Introduction to Tree-based Models

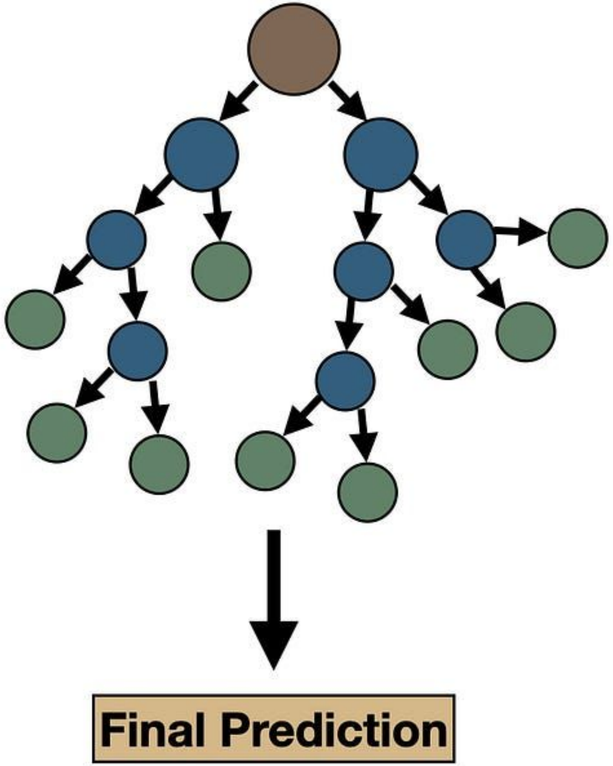
•••

Zikun Zheng
DRP Win24

Outline

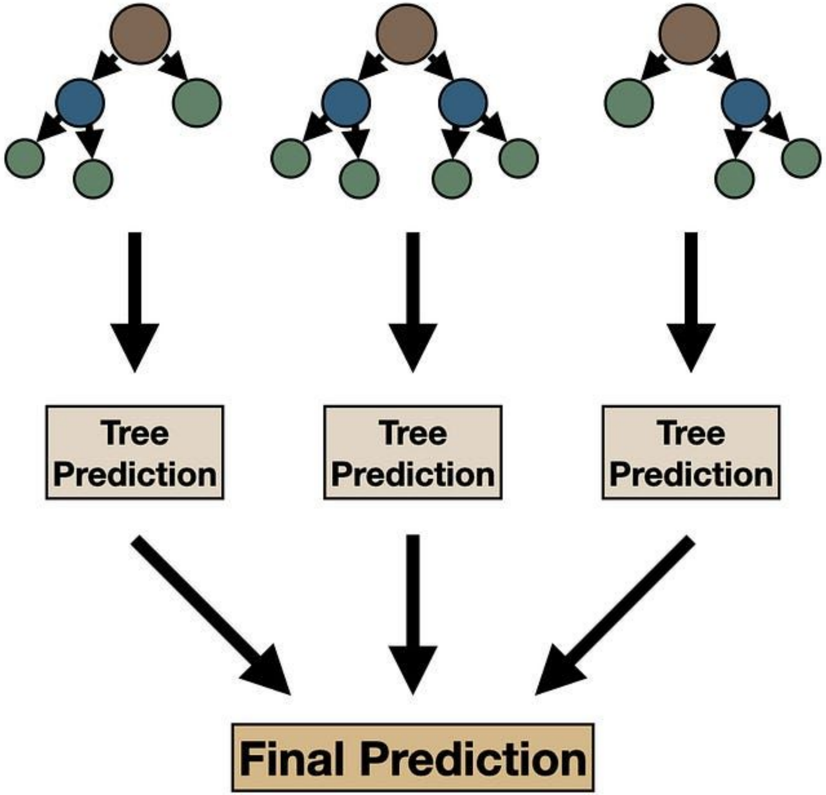
- CART
- Bagging
- Random Forest
- Boosting
- Deep Forest

