

Business Data Analysis

MGM509

Lecture 2

Predictive Modeling Foundations:
Correlation to Segmentation

Prepared by
Eng. Ahmed Hussein, MBA, PfMP®,
PMP®, CMP APMG

Lecture
2

What we will cover

1. The purpose and power of predictive models
2. What makes an attribute “informative”
3. Measuring data uncertainty
4. How decision trees use data to create business insights

Lecture
2

What is Predictive Modeling?

Using Data to Predict Outcomes

- Creating models to forecast unknown outcomes, based on historical data with known outcomes
- Used in many business problems as:
 - Churn prediction
 - Fraud detection
 - Credit scoring
 - Demand forecasting

It's not about explaining the past — it's about guiding future decisions.

Lecture 2

Predictive vs. Descriptive Modeling

Feature	Predictive Modeling	Descriptive Modeling
Purpose	Forecast outcomes	Understand Past Patterns
Uses Labeled Data	Yes	Not necessarily
Common in	Business decision support	Reporting & Analysis
Output	Predictions/scores	Insights/segments

Predictive modeling supports action, not just understanding.

Lecture
2

Models, Induction, and Prediction

- Model: A simplified representation of reality
- A Predictive Model: A formula for estimating the unknown value of interest (**the Target**). Formula could be Mathematical or Logical or a hybrid
- Model Induction: Creation of models from data, or generalizing from specific cases to general rules).
- Prediction: Using a model to estimate an unknown value
Example:
 - Data: 1000 past customers
 - Induction: Customers with < 1 year tenure are likely to churn
 - Prediction: New customer with 8 months tenure is likely to churn

Lecture
2

Supervised Learning

Learning from Examples

- Model creation where the model describes a relationship between a set of **selected variables (attributes or features)** and a predefined variable called the **target variable**
- Train a model using data with known outcomes.
- Components:
 - Instance or Feature vector (example, data point, row)
 - Attributes (Features)
 - Target Variable (Label)

Example

Attribute	Target Variable (Churn?)
Monthly Charges: \$100	Yes
Contract Type: One-year	No

Lecture
2

The Target Variable

What are you trying to predict

- Central to supervised learning
- Must be clearly defined and measurable
- Examples:
 - Customer will churn (Yes/No)
 - Loan will default (Yes/No)
 - Expected monthly revenue (numeric)

The model's job is to learn how attributes relate to the target!

Lecture
2

Predictive Modeling as Supervised Segmentation

Think of Predictive Modeling as Supervised Segmentation

- How can we segment the population into groups that differ from each other with respect to some quantity of interest (or something that we would like to predict or estimate).
- The target of interest is unknown at decision making time.
- Finding or selecting important **informative variables** or “**attributes**” of the entities described by the data, or finding knowable attributes that correlate with the target of interest and reduce our uncertainty in it.

Informative means something that reduces uncertainty

Lecture
2

What is Supervised Segmentation?

Dividing data to improve prediction

- Split the population into segments where behavior is more consistent using attribute values to define segments
- Breaks down data into predictable subgroups
 - Uses labeled data to guide the process
 - Sets the stage for:
 - Attribute selection
 - Decision tree construction
 - Rule generation
- Goal: Identify areas where prediction is more certain

Segmentation helps uncover meaningful patterns in messy data.

Lecture
2

Why Segment?

