3050571 Practical Clin Data Sci

Session 10: Machine learning framework

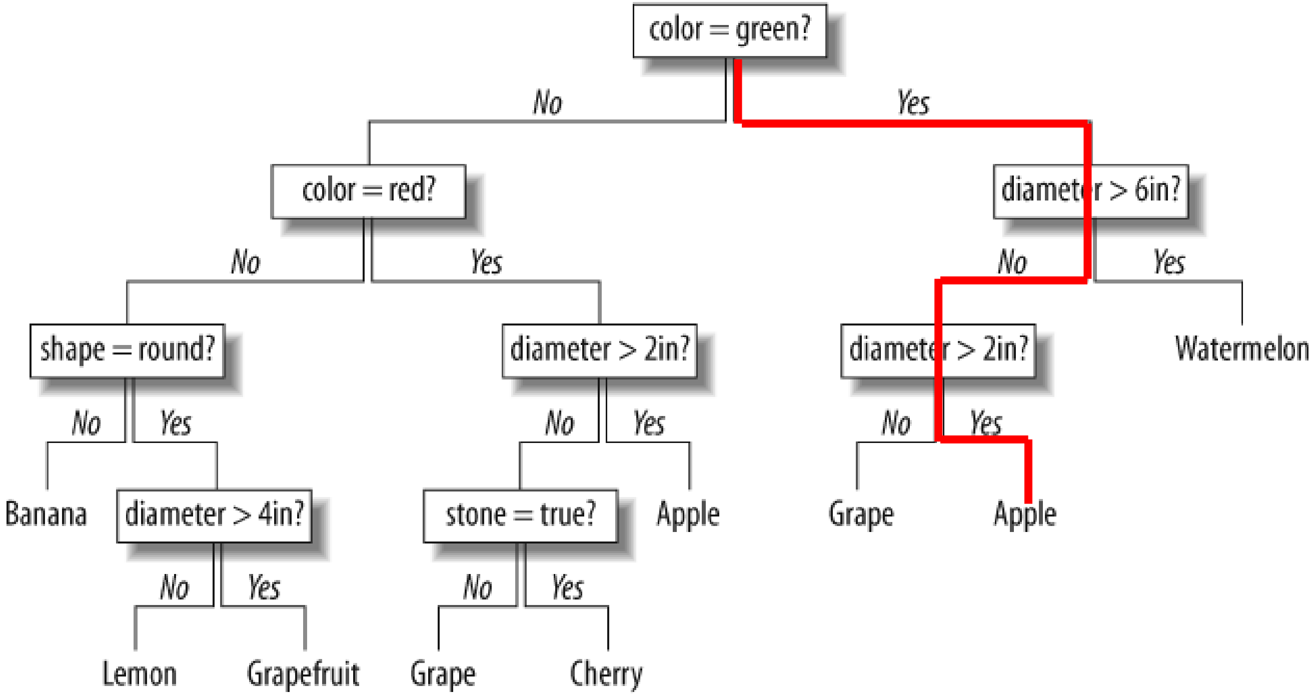
February 15, 2024

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

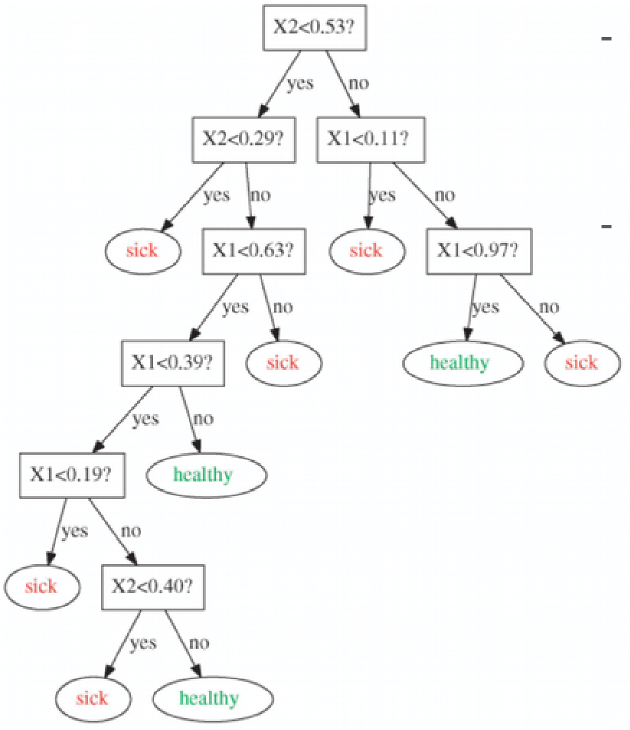
Decision tree

Decision tree

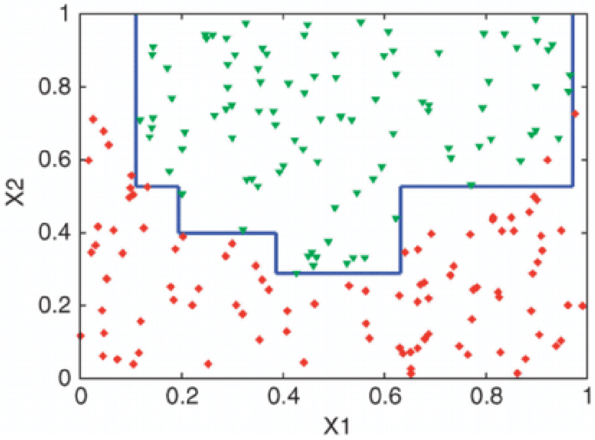


Source: Programming collective intelligence by Toby Segaran

Decision tree behaviors



- Each decision is a threshold on each feature
  - Piecewise linear
  - Parallel to an axis
- Good for criteria-based classification



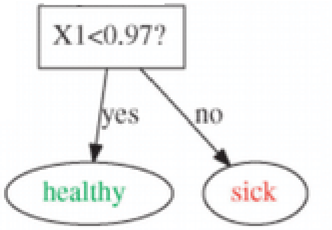
Miller, C. "Screening meter data: Characterization of temporal energy data from large groups of non-residential buildings"

Splitting quality

100 healthy, 100 sick

85 healthy, 5 sick

45 healthy, 51 sick



