

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

BC2406 Business Analytics I: Predictive Techniques

Seminars 7

Classification and Decision Tree

Instructor: Prof. Lee Gun-woong

Nanyang Business School

Classification Techniques

- Decision Tree based Methods
- Neural Networks
- Naïve Bayes and Bayesian Networks
- Support Vector Machines
- Logistic Regression
- Discriminant Analysis

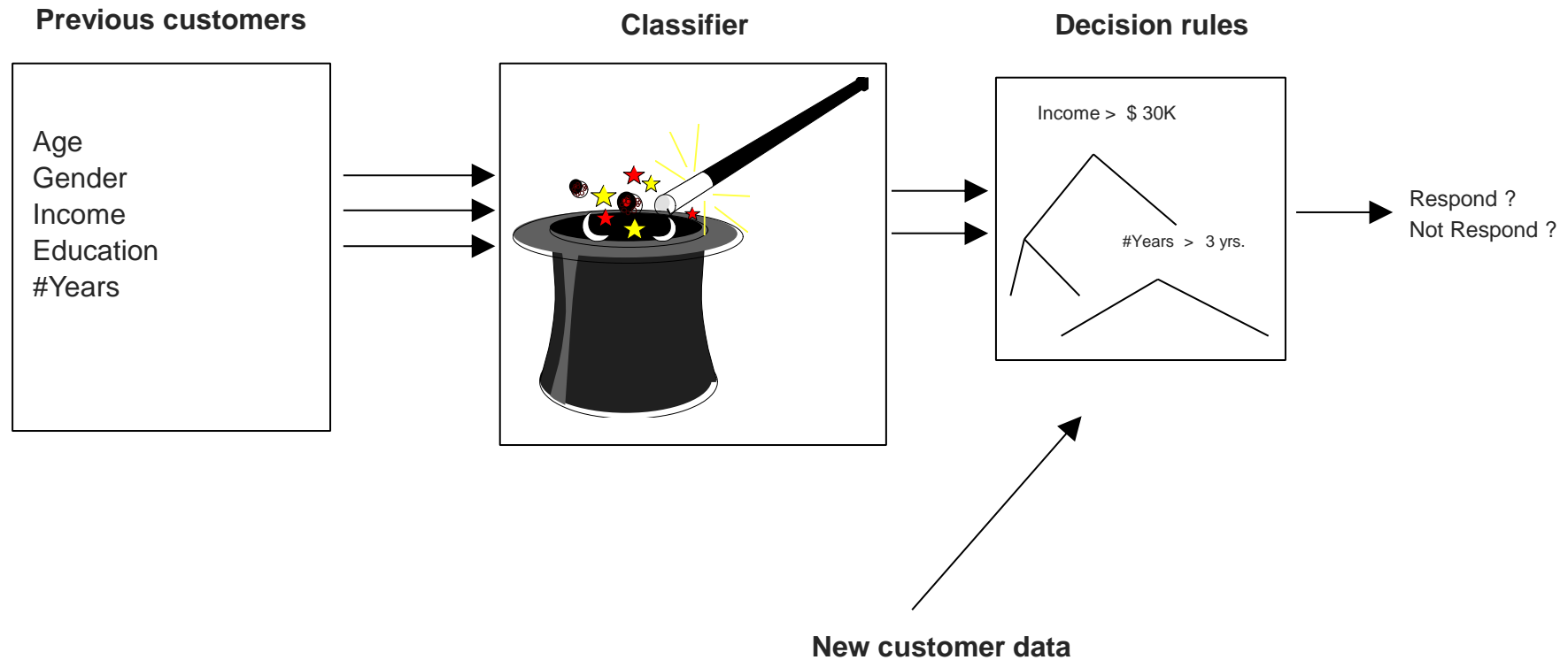
Classification: Definition

- Given a collection of records
 - Each record contains a set of *attributes/variables*
 - One of the attributes is the *class*
 - An outcome/target variable including classes (e.g., Yes / No)
- A classification analysis is to find a *model* for the class attribute as a function of the values of other attributes.
 - Marketing studies of customer behavior such as satisfaction or churn
 - Diagnosis of medical conditions based on lab measurements and symptoms

Key Procedures in a Classification Analysis

- Model Construction
 - Model the relationships among the features/attributes and the potential outcomes
- Model Evaluation
 - Evaluate model predictive performance (accuracy)
- Model Use
 - **Classify unseen instances (examples/records/observations)** to predict the class labels for new unclassified instances

Which customers are likely to respond to a given offer?"

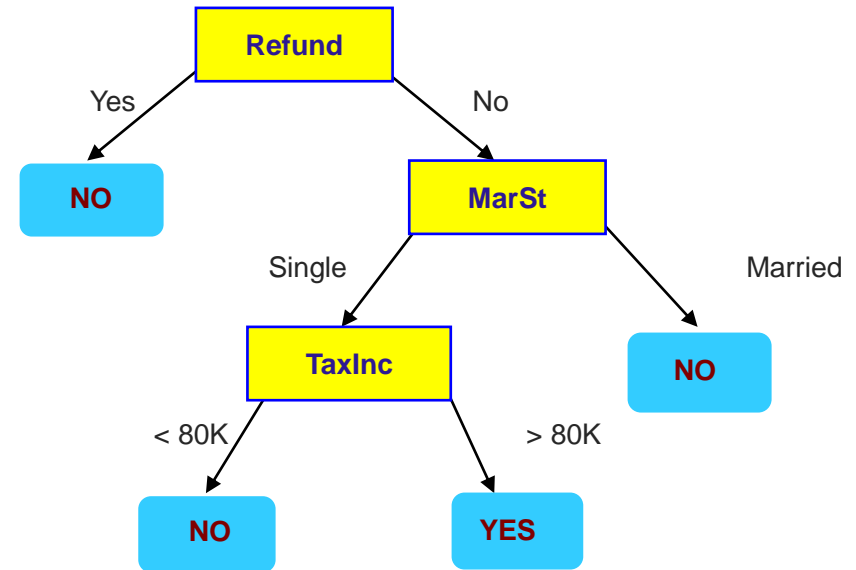


Decision Tree

Model Construction

Who are likely to be tax evaders?

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Single	95K	Yes
6	No	Married	60K	No
7	Yes	Single	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



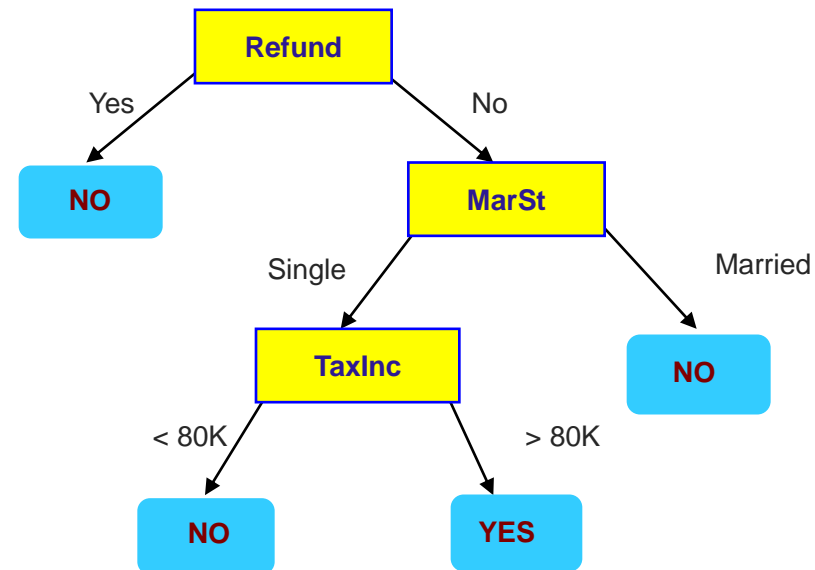
Model: Decision Tree

Training Data

Training Data: The set of all instances used for learning the model is called **training set**