Single Decision Tree



Final Prediction

Decision Tree Ensemble

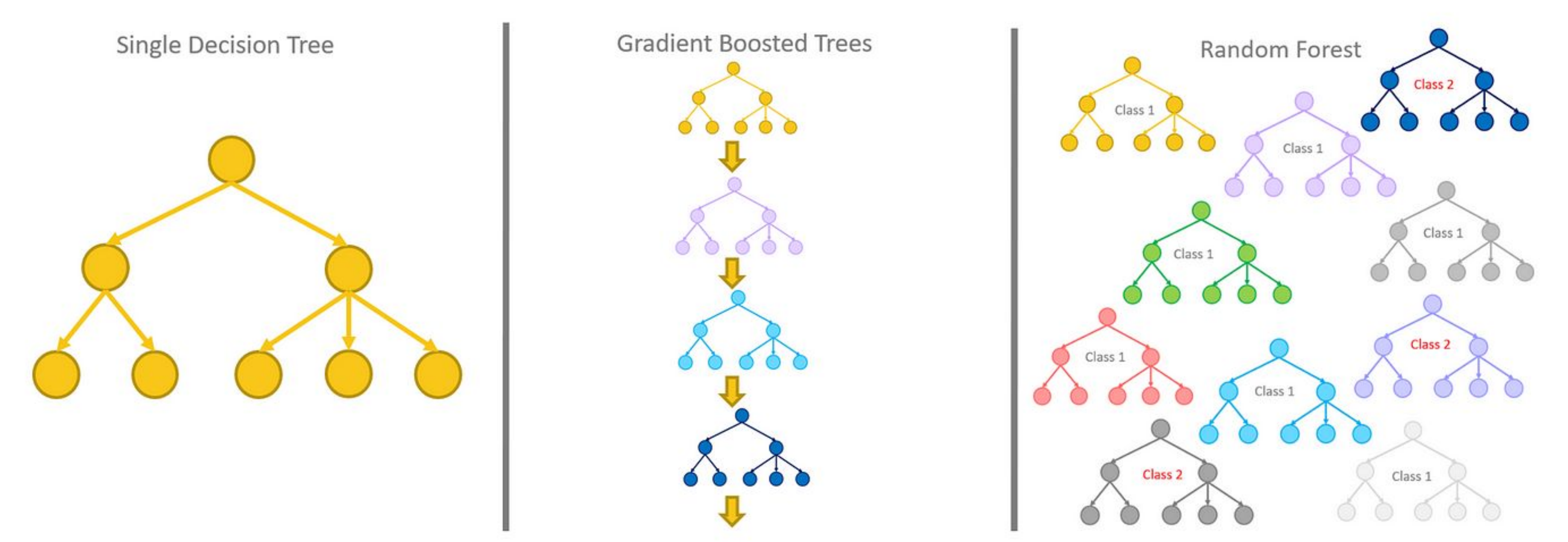


Tree Prediction

Tree Prediction

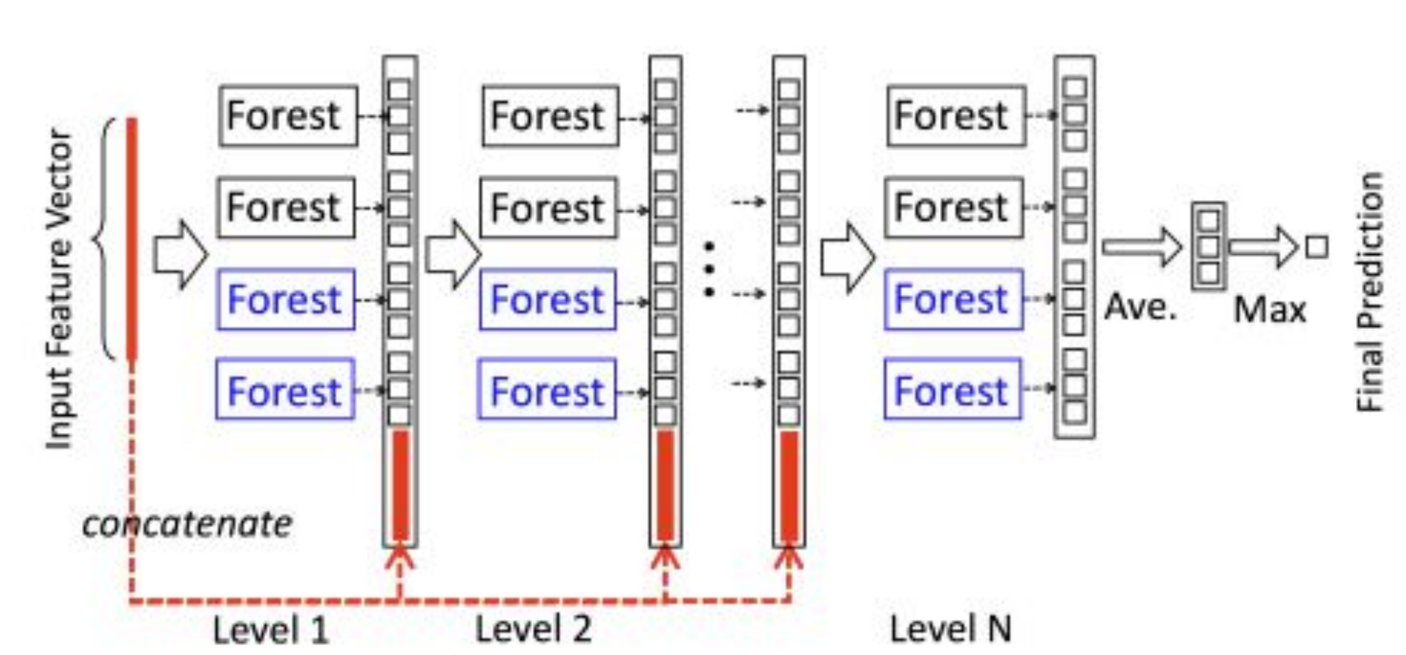
Tree Prediction

Final Prediction


Single Decision Tree
Gradient Boosted Trees
Random Forest
Class 2
Class 1
Class 1
Class 1
Class 1
Class 2
Class 1
Class 1
Class 2
Class 1

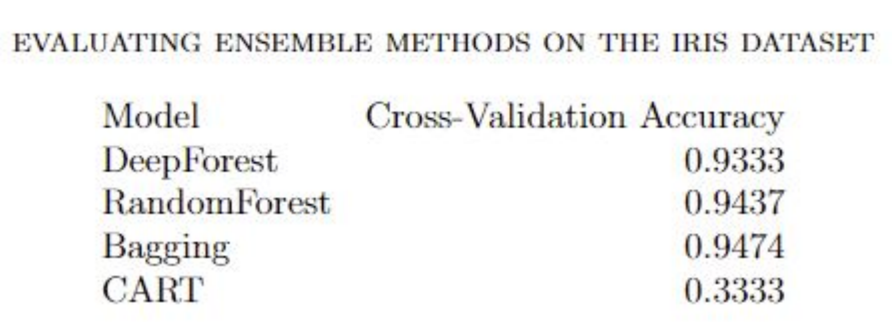
Motivation

- CART prones to overfitting (big variance)
- Bagging (bootstrap sampling)
- Random Forest (further reducing correlation)
- Boosting (learning residuals)
- Deep Forest (depth and width)


Input Feature Vector
Forest
Forest
Forest
Forest
Forest
Forest
Forest
Forest
Forest
Forest
Forest
Forest
Ave.
Max
Final Prediction
concatenate
Level 1
Level 2
Level N

