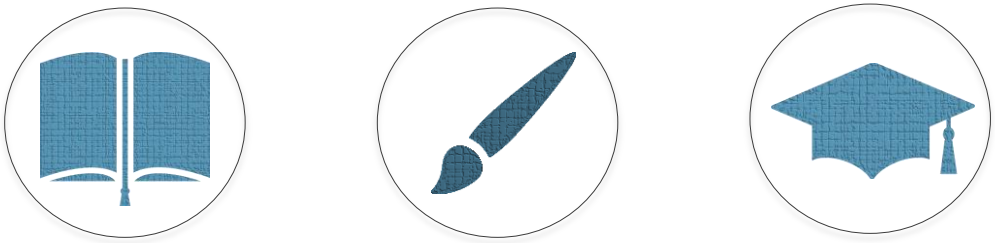


Data Science Concepts

Lesson04–Decision Tree Concepts

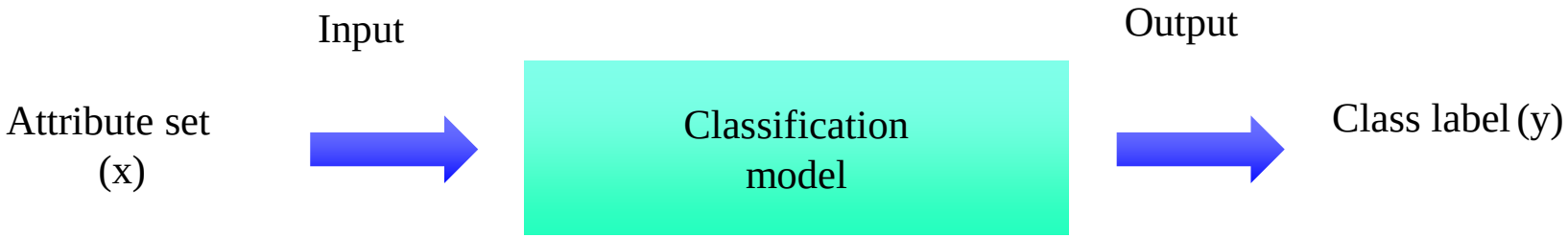
Objective

After completing this lesson you will be able to:

- Explain Decision Trees and its applications
- Explain the various parameters which are used to evaluate the outcome of the decision trees.

Decision Trees

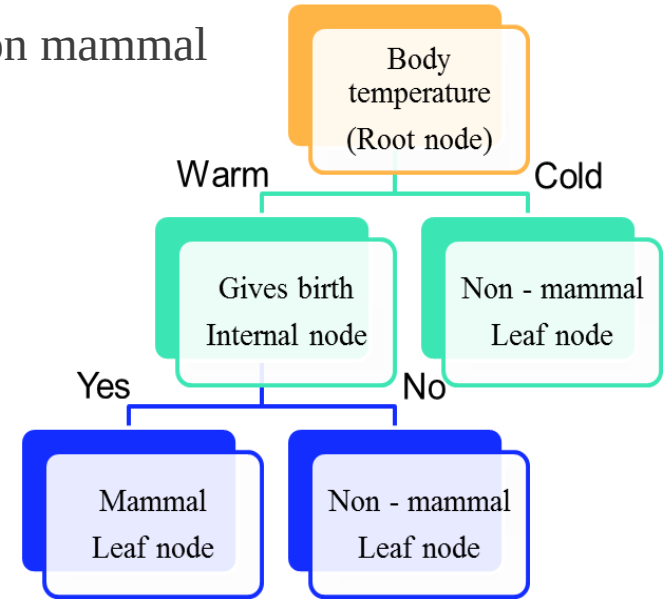
- Classification is a task of assigning objects to one of the several pre-defined categories.
  - Descriptive modelling: Can be used as an explanatory tool to distinguish between objects of different classes.
  - Predictive modelling: Can be used to predict the class label of unknown records.



- Objective is to build a learning algorithm with good generalization capability.