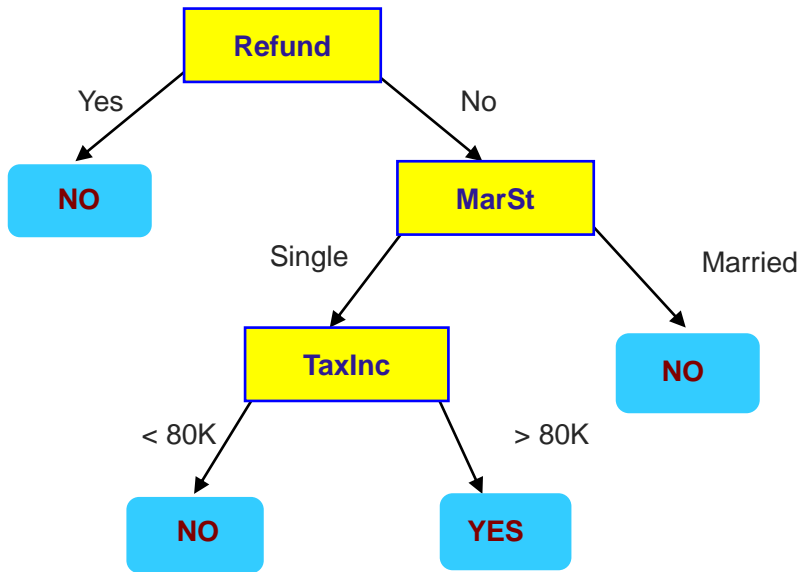
Decision Tree

- Root Node
 - Has no incoming edges
 - Has many outgoing edges
- Internal Nodes
 - Has exactly one incoming edge and two or more outgoing edges
 - Contain conditions to separate records
- Leaf (terminal) Nodes
 - Has exactly one incoming edge and no outgoing edge
 - Each leaf node is assigned a class label



Model: Decision Tree

Rules

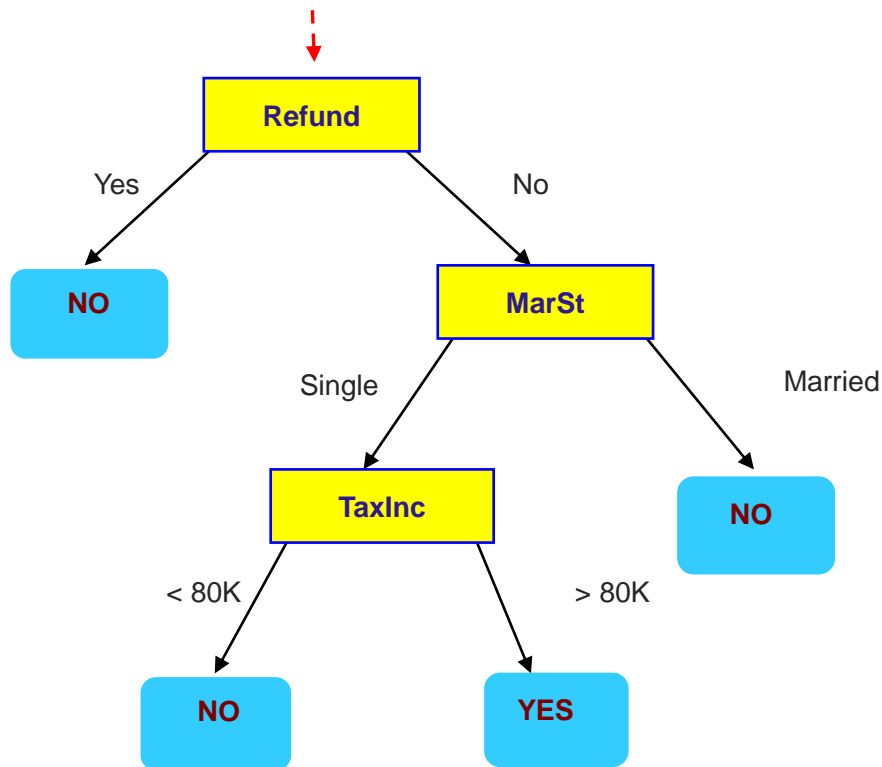


Model: Decision Tree

- If (*Refund* = Yes) Then No
- If (*Refund* = No) And (*MarSt* = Single) And (*TaxInc* < 80K)
Then Class = No
- Any other rules?

Model Use

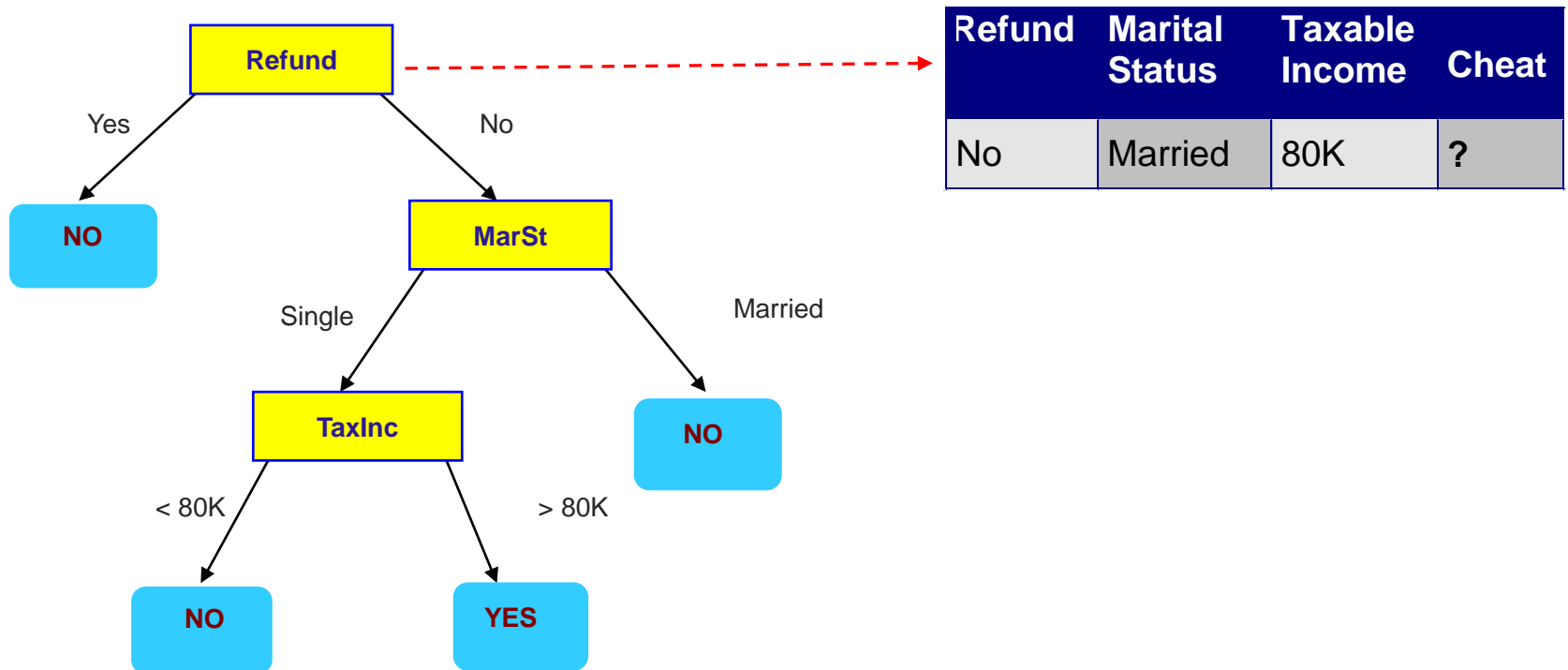
Start from the root of tree.



