IMS Health & Quintiles are now

IQVIA™

Day - 4
Classification Tree

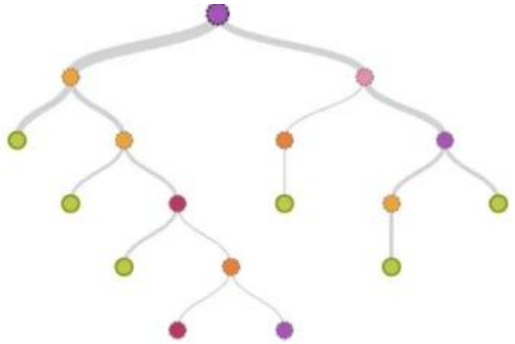
January 11, 2018

What will I learn?

+ Concept behind classification tree and random forest
+ When to use logistic regression and when to use classification tree?
+ How to develop classification tree and random forest in R?

An overview of Classification tree

- It is a non-parametric tree based algorithm used for prediction of discrete variable using a mix of continuous and discrete predictors



Examples:

a. Does a customer default on credit card payment or not

b. To understand whether the HCPs will respond to the sales force campaign (i.e. Response = Yes) or not (i.e. Response = No)

c. Will a student get admission into business school (i.e. Response = Yes) or not (i.e. Response = No)

All of these are examples for categorical outcome variable

Example of a Decision Tree (1/2)

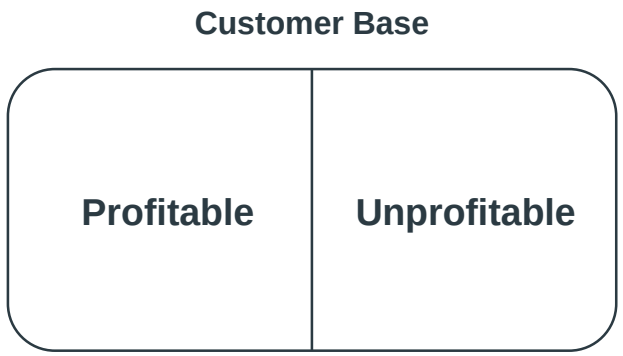
ILLUSTRATIVE

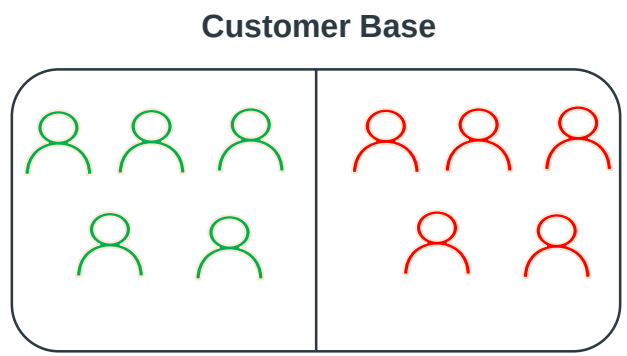
In order to understand decision tree let us start with an example.

Let us take a credit card company which has a set of customers. Some of them are profitable and some of them are unprofitable

- **Unprofitable customers :** Do not use credit cards frequently / use card but pay on time
- **Profitable customers :** Do not make payment in full (carry balance on card) or on time

