Machine Learning

Kelli Okumura, Zach Higa and Kody Kimoto

Agenda

- Overview of Machine Learning
- Strengths and Limitations of Machine Learning
- Foundations of Supervised Learning
- Decision Trees and Random Forest
- UHCC Retention Model
- Discussions

Overview of Machine Learning

Primary Methods of Machine Learning

Supervised Learning

Learning more about this today!

Unsupervised Learning

Deals with unlabeled data.

Identifies patterns and relationships without specific output labels.

Uses techniques like clustering and dimensionality reduction.

Reinforcement Learning

Learns by performing actions and receiving rewards or penalties.

Goal-oriented, used for learning sequences of actions to achieve specific objectives.

Algorithms for Unsupervised Learning

- Clusters
  - K-Means
  - DBSCAN
  - Hierarchical Cluster
- Anomaly or novelty detection
  - One-class SVM
  - Isolation Forest
- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE)
  - T-distribution Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
  - Aprioiri
  - Eclat

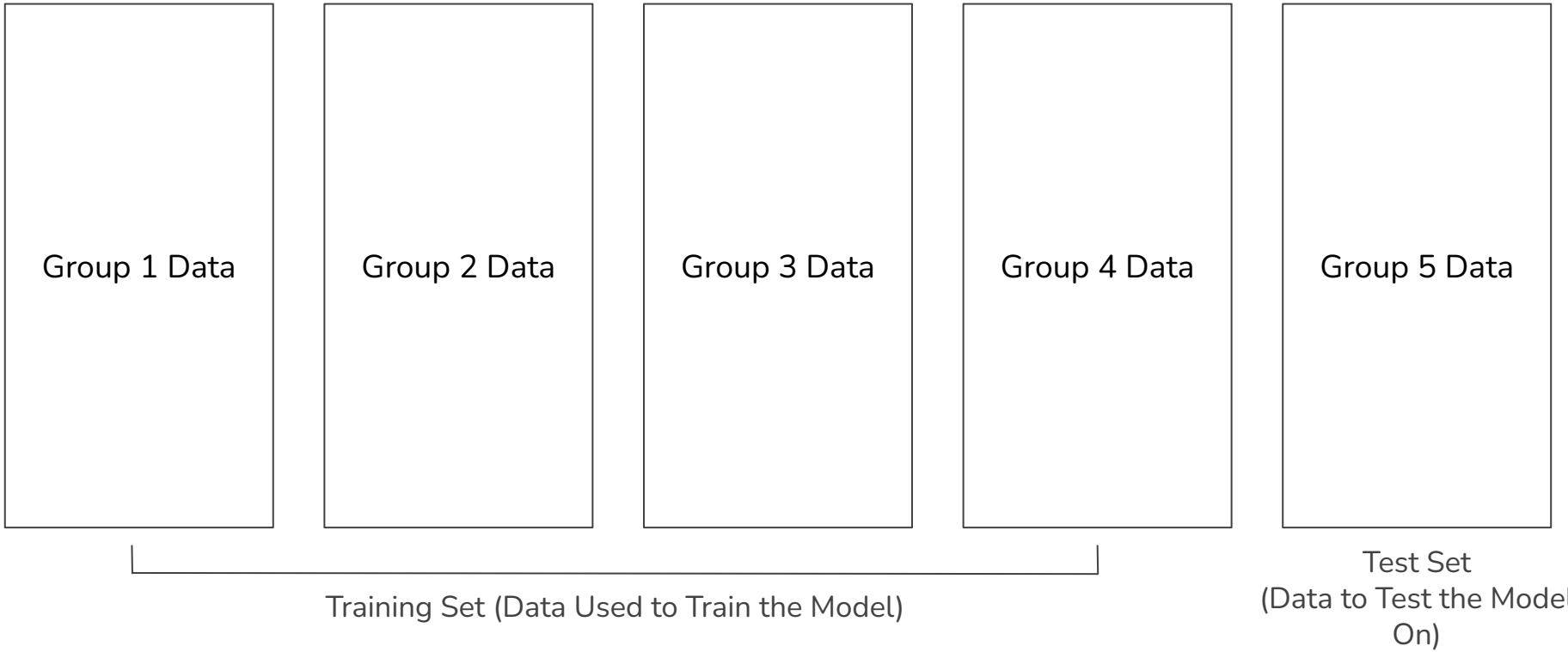
Strengths and Limitations of Machine Learning

