Classification deals with categorical target variables.

You are admissions offer deciding admissions to the university.

Binary classification yes/no true or false

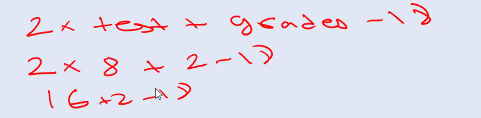
Multiclass example is sports 2+ categories.



We cannot determine from two datapoints.

Grades tests exploratory variables.

We draw a line separating the admission criteria.



Line properly 18-18=0 point that lies on has a score of 0.

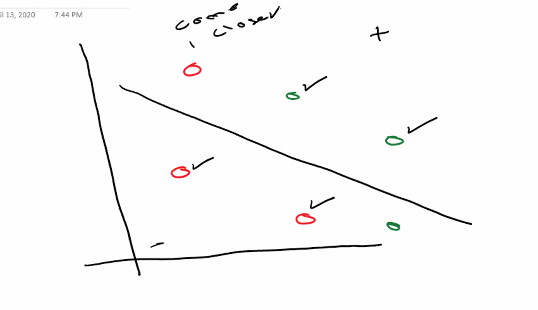




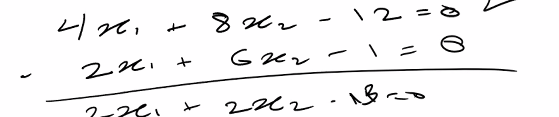
Points below line green y^=0 in actual it should have been 1

Points red above line y^=1 in actual is should have been 0

We have to compare the y with y^ to find the error.



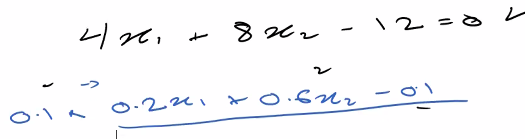
First line drawn is random



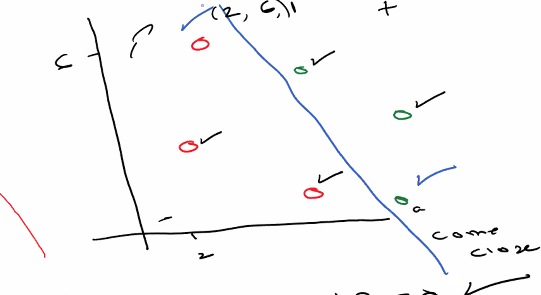


Movement should be very slow.

Learning rate multiples these points with a very low value.



Substraction will be using these blue points. With learning rate of 0.1



At the end line will be correctly classified

We take derivative with respect to error and line movement will continue until error is minimum.

If we take gradient descent the step function becomes unable to evaluate itself.We cannot calculate the confidence using step function.We cannot find the exact percentage.Every point has a confidence level.

Sigmoid function changes into probabilities.By using the values calculated from the line.

In step function it only considers 0 or 1.It does not consider the result only consider whether its greater than 0 or not.

On the line sigmoid function gives 50% value.

On the line step function gives 0 value.

Predictions for getting admissions.

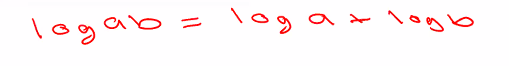
0.7 for getting admissions and 0.3 for not getting admissions.

Red means not getting admissions and blue means getting admisions

We will consider inverse value for misclassified point.

Maximum likehood means after multiplying the one which gives the best result is selected.

Multiplication increases computing cost and changing one value has a lot of effect.So we can use sum method.



Cross entropy is inverse the model with less value is better.

Logistic regression is used for classification.

Sns.set() we include default arguments

For numerical scaling we used minmax scalar() for the categorical variables we use one hot encoding to change into proper format.

first column 0 probability second column one probability

we only need 1 column not 0 column

hot topics .

