Feature Importance in Decision Trees:

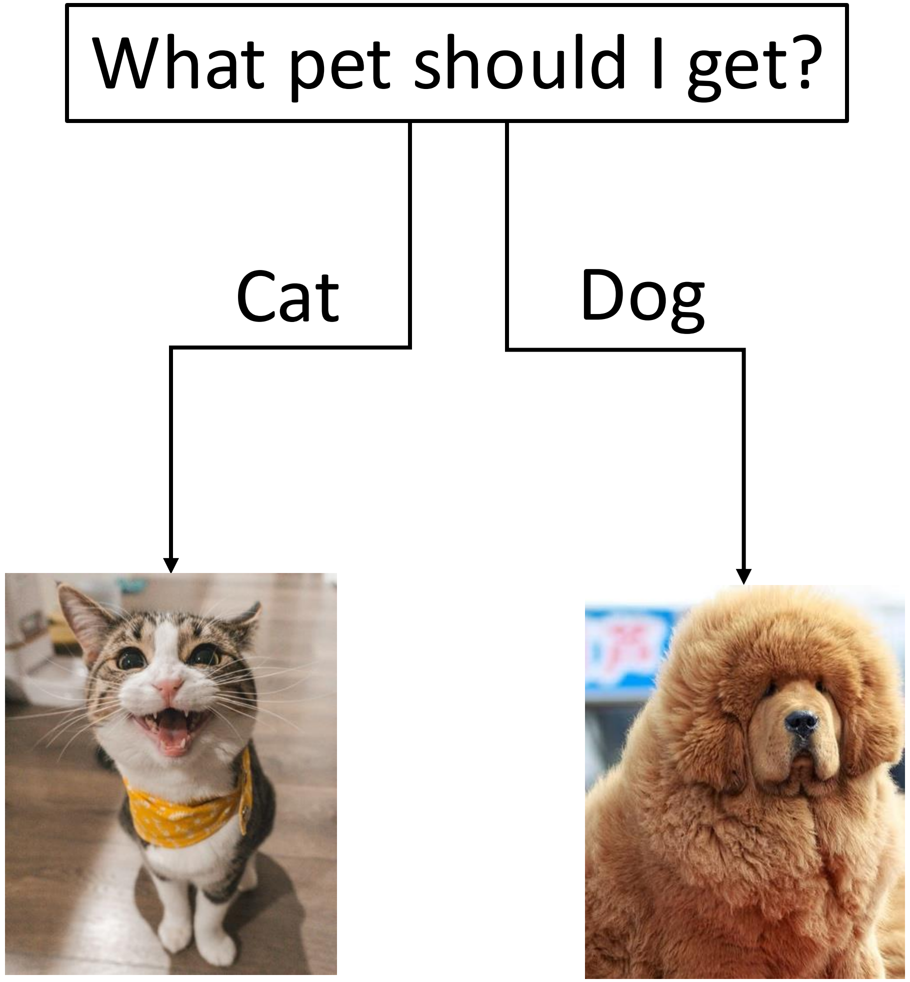
Impurity-based Importance Calculations & Explainable AI

Dr. Franziska Boenisch

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

SprintML

You already interact(ed) with decision trees!



May 11, 2023 2m read

Random Forests: Netflix Customer Recommendations Improved by 20%

… watching Netflix …

Spotify — Decision Trees with Music Taste

7 min read · Nov 26, 2020

J Jinkim Follow

… or listening to music.

Outline for Today

- Intuition on classification with decision tree
- Impurity-based feature importance metrics
- Building decision trees based on impurity reduction
- Feature importance and explainable AI

