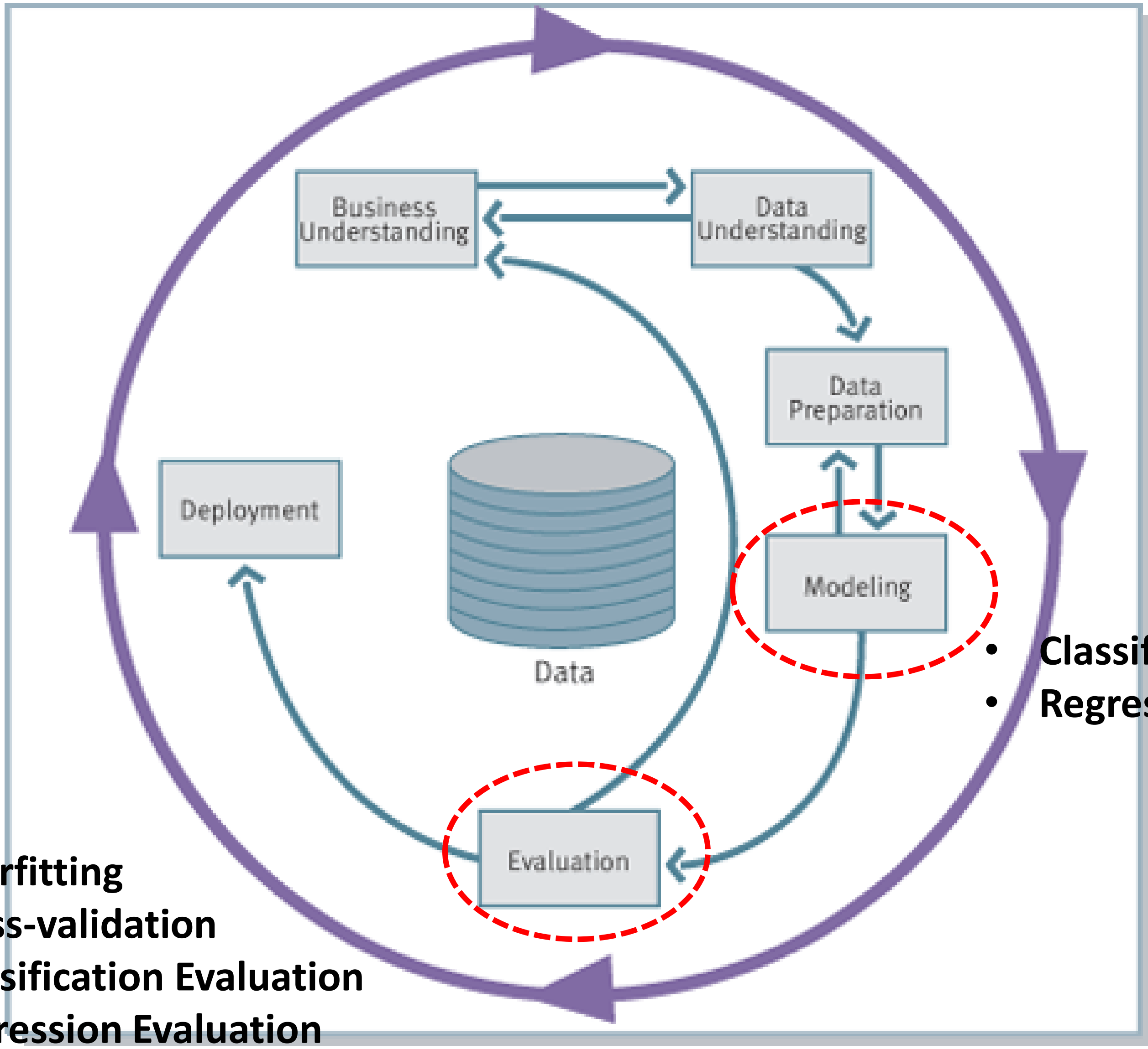
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Business Analytics (MM 5425)

