



Data Science and Business Intelligence

BU.330.780

Session 3

Instructor: Changmi Jung, Ph.D.



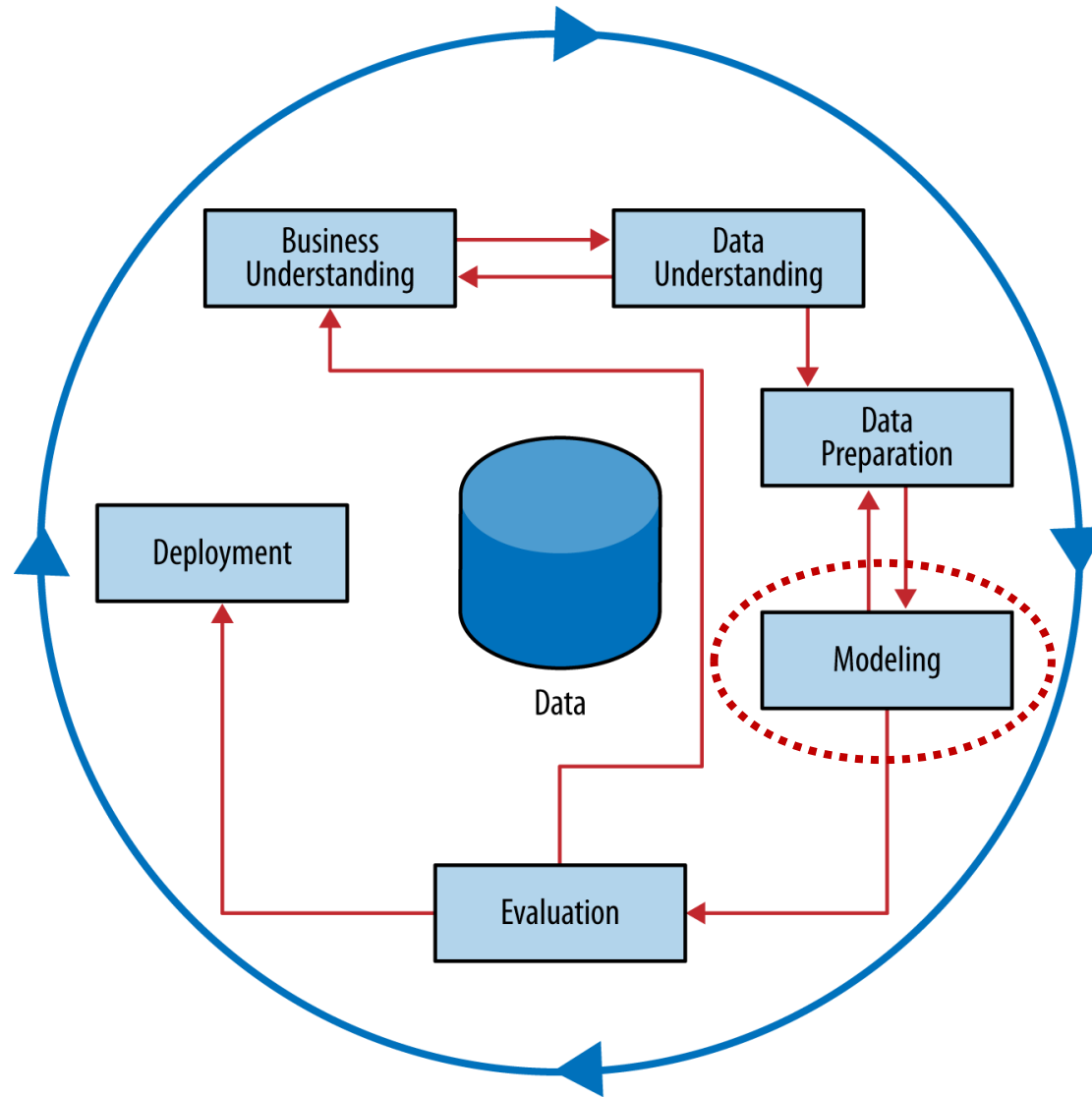
Announcement

» Assignment #2 is due next week

» Quiz #1

- Onsite test: closed-book, closed-notes
- Covers content from Week 1 through Week 3
- Format: around < 20 multiple choice questions (including True/False)
- Administered via Respondus Lockdown Browser
- Logistics: After checking your network, we will take the quiz for the first 35 - 45 minutes of class time. After submitting the quiz, you may leave the classroom and return to join the class by the specified time.
- Sample questions: Modules > Week 4 > “Quiz 1 Study Guide and Sample Questions.pdf”

Data science as a process





Today's agenda

- » Introduction to classification (a.k.a. supervised segmentation)
- » Classification tree model ✓
- » Classification tree with R ✓
- » Probability estimation → next week
- » Random forest with R → next week



Supervised Segmentation : Checkpoint

A classification example: Credit card default



- » We want to detect customers likely to default before they actually default on their credit card balances.
- » What would be a target for the credit card default problem?
- » How should we proceed?





Key 1: What are we predicting? What is the target?

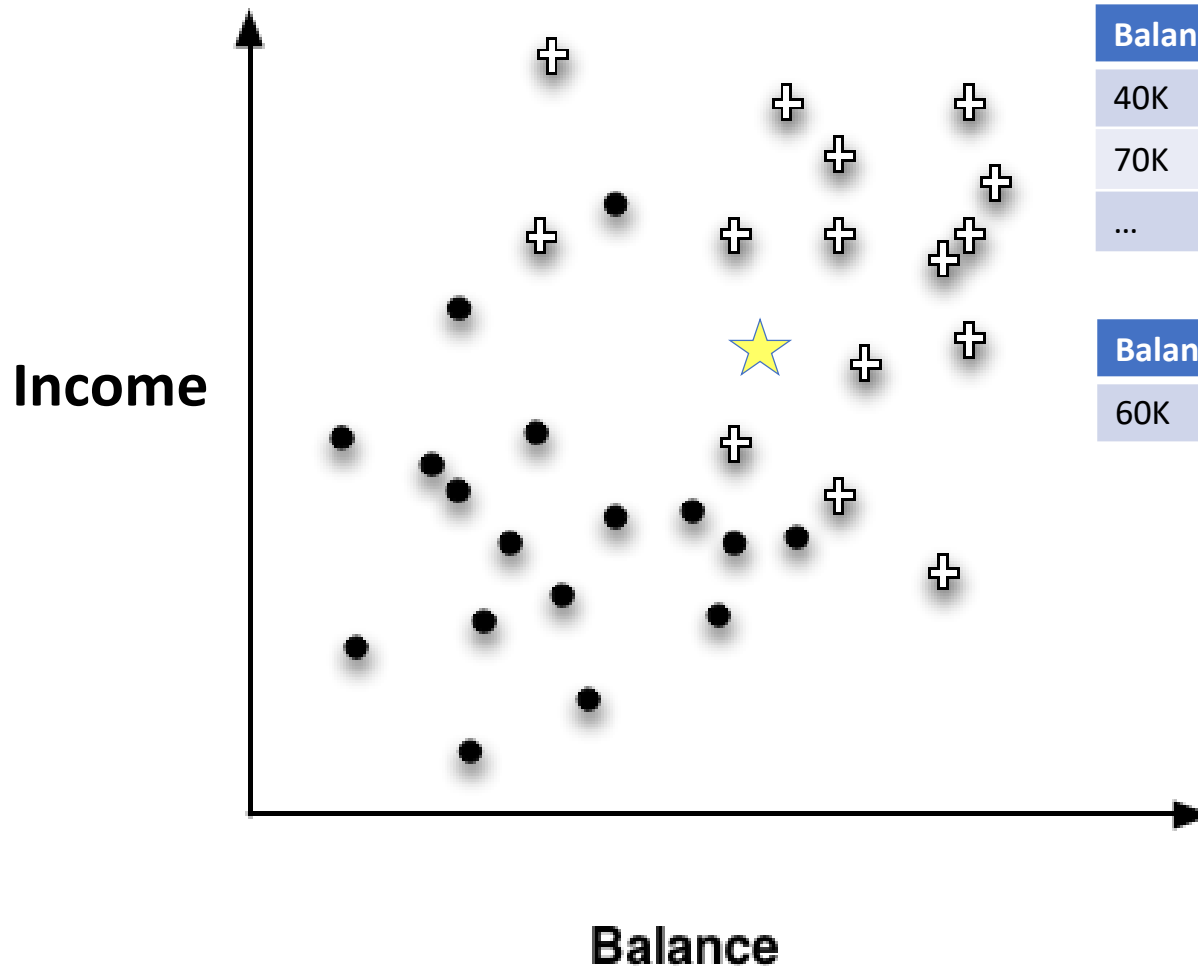
» Is there a specific, quantifiable target we are interested in or trying to predict?

» Examples:

- What will be the Google stock price tomorrow? (o) ✓
- Will this prospect default on her loan? (o)
- Do my customers naturally fall into different groups? (x)
 - No objective target → Unsupervised



What are we predicting?



Balance	Income	Default?
40K	500K	No
70K	600K	Yes
...

Balance	Income	Default?
60K	500K	?

● Not Default – 17 cases
+ Default – 15 cases



Key 2: Do we have good (enough) data?

» Checkpoint 2: do we have data on this target? Do we have the same/similar **phenomenon** and **context**?

What? Data on the target variable?



Age	Student	Income	Balance	Default?
40	No	44K	3K	No
32	Yes	12K	1.5K	No
27	No	31K	1K	Yes
51	No	35K	0.5K	No



Classical Pitfalls

- » We don't need the exact data on the target variable (e.g., whether the new customers' future default status)

I can't have them unless I'm a seer.



- » But we need data for **the same or a related phenomenon** (e.g., the default status of customers from last quarter) because we will use this data to build a model to predict the phenomenon of interest. → check if the 'phenomenon' is the same/similar
- » The training data (a dataset used to build a model) should be **as similar as possible to the USE data**; collect the training data from the context that is as similar as possible to the context where the model will be applied → check the **context**

Q: Classic Pitfalls



- » Is the phenomenon of “who did not pay off the debt” last quarter the same as the phenomenon of “who will not pay off the debt” this quarter?
- » Who might buy this completely new type of product we have never sold before? Let’s predict it by using our current customer data.



Bad Example: “Survivorship issues”

Lending Club wants to develop a model to automate the initial loan application screening process. The goal is for the model to identify and reject applications that are likely to default.

The company has provided historical data on past loans and their outcomes. They plan to use this dataset to train the model.

»» Is it a good practice?

»» Why?

Bad Example: Different Sources



- » Things go really bad if the positive (treated) and negative cases (control) are sourced differently
- » Example: Looking for drivers of diabetes. How do you assemble the training data?
- » Bad practice:
 - Go to a specialized hospital and get records from people treated for diabetes
 - Go somewhere else (swimming club) to get records for healthy people
- » Why?
 - The sick people came from a very artificial subset



Bad Example: “Looking under the streetlight”

» Target Proxy

- I do not see if a person bought the book after seeing an ad, so let's model with the number of clicks...

» Sample Proxy

- I want to run a campaign in Spain but only have data on US customers

» Why bad?

- Your target is NOT really your target!
- No way of telling how the model will perform
- No way of testing either when the data itself has issues

Q: Plumbing Inc.



» Plumbing Inc. has been selling plumbing supplies for the past 20 years. The owner, Joe, decides that it is time to diversify by adding gardening tools to the products. Having had success using customer data to build predictive models to guide direct mail campaigns for special plumbing offers, he considers that data mining could help him to identify a subset of customers who should be good prospects for his new set of products. What would you suggest as the target variable? Is Joe ready to solve this as a supervised learning problem?

It is important to ensure the data quality and also, understand the limitation of the data, so that we can avoid bad practices of modeling based on bad data.

