
upGrad
Raho Ambitious

upGrad

#LifeKoKaroLift

SGC Coaching Session :18

RANDOM FORESTS

Session By:Trinadh Veeramachaneni

Agenda:

upGrad

From decision trees to random forests

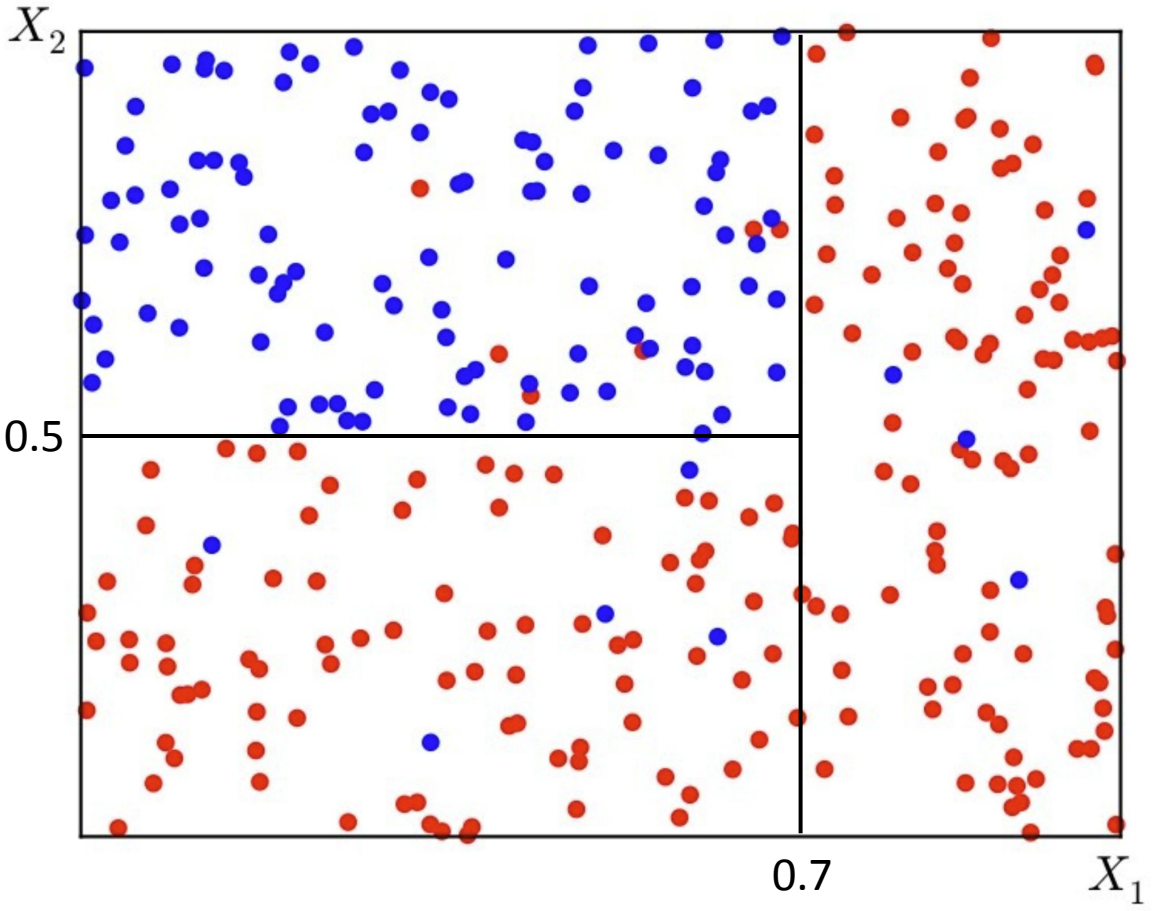
Bootstrapping, Concepts of Random Forest, Ensemble learning

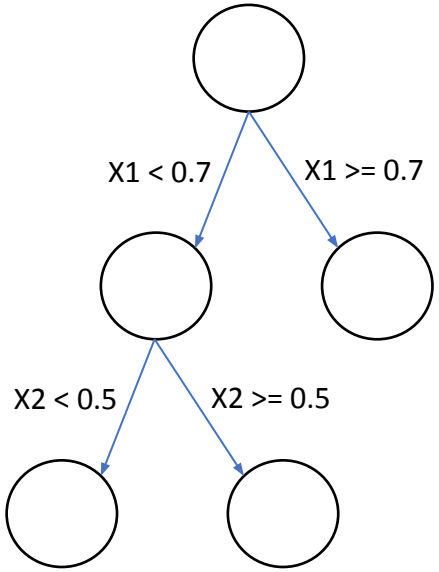
Applications

Case study/Examples

Personalized Feedback and Doubt Resolution

Let's recall decision tree ...