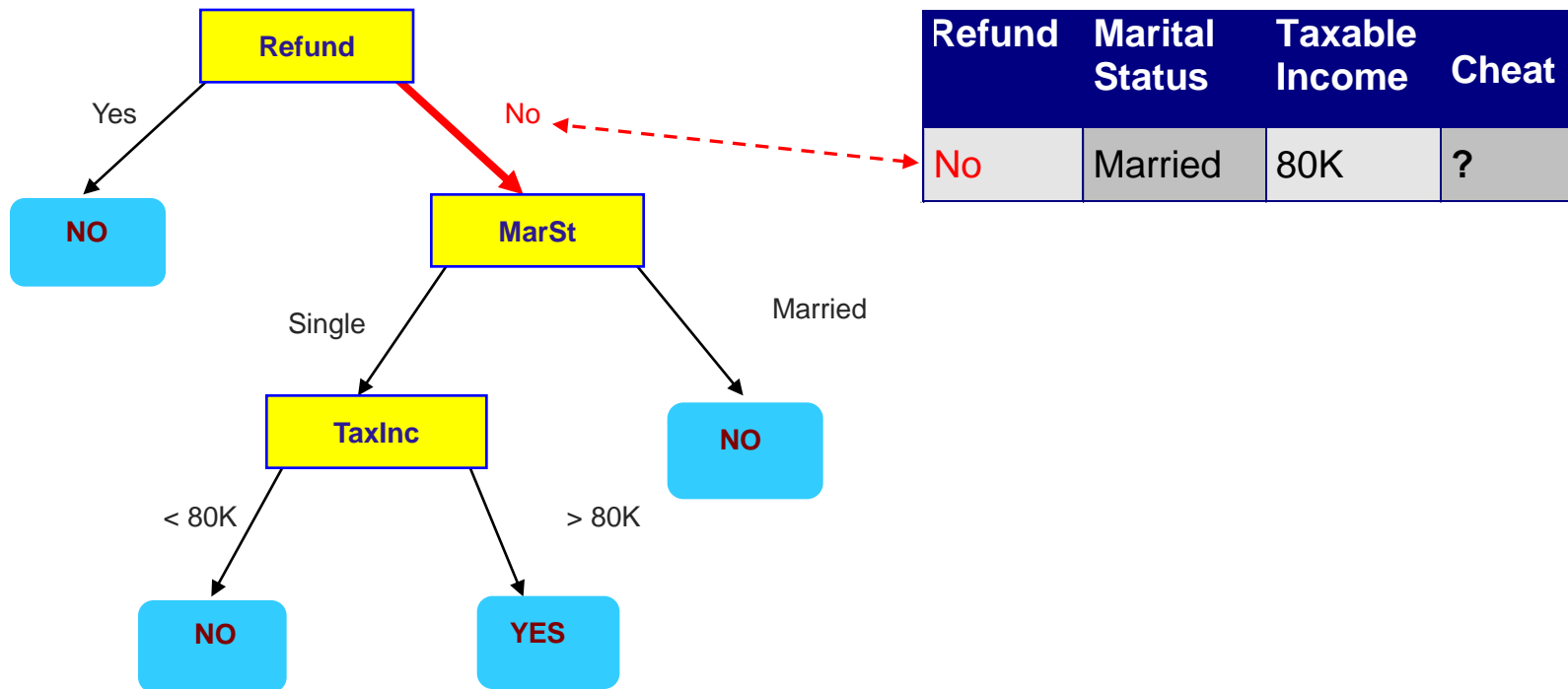
New Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