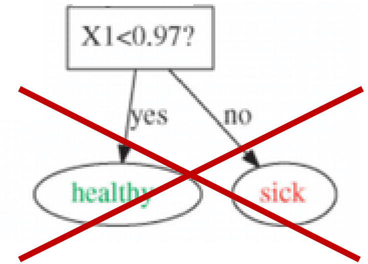
15 healthy, 95 sick

55 healthy, 49 sick

- Gini impurity: $\sum p(1-p)$
- Entropy: $-\sum p \ln(p)$
  - Minimal at p = 0 or 1 → Perfect split
  - Maximal at p = 0.5 → 50-50 split
- Search for feature and cutoff that yield lowest impurity or entropy

Control mechanisms for tree building

1. Too few samples to make a split



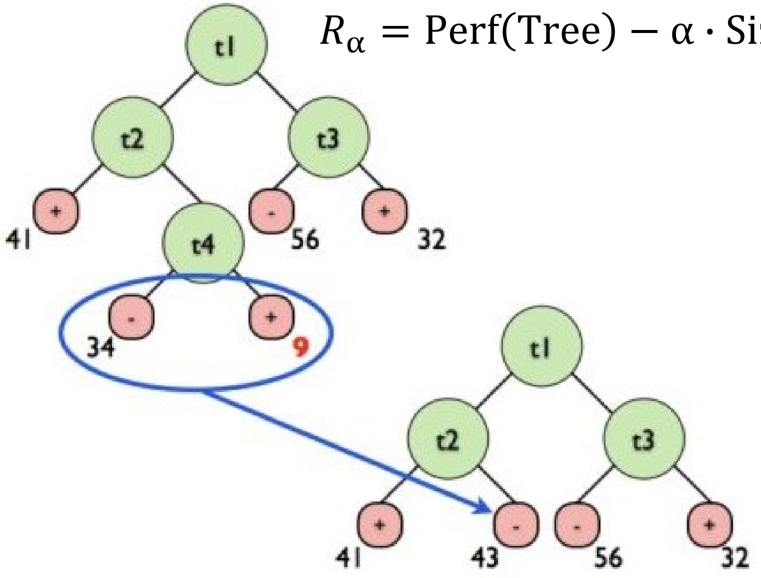
3. Impurity or entropy does not change much after the split

2. Too few samples on either branch

- Limit the tree size
- Limit the improvement in quality
- Limit the number of samples that support a split

Tree pruning (post-processing)

$$R_\alpha = \mathrm{Perf(Tree)} - \alpha \cdot \mathrm{Size}$$



Patel, N. and Upadhyay, S. “Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA” IJCA 2012