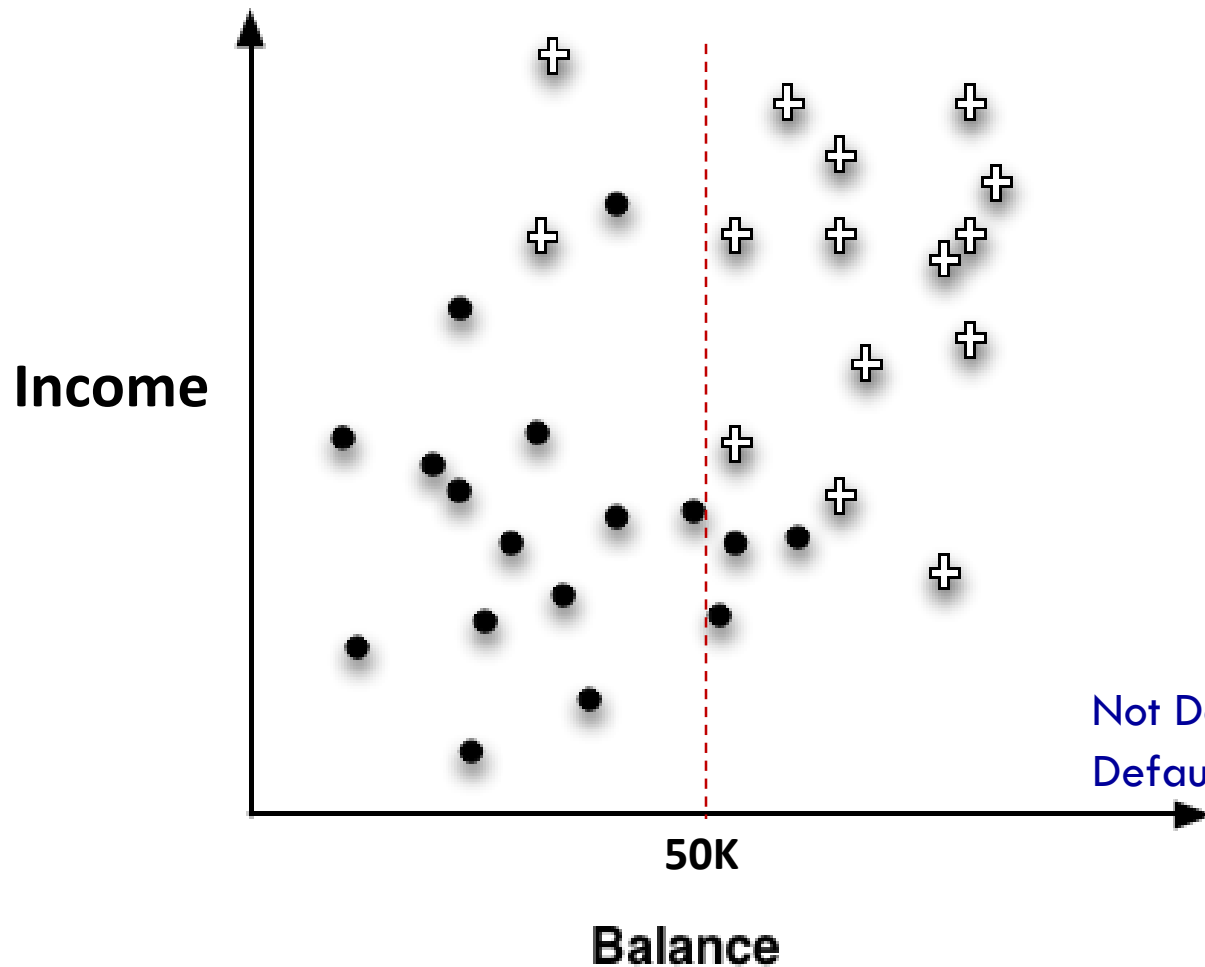


Key 3: What is a model?

- » The result of supervised data mining is a MODEL predicting some quantity of given data
- » A model is **a formula for estimating the unknown value of interest**, the target.
- » There are different sorts of classification models. Here are just two examples:
 - Tree/Rule
 - if** (income < 50K)
 - then** Default
 - else** no Default
 - Numeric function
 - $P(\text{Default} | x) = f(x_1, x_2, \dots, x_k)$

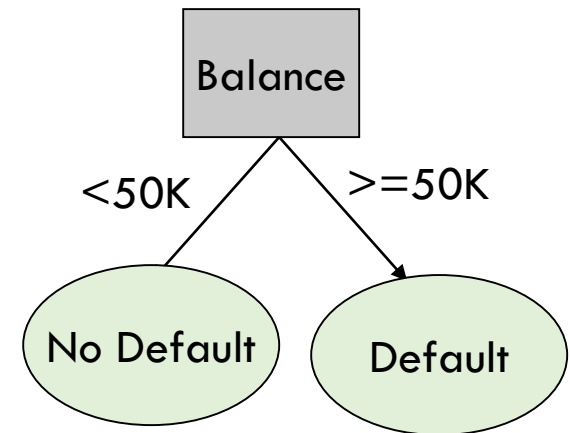


Classification Model Examples: Tree-based models



● Not Default – 17 cases
+ Default – 15 cases

Not Default – 17 cases
Default – 15 cases

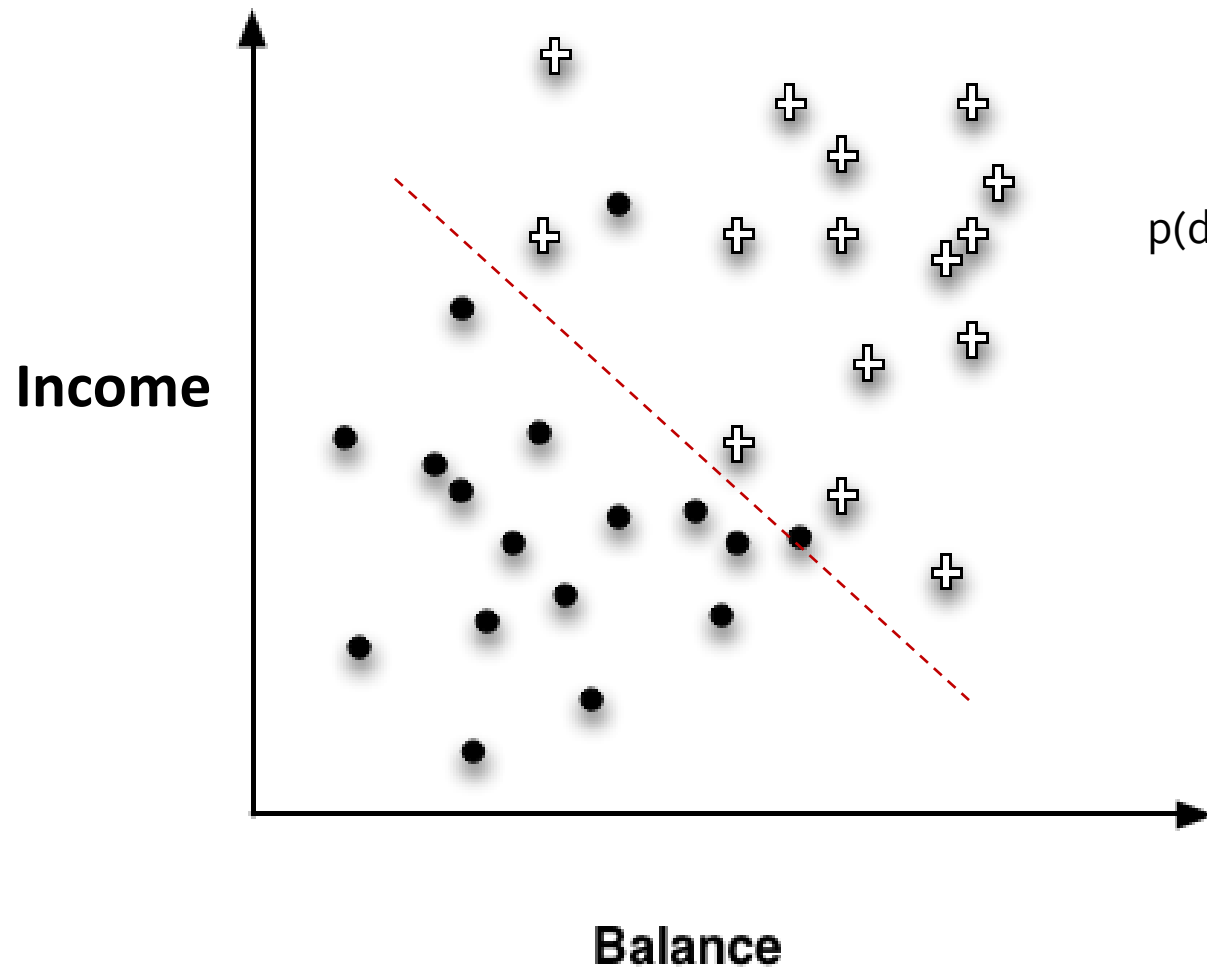


Not Default 14
Default 2

3
13



Classification Model Examples: Linear models



Logistic Regression

$$p(\text{default}|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

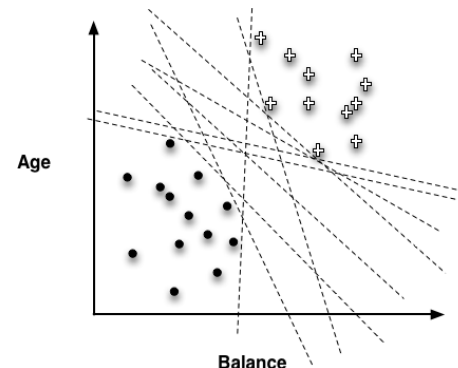
$$\begin{aligned}\beta_0 &= 123 \\ \beta_1 &= -1.3\end{aligned}$$

We will talk about this next week

Key Terminologies

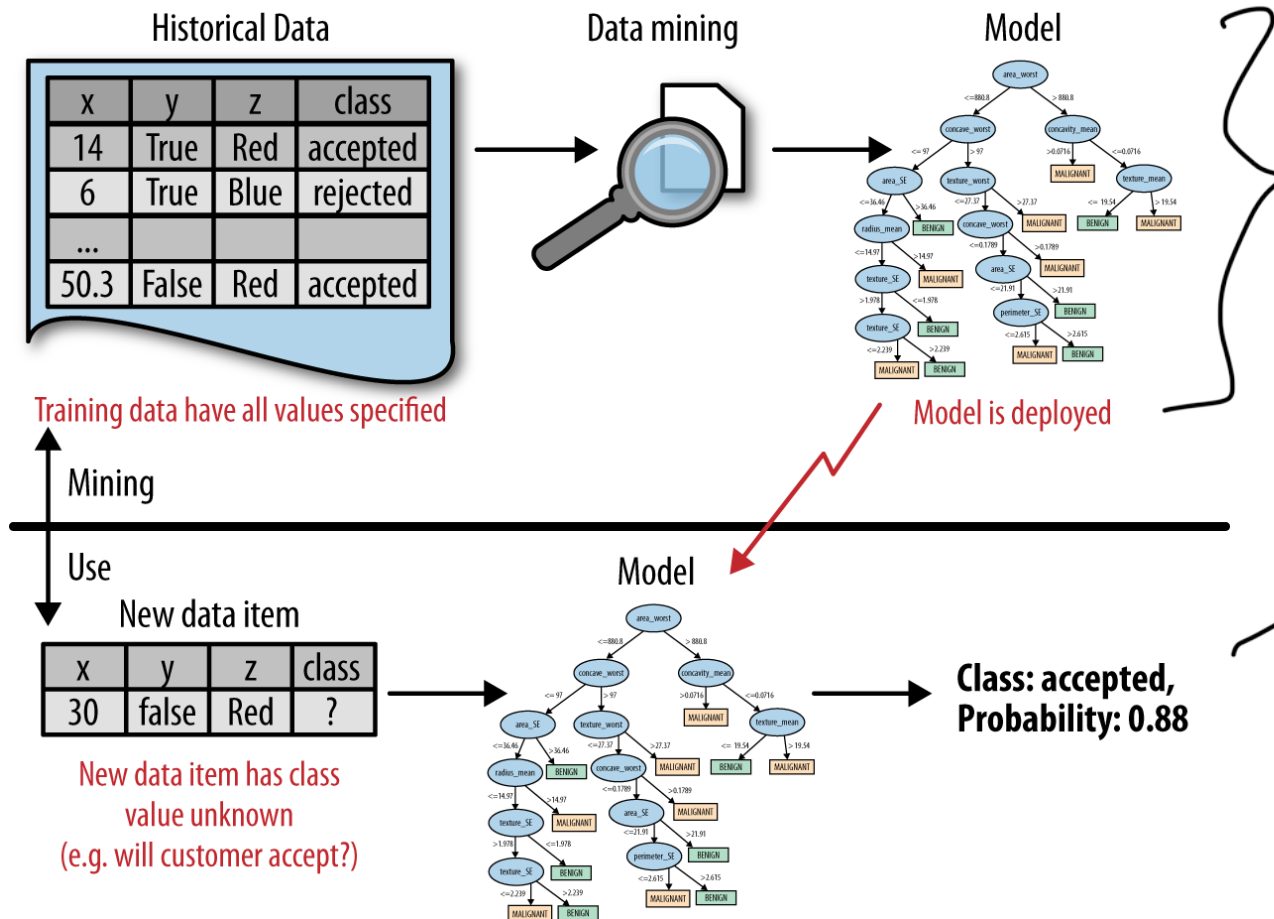


- » **Classification**: learning to predict categories
- » **Decision boundary**: the surface separating different predicted classes
- » **Linear classifier**: a classifier that learns linear decision boundaries
- » **Linearly separable**: a data set can be perfectly explained (separated) by a linear classifier