Questions

How do we select the variable on which splitting is done?

How do we select the bins/cutoffs for the selected variable?

How do we evaluate a

Why not decision tree ...

upGrad

- Overfits when data is less
- High variance model
- Gives more importance categorical variables with more levels

Questions

What different things can you do with DT to get better accuracy when the data is less, and your tree is overfitting?

What different hyperparameters should you adjust ?

Bagging (Bootstrap Aggregation)

upGrad

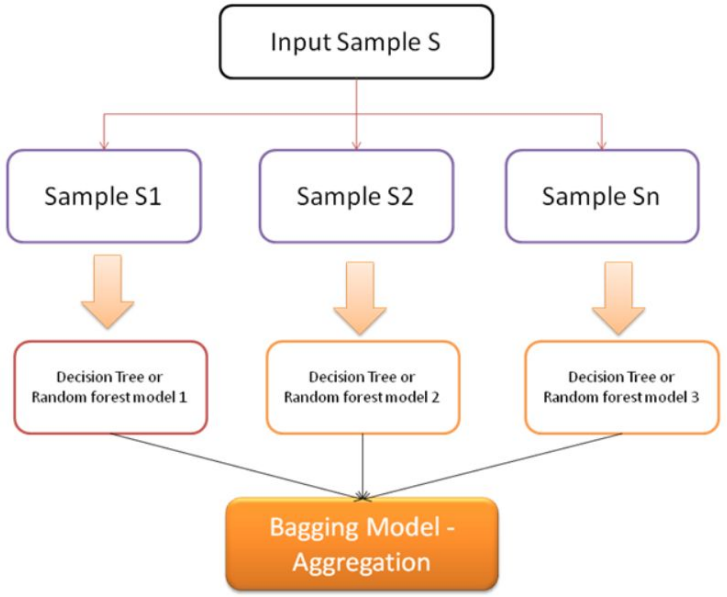
Bootstrapping is a process of creating random samples with replacement for estimating sample statistics. With replacement means , the sample might have duplicate values from the original set.Once bootstrap samples are created, model classifier is used for training or building a model and then selecting model based on popularity votes.

In case of a classification model, a label with maximum votes will assigned to the observations.

In case of a regression model - average value is used.