- Increases prediction accuracy

Enables targeted actions

Simplifies complex problems

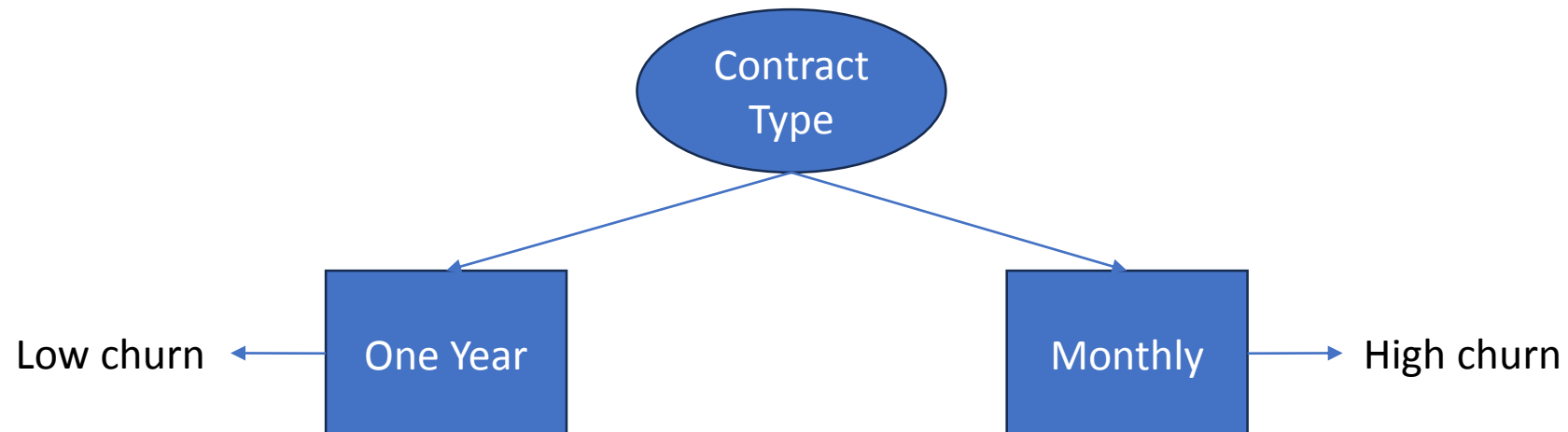
Examples:

- Churn risk may depend heavily on tenure for low-data users, but not for heavy users
- Loan default may depend on age, employment status, and loan amount, but not on location or marital status.

Lecture
2

Supervised Segmentation in Practice

- Begins with a target variable and searches for attributes that best split the data by the target
- Recursive: Splits again within each segment to refine predictions
- Example (Telco Customers Dataset):



Lecture
2

How Segmentation Helps Prediction?

- **Key Question: Which attribute to split on?**
- We need a way to rank attributes by how useful they are
- Look for attributes that reduce uncertainty about the target
- **Goal:** Find the attribute (**Informative Attribute**) that gives the cleanest separation between churners and non-churners



Lecture
2

What Makes an Attribute Informative?

- Strongly correlates with the target variable

Helps predict the target better than chance

Not all correlations are helpful — we care about predictive power

- Examples:
 - Useful: “Tenure” for churn prediction
 - Useless: “Customer ID” (unique for each person)

Lecture
2

Goal of Attribute Selection

Choosing the Right Splitting Attribute

- We want to reduce uncertainty about the target variable
- Best attribute = the one that creates the purest subgroups
- Evaluate based on how well an attribute separates classes

The more predictable the subgroup, the better the split.

Lecture
2

Measuring Uncertainty using Entropy

- **Entropy**: A measure of disorder or unpredictability in data
- “How uncertain we are before making a prediction”
- High entropy = mixed outcomes (impure), Entropy is highest when split is 50/50 (pure uncertainty)
- Low entropy = mostly one class (pure), Entropy is 0 when all examples are the same (pure certainty)
- Examples:
 - Dataset with 50% churners = high entropy
 - Dataset with 95% churners = low entropy

Lecture
2

Measuring Entropy

- Entropy(S) = $- P_1 \log (P_1) - P_2 \log (P_2) - \dots$
- Where: P_i is the probability (the relative percentage) of property (i) in the dataset
- $P_i = 1$ when all members of the set have property (i),
- and $P_i = 0$ when no members of the set have property (i).

Lecture
2

Information Gain

- How Much Does an Attribute Reduce Uncertainty?
- Information Gain (IG) = Reduction in entropy after the split
- Chooses the attribute that gives the greatest information gain
- Formula:

Information Gain (IG) = Entropy (before split) – Weighted Entropy (after split)

Or

Information Gain (parent, children) = $\text{entropy}(\text{parent}) - P(c1) \times \text{entropy}(c1) - P(c2) \times \text{entropy}(c2) - \dots$

Lecture 2

Example – Calculating Gain

Attribute: Contract Type	# Churn	# No Churn	Total
Month-to-Month	80	20	100
One-Year	10	90	100
TOTAL	90	110	200

- Entropy Calculation:

	P (Churn)	P (No Churn)	Entropy
TOTAL (before split)	$90/200 = 0.45$	$110/200 = 0.55$	$-0.45 \log (0.45) - 0.55 \log (0.55) = \text{approx } 0.991$
Month-to-Month	$80/100 = 0.8$	$20/100 = 0.2$	$-0.8 \log (0.8) - 0.2 \log (0.2) = \text{approx. } 0.722$
One-Year	$10/100 = 0.1$	$90/100 = 0.9$	$-0.1 \log (0.1) - 0.9 \log (0.9) = \text{approx. } 0.469$

Lecture
2

Example – Calculating Gain

- Entropy(after split) = $100/200 \times 0.722 + 100/200 \times 0.469 = 0.5955$
- IG= Entropy(total) – Entropy(after split) = $0.991 - 0.5955 = 0.3955$
- **What does this mean?**
- Contract Type is a useful predictor of whether a customer will churn:
 - Customers with Month-to-Month contracts are more likely to churn.
 - Customers with One-Year contracts are more likely to stay.
- The Information Gain of 0.396 means that Contract Type gives you a moderately strong signal that helps reduce uncertainty about churn behavior.