Key 4: What is it for? To understand or to predict

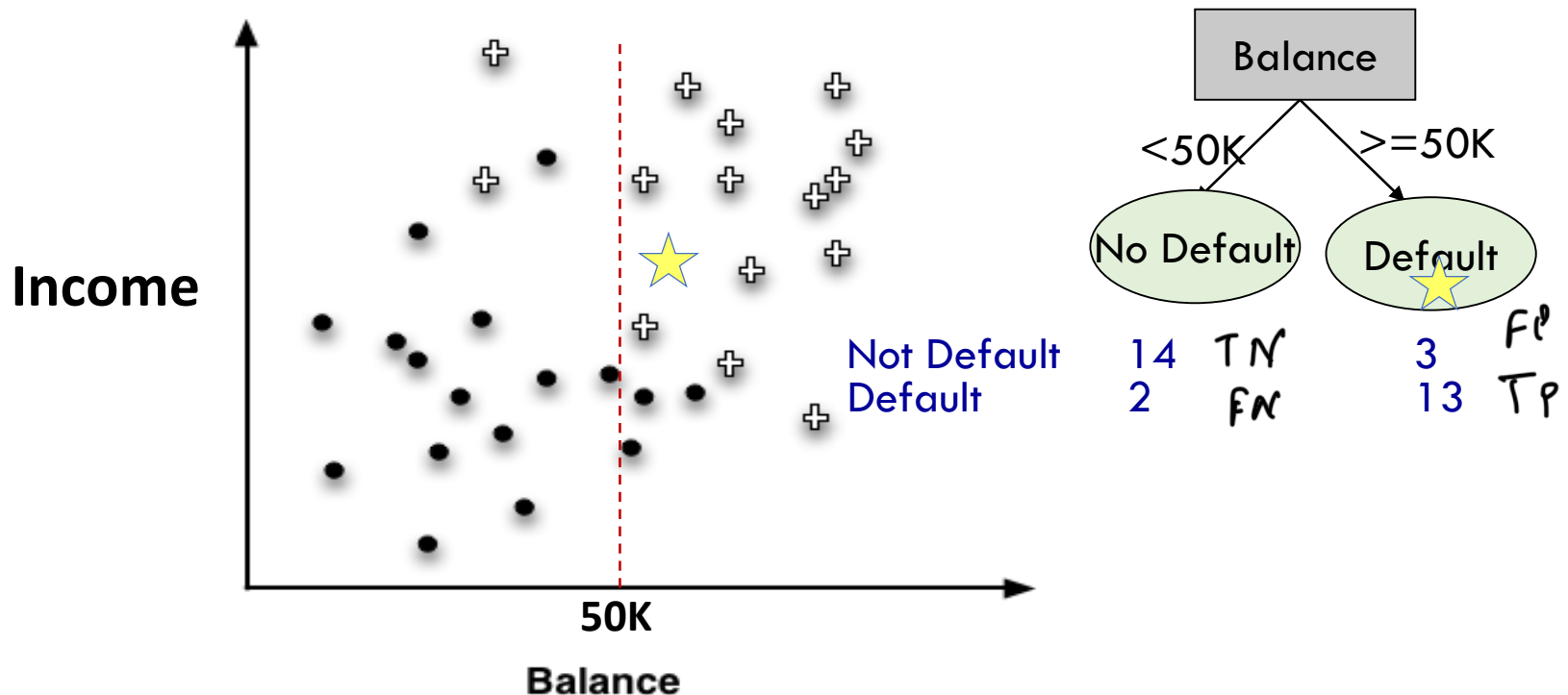
» A data-driven model can either be used to **understand** or to **predict**





USE of models - to understand or to predict

» Result of Supervised: you can apply this rule to any customer and it gives you prediction



● Not Default – 17 cases
+ Default – 15 cases

Likely to default? YES



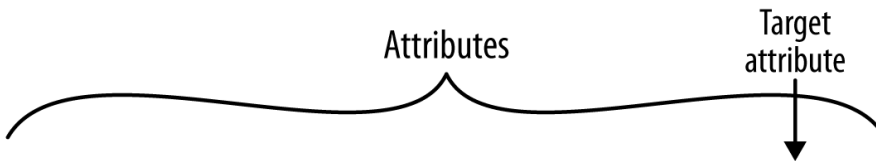
Tree-based Models (Induction Models)

- Classification Trees
- Random Forest

Terminology: Induction

» Induction (a.k.a. learning, inductive learning, model induction) —

- A process by which a pattern/model is generalized from factual data —
- A method or algorithm used to **generalize** a model or pattern **from a set of examples**



Name	Balance	Age	Employed	Write-off
Mike	\$200,000	42	no	yes
Mary	\$35,000	33	yes	no
Claudio	\$115,000	40	no	no
Robert	\$29,000	23	yes	yes
Dora	\$72,000	31	no	no



Learner:

Induces a model
from examples



If Balance \geq 50K and Age $>$ 45
Then Default = 'no'
Else Default = 'yes'

Pattern/Model?



Attributes

Target attribute

Name	Balance	Age	Employed	Write-off
Mike	\$200,000	42	no	yes
Mary	\$35,000	33	yes	no
Claudio	\$115,000	40	no	no
Robert	\$29,000	23	yes	yes
Dora	\$72,000	31	no	no

Good vs bad patterns?

Pattern 5:

If **Balance** \geq 50K and **Age** $>$ 45
Then **Default** = 'no'
Else **Default** = 'yes'

Pattern 1:

If **Names** starts with M
Then **Default** = 'yes'
Else **Default** = 'no'

Pattern 2:

Age is inversely proportional to the
alphabetical order of the name }

Pattern 3:

Young people are more likely to default

Pattern 4:

If **Names** ends with 'e'
Then **Balance** $>$ 100000
Else **Balance** $<$ 100000 /



Selecting informative variables/attributes

- » How can we (automatically) judge whether a variable contains important information about the target variable?
 - i.e., What variable gives us the most information about the future default rate of the population?
 - What exactly does “informative” mean?

- » The most basic predictive modeling/data mining technique
 - Can be used as preprocess to many other data mining techniques
 - The basis for one of the most popular, more-complex data mining techniques

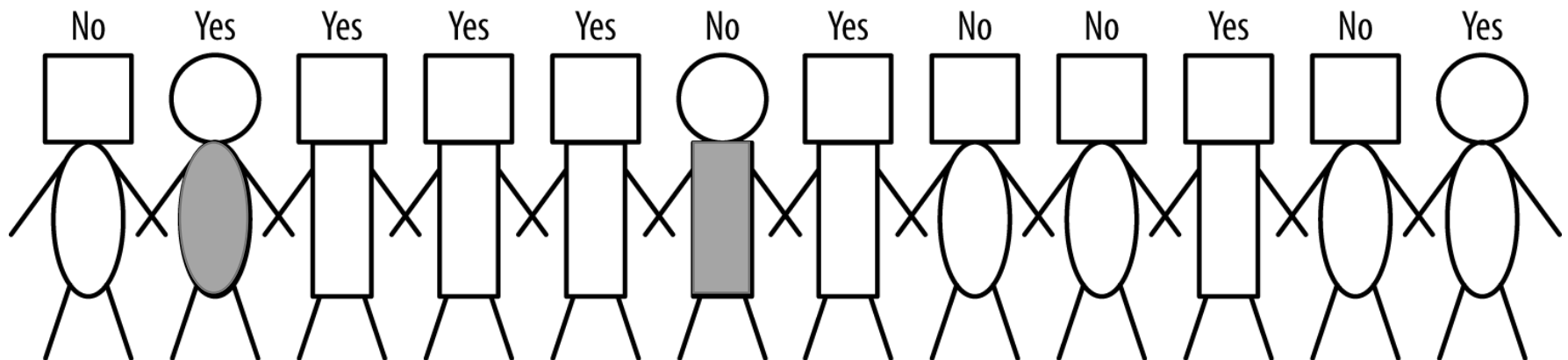
Simple Example



» Let's suppose we have the following attributes of customers from a credit card company

- Write-off or Not: Yes, No
- Head shape: round, square
- Body shape: oval, rectangular
- Body color: grey, white

Yes: 7
No: 5

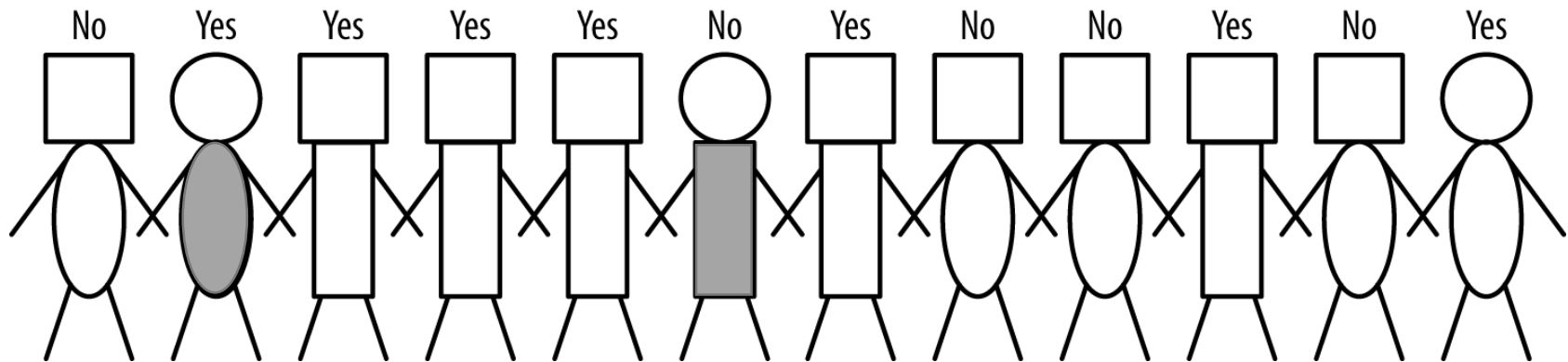




Selecting informative attributes

» Which attributes would best segment these people into groups so that write-offs (Yes) will be distinguished from non-write-offs (No)?

- The resulting groups should be as **pure** as possible!



» We know three attributes: head-shape, body-shape, body-color

Yes: 7
No: 5

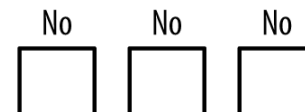
» Let's segment the set based on body-shape.



Oval Bodies



Default? 1/3



How informative is body shape in predicting write-offs?

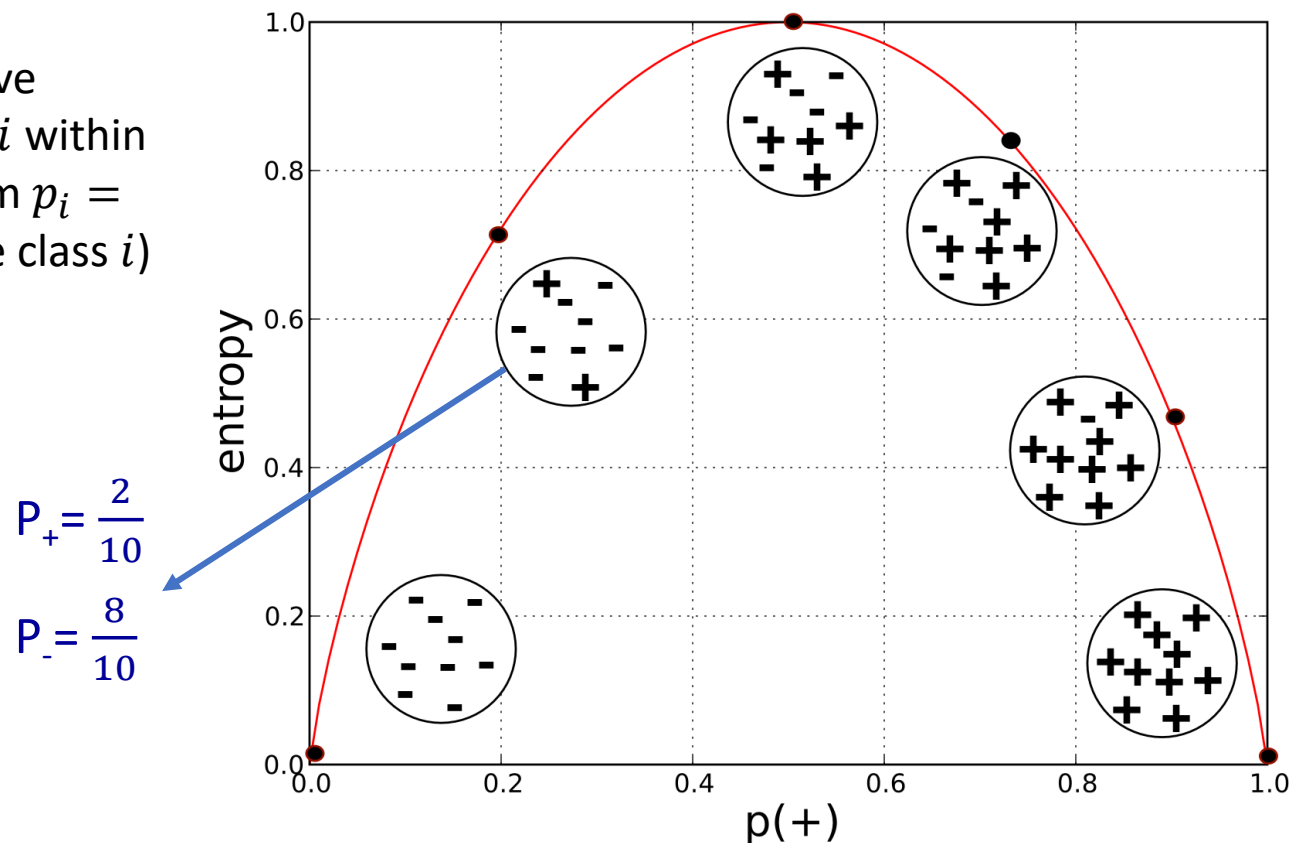


Calculating Impurity (as Entropy)