Company's customer base



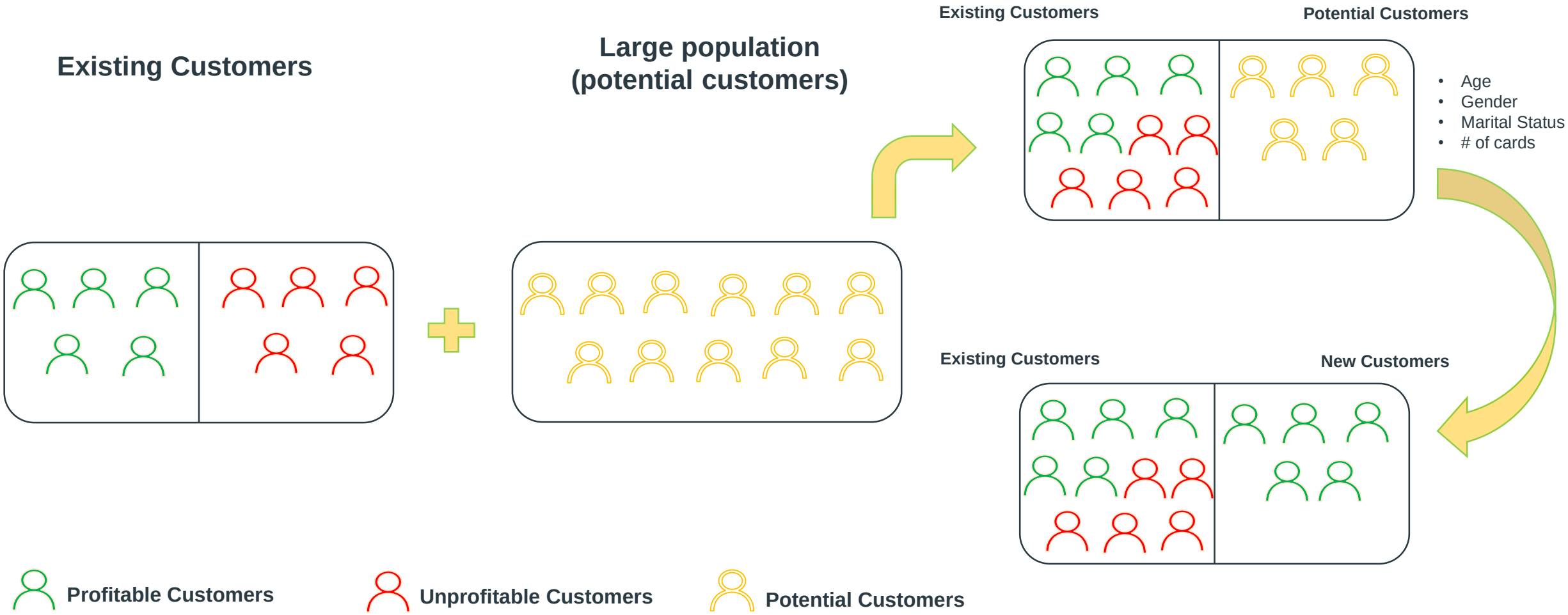


Profitable Customers

Unprofitable Customers

Example of a Decision Tree (2/2)

ILLUSTRATIVE



*** Potential Customers are the people who are not the customers of this company but the company can market to these customers so they have the potential to be its customers**