Linear Regression & Logistic Regression

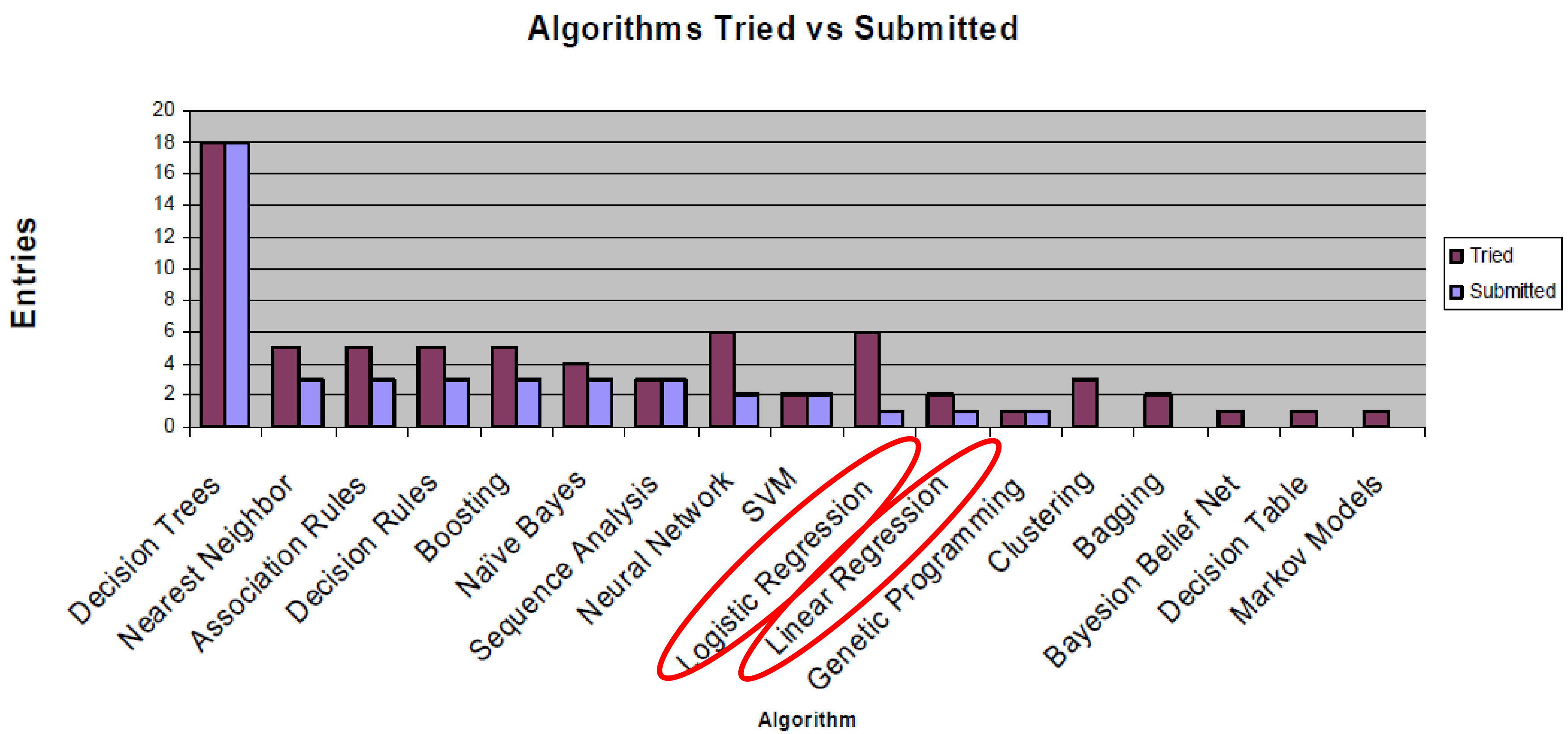
Dr. Yue (Katherine) FENG



- **Classification Tree**
- **Regression Tree**

- **Overfitting**
- **Cross-validation**
- **Classification Evaluation**
- **Regression Evaluation**

Commonly Used Algorithms



Agenda

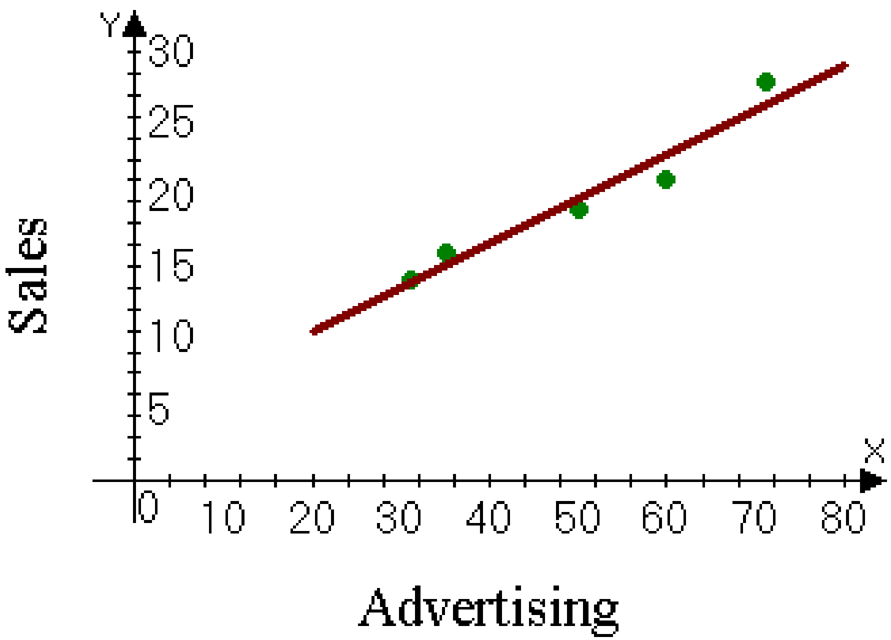
I. Linear Regression - A Brief Intro

II. Logistic Regression - Model Setup

III. Decision Tree vs. Logistic Regression

Linear Regression

Linear Regression



$$h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$$

The coefficients we want to know.

