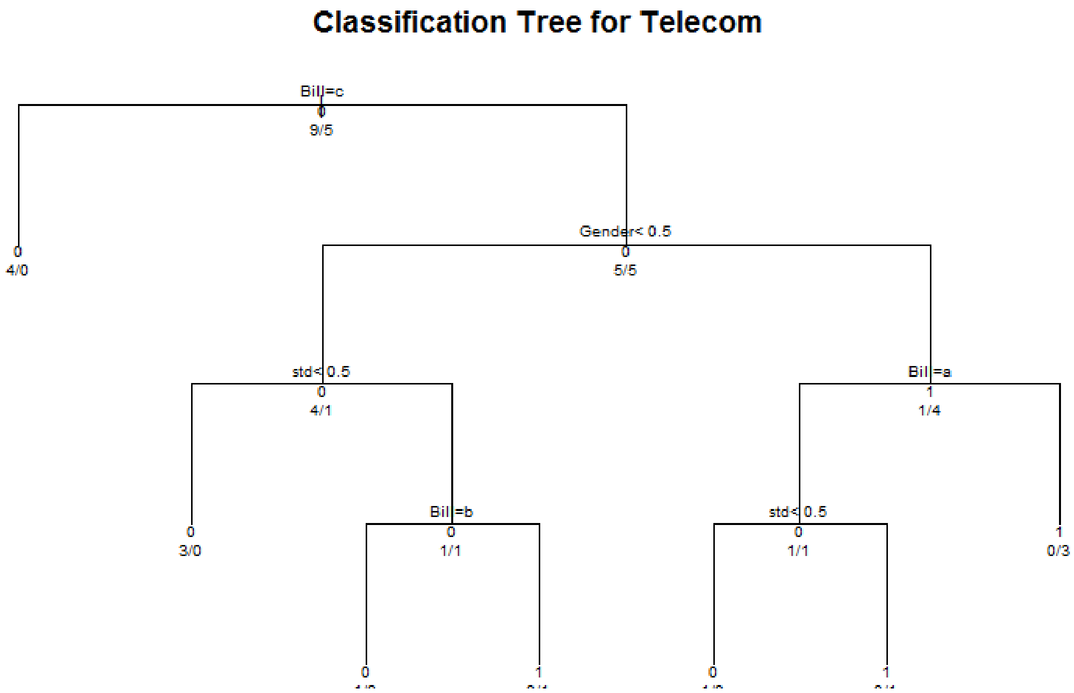
12 July 2018
21:54

Decision Tree in R - A Telecom Case Study



So we have got the decision tree, now let's see how to interpret the same and also understand how R or any other software draw decision tree, using Entropy and Information gain base algorithm.