Limitations of Machine Learning

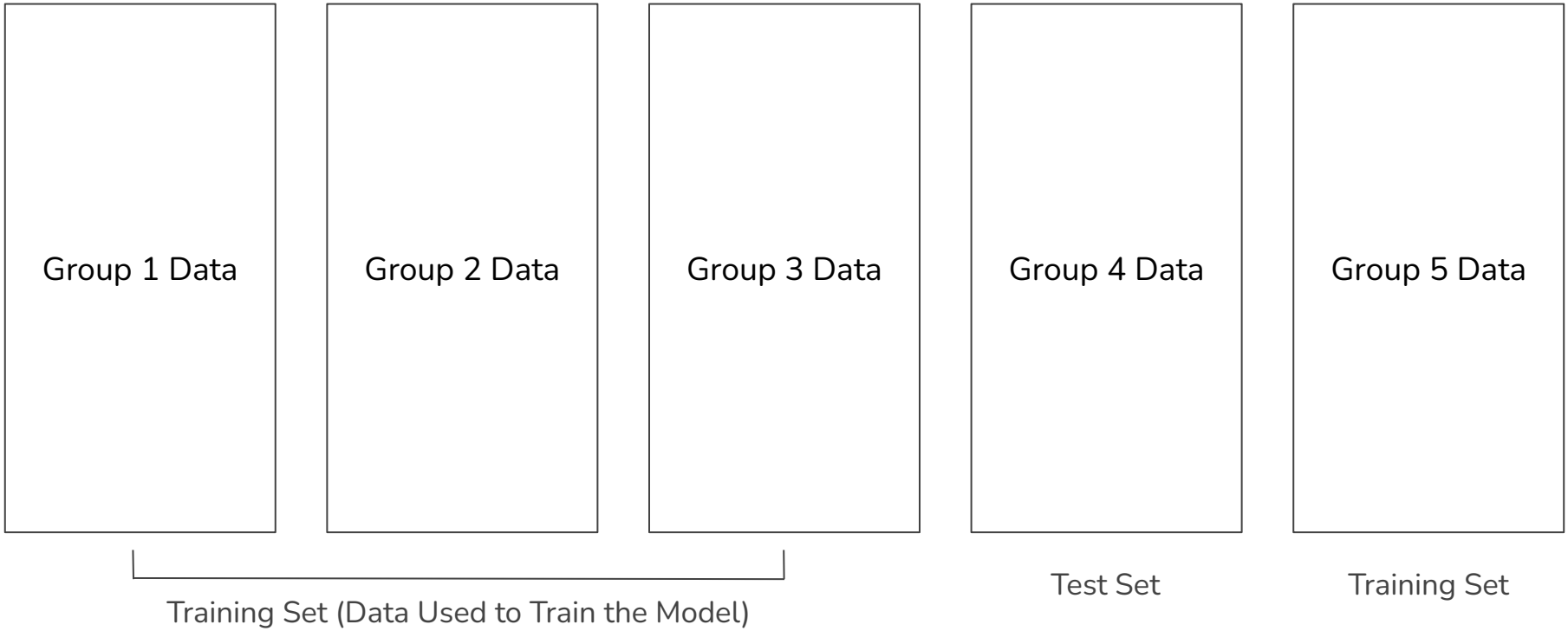
- Limited by the quality of the data
  - Data validation still required
  - More useful with larger datasets
  - Potential for inaccuracy due to outliers and anomalies
- Overfitting
  - Models are trained on historical data, models can sometimes provide accurate predictions for the training data but not on new data
  - Reducing Overfitting
    - k-fold Cross-Validation (covered in following slides)
    - Hyperparameter Tuning (e.g., determining number of trees, max number of features, etc. in random forest)
- Ethical Concerns of Machine Learning

K-fold Cross-Validation to Reduce Overfitting

Breaking dataset into five groups (folds):