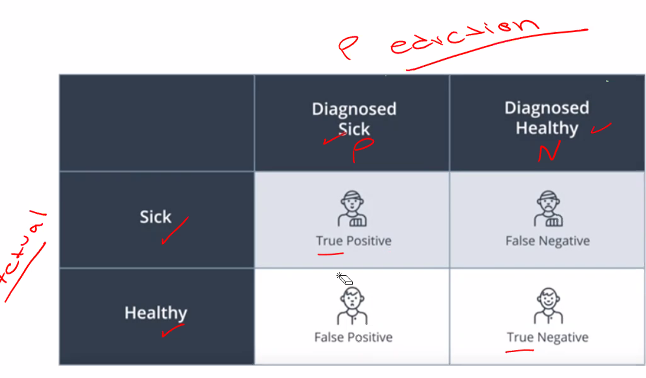


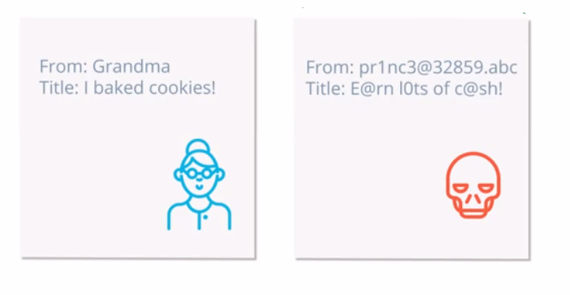
Model decides whether healthy or sick.

Machine learning algorthms throws fake emails in spam.

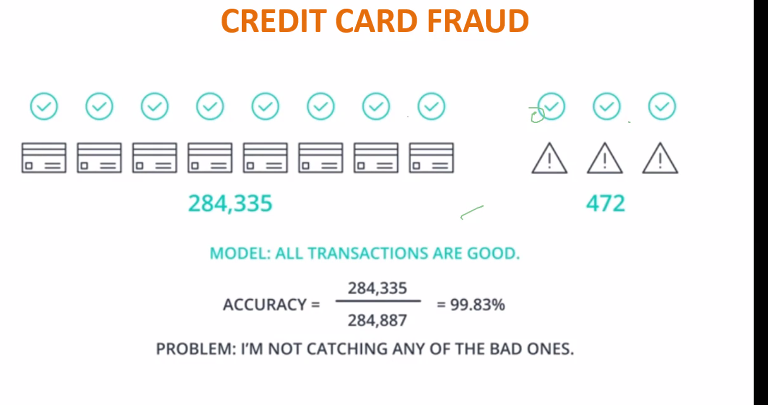
Row=actual confusion matrix

Columns=prediction.





Divide true by the total.



If data is skewed accuracy becomes biased.

Accuracy is not a good parameter when there is uneven distribution.



False negative is critical error in this case(medical).

In email one false positive is the critical.

Precision Quiz 6/8 and 5/6

Recall quiz.6/7. And 5/7 for the negative ones.



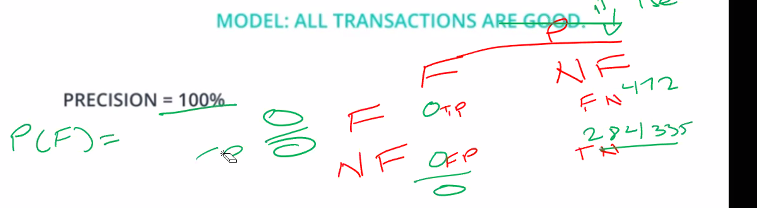
We cannot do average because we cannot assign equal weight to these.

|  |  |  |
| --- | --- | --- |
|  | Model Says Fraud | Model Says Good |
| Fraud | 472 | 0 |
| Good | 284335 | 0 |

Model says all are fraud precision=472/284335+472=0.16 Recall=472/472=1

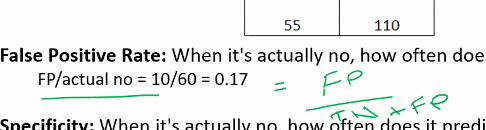
F1score is the harmonic mean of precision and recall.

Ideally both precision and recall should be high.



0/0 is infinity that is why 100% precision.

We reversed confused matrix.



**ROC CURVE.**

You only need one variable minimum for classification and regression.

We repeat these values at multiple points when we are splitting we plot these points.We plot the true positive and false positive rates.

Complete square area under the curve=1.

Nearer to 1 is good model.