- » Entropy measures the general disorder of a set (level of data impurity).

$$\text{Entropy} = \sum -p_i \log_2 p_i = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - \dots$$

with p_i as the relative percentage of class i within the set, ranging from $p_i = 0$ to $p_i = 1$ (all have class i)

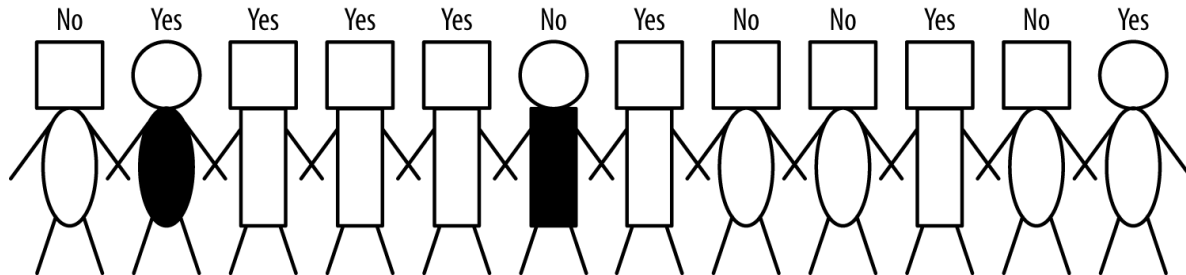




Entropy of initial population

» Entropy = $\sum -p_i \log_2 p_i = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - \dots$

» Our initial population is composed of 7 cases of class “Yes” and 5 cases of class “No”



$$\underline{p_y = \frac{7}{12}}, \quad \underline{p_n = \frac{5}{12}}$$

» Entropy (entire population of examples) =

$$-\left(\frac{7}{12} \cdot \log_2 \frac{7}{12}\right) - \left(\frac{5}{12} \cdot \log_2 \frac{5}{12}\right) = 0.98$$



Entropy after segmentation

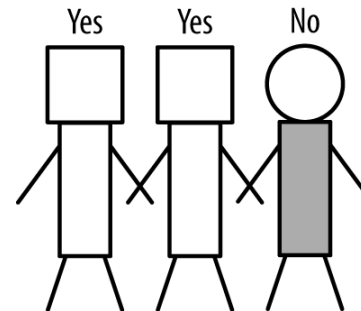
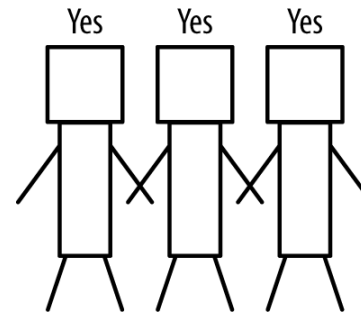
» Entropy (Rectangular):

$$\begin{aligned} &= - \left[\left(\frac{5}{6} \right) \times \log_2 \left(\frac{5}{6} \right) + \left(\frac{1}{6} \right) \times \log_2 \left(\frac{1}{6} \right) \right] \\ &\approx -[0.83 \times -0.26 + 0.17 \times -2.58] \\ &\approx 0.65 \end{aligned}$$

» Entropy (Oval) :

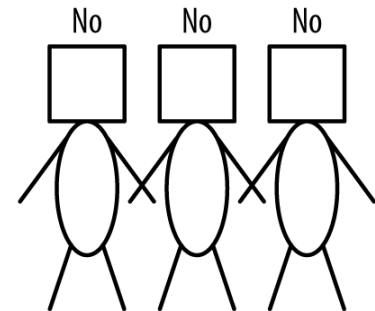
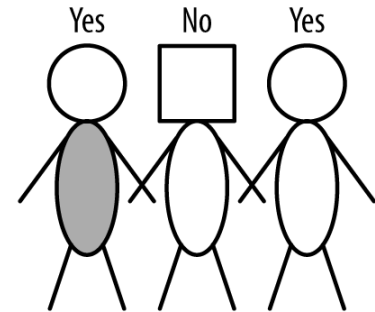
$$\begin{aligned} &= - \left[\left(\frac{2}{6} \right) \times \log_2 \left(\frac{2}{6} \right) + \left(\frac{4}{6} \right) \times \log_2 \left(\frac{4}{6} \right) \right] \\ &\approx -[0.33 \times -1.58 + 0.66 \times -(0.58)] \\ &\approx 0.92 \end{aligned}$$

Rectangular Bodies



Yes: 5
No: 1

Oval Bodies



Yes: 2
No: 4



Information Gain (IG)

» Information Gain

- Measures how much an attributes improves (decreases) entropy over the whole segmentation it creates.

» $IG(\text{parent}, \text{children})$

$$= \text{entropy}(\text{parent}) - [\text{entropy}(\text{children})]$$

$$= \text{entropy}(\text{parent}) - [p(c_1) * \text{entropy}(c_1) + p(c_2) * \text{entropy}(c_2) + \dots]$$

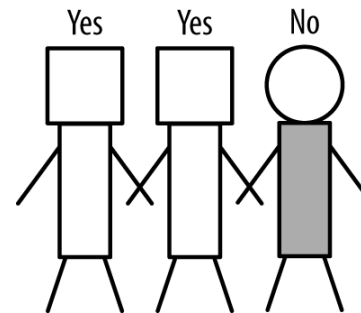
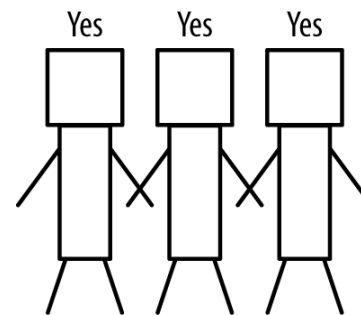
- The entropy for each child c_i is weighted by the proportion of instances belonging to that child



Entropy after segmentation

- » Entropy (all population) ≈ 0.98
- » Entropy (Rectangular) ≈ 0.65
- » Entropy (Oval) ≈ 0.92
- » Information Gain
$$= 0.98 - (0.5 \times 0.65 + 0.5 \times 0.92)$$
$$= 0.196$$
- » This split reduces entropy substantially \rightarrow the body shape provides a lot of information on the value of target (write-offs).

Rectangular Bodies

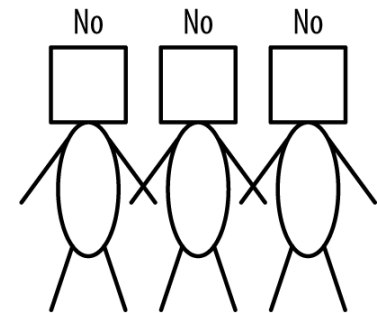
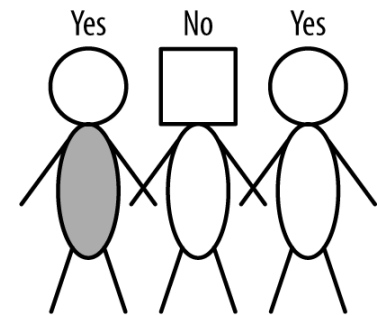


6 out of 12



$$P(c_1) = 6/12 = 0.5$$

Oval Bodies



6 out of 12

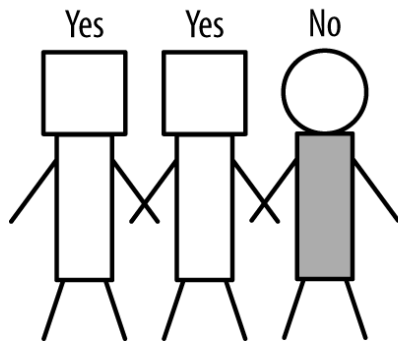
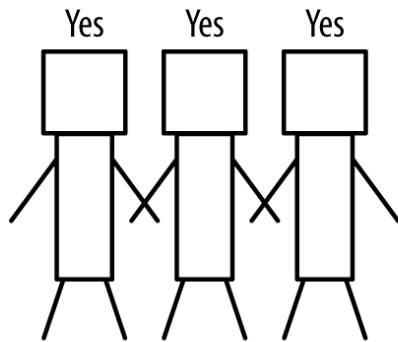


$$P(c_2) = 6/12 = 0.5$$

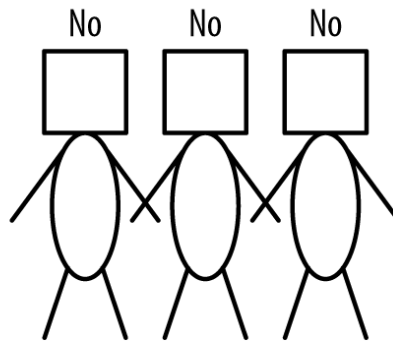
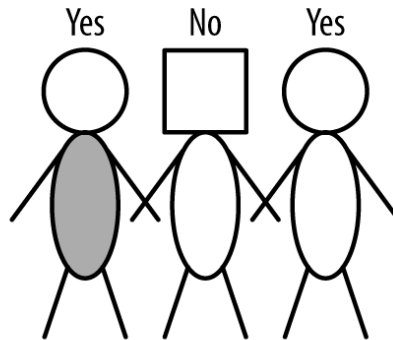


How to build a decision tree (I)

Rectangular Bodies



Oval Bodies



Now, let's take each subset as a population and repeat the process

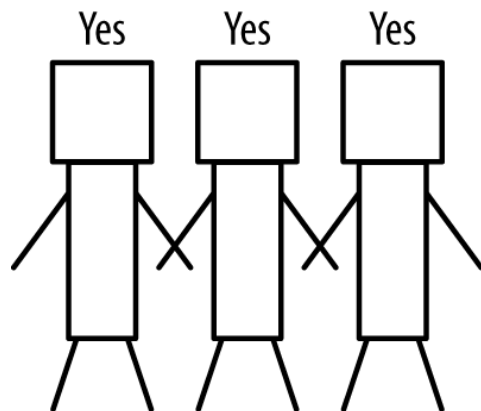
→ Select the most informative attribute in each subset

Shall we start with Rectangular body group?