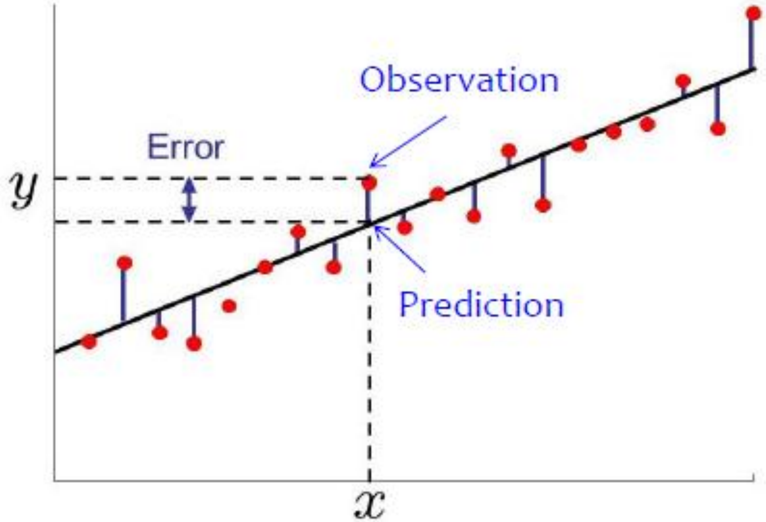
Linear regression models a **linear relationship** between target variable and predictor variables.

Ordinary Least Squares (OLS) for Linear Regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

Objective function: $$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2$$

- Minimizes sum of squared errors
- One unit of change in xi is associated with θi change in the value of y.

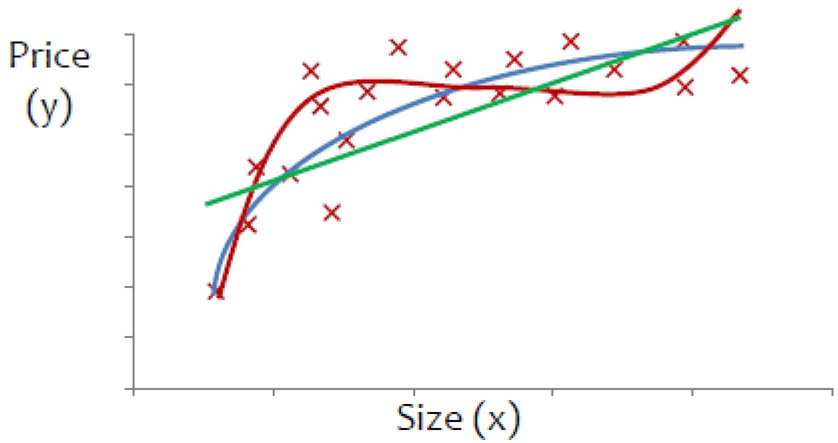


Linear Regression (Optional)

> "Linear regression" = **linear in parameters (θ)**

> The inputs for linear regression can be

- Transformation of quantitative inputs: e.g., log, square root, square
- Polynomial transformation: e.g., 1, x, x^2, …
- Interactions between variables: e.g., $x3 = x1 \cdot x2$



$$
\begin{aligned}
h_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \theta_0 + \theta_1 x \\
&= \theta_0 + \theta_1 x + \theta_2 x^2 \\
&= \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3
\end{aligned}
$$

Linear vs. Logistic Regression

Response		Analysis
Continuous	→	Linear Regression Analysis
Categorical	→	Logistic Regression Analysis

Logistic Regression – Model Setup

What Does Logistic Regression Do?

> Logistic regression is a **classification** model.

> The values of the target variable in the data are categorical.

> The model produces a numeric estimate - probability of a specific class.

Classification Problems: Revisit

> Churn in cellular services: Stay / Leave?

> Email: Spam / Not Spam?

> Online Transactions: Fraudulent (Yes / No)?

$$y \in \{0, 1\} \begin{cases} 0: \text{“Negative Class”} \\ 1: \text{“Positive Class”} \end{cases}$$

Linear vs. Logistic Regression

> Simple linear regression:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

> Since *y* is categorical now, we are interested in its **probability (*P*)**

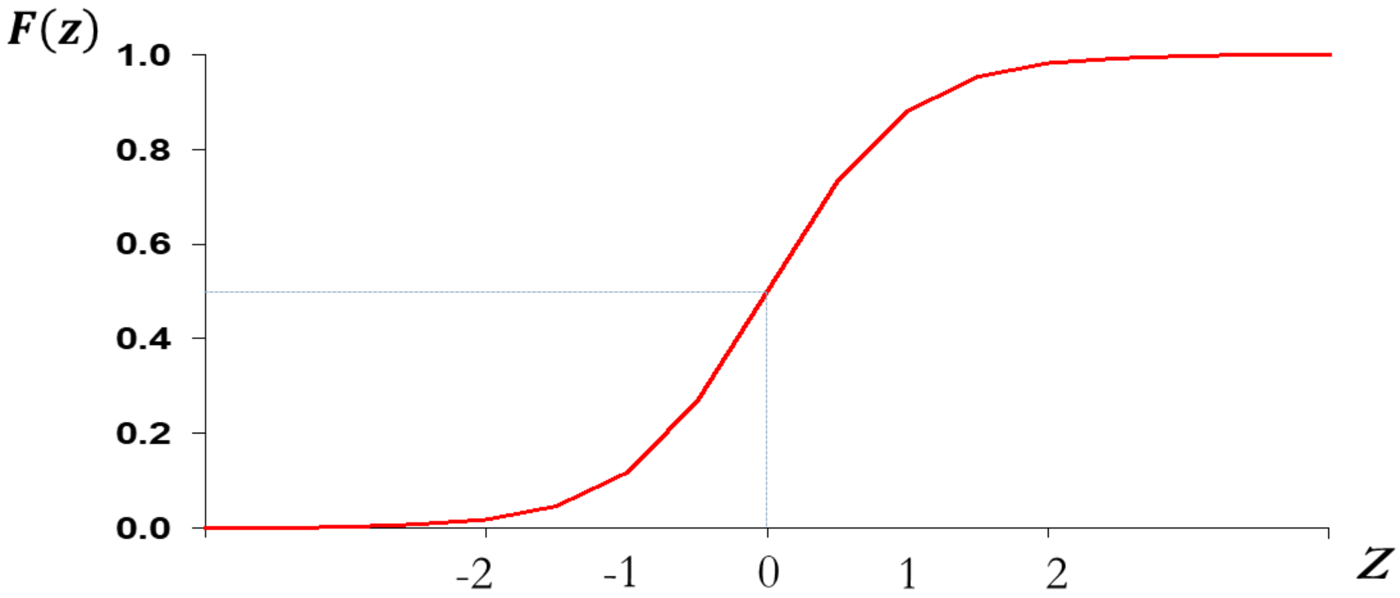
- *P*: probability of a data instance to be a specific class **(*y*=1)**