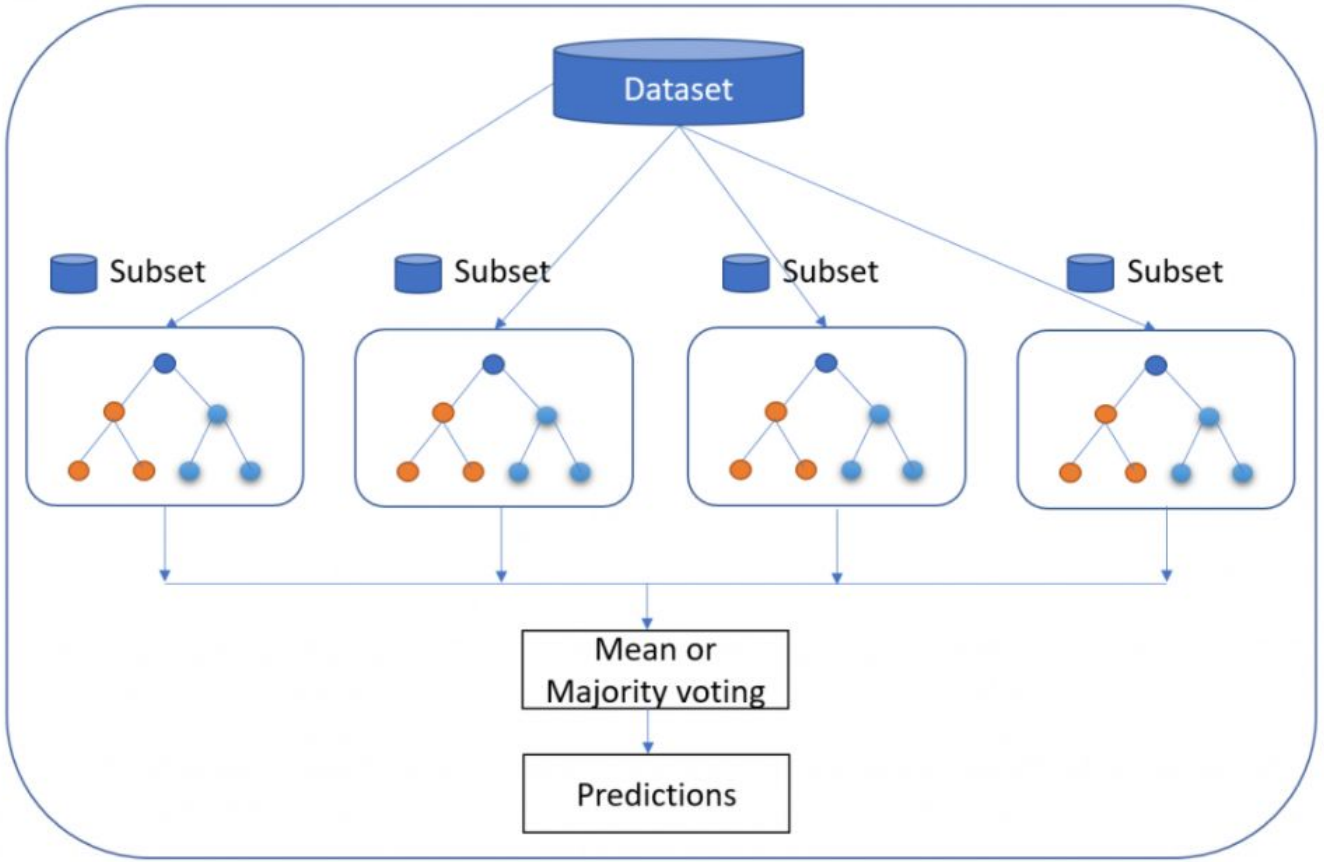
Bagging is an ensembling process – where a model is trained on each of the bootstrap samples and the final model is an aggregated models of the all sample models.

```
Sample S = {10,23,12,11,34,11,1,4,2,14}

Bootstrap sample 1:              {10, 23, 11, 4, 2, 14, 11}

Bootstrap sample 2               {23, 10, 12, 11, 14, 2, 14} - 14 is duplicate (which means with replacement in
bootstrap sample set)

Bootstrap sample n :   {10, 1, 2, 2, 14, 1, 23}
```