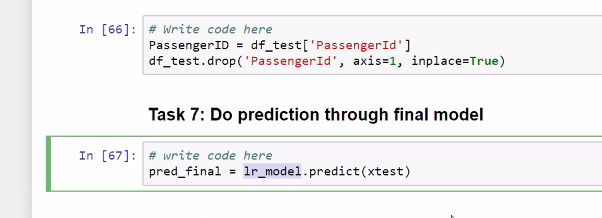
Confusion matrix in code is not given in correct format.We need to convert it into the correct format.

IT shows inverse format.

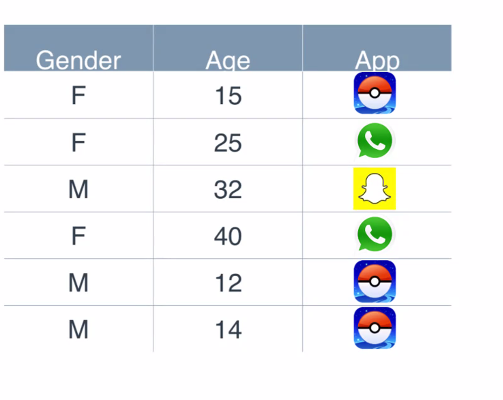


Ravel() converts 2D array to 1D.

Every evaluation function requires ytest and pred.



Decision Tree algorithm



Recommendation engine recommend app on the basis of two exploratory features Sex and age.Y is feature is that we want to predict(app).

Decisive feature is Age. It is the parent node.Root node.

We need to create homogenous set.

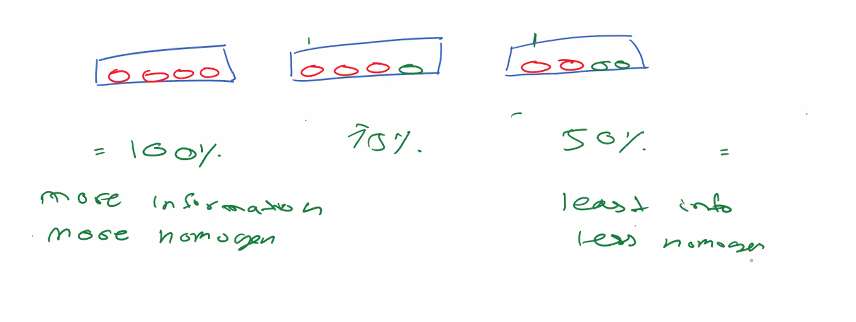
Decision node must have two children.

We have algorithms to decide root node and other nodes.

Scores will be compared at the end and the one with the best score with be the parent/root node.

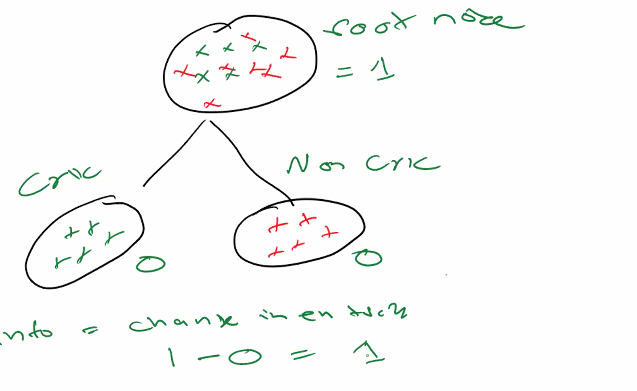
P=success q=failure.

Distribution is not equal female and males are not equal.



More information more homogenous=less entropy.

Information gain=change in entropy.



The one which has the most difference in entropy is the best node.

Information gain requires two nodes in order to find gain. You cannot find it using one node.

1-0.86=difference more (so this is parent node).

1-0.99=difference less

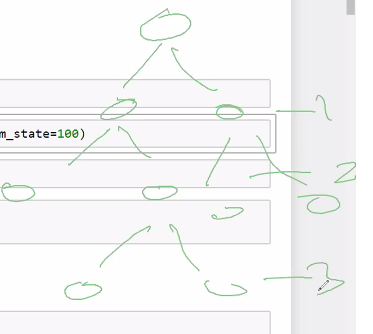
Chi square is also used for feature selection.

Chi square expects child to have same distribution as parent.

Formalas are the important things to remember.

In chi square the biggest the value the better it is.

**Do tasks till decision use gini.**



Height=3 max depth=3

For information gain it would have been entropy instead of gini

True false is due to one hot encoding 1 and 0.