Apply Model to Test Data

ILLUSTRATIVE

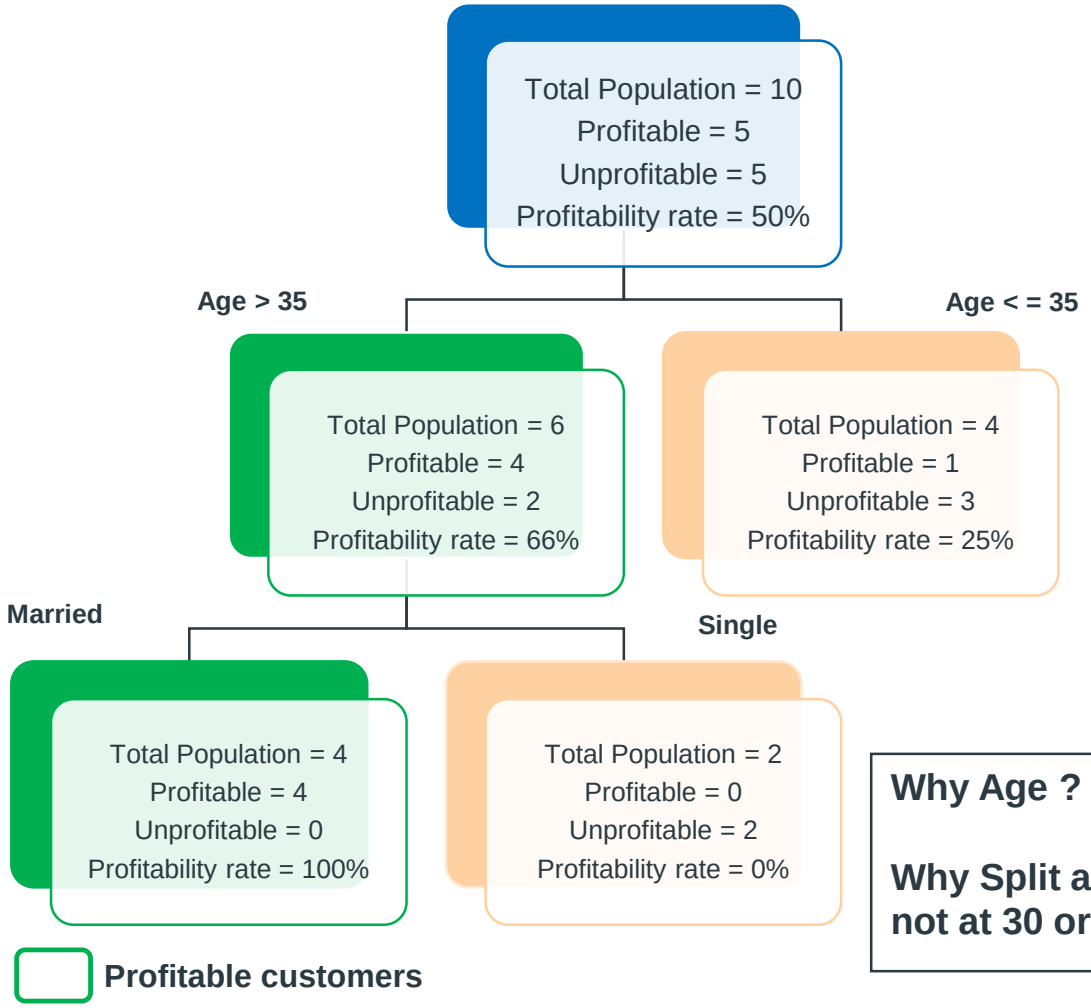
Existing customer data base

S. No	Age	Gender	Marital Status	Profitability
1	36	M	M	P
2	32	F	S	U
3	38	M	M	P
4	40	F	S	U
5	44	F	M	P
6	56	F	M	P
7	58	M	S	U
8	30	M	S	P
9	28	M	M	U
10	26	M	M	U

Profitable Customers

Unprofitable Customers

Decision tree output (Historical Data)



Why Age ?

Why Split at 35 and not at 30 or 45?

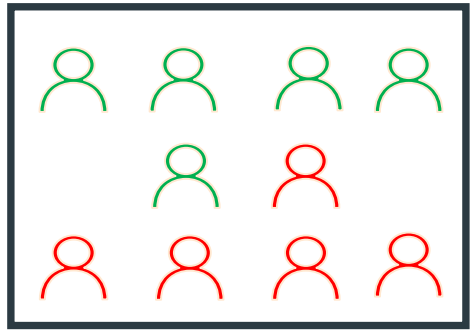
How to determine the Best Split (1/2)

ILLUSTRATIVE

Gini Coefficients = $(p(green))^2 + (p(red))^2$
Where p(green) is the proportion of green in the data

