Decision Trees & Gini Impurity

Failure Risk Prediction – Class Notes

Problem Setup: Failure Risk Prediction

- Goal: Predict Failure_Risk of a component.
- Target (Y): Failure_Risk → 0 = Low Risk, 1 = High Risk.
- Features (X): Temperature, Vibration, Pressure, Age.
- We split data into train/test and train a DecisionTreeClassifier with max_depth = 3.

What is a Decision Tree?

- A model that makes predictions by asking a sequence of yes/no questions.
- Internal node: condition such as Age <= 48.5.
- Branch: True/False path from a condition.
- Leaf node: final prediction (Low Risk or High Risk).
- Tree can be visualized as a flowchart of if–else rules.

Gini Impurity: Intuition

- Each node is a group of samples (e.g., components).
- Some samples are Low Risk, some are High Risk.
- Gini measures how mixed the node is.
- Gini = 0 → perfectly pure (all are the same class).
- Higher Gini → more mixed; less certainty for a single decision.

Gini Impurity: Formula & Simple Examples

- For binary classes: $\text{Gini} = 1 - (p0^2 + p1^2)$.
- Example 1: 90 Low, 10 High → p0 = 0.9, p1 = 0.1 → Gini ≈ 0.18 (fairly pure).
- Example 2: 50 Low, 50 High → p0 = 0.5, p1 = 0.5 → Gini = 0.5 (maximally mixed).
- Lower Gini means a cleaner segment and more confident decisions.

Gini on Our Tree: Real Node Example

- Node path example: Age ≤ 48.5 → Vibration ≤ 7.934 → Vibration ≤ 0.521.
- Node values: samples = 508, value = [486 Low, 23 High].
- Proportions: p Low ≈ 0.957, p High ≈ 0.045.
- Gini $\approx 1 - (0.957^2 + 0.045^2) \approx 0.083$ → very pure node.
- Interpretation: young machines with very low vibration are almost always Low Risk.

How Splits Happen in a Decision Tree

- At each node, the algorithm considers all features and many thresholds.
- For each candidate split:
- • Split into left (feature ≤ threshold) and right (feature > threshold).
- • Compute Gini for left and right nodes and their weighted average.
- The split with the largest Gini reduction (Gini gain) is chosen.
- This process repeats recursively until max_depth or purity/size limits are reached.

Example: First Split in Our Tree

- Root node: samples = 800, value = [574 Low, 226 High], Gini ≈ 0.405.
- Split chosen: Age ≤ 48.5.
- Left node (Age ≤ 48.5): samples = 632, value = [561 Low, 71 High], Gini ≈ 0.199.
- Right node (Age > 48.5): samples = 168, value = [13 Low, 155 High], Gini ≈ 0.143.
- Weighted Gini after split ≈ 0.187 → Gini gain ≈ 0.218.
- This strong reduction in impurity makes Age ≤ 48.5 the best first question.

Business Interpretation of Gini & Nodes

- Low Gini, Low-Risk node: segment behaves consistently as Low Risk → de-prioritize for urgent maintenance.
- Low Gini, High-Risk node: segment behaves consistently as High Risk → prioritize for inspections and preventive actions.
- Higher Gini nodes: mixed behaviour (grey zones) → need more data or careful policies.
- Gini helps us see where we can take clear, confident business actions for each segment.

Role of max_depth in the Tree

- In our code: max_depth = 3 for DecisionTreeClassifier.
- Limits the number of levels of questions the tree can ask.
- Smaller depth → simpler, more general model; less risk of overfitting and easier to explain.
- Larger depth or no limit → more detailed splits but higher risk of memorizing noise.
- Depth = 3 is a good balance for teaching and interpretability.

Summary: Key Takeaways

- Decision Trees are flowcharts of if–else questions that lead to a prediction.
- Gini impurity measures how mixed a node is; lower Gini means purer segments.
- The tree chooses splits that reduce Gini impurity the most at each step.
- In our failure risk example, Age, Vibration, and Temperature create meaningful risk segments.
- Use low-Gini segments to drive clear business decisions in maintenance planning.