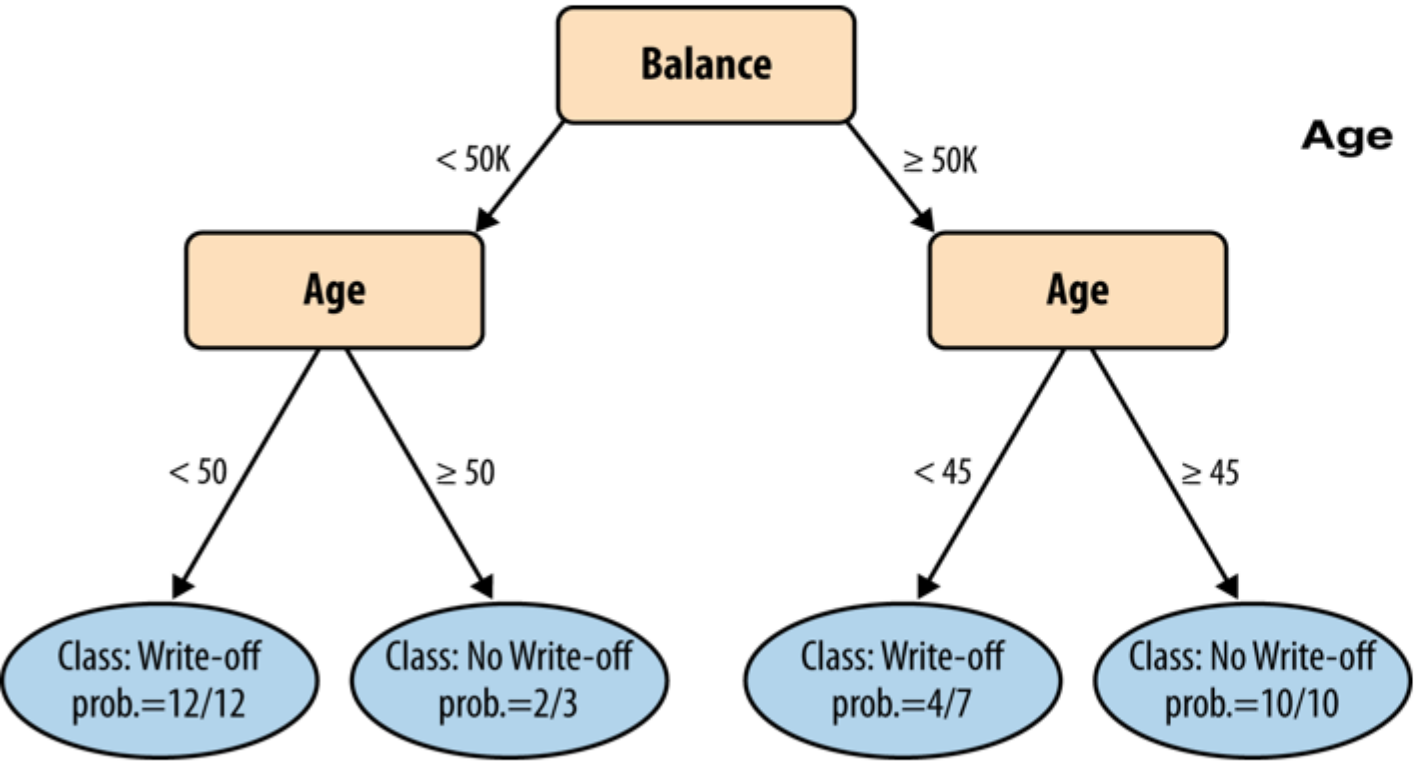
- It is instructive to visualize exactly how a classification tree partitions the instance space.
- The instance space is simply the space described by the data features.

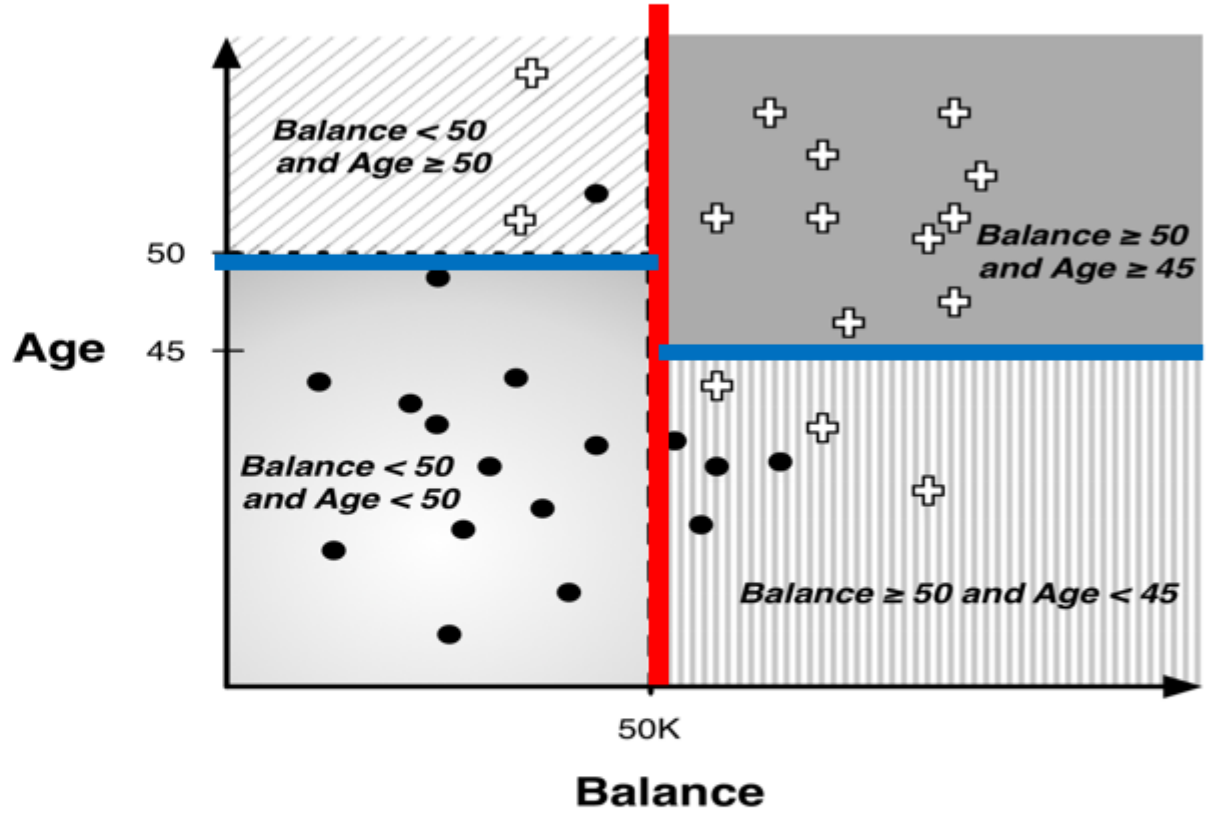
- A common form of instance space visualization is a scatterplot on some pair of features, used to compare one variable against another to detect correlations and relationships.