Decision Tree–Concept Development

Classifying species as mammal or non mammal



CART	C5.0	CHAID
Hunt's algorithm	Hunt's algorithm	CHAID algorithm
Split: Gini Index	Split: Entropy	Split: x^2 test

Criteria for comparing different methods: Predictive accuracy, speed, robustness, scalability, Interpretability

Decision Tree - CHAID

The hypothesis being tested is:

- H0: There is no relationship between the two variables (Y and one of the X's which is selected)
- Ha: There is a relationship between the two variables (dependent)

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right); \textbf{Expected frequency} = \frac{rowsum * columnSum}{totalSum}$$

- **Steps in the test**
  - Examine each predictor variable for its statistical significance with the dependent variable.
  - Determine the most significant predictors using p-value (smallest P-Value).
  - Divide the data by levels of the most significant predictors. Each of these groups will be examined individually further.
  - For each sub-group, determine the most significant variable from the remaining predictor and divide the data again.

Decision Tree - CHAID

The calculation is repeated for every X w.r.t Y.

X with the smallest P value is picked for first split.

The steps are repeated for second level tree

Contingency Table

Gender (X)	NPA Status (Y)		Total
	1	0	
Male (0)	135	139	274
Female(1)	165	561	726
Total	300	700	1000

Calculation Table

	Observed	Expected	$(O-E)^2/E)$
M - (Status1)	135	(274*300)/1000 = 82.2	33.91
M - (Status0)	139	(274*700)/1000=191.8	14.53
F - (Status1)	165	(726*300)/1000=217.8	12.8
F - (Status0)	561	(726*700)/1000=508.2	5.48
Total	1000	1000	66.73
		P – value	3.106E-16

Decision Tree - CART

- CART (Classification and Regression Tree) always performs binary splits.
  - Gini Index is a measure of impurity at the node. If sample is completely homogenous then less impurity. If sample is equally divided then more impurity.

$$\boldsymbol{i(t) = Gini(t) = \sum_{j=1}^{J} P(j \mid t) * \left(1 - P(j \mid t)\right)}$$

$$where P(j|t)\ is\ the\ proportion\ of\ cateogy\ j\ at\ node\ t.$$

$$Change\ in\ impurity = \left[\boldsymbol{i(t) - P_L * i\left(t(L)\right) - P_R * i\left(t(R)\right)}\right]$$

$$P_L = Proportion\ of\ obs\ in\ left\ branch$$

$$P_R = Proportion\ of\ obs\ in\ left\ branch$$

  - The variable which maximizes the change in impurity is picked up for building decision tree
  - In case of a two category, minimum value of Gini is 0 and Maximum value of Gini can be 0.5 (50% zeros and 50% ones as the two categories).
  - If a variable has more than two classes, the classes are combined and then Gini index is computed:

$$No\ of\ combinations = 2^{k-1} - 1$$

Decision Tree - CART

- Entropy is another measure to select the best split

$$\boldsymbol{Entropy(t) = -\sum_{j=1}^{J} P(j \mid t) * log_2\big(P(j \mid t)\big)}$$

$$where P(j|t)\ is\ the\ proportion\ of\ cateogy\ j\ at\ node\ t.$$

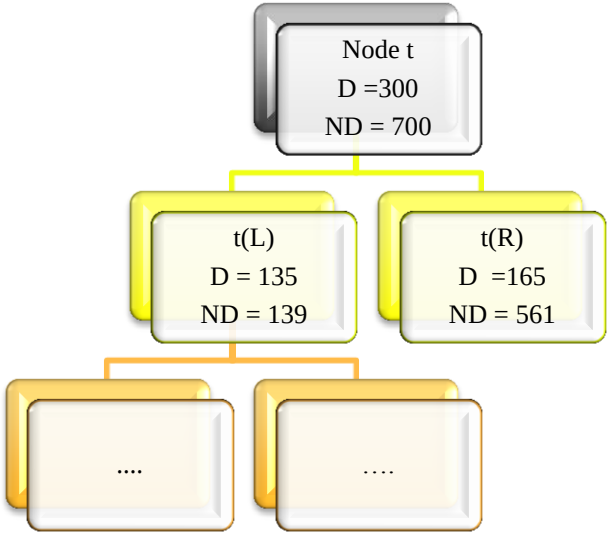
  - The variable which maximizes the change in impurity is picked up for building decision tree
  - In case of a two category, minimum value of entropy is 0 and Maximum value is 1 (50% zeros and 50% ones as the two categories).
  - If a variable has more than two classes, the classes are combined and then Gini index is computed:

$$No\ of\ combinations = 2^{k-1} - 1$$

Decision Tree - CART

Contingency Table

Gender (X)	NPA Status		Total
	1	0	
Node (t(L): Male	135	139	274
Node (t(R): Female	165	561	726
Total	300	700	1000