> What if we apply linear equation to predict *P*(y=1)?

$$P(y=1) = \theta_0 + \theta_1 x \notin [0,1]$$

Logistic Function

$$p = F(\theta_0 + \theta_1 x\,)$$ ***F(.)*** **is the logistic function**



$$p = F(z) = \frac{1}{1 + e^{-z}}$$

$$z = \theta_0 + \theta_1 x$$

Logistic Regression

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

$$P(y = 1|x; \theta) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k)}}$$

- Probability of y=1, given features $x_1, x_2, \cdots x_k$ and θ.
- Note that $P(y = 0|x; \theta) = 1 - P(y = 1|x; \theta)$

Parameter Estimation (Optional)

> We use the training data to fit model and estimate the parameters.

> Linear regression: Ordinary Least Square (OLS)

> Logistic regression:

- Maximum Likelihood Estimate (MLE) in Statistics
- Minimize objective function **(cost function)**

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

Where $h_\theta\left(x^{(i)}\right) = P\left(y = 1 \middle| x^{(i)}; \theta\right)$ and m indicates the number of instances.

Interpret Logistic Regression

- A positive parameter means that the positive event is more likely to take place given one unit increase in the predictor variable.

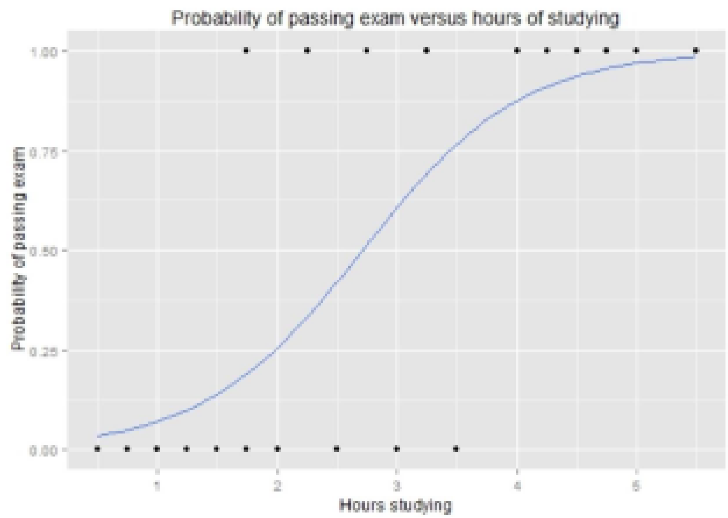
- A negative parameter means that the positive event is less likely to take place given one unit increase in the predictor variable.

An Example

- A group of 20 students spend between 0 and 6 hours studying for an exam. Can we predict whether a student will pass an exam based on the hours studying for the exam?

0: failed; 1: passed

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



If a student studies for 2 hours, estimated probability of passing the exam of 0.26;