Not Mr=true

Blue color shows survived red orange shows not survived.

The genie index process is repeated node per node highest one is made the parent node in each case.

Values show how many 0s and 1s are there.[0 1] format.

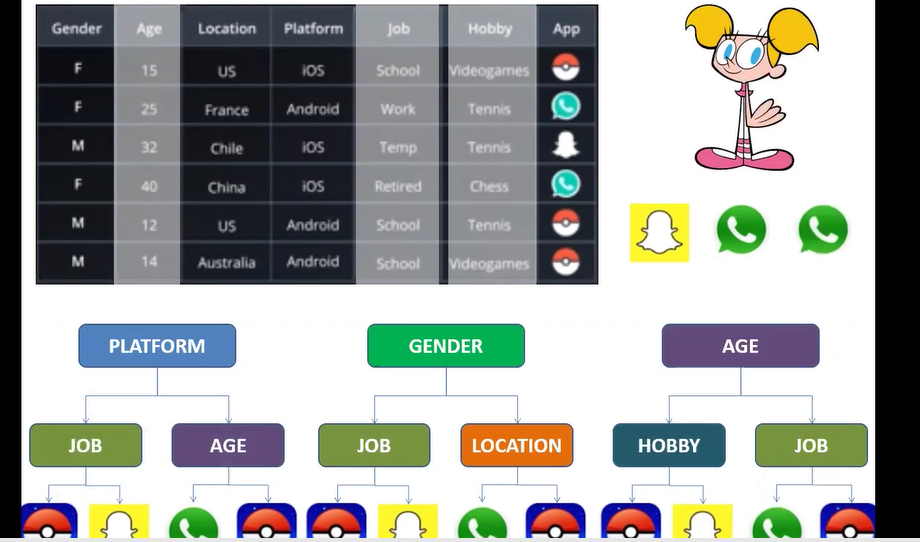
When depth increases no of features’s importance increases.

When features are large it overfits and memorises the problem.There is no generalized rule.

For feature selection use decision tree and use those features in another algorithm good jugar.

Randomly select three columns.

The prediction is done through polling.Many decision trees are made and the decision is made by majority.If 3 say survive and 1 says non survive survived is selected.



Majority is whatsapp so whatsapp will be recommended.

N estimator decides no of subtrees.(in this case 3)

Max features decides how many features to use. In this case three features platform,job,age

SQL cannot be used to train data.

Out of time testing.

Model made on june data testing accuracy 80%.We will use model on march we quote out of time testing to the business.To cover the seasonality impact.

More data random forest gives issues.

Hyperparameters effect model performance(increasing or decreasing).

We have kept parameters by default.we don’t know which to pick.

More features since 10 trees and all trees have different exploratory features.

Grid Search CV finds optimal hyper parameter.

The values of accuracy in the grid are calculated by using cross validation.Cross validated values are put into the grid. For example with a cv=5.Score we find Is the cross validation score.

We keep min\_samples leaves at 1 and vary depth then we keep it at 2 and vary depth and repeat the process.It works like a nested for loop.Algorithm run 25 times.

Grid Search View uniform flow.

Random Search selects random values for example depth 26 min\_samples \_leaves 5 etc..If you do not the optimum point Random search helps to pinpoint that.

The grid will run like 9 1 1 then 9 1 2 in a for loop format.

Best params will be returned.

N\_jobs is parallel threading it allows to run 4 jobs at once instead of waiting for one to complete it runs four jobs in parallel with each other which speeds up the process.

Scoring criteria is accuracy for now it can be precision or recall etc..

Might be that it maybe between these ranges so in next code we give values near the sweet spot discovered above for example for n estimators around 18 for depth around 10 for min leaves around 2.

Degree is hyper parameter in case of linear and logistic regression.

10 numbers between 10 and 1000. We give hardcoded values for min split.

Verbose shows backend computation.

In real life we use multiple algorithms then we do model comparison based on business knowledge which is more important recall ,precision etc.

Bagging and boosting are those types of algorithms that work when nothing is working so important.

Jugar find the range using random search then group GridSearch.