Calculation Table

Node	Proportion of the Class	i(t) = P(j | K)*(1-P(j | K))		Proportion	$\Delta i(t)$
t():	P(D | t) = 300/1000 P(ND | t) = 700/1000	0.30*(1-0.30) = 0.21 0.70*(1.0.70) = 0.21	**0.42**		
t(L): Male	P(D | t(L)): 135/274 = 0.49 P(ND | t(L)): 139/274 = 0.51	(0.49)*(1-0.49) = 0.25 (0.51)*(1-0.51) = 0.25	**0.50**	274/1000= 0.274	[0.42 − (0.27∗0.50) − (0.726∗0.34)]= 0.038
t(R): Female	P(D | t(R)): 165/726 = 0.23 P(ND | t(R)): 561/726 = 0.77	(0.23)*(1-0.23) = 0.17 (0.77)*(1-0.77) = 0.17	**0.34**	726/1000= 0.726	

Decision Tree–Classification Matrix

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) = \frac{4}{7} = 57.1\%$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) = \frac{17}{17} = 100\%$$

Classification matrix		
	Predicted	
Observed	Class=1 (Positive)	Class=0 (Negative)
Class =1 (Positive)	f_{11}= 4 [TP]	f_{10}= 3 [FN]
Class =0 (Negative)	f_{01}= 0 [FP]	f_{00}= 17 [TN]

$$Model\ accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) = \frac{21}{24} = 87.5\%$$

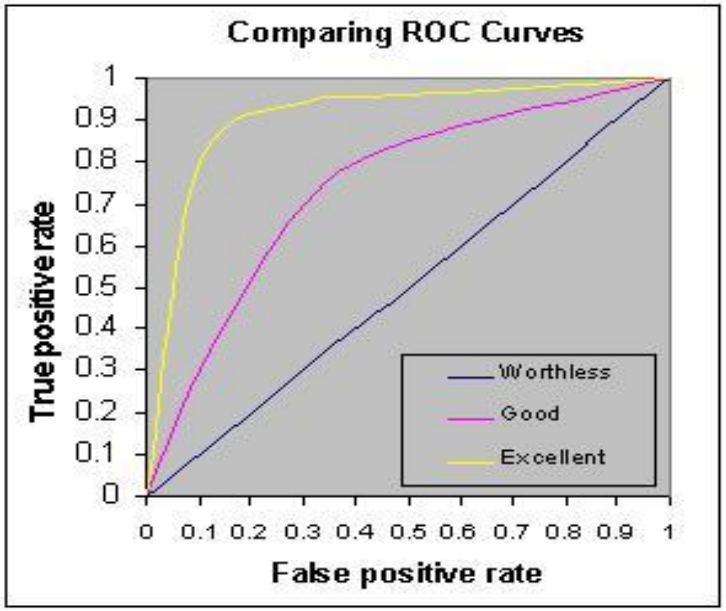
Sensitivity is the probability that predicted class is 1 when observed class is 1.
Specificity is the probability that the predicted class is 0 when the observed class is 0.

Decision Tree–ROC Curve

- Receiver operating characteristics (ROC) Curve is a useful way to determine cut-off point which maximizes sensitivity and specificity.
- Sensitivity and specificity measures are computed based on a sequence of cut-off points to be applied to the model for predicting observations into Positive or Negative.

An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). AUC values between:

- 0.9-1 indicate perfect sensitivity and specificity,
- 0.8-0.9 indicate good sensitivity and specificity,
- 0.7-0.8 indicate fair sensitivity and specificity,
- 0.6-0.7 is poor
- 0.6 and below indicate by chance outcome



Decision Tree–Gain Chart and Lift Chart

- Lift and Gain chart measure how much better one can expect to do with the model comparing without a model.
- In contrast to the confusion/classification matrix that evaluates models on the whole population, gain or lift chart evaluates model performance in a portion of the population.