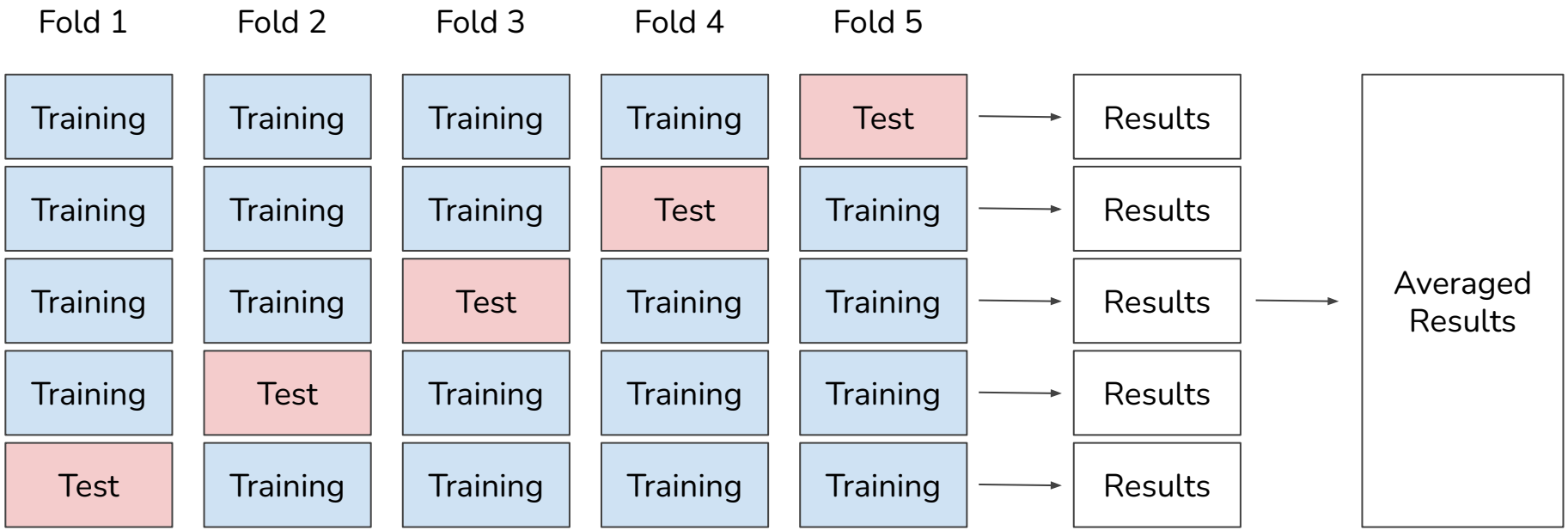


What if we change the Test and Training Sets?



Yields Different Results!

5-Fold Cross Validation to Reduce Overfitting



Strengths of Machine Learning

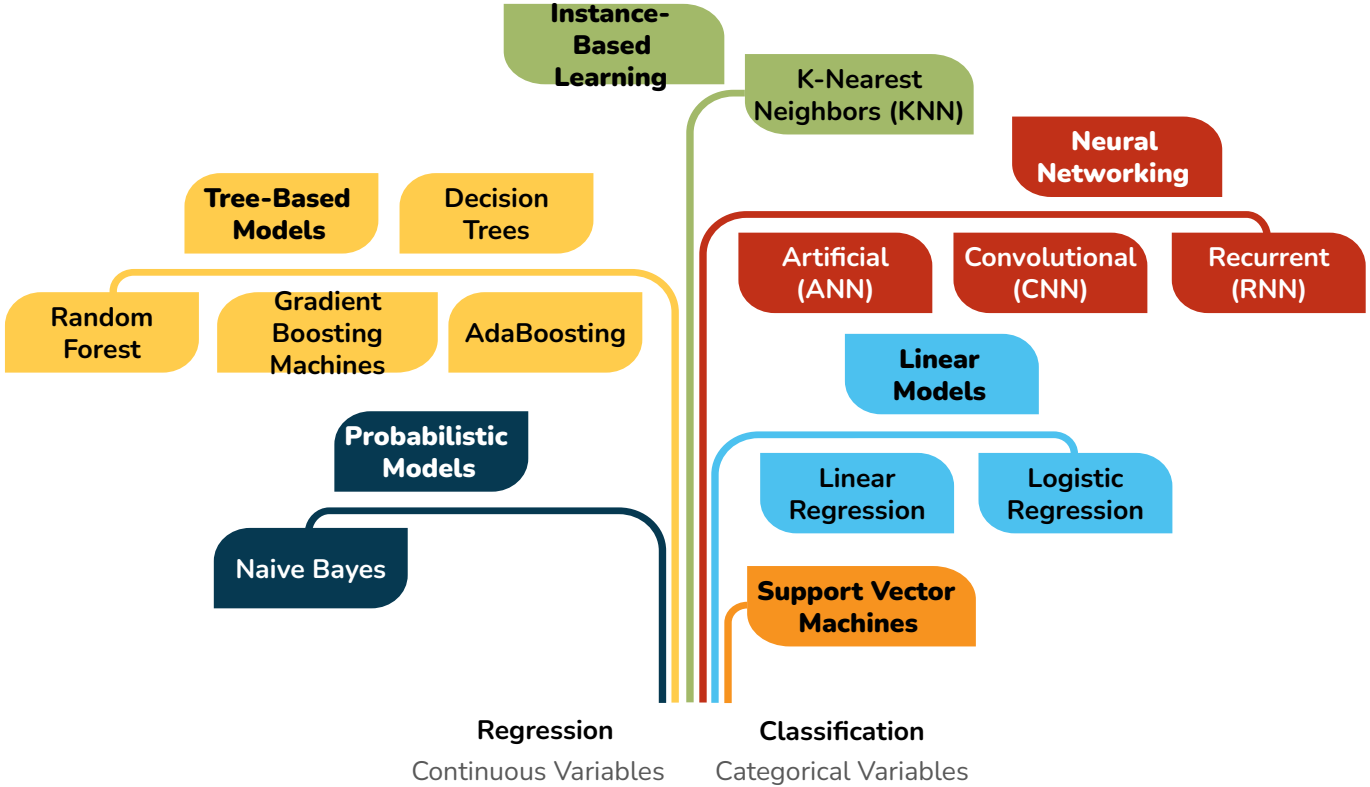
- Process Automation
  - Useful in automating repetitive tasks
    - Data cleaning
    - Yearly Analysis
- Efficiency in Identifying Trends and Patterns
- Predictive Analytics
  - Forecasting Future Outcomes and Trends
  - Ex: UHCC Retention Model
- Providing Insights on a Record/Individual-Level