Random Forest

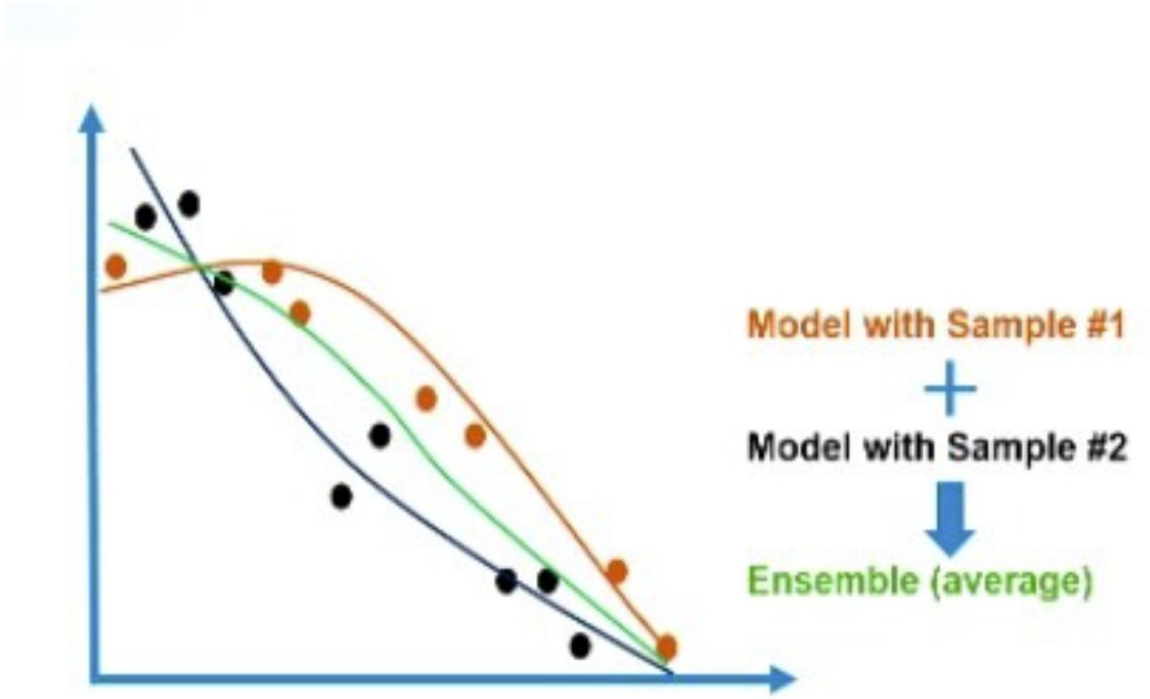
upGrad



Bagging is averaging

aka bootstrap aggregation

Why do we do it ? How does bagging help ?
How do we do it ?

