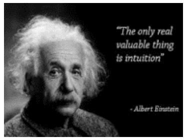
Introduction to Classification Models

YC(×) – Y is not continuous, but categorical

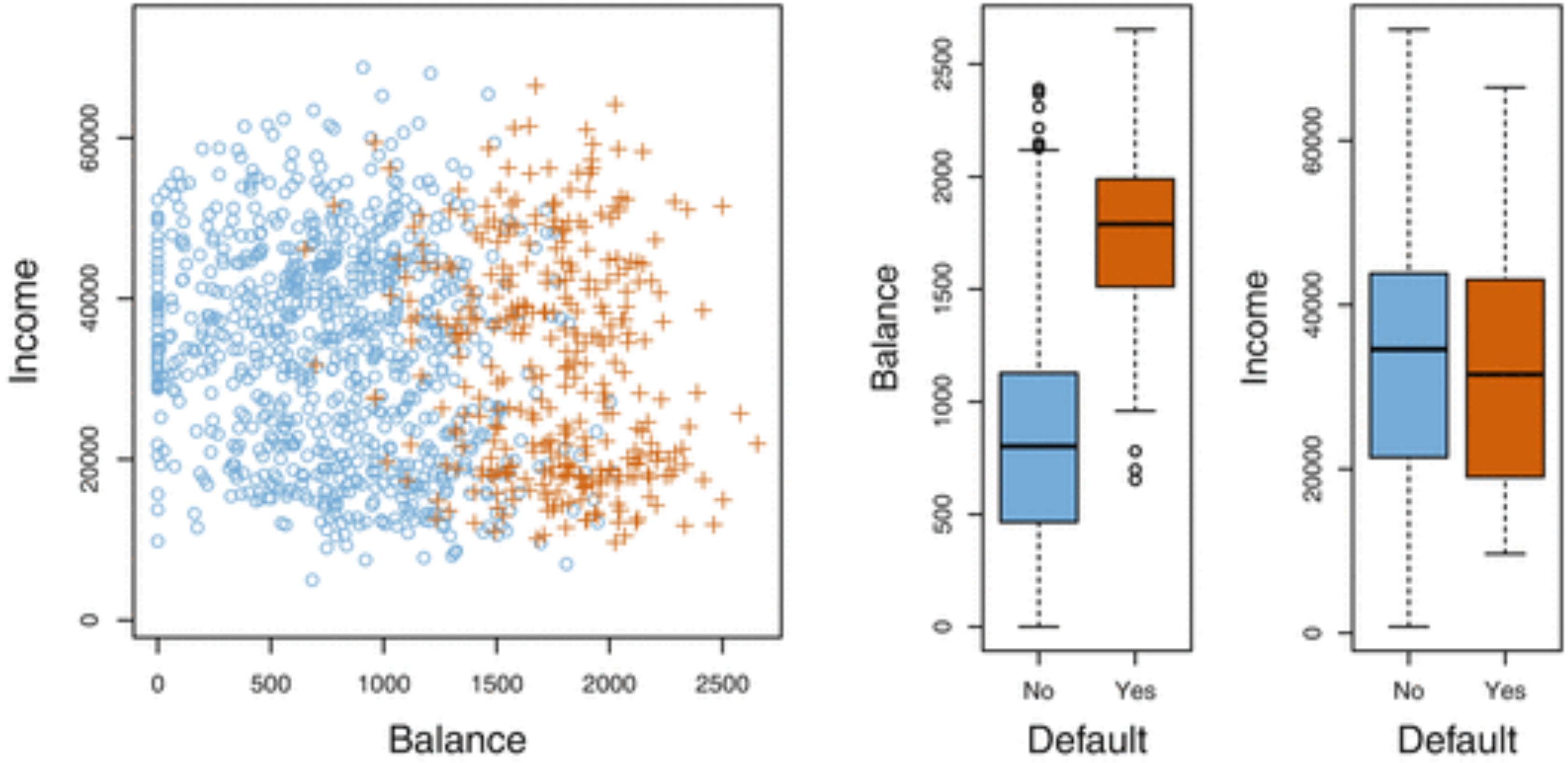


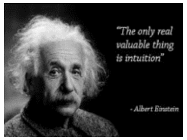
Classification Models: Intuition

- A classification model is one in which the **outcome** variable is **categorical**, and the model aims at predicting when an observation will fall in one category or another.
- In essence, **predicting** a **qualitative response** is **equivalent** to **classifying** that observation to a category or class
- **Examples:** is an e-mail message spam? is a loan customer likely to default? is a given purchase transaction fraudulent?
- **Another example:** how would you predict who will make an A in the class? Popular approaches to this problem include:
  - **Logistic Regression** – Binomial and Multinomial
  - **Linear & Quadratic Discriminant Analysis**
  - **K Nearest Neighbors**
  - **Decision Tree Models**
  - **Support Vector Machines**

Classification Illustration

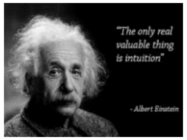
The classification method is **visually** illustrated below. In this example, it is abundantly clear that a person's outstanding balance is strongly associated with whether he/she defaults on a loan or not, whereas the person's income is not.





Binomial vs. Multinomial Outcome

- A **categorical outcome** has various possible **classifications**
- Some categories are **ordinal** – e.g., bad, acceptable, good
- Some are **categorical** – e.g., rural, suburban, urban
- Most categorical outcomes, even if ordinal, are difficult to quantify, so traditional regression methods cannot be used
- If the outcome variable is **binomial** – i.e., has two possible outcomes, it can be easily quantified with **dummy variables** (e.g., 0 – no loan default; 1 – loan default)
  - ➢ Most binomial models (e.g., logistic regression, decision trees) will then predict the **probability** of the outcome being a 0 or a 1
- If the outcome is **multinomial** – i.e., has more than 2 possible outcomes (e.g., stroke, heart attack, no illness), then the outcomes **cannot** be **easily quantified**
  - ➢ There are **models specifically tailored** to handle multinomial classification



Binomial Classification and MLE

- **MLE** is a popular estimation method for **binomial classification**
- The detailed math is covered in the textbook, but we present a **simple** explanation **MLE** using **binomial classification** (e.g., logistic regression) with a **single predictor**.
- For **2 independent** observations A and B, the probability of both happening is equal to the probability of A **times** the probability of B
- So, if an outcome variable can take a value of **0** or **1**, the probability of all observations x_i being classified correctly is equal to the **product** of the **probability** of **each** observation being classified correctly, either as **1** → $P(x_i)$; or as **0** → $1\text{-}P(x_i)$:

$$Likelihood(\beta_0,\beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

- The **MLE** method computes this likelihood for several values of β_0 and β_1 using **algorithms** and selects the ones where the likelihood of a correct classification is the **largest**

KOGOD SCHOOL
of
BUSINESS