Recap

31 July 2023 15:16

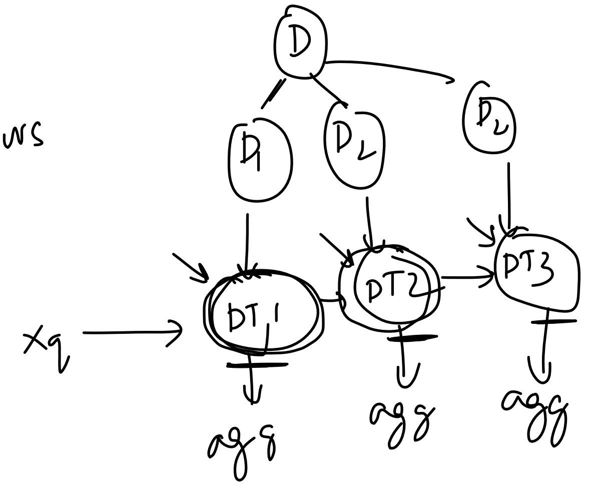
Wisdom of the crowd (maths principle)

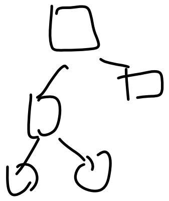
randomnes

Decision Tree

1) Bootstraping → rows

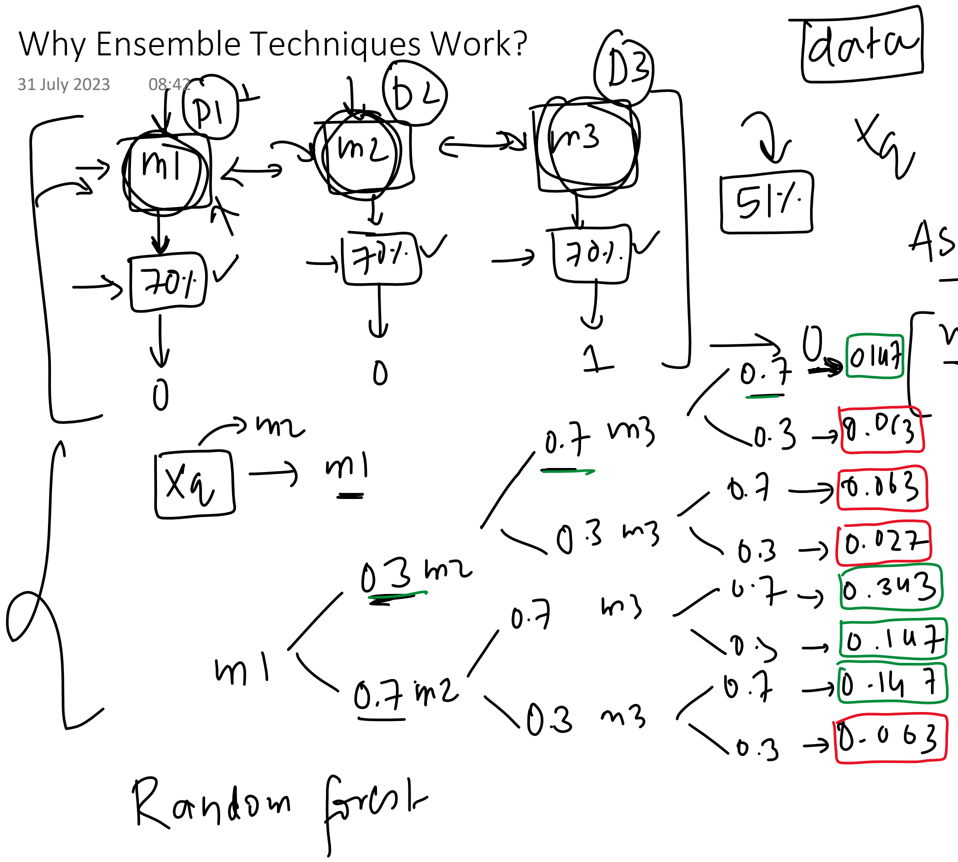
decorrelated

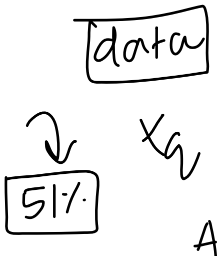




Why Ensemble Techniques Work?