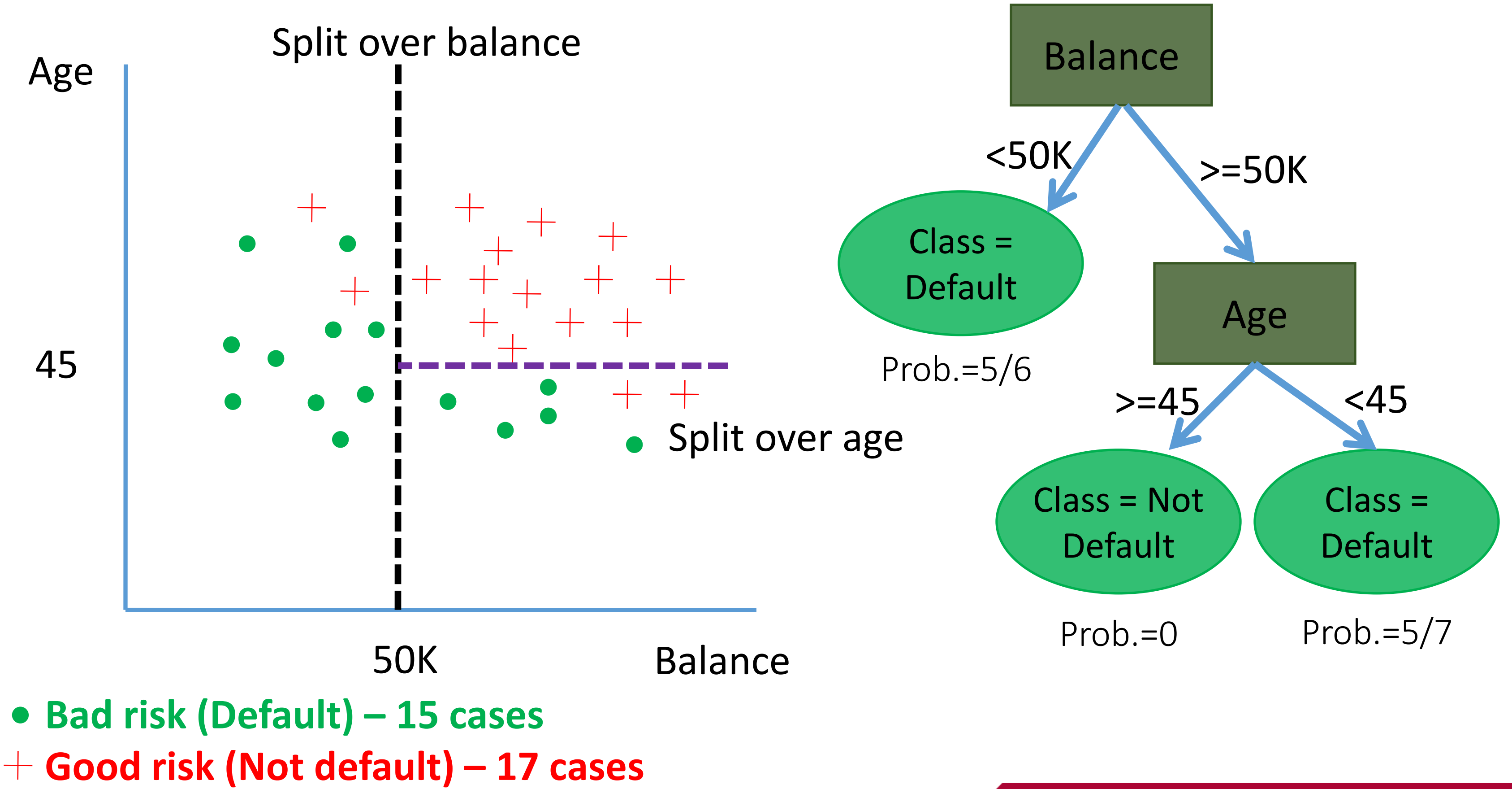
If a student studies for 4 hours, estimated probability of passing the exam is 0.87.

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot \text{Hours} - 4.0777\right)\right)}$$

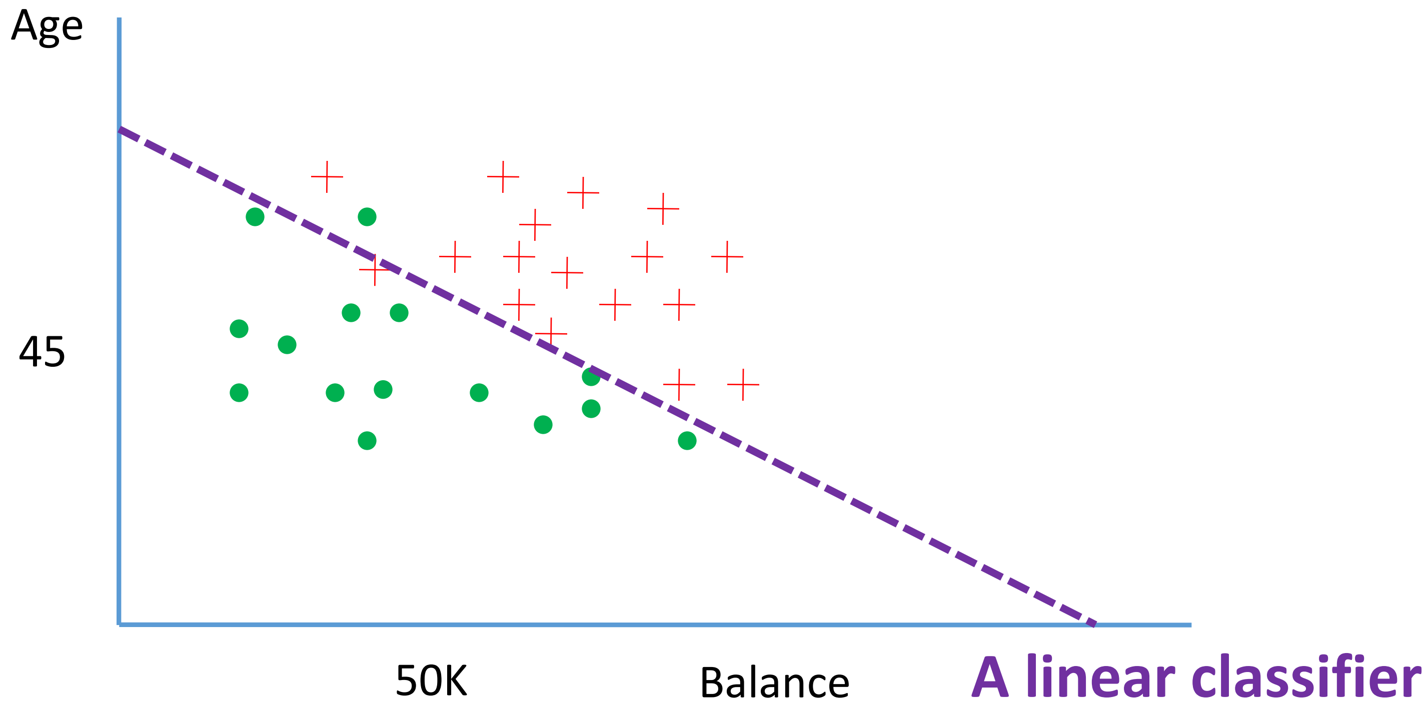
Decision Tree
vs. Logistic Regression

Classification by Decision Tree

> Classification tree partitions space of examples with axis-parallel decision boundaries