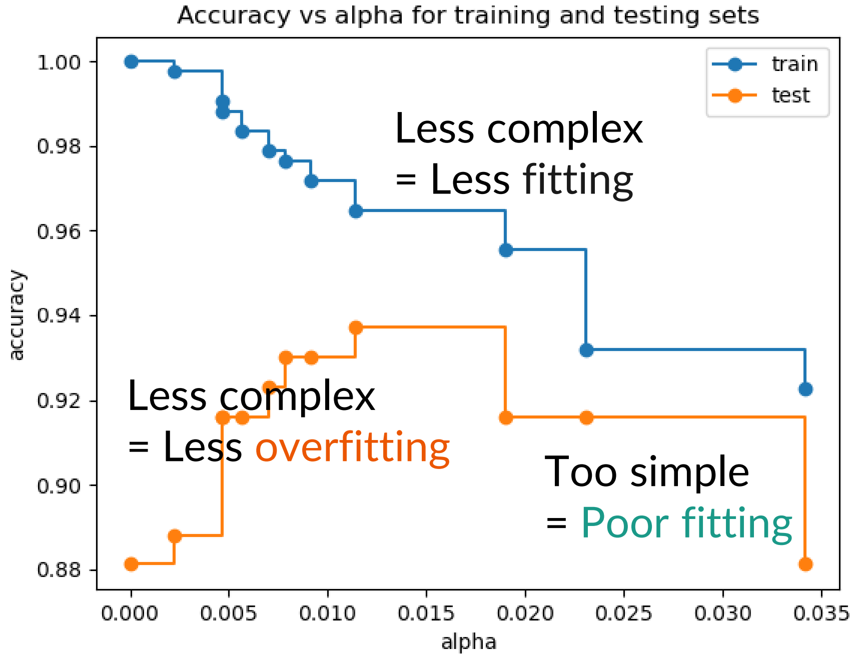
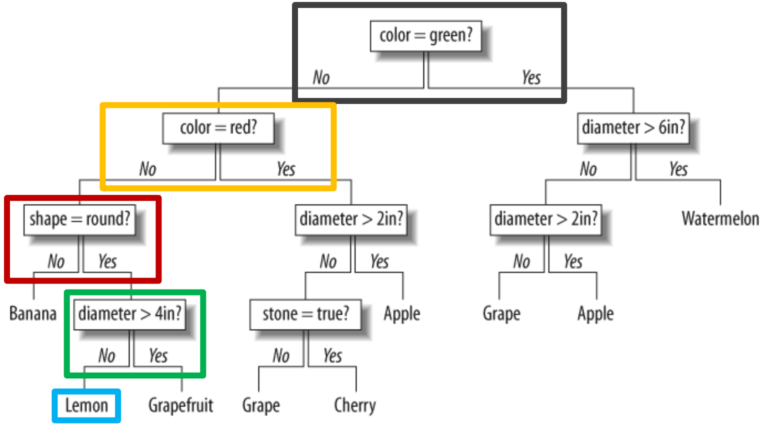


Image from scikit-learn,org

Regularization on features

- **Linear model:** $\hat{y}_i = b_0 + b_1 x_{i,1} + \cdots + b_n x_{i,n}$
  - LASSO
- **Tree model:**
  - Repeatedly using the same feature
  - Early decision affects the rest
- Feature bagging
  - Look at only *N* features at each step
  - Force model to use diverse features