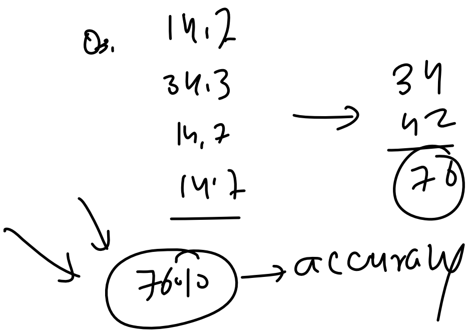
31 July 2023 08:42





Assumption

m1 m2 and m3 are independent models



Random forest

Random Forest Hyperparameters

31 July 2023 08:44

sqrt(p) fully grown 5 sqrt(5) tree

Forest Level HP	Tree Level HP	Miscellaneous HP
N_estimators	Criterion	Oob_score
Max_features	Max_depth	N_jobs
Bootstrap	Min_Samples_split	Random_state
Max_samples	Min_samples_leaf	verbose
Bootstrap_feature	Min_weight_fraction_leaf	Warm_start
	Max_leaf_nodes	Class_weight
	Min_impurity_decrease	
	Ccp_apha	

gini, entropy, logloss

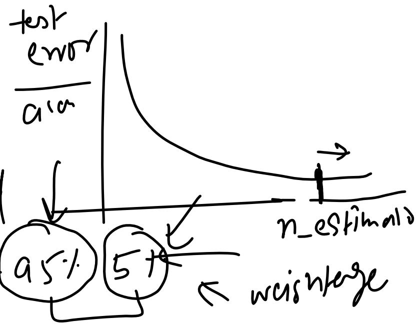
Bootstrap → True

pre

post

Random_state → 42

Class_weight → imbalanced



row sampling with replace without

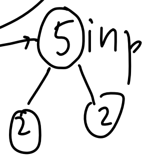
how many dt 100 → deep trees → low 5, 10, overfitting

