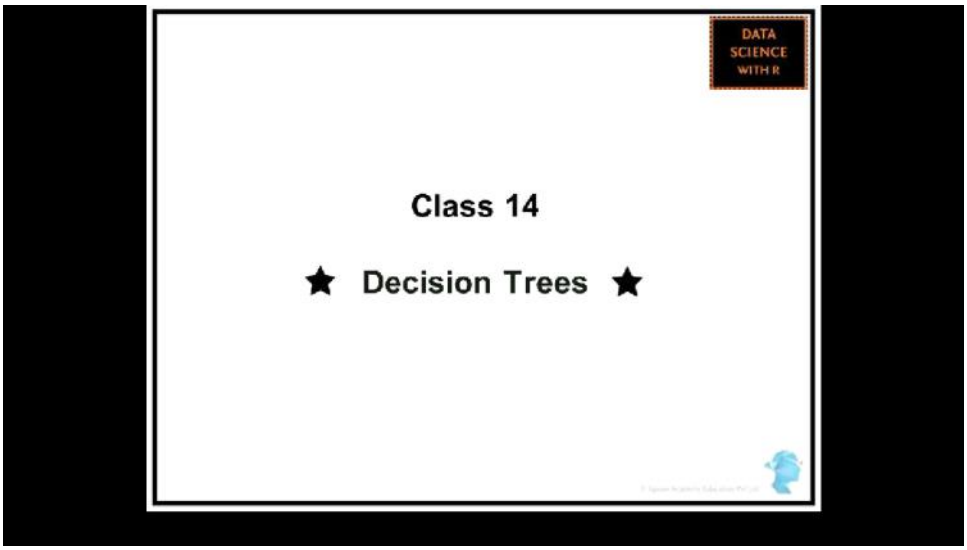
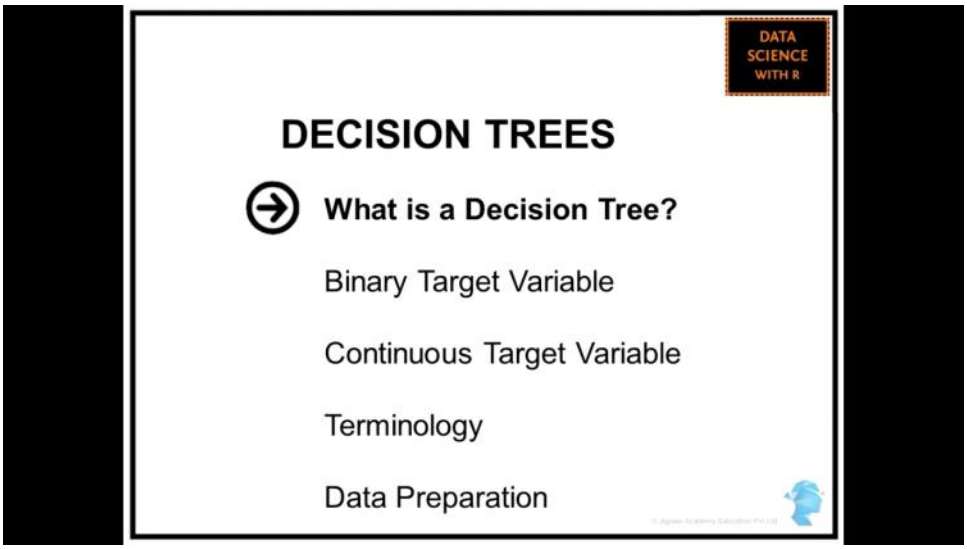
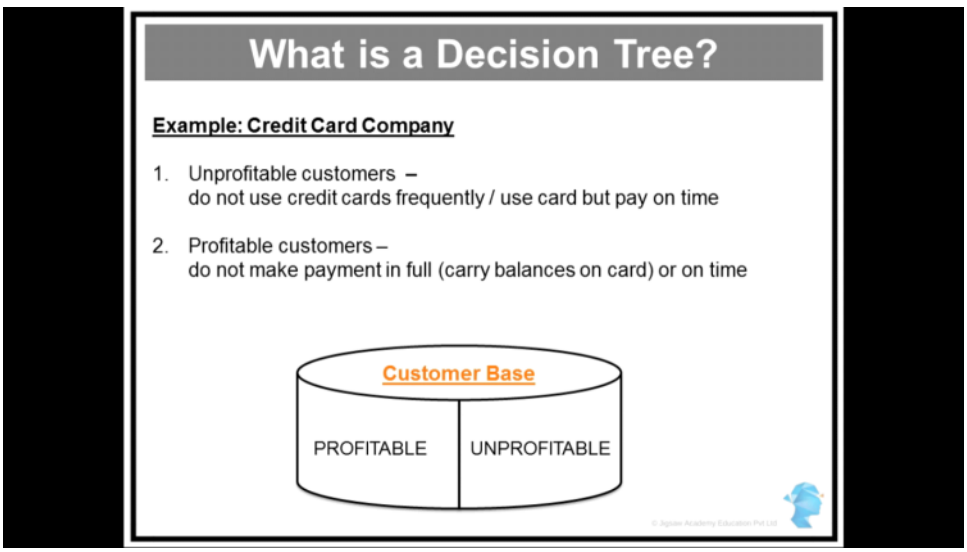
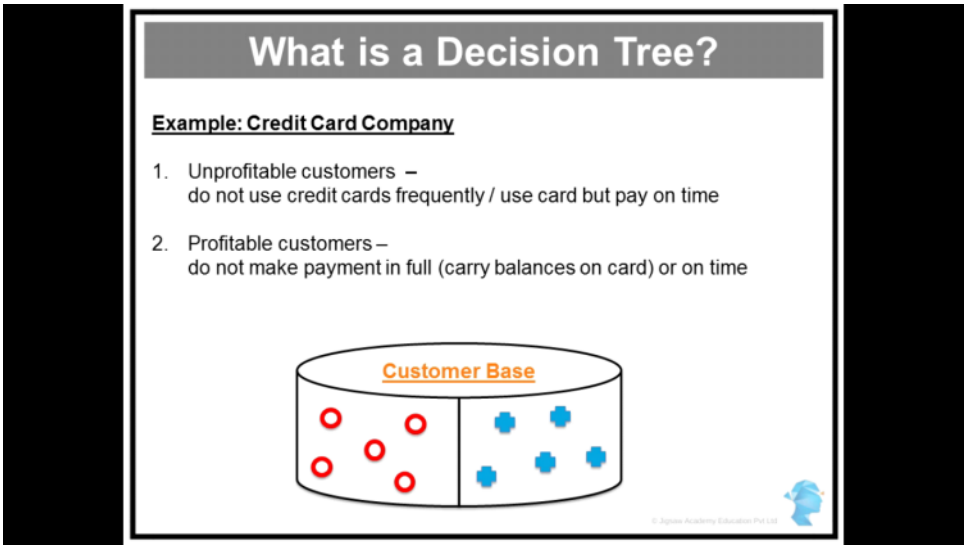
Decision Tree

Wednesday, November 30, 2016 5:54 AM

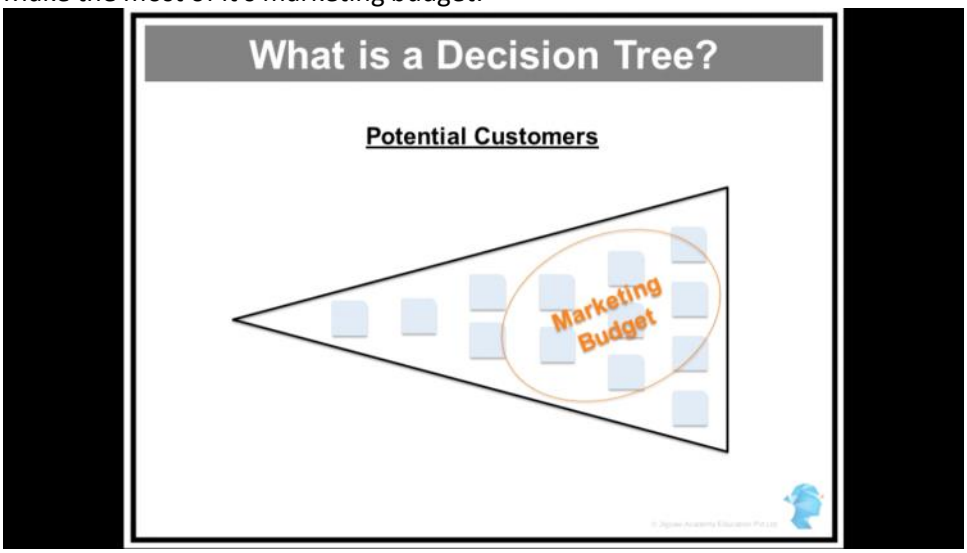


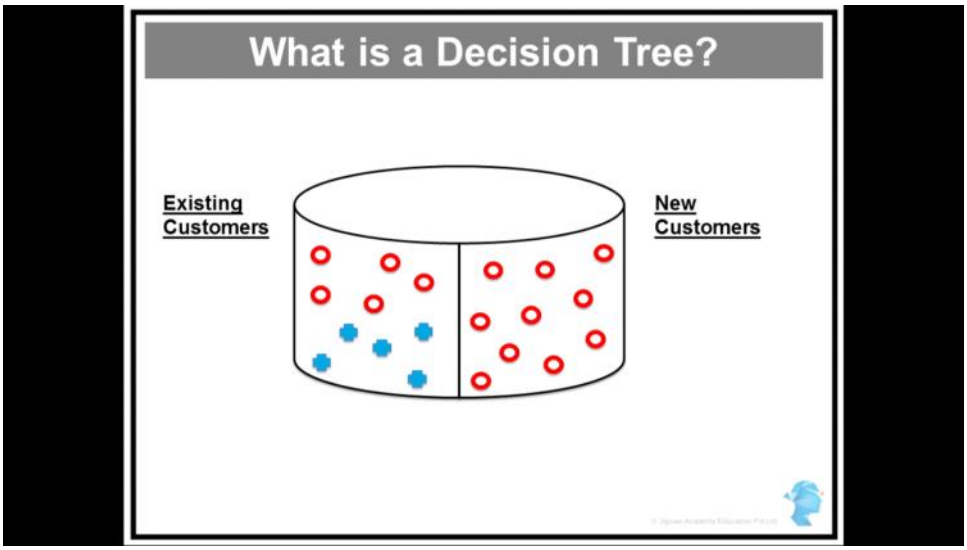




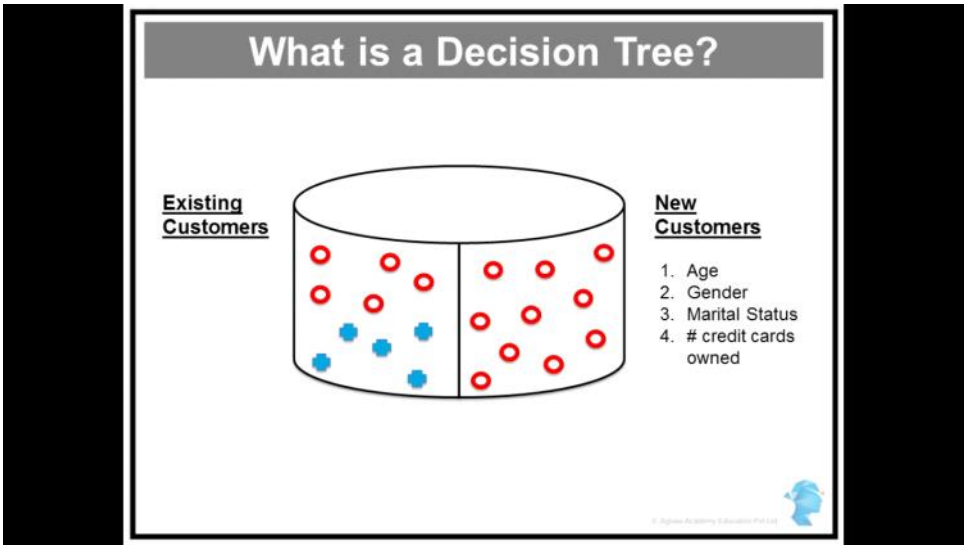


Make the most of it's marketing budget:

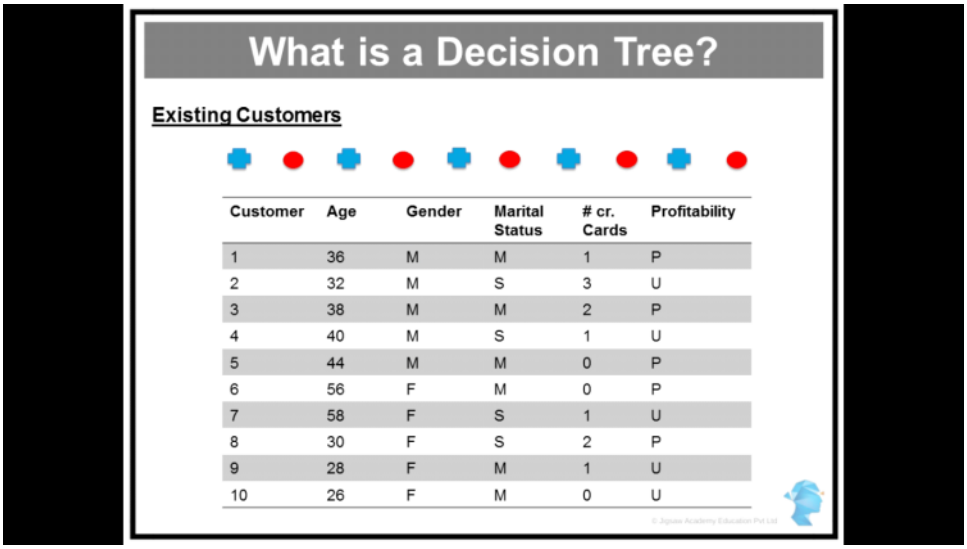




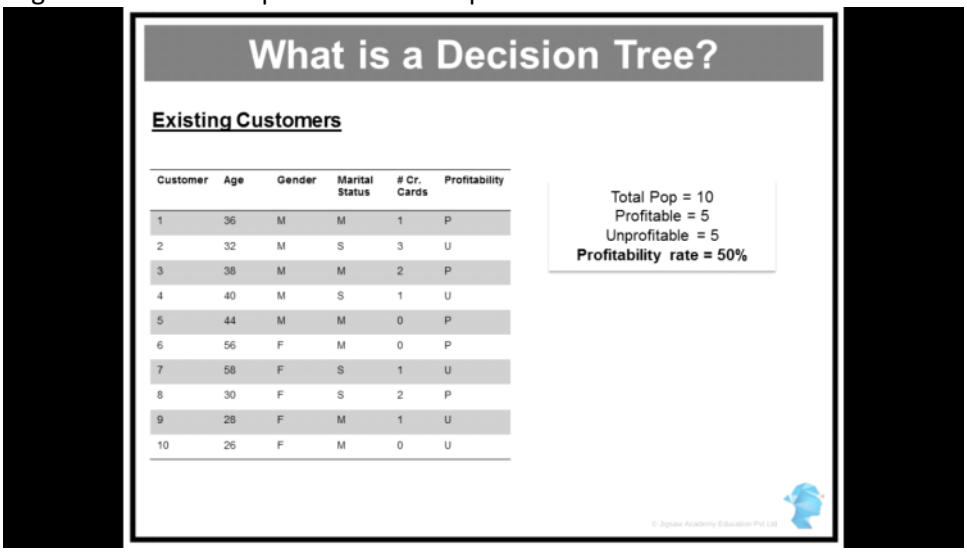
Focus of attracting profitable customers.

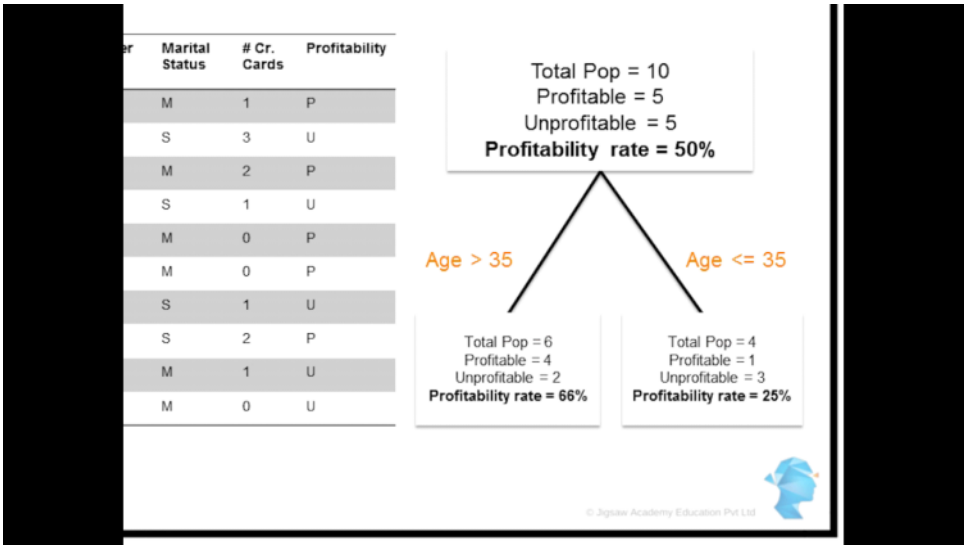


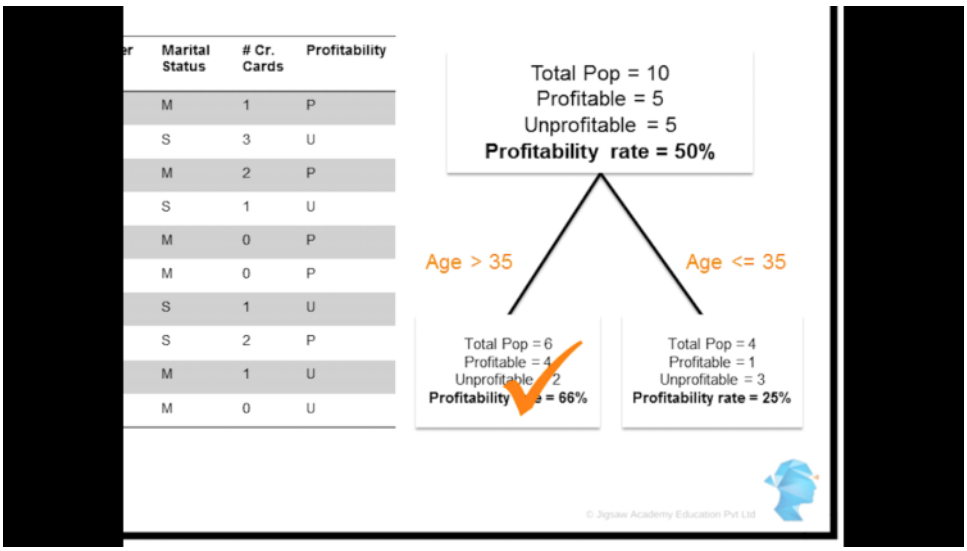
How can we use these variables be used to predict the profitablity of these customers.

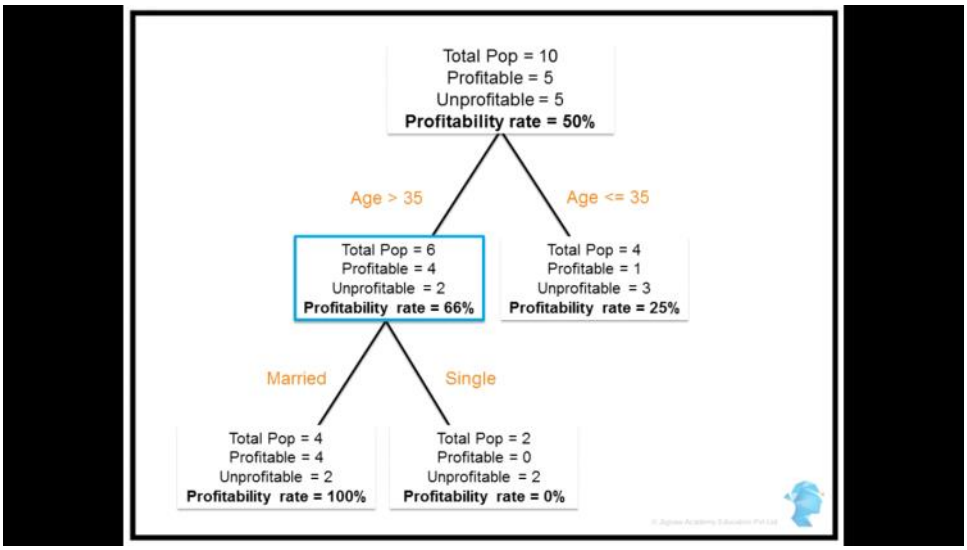


Tag the customers as profitable and unprofitable.




Marital Status
Cr. Cards
Profitability
M 1 P
S 3 U
M 2 P
S 1 U
M 0 P
M 0 P
S 1 U
S 2 P
M 1 U
M 0 U
Total Pop = 10
Profitable = 5
Unprofitable = 5
Profitability rate = 50%
Age > 35
Age <= 35
Total Pop = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%
Total Pop = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%


Marital Status
Cr. Cards
Profitability
M 1 P
S 3 U
M 2 P
S 1 U
M 0 P
M 0 P
S 1 U
S 2 P
M 1 U
M 0 U
Total Pop = 10
Profitable = 5
Unprofitable = 5
Profitability rate = 50%
Age > 35
Age <= 35
Total Pop = 6
Total Pop = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%


Total Pop = 10
Profitable = 5
Unprofitable = 5
Profitability rate = 50%
Age > 35
Age <= 35
Total Pop = 6
Profitable = 4
Unprofitable = 2
Profitability rate = 66%
Total Pop = 4
Profitable = 1
Unprofitable = 3
Profitability rate = 25%
Married
Single
Total Pop = 4
Profitable = 4
Unprofitable = 2
Profitability rate = 100%
Total Pop = 2
Profitable = 0
Unprofitable = 2
Profitability rate = 0%

What is a Decision Tree?