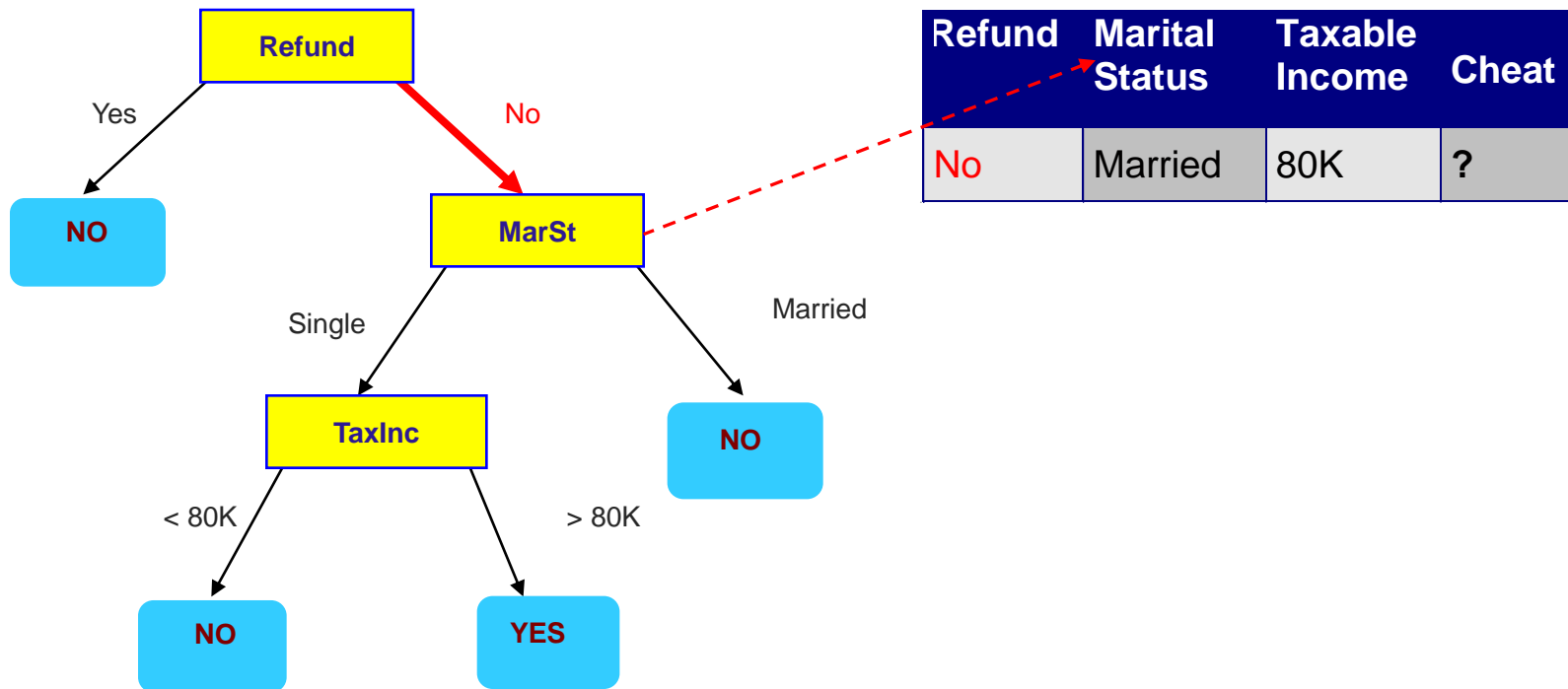
Model Use



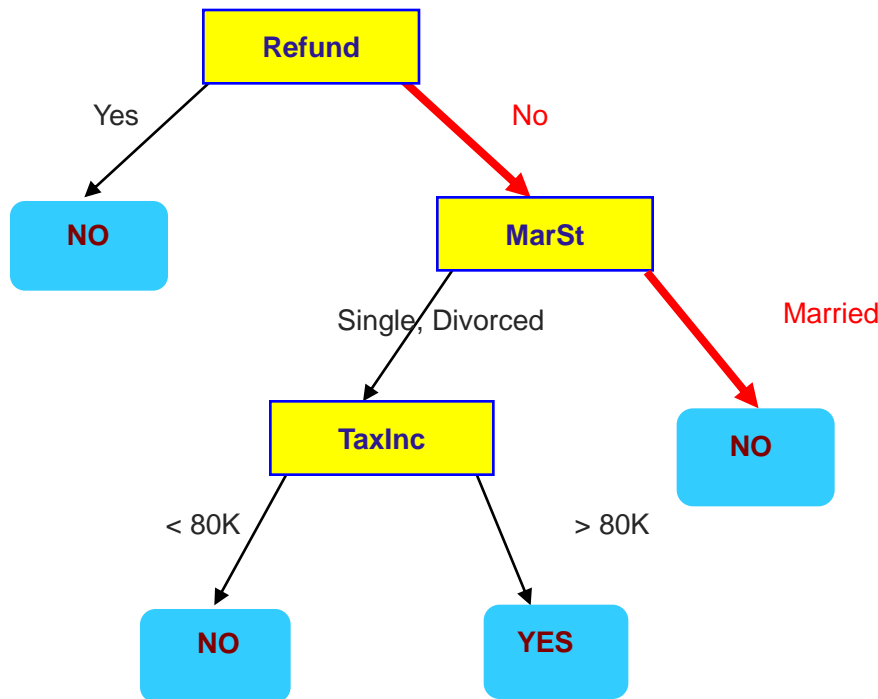
Model Use



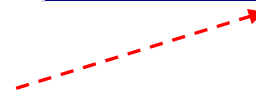
Model Use



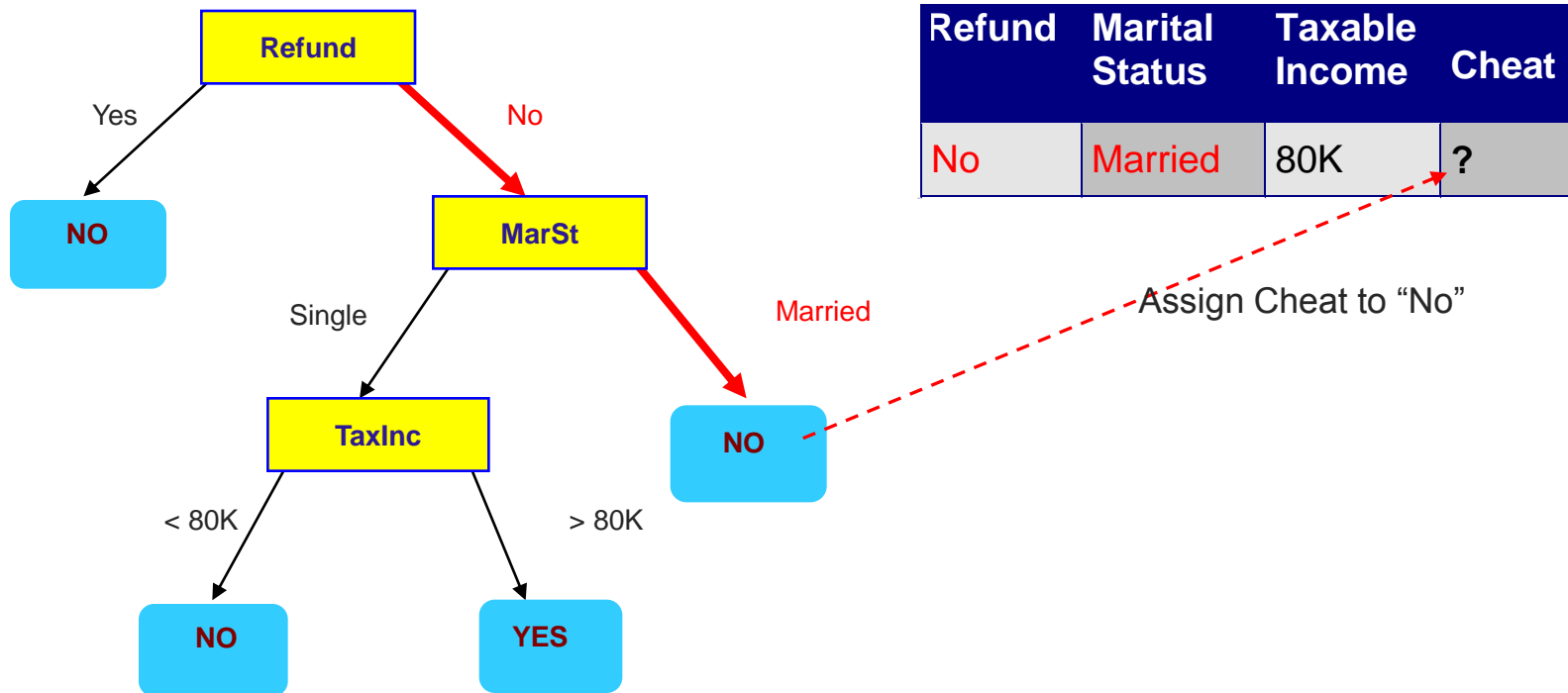
Model Use



Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Model Use



Tree Construction

Basic Steps in Tree Construction

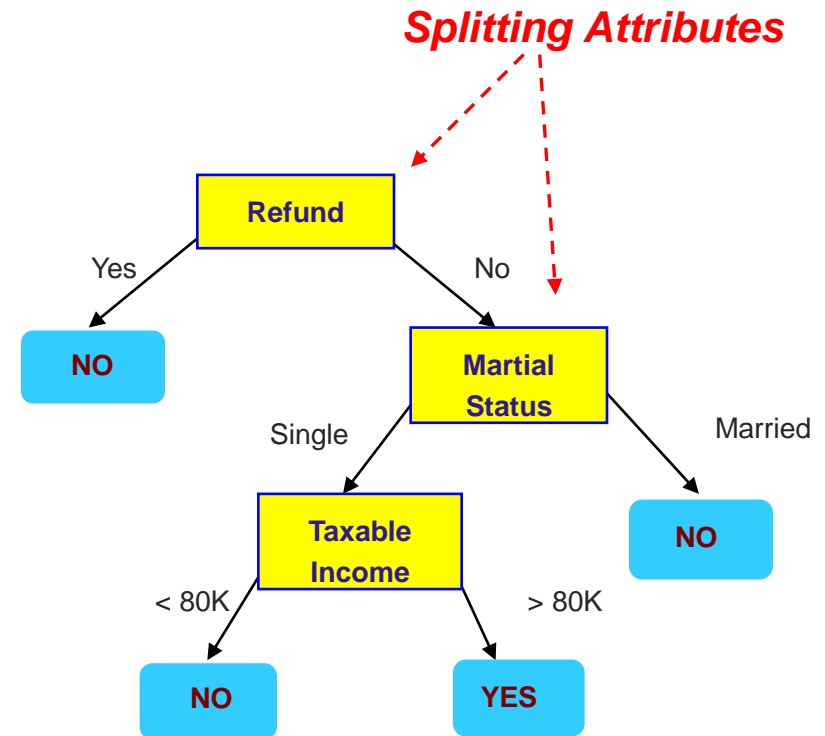
- Tree starts a single node
- Selects an attribute that best splits the instances at every stage
- Stops when
 - All (or nearly all) of the instances at the node have the same class.
 - There are no remaining attributes to distinguish among the instances.
 - The tree has grown to a pre-defined size limit.
- Also called recursive partitioning
 - The decision tree algorithm works by splitting the dataset recursively.
 - The subsets that arise from a split are further split until a pre-defined termination criterion is reached.

