Methods and Algorithms – MSc Data Science

By the end of this lecture, you will be able to:

1. Explain how decision trees partition feature space
2. Implement Random Forests using bagging and feature randomization
3. Interpret feature importance and out-of-bag error
4. Apply ensemble methods to fraud detection problems

Finance Application: Fraud detection with interpretable feature importance

From single models to ensemble methods that combine many weak learners

Fraud Detection Challenge

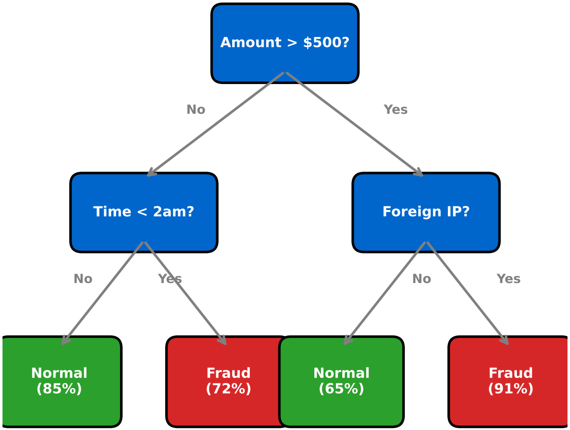
- Need high accuracy: fraudulent transactions cost millions
- Need interpretability: explain why transaction flagged
- Complex patterns: fraud evolves and adapts

Why Random Forests?

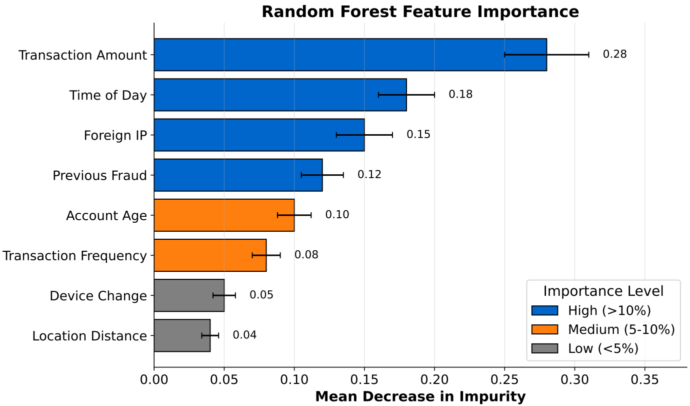
- Combines many trees for robust predictions
- Built-in feature importance ranking
- Handles non-linear relationships naturally

Ensemble methods: "wisdom of crowds" for machine learning

Decision Tree for Fraud Detection

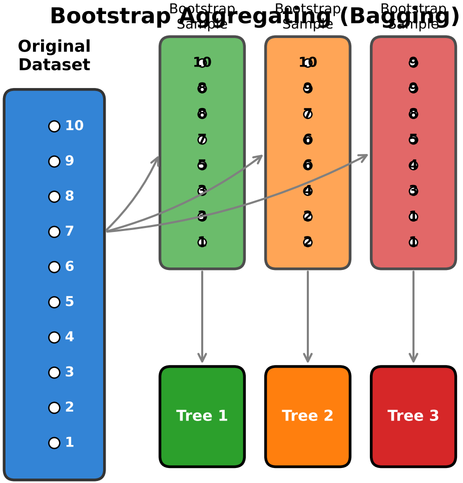


Trees split data using simple rules at each node until reaching a prediction



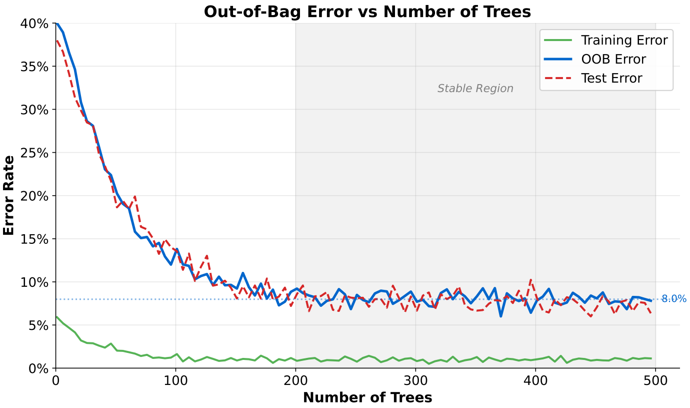
Random Forests automatically rank which features matter most for prediction