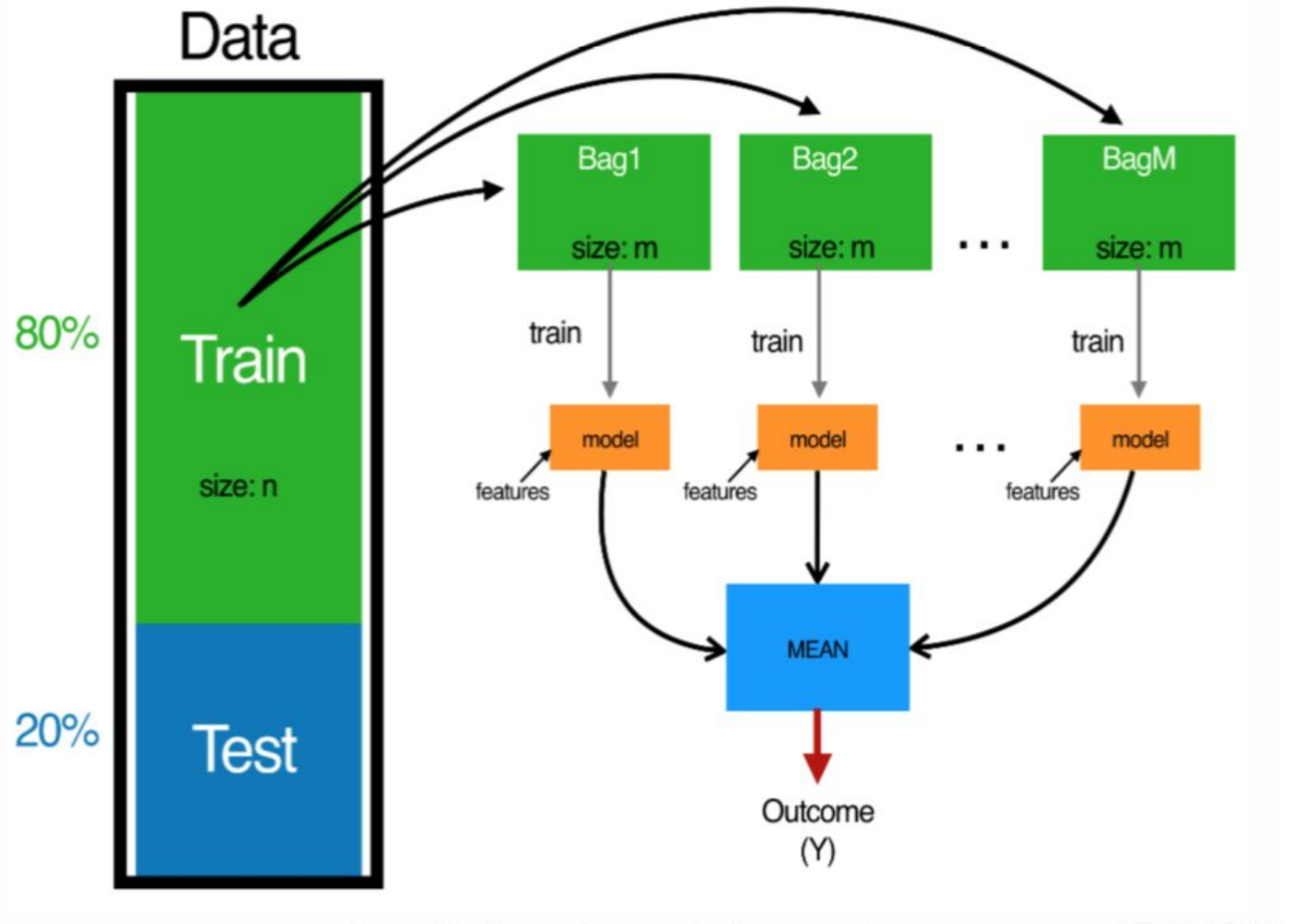


Which colour line will give you better predictions?

Think about you deciding to go to a movie *by asking one of your friend* Vs *asking all your friends* and aggregating their inputs

Understanding bagging... 2/2

upGrad



How do we do it ?

1. Split your data to train and test
2. Do bootstrap sampling on your training dataset. Randomly sampling *m* records from *n* records for *M* times. You will get *M* bags.
3. Now train your model separately on each of the *M* datasets. You will get *M* models.
4. These *M* models will give you *M* predictions
5. The final prediction will be an average of all these *M* predictions

Now, from decision trees to random forests ...

upGrad

Random forest (RF) – a collection of decision trees that **run parallel**

Each decision tree is made of a **different dataset – different features, different records** which can overlap

Doing EDA for random forests is easy ...