Foundations of Supervised Learning

Supervised Learning

- Goal: learn a mapping from inputs to outputs based on example input-output pairs
- Strengths:
  - Handles large amounts of data
  - Provides a clear and interpretable output
  - Can be used in a wide variety of applications
- Weakness:
  - Requires a large amount of categorical data
  - Can be prone to overfitting if the model is too complex
  - Performance depends on the quality of the data

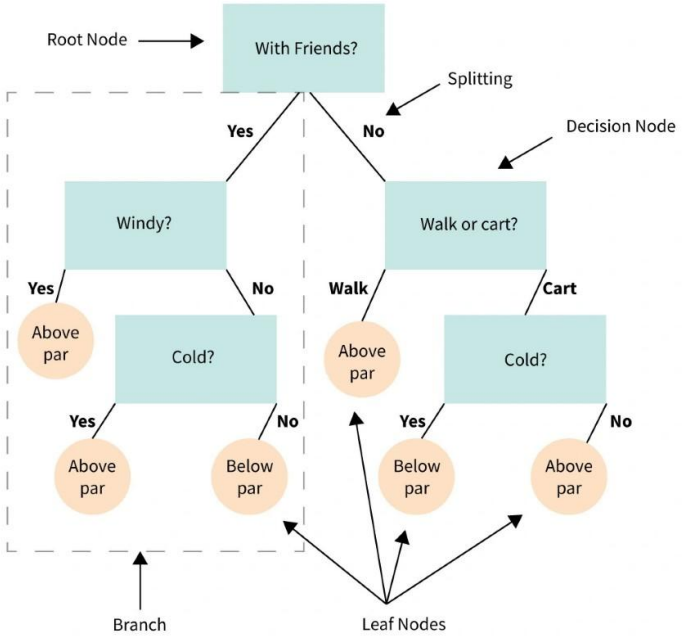
Types of Supervised Learning



Decision Trees and Random Forest

Decision Trees

- Decision Trees examine a variety of features
- Sets rules from these features to make classifications
- Arrives at conclusion based on rules