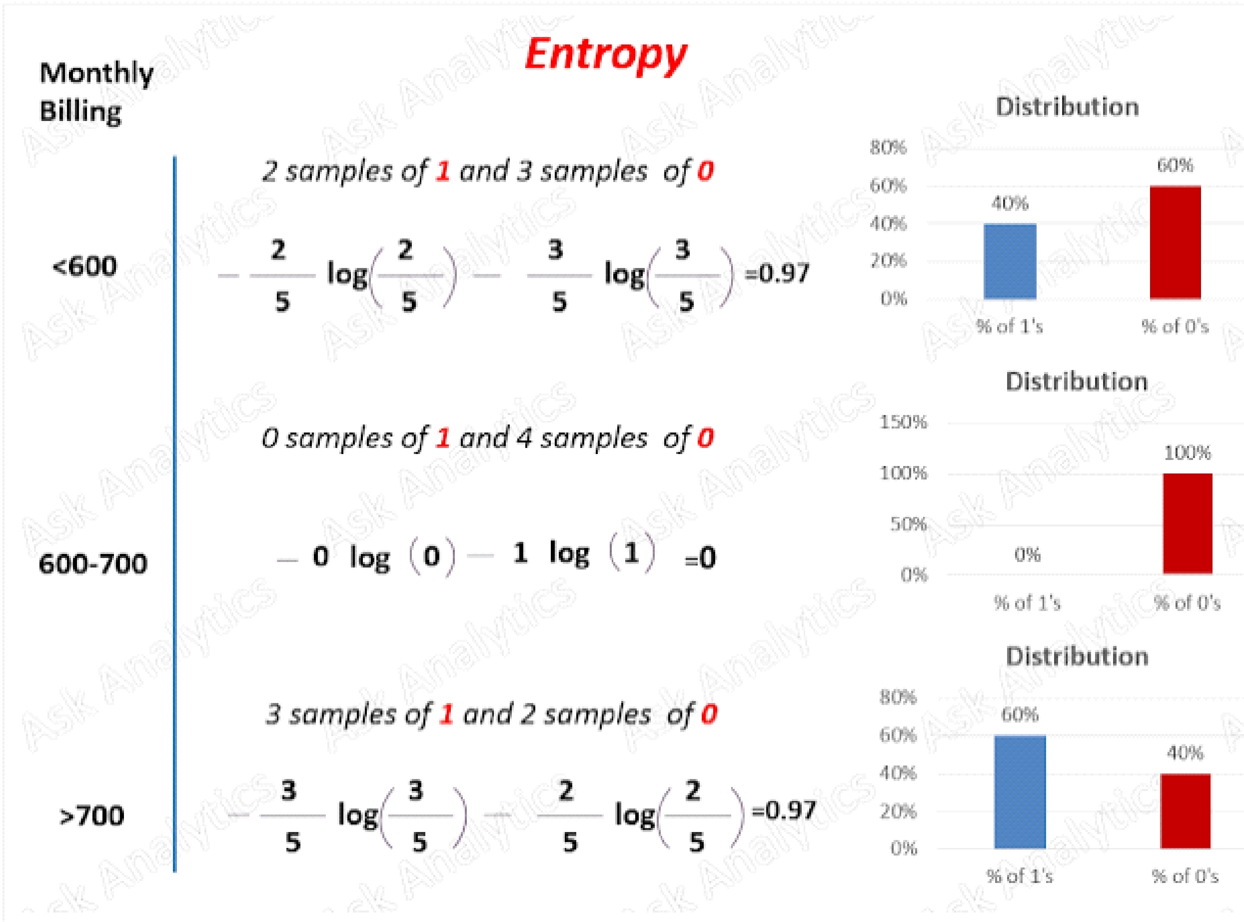
Our data looks like >>

There are four variables given in the data:

Monthly Billing : monthly bill of each individual
Gender : 1- Male , 0-female
Std : 1- taken std facility, 0 - has not taken std facility
Leave service : 1 - Customer has moved to other telecom operator, 0- continuing services with same operator

It first calculates the entropy of each variable for every bucket :

Entropy of Monthly Billing variable:



Entropy of Gender variable :

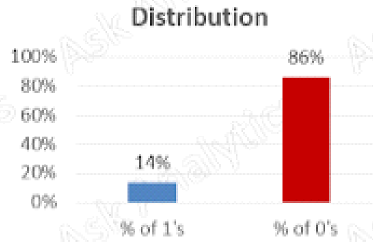
Gender

Entropy

0(Female)

*1 samples of **1** and 6 samples of **0***

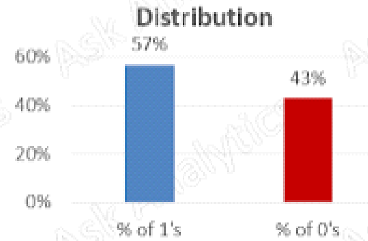
$$-\frac{1}{7}\log\left(\frac{1}{7}\right)-\frac{6}{7}\log\left(\frac{6}{7}\right)=0.5916$$



1(Male)

*4 samples of **1** and 3 samples of **0***

$$-\frac{4}{7}\log\left(\frac{4}{7}\right)-\frac{3}{7}\log\left(\frac{3}{7}\right)=0.9852$$



Entropy of Std variable:

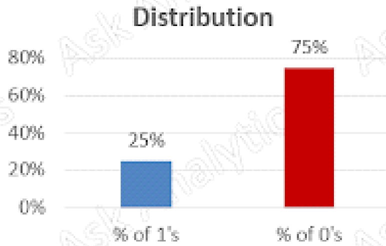
Std

Entropy

0(Do not use STD)

*2 samples of **1** and 6 samples of **0***

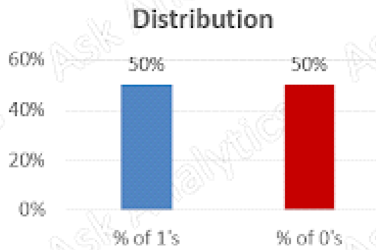
$$-\frac{2}{8}\log\left(\frac{2}{8}\right)-\frac{6}{8}\log\left(\frac{6}{8}\right)=0.8112$$