Lecture
2

Other Measures of Informativeness

- Alternatives to information gain:
 - Gini Index (used in CART)
 - Chi-square statistic
 - Variance reduction (for numeric targets)

All aim to select the attribute that leads to the most useful partitioning

Lecture
2

Attribute Selection – Real-World Churn Example

- Dataset: Telecom customer data with known churn outcomes
- Candidate attributes:
 - Tenure
 - Monthly charges
 - Contract type
 - Customer service calls
 - Data usage
- Which attributes best predict whether a customer will churn?

Lecture
2

Ranking Attribute Informativeness

Attribute	Information Gain
Contract Type	0.25
Customer Service Call	0.18
Tenure	0.12
Monthly Charges	0.05
Gender	0.0

- Contract Type is most informative
- Gender has no predictive power
- Not all available data is useful for prediction

Lecture
2

What Makes “Contract Type” SO Useful?

- Clear behavioral patterns:

• Month-to-month
= higher churn

One-year or two-
year = lower churn

Strong, non-
random split

Easily actionable
insight for
retention
strategies

Lecture
2Feature Importance – Interpret with
Caution

High information gain $\neq$ causal relationship

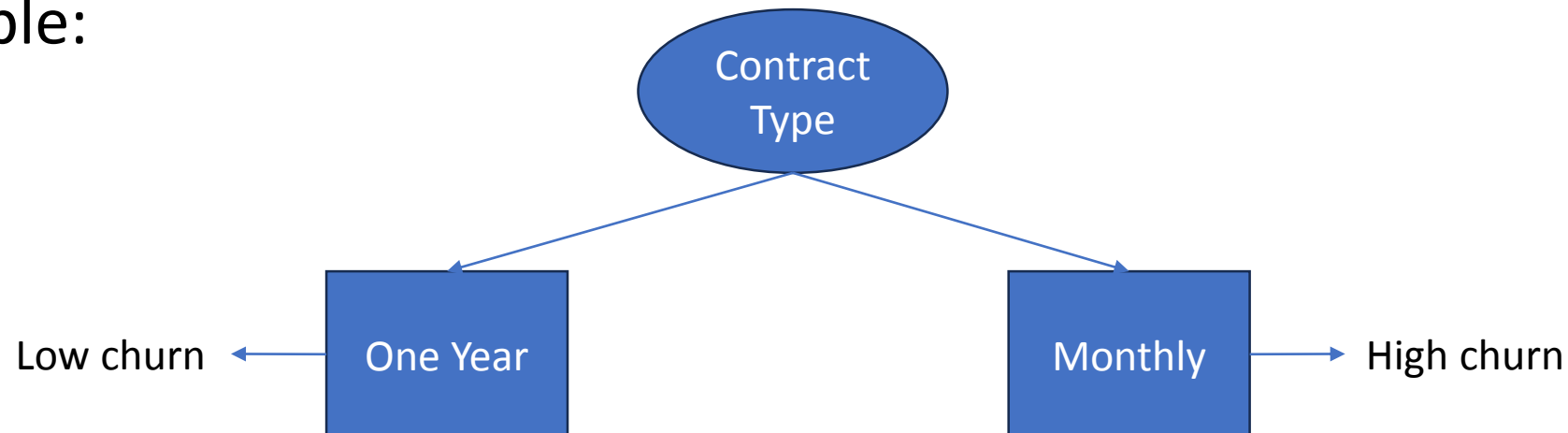
- Beware of:
 - Correlated attributes
 - Leakage variables (e.g., refund requested after cancellation)
 - Use domain knowledge to validate statistical findings

Good features are predictive, not necessarily explanatory

Lecture
2

What is a Decision Tree?