Steps to build Gain / Lift:

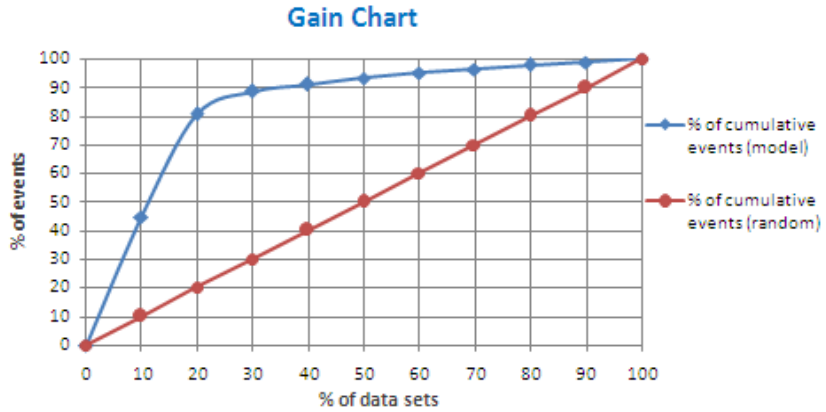
1. Randomly split data into two samples (say): 80% = training sample, 20% = validation sample.
2. Score (predicted probability) the validation sample using the response model (training sample).
3. Rank the scored file, in descending order by probability.
4. Split the ranked file into 10 sections (deciles). Count the number of events in each section.

Cumulative gains and lift charts are a graphical representation to depict the advantage of using a predictive model to choose which customers to contact.

Decision Tree–Gain Chart

- Gain at a given decile level is the ratio of cumulative number of targets (events) up to that decile to the total number of targets (events) in the entire data set.

	Input Values					
Decile	Number of Cases	Number of Responses	Cumulative Responses	% of events	Gain	Cumulative Lift
1	2500	2179	2179	44.71	44.71	4.47
2	2500	1753	3932	35.97	80.67	4.03
3	2500	396	4328	8.12	88.80	2.96
4	2500	111	4439	2.28	91.08	2.28
5	2500	110	4549	2.26	93.33	1.87
6	2500	85	4634	1.74	95.08	1.58
7	2500	67	4701	1.37	96.45	1.38
8	2500	69	4770	1.42	97.87	1.22
9	2500	49	4819	1.01	98.87	1.10
10	2500	55	4874	1.13	100.00	1.00
	25000	4874				



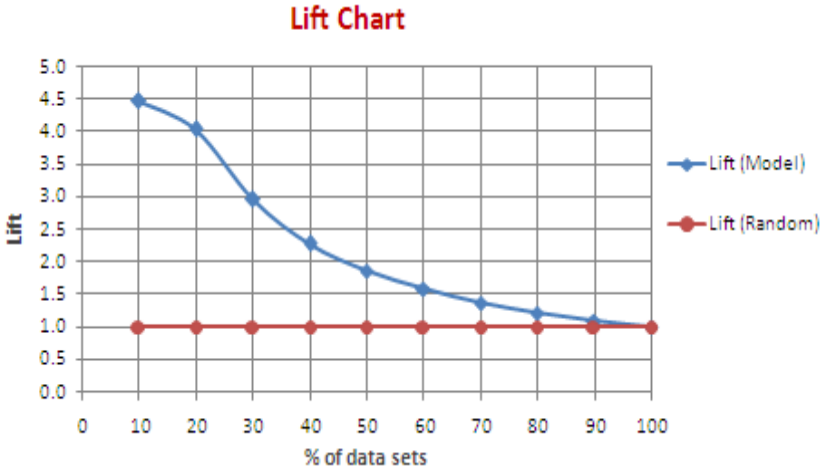
Source: http://www.listendata.com/2014/08/excel-template-gain-and-lift-charts.html

Decision Tree–Lift Chart

- Lift measures how much better one can expect to do with the model comparing without a model.
- It is the ratio of gain % to the random expectation at a given decile level. The random expectation at the x^{th} decile is x%.