100, 500, 1000

1000 rows

↳ 100 dt → 100 datasets sampling

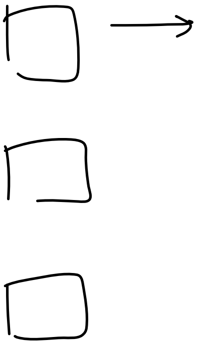
5 input cols



col samplin

row samp

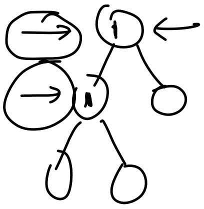
max_feature = 2



1000

500

100



node level col samp

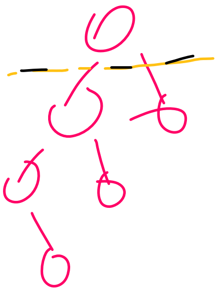
LBHV → deep decision tree

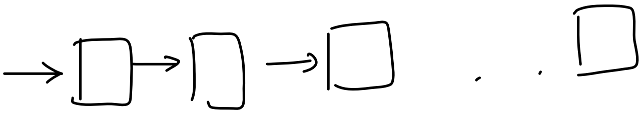
↓

LBLV

LB

HB





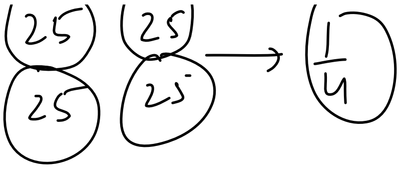
↓

multi-core → quad process

25 25 → 1/4

↓

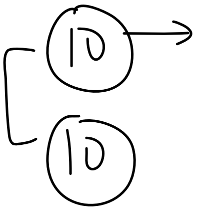
multi-core → quad process

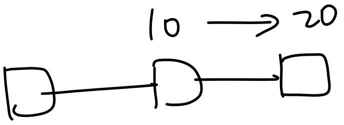


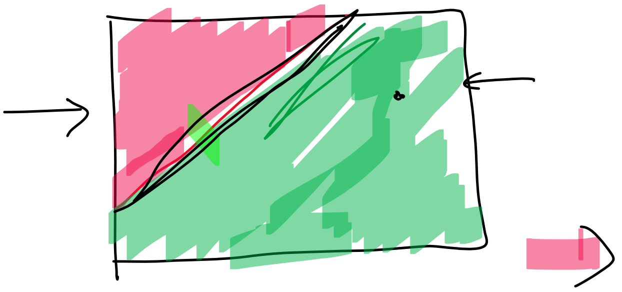
warm_start → training time is high

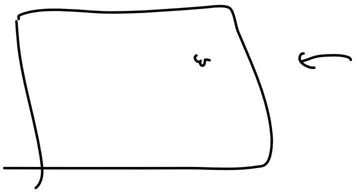
True → false

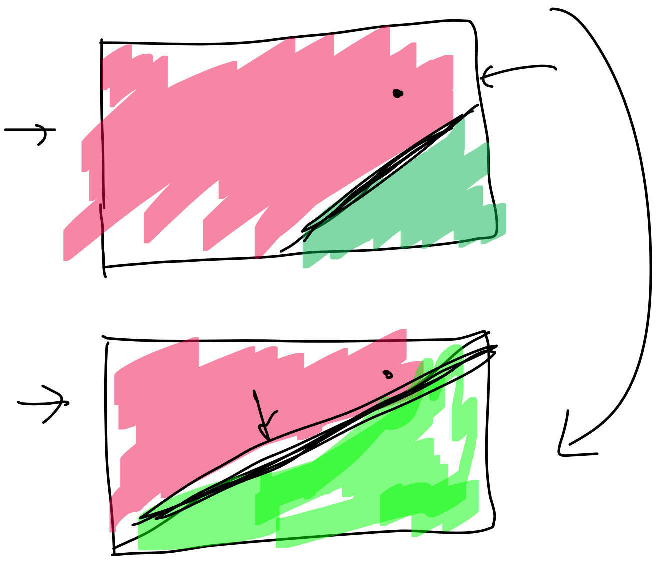
→ RF(n_estimators=10) 20 → 50 →

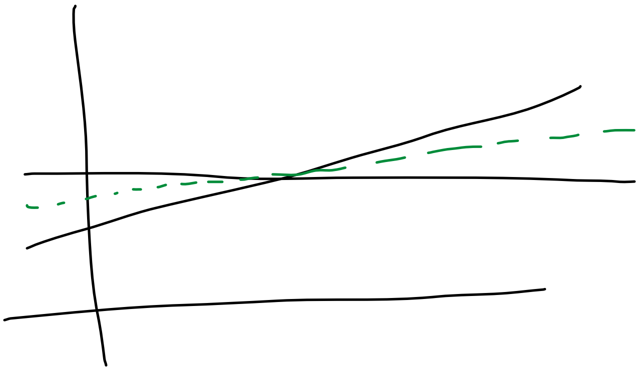






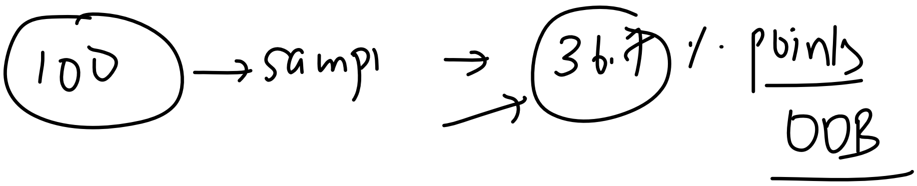






OOB Score

31 July 2023 08:47

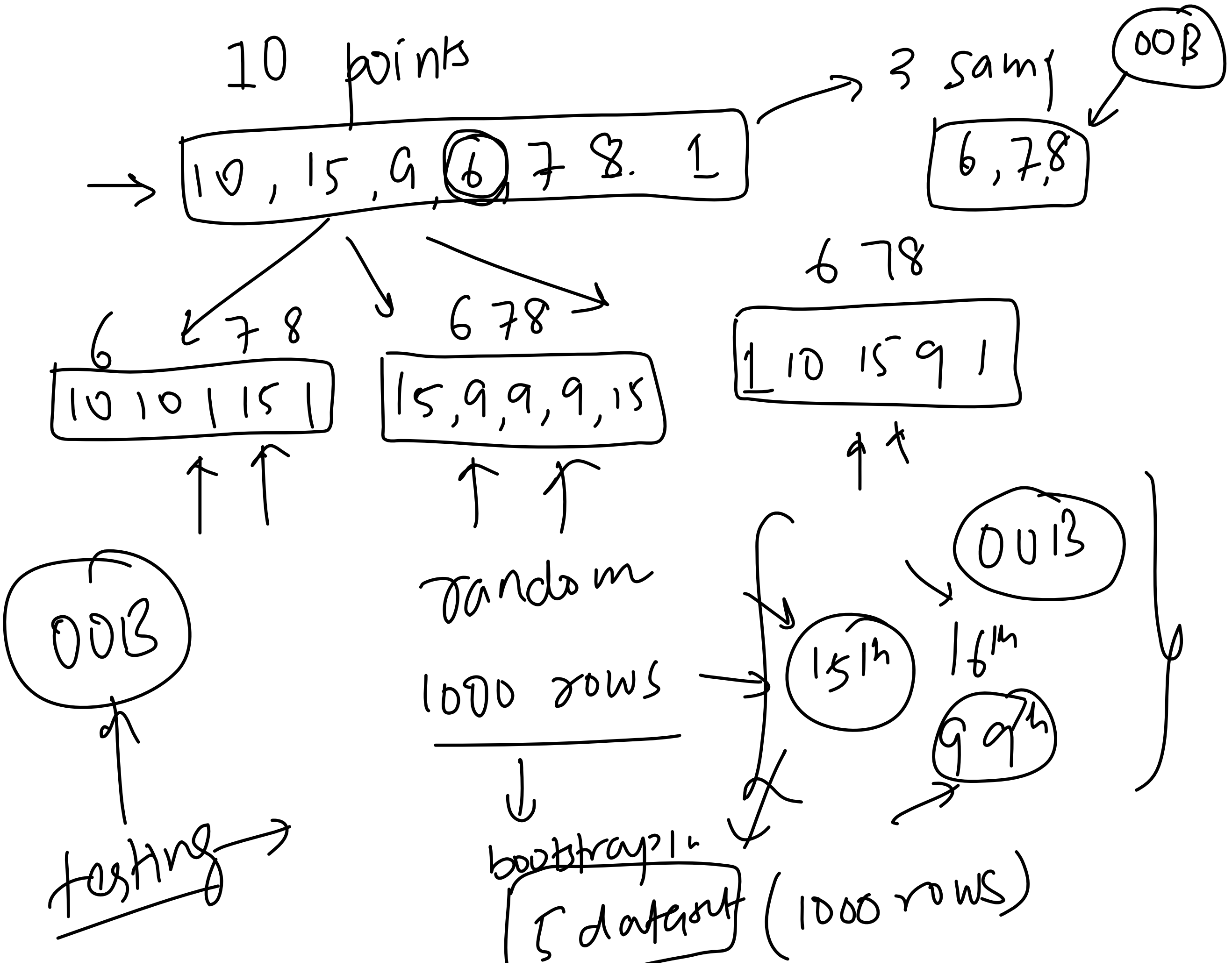


"OOB" stands for "out-of-bag". In the context of machine learning, an out-of-bag score is a method of measuring the prediction error of random forests, bagging classifiers, and other ensemble methods that use bootstrap aggregation (bagging) when sub-samples of the training dataset are used to train individual models.

Here's how it works:

1. Each tree in the ensemble is trained on a distinct bootstrap sample of the data. By the nature of bootstrap sampling, some samples from the dataset will be left out during the training of each tree. These samples are called "out-of-bag" samples.
2. The out-of-bag samples can then be used as a validation set. We can pass them through the tree that didn't see them during training and obtain predictions.
