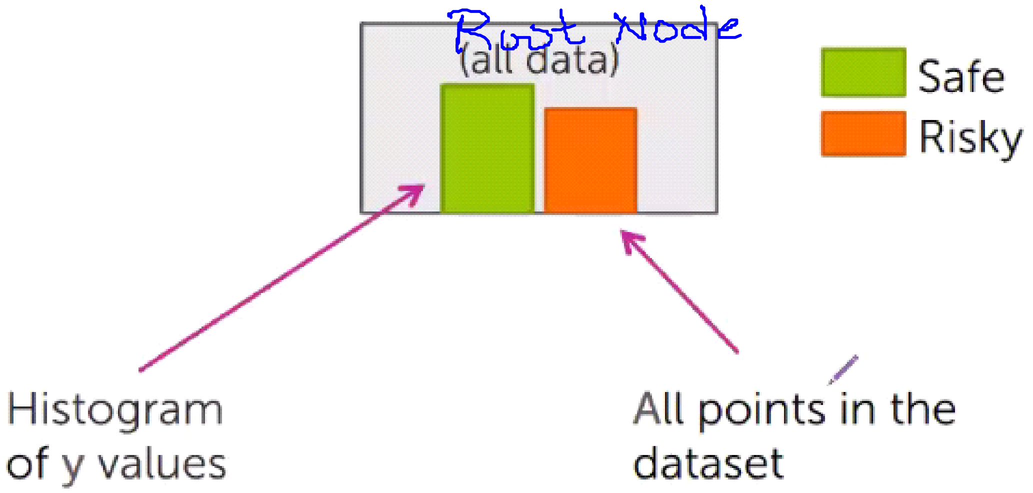
08 March 2018
18:57

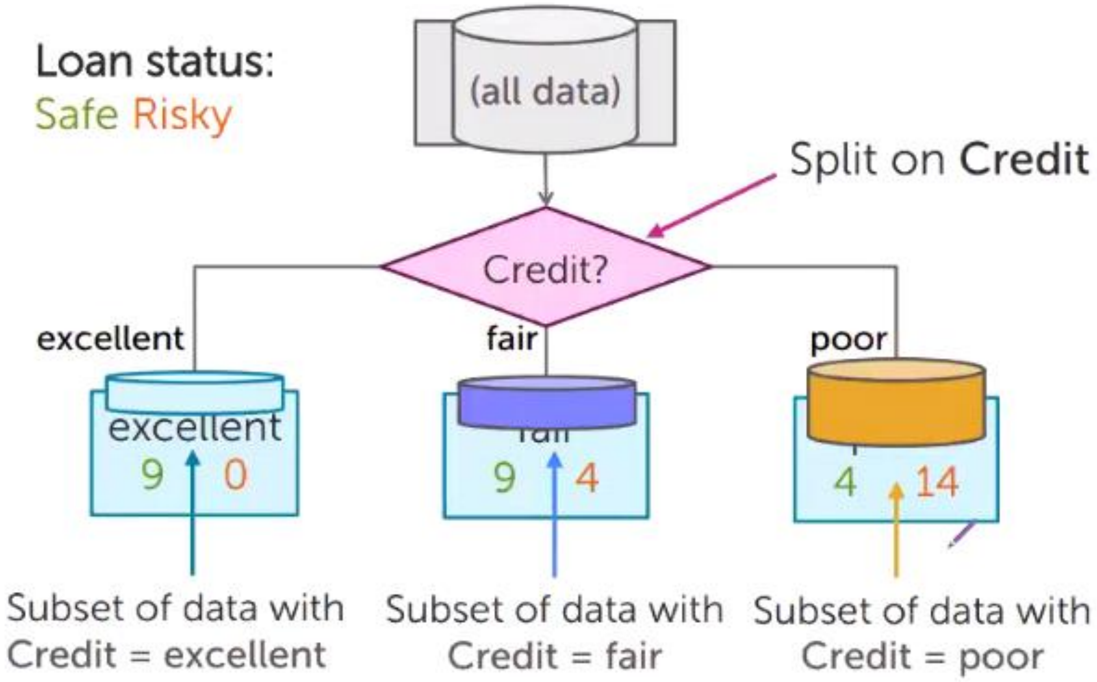
Let say's there are 40 observations in training data.
Green colour are Safe loans contains 22 and Orange colour contains Risky contains of 18

Step 1: Start with an empty tree



? How It is going to predict the out put of that Node is?
A: Majority of the data points belongs to that particular class. That particular class is going to be predicted.