Alternative Partitioning

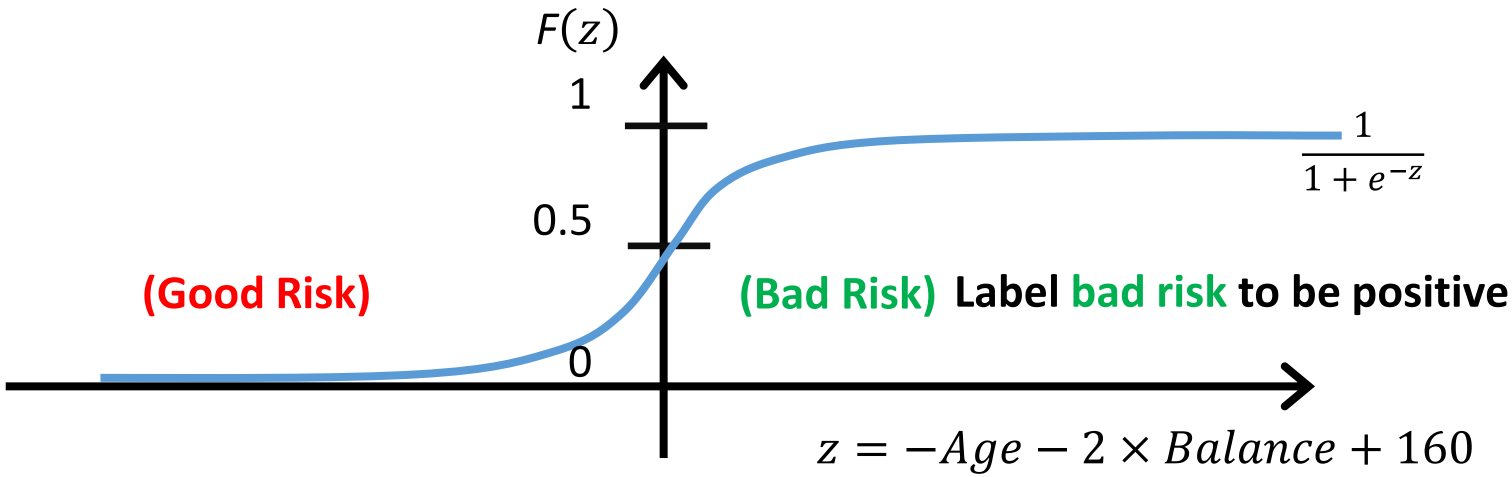


- Bad risk (Default) – 15 cases
- Good risk (Not default) – 17 cases

Classification by Logistic Regression

By default, we take the threshold as **0.5**:

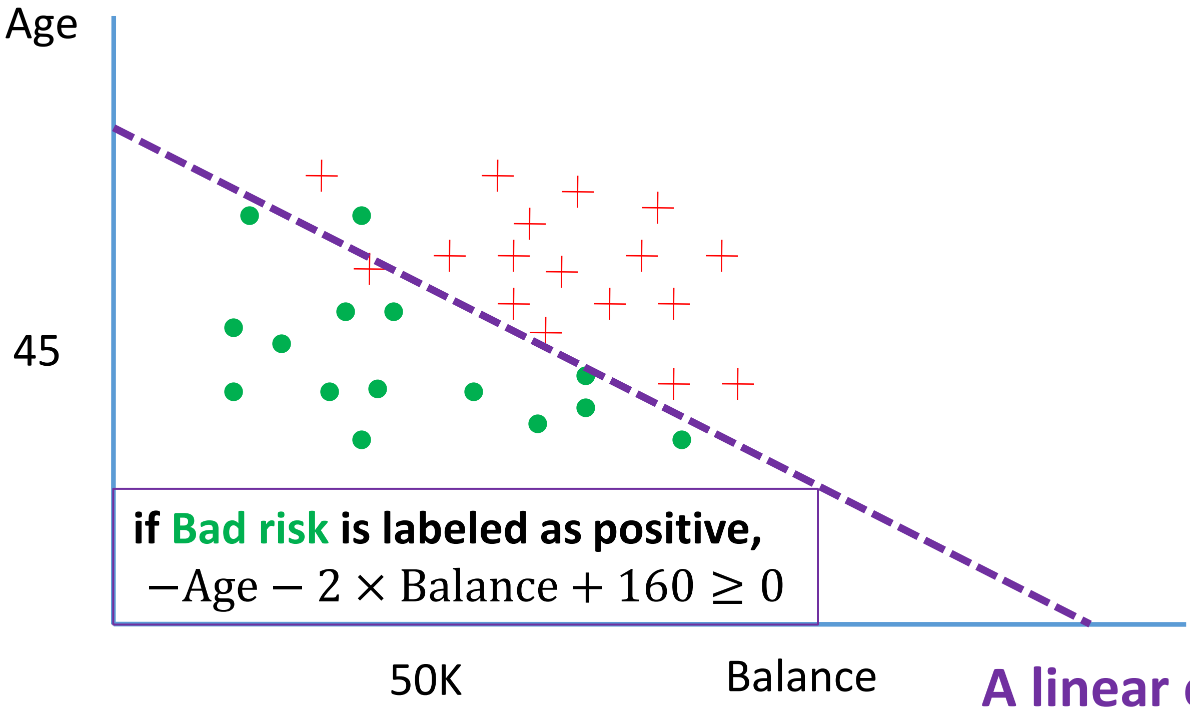
If $p = F(z) \geq 0.5$, predict "y = 1"; If $p = F(z) < 0.5$, predict "y = 0"