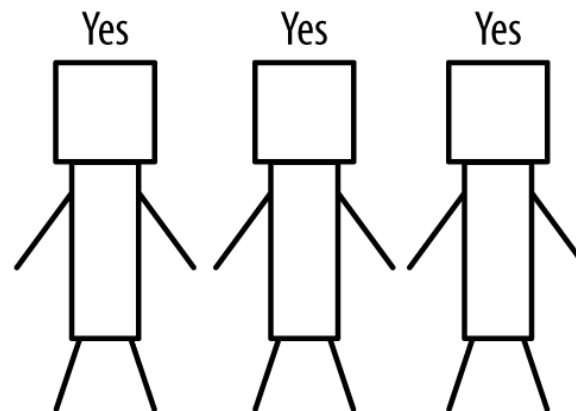
How to build a decision tree (II)



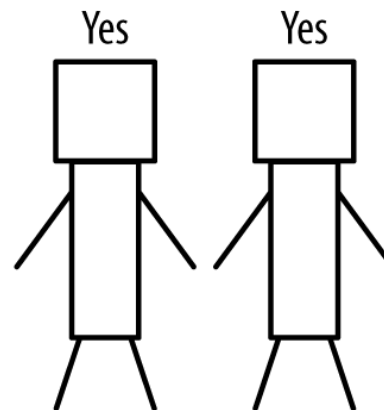
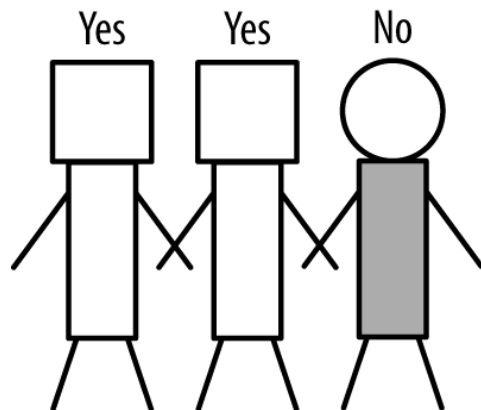
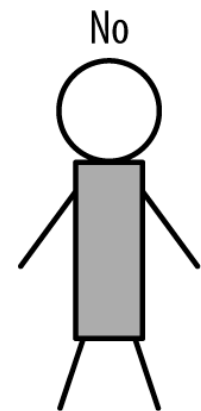
Rectangular Bodies



Rectangular Body and White



Rectangular Body and Gray



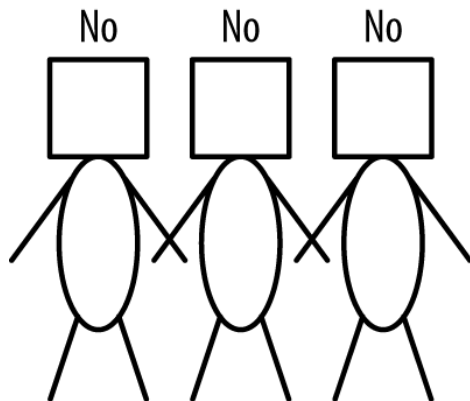
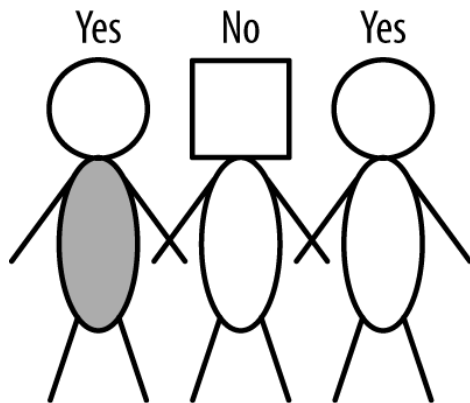
Let's choose body-color to segment this group!

Are the resulting groups pure?

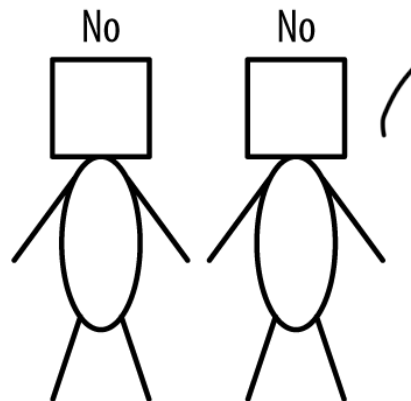
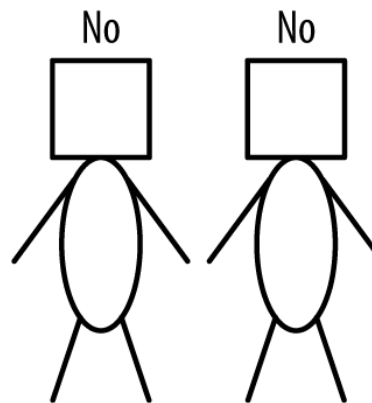


How to build a decision tree (III)

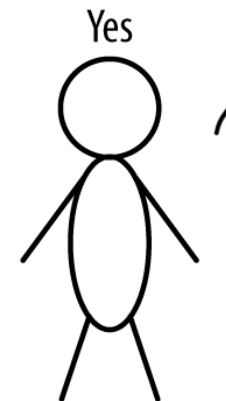
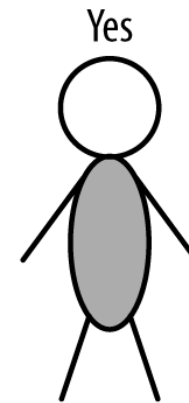
Oval Bodies



**Oval Body and
Square Head**



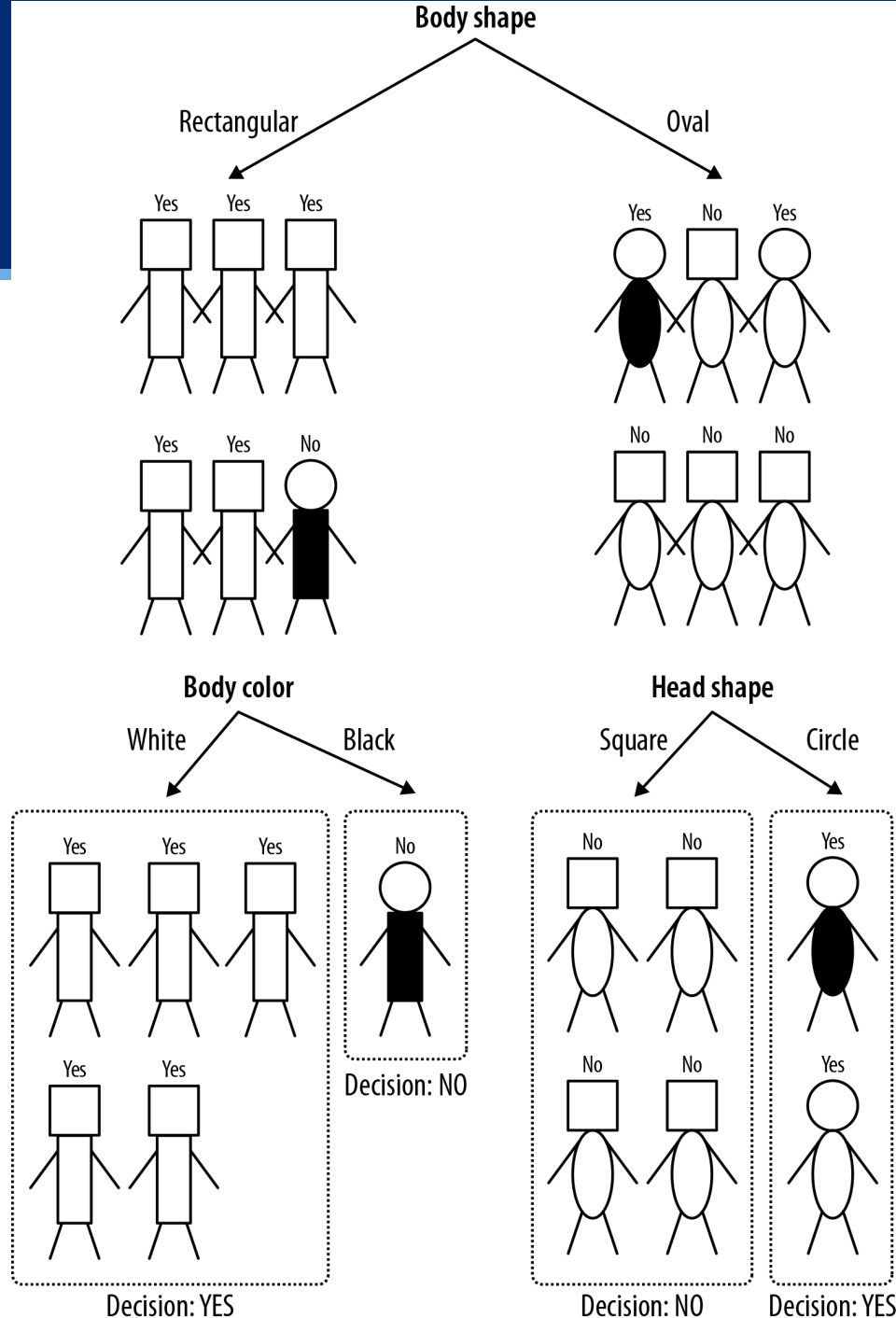
**Oval Body and
Circular Head**



How to build a decision tree (IV)

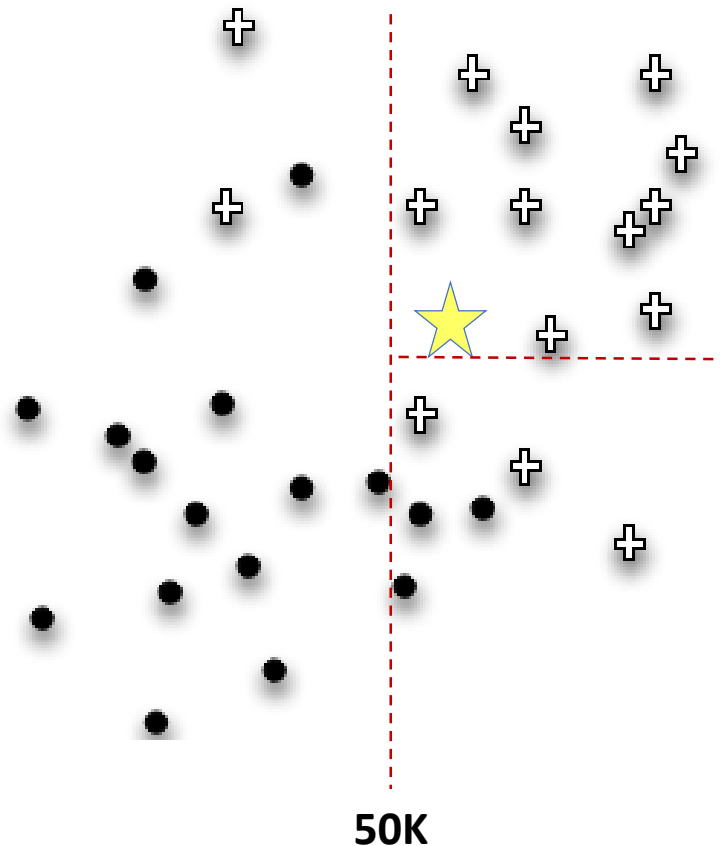
» Recursively apply **attribute selection** to find the best attribute to partition the data set

» The goal at each step is to select an attribute to partition the current group into subgroups that are as pure as possible w.r.t. the target variable





Credit Card Default Case

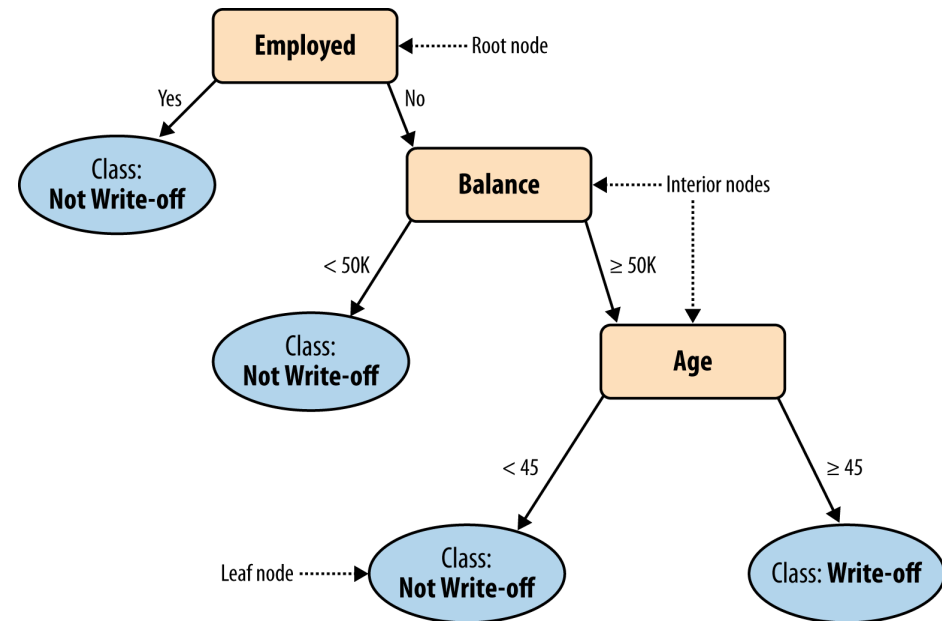


$$\text{Entropy}(\text{population}) - \text{Entropy}(\text{After split})$$

Information Gain

- Not Default – 17 cases
- + Default – 15 cases

Tree-Structured Models

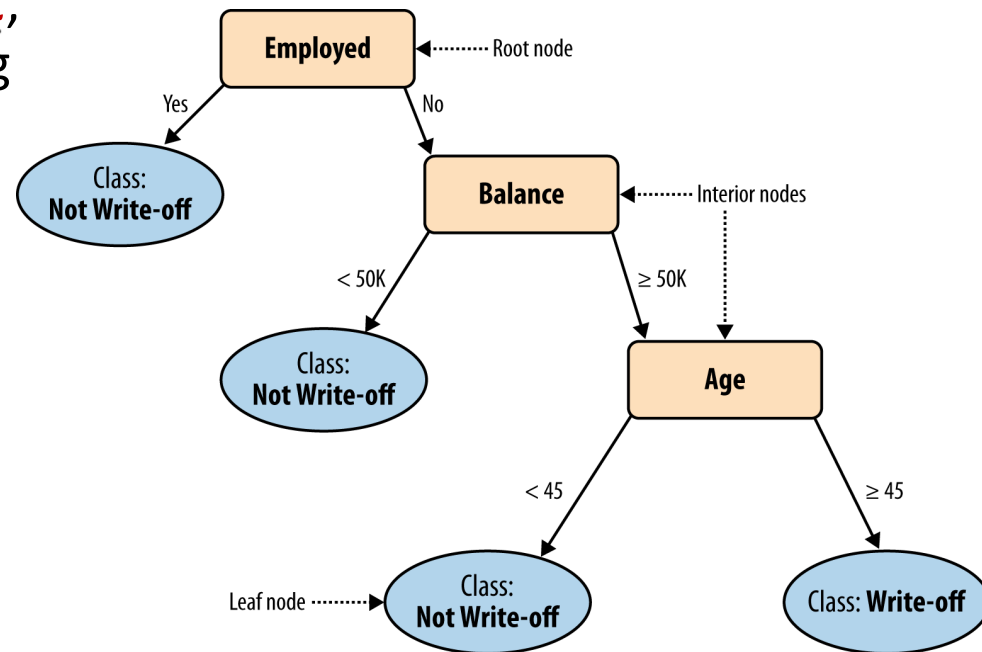


So, now we know where the name of classification tree came from: upside-down tree image because we branches out from top to bottom.