- **Tree-Structured Predictive Models:** Recursive segmentation of data based on attribute values
- Each node splits the data by the most informative attribute
- Ends in leaves that assign a prediction or probability
- Example:



Lecture
2

How Decision Trees are Constructed

Recursive Binary Splitting

• Choose the attribute with highest information gain



Split the dataset based on that attribute



Repeat the process on each branch

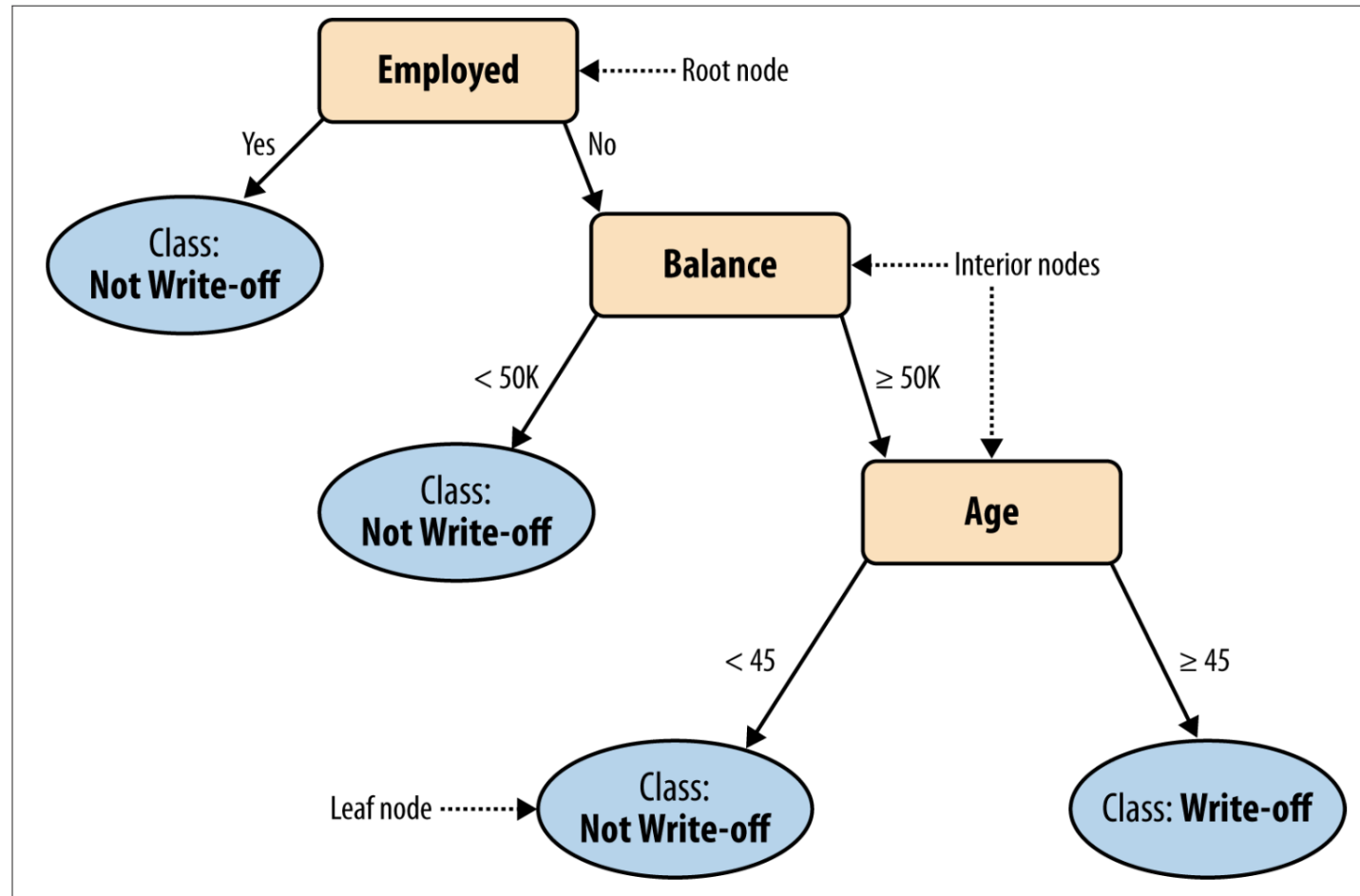


Stop when:

- No more informative attributes, or
- Segment is pure enough, or
- Tree is too deep (risk of overfitting)

Lecture 2

Sample Decision Tree



Lecture
2

Overfitting – A Common Pitfall

- Deep trees may fit the training data too well but Performs poorly on new, unseen data
- Too specific to training data
- Poor generalization
- False sense of performance
- **Solution:**
 - Use validation data
 - Prune trees
 - Apply complexity penalties