Decision tree for regression

Predictors				Target
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

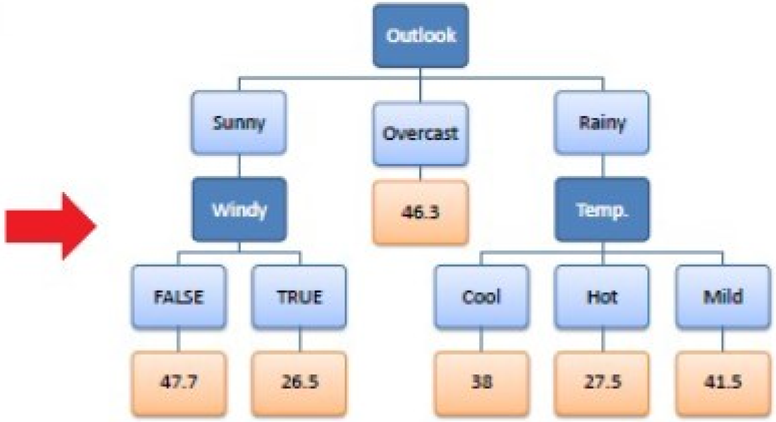


Image from saedsayad.com

- Predict an average of samples in the same branch

Decision tree is a piecewise constant function

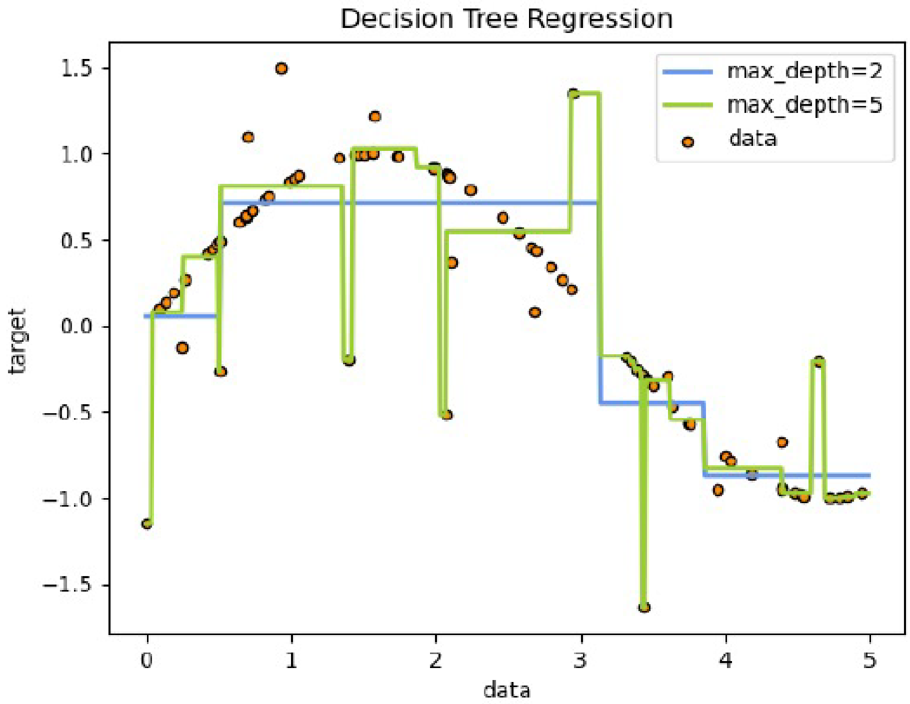


Image from scikit-learn,org

Linear-Tree model

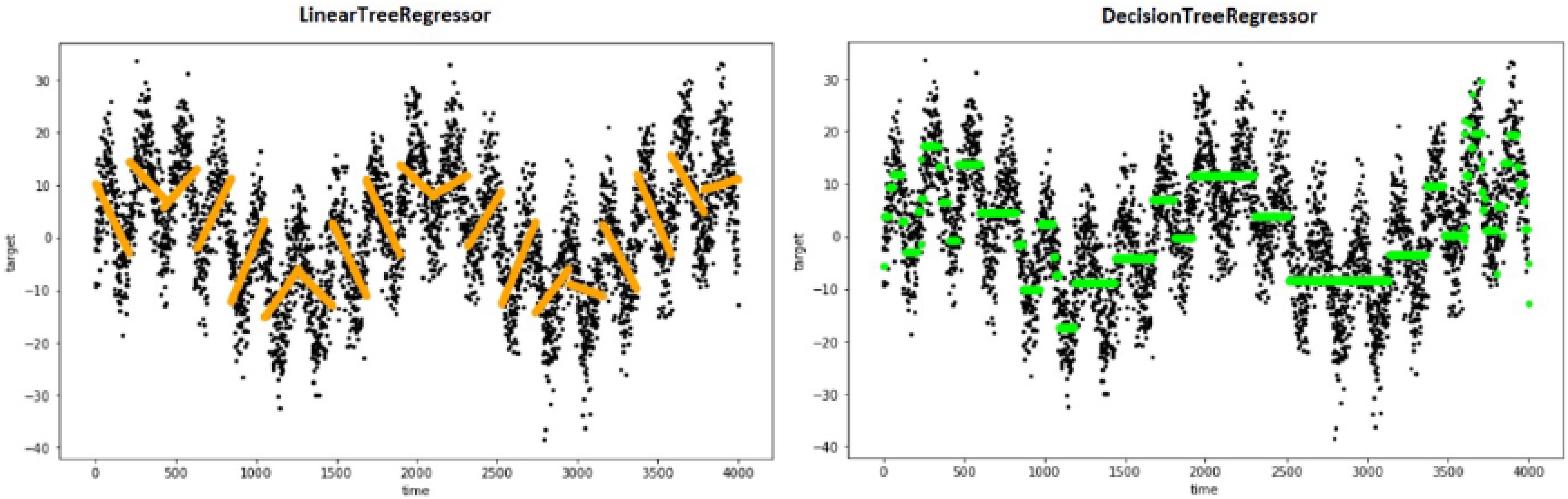
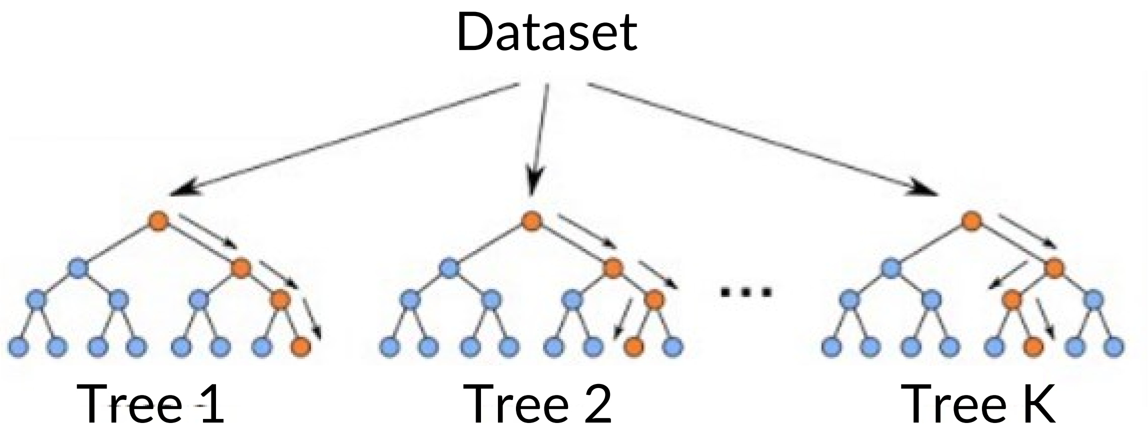