Interpretation:

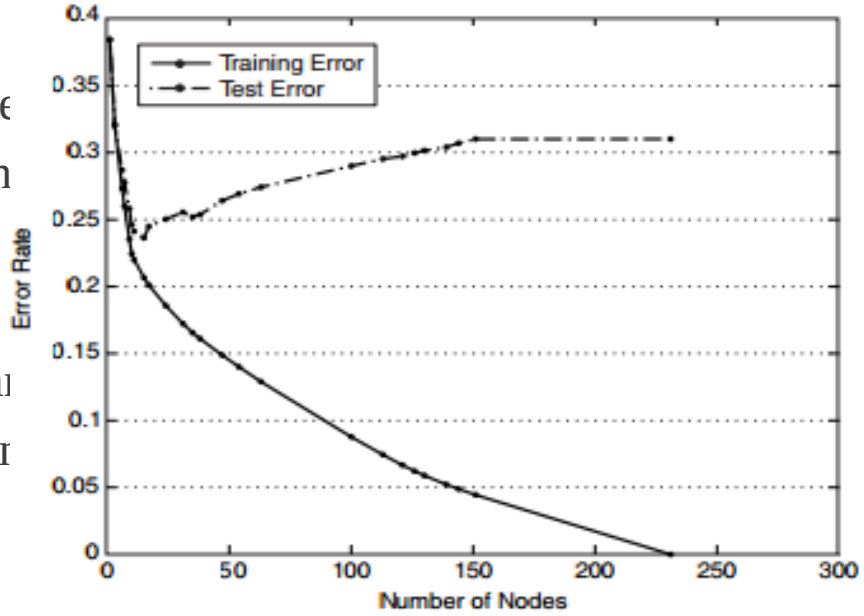
By contacting only 10% of customers, 4.5 times customers may respond.



To build Lift and Gain Chart in R. Refer to
https://heuristically.wordpress.com/2009/12/18/plot-roc-curve-lift-chart-random-forest/

Decision Tree–Under Fitting and Over Fitting

- Model under fitting:
  - Model did not learn from the training set due
  - Training and test error rate are large when th
- Model overfitting:
  - Model has learned too much from the data a
  - As the number of nodes increases, the trainir increase
  - More complex trees than needed.



Model under fitting or over fitting leads to lack of generalizability and thus such decision tree models may not be useful in correct classification on unknown cases.

Decision Tree–Pruning

Pruning is applied to overcome the under fitting or over fitting issues in the decision tree model

Pre-pruning

Stop the algorithm before it becomes a fully grown tree:

- Stop if number of instances is less than some user specified threshold.
- Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain) by at least some threshold

This is more efficient but less accurate.

Post Pruning

Grow decision tree to its entirety. Trim the nodes of the decision tree in a bottom-up fashion

- If generalization error improves after trimming, replace sub-tree by a leaf node.
- Class label of leaf node is determined from majority class of instances in the sub-tree

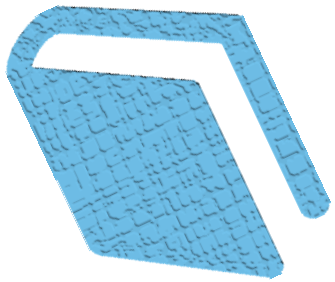
This is more accurate but less efficient.

Misclassification error pruning: Decision tree pruning stops when number of cases in a terminal node becomes less than a threshold

Decision Tree in R Using an Example

Summary

Summary of the topics covered in this lesson:



- Decision Tree is one of the most widely used data mining technique.
- The outcome of decision tree can be used for exploration of data as well as to build in predictive model.
- Unlike regression and logistic regression model, there are no statistical attributes which can suggest that the decision tree model is good and generalizable.

QUIZ TIME

Quiz Question 1

Quiz 1

Which of the below is a correct statement?

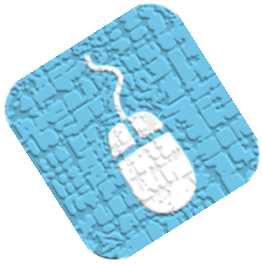
Select all that apply?

a. Sensitivity is the probability that predicted class is 1 when observed class is 1.

b. Specificity is the probability that the predicted class is 1 when the observed class is 0.

c. Specificity is the probability that the predicted class is 0 when the observed class is 0.

d. Sensitivity is the probability that predicted class is 0 when observed class is 1.



Quiz Question 1

Quiz 1

Which of the below is a correct statement?

Select all that apply?

a. Sensitivity is the probability that predicted class is 1 when observed class is 1.

b. Specificity is the probability that the predicted class is 1 when the observed class is 0.

c. Specificity is the probability that the predicted class is 0 when the observed class is 0.

d. Sensitivity is the probability that predicted class is 0 when observed class is 1.

Correct answer is: b & d are incorrect statements.

a & c

End of Lesson04–Decision Tree Concepts