Lecture
2

When to Stop Splitting

Prevent Overfitting

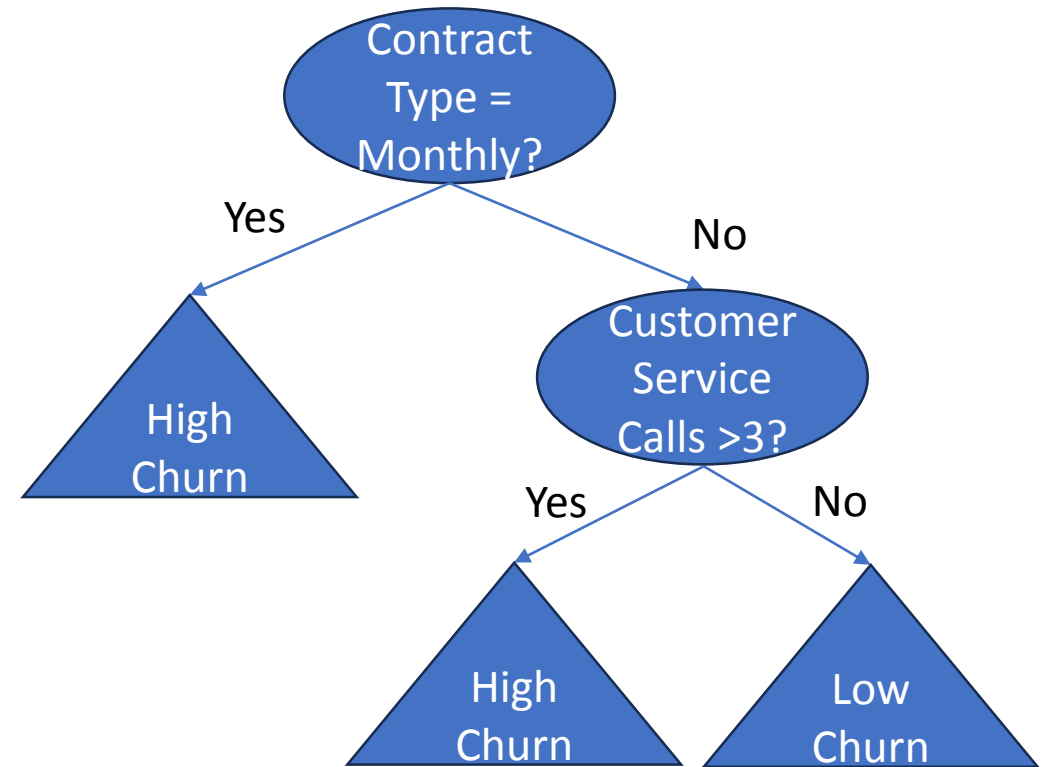
- Stop tree growth when:
 - Nodes are too small
 - Splits don't improve prediction
 - Maximum depth is reached
 - Or use post-pruning to cut back a large tree

Simpler trees are often more robust

Lecture
2

Visualizing a Decision Tree

- A Map of Predictive Segmentation
- Tree diagrams show:
 - Nodes = decision points
 - Branches = attribute-based splits
 - Leaves = predictions



Lecture
2

Trees as Sets of Rules

Each Path = One IF–THEN Rule

- A path from root to leaf becomes a decision rule
- Examples:

IF Contract $\neq$ Month-to-Month AND CustServCalls > 3 THEN Predict
Churn = Yes

IF Contract = Month-to-Month THEN Predict Churn = Yes

Lecture
2

Benefits and Limitations of Rule-Based Interpretation

Benefits

Human-readable

Easy to explain to stakeholders

Easy to convert into:

- Business logic
- SQL filters
- Operational policies

Helps in:

- Auditing model logic
- Regulatory compliance
- Building trust in the model

Limitations

Can become large and complex

Deep trees are:

- Harder to interpret
- More prone to overfitting
- May miss interactions that ensemble models can capture

Balance interpretability with performance needs

Lecture
2

Beyond Classification – Estimating Probability

Trees Don't Just Say “Yes” or “No”

- Leaves can output probabilities, not just labels
- Based on training data in each leaf:

Example: 75 out of 100 customers in a leaf churned → Churn probability = 75%

Probabilistic predictions support risk-based decision-making

Lecture
2

Why Use Probabilities?