4 variables, 10 records

1. Age
2. Gender
3. Marital Status
4. # Credit Cards

- Why Age?
- Why split at 35?
- What if there are hundreds of variables and millions of records?

Decision Trees

- **Gini**
- **Entropy**
- **Chi Square**
- **Reduction of Variance**

DATA SCIENCE WITH R

DECISION TREES

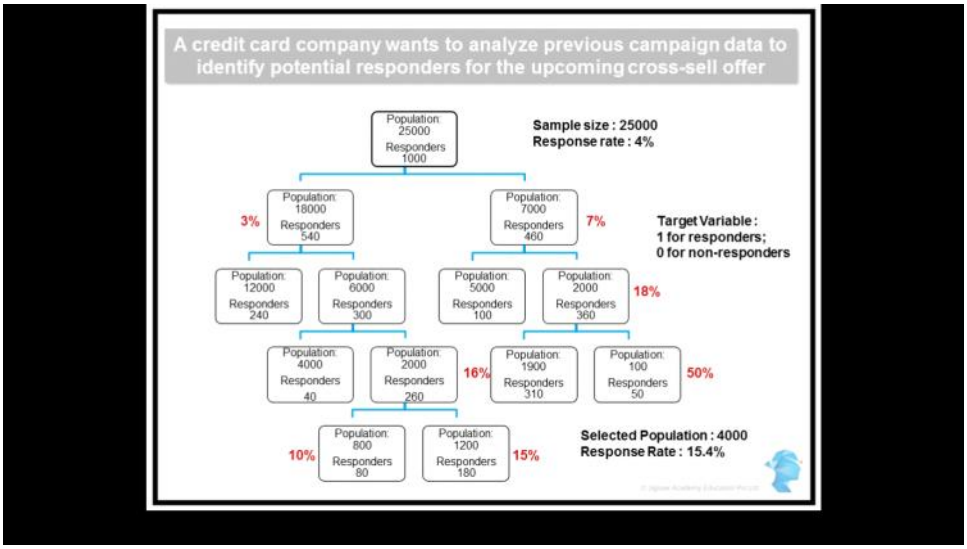
What is a Decision Tree?

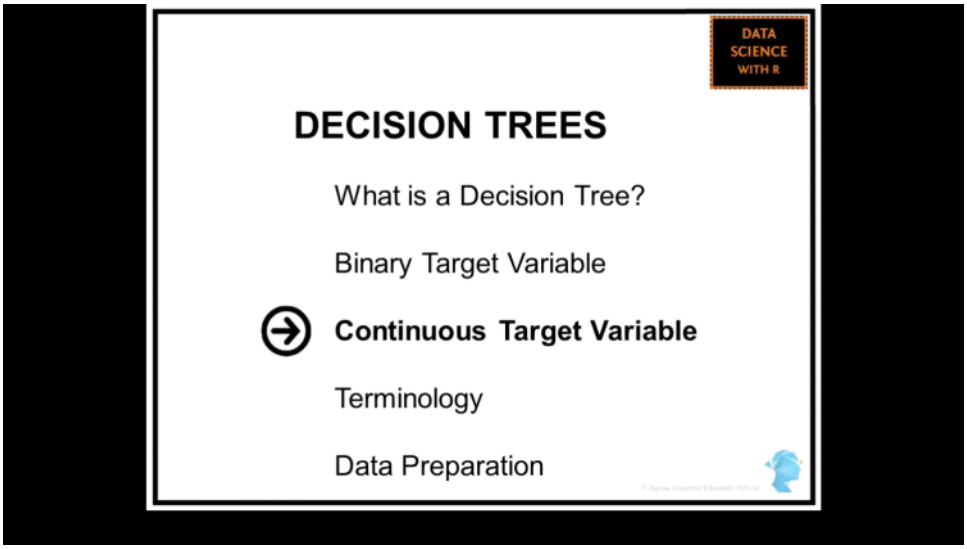
Binary Target Variable

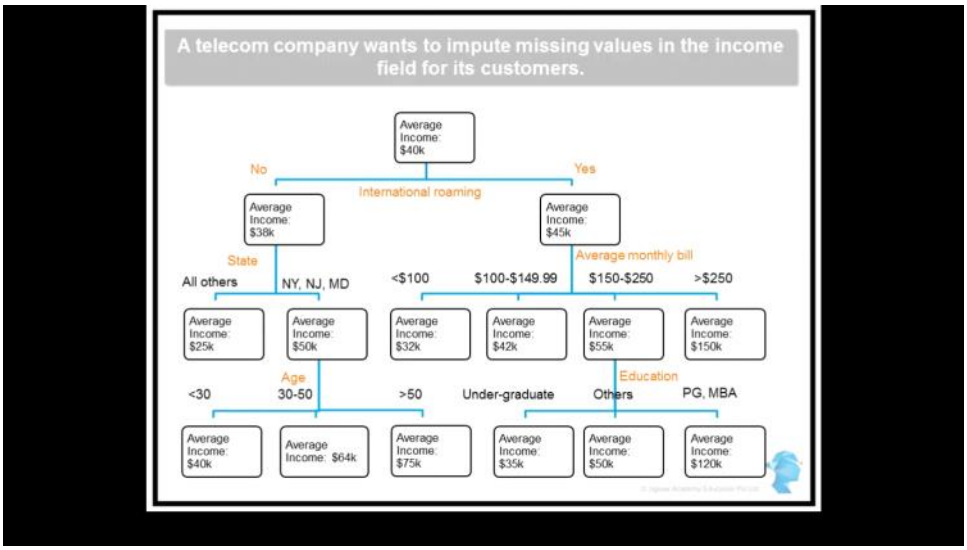
Continuous Target Variable

Terminology

Data Preparation


A credit card company wants to analyze previous campaign data to identify potential responders for the upcoming cross-sell offer
Population: 25000 Responders 1000
Sample size : 25000
Response rate : 4%
Population: 18000 Responders 540
3%
Population: 7000 Responders 460
7%
Target Variable :
1 for responders;
0 for non-responders
Population: 12000 Responders 240
Population: 6000 Responders 300
Population: 5000 Responders 100
Population: 2000 Responders 360
18%
Population: 4000 Responders 40
Population: 2000 Responders 260
16%
Population: 1900 Responders 310
Population: 100 Responders 50
50%
10%
Population: 800 Responders 80
Population: 1200 Responders 180
15%
Selected Population : 4000
Response Rate : 15.4%

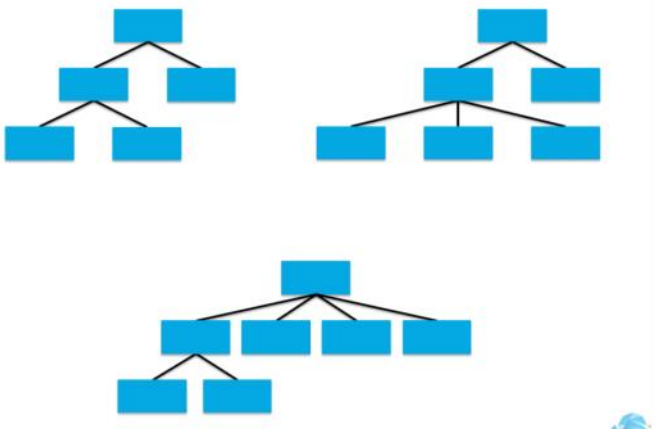

DATA SCIENCE WITH R
DECISION TREES
What is a Decision Tree?
Binary Target Variable
Continuous Target Variable
Terminology
Data Preparation


A telecom company wants to impute missing values in the income field for its customers.
Average Income: $40k
No
International roaming
Yes
Average Income: $38k
Average Income: $45k
State
All others
NY, NJ, MD
Average monthly bill
<$100
$100-$149.99
$150-$250
>$250
Average Income: $25k
Average Income: $50k
Average Income: $32k
Average Income: $42k
Average Income: $55k
Average Income: $150k
Age
<30
30-50
>50
Education
Under-graduate
Others
PG, MBA
Average Income: $40k
Average Income: $64k
Average Income: $75k
Average Income: $35k
Average Income: $50k
Average Income: $120k

Each split in the tree is chosen to decrease the variance in the values of the target variable within each child node

Continuous Target Variable

1. A decision tree is more suited to the prediction of discrete or categorical variables
2. To estimate a continuous variable, use a continuous function
3. Regression and neural nets are usually better for estimation of continuous variables



DATA SCIENCE WITH R

DECISION TREES

What is a Decision Tree?

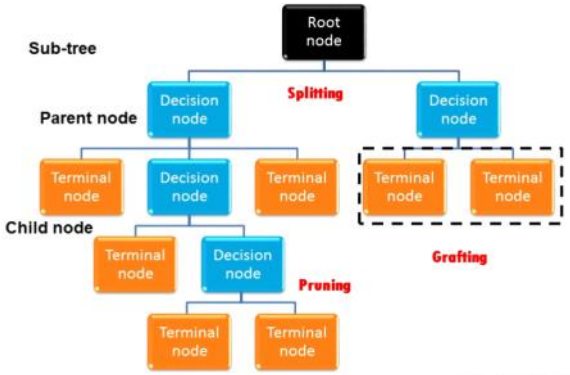
Binary Target Variable

Continuous Target Variable

Terminology

Data Preparation

Decision Tree Terminology



DATA SCIENCE WITH R

DECISION TREES

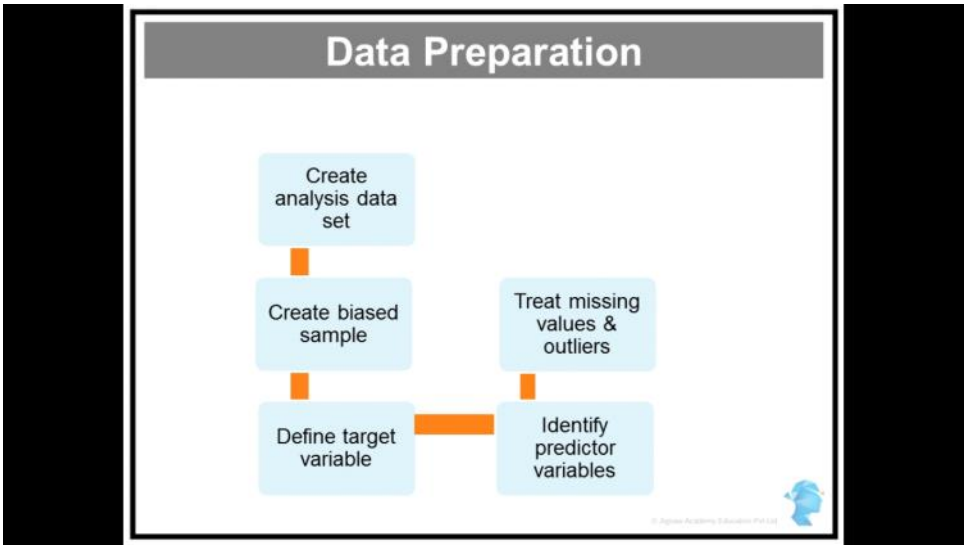
What is a Decision Tree?

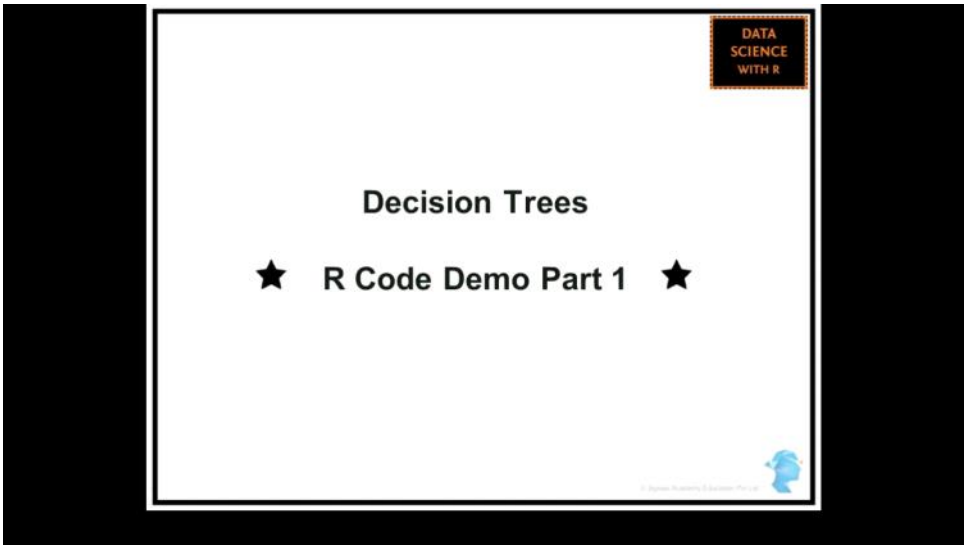
Binary Target Variable

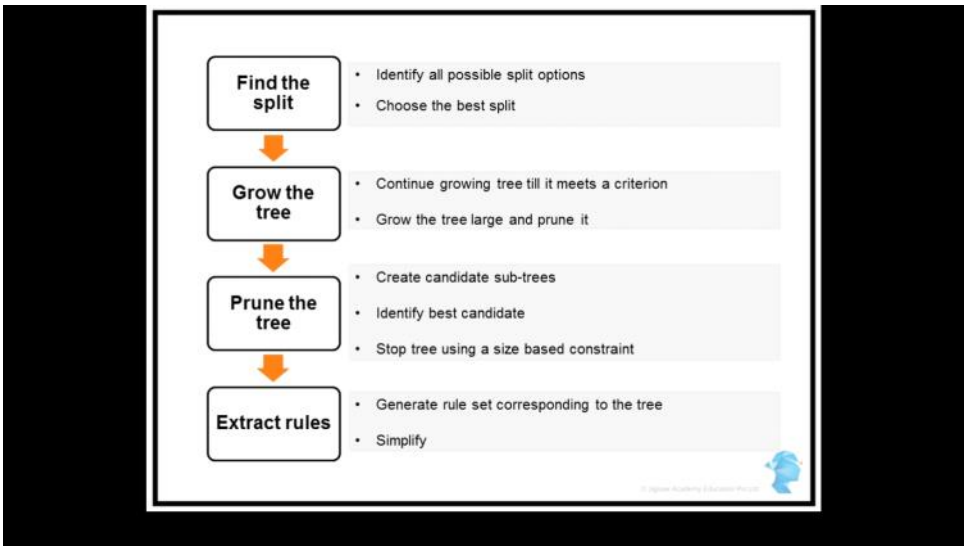
Continuous Target Variable

Terminology

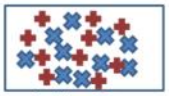
Data Preparation


Data Preparation
Create analysis data set
Create biased sample
Define target variable
Treat missing values & outliers
Identify predictor variables


DATA SCIENCE WITH R
Decision Trees
R Code Demo Part 1


Find the split
Identify all possible split options
Choose the best split
Grow the tree
Continue growing tree till it meets a criterion
Grow the tree large and prune it
Prune the tree
Create candidate sub-trees
Identify best candidate
Stop tree using a size based constraint
Extract rules
Generate rule set corresponding to the tree
Simplify

Finding the Split



Training Data

3 variables

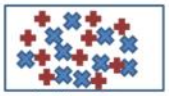
Male or Female

Low Income / Medium Income / High Income

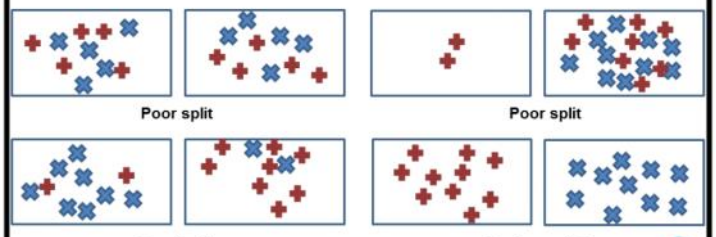
Karnataka or UP or Tamil Nadu. etc

2*3*25 = 150 options

Finding the Split



Training Data



Poor split

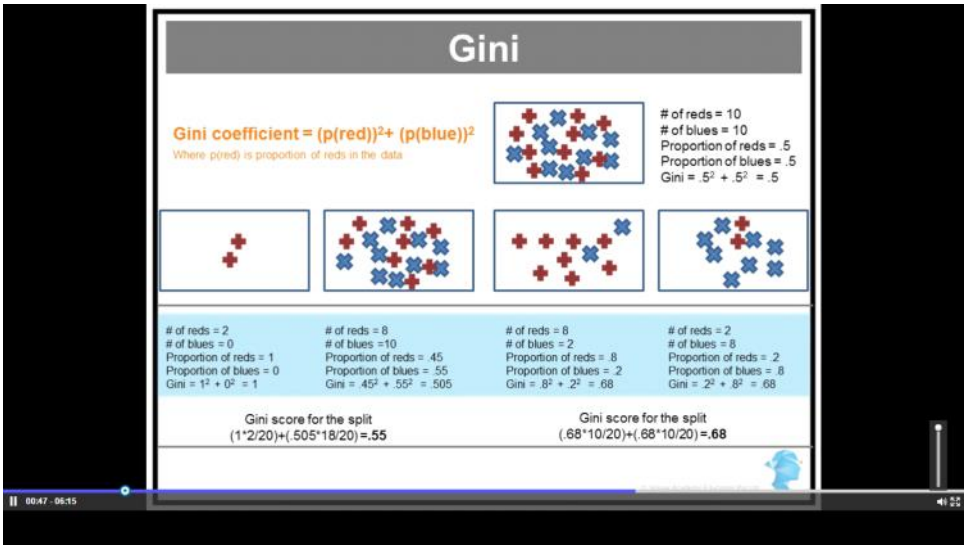
Poor split

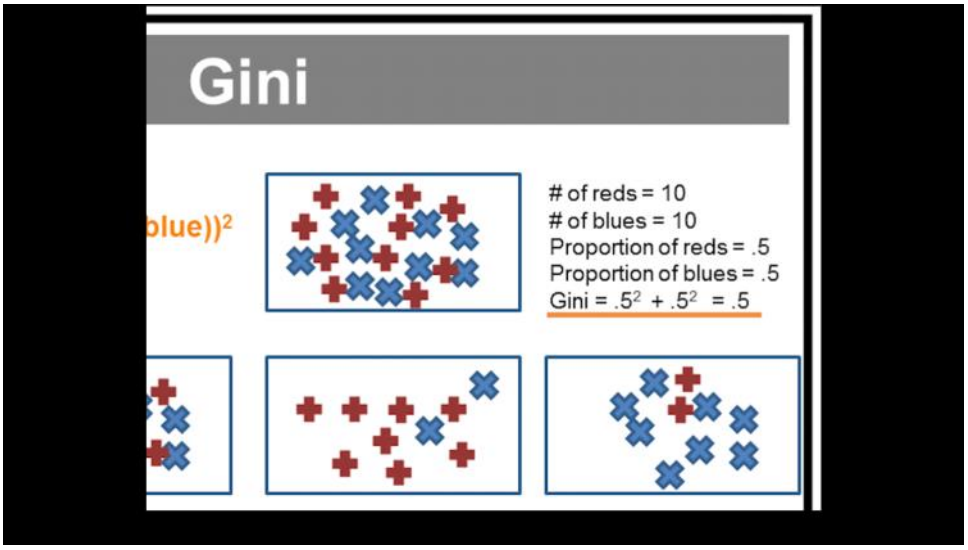
Good split

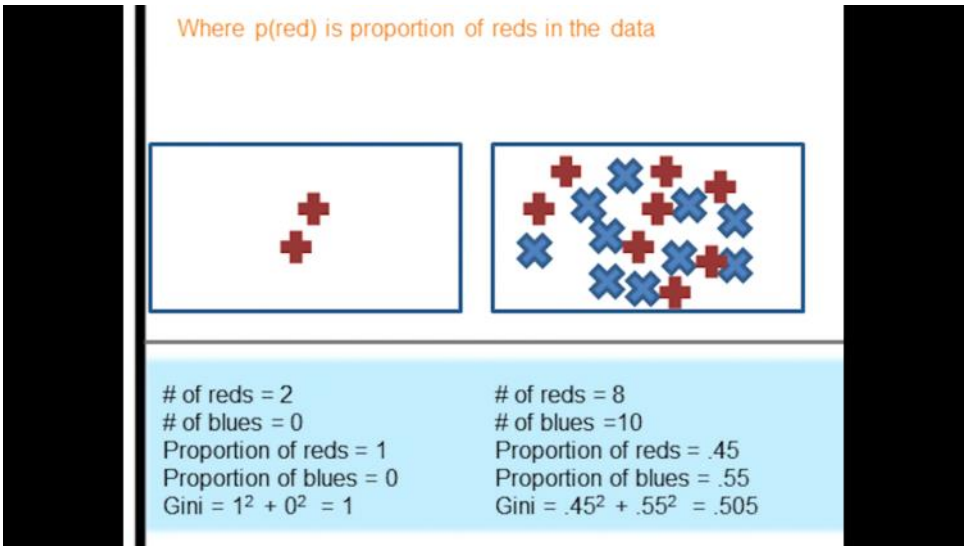
Perfect split?