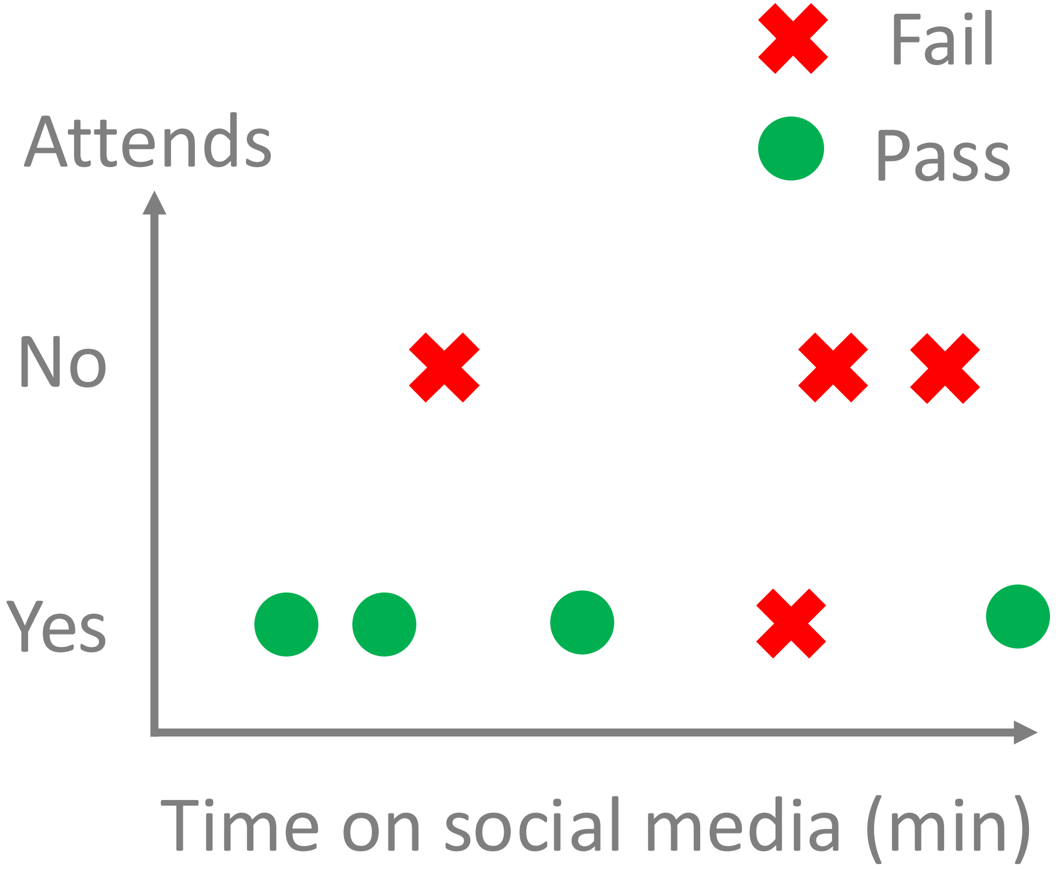
Datasets and Tree-based Classification

Dataset from my Lecture

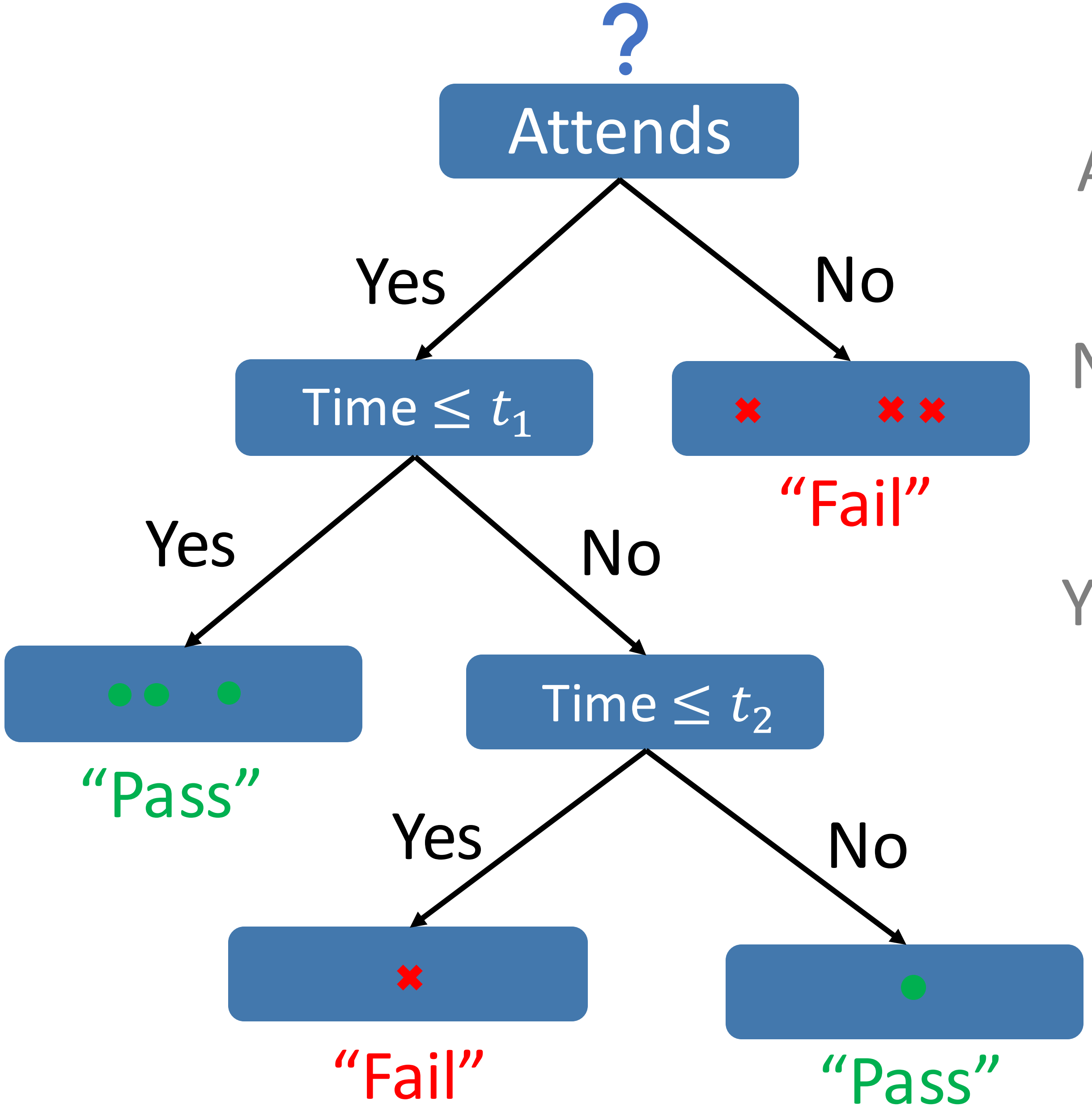
Social Media Time (min)	Attends Class	Passed the Midterm
30	Yes	Pass
80	Yes	Pass
140	Yes	Pass
50	Yes	Pass
110	No	Fail
60	No	Fail
100	Yes	Fail
120	No	Fail