Tree-Structured Models - Terminology

- » The tree creates a **segmentation** of the data
- » Each **node** in the tree contains a test of an attribute
- » Each **path** eventually terminates at a leaf
- » Each **leaf** corresponds to **segment**, and the attributes and values along the path give the characteristics
- » Each leaf contains a value for the target variable



Q: Using Tree-Structured Models

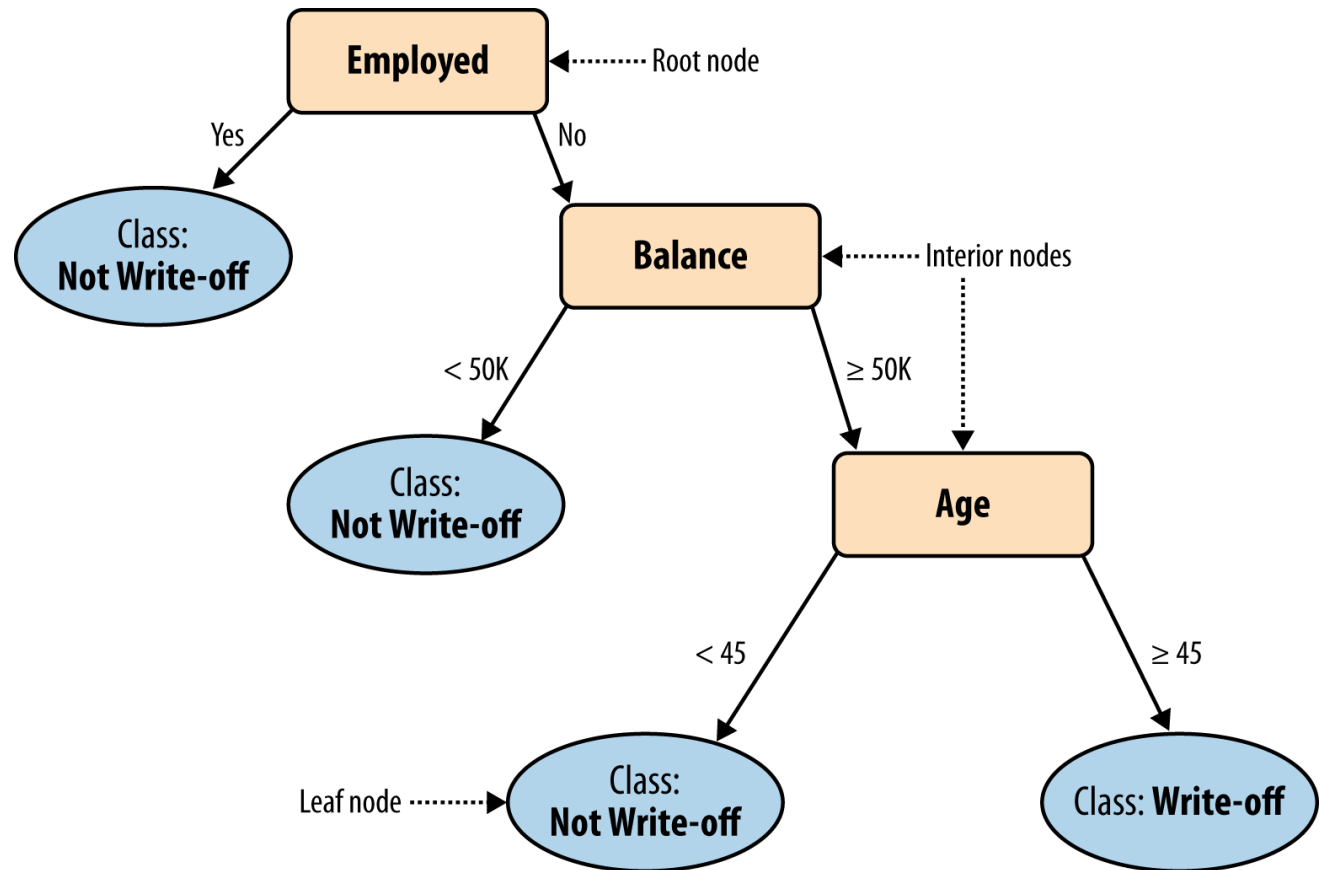


- Classify 'John Wick'

- Balance=115K

- Employed=No

- Age=40





When do we stop? – Pruning Trees

» When to stop?

- if all nodes are pure ✓
- if there are no more variables to be considered, or ✓
- even earlier (over-fitting – to be continued..) ✓

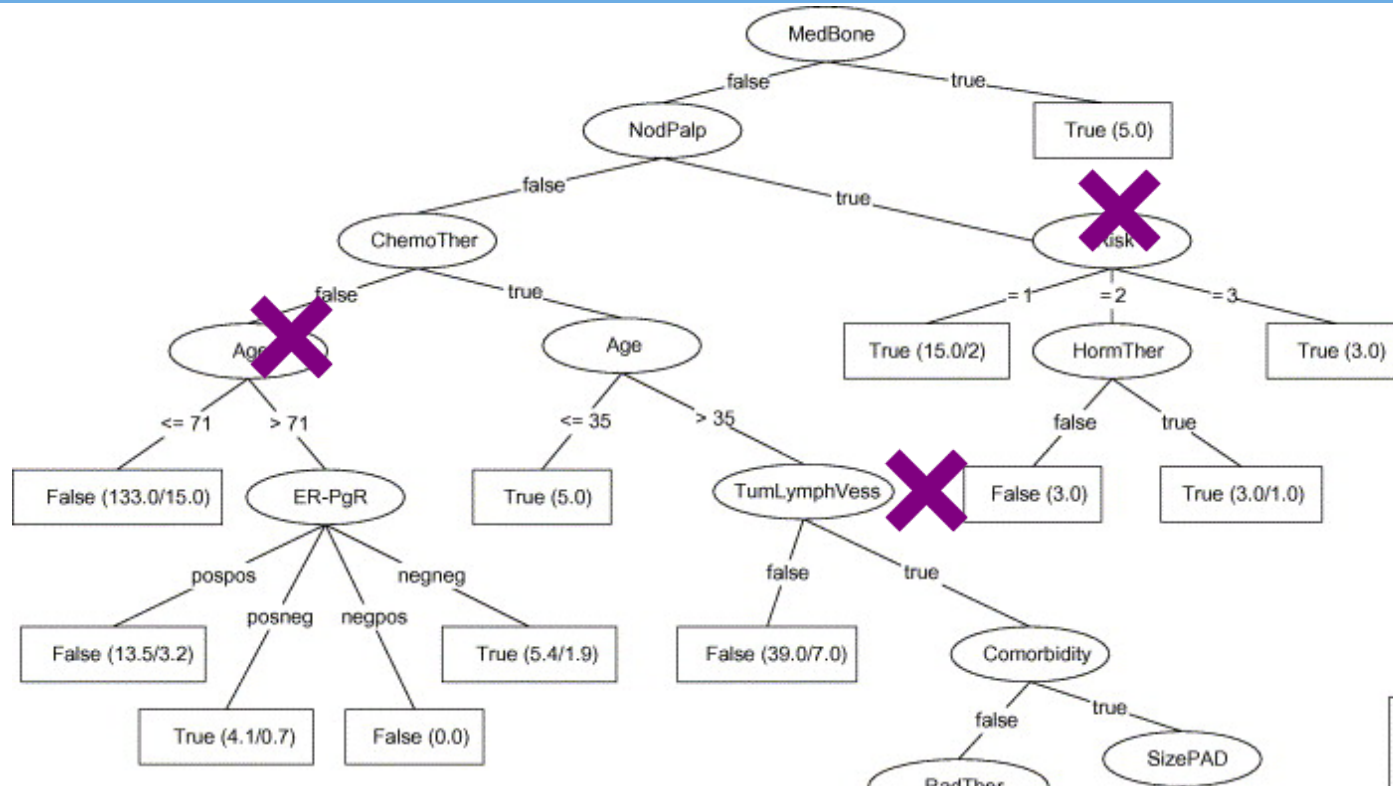
» **Pruning** simplifies a classification tree to prevent **over-fitting** to noise in the data ✓

- Post-pruning: takes a fully-grown decision tree and discards unreliable parts
- Pre-pruning: stops growing a branch when information becomes unreliable ✓

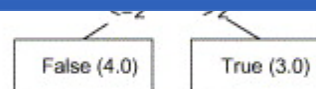
» Post-pruning preferred in practice



Post-pruning a tree (example)



We will discuss tree pruning in Week 5.