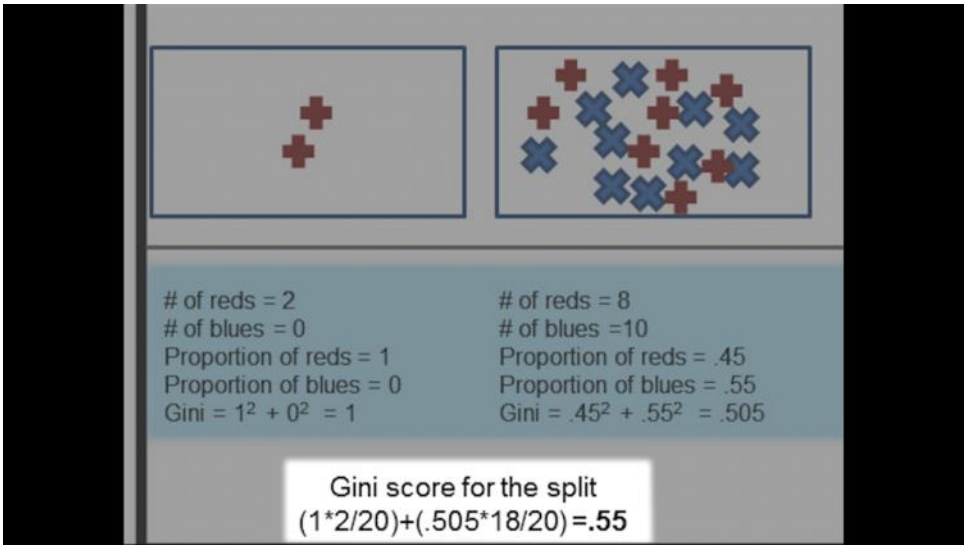
Effectiveness of Split

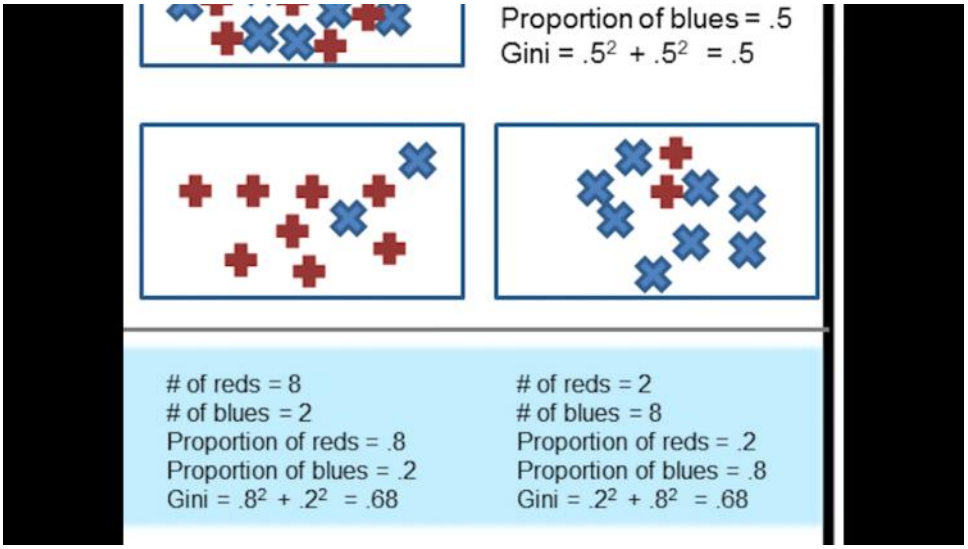
1. **Categorical target variable**
   - Gini
   - Chi-square
   - Information gain
2. **Continuous target variable**
   - Reduction in variance
3. **Other methods**
   - Information gain ratio
   - F test

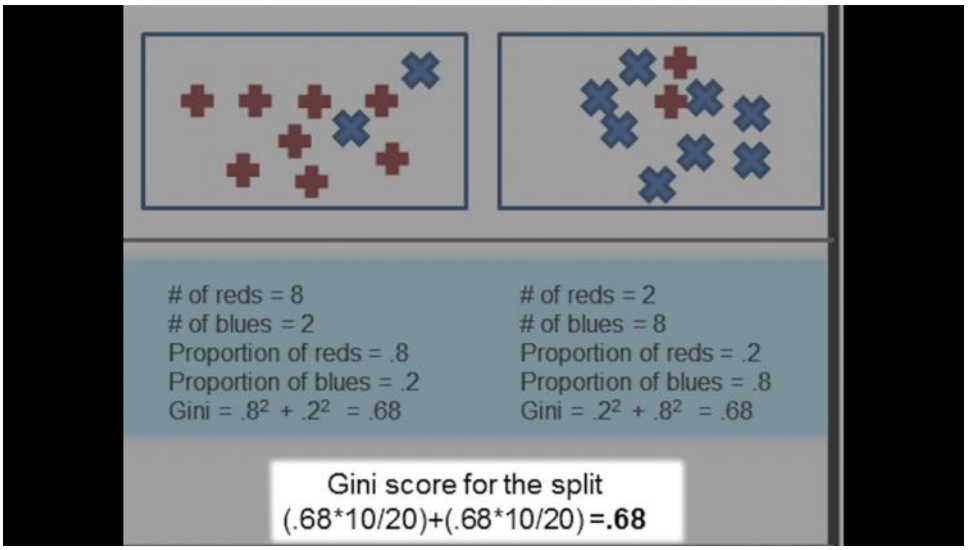

Gini
Gini coefficient = (p(red))²+ (p(blue))²
Where p(red) is proportion of reds in the data
# of reds = 10
# of blues = 10
Proportion of reds = .5
Proportion of blues = .5
Gini = .5² + .5² = .5
# of reds = 2
# of blues = 0
Proportion of reds = 1
Proportion of blues = 0
Gini = 1² + 0² = 1
# of reds = 8
# of blues =10
Proportion of reds = .45
Proportion of blues = .55
Gini = .45² + .55² = .505
# of reds = 8
# of blues = 2
Proportion of reds = .8
Proportion of blues = .2
Gini = .8² + .2² = .68
# of reds = 2
# of blues = 8
Proportion of reds = .2
Proportion of blues = .8
Gini = .2² + .8² = .68
Gini score for the split
(1*2/20)+(.505*18/20)=.55
Gini score for the split
(.68*10/20)+(.68*10/20)=.68

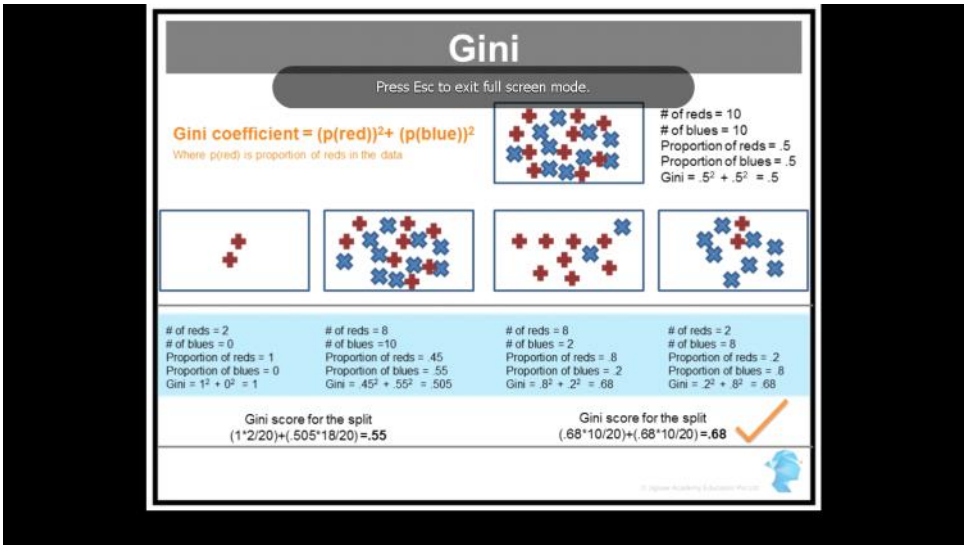

Gini
blue))²
# of reds = 10
# of blues = 10
Proportion of reds = .5
Proportion of blues = .5
Gini = .5² + .5² = .5

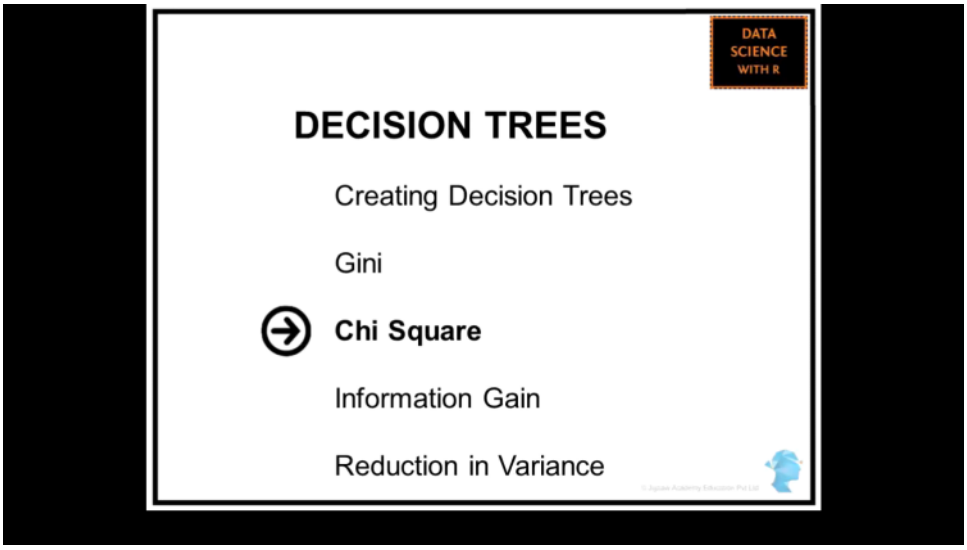

Where p(red) is proportion of reds in the data
# of reds = 2
# of blues = 0
Proportion of reds = 1
Proportion of blues = 0
Gini = 1² + 0² = 1
# of reds = 8
# of blues =10
Proportion of reds = .45
Proportion of blues = .55
Gini = .45² + .55² = .505

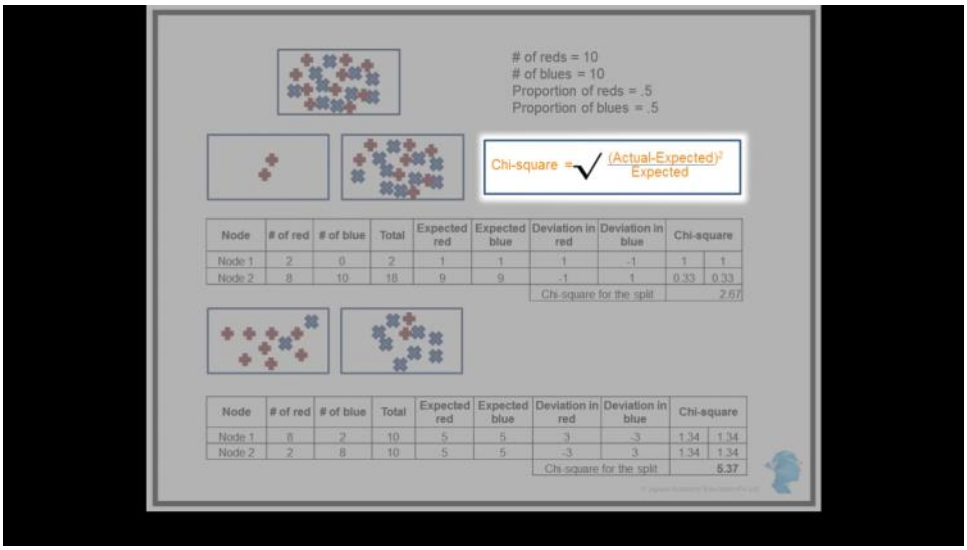

# of reds = 2
# of blues = 0
Proportion of reds = 1
Proportion of blues = 0
Gini = $1^2 + 0^2 = 1$
# of reds = 8
# of blues =10
Proportion of reds = .45
Proportion of blues = .55
Gini = $.45^2 + .55^2 = .505$
Gini score for the split
(1*2/20)+(.505*18/20)=.55


Proportion of blues = .5
Gini = $.5^2 + .5^2 = .5$
# of reds = 8
# of blues = 2
Proportion of reds = .8
Proportion of blues = .2
Gini = $.8^2 + .2^2 = .68$
# of reds = 2
# of blues = 8
Proportion of reds = .2
Proportion of blues = .8
Gini = $.2^2 + .8^2 = .68$

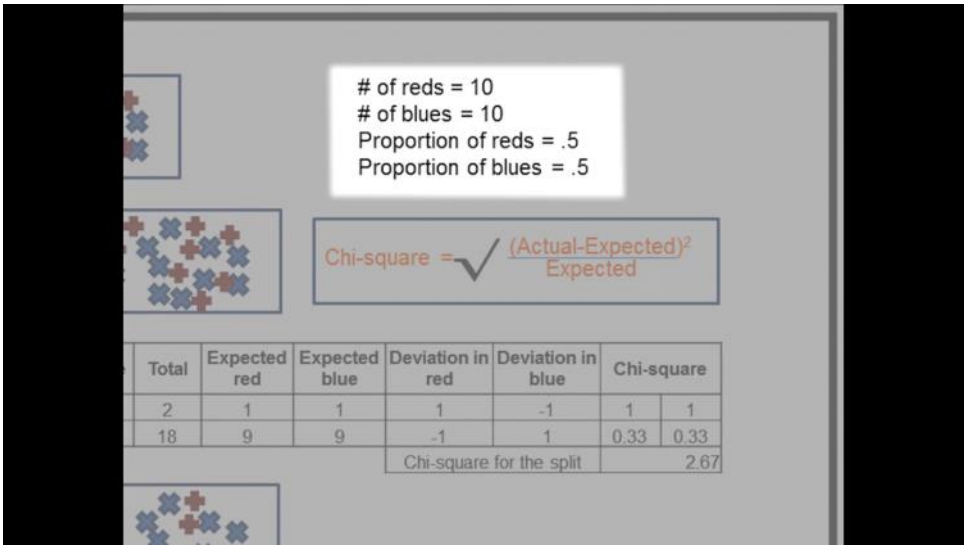

# of reds = 8
# of blues = 2
Proportion of reds = .8
Proportion of blues = .2
Gini = $.8^2 + .2^2 = .68$
# of reds = 2
# of blues = 8
Proportion of reds = .2
Proportion of blues = .8
Gini = $.2^2 + .8^2 = .68$
Gini score for the split
(.68*10/20)+(.68*10/20)=.68



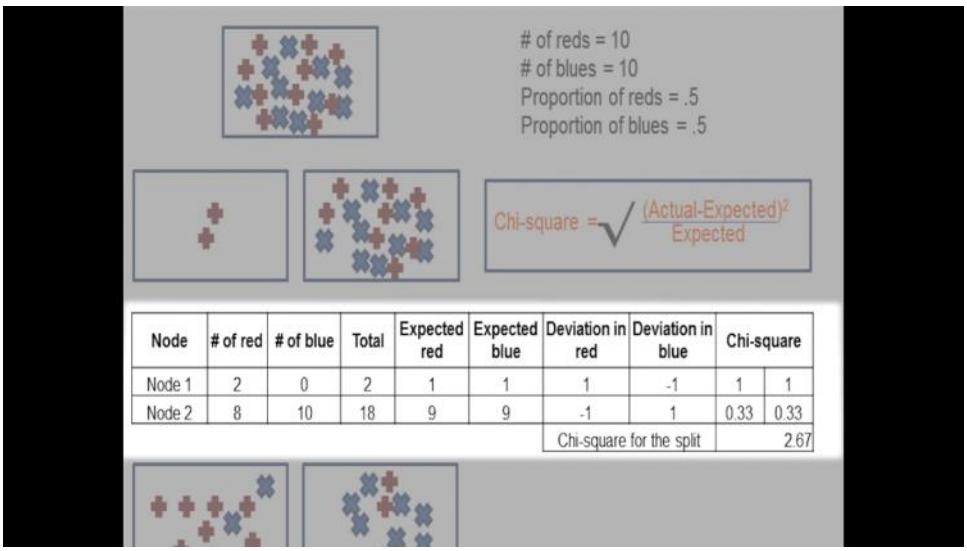




Higher values of chi square means that the variance is more significant and not by mere chance

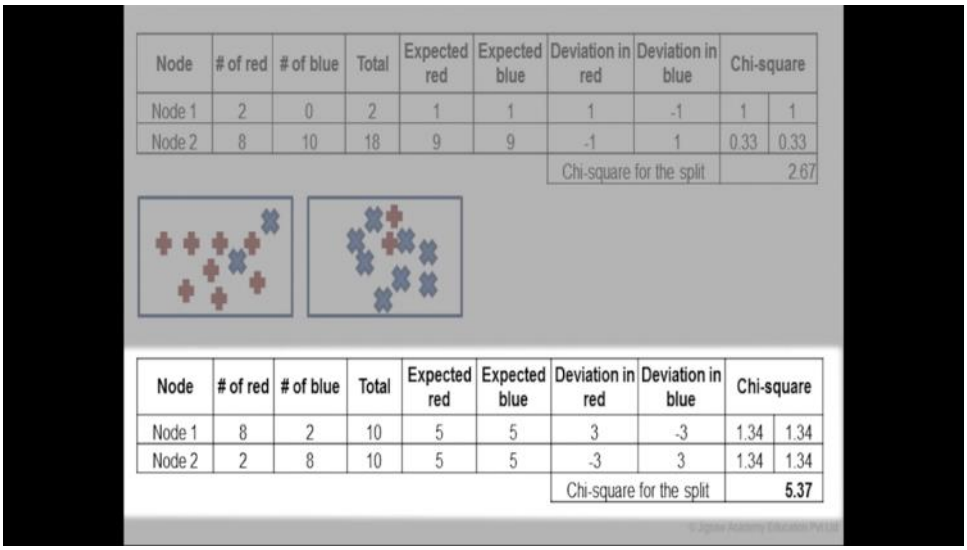


	Total	Expected red	Expected blue	Deviation in red	Deviation in blue	Chi-square	
	2	1	1	1	-1	1	1
	18	9	9	-1	1	0.33	0.33
				Chi-square for the split			2.67



Node	# of red	# of blue	Total	Expected red	Expected blue	Deviation in red	Deviation in blue	Chi-square	
Node 1	2	0	2	1	1	1	-1	1	1
Node 2	8	10	18	9	9	-1	1	0.33	0.33
						Chi-square for the split			2.67

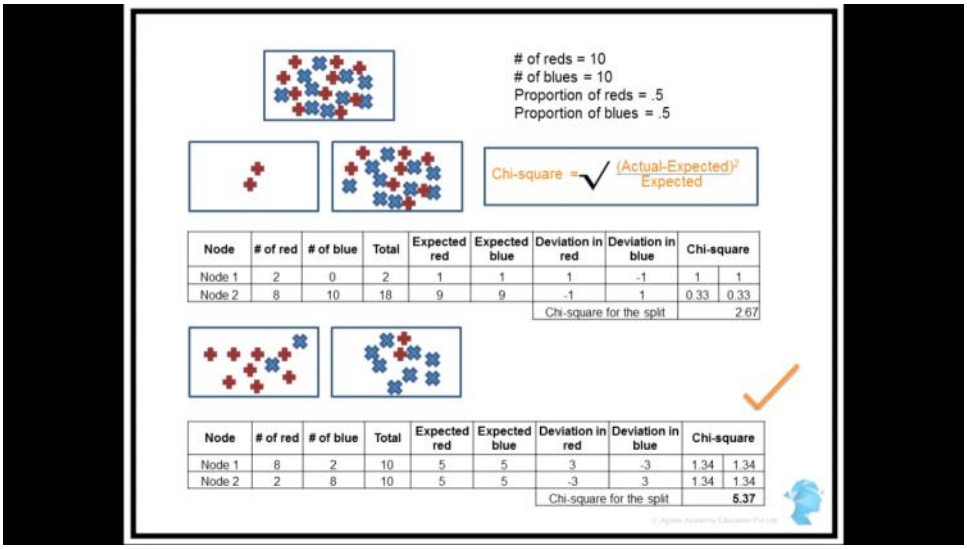
The chi square for the split is the sum of all the chi squares calculated.