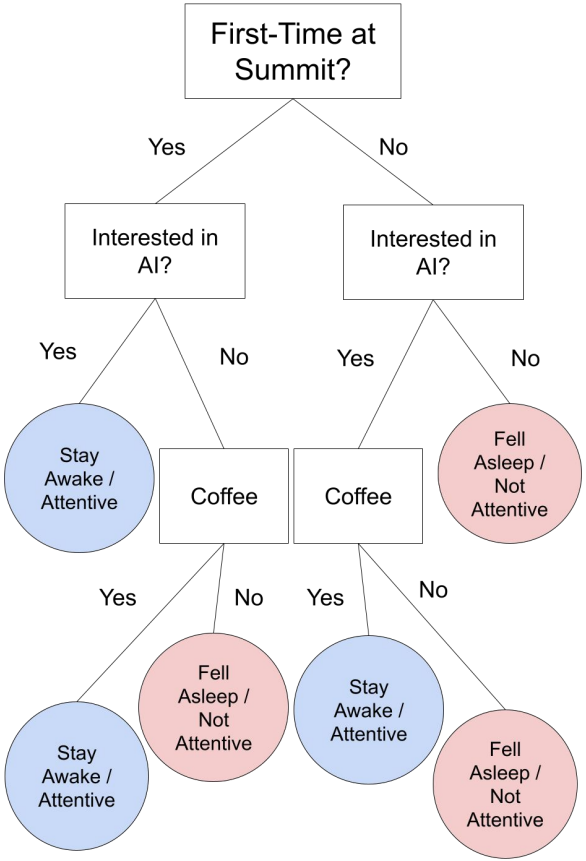


Let’s Dive into an Example on Decision Trees

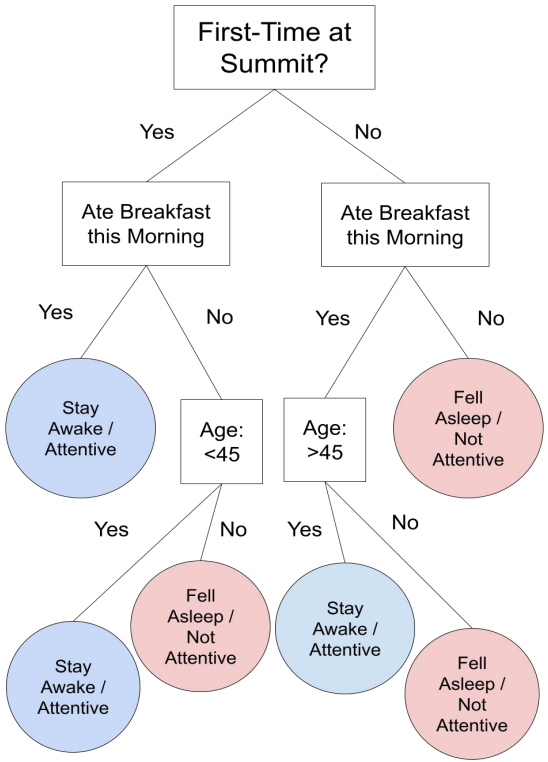
Ex: Predictive Attention Model

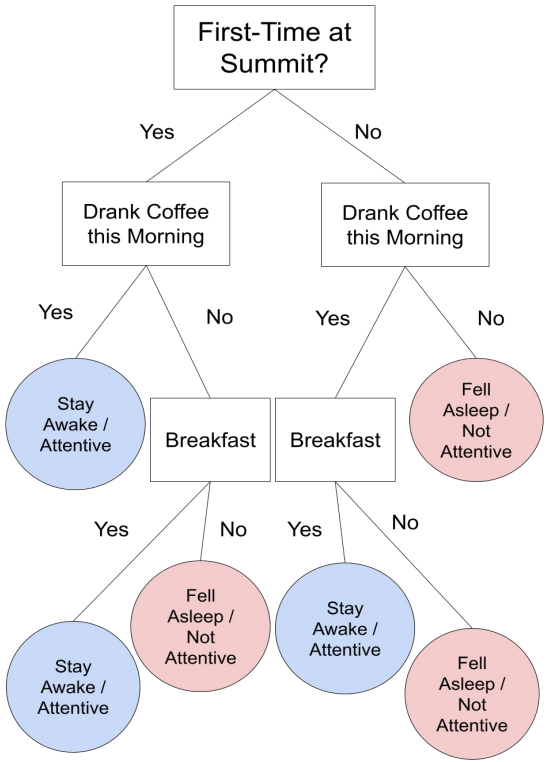
- Goal: Predict the “attentiveness” of SWIR Summit Attendees to provide support to those who may struggle to pay attention
- Define Attentiveness/Success Metric(s):
  - Attendee does not fall asleep during the presentations
  - Audience Participation
- Define Features that may impact Attentiveness:
  - Drank Coffee/Ate Breakfast this Morning
  - Demographic features: Age, Gender, etc.
  - First-Time at Summit
- Use historical data to make predictions for the upcoming summit

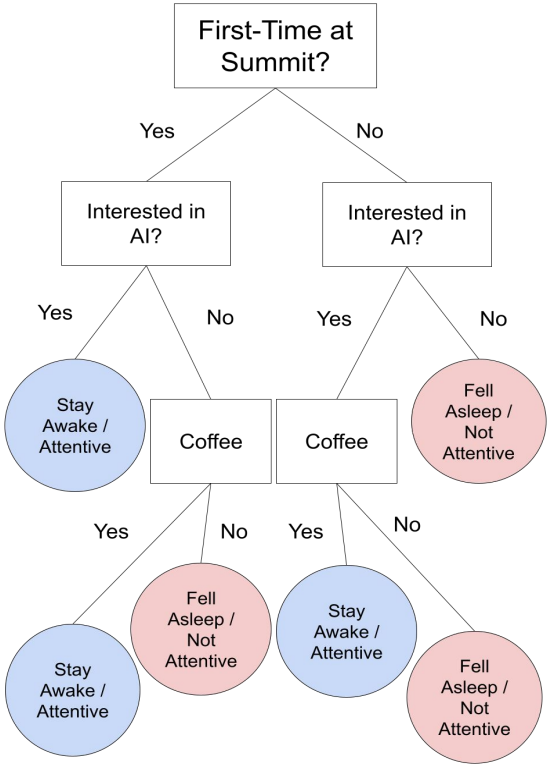
Ex: Predictive Attention Model Decision Tree