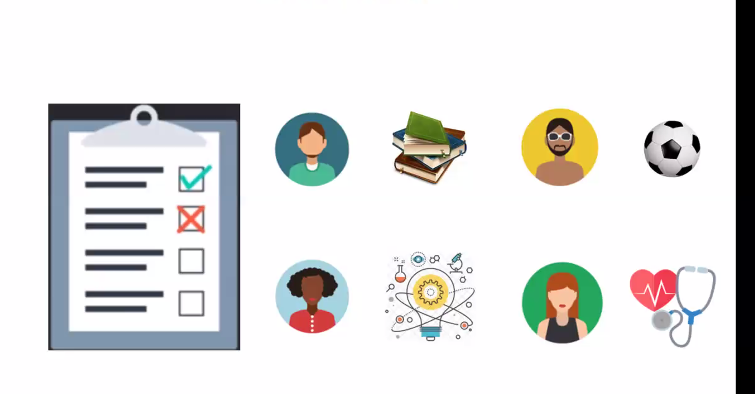
Multiple outcomes combined create output.=ensemble methods



A person had quiz he passed it friends friends help to find answer.All friends solve the question paper and majority polling occurs

Majority says true first is true majority says false second is false.

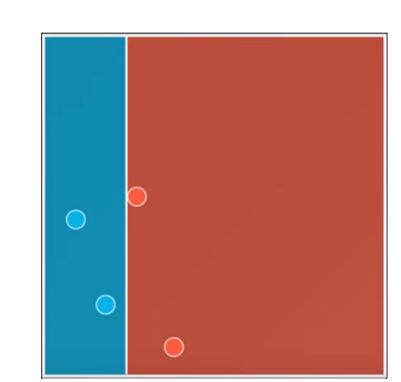
Boosting is we let people who are specialist answer the question.We let those people answer the question who are good at it.

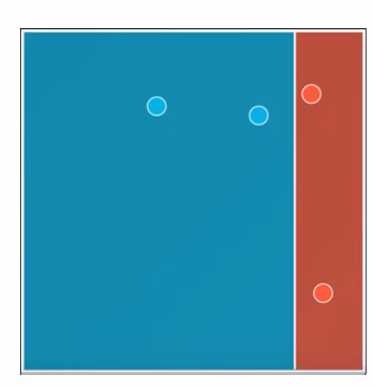


Friends are weak learners.The result aggregated is strong learner.Weak learners combined create strong learner.

Bagging will divide the overall data point into subsets.

Then it will draw a line to separate the points.positive and negative.

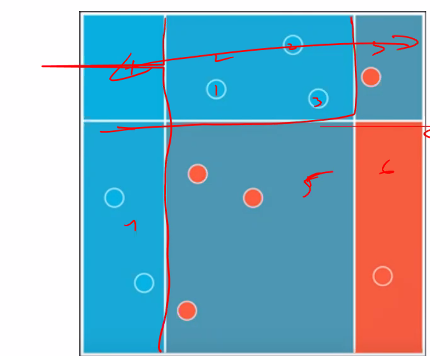


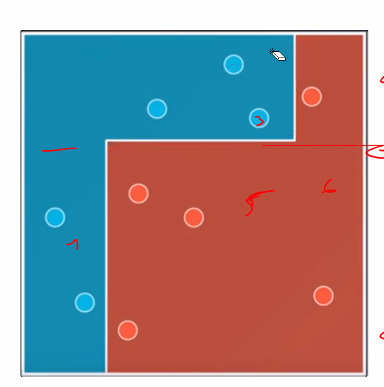


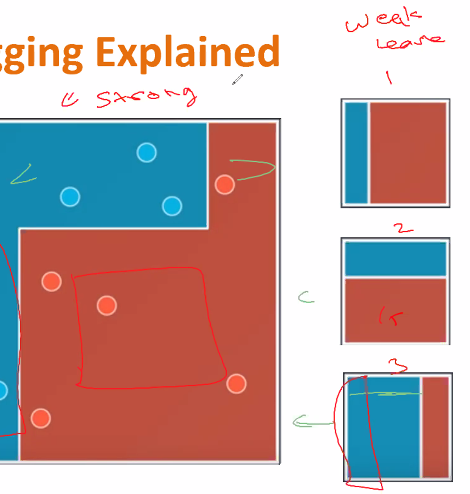


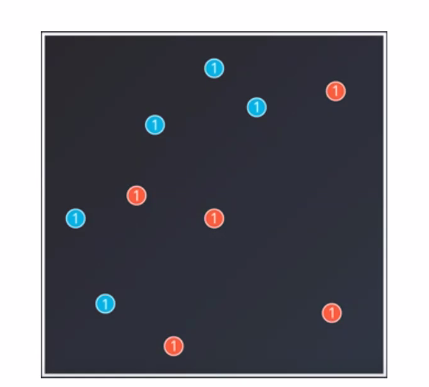
These subsets get saved in memory.