Node	# of red	# of blue	Total	Expected red	Expected blue	Deviation in red	Deviation in blue	Chi-square	
Node 1	2	0	2	1	1	1	-1	1	1
Node 2	8	10	18	9	9	-1	1	0.33	0.33
						Chi-square for the split			2.67

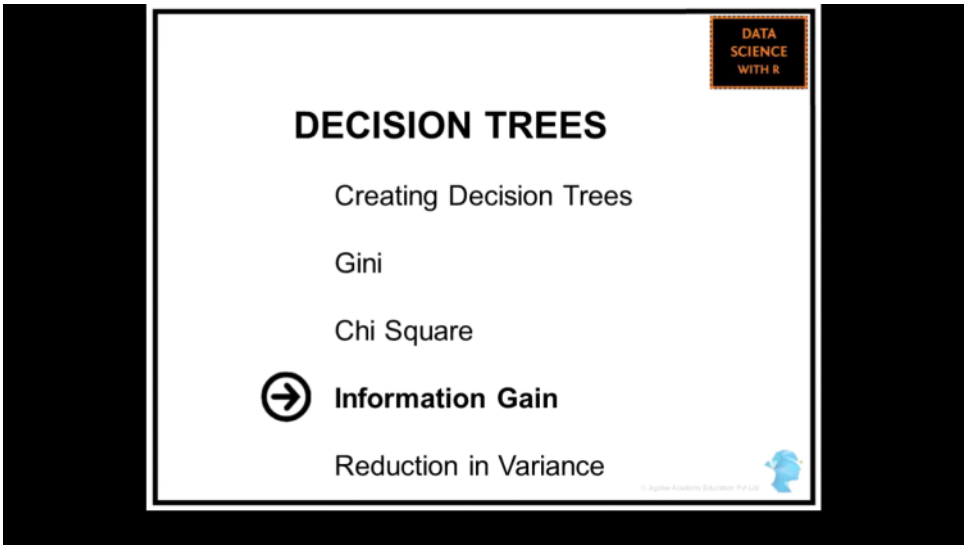
Node	# of red	# of blue	Total	Expected red	Expected blue	Deviation in red	Deviation in blue	Chi-square	
Node 1	8	2	10	5	5	3	-3	1.34	1.34
Node 2	2	8	10	5	5	-3	3	1.34	1.34
						Chi-square for the split			**5.37**

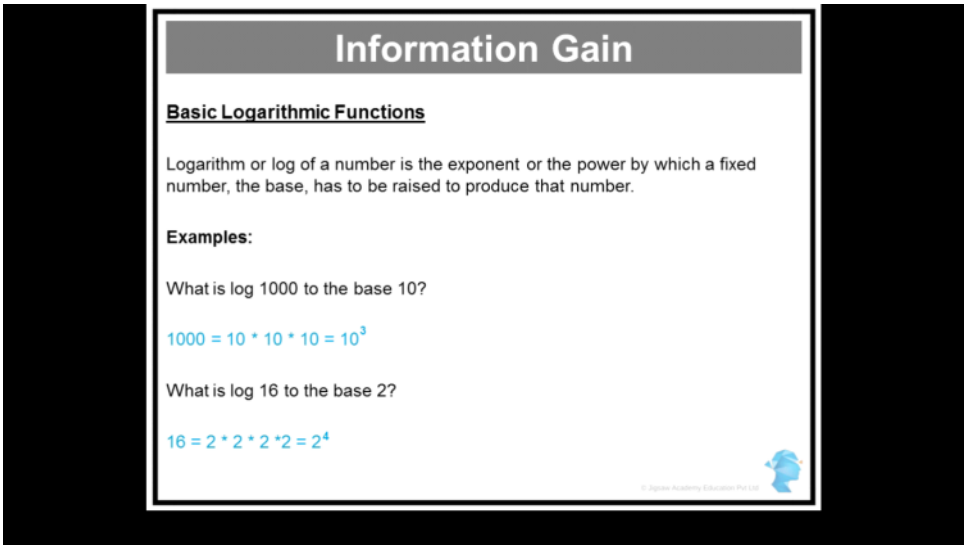
High score means a successful split with significantly population distribution.

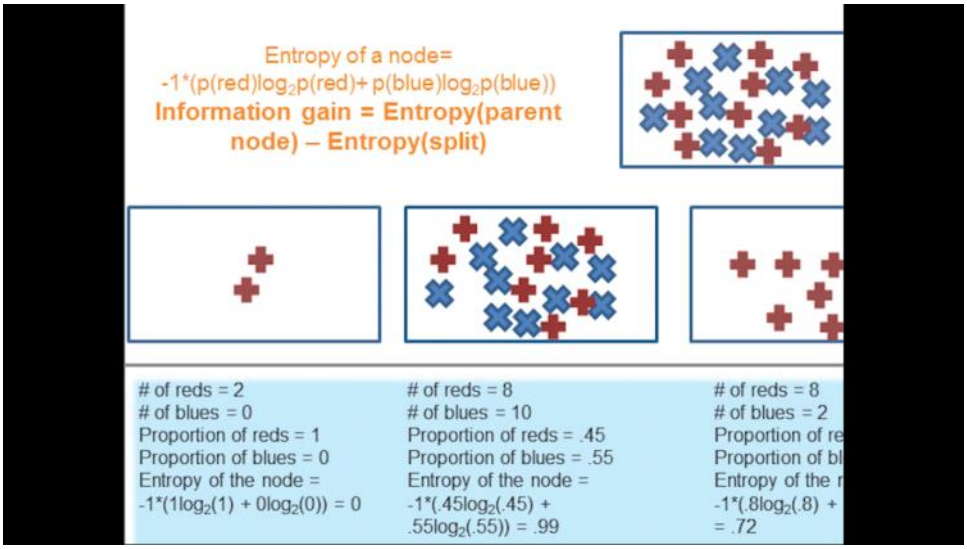


Node	# of red	# of blue	Total	Expected red	Expected blue	Deviation in red	Deviation in blue	Chi-square	
Node 1	2	0	2	1	1	1	-1	1	1
Node 2	8	10	18	9	9	-1	1	0.33	0.33
						Chi-square for the split			2.67

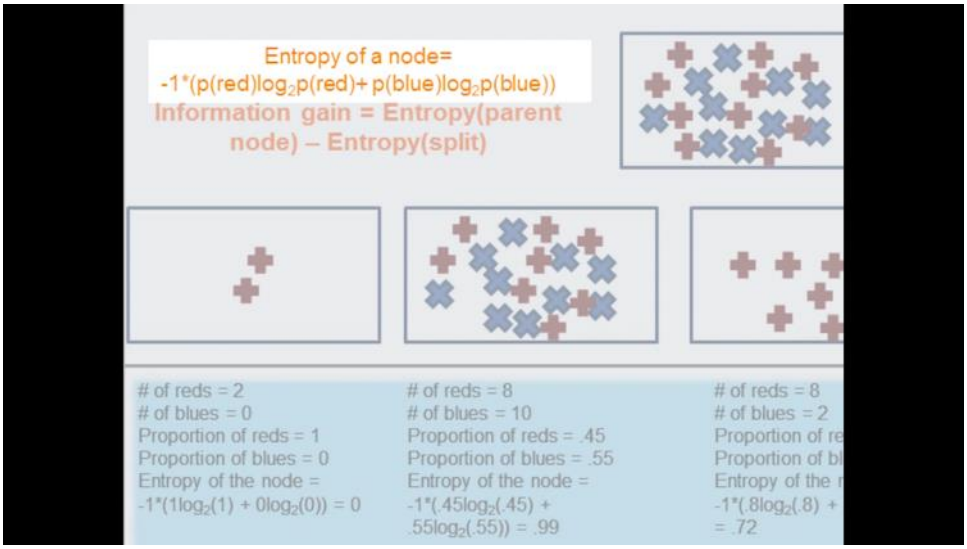
Node	# of red	# of blue	Total	Expected red	Expected blue	Deviation in red	Deviation in blue	Chi-square	
Node 1	8	2	10	5	5	3	-3	1.34	1.34
Node 2	2	8	10	5	5	-3	3	1.34	1.34
						Chi-square for the split			**5.37**

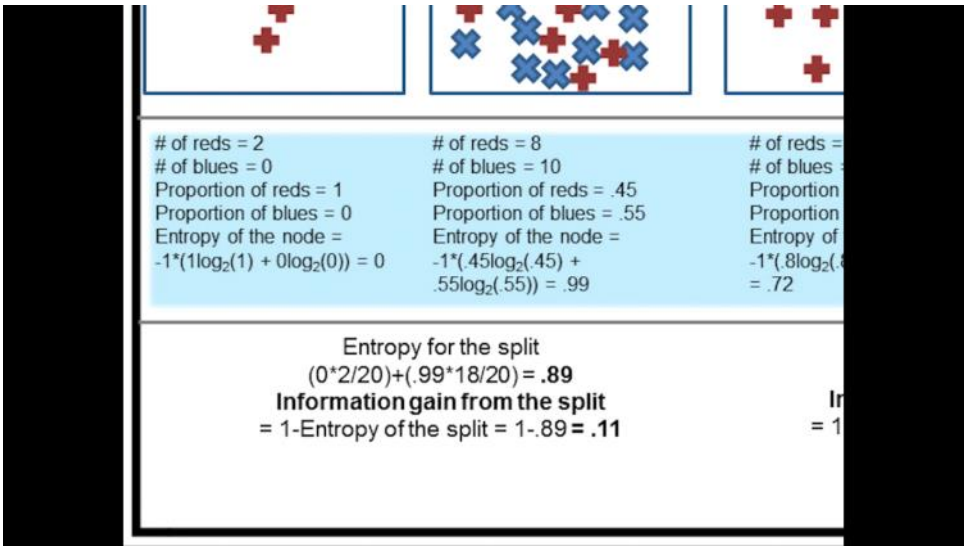


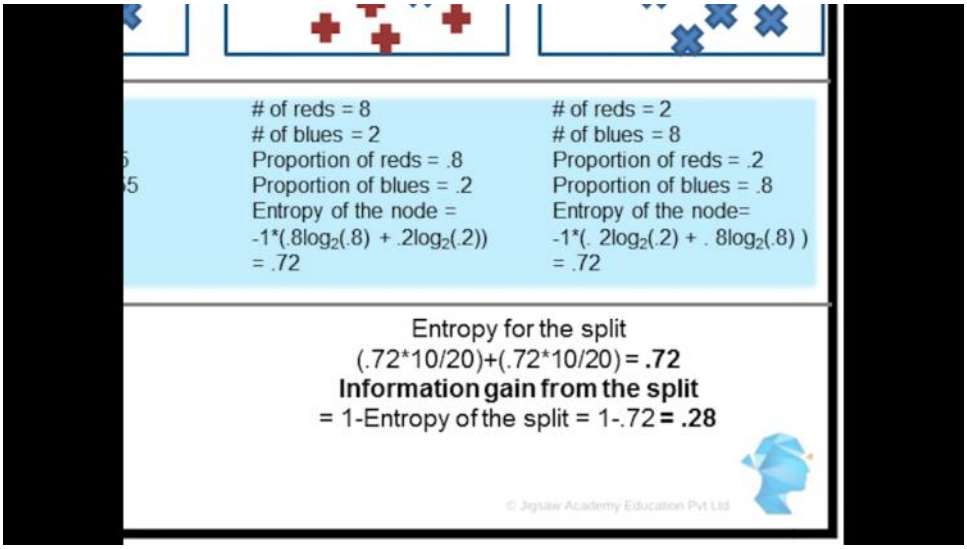


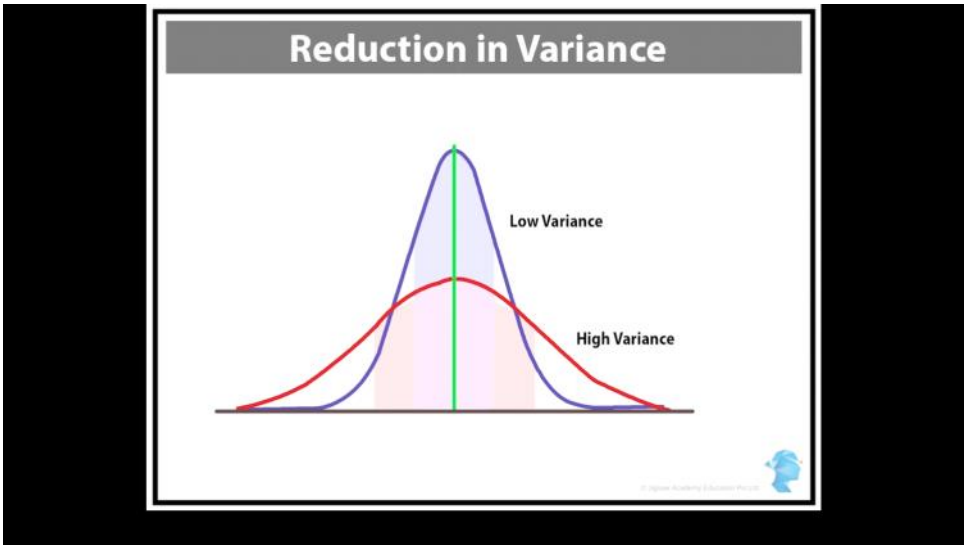


Information gain is the same as entropy reduction. Both a re used for decision tree algorithm

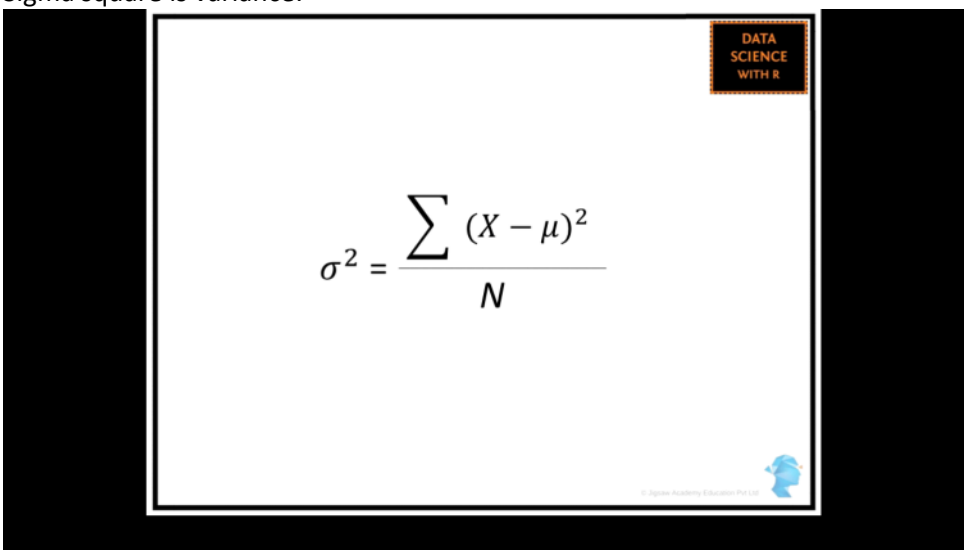


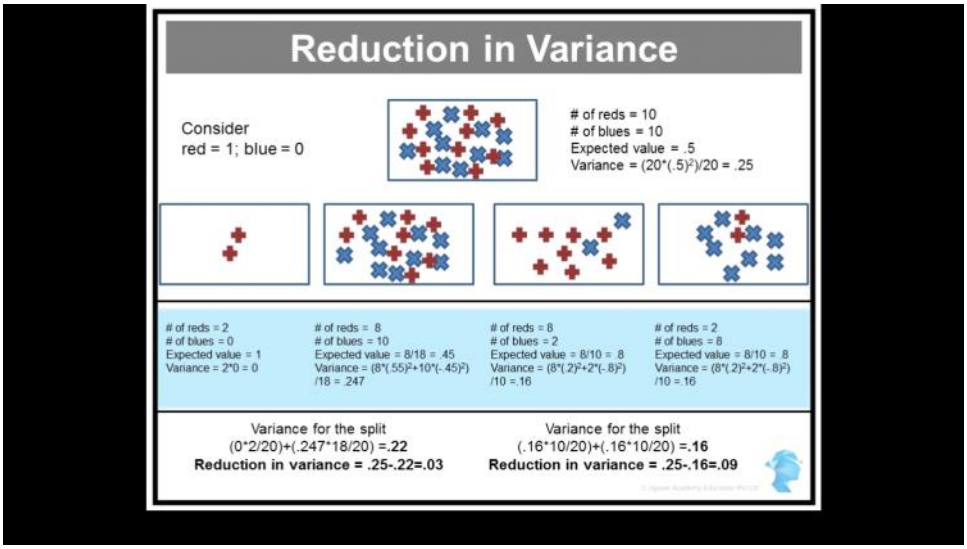


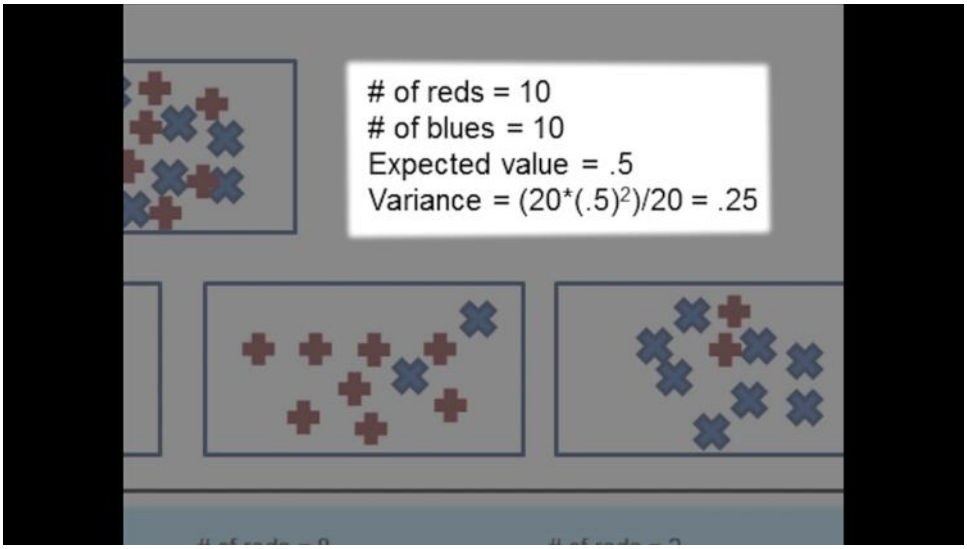


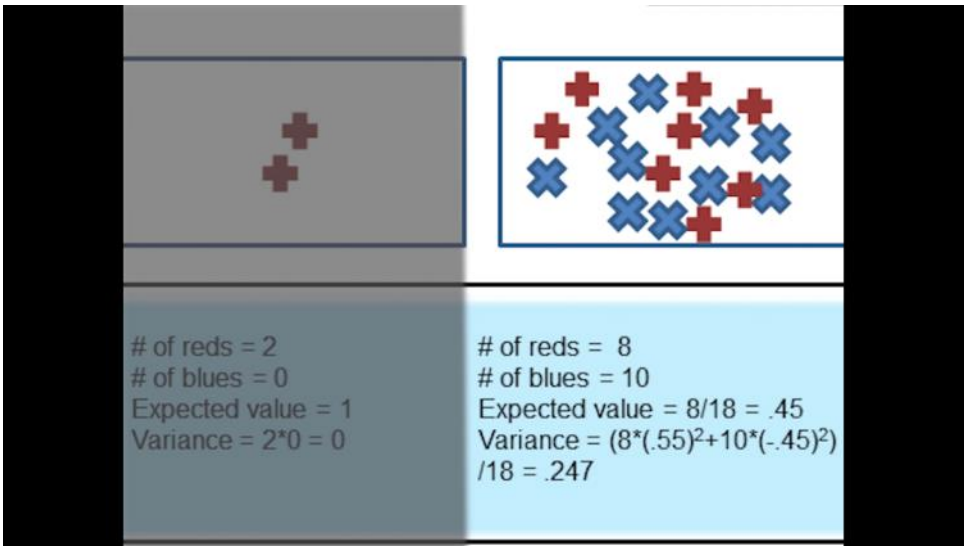


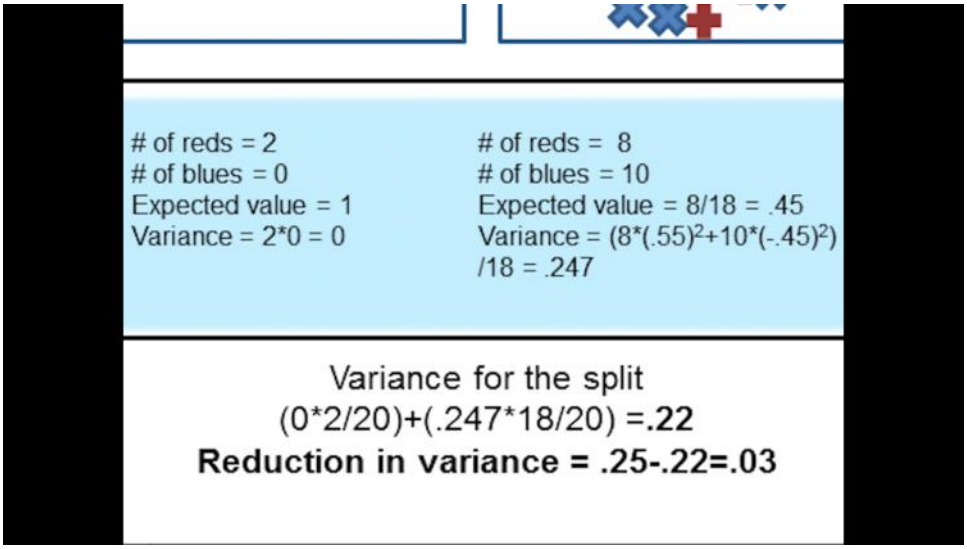
Sigma square is variance:

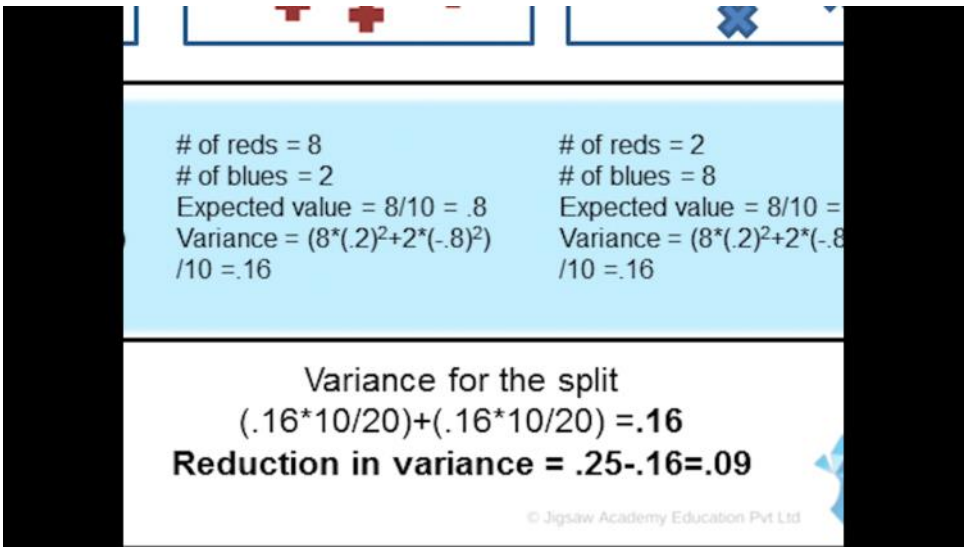


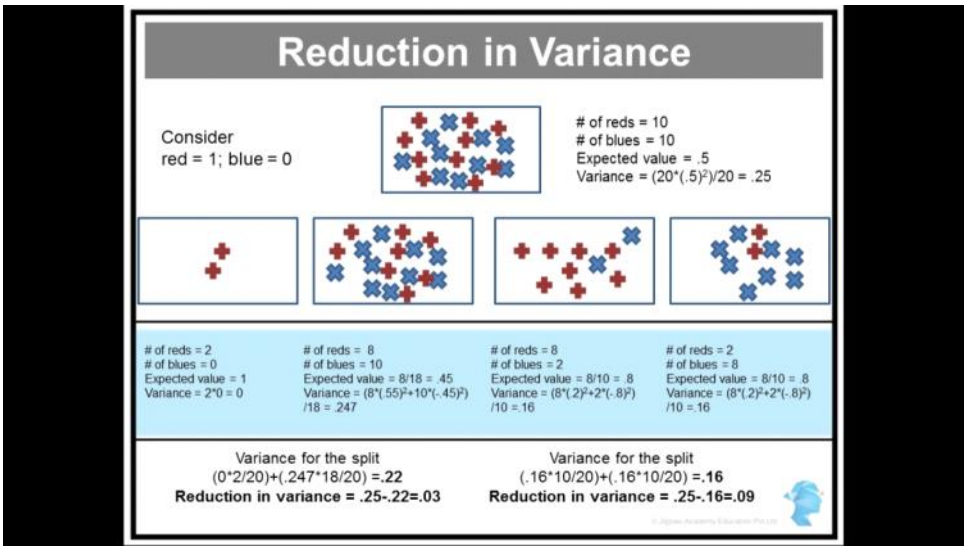

Reduction in Variance
Consider
red = 1; blue = 0
# of reds = 10
# of blues = 10
Expected value = .5
Variance = (20*(.5)²)/20 = .25
# of reds = 2
# of blues = 0
Expected value = 1
Variance = 2*0 = 0
# of reds = 8
# of blues = 10
Expected value = 8/18 = .45
Variance = (8*(.55)²+10*(-.45)²) /18 = .247
# of reds = 8
# of blues = 2
Expected value = 8/10 = .8
Variance = (8*(.2)²+2*(-.8)²) /10 = .16
# of reds = 2
# of blues = 8
Expected value = 8/10 = .8
Variance = (8*(.2)²+2*(-.8)²) /10 = .16
Variance for the split
(0*2/20)+(.247*18/20) =.22
Reduction in variance = .25-.22=.03
Variance for the split
(.16*10/20)+(.16*10/20) =.16
Reduction in variance = .25-.16=.09


# of reds = 10
# of blues = 10
Expected value = .5
Variance = (20*(.5)²)/20 = .25


# of reds = 2
# of blues = 0
Expected value = 1
Variance = 2*0 = 0
# of reds = 8
# of blues = 10
Expected value = 8/18 = .45
Variance = (8*(.55)²+10*(-.45)²) /18 = .247


# of reds = 2
# of blues = 0
Expected value = 1
Variance = 2*0 = 0
# of reds = 8
# of blues = 10
Expected value = 8/18 = .45
Variance = (8*(.55)²+10*(-.45)²) /18 = .247
Variance for the split
(0*2/20)+(.247*18/20) =.22
Reduction in variance = .25-.22=.03

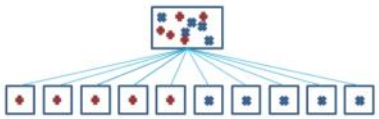

# of reds = 8
# of blues = 2
Expected value = 8/10 = .8
Variance = (8*(.2)²+2*(-.8)²) /10 =.16
# of reds = 2
# of blues = 8
Expected value = 8/10 =
Variance = (8*(.2)²+2*(-.8
/10 =.16
Variance for the split
(.16*10/20)+(.16*10/20) =.16
Reduction in variance = .25-.16=.09
© Jigsaw Academy Education Pvt Ltd


Reduction in Variance
Consider
red = 1; blue = 0
# of reds = 10
# of blues = 10
Expected value = .5
Variance = (20*(.5)²)/20 = .25
# of reds = 2
# of blues = 0
Expected value = 1
Variance = 2*0 = 0
# of reds = 8
# of blues = 10
Expected value = 8/18 = .45
Variance = (8*(.55)²+10*(-.45)²) /18 = .247
# of reds = 8
# of blues = 2
Expected value = 8/10 = .8
Variance = (8*(.2)²+2*(-.8)²) /10 =.16
# of reds = 2
# of blues = 8
Expected value = 8/10 = .8
Variance = (8*(.2)²+2*(-.8)²) /10 =.16
Variance for the split
(0*2/20)+(.247*18/20) =.22
Reduction in variance = .25-.22=.03
Variance for the split
(.16*10/20)+(.16*10/20) =.16
Reduction in variance = .25-.16=.09

Overfitting and Tree Pruning

Overfitting:

Greed for accuracy leading to over-complex trees



Pruning algorithms:

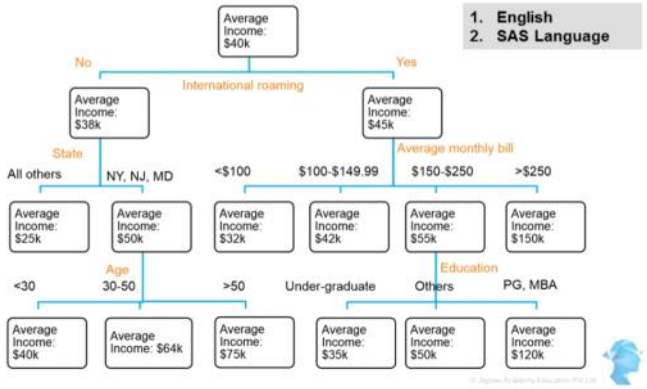
- **CART** – Prunes the tree by imposing a complexity penalty based on number of leaves in the tree.
- **C5** – Assumes a higher rate of error than what is seen on the training data. The smaller the node, the more the increase over observed. When the child node estimate is higher than the parent node, tree is pruned.

Manual pruning:

- Analyst needs to study the tree in detail. Any node that looks unstable should be pruned.

Decision Tree Rules

A telecom company wants to impute missing values in the income field for its customers.

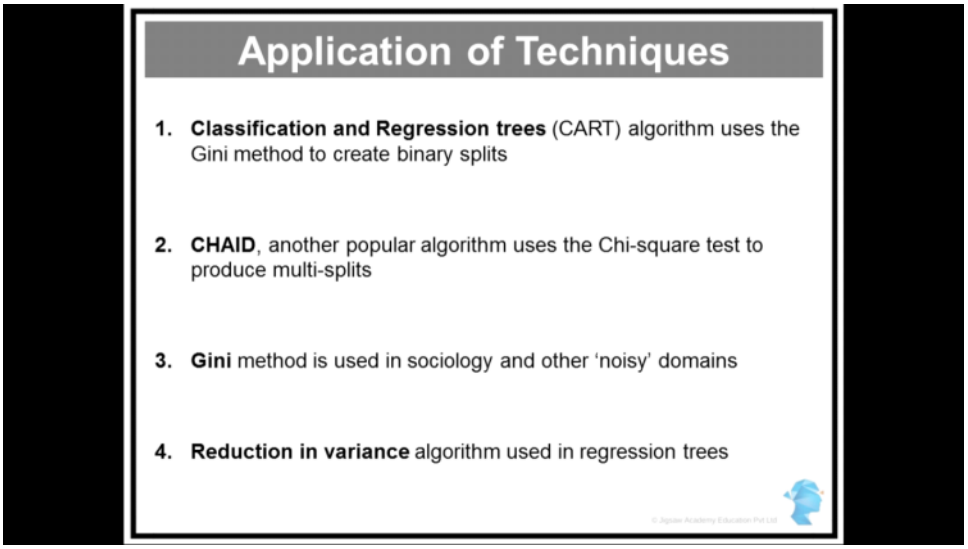


1. **English**
2. **SAS Language**

DATA SCIENCE WITH R

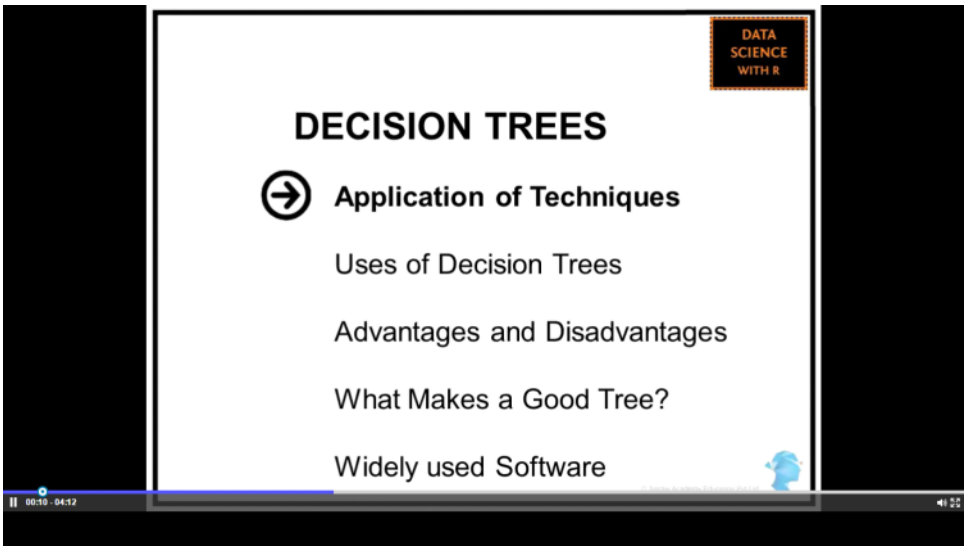
DECISION TREES

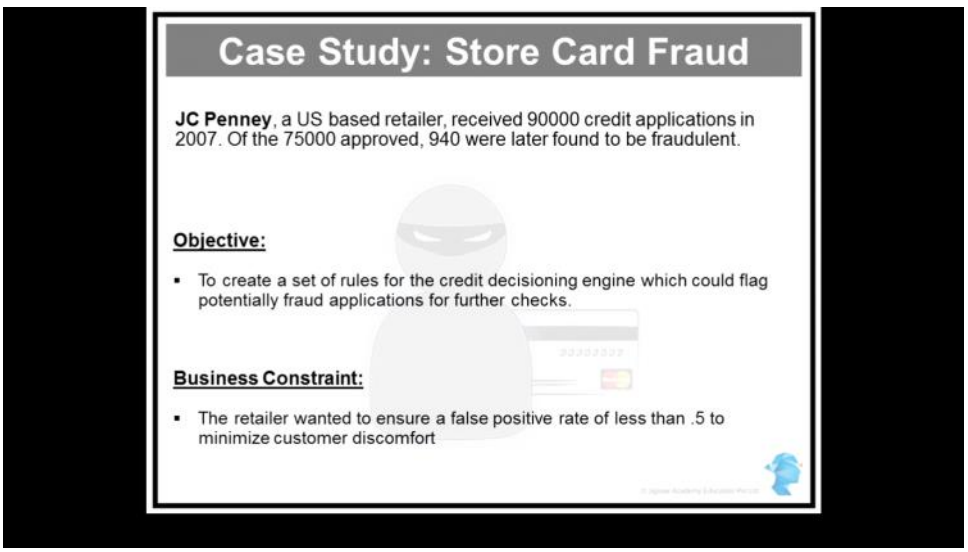
- **Application of Techniques**
- Uses of Decision Trees
- Advantages and Disadvantages
- What makes a Good Tree?
- Widely used Software



Thumbrules:

- Binary target --> cart
- Multiple splits at each level(tree) --> CHAID
- Lot of noise --> CART
- F test --> for numeric target variables i.e. reduction in variance test





Case Study: Store Card Fraud

Methodology:

- A random set of 8460 non-fraudulent credit applications was combined with the 940 valid fraudulent applications
- Biased sample set created with 10% fraud rate
- Data partitioned into test (70%) and validation (30%)
- Multiple decision trees created using different algorithms
- Focus on maximizing fraud detection and minimize false positives

Case Study: Store Card Fraud

Methodology:

- Final tree selected based on the performance on validation data set
- Rules extracted and simplified. Converted to SQL
- Rules implemented on the client engine

Results:

Model performance was tracked over the next 3 months. Out of 17340 applications coming in, the model flagged 585 applications of which 320 were later denied credit. Fraud rate for JC Penney came down from 1.25% to .7%.

DATA SCIENCE WITH R

DECISION TREES

Application of Techniques

Uses of Decision Trees

Advantages and Disadvantages

What Makes a Good Tree?

Widely used Software

Uses of Decision Trees

- ❑ Variable selection
- ❑ Interaction detection
- ❑ Missing value imputation
- ❑ Model interpretation
- ❑ Predictive modeling

DATA SCIENCE WITH R

DECISION TREES

Application of Techniques

Uses of Decision Trees

Advantages and Disadvantages

What makes a Good Tree?

Widely used Software

Advantages

- ✓ Easy to use and understand
- ✓ Graphical representation aids visual understanding
- ✓ Require less data preparation than other methods
- ✓ Non parametric method

Advantages

- ✓ Catch interactions and non-additive behavior between variables
- ✓ Result in form of rules. Easy to interpret.
- ✓ Deal well with outliers and missing values
- ✓ Combine well with other modeling techniques

Disadvantages

1. Tendency to over-fit
2. Do not make use of information in data as well as neural networks or regression
3. May results in too many splits on variables with many values
4. Time series data requires extensive preparation

DATA SCIENCE WITH R

DECISION TREES

Application of Techniques

Uses of Decision Trees

Advantages and Disadvantages

What Makes a Good Tree?

Widely used Software

What Makes a Good Tree?

Most desirable trees are not necessarily the ones that do the best on the training data

Features of a good tree:

- Accurate
- Use few variables
- Easy to visualize and interpret
- Make sense intuitively

When To Use?

- Very popular classification technique especially when problem is binary
- Not the first choice for estimating continuous values
- Best option when the interpretation of results is more important than the accuracy
- Use when time is of the essence
- Whenever available, use for initial data exploration in any project

DATA SCIENCE WITH R

DECISION TREES

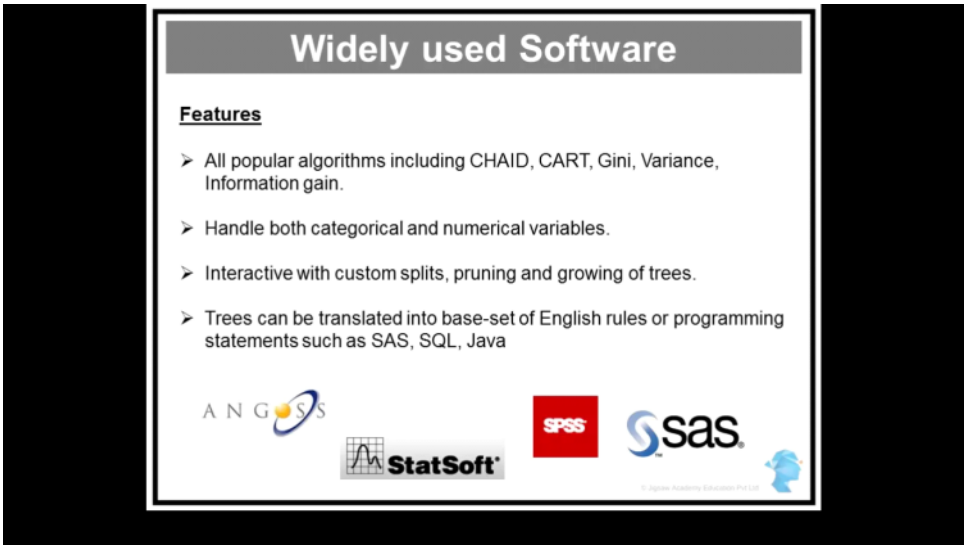
Application of Techniques

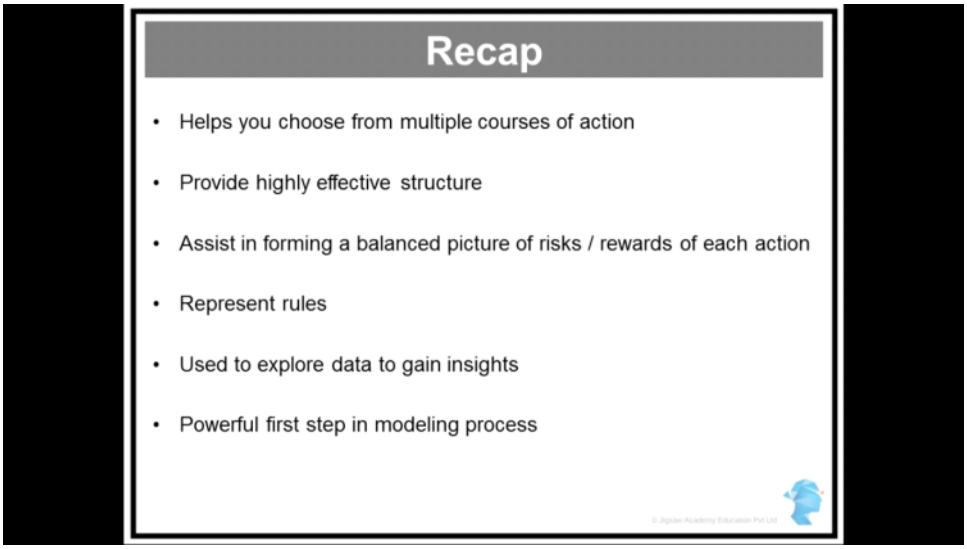
Uses of Decision Trees

Advantages and Disadvantages

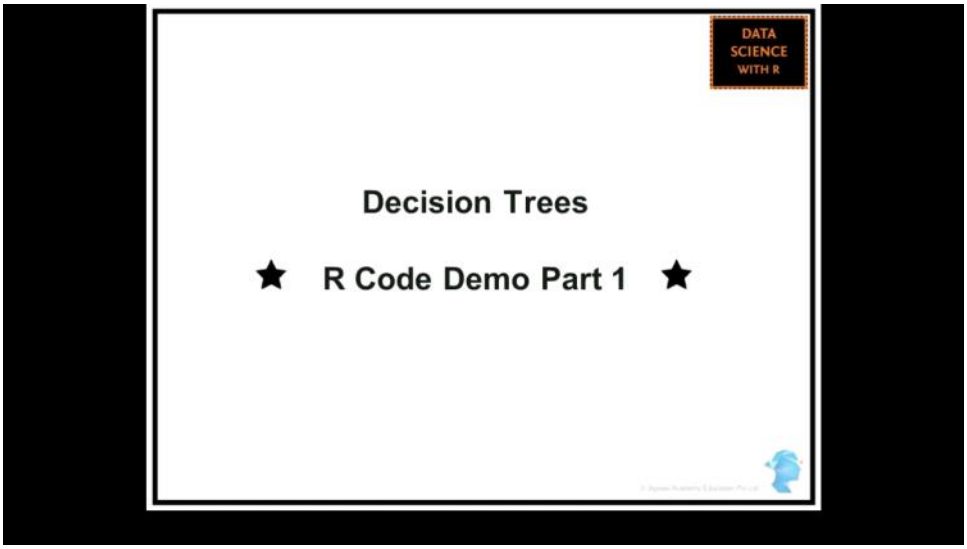
What Makes a Good Tree?