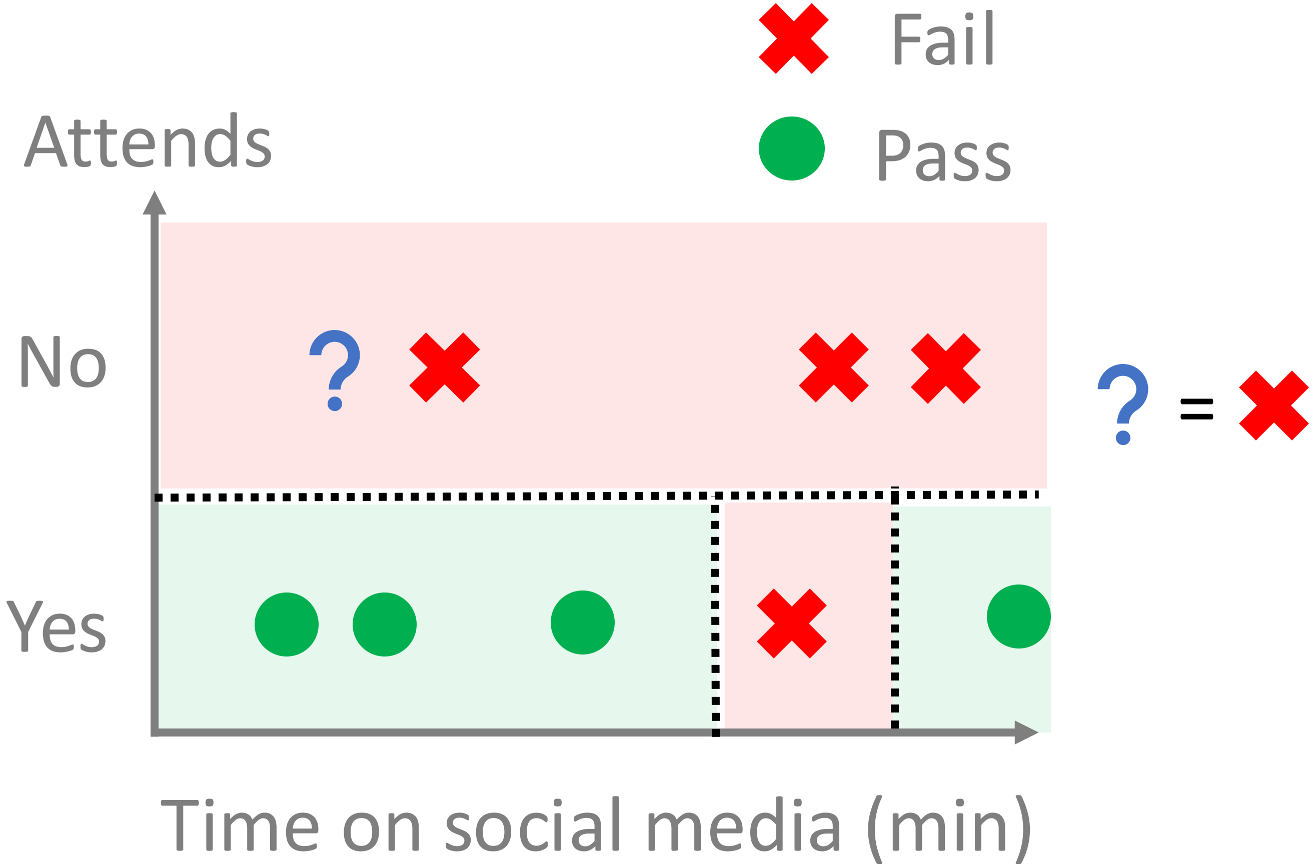
Continuous Categorical

Features Label



Datasets and Tree-based Classification



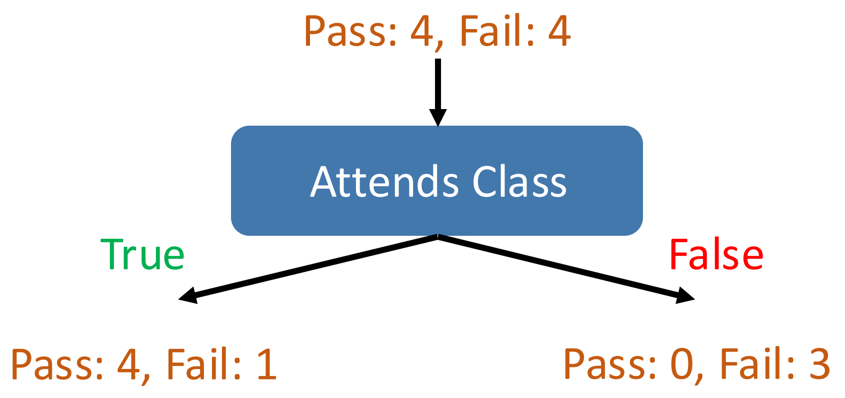


Successively find splits of the data to create decision regions

Greedy, recursive partitioning

Finding the Best Split Criterion

Social Media Time (min)	Attends Class	Passed the Midterm
30	Yes	Pass
80	Yes	Pass
140	Yes	Pass
50	Yes	Pass
110	No	Fail
60	No	Fail
100	Yes	Fail
120	No	Fail



Impure split

