L04: Random Forests

Ensemble Learning for Robust Predictions

Methods and Algorithms

Spring 2026

Outline

Learning Objectives

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

1. **Derive** the variance reduction formula for bagging and analyze the role of tree correlation
2. **Evaluate** Random Forest vs. gradient boosting for a given prediction task using bias-variance tradeoff
3. **Analyze** feature importance using permutation importance, MDI, and SHAP values
4. **Critique** ensemble methods for regulatory compliance in financial applications

Finance Application: Fraud detection with interpretable feature importance

---

Bloom's Level 4–5: Analyze, Evaluate, Critique — MSc-level objectives

The Business Problem

Fraud Detection Challenge

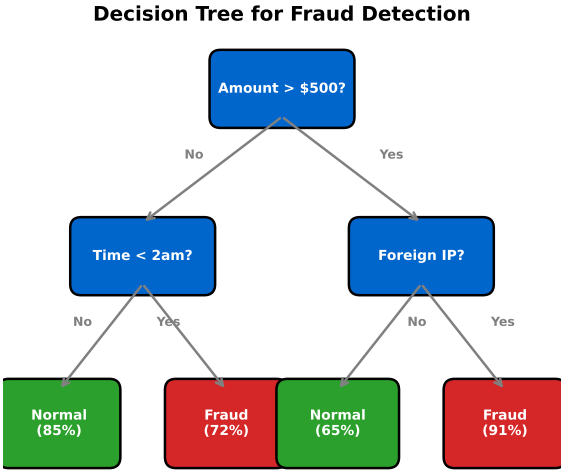
- Need high accuracy: fraudulent transactions cost millions
- Need interpretability: explain why transaction flagged
- Complex patterns: fraud evolves and adapts

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

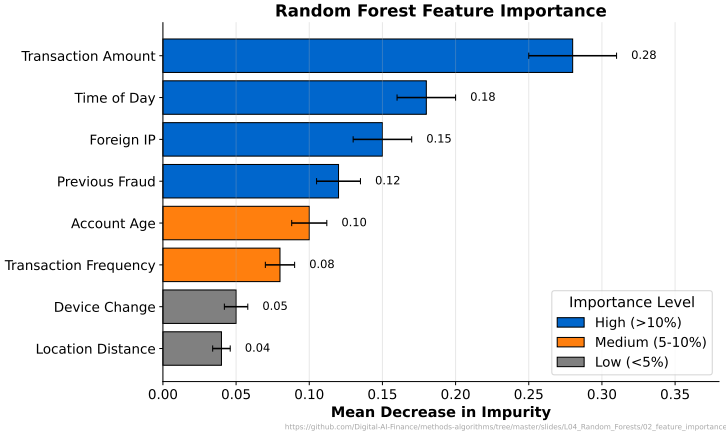
Decision Tree Structure



https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/L04_Random_Forests/01_decision_tree

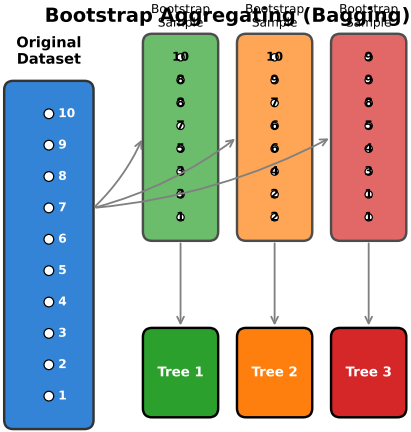
Trees split data using simple rules at each node until reaching a prediction

Feature Importance



Random Forests automatically rank which features matter most for prediction

Bootstrap Aggregating (Bagging)



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

https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/L04_Random_Forests/03_bootstrap

Each tree trains on a random sample, reducing overfitting

Key Equations: Random Forest Theory

Gini Impurity (split criterion):

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

Bagging Variance Reduction (with correlated trees):

$$\text{Var}(\bar{f}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

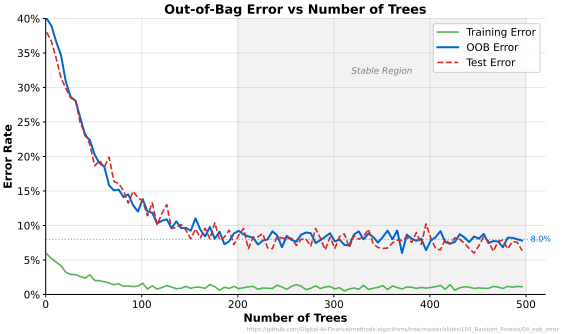
where ρ = pairwise tree correlation, B = number of trees

Random Forest Decorrelation: Feature subsampling with $m \approx \sqrt{p}$ features per split reduces ρ, amplifying the variance reduction effect.

OOB Error: Each sample is excluded from $\approx 36.8\%$ of trees ($P(\text{not selected}) = (1 - 1/n)^n \to 1/e$). OOB predictions aggregate only from those excluded trees.