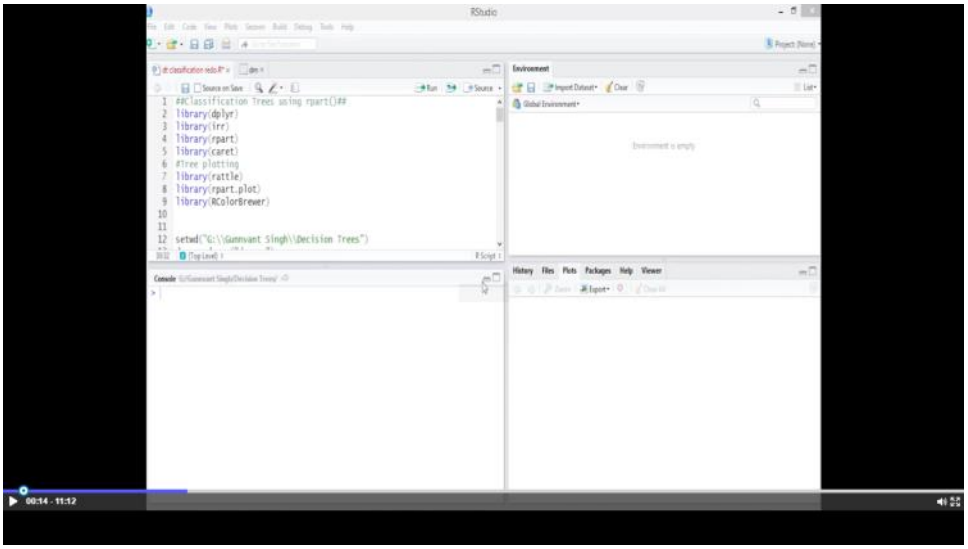
Widely used Software



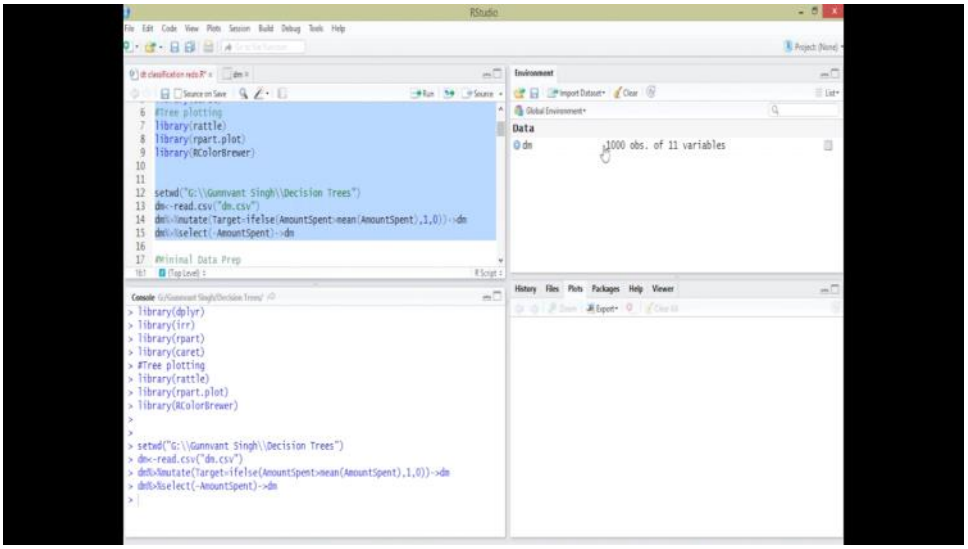


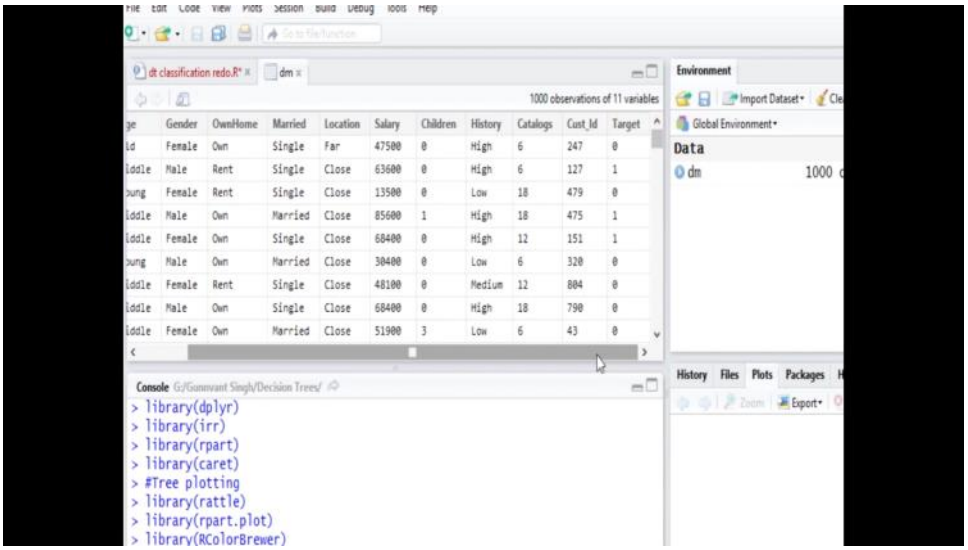
- Data mining techniques by Berry and Linoff



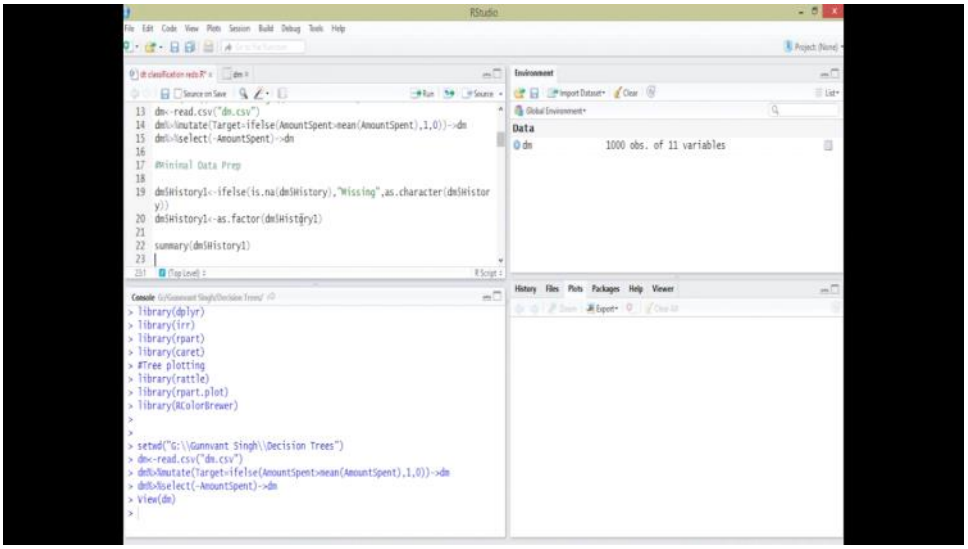


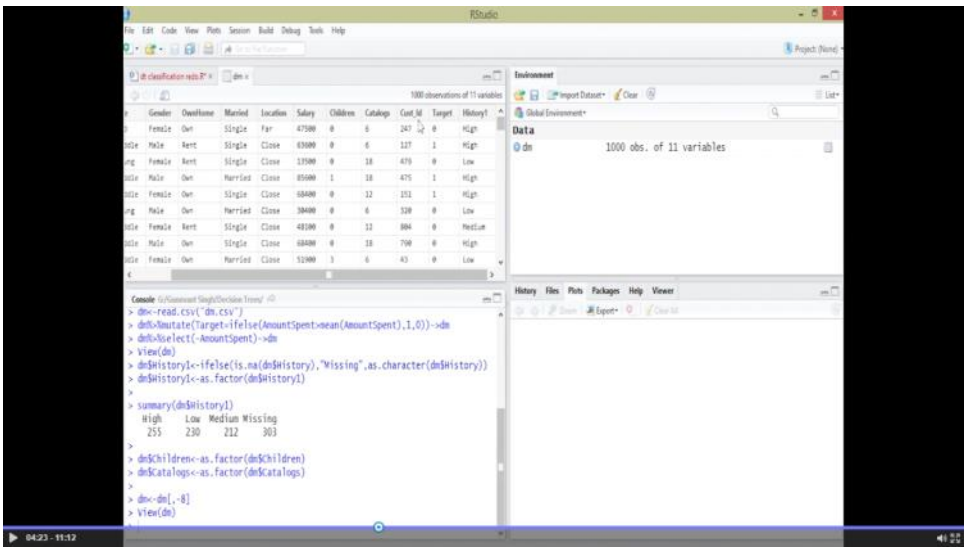
We will be using library (rpart) for classification and decision trees

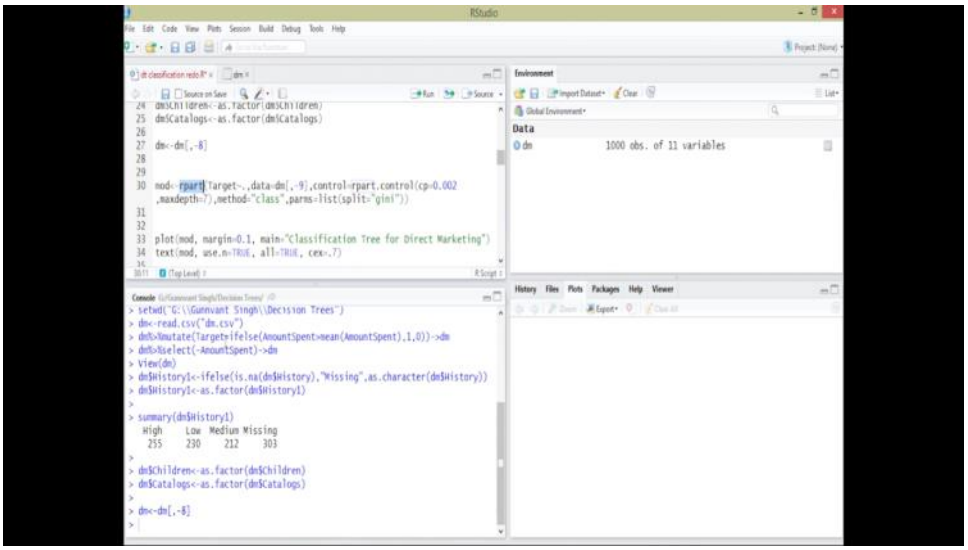




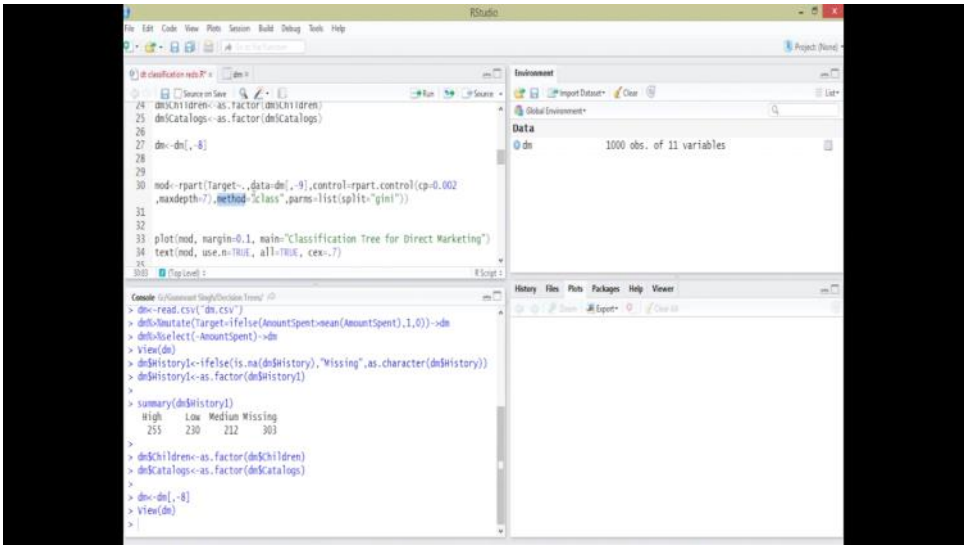
Here is the data prep stage. It is minimal data prep. Substituting NA and converting to factors

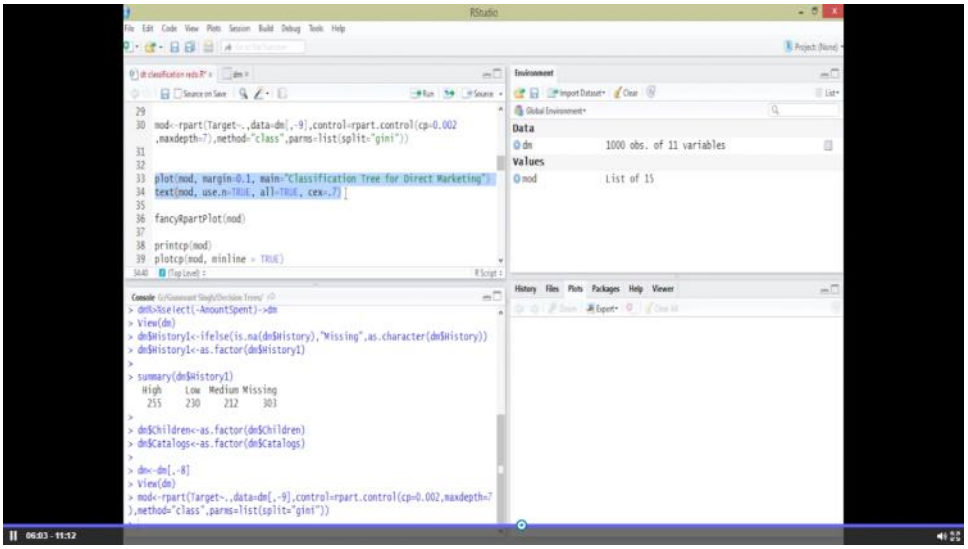


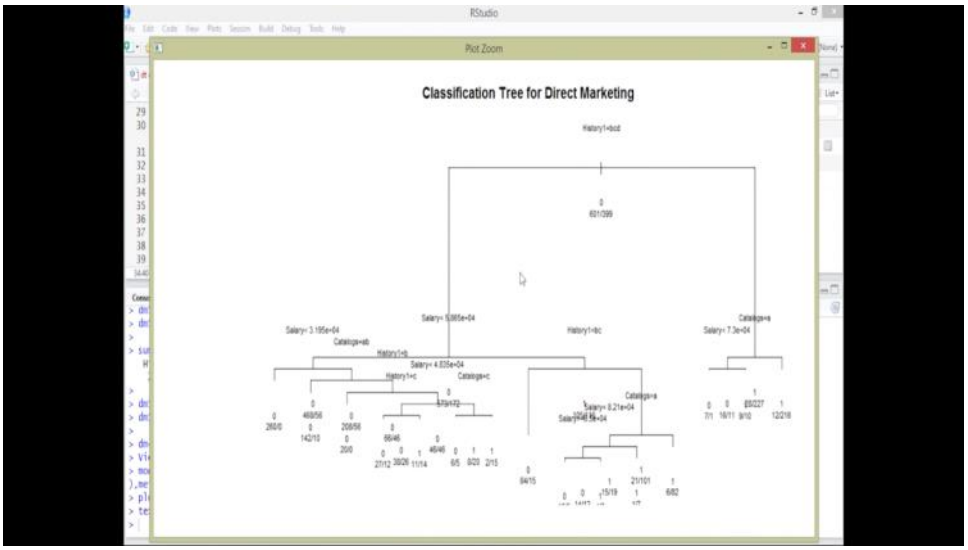




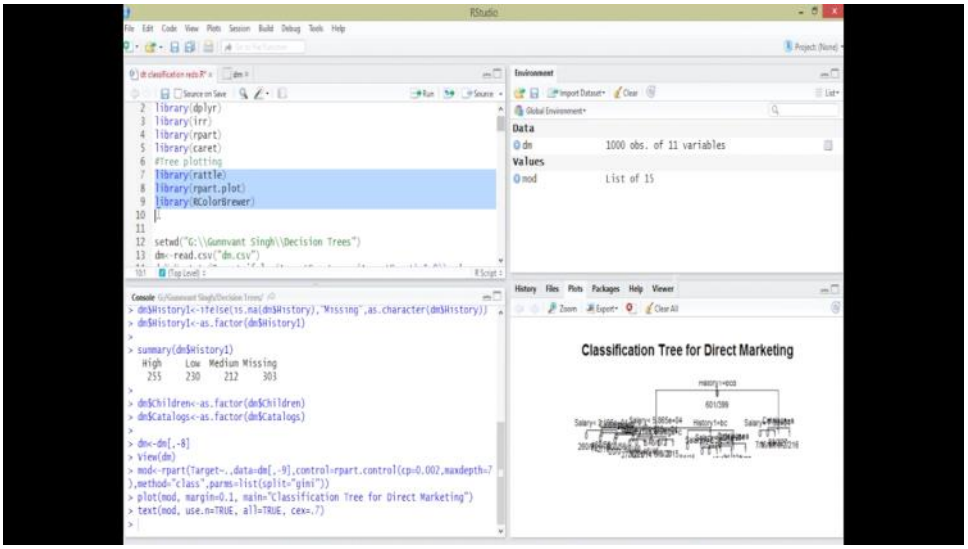
Cp --> cost parameter, maxdepth --> levels in the tree, method --> class as this is a classification model.

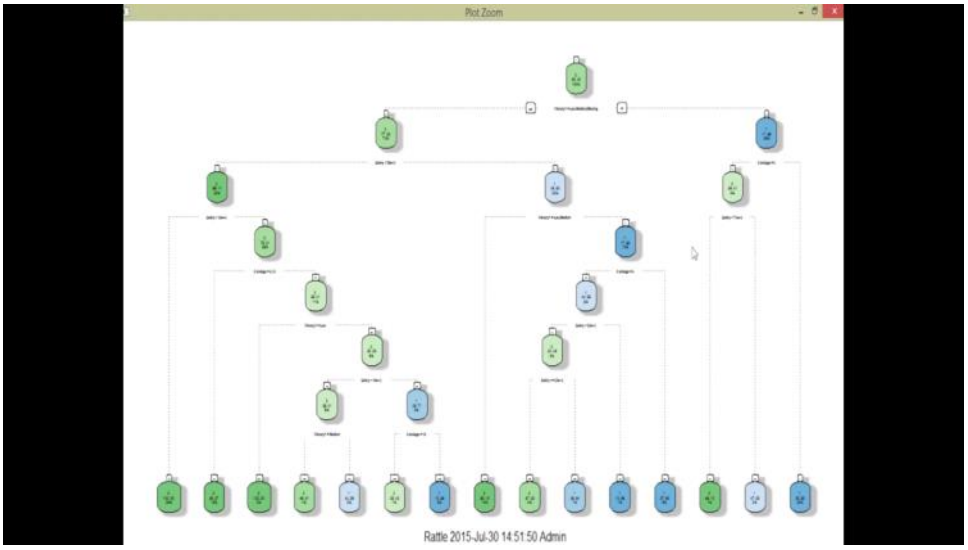




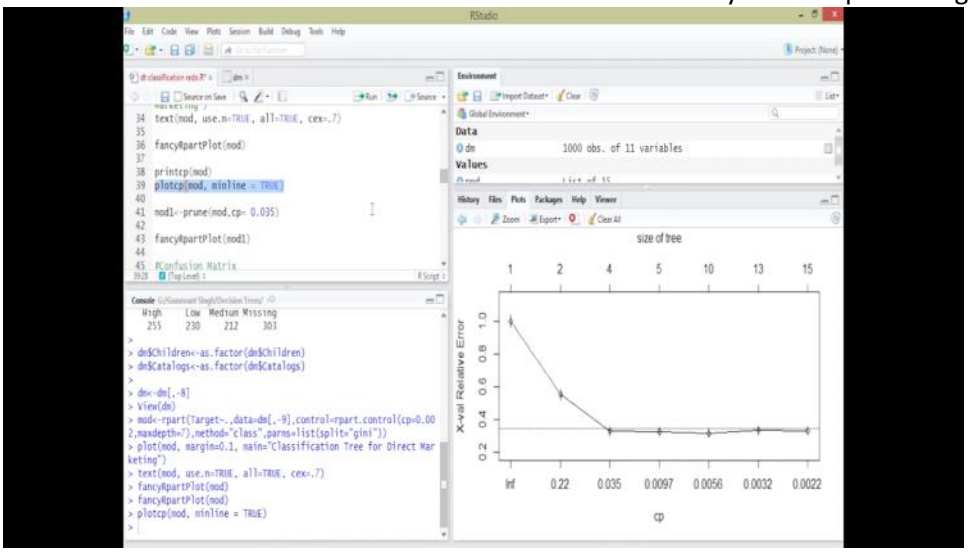


We can make it prettier by using the "fancyrpartplot(model)
This uses the following libraries --> rattle, rpart.plot, RColorBrewer

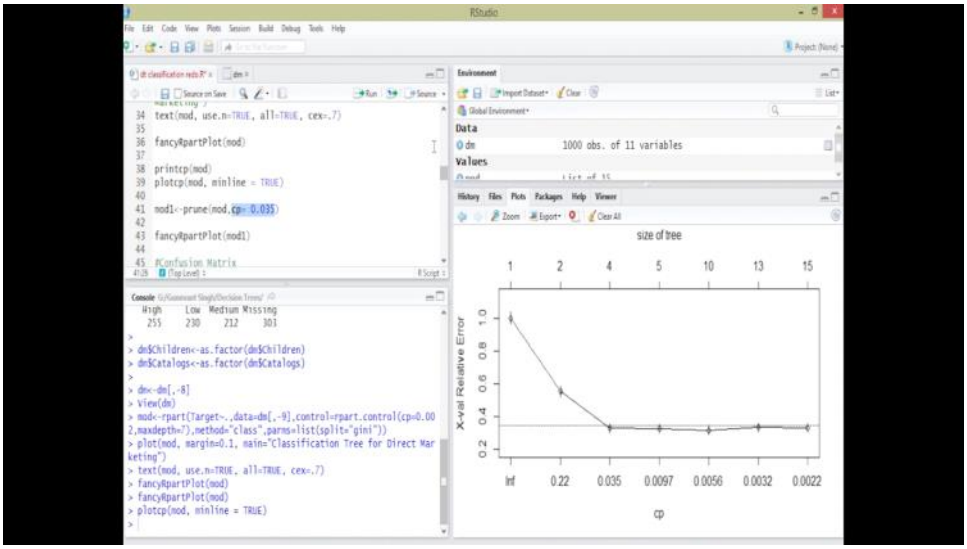




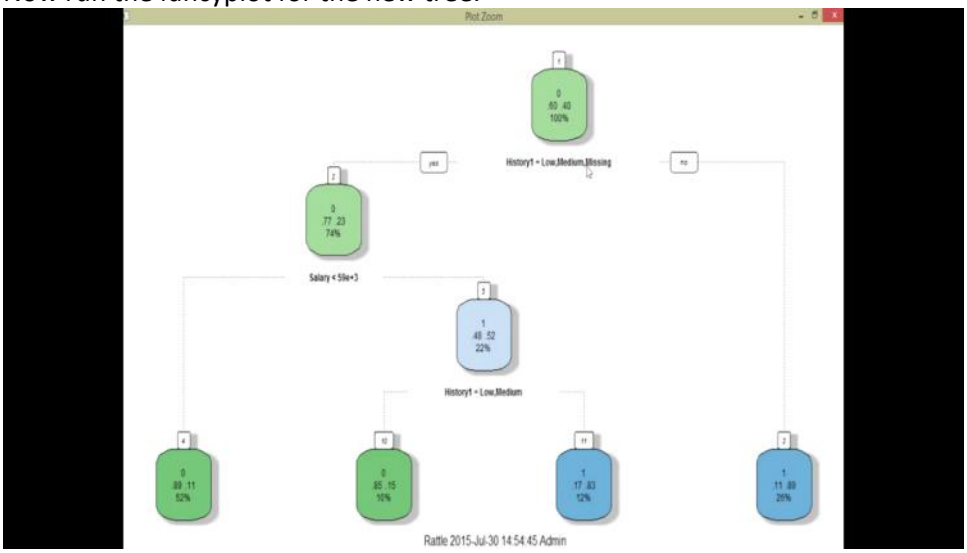
To determine where to cut the decision tree is determined by a screenplot that gives a breakdown:



Use the CP value where the curve drops below the line to determine where to prune.