Profitable Customers

Unprofitable Customers

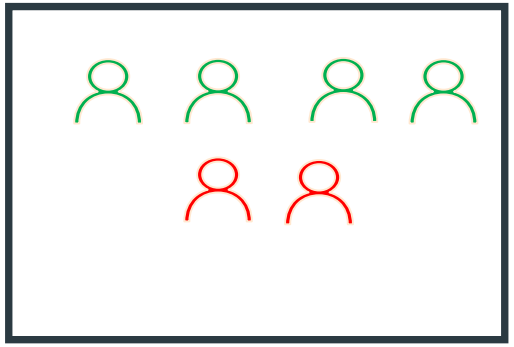


# of greens = 5
# of reds = 5

Proportion of greens = 0.50
Proportion of reds = 0.50

Gini = 0.25 + 0.25 = 0.50

Age > 35

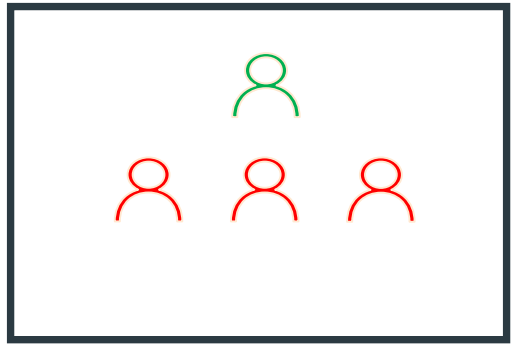


# of greens = 4
# of reds = 2

Proportion of greens = 0.67
Proportion of reds = 0.33

Gini = $(0.67)^2 + (0.33)^2 = 0.56$

Age < = 35



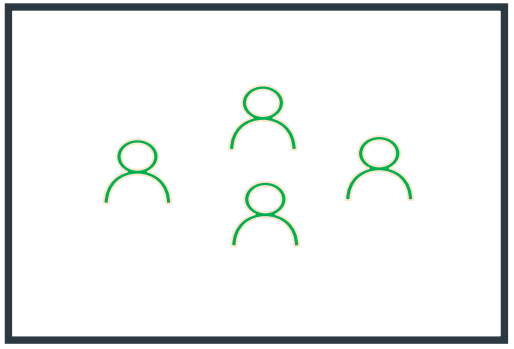
# of greens = 1
# of reds = 3

Proportion of greens = 0.25
Proportion of reds = 0.75

Gini = $(0.25)^2 + (0.75)^2 = 0.62$

Gini score for the split
(4/10)*0.56 + (1/10)*0.62 = **0.84**

Married

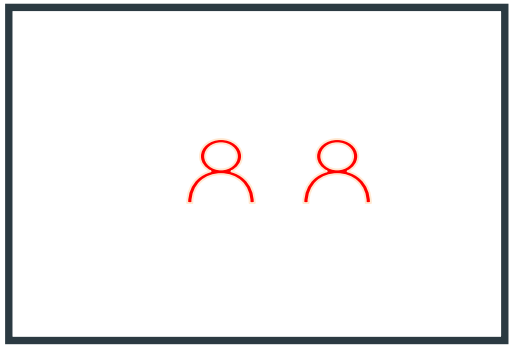


# of greens = 4
# of reds = 0

Proportion of greens = 1.0
Proportion of reds = 0

Gini = $(1.0)^2 + (0)^2 = 1.0$

Single



# of greens = 0
# of reds = 2

Proportion of greens = 0
Proportion of reds = 1.0

Gini = $(0)^2 + (1.0)^2 = 1.0$

Gini score for the split
(4/6)*1.0 + (0/6)*1.0 = **0.67**

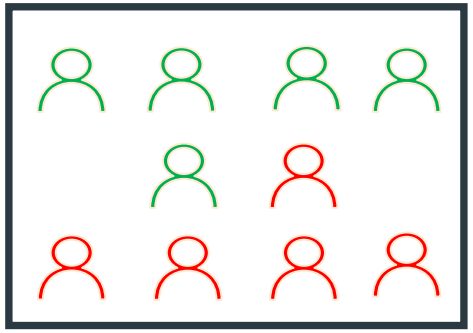
How to determine the Best Split (2/2)

ILLUSTRATIVE

Gini Coefficients = $(p(green))^2 + (p(red))^2$
Where p(green) is the proportion of green in the data