Image from towarddatascience.com

- Fit a different linear model in each branch

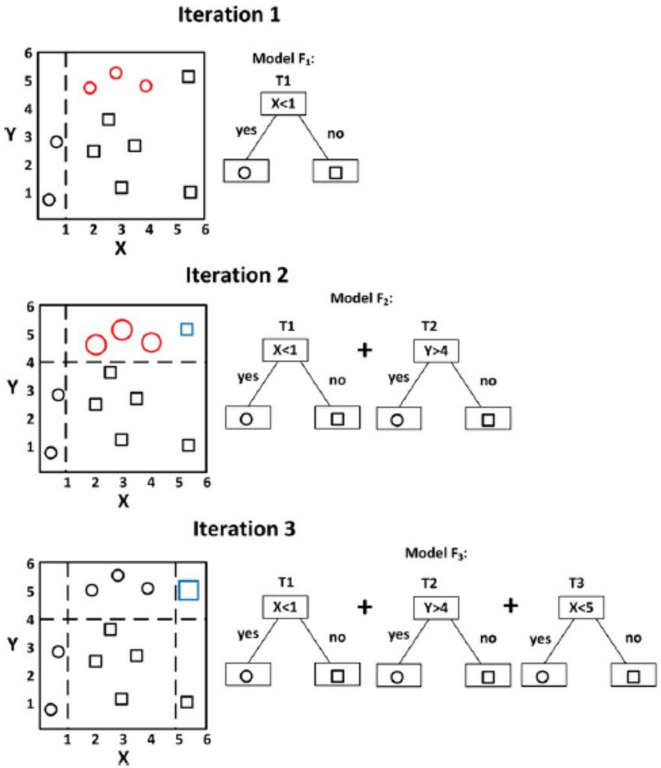
Ensemble approaches

Bagging: Random forest



- Sample 80% of the dataset to train each decision tree
- Each tree may overfit to different part of the dataset
- But the consensus should be correct

Boosting for classification



- The first model made some mistakes
- The mistakes were assigned higher weights for the subsequent models
- As more models are added, the ensemble should make less errors
- Ensemble = $w_1 f_1(x) + \cdots + w_n f_n(x)$