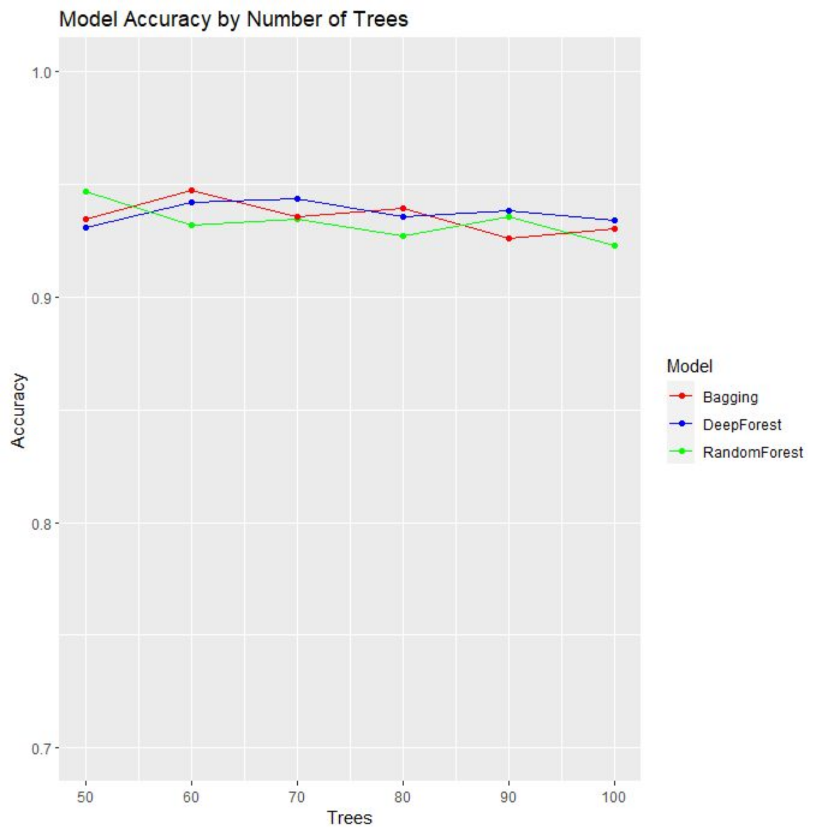
Application

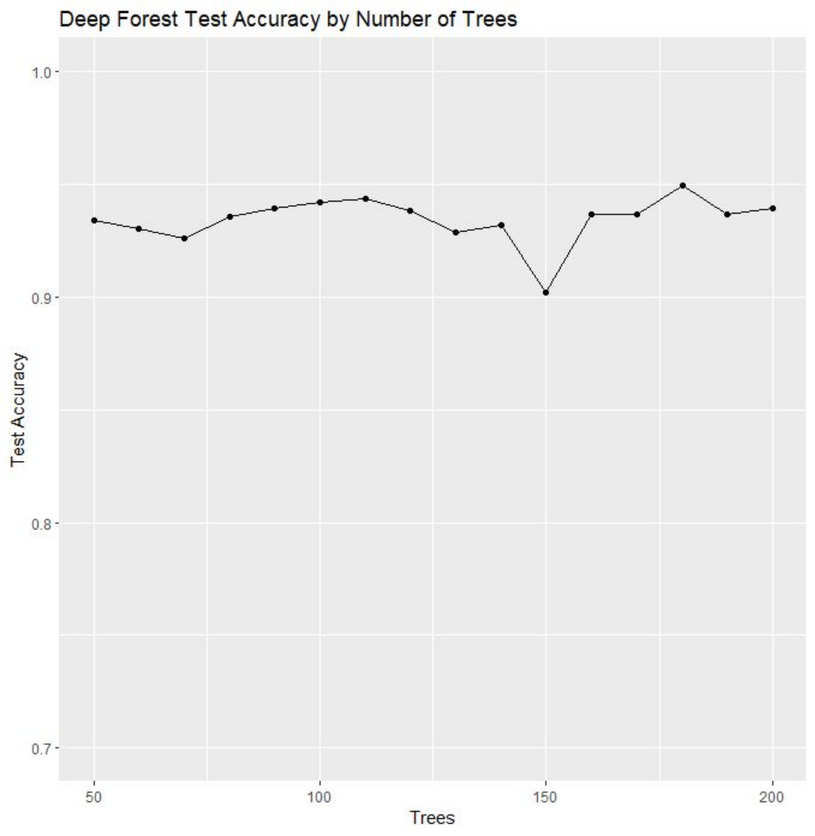
- Iris dataset (https://archive.ics.uci.edu/dataset/53/iris)
- Cross validation (10 folds)



EVALUATING ENSEMBLE METHODS ON THE IRIS DATASET

Model	Cross-Validation Accuracy
DeepForest	0.9333
RandomForest	0.9437
Bagging	0.9474
CART	0.3333


Model Accuracy by Number of Trees
Accuracy
1.0
0.9
0.8
0.7
50
60
70
80
90
100
Trees
Model
Bagging
DeepForest
RandomForest


Deep Forest Test Accuracy by Number of Trees
Test Accuracy
1.0
0.9
0.8
0.7
50
100
150
200
Trees

Conclusion

- Ensemble methods generally outperform CART in accuracy due to their ability to aggregate multiple models and reduce overfitting.
- Deep Forest models leverage ensemble learning's power, layering Random Forests and Bagging to handle complex data.

Reference

https://medium.com/analytics-vidhya/ensemble-models-bagging-boosting-c33706db0b0b

https://towardsdatascience.com/10-decision-trees-are-better-than-1-719406680564

https://arxiv.org/pdf/1702.08835.pdf