1(STD Users)

*3 samples of **1** and 3 samples of **0***

$$-\frac{3}{6}\log\left(\frac{3}{6}\right)-\frac{3}{6}\log\left(\frac{3}{6}\right)=1$$



Average Entropy

Variable(Monthly Billing) :=.357*.97+.285*0+. 357*.97 =.6925

Variable(Gender) :=.50*.5916+.50*.9852=.7884

Variable(Std) :=.57*.8112+.43*1=.892384

Entropy(Sample) :

36% samples of 1 and 64% samples 0

E(S)=-(.36)*log(.36)-(.64)*log(.64) = .94268

Then it calculates the information gain:

Information Gain : Entropy(Sample) – Average Entropy(Variable)

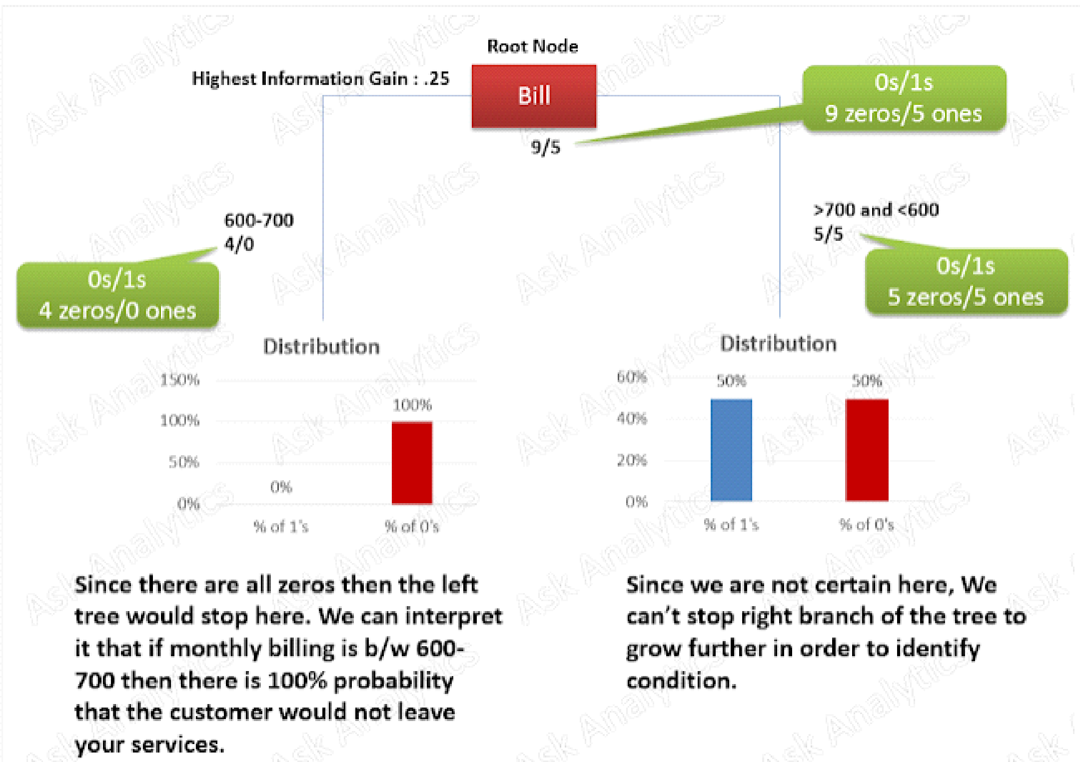
Information Gain(Monthly Billing)= .94268 - .6925 = .2501

Information Gain(Gender)= .94268 - .7884 = .154283

Information Gain(Std)= .94268 - .892384 = .0502

Since the monthly billing has maximum information gain value, it simply means that this variable has maximum ability to reduce the uncertainty and has best prediction ability.

So, monthly billing would be the root variable in decision tree.



Now we have to analyse only observations in which monthly billing is either >700 or <600.

We need to again calculate the information gain to further decide tree node.