Boosting for regression

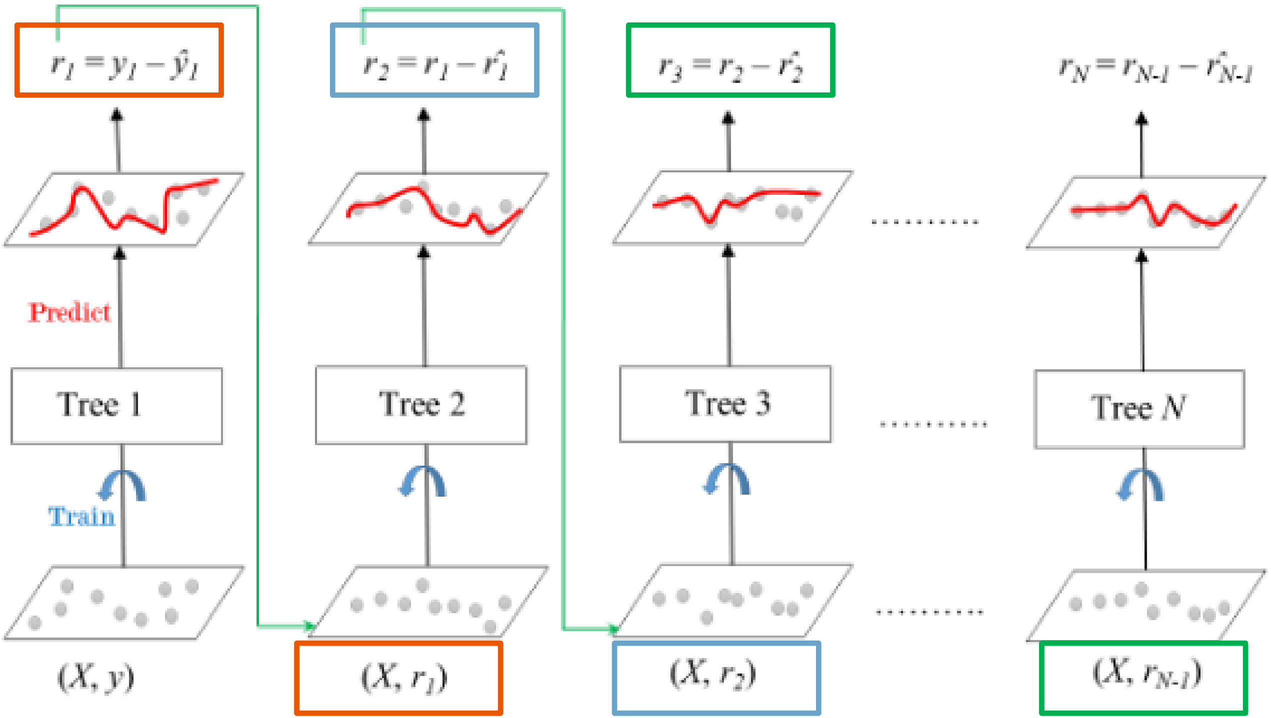
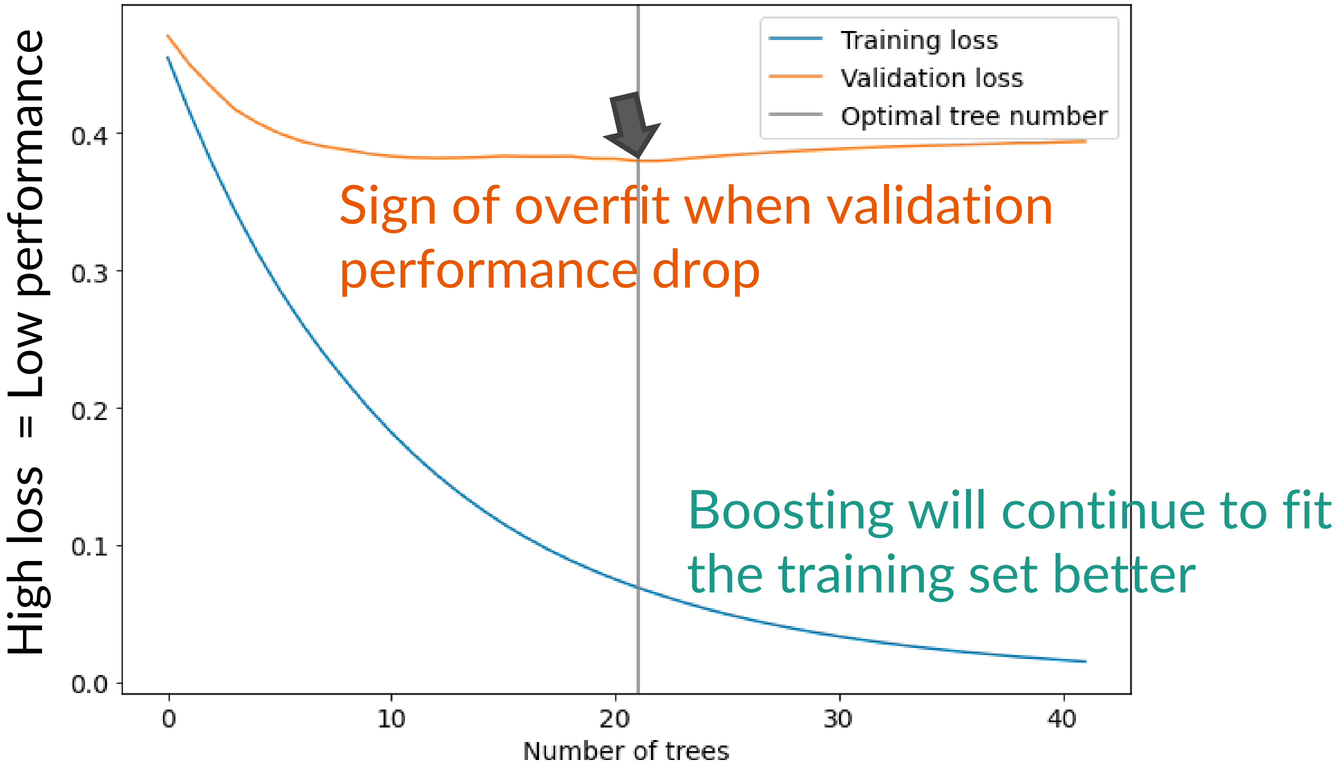