Assume that it is derived from the existing data. We are trying to predict the values for each node.


Loan status:
Safe Risky
(all data)
Split on Credit
Credit?
excellent
fair
poor
excellent
9 0
fair
9 4
4 14
Subset of data with Credit = excellent
Subset of data with Credit = fair
Subset of data with Credit = poor

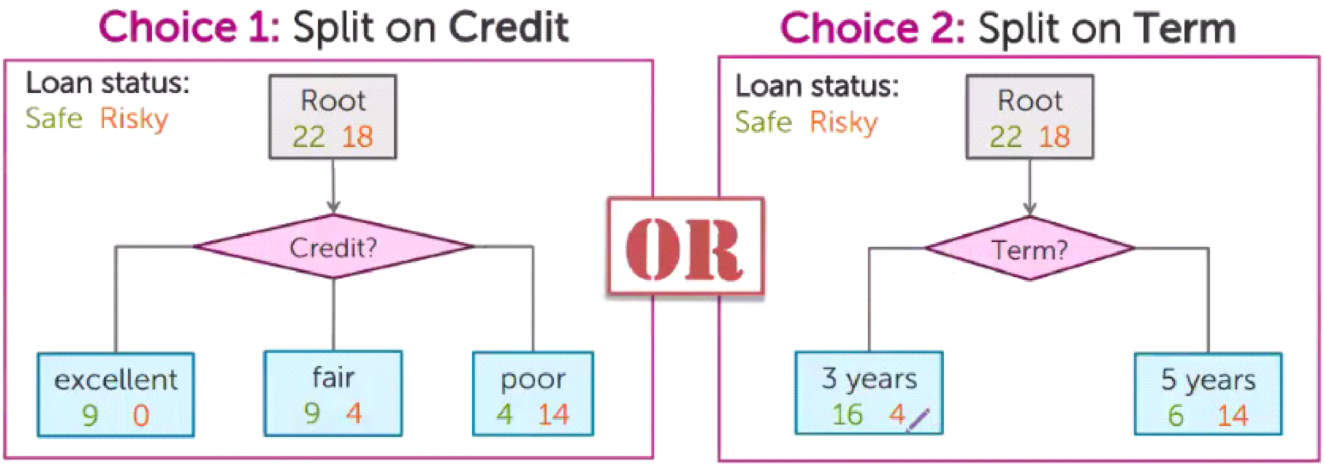
Predicted response:

Safe =9 Safe =13 Risky=18

So, though 0,4,4 are actually responses they are Risky, Risky, Safe
but I am predicted them as Safe, Safe, Risky Which is a Miss Classification Error.

Split on Credit : Miss Classification Error = $\frac{0 + 4 + 4}{40}$ = 0.2

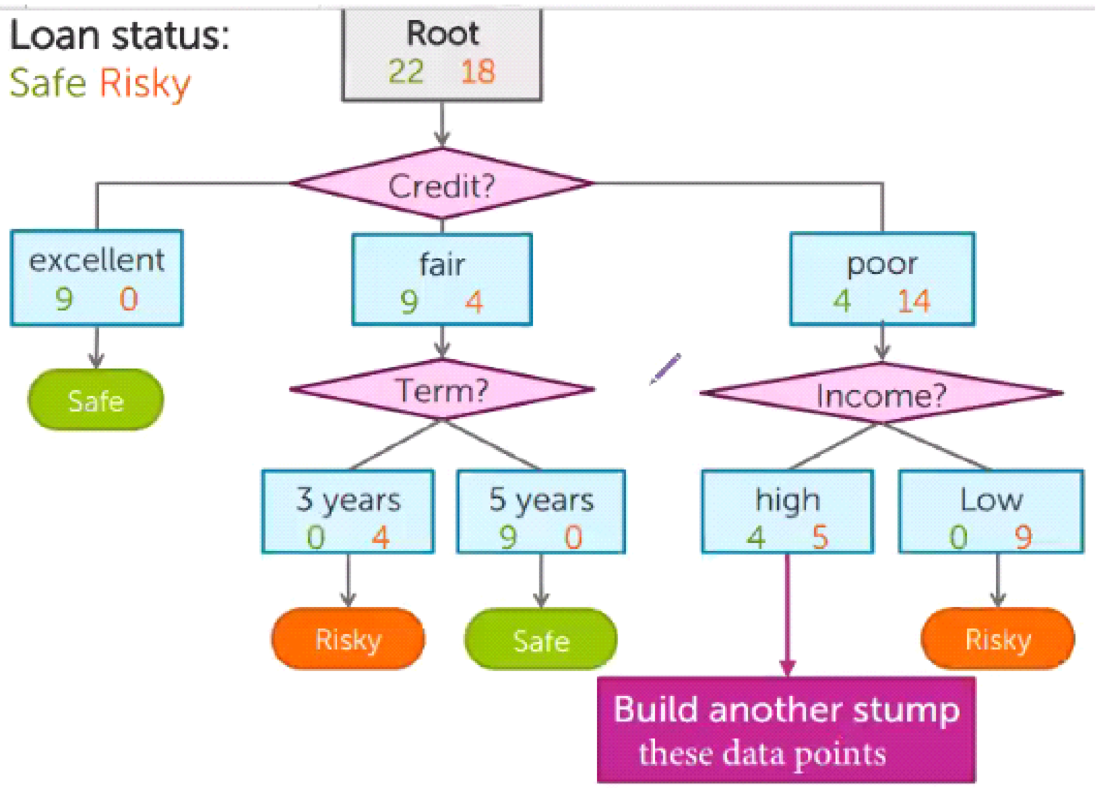
⭐ The same miss classification error is calculated for all the variables, so which ever is less it is finalized.