we merge these.The area that comes two times is taken as positive but the area that is positive only once is taken as negative.The areas that more positive are due to the fact that they are voted blue twice.



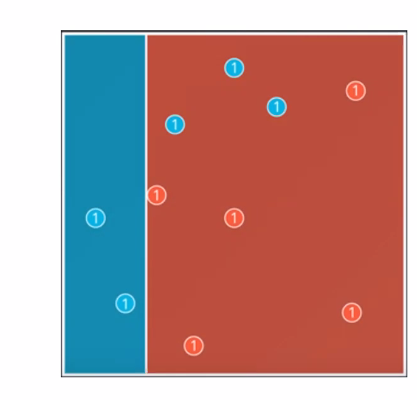




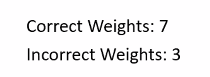


Boosting

Starting all data points assigned weight 1.

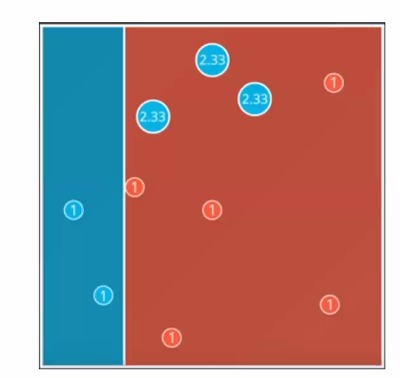


Then we draw a model to separate the blue and red.



We increase the miss classified point’s weight.

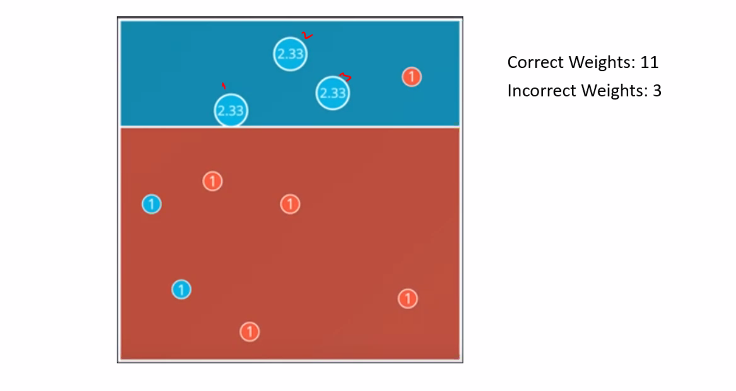
7/3=2.33 we make it 50 50 so that in next model line will be accurate.



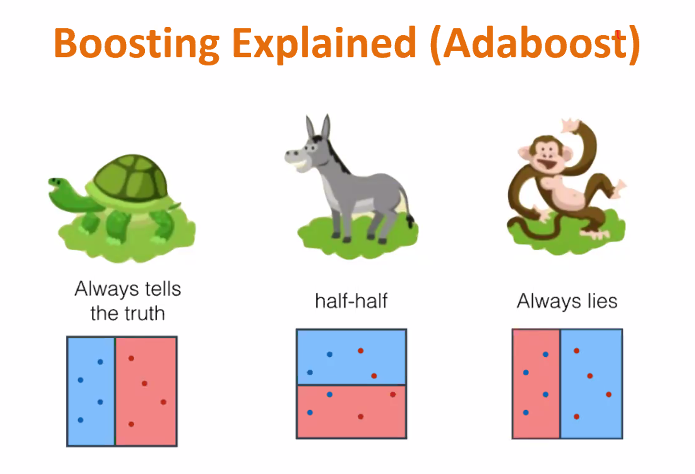
In the next model these lines given more importance.

Weight=7

Incorrect=7



In next model these are covered but some other are missclassifed.