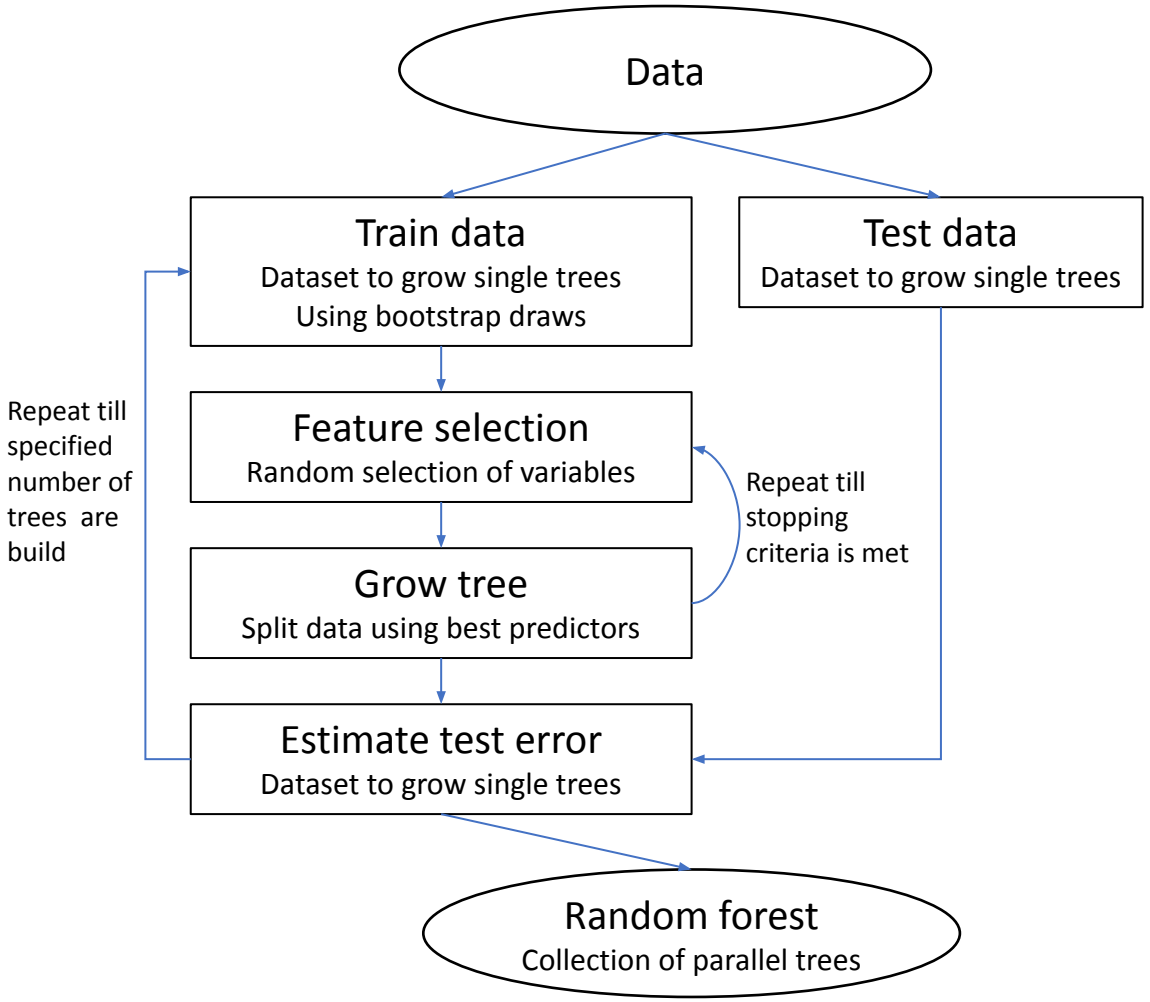
upGrad

- ✔ Remove variables with zero variance. They are of no use.
- ✔ Very highly correlated variables (> 0.8) can be removed.
- ✔ Try doing variable selection using variable importance feature available with RF
- ✔ RF are robust to outliers and missing values. Remove them if they are aberrant observations (e.g., due to recording errors)
- ✔ Standardization/Normalization of variables not required

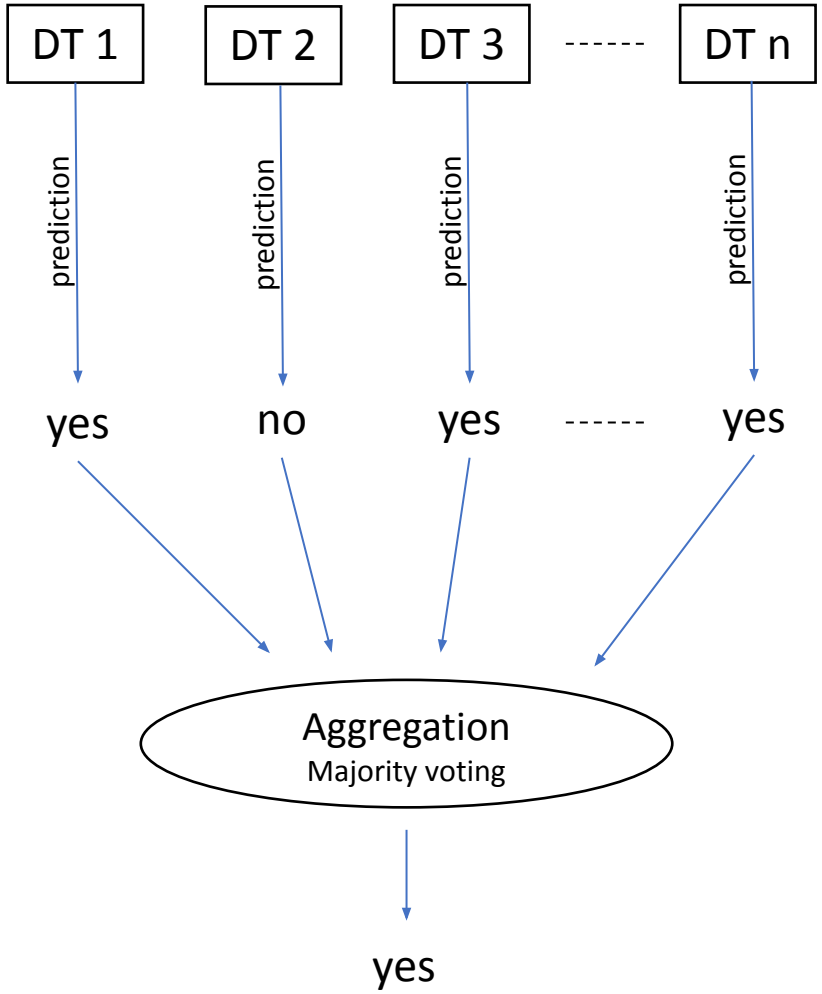
Training & predicting Random Forest model...