Customer	Bill	Gender	Std Calls	Leave Service
1	>700	1	0	1
2	>700	1	1	1
[redacted]				
4	<600	1	0	0
5	<600	0	0	0
6	<600	0	1	1
[redacted]				
8	>700	1	0	1
9	>700	0	0	0
10	<600	0	0	0
11	>700	0	1	0
[redacted]				
14	<600	1	1	1

Entropy

Monthly Billing

	0	1
<600	3	2
>700	2	3

=-(3/5)*log(3/5)-(2/5)*log(2/5)=.97

=-(3/5)*log(3/5)-(2/5)*log(2/5)=.97

Average Entropy

=.5*.97+.5*.97 =.97

Gender

	0	1
0(Female)	4	1
1(Male)	1	4

=-(4/5)*log(4/5)-(1/5)*log(1/5)=.7219

=-(1/5)*log(1/5)-(4/5)*log(4/5)=.7219

Average Entropy

=.5*.7219+.5*.7219 =.7219

Std

	0	1
0(Female)	4	2
1(Male)	1	3

=-(4/6)*log(4/6)-(2/6)*log(2/6)=.9182

=-(1/4)*log(1/4)-(3/4)*log(3/4)=.8112

Average Entropy

=.6*.9182+.4*.8112 =.875

Entropy(Sample) :
50% samples of 1 and 50% samples 0

E(S)=-(.5)*log(.5)-(.5)*log(.5) = 1

Information Gain : Entropy(Sample) – Average Entropy(Variable)

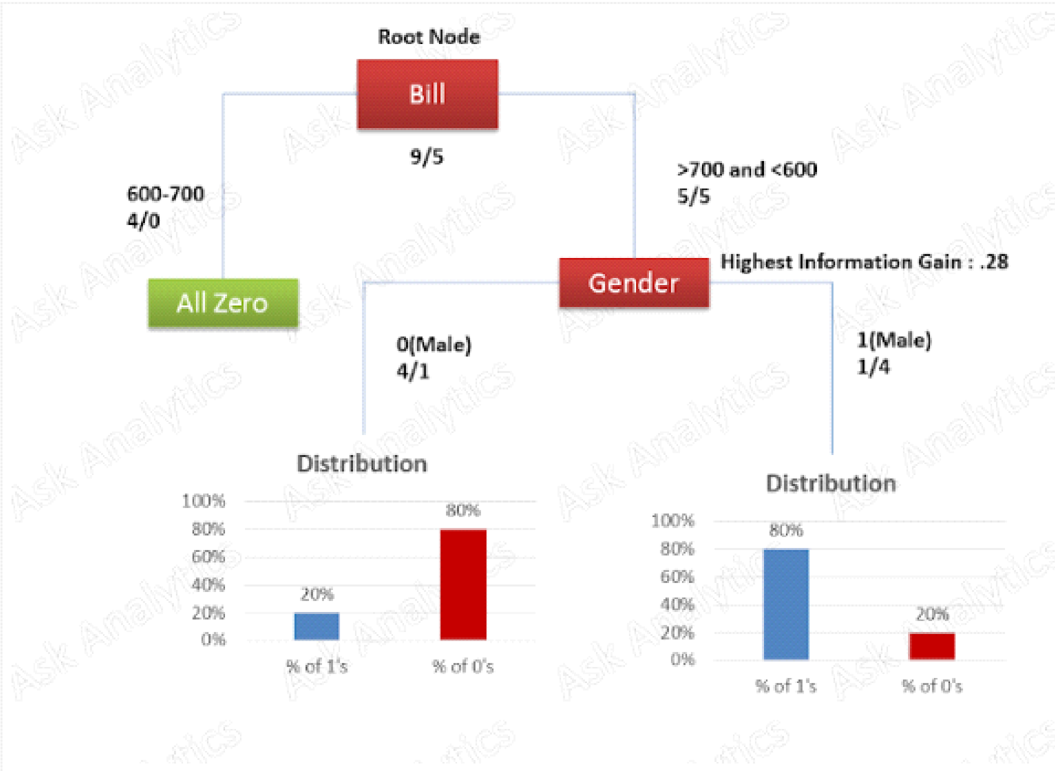
Information Gain(Monthly Billing)= 1 - .97 = .03

Information Gain(Gender)= 1 - .7219 = .28

Information Gain(Std)= 1 - .875 = .13

This time Gender variable has the maximum information gain therefore gender variable would better split the tree node. Hence the tree would be

This time Gender variable has the maximum information gain therefore gender variable would better split the tree node. Hence the tree would be



We will continue this process at each node to reach to the best separation of 1 and 0.

The final tree after this process would be