Visualization of segmentation

- Though data may contain dozens or hundreds of variables, it is only really possible to visualize segmentations in two or three dimensions at once
- Visualizing models in instance space in a few dimensions is useful for understanding the different types of models because it provides insights that apply to higher dimensional spaces as well

Visualizing Segmentations: An Example





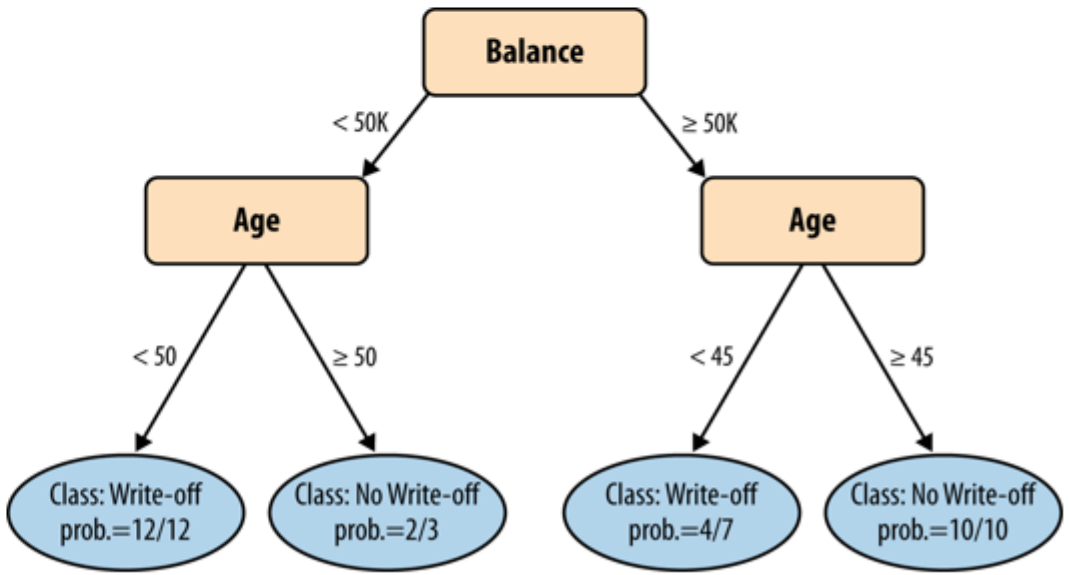
*The black dots correspond to instances of the class Write-off.
*The plus signs correspond to instances of class non-Write-off.

Trees as a set of rule

- You classify a new unseen instance by starting at the root node and following the attribute tests downward until you reach a leaf node, which specifies the instance's predicted class.
- If we trace down a single path from the root node to a leaf, collecting the conditions as we go, we generate a rule.
- Each rule consists of the attribute tests along the path connected with AND.

Trees as Sets of Rules

- IF (Balance < 50K) AND (Age < 50) THEN Class=Write-off
- IF (Balance < 50K) AND (Age ≥ 50) THEN Class=No Write-off
- IF (Balance ≥ 50K) AND (Age < 45) THEN Class=Write-off
- IF (Balance ≥ 50K) AND (Age < 45) THEN Class=No Write-off



Trees as Sets of Rules

- The classification tree is equivalent to this rule set.
- Every classification tree can be expressed as a set of rules this way.

Tree as a set of rule : Example2

- A researcher is trying to identify the root node to design a decision tree classifier to classify the students based on scores greater than or equal to 50% which is a pass and less than 50% which is a fail. The data is given in the below table.

Student ID	No. of Assignments Completed	Number of Hours Studied	Attendance in %	Marks out of 100
1	Less than 5	Less than 3 hours	Less than 70%	30
2	Less than 5	Less than 3 hours	Less than 70%	45
3	Less than 5	More than 3 hours	Less than 70%	50
4	More than 5	More than 3 hours	Less than 70%	70
5	More than 5	More than 3 hours	More than 70%	75
6	More than 5	More than 3 hours	More than 70%	85