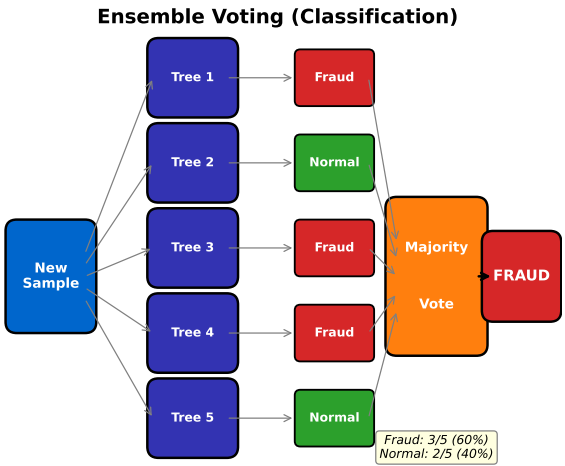
These equations formalize why ensembles of decorrelated trees outperform single models

Out-of-Bag Error



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

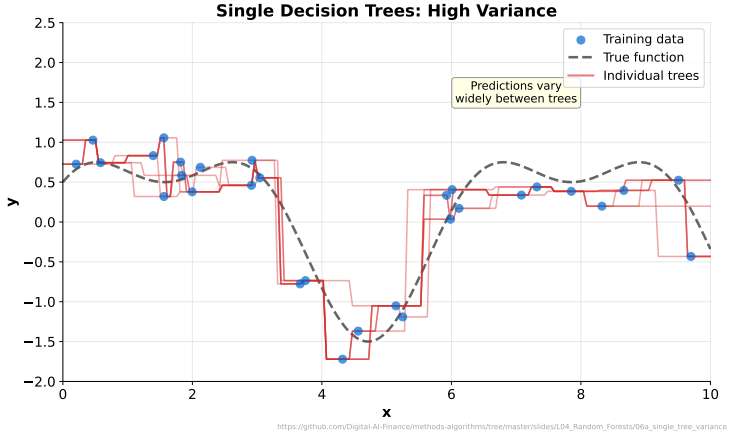
Ensemble Voting



https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/L04_Random_Forests/05_ensemble_voting

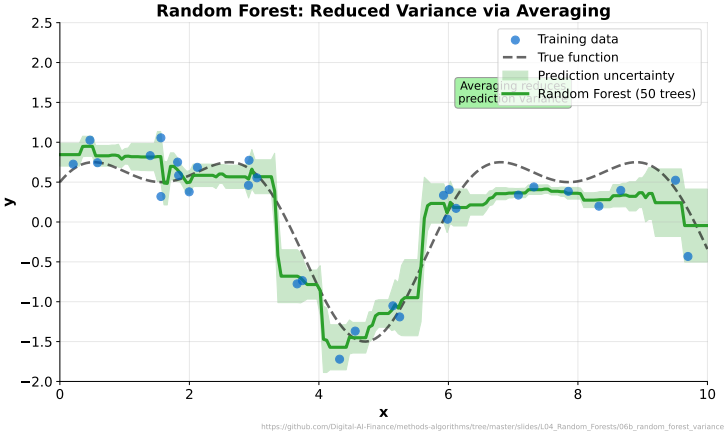
Final prediction combines votes from all trees (majority for classification)

Single Trees: High Variance



Each tree trained on different bootstrap sample produces different predictions

Random Forest: Reduced Variance



Averaging many high-variance trees produces low-variance ensemble

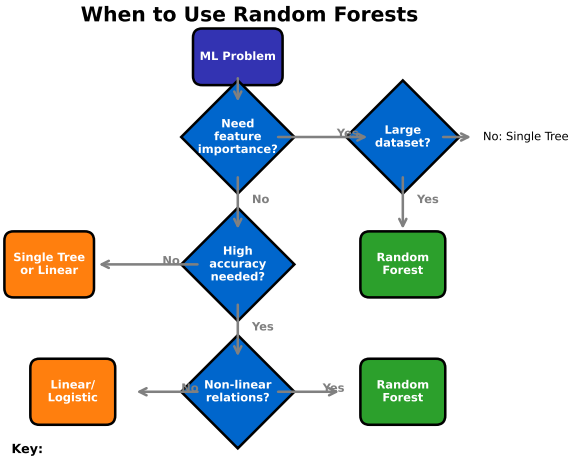
Hands-on Exercise

Open the Colab Notebook

- Exercise 1: Train a decision tree on credit data
- Exercise 2: Build a random forest and analyze feature importance
- Exercise 3: Tune hyperparameters with cross-validation

Link: `https://colab.research.google.com/` See course materials for Colab notebook

Decision Framework



https://github.com/Digital-AI-Finance/methods-algorithms/tree/master/slides/L04_Random_Forests/07_decision_flowchart

Random Forests excel when accuracy and feature importance both matter

References

- Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
- James et al. (2021). *Introduction to Statistical Learning*. `https://www.statlearning.com/`
- Hastie et al. (2009). *Elements of Statistical Learning*. `https://hastie.su.domains/ElemStatLearn/`