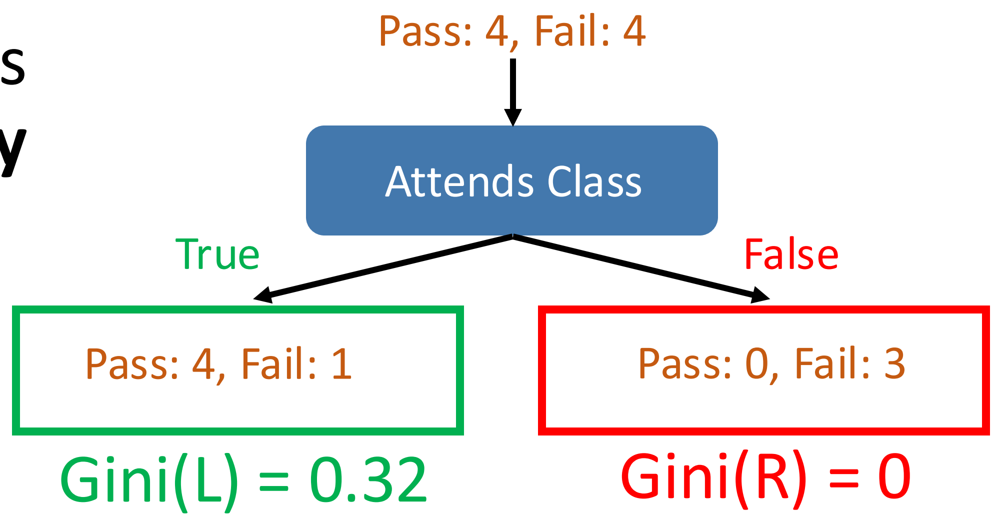
Pure split

Is attendance our best spilt?

Gini Impurity: Definition

Given a node with K classes and class probabilities $p_1, \dots, p_k$. The **Gini Impurity** is defined as

$$Gini(Node) = 1 - \sum_{k=1}^{K} p_k^2 .$$



Here: $1 - \left(p_{pass}^2 + p_{\text{fail}}^2\right)$

$p_{pass} = \frac{4}{5}, p_{fail} = \frac{1}{5}$

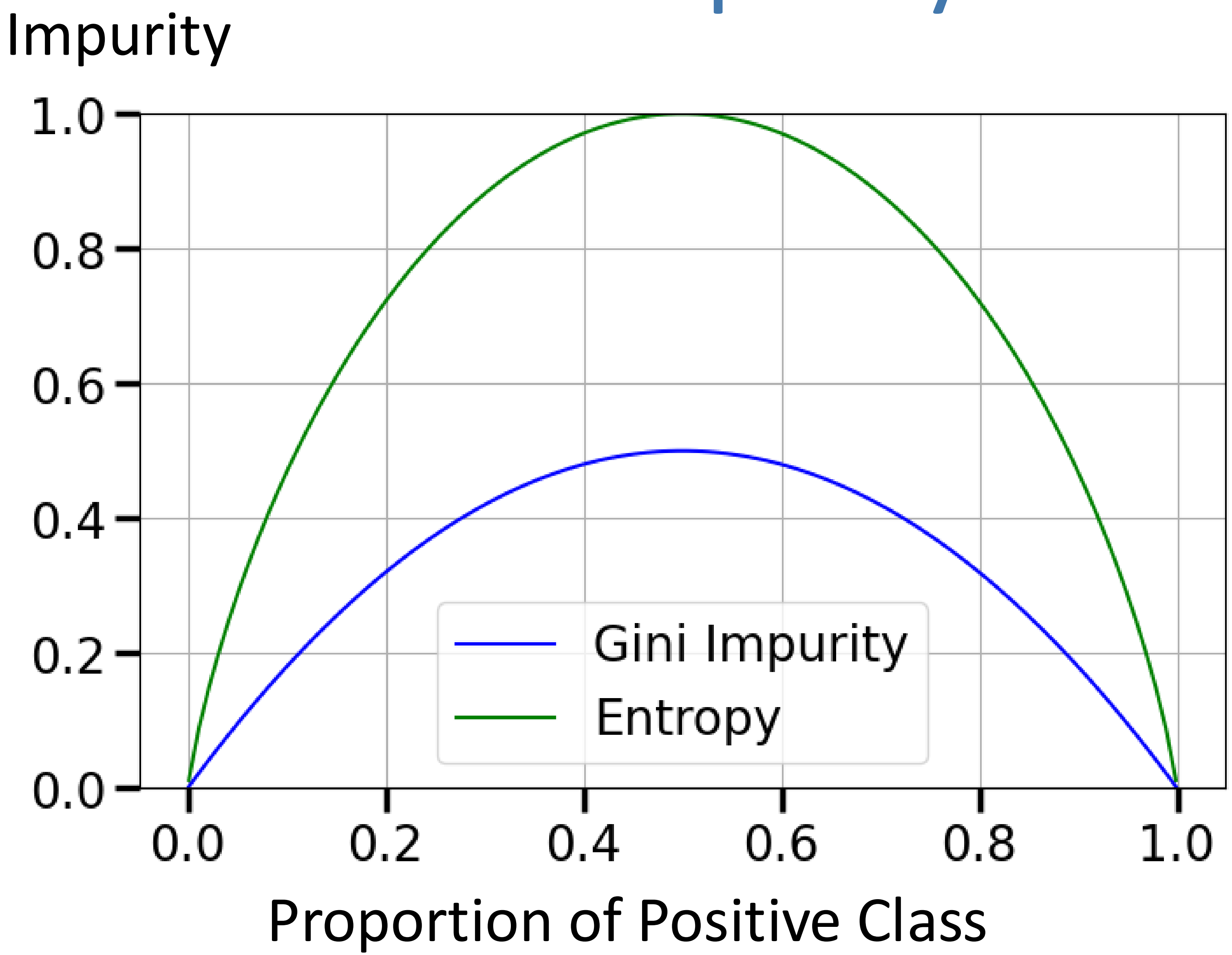
$p_{pass} = \frac{0}{3}, p_{fail} = \frac{3}{3}$

