Ensemble : Using more than one algorithm.

In ML Pipeline – we have huge data – we will create cluster using un supervised data and apply each cluster different technique.

May be one cluster – DT , another cluster may be –RF like that.

Custom ensemble Two types:

1. Bagging : Random Forst (a) Classifier (b) Regression
2. Boosting: Adaboost, Gradient Boost, XTREAM GRADIENT boost.

Difference between Bagging and boosting:

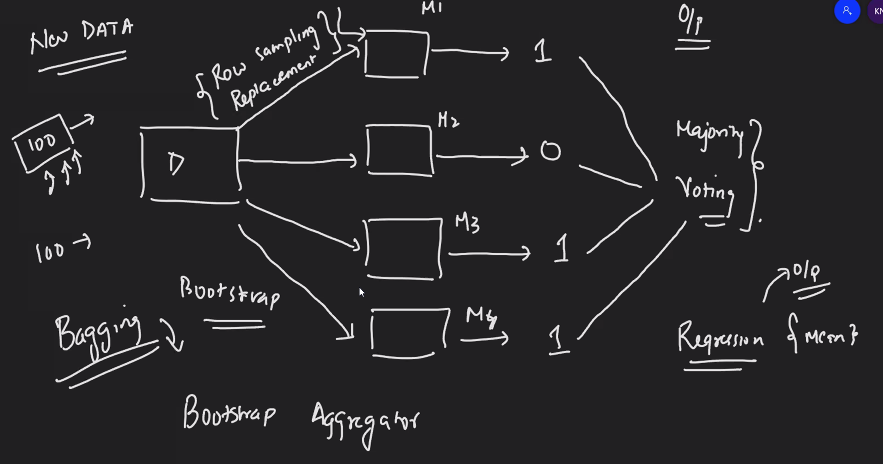
Bagging: From Data we will take some row( Row sampling) Trained with different different models.

New data given to each model.

Assume Model 1 give 1 ,Model 2 give 0, Model 3 give 1 and model give 1

So the output will be 1 ( Major Voting) ( Real use case may be 100 data)

For Regression will consider mean of each model.



Boostrap mean : The data row sampling (random rows ) input to multiple model and get voting.

Bagging is parallel.

Boosting: we are combine the weak learner in sequential.

The weak outpout is input to next model. So that weak is boosted.

**Random Forest:**

Use DT.

Whenever we apply DT low bias and high variance. ( Mean High Training Accuracy & Test Accuracy is low)

To reduce the High Variance into Low variance we will use Hyper Parameter or Random Forest.

Random Forest will create multiple DT and Final result will be majority voted.

Feature sampling, Selected Separate feature given to each model.

May be few feature and row will repeat, But we are creating the model using different different feature and rows, Then the output will be aggregated.

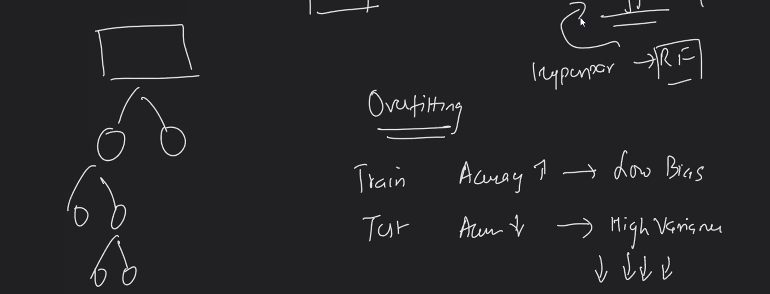
Row sample and Feature sample only for Training data.

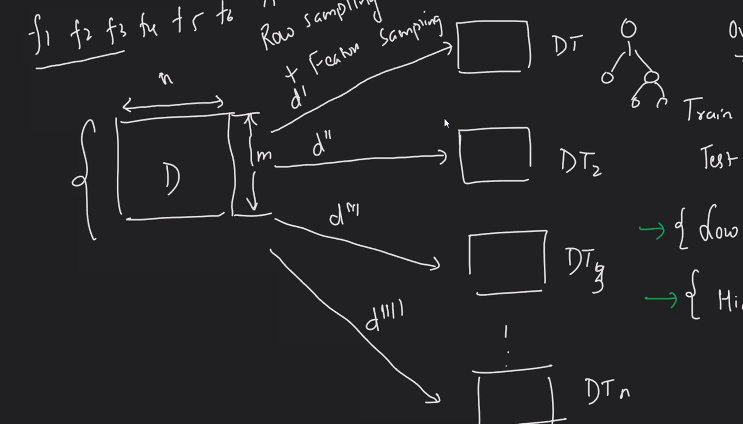
Most of the Kaggle competition will use this type of algo to solve.

Finally the model will be Low bias & Low variance ( Model Genralized)

Low bias – Training good data

Low Variance – Test accuracy are good.





Interview:

2.What is the difference White box and Black box model?

* White box – DT – we can interpret -Lime,Shapash lib will help to interpret the model.

SVM, Logistic, Linear, Nave Biays, SVR

Black box : All NN Model, RF

3.Is the random forest algorithm will impact with outlier? No

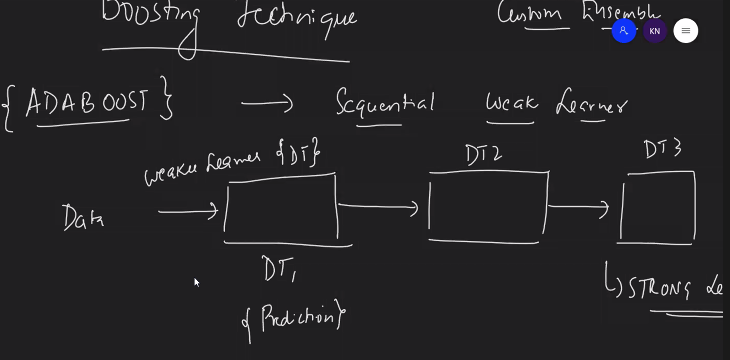
4.Do we need standardization and Normalization – No Because internally will use DT & will split the feature.

5. Random Forest will over fit : No

6. But DT is Low bias or High Variance so prone to overfit.

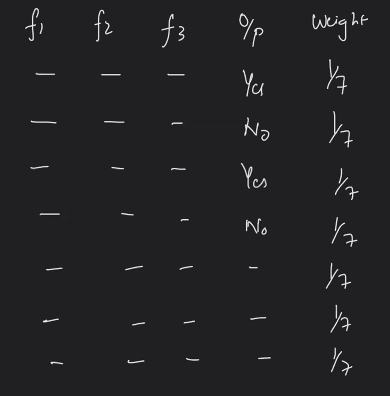
7. Time series will used where?? Forcasting.

8. Is there any other bagging algo –Yes RF and custom ensemble.



Mistake ( wrong prediction) will pass to subsequent model.

Steps:



All feature assigned to same weight.

Step 2:

Stump is one level Decision Tree. Based on that we can calculate the Entorpy.

Whether ensemble suitable for High data set? No will take more time each time 100 Tree will create.

For regression more MSE will passed to next stump