Profitable Customers

Unprofitable Customers

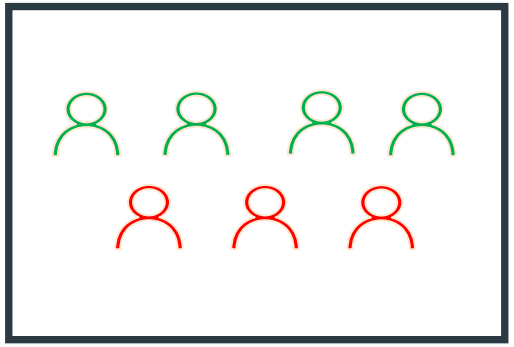


# of greens = 5
# of reds = 5

Proportion of greens = 0.50
Proportion of reds = 0.50

Gini = 0.25 + 0.25 = 0.50

Age > 30

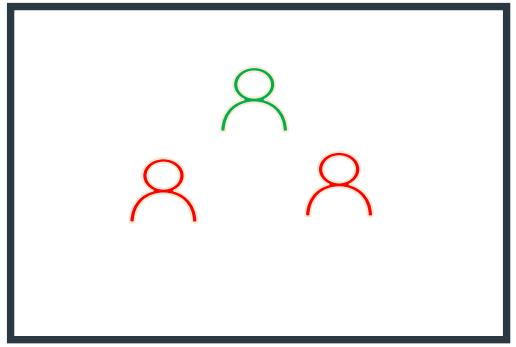


# of greens = 4
# of reds = 3

Proportion of greens = 0.57
Proportion of reds = 0.43

Gini = $(0.57)^2 + (0.43)^2 = 0.51$

Age < = 30



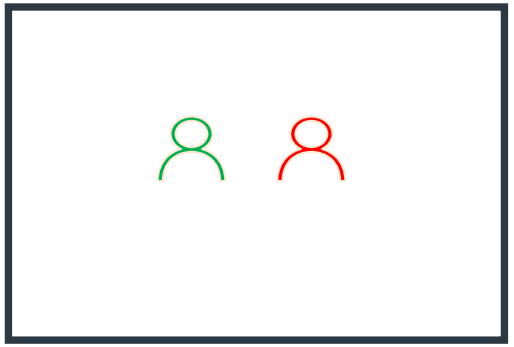
# of greens = 1
# of reds = 2

Proportion of greens = 0.33
Proportion of reds = 0.67

Gini = $(0.33)^2 + (0.67)^2 = 0.56$

Gini score for the split
(4/10)*0.51 + (1/10)*0.56 = **0.75**

Age > 45

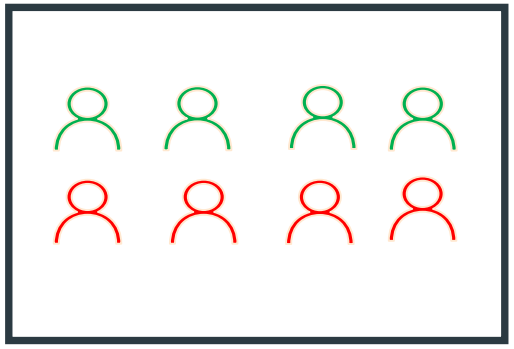


# of greens = 1
# of reds = 1

Proportion of greens = 0.5
Proportion of reds = 0.5

Gini = $(0.5)^2 + (0.5)^2 = 0.5$

Age < = 45



# of greens = 4
# of reds = 4

Proportion of greens = 0.5
Proportion of reds = 0.5

Gini = $(0.5)^2 + (0.5)^2 = 0.5$

Gini score for the split
(1/10)*0.5 + (4/10)*0.5 = **0.25**

Performance evaluation vs model validation

Model performance evaluation : It is an assessment of how accurate the model is, and how well it answers the business question framed

Statistical evaluation
- How well is the model “predicting”/”explaining” ?
- **Metric :** Classification table / Confusion matrix

Business evaluation
- Are the relationship captured by the model intuitive and explainable?
- **Metric :** Look for business explanation

Model Validation : It is assessment of how valid and applicable the model is, beyond the sample on which it was generated

Training dataset
- Typically models should be **build on the training data set**

Test dataset
- Developed model should be used on the test data set **to ensure the general applicability of the model**

Performance evaluation technique

Performance of logistic regression model can be accessed through classification table / confusion matrix

Confusion matrix looks as shown below :