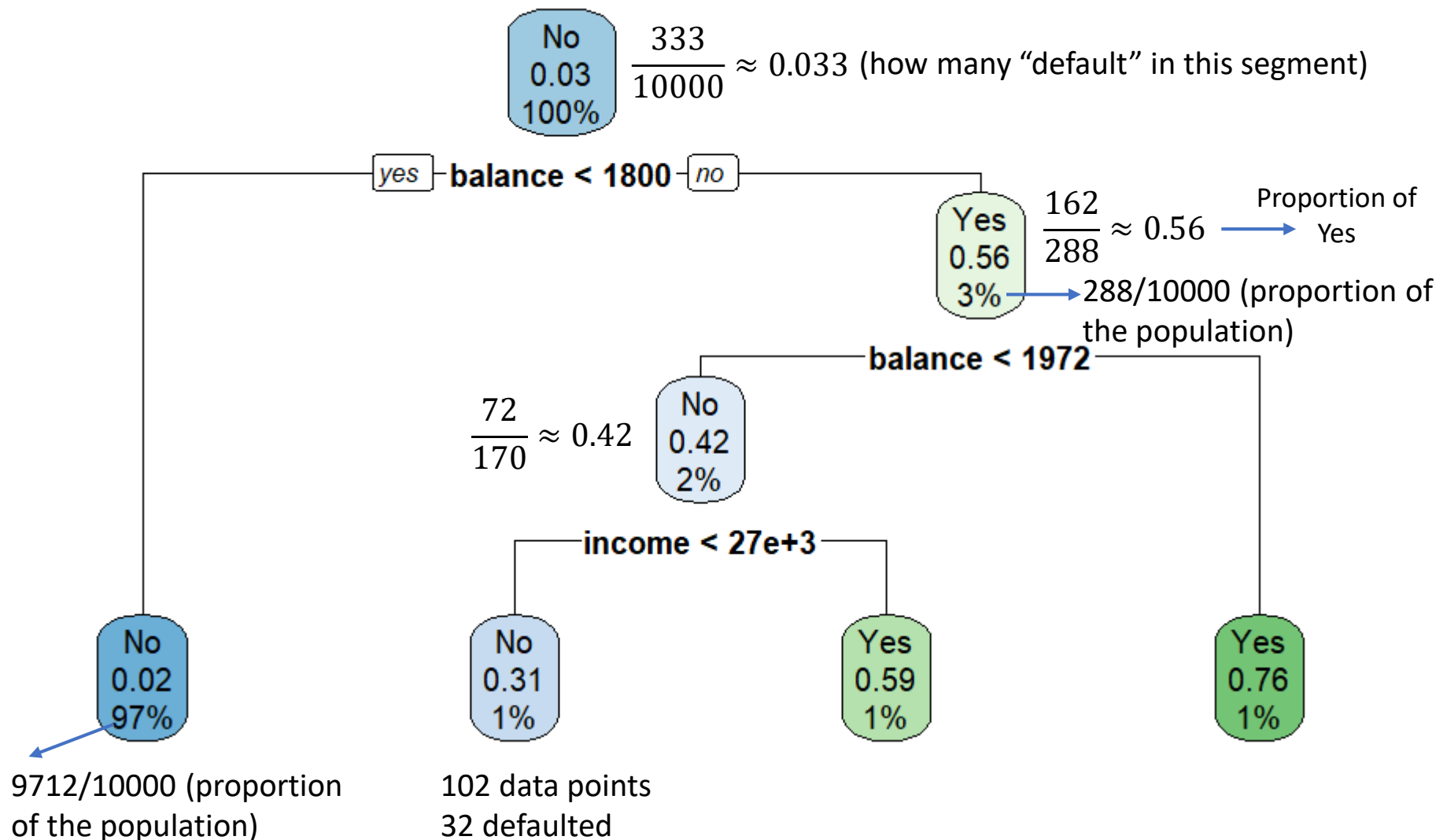
Sensitivity 0.40
Specificity 0.91



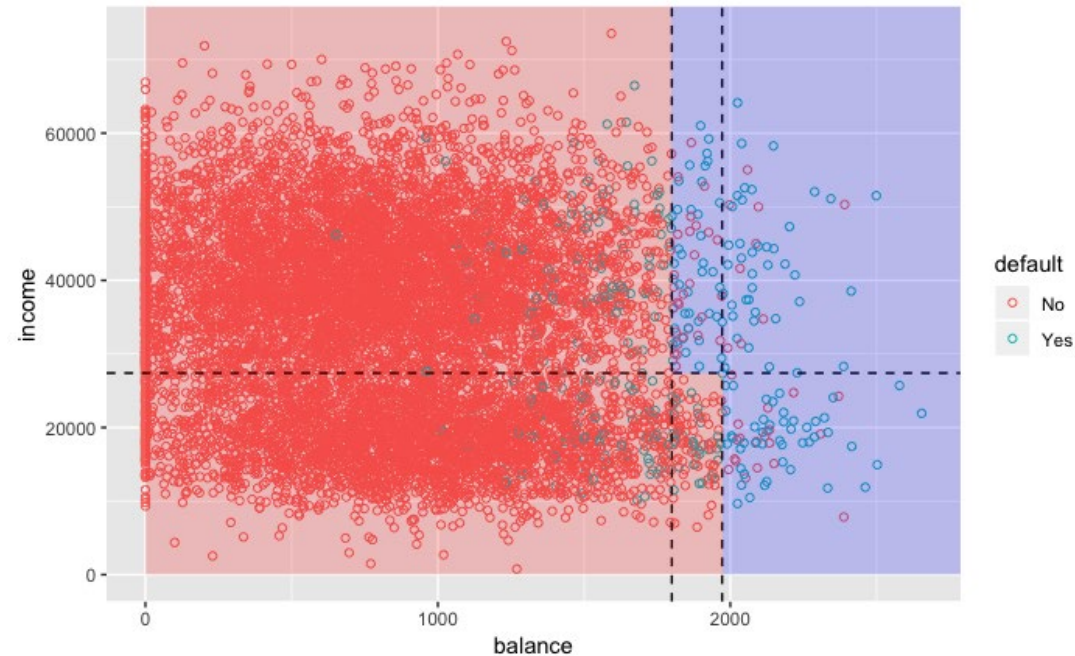
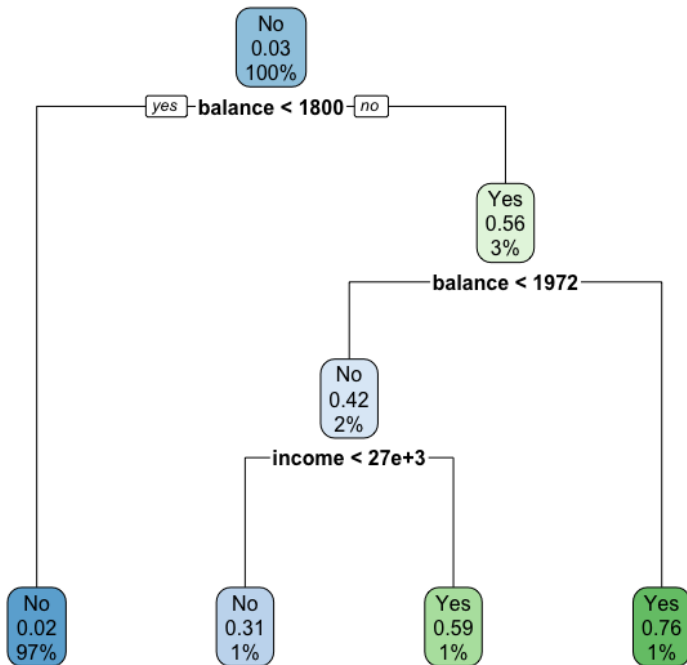
Classification Tree

R exercise

Classification Trees with R



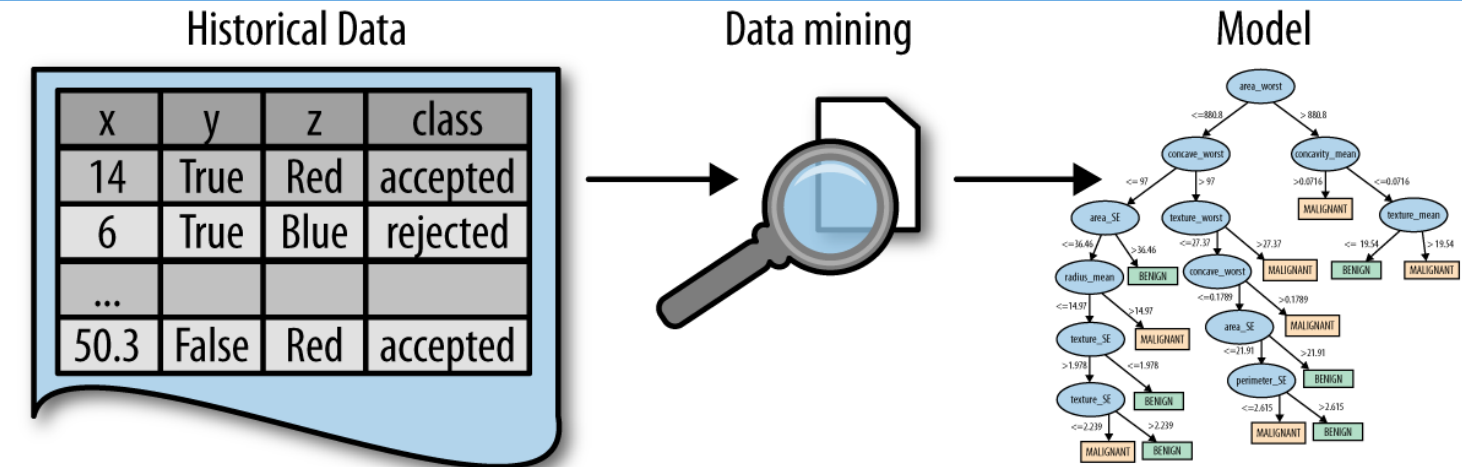
Visualizing Segmentation



What will you do with this result?

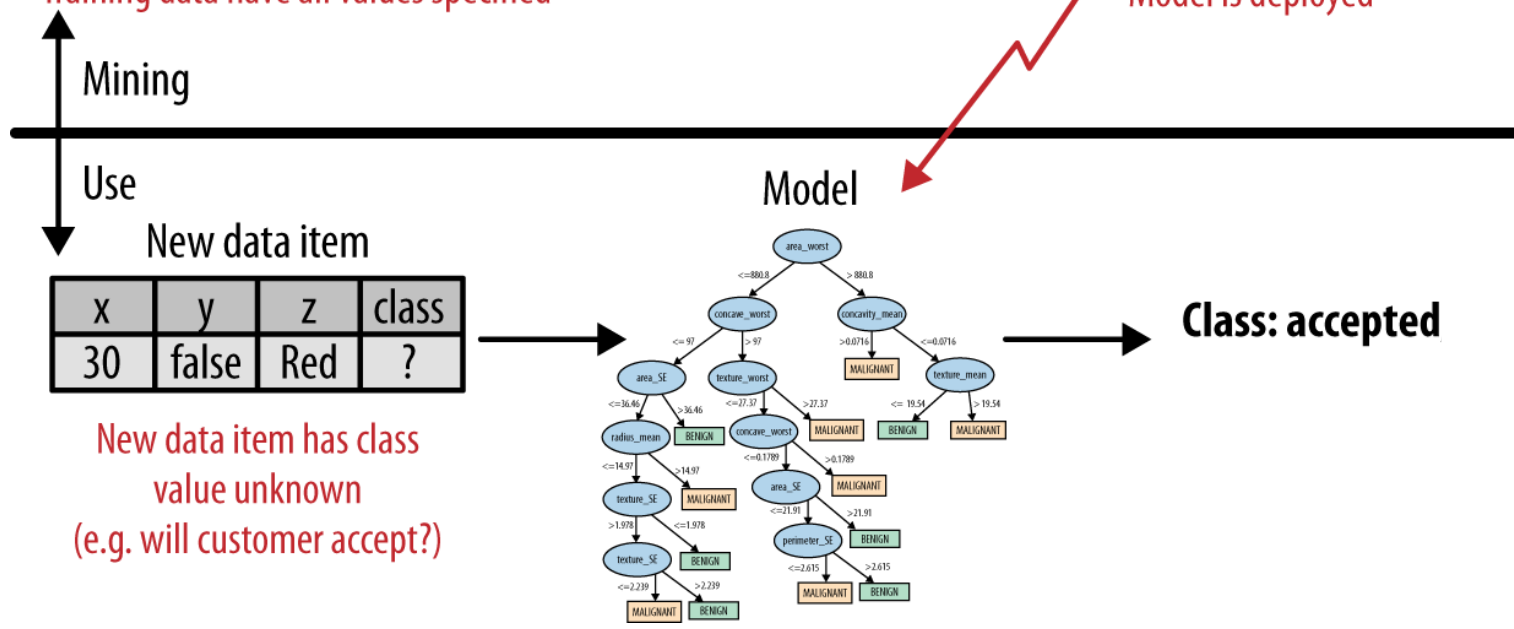


Data mining vs. Use of the model (1/2)

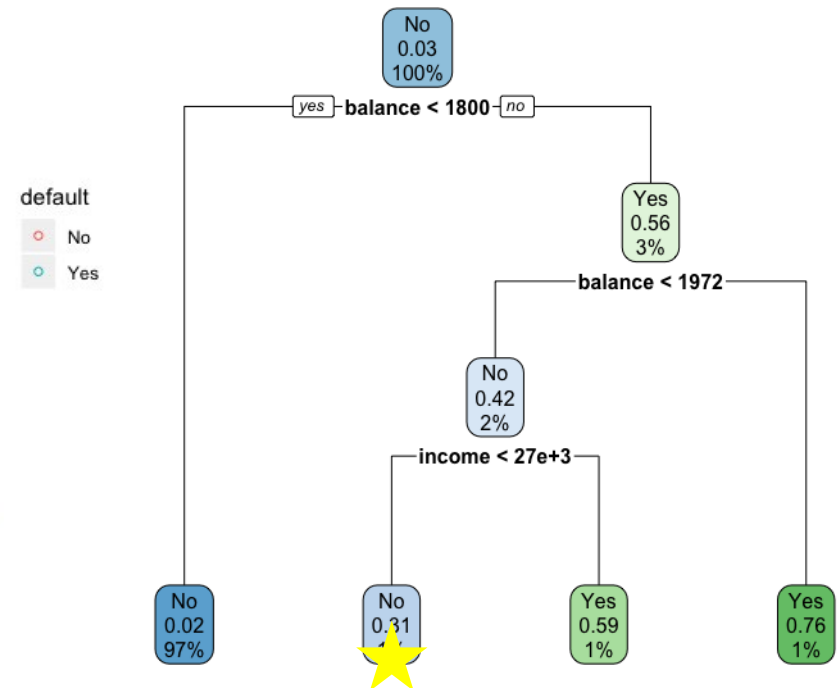
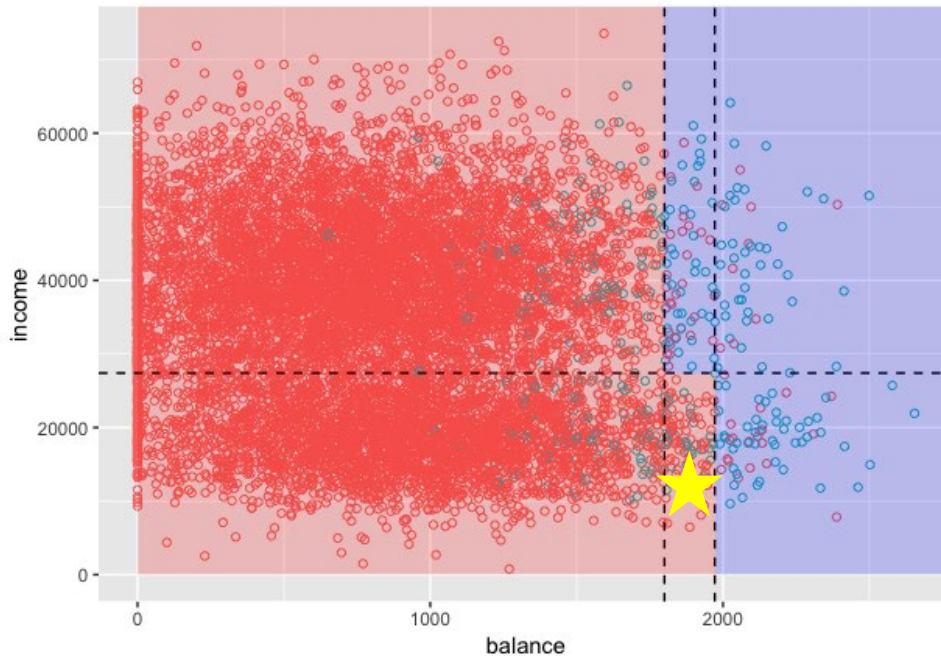


Training data have all values specified

Model is deployed



Data mining vs. Use of the model (2/2)



Less likely to default!