Select Attributes to Split

categorical categorical continuous class

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Single	95K	Yes
6	No	Married	60K	No
7	Yes	Single	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Attribute Selection

- We want to find the most “**useful**” attribute in classifying a sample.
- **Purity**: the degree to which a subset of instances contains only a single class
 - If every instance of a node has the same value (class) for the target, then the node is pure
- Need a measure of node **impurity**
 - Gini Index
 - Entropy

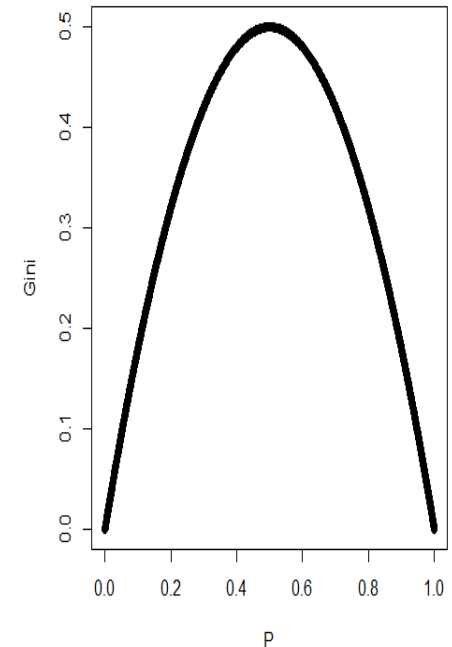
GINI Index

```
> P <- seq(0,1,0.0001)
> Gini <- 1-(P^2 + (1-P)^2)
> plot(P,Gini)
```

- Gini Index for a given node:

$$GINI = 1 - \sum_j [p(j)]^2$$

- $p(j)$ is the relative frequency of class j
 - The proportion of instances into class j



Three candidate variables

Gender	Car ownership	IncomeLevel	Transportation mode
Female	0	Low	Bus
Male	0	Low	Bus
Male	1	Medium	Bus
Male	1	Medium	Bus
Female	1	Medium	Train



Gender	Transportation mode
Female	Bus
Male	Bus
Male	Bus
Male	Bus
Female	Train



Car ownership	Transportation mode
0	Bus
0	Bus
1	Bus
1	Bus
1	Train



IncomeLevel	Transportation mode
Low	Bus
Low	Bus
Medium	Bus
Medium	Bus
Medium	Train

Gini Index

$$p(j) = p(\text{Bus}) = 4/5 = 0.8$$

$$p(j) = p(\text{Train}) = 1/5 = 0.2$$

$$\text{Gini} = 1 - 0.8^2 - 0.2^2 = 0.32$$

Transportation mode
Bus
Bus
Bus
Bus
Train

- Female

$$p(j|t) = p(\text{Bus} / \text{Female}) = 1 / 2 = 0.5$$

$$p(j|t) = p(\text{Train} / \text{Female}) = 1 / 2 = 0.5$$

$$\text{Gini}(\text{Female}) = 1 - 0.5^2 - 0.5^2 = 0.5$$

- Male

$$p(j|t) = p(\text{Bus} / \text{Male}) = 3 / 3 = 1$$

$$p(j|t) = p(\text{Train} / \text{Male}) = 0 / 3 = 0$$

$$\text{Gini}(\text{Male}) = 1 - 1^2 - 0^2 = 0$$

Gender	Transportation mode
Female	Bus
Male	Bus
Male	Bus
Male	Bus
Female	Train

Gini Index After Splitting

- When a node p is split into k partitions (children), the quality of split is computed as,

$$GINI_p = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

- n_i = number of records at child i
- n = number of records at node p

$$GINI_p = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

- Gini(Gender)

- number of records at child female: 2
- number of records at child male: 3

Gender	Transportation mode
Female	Bus
Male	Bus
Male	Bus
Male	Bus
Female	Train

$$Gini(Gender) = 2/5 * Gini(Female) + 3/5 * Gini(Male)$$

$$Gini(Gender) = 2/5 * 0.5 + 3/5 * 0 = 0.2$$

$$Gini(Car\ ownership) = 0.27$$

$$Gini(IncomeLevel) = 0.27$$

How to Find the Best Split

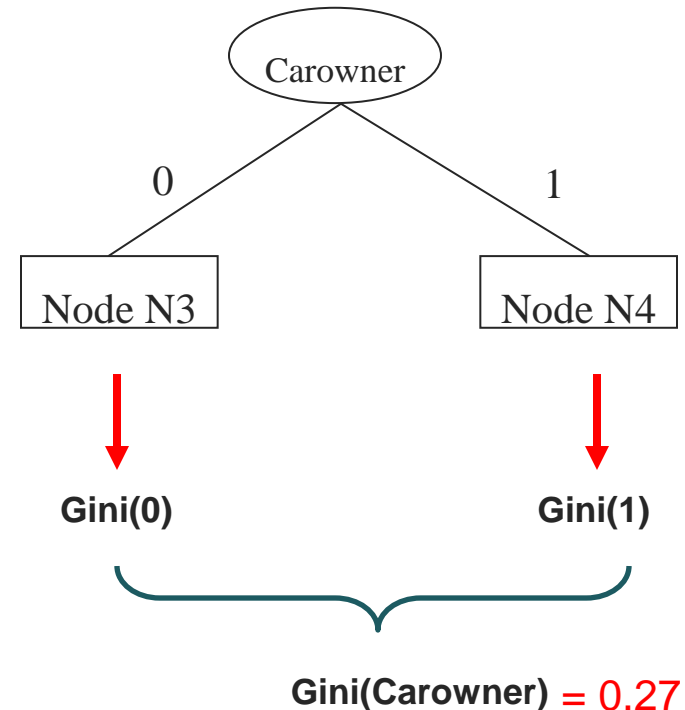
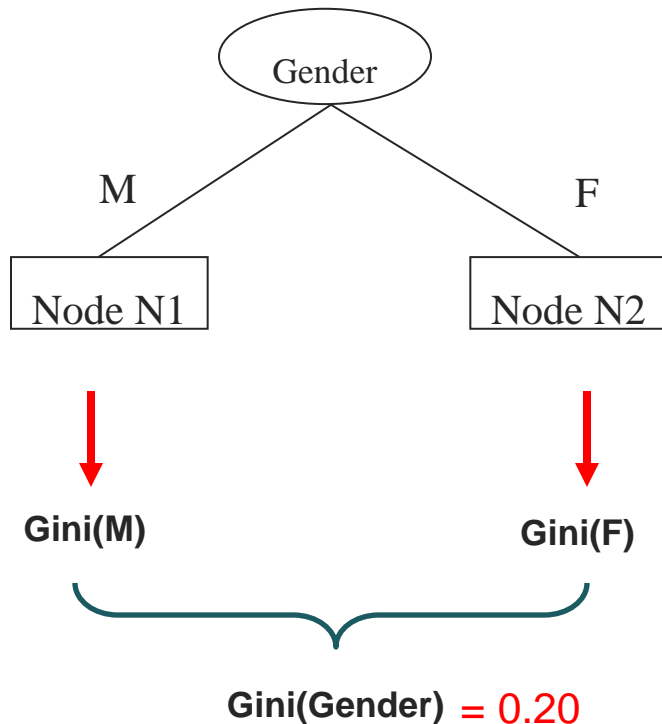
- Compare the degree of impurity of the parent node (before splitting) with the degree of impurity of the child nodes (after splitting).
- The larger their difference, the better the test condition.
- Decision Tree algorithms often choose an attribute which maximizes the gain