Logistic function $F(z) = \frac{1}{1+e^{-z}}$ where $z = -Age - 2 \times Balance + 160$

$$\text{Predict Class} = \begin{cases} \bullet \text{ if } F(z) \geq 0.5 \\ + \text{ if } F(z) < 0.5 \end{cases} \text{ or } \begin{cases} \bullet \text{ if } -Age - 2 \times Balance + 160 \geq 0 \\ + \text{ if } -Age - 2 \times Balance + 160 < 0 \end{cases}$$

Linear Discriminant Analysis

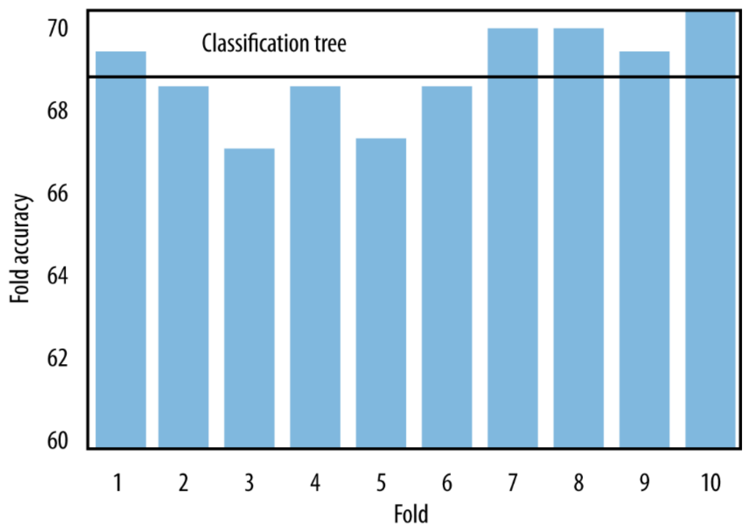


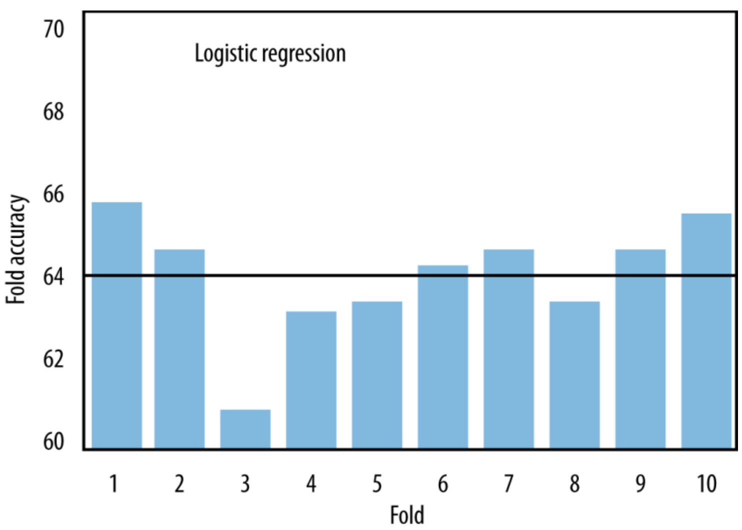
A linear classifier

- **Bad risk (Default) – 15 cases**
- **Good risk (Not default) – 17 cases**

Revisit: Cross-validation by Decision Tree and Logistic Regression

Which customers should TelCo target with a special offer, prior to contract expiration?