- Probabilities allow businesses to act based on confidence, not just classification
- Enables graded responses:
 - 90% churn → Offer retention deal
 - 10% churn → No action needed
- Useful for:

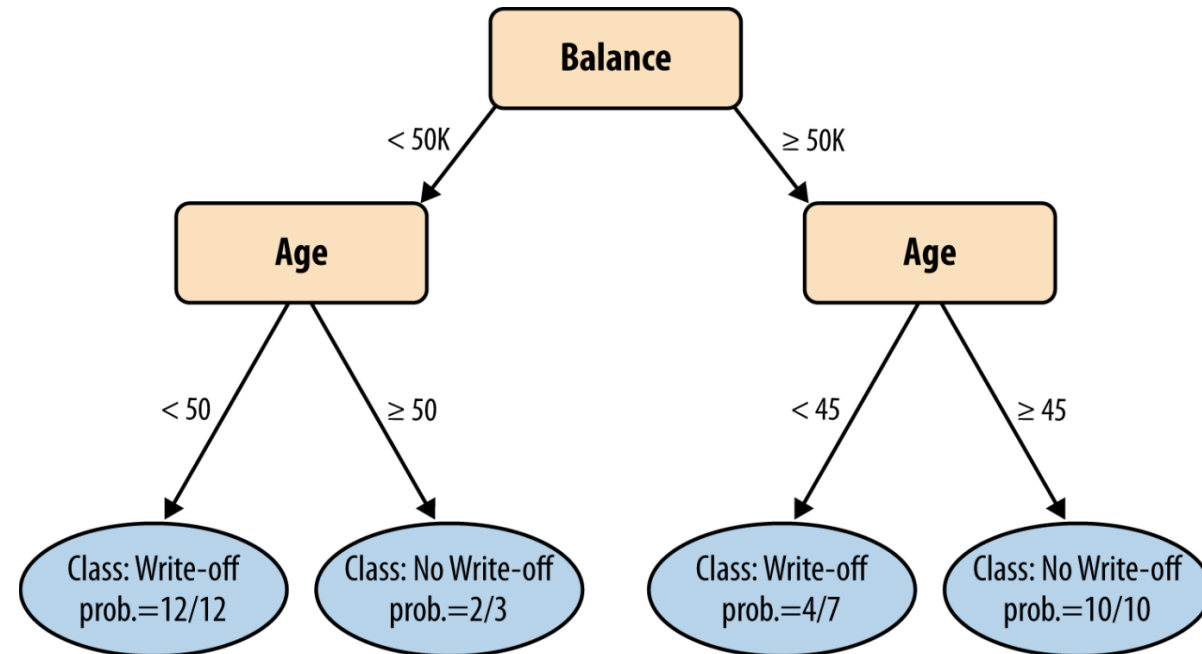
- Prioritization

- Cost-sensitive decisions

- Marketing campaigns

Lecture
2Example on Probabilities –
Which Loans to Write-off?

- New Customer, whose age is >45 and balance $>50K$
- Based on model:
 - What is the prediction?
 - What is the probability?



Lecture
2

How Trees Estimate Probability

- Probability estimates from trees are easy to interpret and use
- Estimate = Proportion of class labels in the leaf
- Can be improved with:
 - Laplace correction (smoothing)
 - Pruning to ensure stable estimates
- Important: More training data in the leaf → More reliable probabilities

Lecture
2

Case Study – Steps to Predict Customer Churn

- Applying a Decision Tree Model
- Business Problem:
“Which customers are likely to cancel their service?”
- Steps Followed:



Lecture
2

Understanding Results from the Tree Model

- Example prediction for New Customer:
 “Customer A” has a 73% probability of churn
- Segment defined by:
 - Month-to-month contract
 - 3 customer service calls
 - High monthly charge
- Interpretation:
 - This customer is at high risk → trigger retention strategy
 - Good starting point, but evaluate business impact too

Lecture
2

Business Implications

Decision trees help bridge analytics and operational action

- Turning Predictions Into Action:
 - Prioritize outreach to high-risk segments
 - Customize retention offers (e.g., loyalty incentives)
 - Track model performance over time

Lecture
2

Key Points

- Predictive modeling supports forward-looking business decisions
- Supervised segmentation divides data into meaningful subgroups
- Informative attributes reduce uncertainty about outcomes
- Decision trees are:
 - Interpretable
 - Easy to deploy
 - Powerful for classification and probability estimation

For Discussion Panel and Next Lecture:

- Use the open-source Telco Customer Churn dataset available on [Kaggle](#).
- Calculate Information Gain for each attribute and identify the most informative attribute(s)
- Visualize the Decision Tree Model

Lecture 2

For Discussion Panel and Next Lecture





Thank You