Bootstrap Aggregating (Bagging)



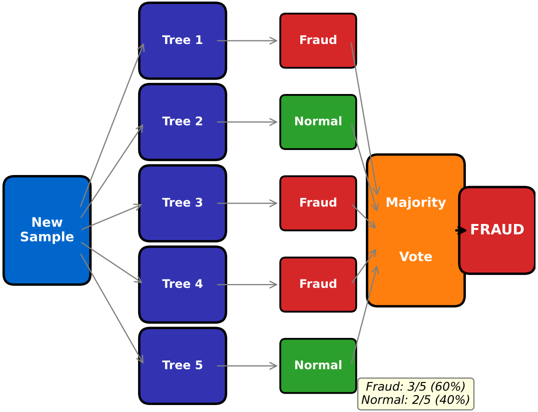
Each tree trained on ~63% unique samples (with replacement)

Each tree trains on a random sample, reducing overfitting

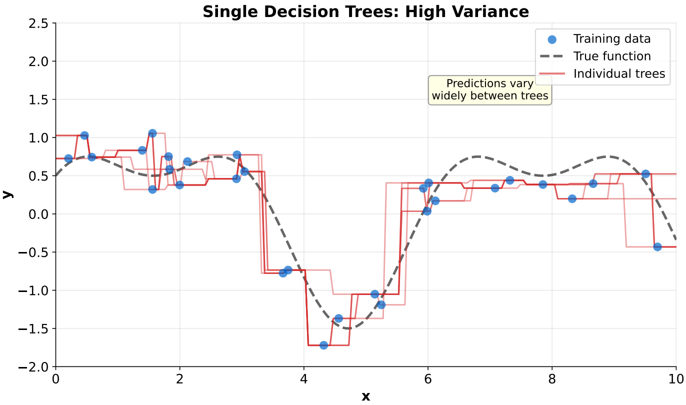


OOB error provides free cross-validation without held-out test set

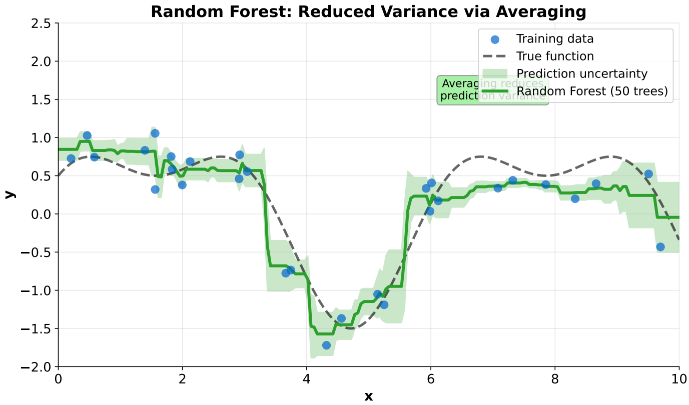
Ensemble Voting (Classification)



Final prediction combines votes from all trees (majority for classification)

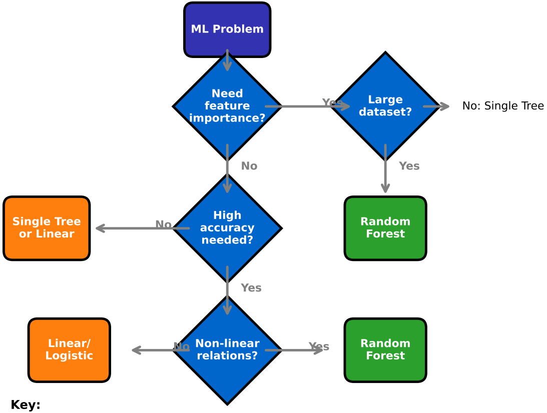


Each tree trained on different bootstrap sample produces different predictions



Averaging many high-variance trees produces low-variance ensemble

When to Use Random Forests



Key:

Random Forest: Best for accuracy + feature importance

Alternative: When interpretability is paramount

Random Forests excel when accuracy and feature importance both matter