In our example:

- True branch: Gini(L) = $1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32$
- False branch: Gini(R) = $1 - (0)^2 - (1)^2 = 0$

Gini of a pure split is zero

Gini Impurity and Entropy



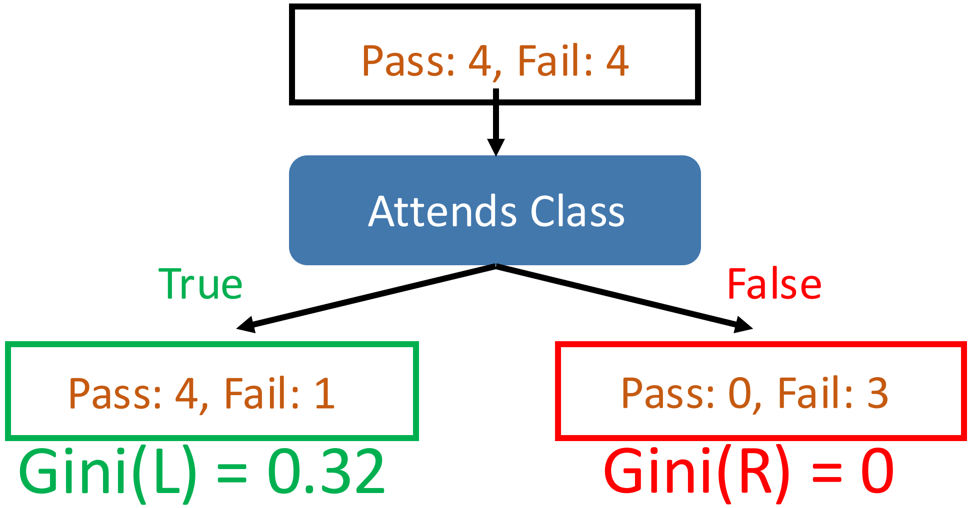
Alternative to Gini:

Entropy $= -\sum_k p_k \log_2 p_k$,

with p_k : proportion of data from class k in the node.

Gini Impurity of the Entire Split

When evaluating a split, we compute the **weighted Gini of the children**:



$$Gini_{split} = \frac{n_L}{n} Gini(L) + \frac{n_R}{n} Gini(R)$$

n : total number of data points in split → 8

n_L: number of points in L(eft) node → 5

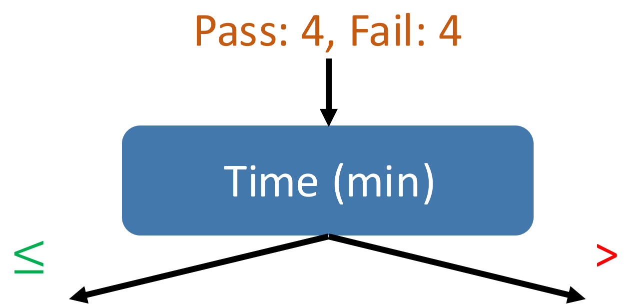
n_R: number of points in R(right) node → 3

$$Gini_{split} = 0.2$$

$$Gini_{split} = \frac{5}{8} * 0.32 + \frac{3}{8} * 0 = 0.2$$

Gini Impurity on Continuous Values

Goal: Identify the best splitting threshold



Social Media Time (min)	Passed the Midterm
30	Pass
50	Pass
60	Fail
80	Pass
100	Fail
110	Fail
120	Fail
140	Pass