upGrad

How a random forest is created



How a random forest is run



Great, you made you first ensemble model !

upGrad

How did we just do?

We created multiple predictors (decision trees) for the same problem and made our final prediction by aggregating all the predictions from individual predictors, leading to better prediction.

Let's finally evaluate our RF model ...

upGrad

Using confusion matrix and cost

Say we are predicting if a mushroom is poisonous …

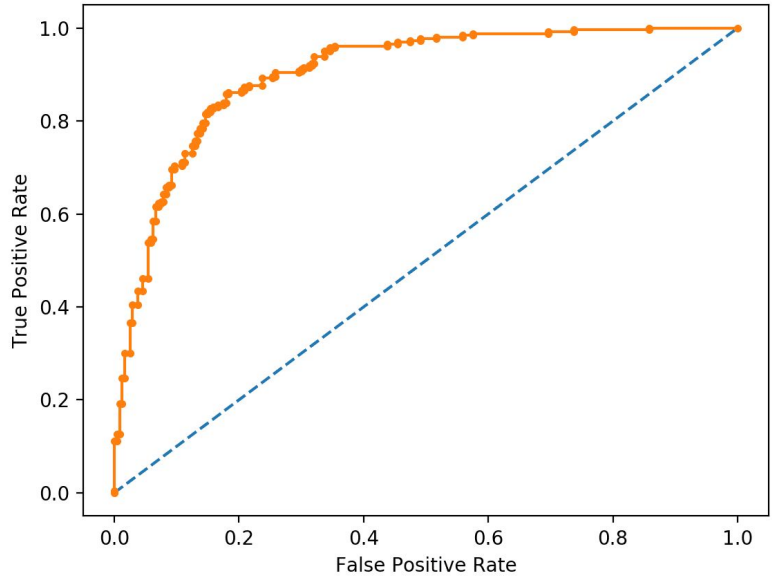
Confusion matrix

Actual \ Predicted	Edible	Poisonous
Edible	TP	FN
Poisonous	FP	TN

Cost matrix

Actual \ Predicted	Edible	Poisonous
Edible	1	-1
Poisonous	-10	4

Using ROC chart



Cost Matrix

upGrad

Computing Cost of Classification problem

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M1	PREDICTED CLASS		
ACTUAL CLASS	sni	+	-
	+	150	40
	-	60	250

Accuracy = 150 + 250/150+250+40+60 = 80%

Cost = 150(-1) + 40(100) + 60(1) + 250(0)

Cost = 3910

Model M2	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	250	45
	-	5	200

Accuracy = 250 + 200/250+200+45+5 = 90%

Cost = 250(-1) + 45(100) + 5(1) + 200(0)

Cost = 4255

There are two models namely Model M1 and M2 both of which are having correct predictions and wrong predictions.

If we compare both the models and if we check their accuracy. Accuracy for Model M2 is higher compare to Model M1, however the cost for Model M2 is higher compare to Model M1.

So it depends on what kind of problem statement we are facing.

If we are focusing on accuracy then we will go with the Model M2 (In this case we need to compromise on cost) , however if we are focusing on cost then we will go with the Model M1 (In this case we need to compromise on accuracy).

Advantages of RF

upGrad

- Low variance & less prone to overfitting – thanks to bagging
- Improved accuracy
- Robust to missing values and outliers
- Fast during training – individual trees can be computed parallelly (downside is that prediction will slow as no. of trees are huge)

upGrad

Questions?