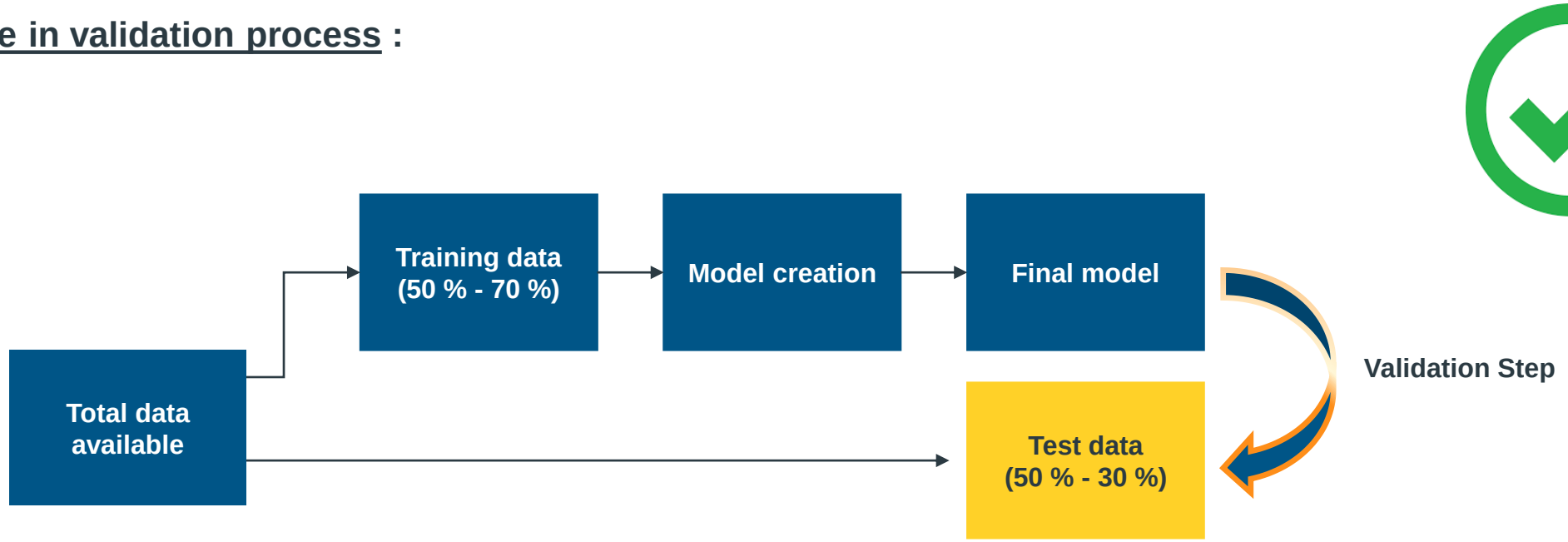
Actual Outcome \ Predicted Outcome	Replayed	Not Replayed
Replayed	True Positive (TP)	False Negative (FN)
Not Replayed	False Positive (TP)	True Negative (TN)