How to Find the Best Split

Before Splitting:

Bus
Train

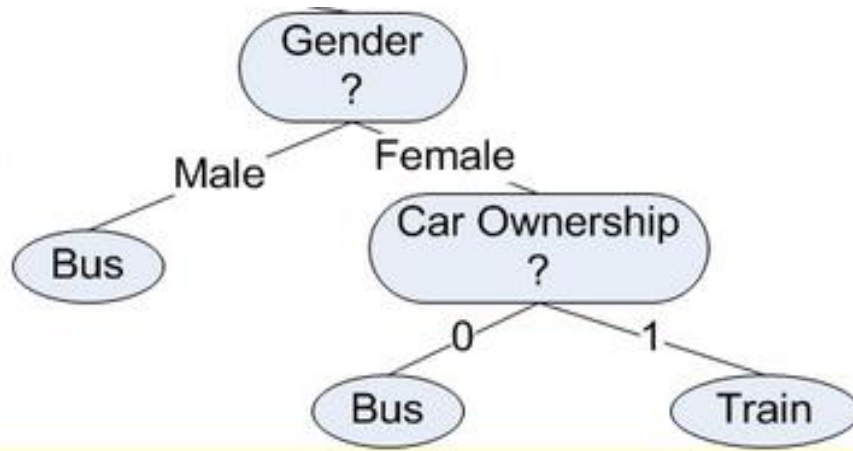
 \longrightarrow $\text{Gini(Before)} = 0.32$



Gain = Gini (Before) – Gini(Gender) vs. **Gain** = Gini(Before)– Gini(Carowner)

$$0.12 = (0.32 - 0.20) > 0.05 = (0.32 - 0.27)$$

The Constructed Tree



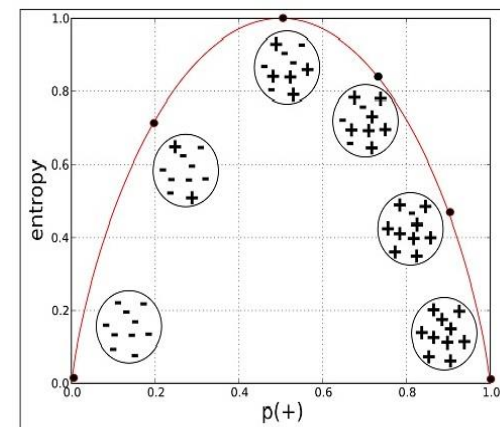
Gender	Car ownership	IncomeLevel	Transportation mode
Female	0	Low	Bus
Male	0	Low	Bus
Male	1	Medium	Bus
Male	1	Medium	Bus
Female	1	Medium	Train

Information Gain

- Entropy at a given node t

$$Entropy(t) = - \sum_j p(j | t) \log p(j | t)$$

- $p(j | t)$ is the relative frequency of class j at node t



- Information Gain:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^k \frac{n_i}{n} Entropy(i) \right)$$

- Measures Reduction in Entropy achieved because of the split. Choose the split that achieves most reduction (maximizes GAIN)

Summary of Tree Generation

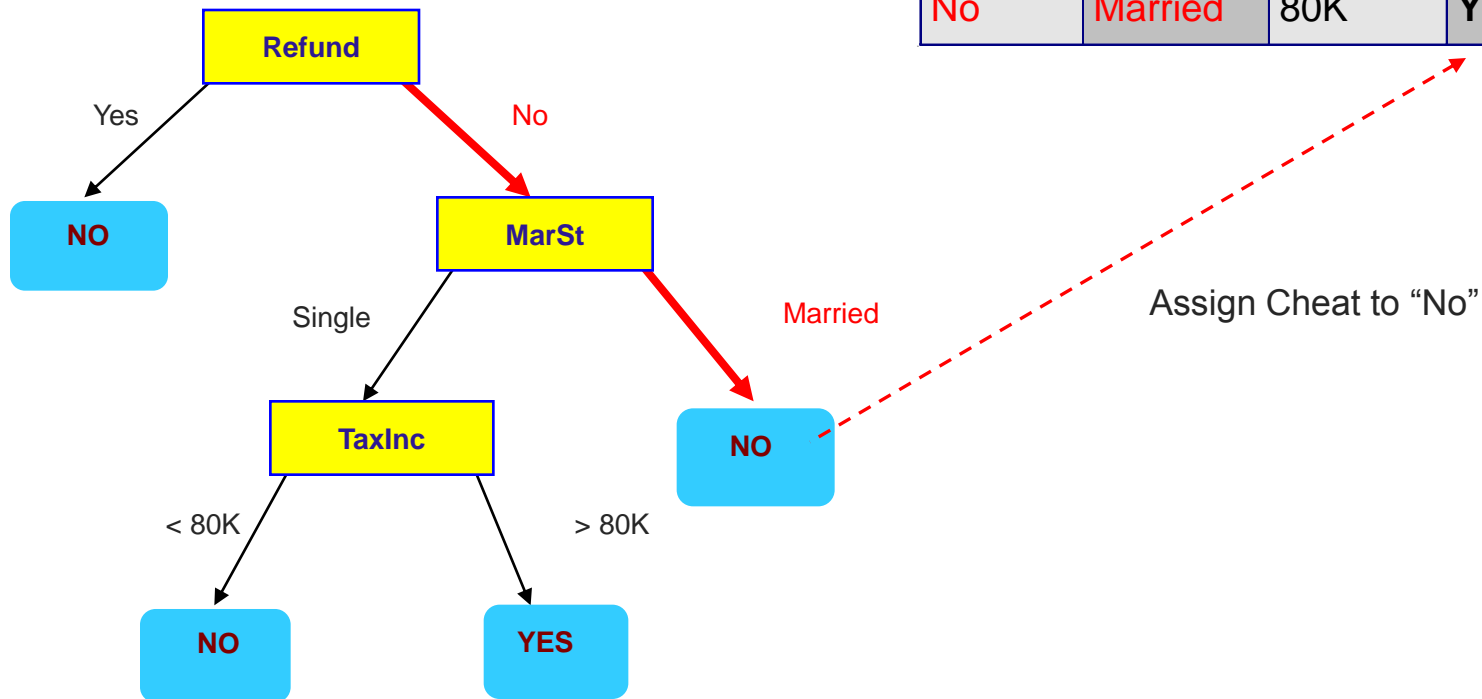
- Choice of impurity measure has little effect on the performance of tree construction
 - Usually, impurity measures are quite consistent with each other.
- The strategy used to **prune** the tree has a greater impact on the final tree