3. These predictions are then compared to the actual values to compute an "out-of-bag score", which can be thought of as an estimate of the prediction error on unseen data.

One of the advantages of the out-of-bag score is that it allows us to estimate the prediction error without needing a separate validation set. This can be particularly useful when the dataset is small and partitioning it into training and validation sets might leave too few samples for effective learning.



Test θ

5 dataset (1000 rows)

they are not a
part of training

Extremely Randomized Trees

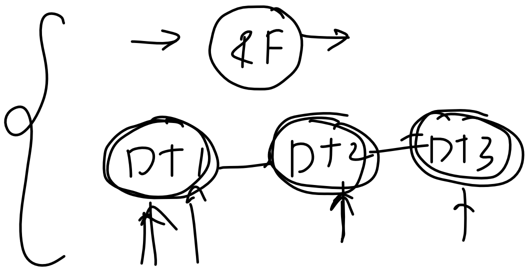
Extra Trees

31 July 2023 11:31

Extra Trees is short for "Extremely Randomized Trees". It's a modification of the Random Forest algorithm that changes the way the splitting points for decision tree branches are chosen.

In traditional decision tree algorithms (and therefore in Random Forests), the optimal split point for each feature is calculated, which involves a degree of computation. For a given node, the feature and the corresponding optimal split point that provide the best split are chosen. On the other hand, in the Extra Trees algorithm, for each feature under consideration, a split point is chosen completely at random. The best-performing feature and its associated random split are then used to split the node. This adds an extra layer of randomness to the model, hence the name "Extremely Randomized Trees".

Because of this difference, Extra Trees tend to have more branches (be deeper) than Random Forests, and the splits are made more arbitrarily. This can sometimes lead to models that perform better, especially on tasks where the data may not have clear optimal split points. However, like all models, whether Extra Trees will outperform Random Forests (or any other algorithm) depends on the specific dataset and task.