Probability estimation from classification models



- » Type of target variable:
 - classification → categorical target
- » But, many classification models **produce continuous values** }
(probabilities, or “ranks”/“scores”)
- » In that case, classification can also be interpreted as a form of probability estimation or ranking



When might a probability be more useful than a class prediction (yes/no)?

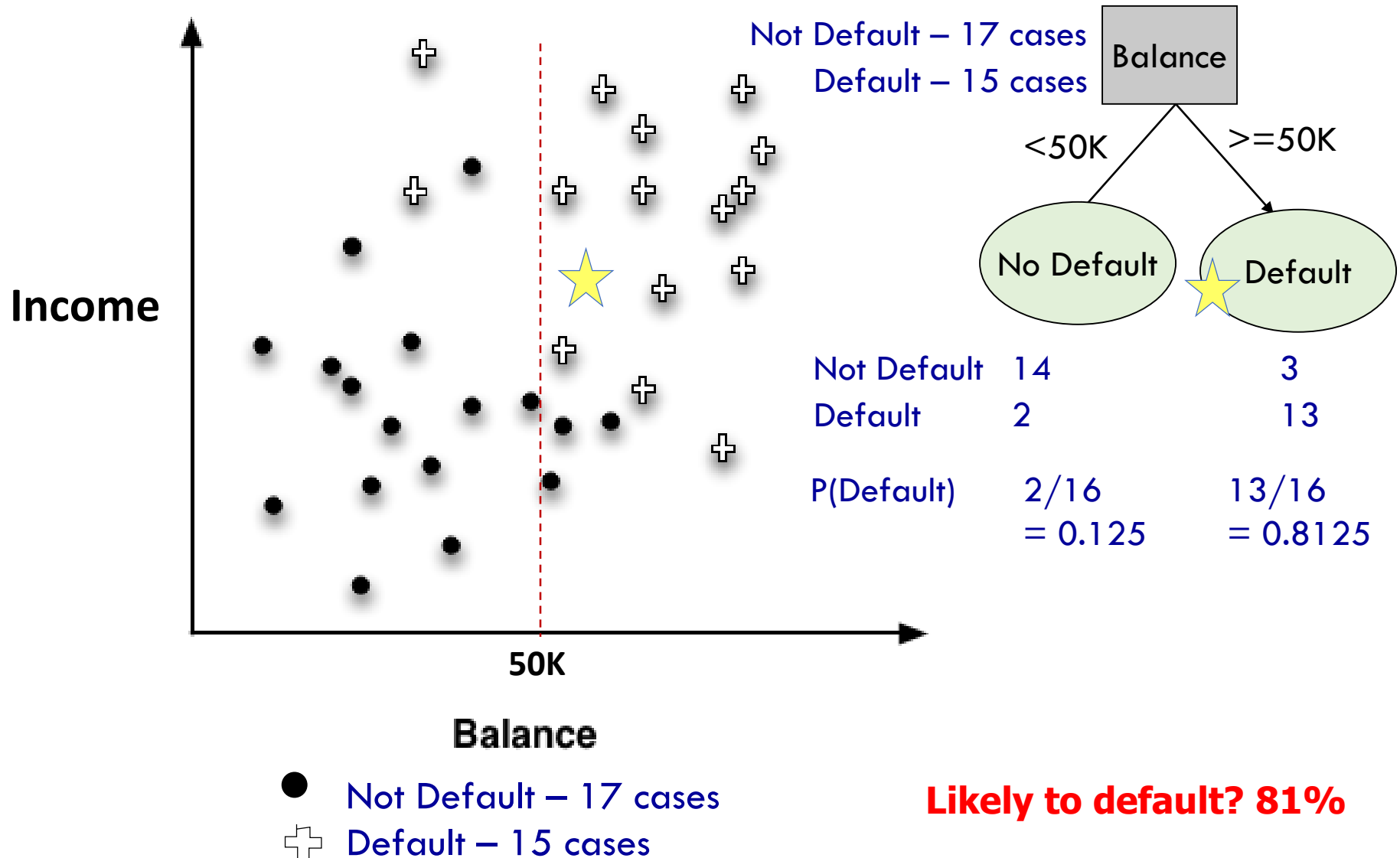
» Credit card 'Default' status prediction?

» Customer churn prediction?

» ?



Example of Probability Estimation





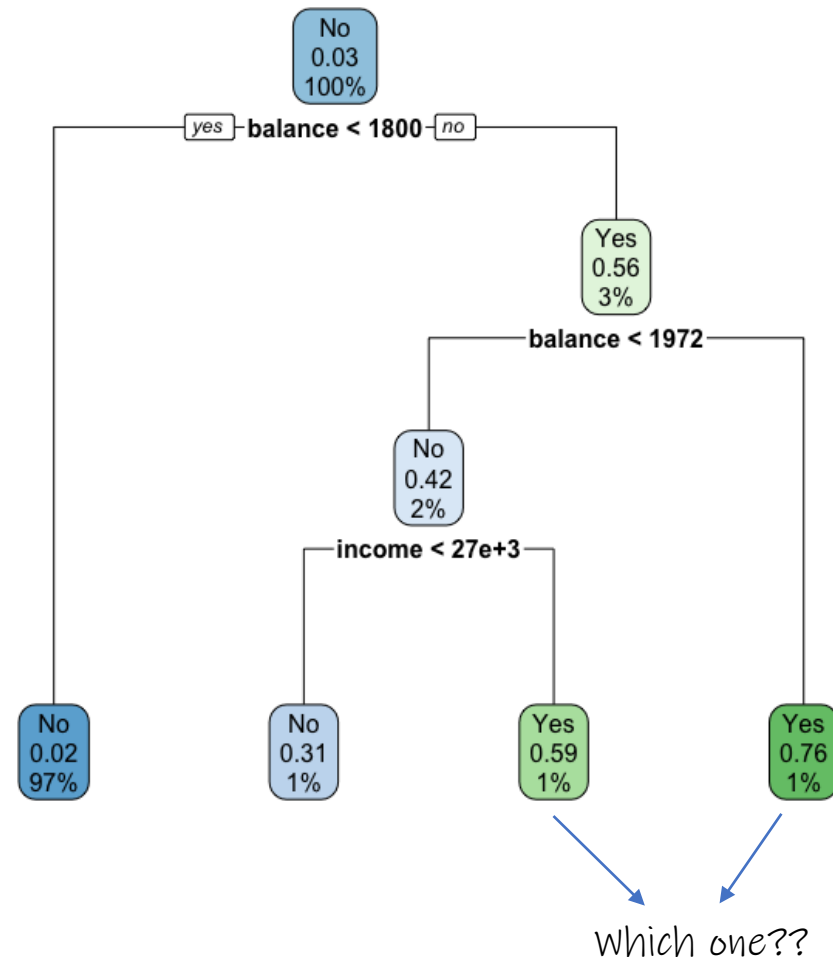
Probability estimation is useful

» We often need a more informative prediction than just a classification

- E.g. allocate your budget to the instances with the highest expected loss
- More sophisticated decision-making process

» Classification may oversimplify the problem

- E.g. if all segments have a probability of < 0.5 for default, every leaf will be labeled “not default”
- We would like each segment (leaf) to be assigned an **estimate of the probability of membership** in the different classes





Classification Tree Advantages

- » For classification modeling, tree induction is one of the most popular data mining tools
- » It is:
 - Easy to understand, interpret, and visualize
 - Easy to use and implement and requires little data preparation
 - Can handle both numerical and categorical features natively
 - [Can handle missing data elegantly](#)
 - Computationally cheap (can be trained quickly on large datasets)
 - Robust to outliers
 - Can model non-linearity in the data
- » Works remarkably well (not the most accurate, by the way)
- » Has advantages for model comprehensibility, which is important for:
 - model evaluation
 - communication to non-DM-savvy stakeholders



Classification Trees Disadvantages

- » Large trees can be hard to interpret
- » Trees have high variance, which causes model performance to be poor
- » Trees overfit easily



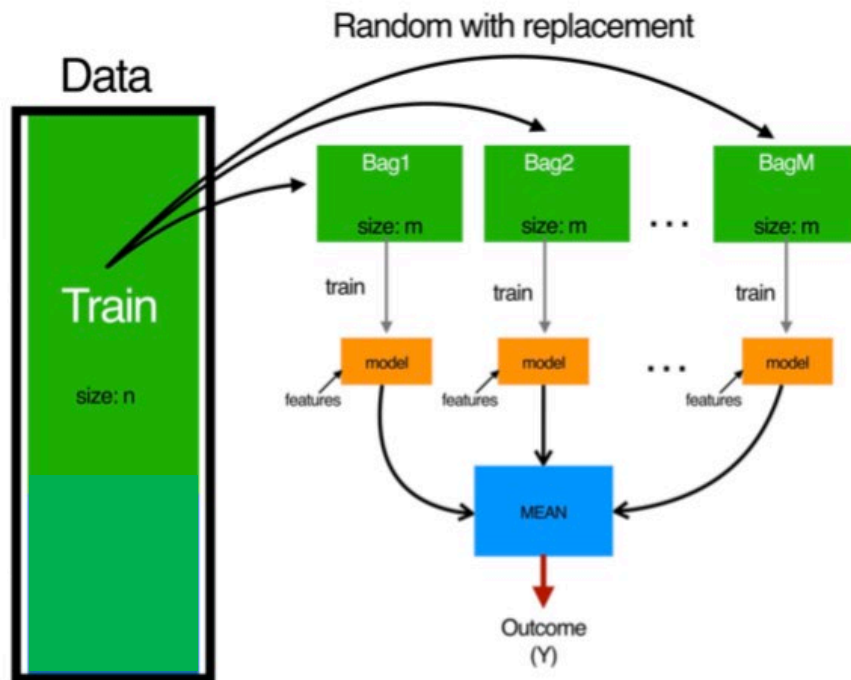
To overcome these disadvantages,
Random Forest model is developed.



Random Forests

- » Examples of “ensemble” methods, “Wisdom of the Crowd”
- » Predictions from many trees are combined – a collection of trees makes the final prediction

How does it work? (1/2)



- » Draw multiple bootstrap resamples of cases from the data
- » For each resample, use a random subset of predictors and produce a tree
- » Combine the predictions/classifications from all the trees (the “forest”)
 - Voting for classification
 - Averaging for regression



How does it work? (2/2)

Original Dataset

Age	Student	Income	Balance	Default
40	No	44K	3K	No
32	Yes	12K	1.5K	No
27	No	31K	1K	Yes
51	No	35K	0.5K	No
35	No	63K	2K	No

Bootstrapped Dataset

Age	Student	Income	Balance	Default
40	No	44K	3K	No
40	No	44K	3K	No
27	No	31K	1K	Yes
51	No	35K	0.5K	No
35	No	63K	2K	No

32	Yes	12K	1.5K	No
----	-----	-----	------	----

Out-Of-Bag Dataset

- » OOB (Out-Of-Bag) datasets are great test data
- » OOB Error: Performance of Random Forest is assessed by the proportion of OOB samples that are classified incorrectly by the model

Random Forests – Strengths and Weaknesses



- » Very good predictive performance, better than single trees (often the top choice for predictive modeling)
- » Cost: loss of rules you can explain and implement (since you are dealing with many trees, not a single tree)
- » However, RF does produce “variable importance scores,” which can be useful.