Image from geeksforgeeks.org

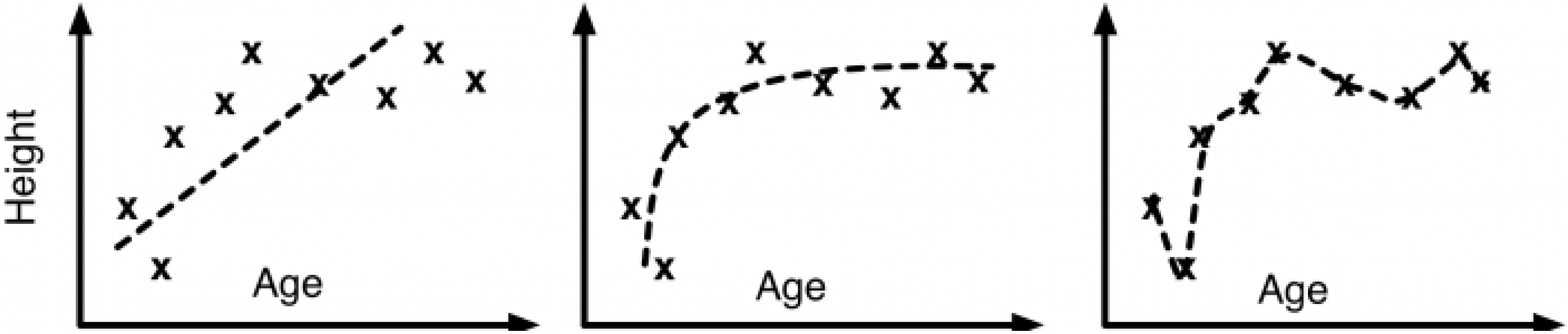
Controlling boosting with learning rate



Impact of ensemble

Boosting solves underfitting

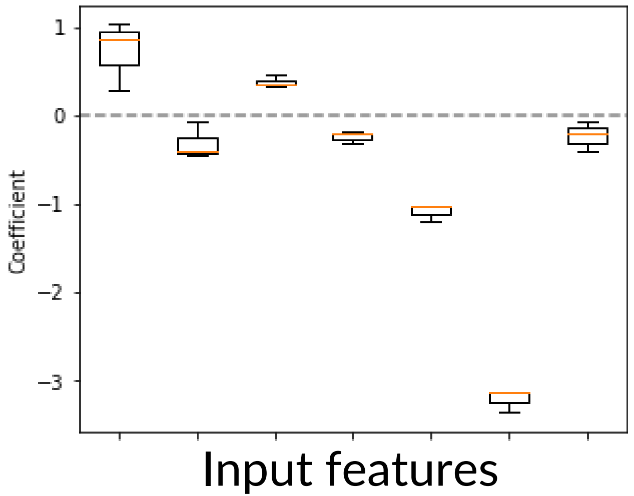
Bagging prevent overfitting

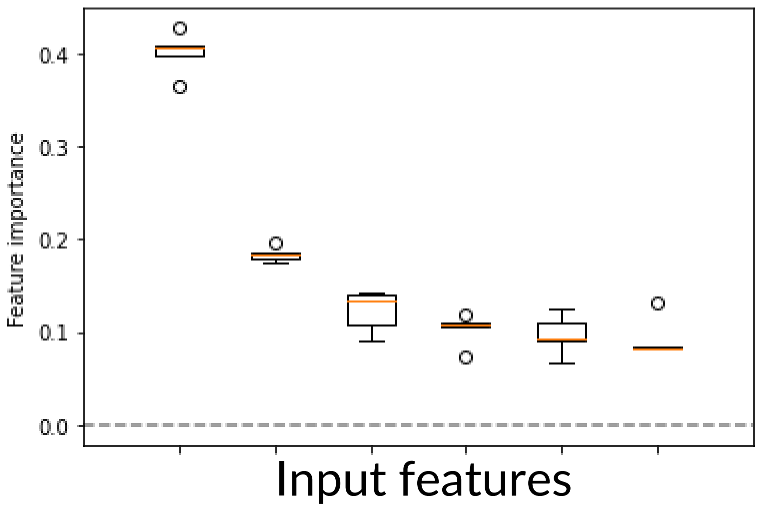


https://realkm.com/2018/04/23/optimization-and-complexity-the-cost-of-complexity-systems-thinking-modelling-series/

Explainability

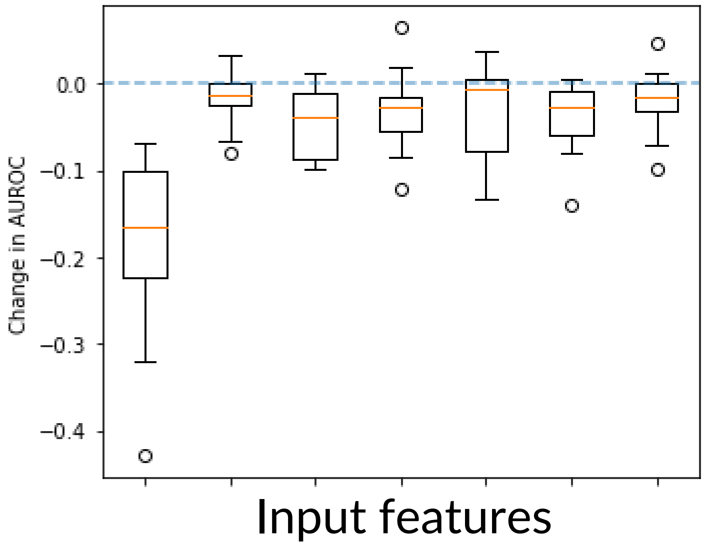
Feature importance





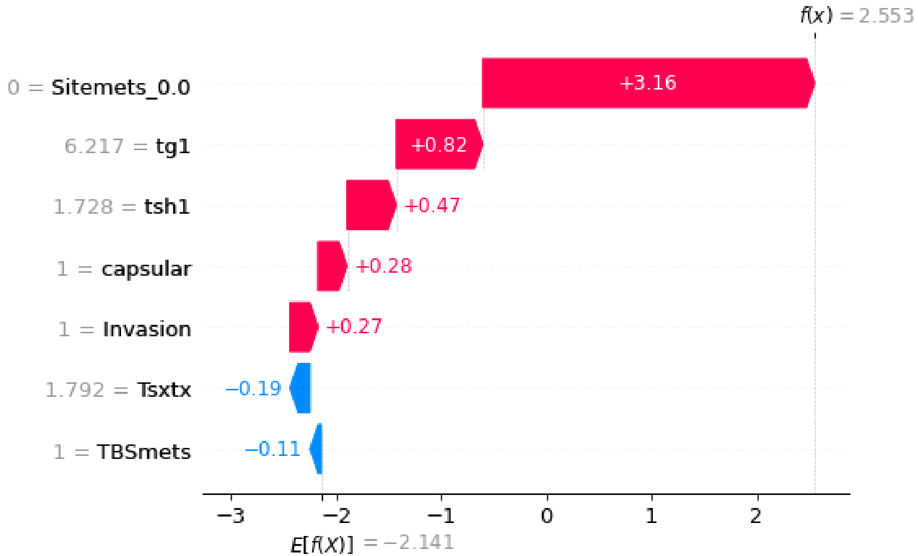
- Coefficients of linear, logistic, and SVM models
- Average improvement in impurity or entropy in tree models
- Model-level explanation

Change in performance after dropping a feature



- Compare performance with and without each input feature
- Big drop = important

Shapley value



- Change in predicted value due to the addition of a feature *i*
  - $\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S \cup \{i\}) - v(S)]$
- Sample-level explanation

Any questions?

See you on February 23rd