Ex: Predictive Attention Model Decision Trees

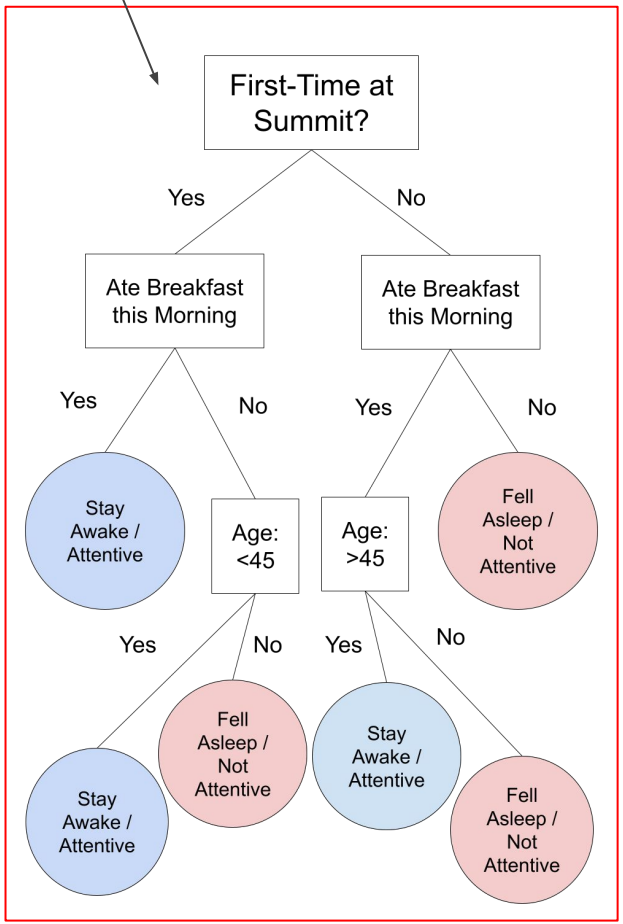




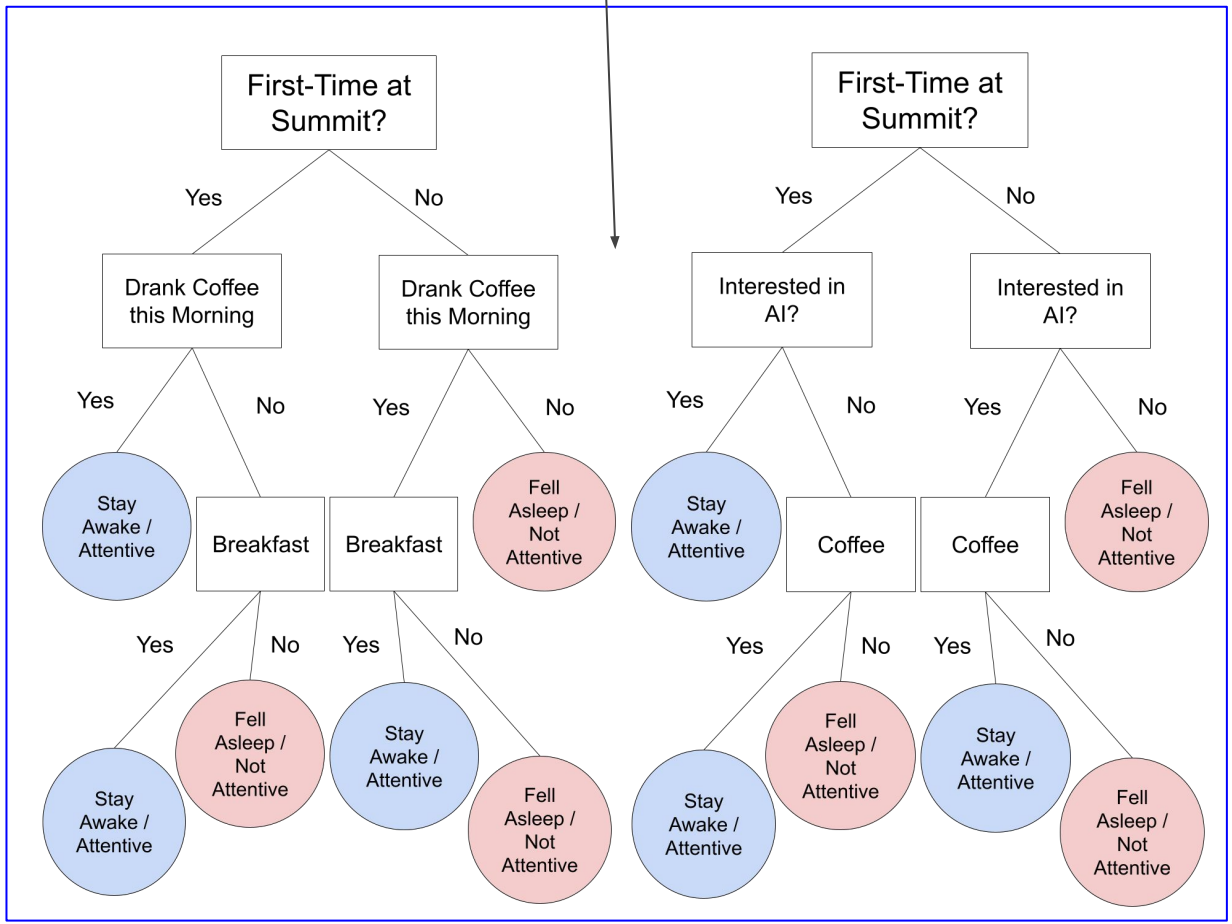


All Yield Different Results!

Not Great at Predicting Attentiveness in Test Set



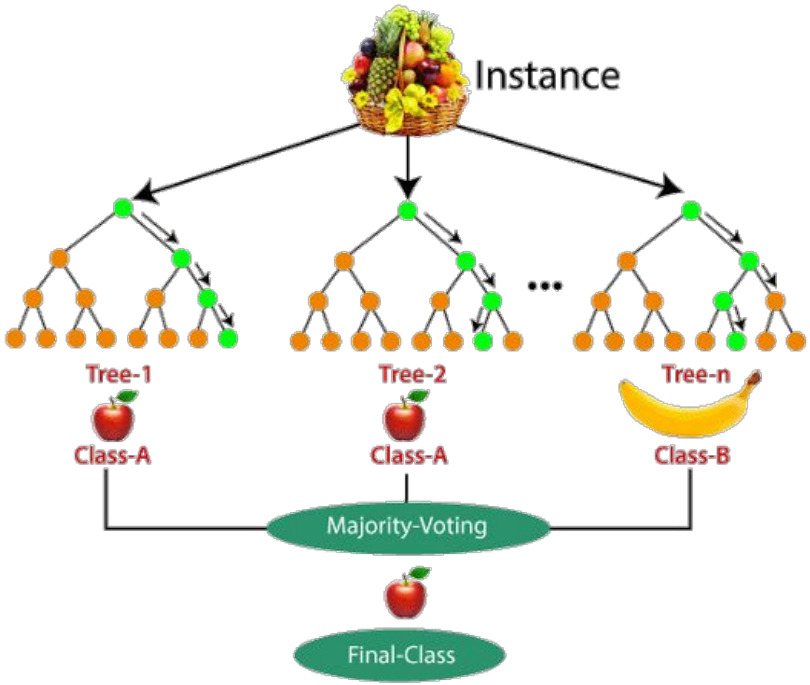
Good at Predicting Attentiveness in Test Set



How do we choose from our Decision Trees?

Random Forest Classifier

- Keeps trees that are good at predicting outcomes
- Final output determined by averaging results of decision trees
- Strengths
  - Allows for more variables
  - Higher score stability
- Weaknesses
  - Relatively slower computation
  - Prone to overfitting
- Hyperparameter Tuning
  - Number of Trees, Max Features, etc.



UHCC Retention Model

Background

- Goal: Use ML to build accurate profiles of our students to drive intervention
- Predict likelihood of student retention/transfer/graduation
- Based on model used by City Colleges of Chicago
- Similar use case as Kentucky Council on Postsecondary Education
- Utilize financial aid to increase chance of retaining high risk students
- Maximize efficiency of dollars spent
- Created two models
    - New students
    - Continuing students