1. Sort ascending

2. Identify where class changes

3. Take average as threshold

≤ 55

≤ 70

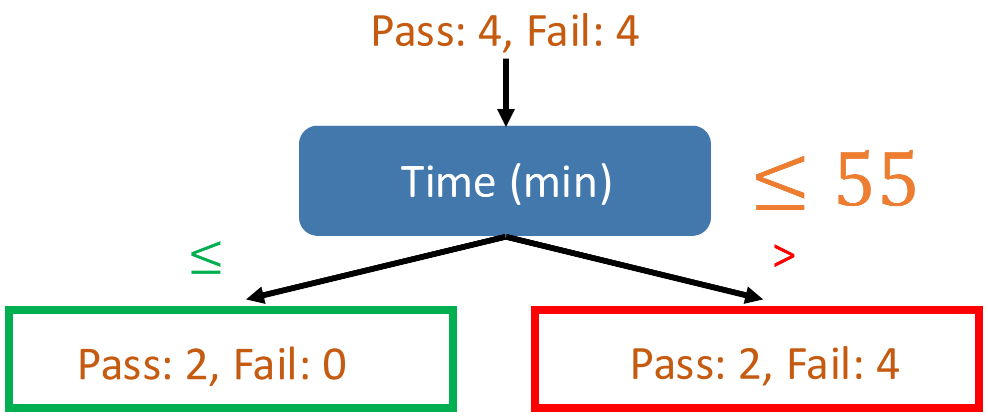
≤ 90

≤ 130

Gini Impurity on Continuous Values

Social Media Time (min)	Passed the Midterm
30	Pass
50	Pass
60	Fail
80	Pass
100	Fail
110	Fail
120	Fail
140	Pass

0.33
0.48
0.38
0.44

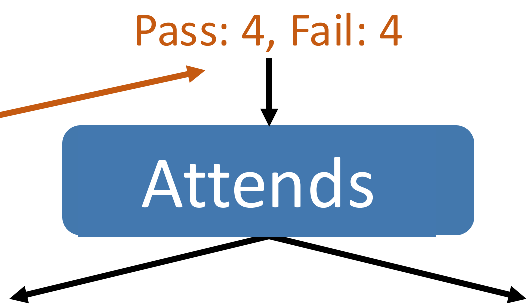


$Gini(L) = 0$ $\quad$ $Gini(R) \approx 0.44$

$$Gini_{split} \approx 0.33$$

Impurity Reduction to Choose the Best Split

Choose the split that causes the maximum **impurity reduction** $\Delta\mathrm{i}(\mathrm{split})$:



$$\Delta i\ (split) = \max(Gini_{parent} - Gini_{split})$$

$$Gini_{parent} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

is fixed → find lowest $Gini_{split}$

Gini impurity over all possible splits:

$$\boxed{Gini_{Attends} = 0.2}$$

$$Gini_{Time \leq 55} = 0.33$$

$$Gini_{Time \leq 70} = 0.48$$

$$Gini_{Time \leq 90} = 0.38$$

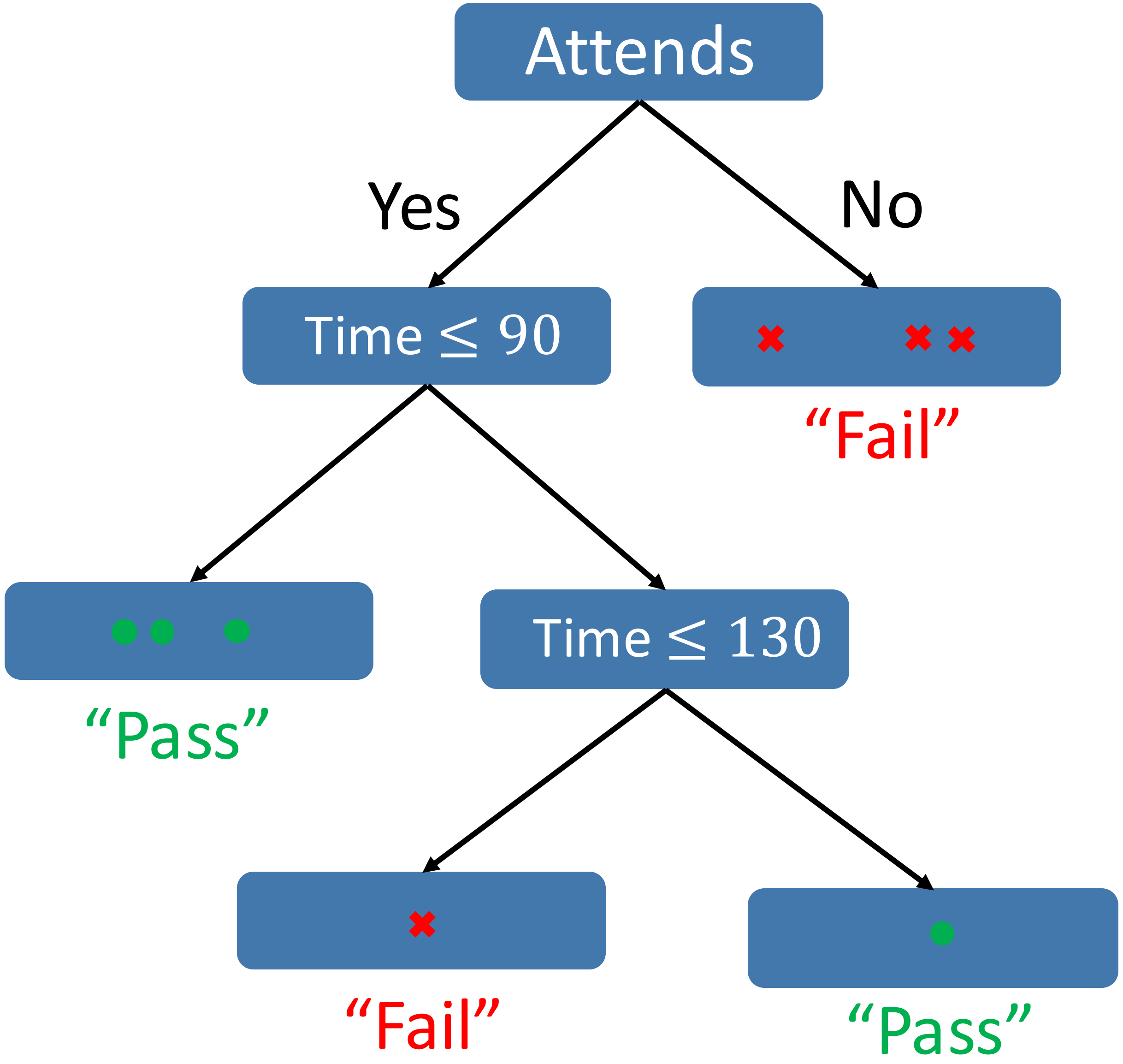
$$Gini_{Time \leq 130} = 0.44$$

$$\Delta i\ (Attends) = 0.5 - 0.2 = 0.3$$

From Trees to Explainable AI

Decisions in the tree are:

- Human-interpretable
- Verifiable
- We can ask "What if?" (Counterfactuals)