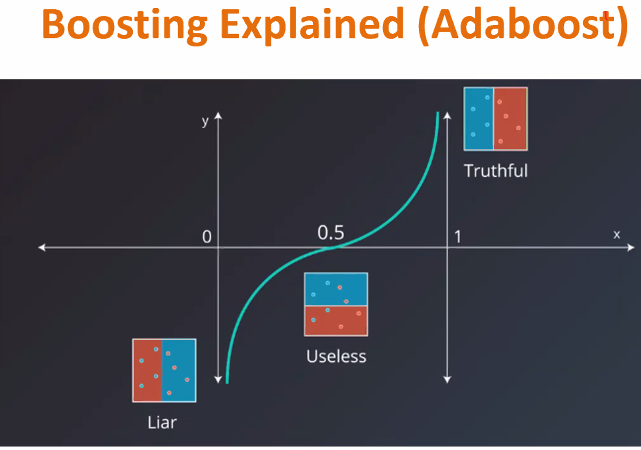
Worst is half half. We cannot derive any information from this .We have no idea.

Always tells the truth best model.

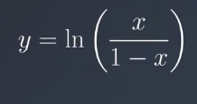
Always lies we can inverse it in ML.

We give the weightage to always truth we give importance in negative terms to always lies and we have to give zero importance to half half.

Always truth=100%=1

Always lies=0=0

Truthful and liar assigned weight.



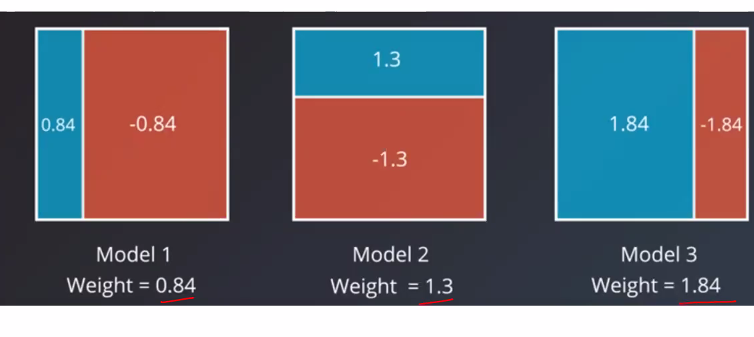
X=accuracy y=weight..

Ln1=0 so mid point useless is 0 weight.

Positive has highly positive weight and liar has negative weight.

Another formula is ln(correct/incorrect)

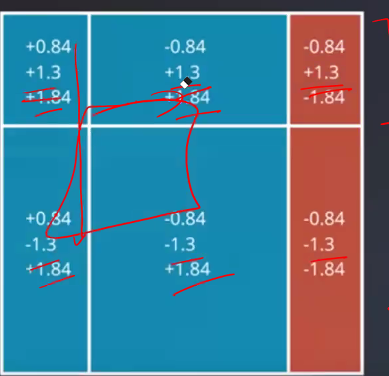
Last model has the highest weight of 1.84.

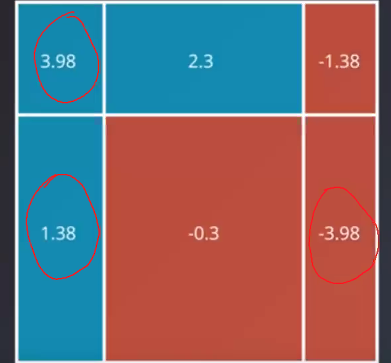


In the end we will merge.









The area with positive weight is assigned positive the area with negative weight is assigned negative.

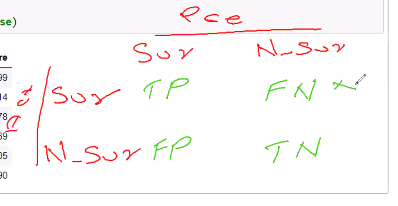
SAMME-R algorithm is responsible for the entire story before.

Gradient Boosting is the most powerful algorithm that is used.

Seed and random\_state do similar work.

We can optimize gradient boosting or adaboost using gridsearchCV

We have to decide which is more important for our model.



We made our model based on survival so it if we consider survived non survived then that is wrong.

So high recall so adaboost.

For complete boosting theory check recording.