$$Model\ Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

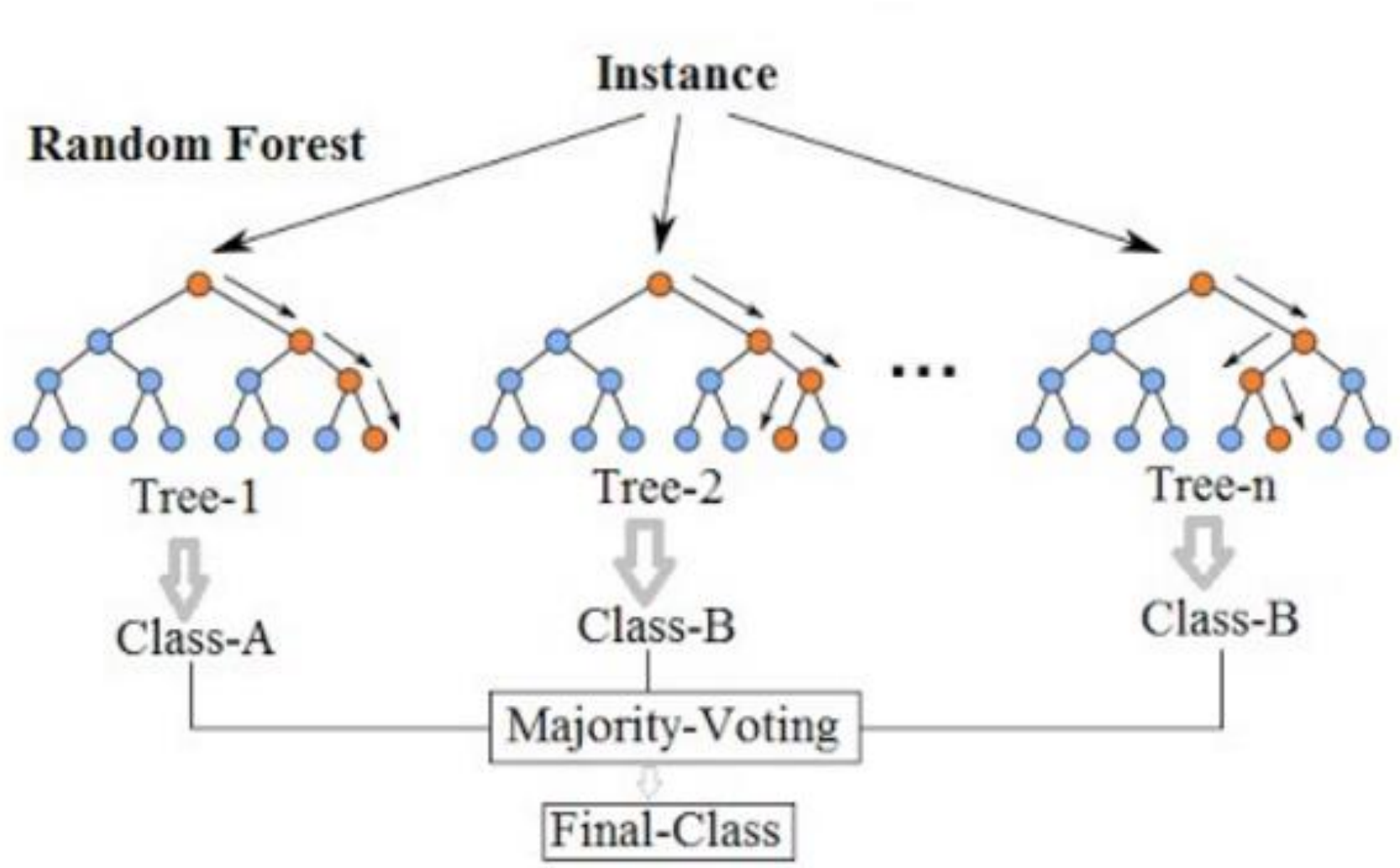
Model validation technique

It is assessment of how valid and applicable the model is, beyond the sample on which it was generated

Steps involve in validation process :



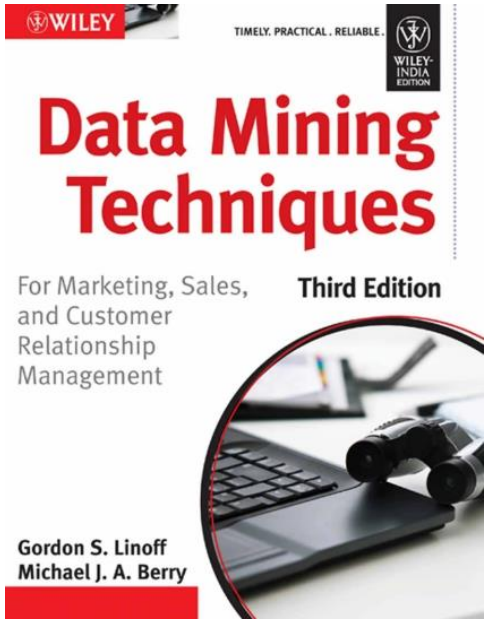
An overview of random forest



IMS Health & Quintiles are now IQVIA™

Reference :

Data Mining Techniques
- Gordon S. Linoff



IMS Health & Quintiles are now IQVIA™

Practical Session – Implementation in R