Tree Pruning

Apply Model to an Observation

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	Yes

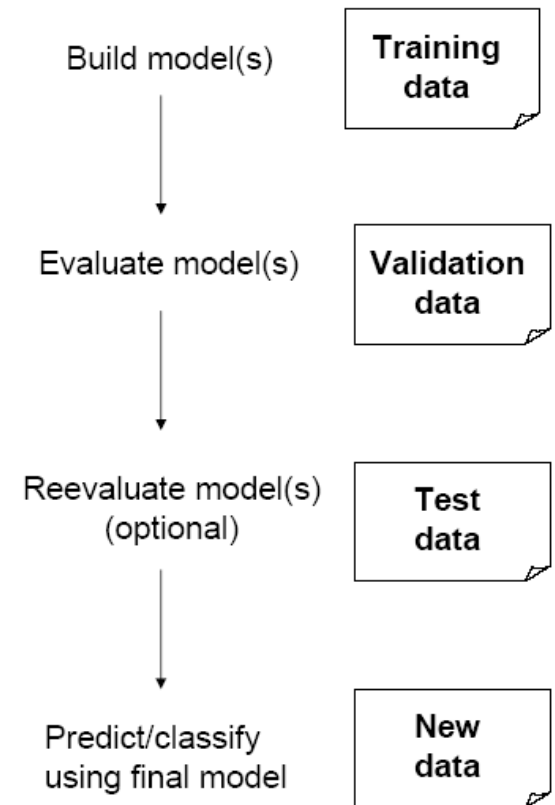


Accuracy

- The known labels of instances are compared with the predicted class from model
 - Same: correct prediction
 - Different: incorrect prediction
- Accuracy of the model
 - Ratio of correct predictions
- Evaluating Accuracy

Partitioning the Data

- **Training data:** develop the model
- **Validation data:** a set of examples used to evaluate performances or tune the parameters of a model
- **Testing data:** a set of examples used only to assess the performance of a fully-trained classifier



Overfitting

- A good classification model must not only fit the training data well, it must also accurately classify records it has never seen before.
- Overfitting leads to low predictive accuracy of new data
 - Perform very well in the training data
 - Perform very poorly in the testing data
- Consequence: Deployed model will not work as well as expected with completely new data.

Two Approaches to Avoid Overfitting

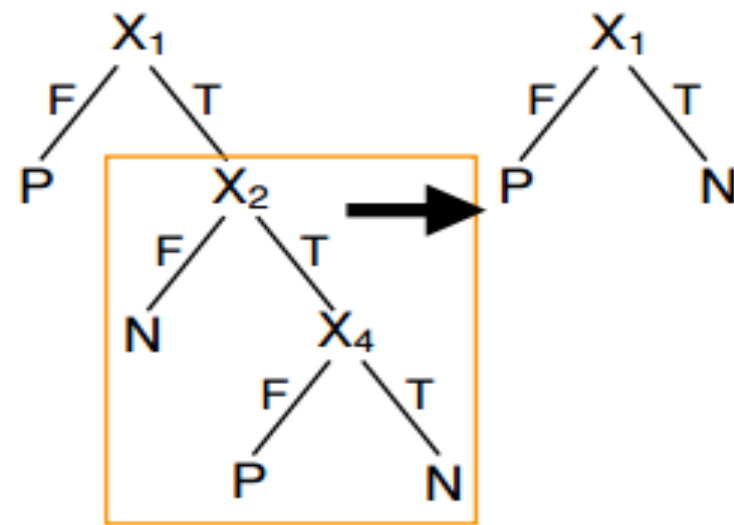
- Pruning
 - Reduces the size of a decision tree such that it generalizes better to unseen data
- Two Approaches
 - Pre-pruning
 - Post-pruning

Pre-pruning

- Halt tree construction early
 - Stop splitting when the observed gain in impurity falls below a certain threshold
 - It is difficult to choose appropriate threshold
 - Splitting stops when purity improvement is not statistically significant
- The tree avoids doing needless work, but there is no way to know whether the tree will miss subtle, but important patterns that it would have learned had it grown to a larger size.

Post-pruning

- Grow complete tree and prune it back
- Consider each node:
 - Evaluate performance with validation data
 - s : performance with subtree at node
 - t : performance with subtree replaced with leaf
 - If $s \leq t$, prune subtree



Model Evaluation Metrics

inst#,	actual,	predicted,	error,	probability	dis
1	1:A	1:A		*0.868	0.132
2	1:A	1:A		*0.868	0.132
3	1:A	1:A		*0.868	0.132
4	1:A	1:A		*0.811	0.189
5	1:A	1:A		*0.868	0.132
6	1:A	1:A		*0.868	0.132
7	1:A	2:B	+	0.176	*0.824
8	1:A	2:B	+	0.143	*0.857
9	1:A	1:A		*0.868	0.132
10	1:A	1:A		*0.868	0.132
11	1:A	1:A		*0.868	0.132
12	1:A	1:A		*0.789	0.211
13	1:A	2:B	+	0.176	*0.824
14	1:A	1:A		*0.868	0.132
15	1:A	1:A		*0.868	0.132
16	1:A	2:B	+	0.25	*0.75
17	1:A	1:A		*0.868	0.132
18	1:A	1:A		*0.811	0.189
19	1:A	1:A		*0.868	0.132
20	1:A	1:A		*0.868	0.132
21	1:A	1:A		*0.868	0.132
22	1:A	1:A		*0.868	0.132
23	1:A	1:A		*0.789	0.211

Metrics for Performance Evaluation

- Confusion Matrix:

	PREDICTED CLASS		
		Class=1	Class=0
	Class=1	a	b
	Class=0	c	d

Confusion Matrix

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

201 1's correctly classified as "1"

85 1's incorrectly classified as "0"

25 0's incorrectly classified as "1"

2689 0's correctly classified as "0"

Metrics for Performance Evaluation

ACTUAL CLASS	PREDICTED CLASS		
		Class=1	Class=0
	Class=1	a	b
	Class=0	c	d

$$\text{Accuracy} = \frac{a + d}{a + b + c + d}$$

Accuracy and Error Rate

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

$$\text{Accuracy} = (201+2689)/3000 = 96.33\%$$

$$\text{Overall error rate} = 1 - \text{Accuracy} = (25+85)/3000 = 3.67\%$$

Limitation of Accuracy

- Consider a 2-class problem
 - Number of Class 0 examples = 9990
 - Number of Class 1 examples = 10

ACTUAL CLASS	PREDICTED CLASS		
		Class=1	Class=0
	Class=1	0	10
	Class=0	0	9990

- If a model predicts everything to be class 0, accuracy is $9990/10000 = 99.9\%$
 - Accuracy is misleading because model does not detect any class 1 example

When One Class is More Important

- In many cases it is more important to identify members of one class
 - Tax fraud (**cheat** or not cheat)
 - Credit default (**default** or not default)
 - Response to a promotional offer (**respond** or not respond)
 - Detecting electronic network intrusion (**intrude** or not intrude)
 - Predicting delayed flights (**delayed** or on time)
- In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention.