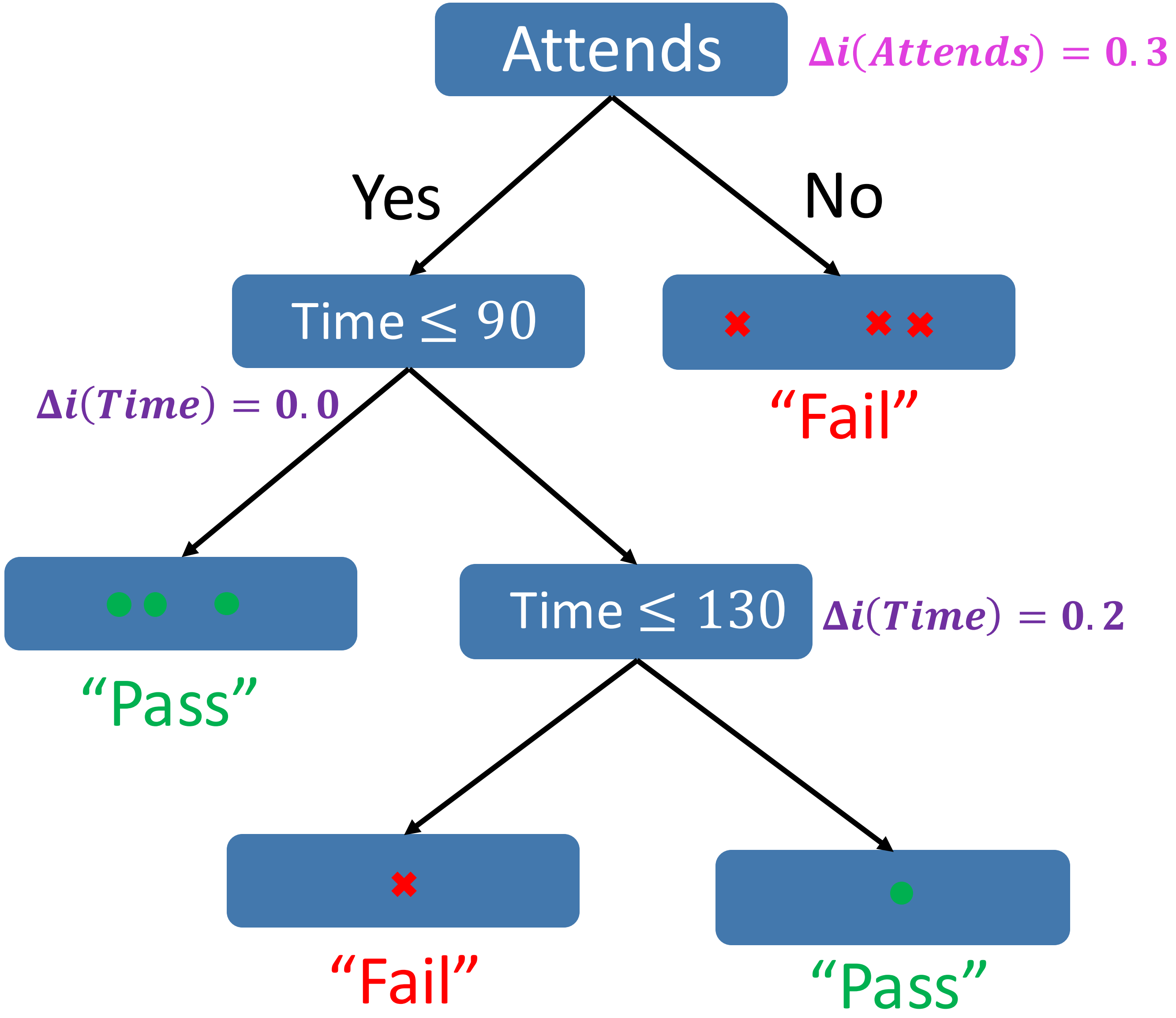
Impurity-based Feature importance

We calculate the **importance** of a feature f in a tree as:

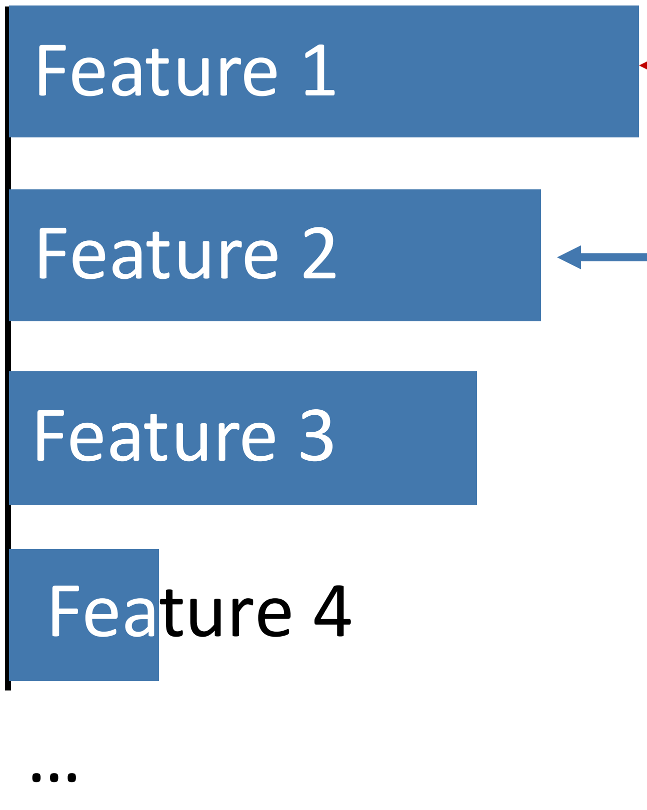
$$Importance(f) = \frac{\sum_{t \in Splits\ on\ f} \Delta i(t)}{\sum_{s \in All\ splits} \Delta i(s)}.$$

$$Importance(Attends) = \frac{0.3}{0.3 + 0.0 + 0.2} = 0.6 = 60\%$$

$$Importance(Time) = \frac{0.0 + 0.2}{0.3 + 0.0 + 0.2} = 0.4 = 40\%$$



Feature Importance for Explainability



ZIP Code for
Credit Application

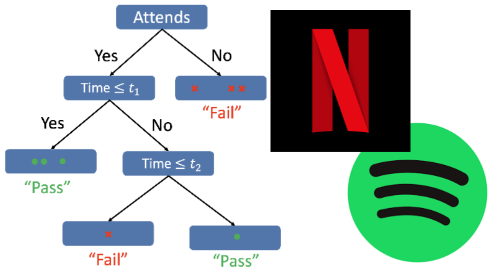
Noisy?
Weak?

Understanding predictions

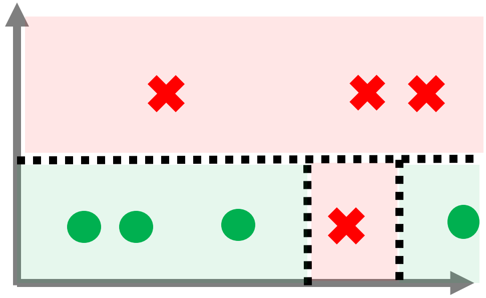
Model debugging

Identifying biases

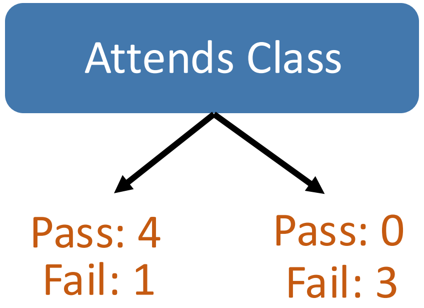
Summary & Lecture Materials



Decision Trees:
Omnipresent



Divide Data
in Regions



Impurity-based
Feature Splits

Serve
Explainable AI

Lecture Materials:

SCAN ME