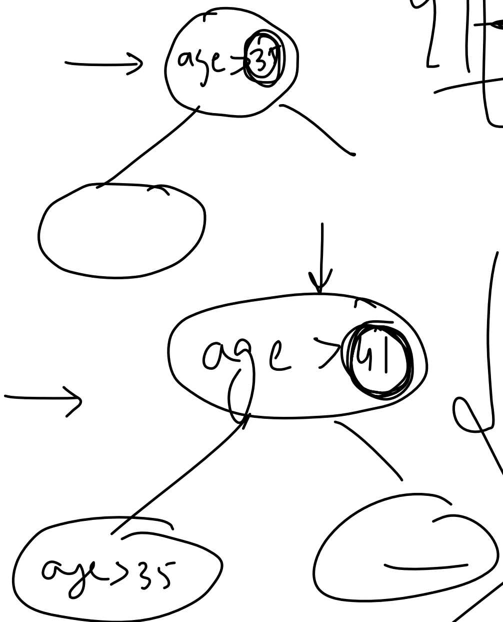
decorrelated

system → randomness

Yes

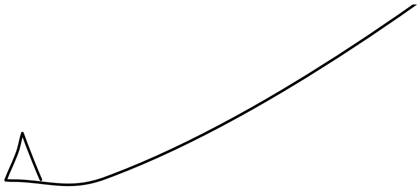
35

age	insurance
27	N
43	Y
35	



high bias

training accur ↓ decrease



Advantages and Disadvantages

31 July 2023 08:53

Advantages

- Robustness to Overfitting: Random Forests are less prone to overfitting compared to individual decision trees, because they average the results from many different trees, each of which might overfit the data in a different way.

- Handling Large Datasets: They can handle large datasets with high dimensionality effectively.

- Less Pre-processing: Random Forests can handle both categorical and numerical variables without the need for scaling or normalization. They can also handle missing values.

- Variable Importance: They provide insights into which features are most important in the prediction.

- Parallelizable: The training of individual trees can be parallelized, as they are independent of each other. This speeds up the training process.

- Non-Parametric: Random Forests are non-parametric, meaning they make no assumptions about the functional form of the transformation from inputs to output. This makes them very flexible and able to model complex, non-linear relationships.

Linar

non-linear

Disadvantages

- Model Interpretability: One of the biggest drawbacks of Random Forests is that they lack the interpretability of simpler models like linear regression or decision trees. While you can rank features by their importance, the model as a whole is essentially a black box.

black

- Performance with Unbalanced Data: Random Forests can be biased towards the majority class when dealing with unbalanced datasets. This can sometimes be mitigated by balancing the dataset prior to training.

- Predictive Performance: Although Random Forests generally perform well, they may not always provide the best predictive performance. Gradient boosting machines, for instance, often outperform Random Forests .If the relationships within the data are linear, a linear model will likely perform better than a Random Forest.

- Inefficiency with Sparse Data: Random Forests might not be the best choice for sparse data or text data where linear models or other algorithms might be more suitable.

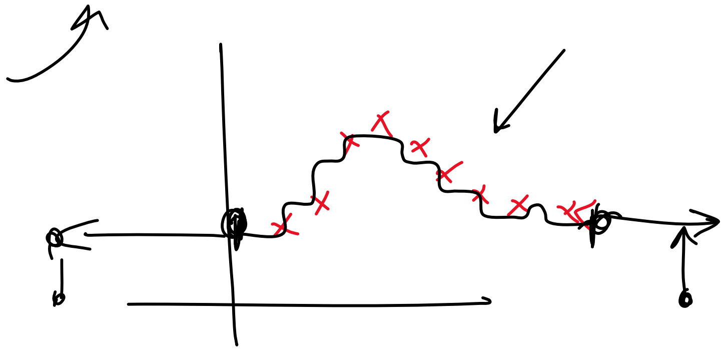
- Parameters Tuning: Although Random Forests require less tuning than some other models, there are still several parameters (like the number of trees, tree depth, etc.) that can affect model performance and need to be optimized.

- Difficulty with High Cardinality Features: Random Forests can struggle with high cardinality categorical features (features with a large number of distinct values). These types of features can lead to trees that are biased towards the variables with more levels, and may cause overfitting.

OHE

- Can't Extrapolate - This is because they do not predict beyond the range of the training data, and that they may not predict as accurately as other regression models.

data, and that they may not predict as accurately as other regression models.



linear