Alternate Accuracy Measures





ACTUAL CLASS	PREDICTED CLASS		
		Class=1 (Important)	Class=0 (Less Important)
	Class=1 (Important)	a (TP)	b (FN)
	Class=0 (Less Important)	c (FP)	d (TN)

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

		PREDICTED CLASS	
ACTUAL CLASS		Tested Positive (Class=1)	Tested Negative (Class=0)
	Has Disease (Class=1)	 True Positive Sensitivity	 FALSE Negative
	Doesn't Have Disease (Class=0)	 FALSE Positive	 True Negative Specificity

Sensitivity (True Positive Rate) = 1: All sick people are tested as being sick

Specificity (True Negative Rate) = 1: All health people have a negative test

Alternate Accuracy Measures

If “ C_1 ” is the important class,

Sensitivity = % of “ C_1 ” class correctly classified

$$\text{Sensitivity} = \frac{a}{a + b} = \frac{TP}{TP + FN}$$

- Accuracy in classifying the important class correctly

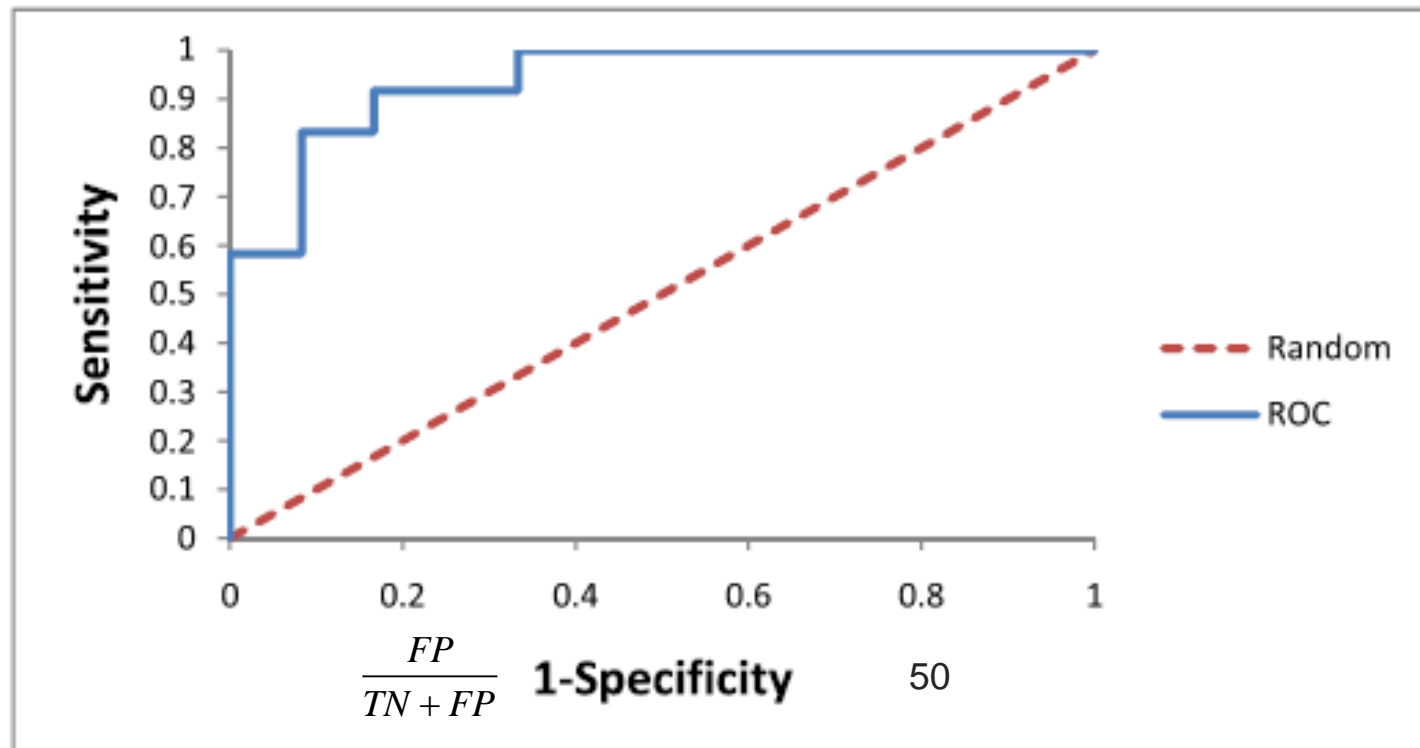
Specificity = % of “ C_0 ” class correctly classified

$$\text{Specificity} = \frac{d}{d + c} = \frac{TN}{TN + FP}$$

- Accuracy in classifying the less important class correctly

ROC Curve

- Receiver Operating Characteristic
- The closer to upper left corner the better



Lift Charts: How to Compute

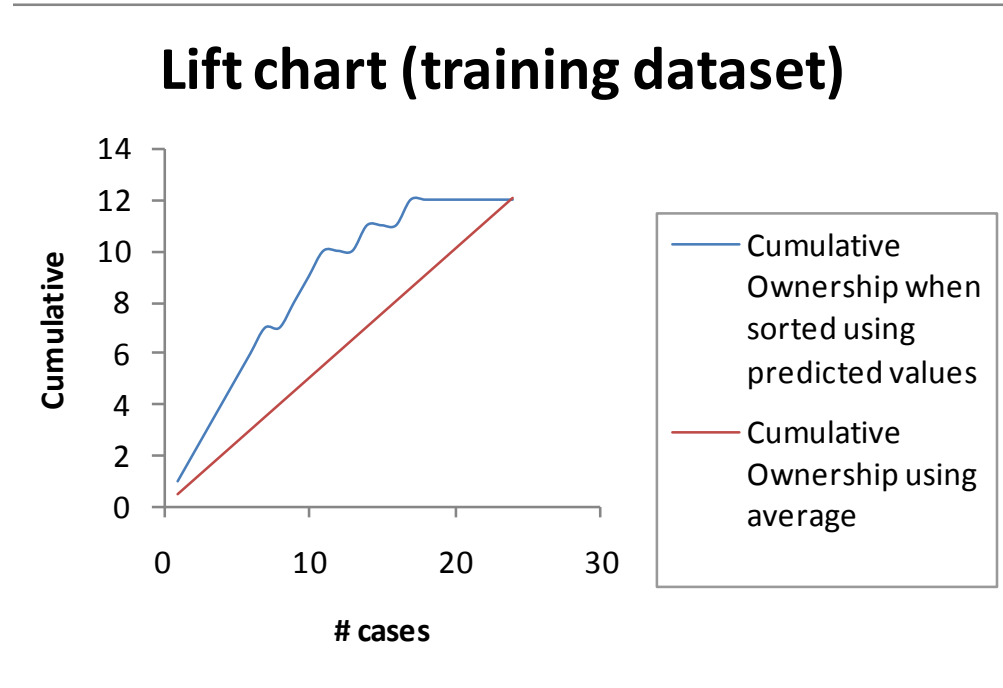
- Using the model's classifications, sort records from most likely to least likely members of the important class
- Compute lift: Accumulate the correctly classified “important class” records (Y-axis) and compare to number of total records (X-axis)

Lift Chart

- Compare performance of DM model to “no model, pick randomly”
 - The diagonal line is the curve for random samples without using sorted data
 - Any good selection will keep the lift curve above the diagonal
 - In the lift chart, we will like to stay towards the upper left corner

Lift Chart – Cumulative Performance

Serial no.	Predicted prob of 1	Actual Class	Cumulative Actual class
1	0.995976726	1	1
2	0.987533139	1	2
3	0.984456382	1	3
4	0.980439587	1	4
5	0.948110638	1	5
6	0.889297203	1	6
7	0.847631864	1	7
8	0.762806287	0	7
9	0.706991915	1	8
10	0.680754087	1	9
11	0.656343749	1	10
12	0.622419543	0	10
13	0.505506928	1	11
14	0.47134045	0	11
15	0.337117362	0	11
16	0.21796781	1	12
17	0.199240432	0	12
18	0.149482655	0	12
19	0.047962588	0	12
20	0.038341401	0	12
21	0.024850999	0	12
22	0.021806029	0	12
23	0.016129906	0	12
24	0.003559986	0	12



After examining (e.g.,) 10 cases (x-axis), 9 owners (y-axis) have been correctly classified

Lift Chart: Goal

- Useful for assessing performance in terms of identifying the most important class
- Helps evaluate, e.g.,
 - How many tax records to examine
 - How many loans to grant
 - How many customers to mail offer to