Data Points

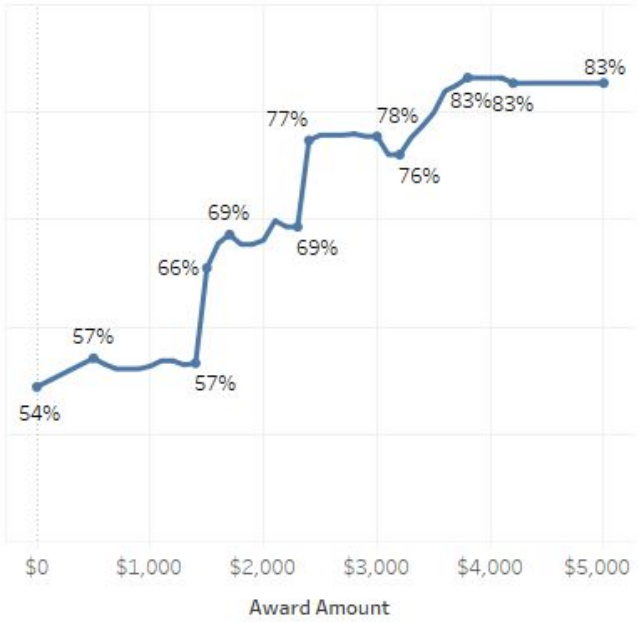
- Student Information
  - Home Campus
  - Full Time/Part Time Status
- High School Data
  - Max Test Score: Maximum of weighted scores among EdReady and Accuplacer
  - Public School: Whether or not the student graduated from a Hawaii public high school
  - GED: Whether or not the student has a GED
- Demographic Data
  - Gender
  - Race
  - Hawaii Residence (based on permanent address zip code)
- Academic Data
  - First Registration: Days between first registration and start of the semester
  - High School English/Math (entering students)
  - Pass English/Math (continuing students)
  - Cumulative GPA (continuing students)
- Student History
  - Early College: Whether or not the student ever did early college
  - Summer School Attendance
- Financial Data
  - Pell Status
  - Direct Unmet Need
  - Total Amount Awarded: Total financial aid awarded in semester

Retention Model Results

- Testing Method
  - UHCC home campus students who submitted a FAFSA
  - Trained on Fall 2021-2022 data
  - Tested on Fall 2023 data
- Each student assigned a retention score
- Creates a graph to show how increased aid affects retention
- Average Error of Continuing Students Model: 3.6%
  - Past student data: Passed college-level English/Math in 1st year
- Average Error of New Student Model: 4.9%
  - Less Data available for new students
  - Increased importance of high school data

Usage of Model

- Cutoff analysis for Financial Aid
- Trained on past student data
- Change Aid received and Unmet need
- Observe difference in scores



Campus	Gender	Age	Race	EdRdy Eng	EdRdy Mth 1	Degree	Resident	High School Type	Major OrgStr1	Registration	FT-PT Status	GPA	Credits	Pass Math	Pass English	Pell	Total Aid	Direct Need	Unmet Need	Score
MAU	M	25+	Hawaiian	64		Associates	1	Public School	General & Pre-Prof Ed	25	PT	3	11	0	1	Y	1160	9259.5	8099.5	0.5442
MAU	M	25+	Hawaiian	64		Associates	1	Public School	General & Pre-Prof Ed	25	PT	3	11	0	1	Y	4160	9259.5	5099.5	0.7775

Q&A and Discussion

- Can you see the value of how machine learning can help your campus?
- Can you name something that you have learned from this presentation?
- What are other practical uses of ML for your campus (how have you used ML already)
- What data do you need to create your model
- What type of model would you use
- If you haven't already used ML in your work, what are some of your apprehensions
- What are some ethical concerns you have with the usage of ML

Additional Resources & Further Reading

Python

Müller, Andreas C., and Sarah. Guido. 2017. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Sebastopol, Calif.: O'Reilly. [ISBN-13: 9781449369415]

Geron, Aurelien. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed. Sebastopol, CA: O'Reilly.

R

Kabacoff, Robert I. 2015. *R in Action: Data Analysis and Graphics in R* (2nd ed.). Shelter Island, NY: Manning. [ISBN-13: 978-1617291388]

Tutorial

Bowne-Anderson, Hugo. (December 21, 2017). "Kaggle Tutorial: EDA & Machine Learning."Links to an external site. Tutorials. DataCamp. https://www.datacamp.com/tutorial/kaggle-machine-learning-eda

Machine Learning Modeling Framework

Training Constraints	Exploratory Data Analysis	Data Preprocessing	Model Development	Tuning	Model Selection	Deployment Considerations
CPU & GPU Training	Discover variable characteristics and relationships	Data Integration	Unsupervised, Supervised, Reinforcement	Hyper-parameter tuning methods	Goodness of fit	Data Visualization
Storage		Imputations for missing values	Regression, Classification			Production Environment
Production Use Case					Scoring Routines	IV&V
Data Type & Availability	Subject Matter Expert Input	Transformation for regression	Feature Selection		Bias vs Variance	Maintenance & Reliability
Privacy		Feature Engineering				Performance & Scalability
Ethical						Security & Accessibility