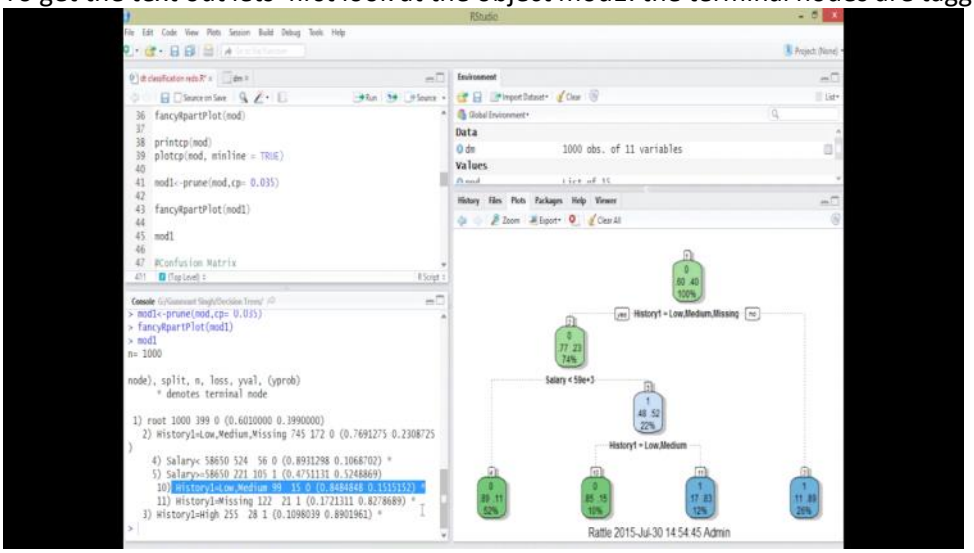


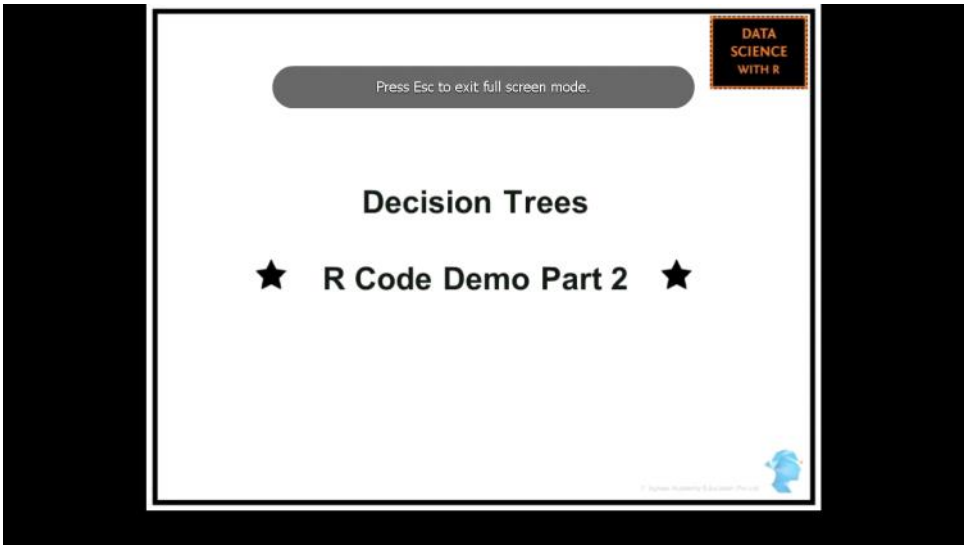
Now run the fancyplot for the new tree.

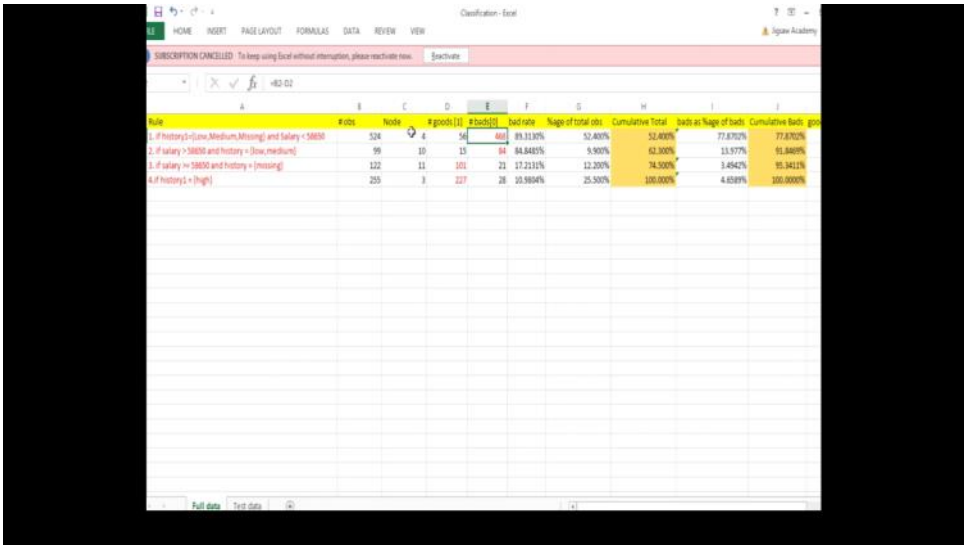


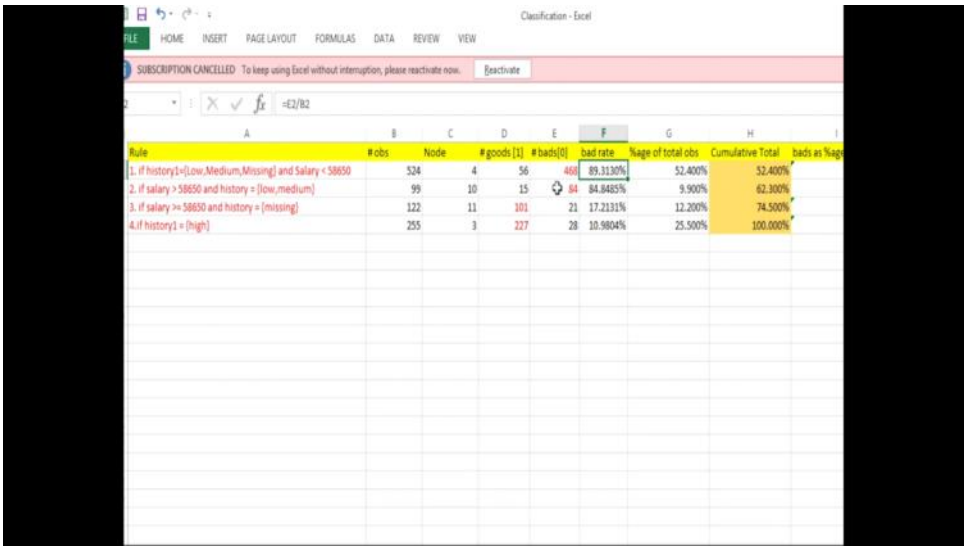
The next step is to derive the rules after pruning .
The rules can either bederieved by interpreting the plot or using the text output of decision tree.

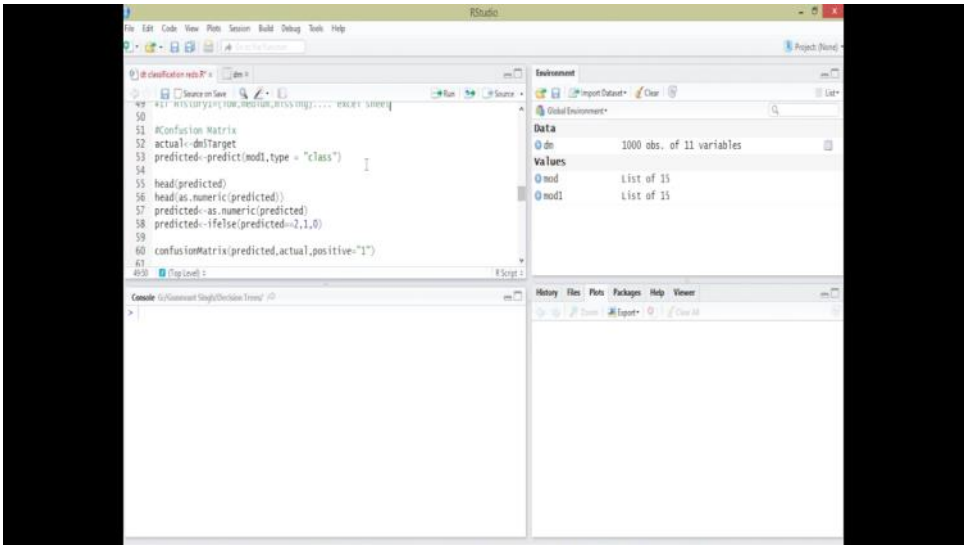
To get the text out lets' first look at the object mod1. the terminal nodes are tagged with *:

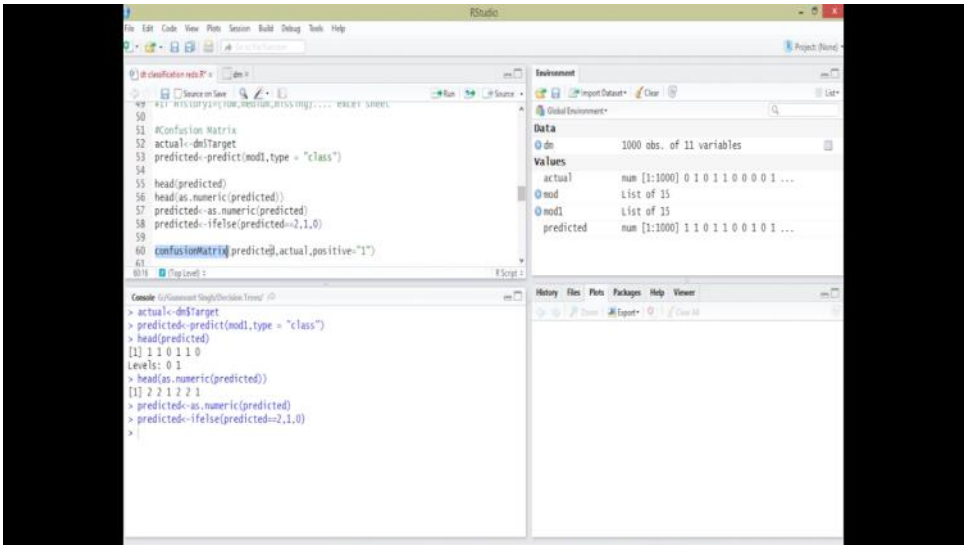



Press Esc to exit full screen mode.
DATA SCIENCE WITH R
Decision Trees
★ R Code Demo Part 2 ★

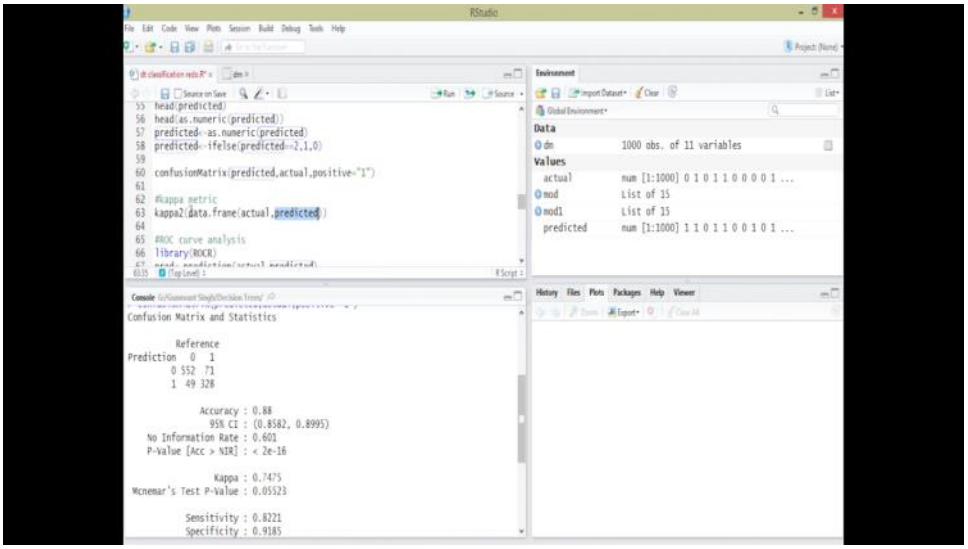

Classification - Excel
HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW
SUBSCRIPTION CANCELLED To keep using Excel without interruption, please reactivate now.
Reactivate
Rule
1. if history1=(Low,Medium,Missing) and Salary < 58650
2. if salary > 58650 and history = (low,medium)
3. if salary >= 58650 and history = (missing)
4.if history1 = (high)
Full data
Test data


Classification - Excel
HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW
SUBSCRIPTION CANCELLED To keep using Excel without interruption, please reactivate now.
Reactivate
=E2/B2
Rule
1. if history1=(Low,Medium,Missing) and Salary < 58650
2. if salary > 58650 and history = (low,medium)
3. if salary >= 58650 and history = (missing)
4.if history1 = (high)

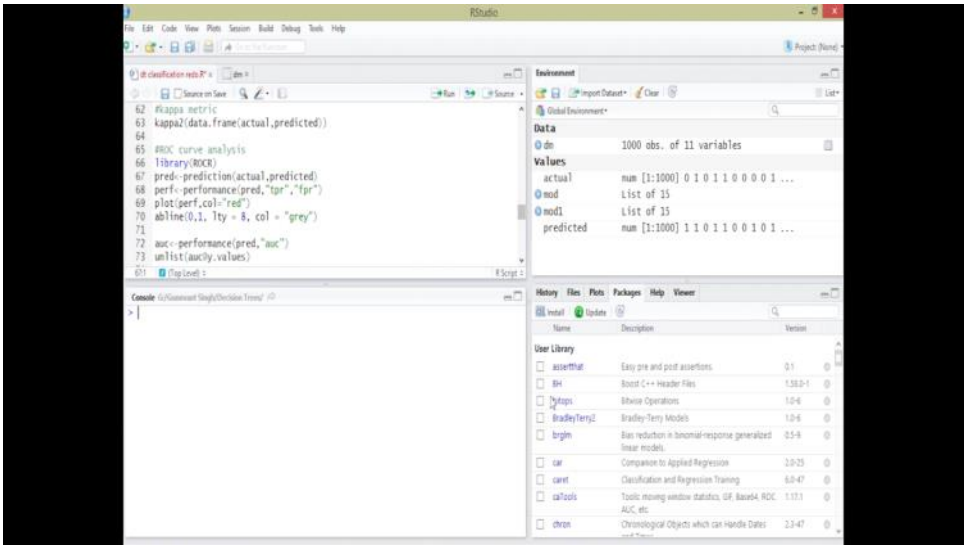




Use confusion matrix from package caret to generate a confusion matrix.



Load the library ROCR to create the ROC(receiver operating) curve:



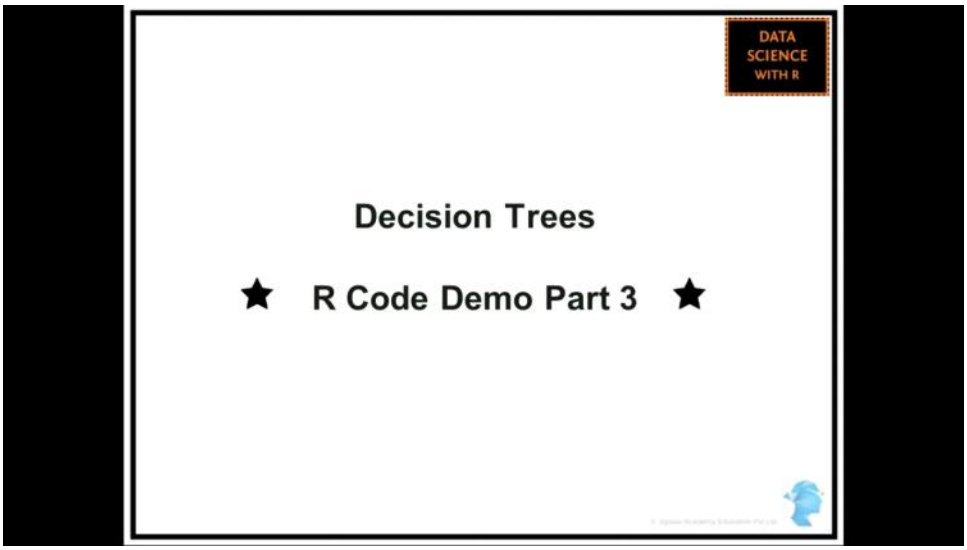
We will use the prediction function as part of ROCR.
Once we have that we will run a performance function with the earlier output.

After getting the output, we plot it in agraph and draw a 45 degree line using abline

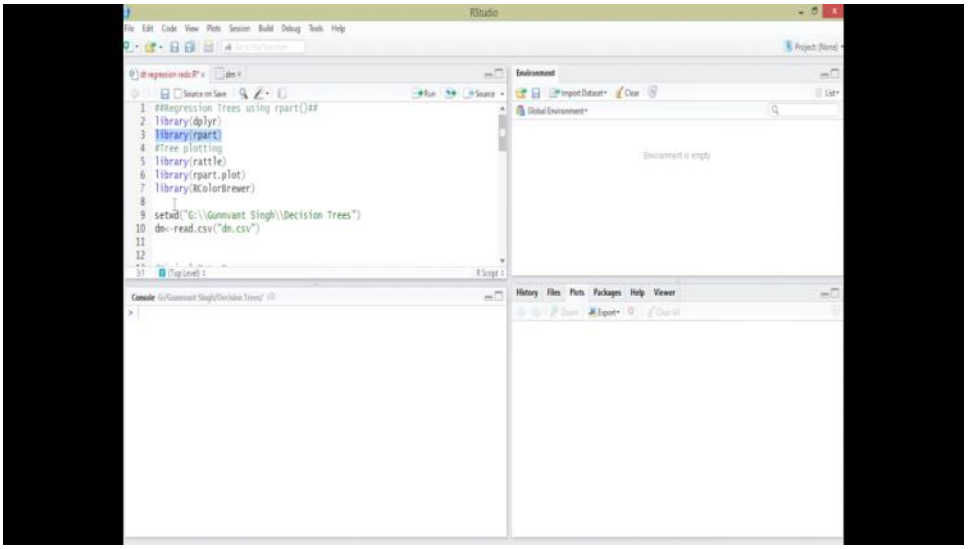
Then we compute the are under the curve which tells about the lift/improvement:
Auc<-performance(pred,"auc")
It is a S4 object.

Since we need data from Y values, we extract it.
Since it is a list object, we will unlist it.