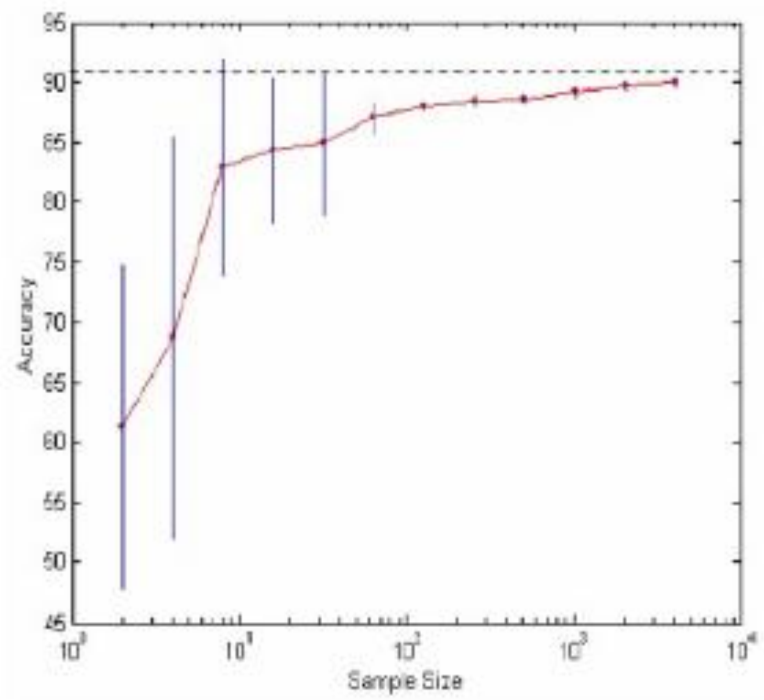


This dataset contains 20,000 examples

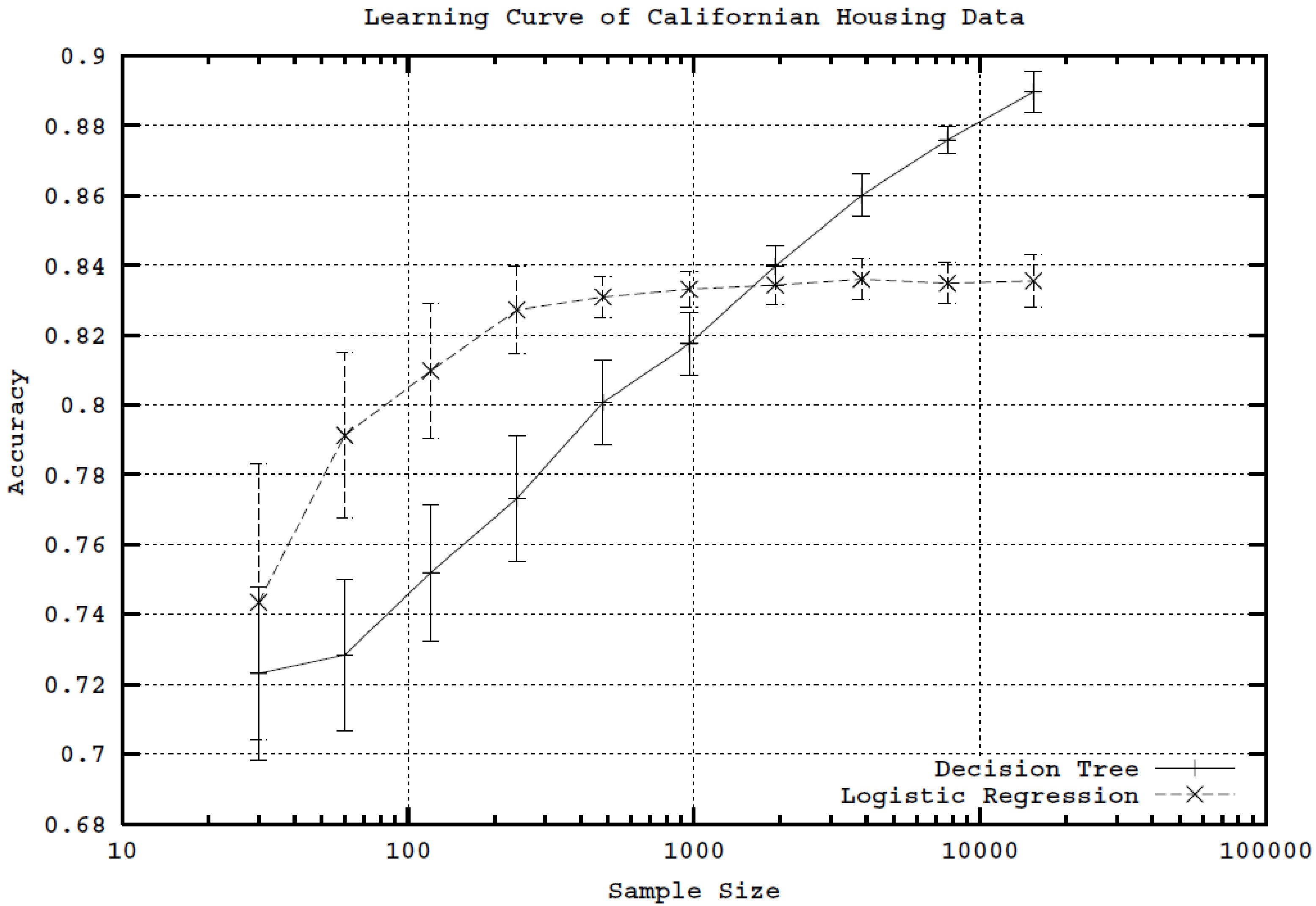
Learning Curve

> Different modeling procedures may have different performance on the same data.

> A learning curve shows how the generalization performance changes with varying **sample size**!



Learning Curve Comparison



Decision Tree vs. Logistic Regression

> For smaller training-set sizes, logistic regression yields **better** generalization accuracy than tree induction

- For smaller data, tree induction will tend to overfit more

> Classification trees are a **more flexible** model representation than linear logistic regression

> Flexibility of tree induction can be an advantage with larger training sets:

- Trees can represent substantially **nonlinear relationships** between the features and the target

Decision Tree vs. Logistic Regression

> What is more **comprehensible** to the stakeholders?
  - Rules? A numeric function?

> **How much data** do you have?
  - There is a key tradeoff between the complexity that can be modeled and the amount of training data available
  - Trees need a lot of data to approximate curved boundaries

> What are the **characteristics** of the data: missing values, types of variables (numeric, categorical), relationships between them, how many are irrelevant, etc.
  - Trees are fairly robust to these complications

> Do you need a good estimate of **class probabilities**?
  - Logistic regression generates probabilities in a more sophisticated way.

Thank You !