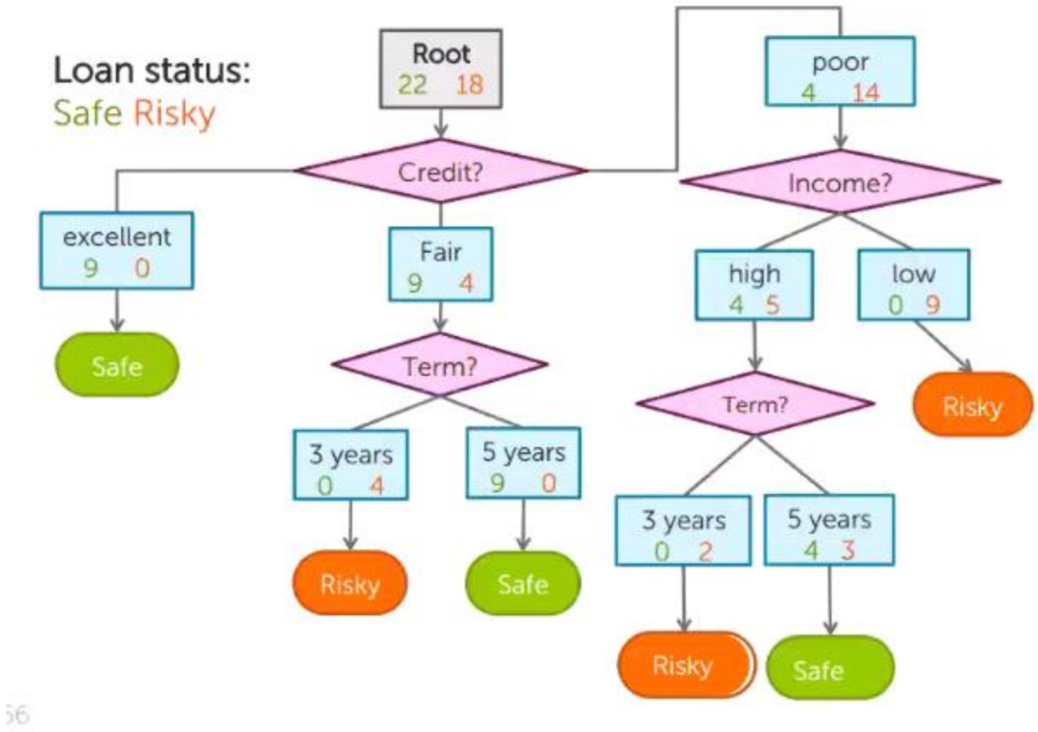


Split on Term : Miss Classification Error = $\frac{4\ \ +6}{40}$ = 0.25

You can conclude that 25% of the times your prediction may wrong.

Now, you can calculate the miss classification error for all the independent variables and decide which is going to be your Root Node.





Note:

- ⭐ Income variable contains of two levels and out of them high is contains some miss classification value whereas Low does not contains. So Low is called "Pure Node"
- ⭐ **Entropy** is also a method to measure miss classification error.
- ⭐ **GINI INDEX ,Chi-square, Reduction of variance** are also a measure to select the variable while splitting the variables.
- ⭐ All the measures gives the almost the same kind of results.