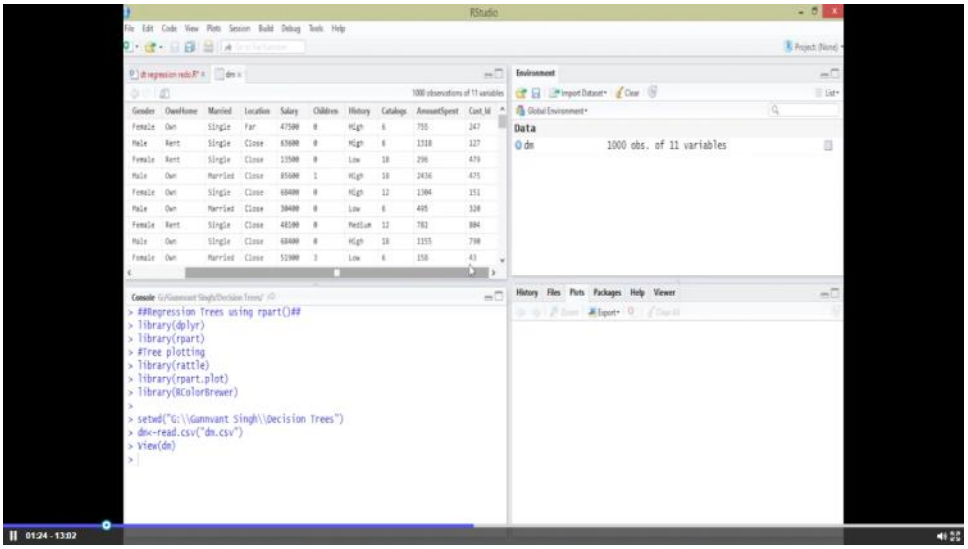
Unlist(auc@y.values)
For a decent classifier our AUC needs to be better than 0.6



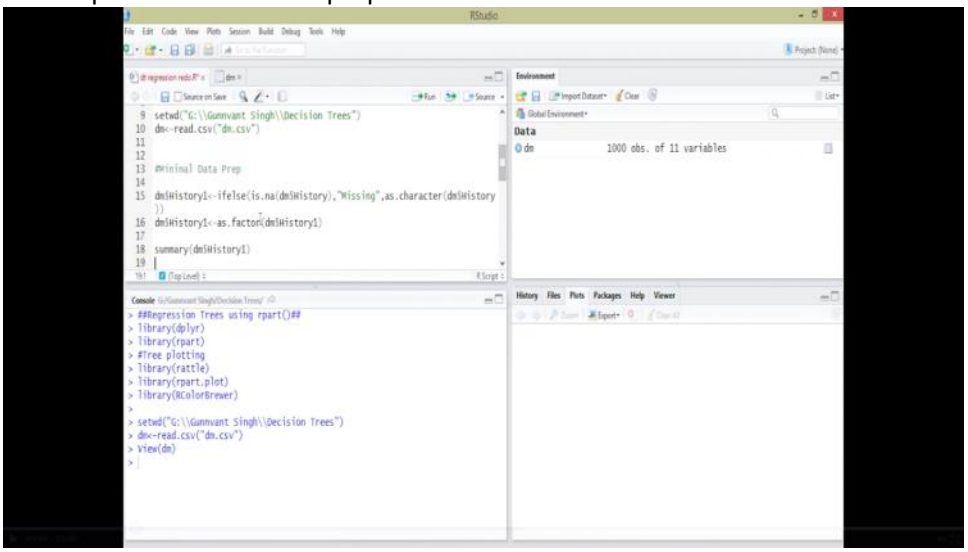
How to build regresion models using decision trees. We will use rpart to build both the classification tree and regression tree:



We will look at predicting amount spent

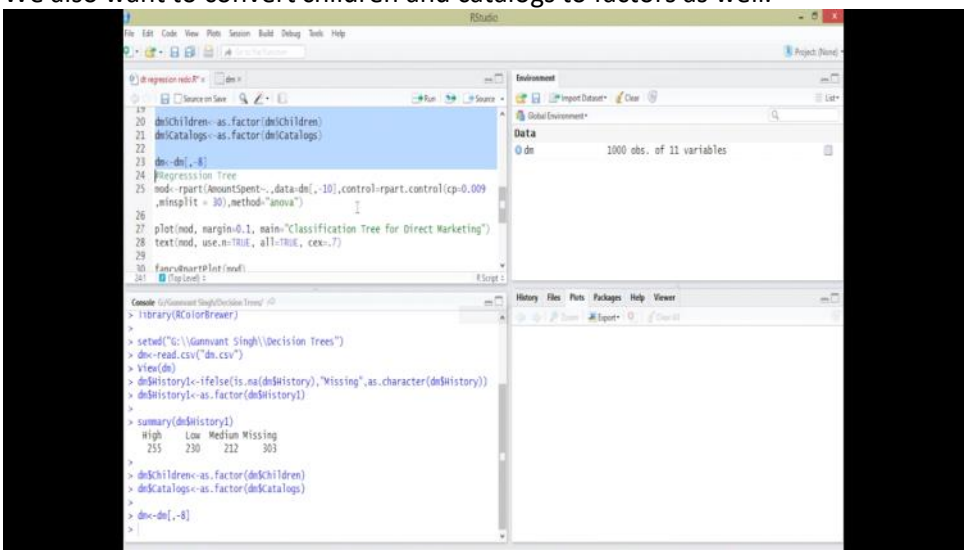


This requires minimal data prep.

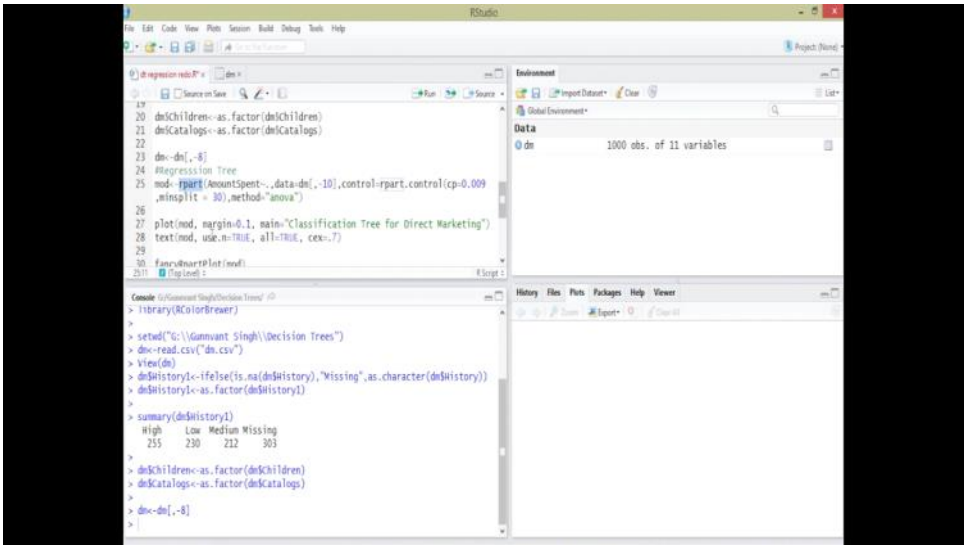


First step is to replace NA's . We do it for history, this is basically an indicator of historic spending behavior

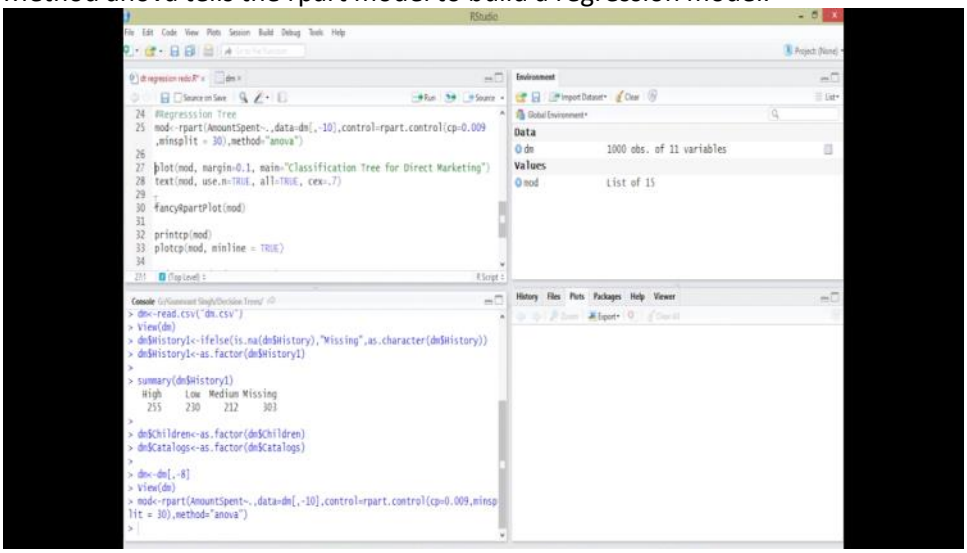
We also want to convert children and catalogs to factors as well.

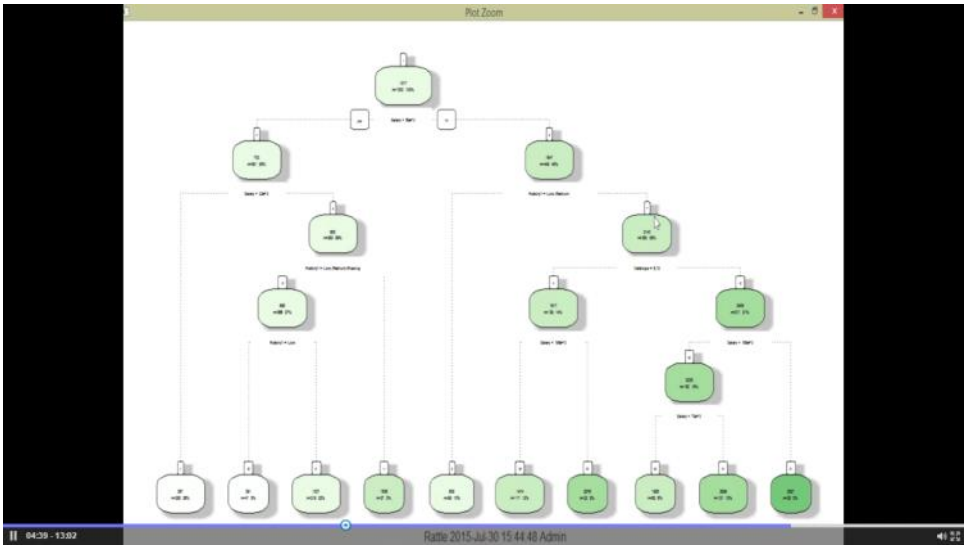


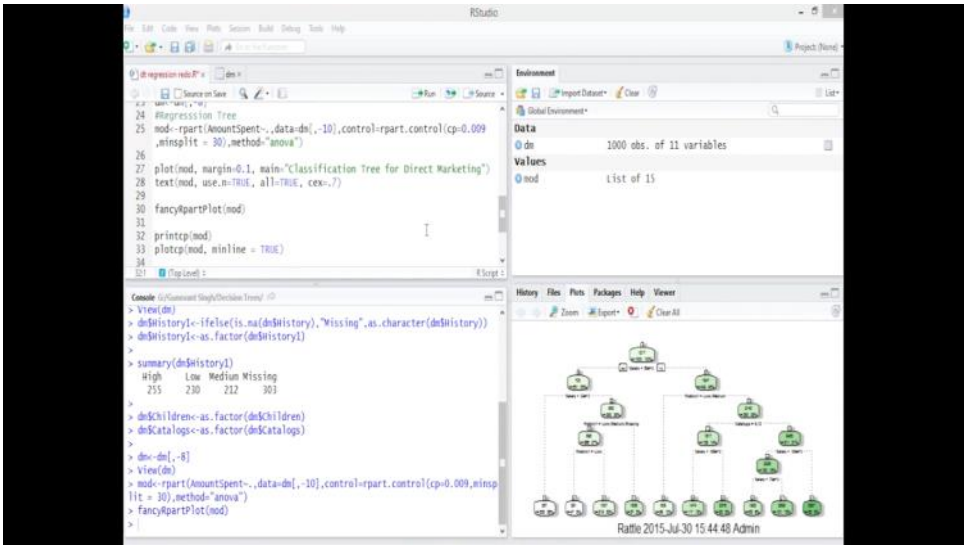
We will use all our variable and run the rpart command to find the Amount spent,excluding the customer id



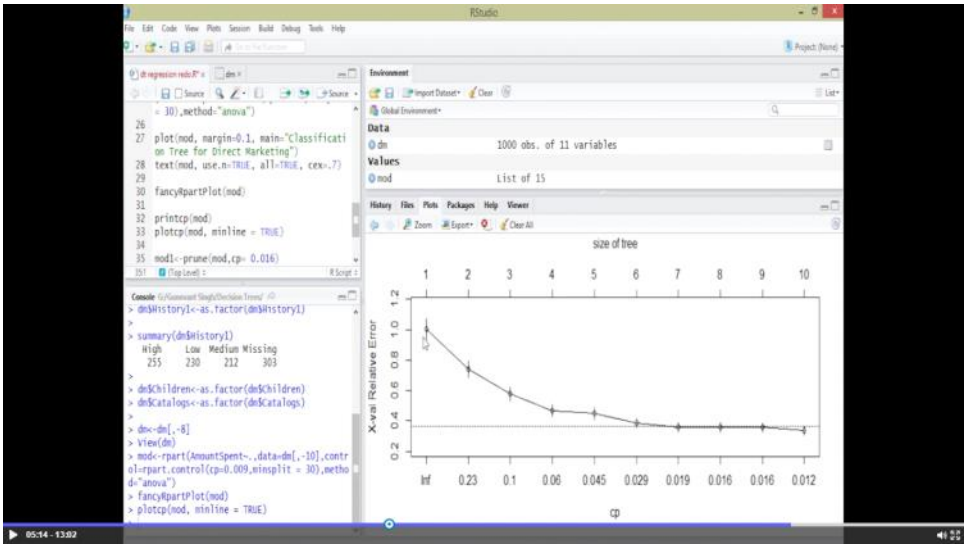
Method anova tells the rpart model to build a regression model.



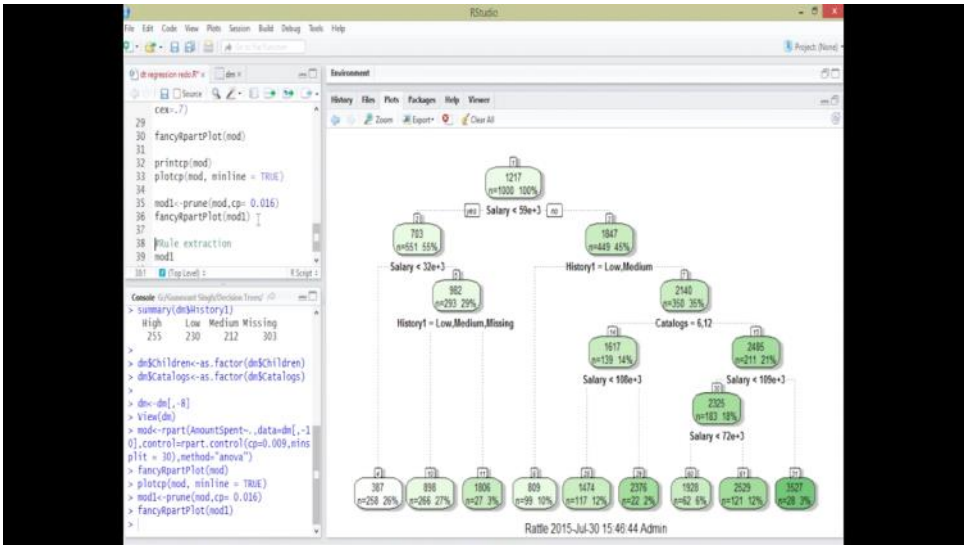


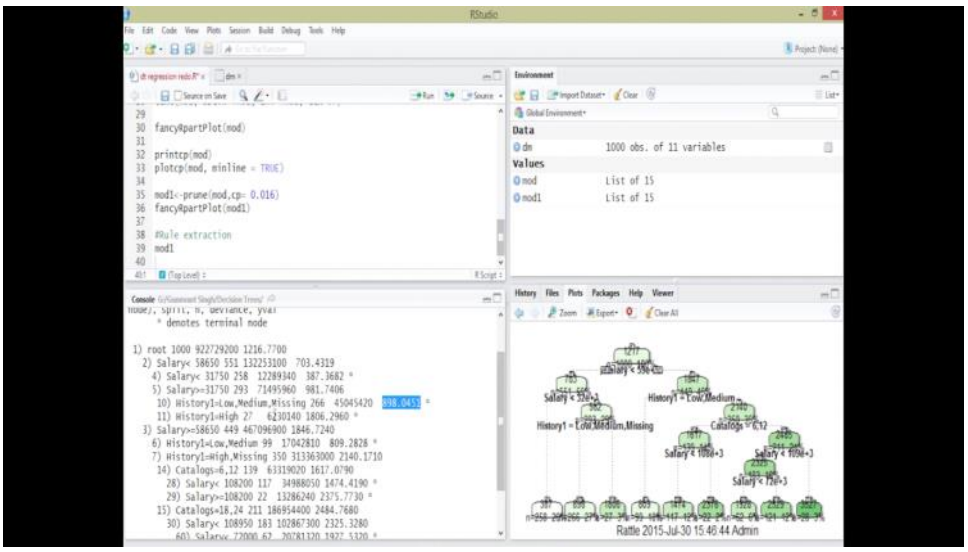


Since this is a big tree we will need to prune it. To do that we will need to have a plot of cost complexity and minimum error. To do that we will use plotCP command

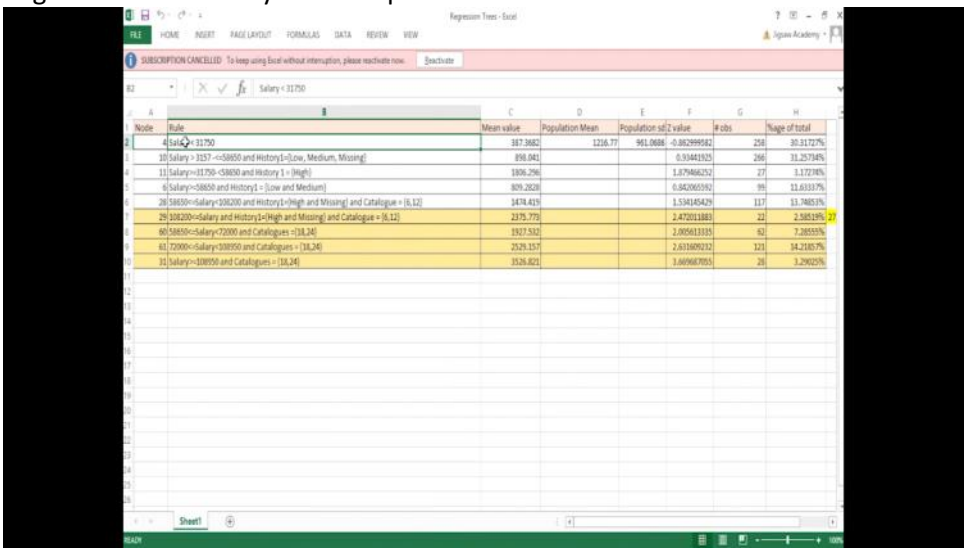


Using this plot we will prune the tree for a value of 0.016. that's where the line intersect

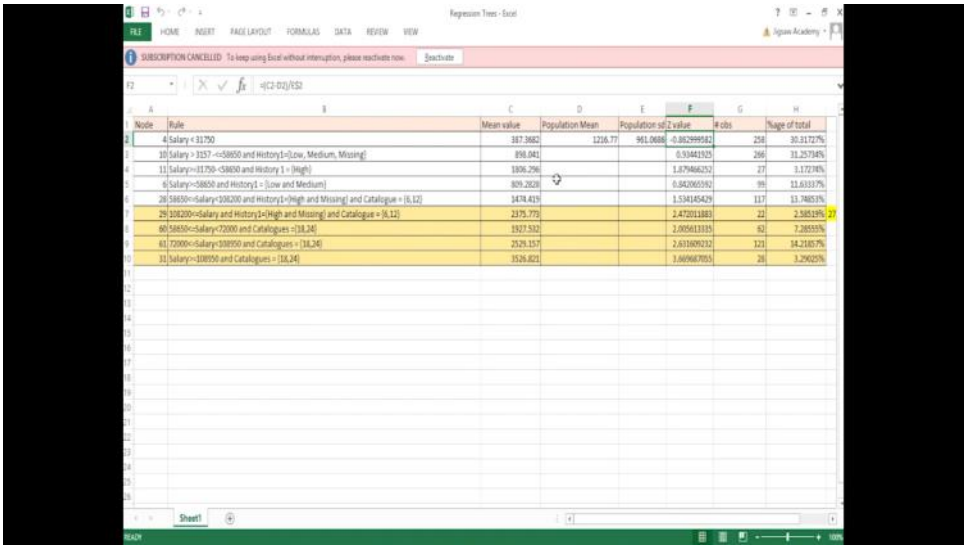




Organize the summary of this otput in an excel sheet.



Calculate the means, population mean, std dev , z values



Node	Rule	Mean value	Population Mean	Population sd	Z value	# obs	%age of total
4	Salary < 31750	387.3682	1216.77	961.0686	-0.862999582	258	30.31727%
10	Salary > 3157 -<=58650 and History1=(Low, Medium, Missing)	898.041			0.93441925	266	31.25734%
11	Salary>=31750-58650 and History 1 = (High)	1806.296			1.879466252	27	3.17274%
6	Salary>-58650 and History1 = (Low and Medium)	809.2828			0.842065392	99	11.63337%
28	58650<=Salary<108200 and History1=(High and Missing) and Catalogue = (6,12)	1474.415			1.534145429	117	13.74853%
29	108200<=Salary and History1=(High and Missing) and Catalogue = (6,12)	2375.773			2.472011883	22	2.585195%
60	58650<=Salary<72000 and Catalogues =(18,24)	1927.532			2.005613335	62	7.28555%
61	72000<=Salary<108950 and Catalogues = (18,24)	2529.157			2.631809232	123	14.21857%
31	Salary>=108950 and Catalogues = (18,24)	3526.821			3.669687055	28	3.29025%

The excel formula should be (C2-D$2)/E$2 for the computation of Z values.

PreviousNext

From <https://jigsawacademy.net/courses/119/pages/video/t11-dot-2-19-r-code-demo-part-3-00-13-02?module_item_id=9670>