**Assignment 07**

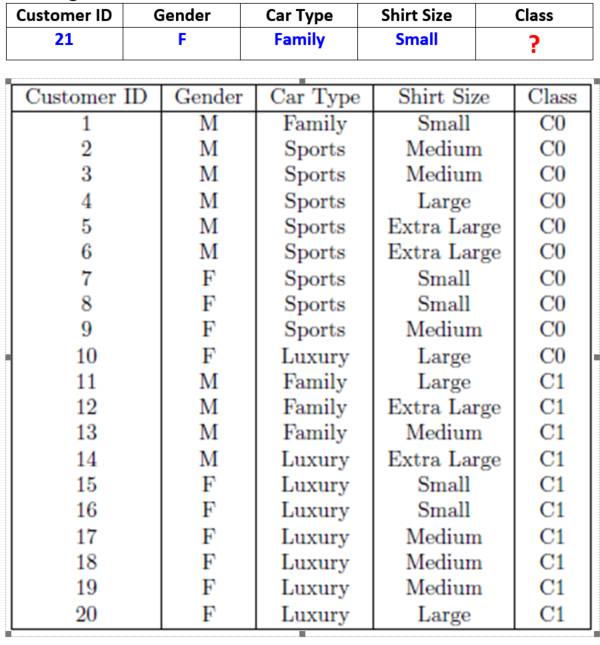
Decision Tree Induction

**Name: Dean D’souza**

**H.U. ID: 168424**

# **Solutions**:

**You are employed as a data mining consultant for some company. Given the following data table, develop a decision tree based solution that can predict the right class for the following customer:**

****

For the first attempt at building the Decision Tree, we follow the method outlined in the lecture (which uses the concept of entropy). Hence, our first step is to calculate the entropy for each variable in the given table, which is done as follows:

**For Customer ID:**

While we could show the calculations for the same, it would not be in our best interests at the moment as we know that this variable is a unique identifier for each case and so we can immediately say that it would not contribute to building the Decision Tree.

**For ‘Gender’:**

As this variable gives the gender of each applicant and has values of ‘M’ and ‘F’ we perform the entropy calculation as follows:

**Before split:** 20 records, 10 C0, 10 C1

* H(S) = -(10/20 log2 10/20 + 10/20 log2 10/20) = 1.0

‘Gender’ = ‘M’: 10 records, 6 C0, 4 C1

* H(S) = -(6/10 log2 6/10 + 4/10 log2 4/10) = -(-0.44-0.53) = 0.97

‘Gender’ = ‘F’: 10 records, 4 C0, 6 C1

* H(S) = -(4/10 log2 4/10 + 6/10 log2 6/10) = -(-0.53-0.44) = 0.97

Expected new Entropy for attribute ‘Gender’:

* 10/20 \* 0.97 + 10/20 \* 0.97 = 0.485 +0.485 = 0.97

**Information Gain = 1.0 – 0.97 = 0.03**

**For ‘Car Type’:**

As this variable gives the type of car driven by each applicant and has values of ‘Sports’, ‘Luxury’ and ‘Family’ we perform the entropy calculation as follows:

**Before split:** 20 records, 10 C0, 10 C1

* H(S) = -(10/20 log2 10/20 + 10/20 log2 10/20) = 1.0

‘Car Type’ = ‘Sports’: 8 records, 8 C0, 0 C1

* H(S) = -(8/8 log2 8/8 + 0/8 log2 0/8) = 0

‘Car Type’ = ‘Luxury’: 8 records, 1 C0, 7 C1

* H(S) = -(1/8 log2 1/8 + 7/8 log2 7/8) = -(-0.38-0.17) = 0.55

‘Car Type’ = ‘Family’: 4 records, 1 C0, 3 C1

* H(S) = -(1/4 log2 1/4 + 3/4 log2 3/4) = -(-0.5-0.31) = 0.81

Expected new Entropy for attribute ‘Car Type’:

* 8/20 \* 0 + 8/20 \* 0.55 + 4/20 \* 0.81 = 0 + 0.22 +0.162 = 0.382

**Information Gain = 1.0 – 0.382 = 0.618**

**For ‘Shirt Size’:**

As this variable gives the size of the shirt worn by each applicant and has values of ‘Small’, ‘Medium’, ‘Large’ and ‘Extra Large’ we perform the entropy calculation as follows:

**Before split:** 20 records, 10 C0, 10 C1

* H(S) = -(10/20 log2 10/20 + 10/20 log2 10/20) = 1.0

‘Shirt Size’ = ‘Small’: 5 records, 3 C0, 2 C1

* H(S) = -(3/5 log2 3/5 + 2/5 log2 2/5) = -(-0.44-0.53) = 0.97

‘Shirt Size’ = ‘Medium’: 7 records, 3 C0, 4 C1

* H(S) = -(3/7 log2 3/7 + 4/7 log2 4/7) = -(-0.52-0.46) = 0.98

‘Shirt Size’ = ‘Large’: 4 records, 2 C0, 2 C1

* H(S) = -(2/4 log2 2/4 + 2/4 log2 2/4) = -(-0.5-0.5) = 1.0

‘Shirt Size’ = ‘Extra Large’: 4 records, 2 C0, 2 C1

* H(S) = -(2/4 log2 2/4 + 2/4 log2 2/4) = -(-0.5-0.5) = 1.0

Expected new Entropy for attribute ‘Shirt Size’:

* 5/20 \* 0 .97+ 7/20 \* 0.98 + 4/20 \* 1.0 + 4/20 \* 1.0 = 0.243 + 0.343 +0.2 + 0.2 = 0.986

**Information Gain = 1.0 – 0.986 = 0.014**

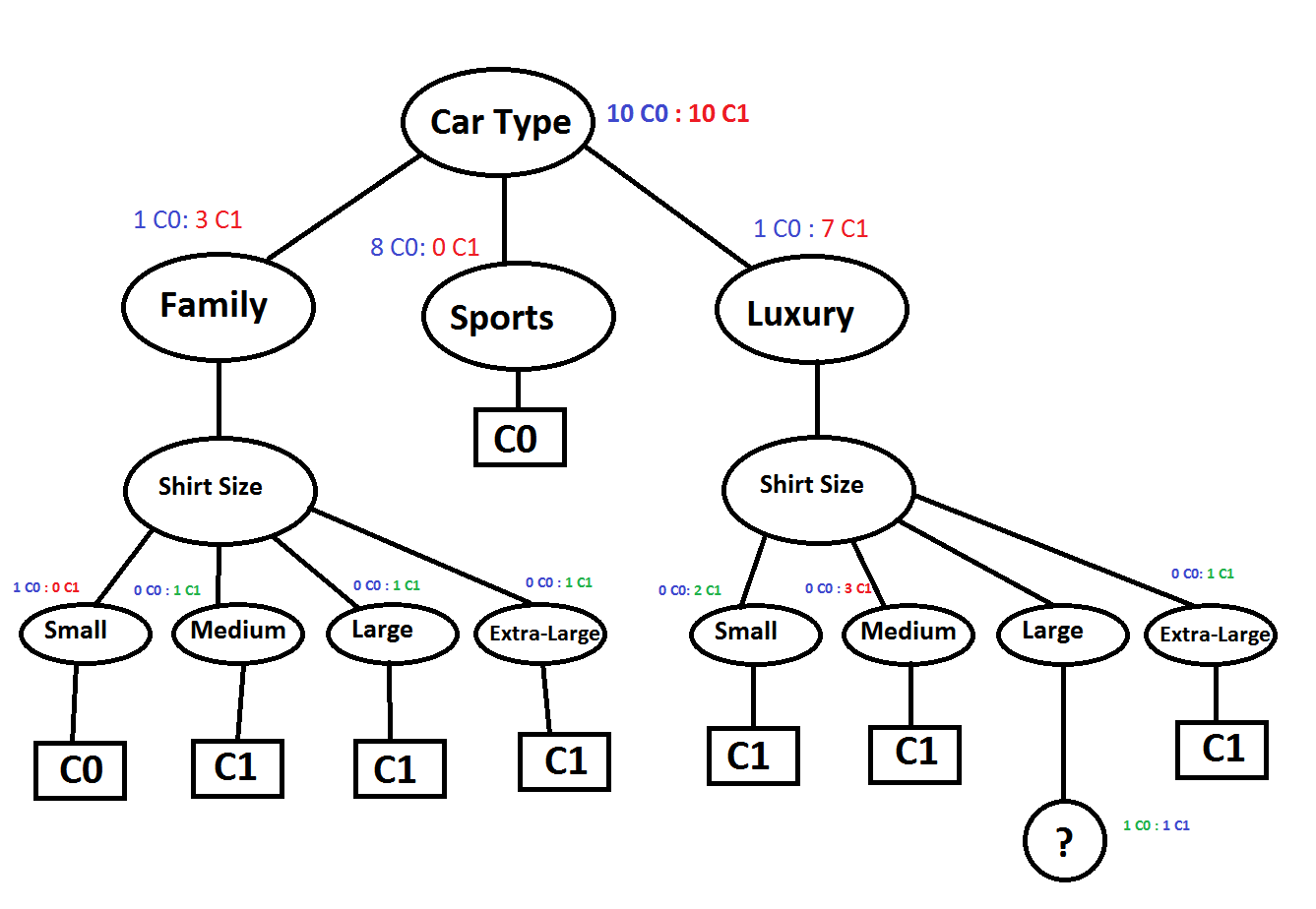
Based on the above calculations, we arrange the attributes in the following order to determine how to construct the Decision Tree:

IG ‘Car Type’ = 0.618

IG ‘Gender’ = 0.03

IG ‘Shirt Size’= 0.014

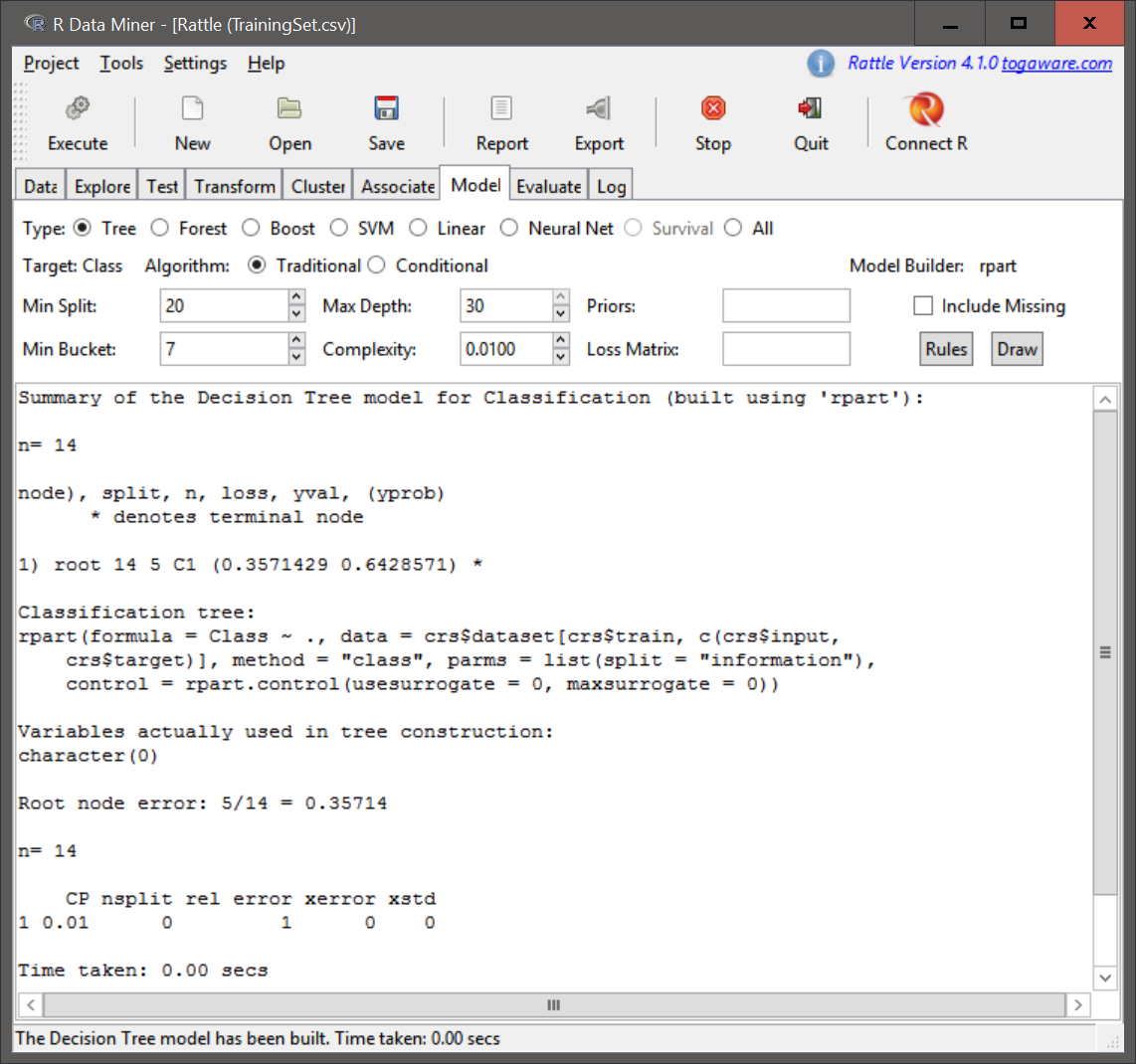
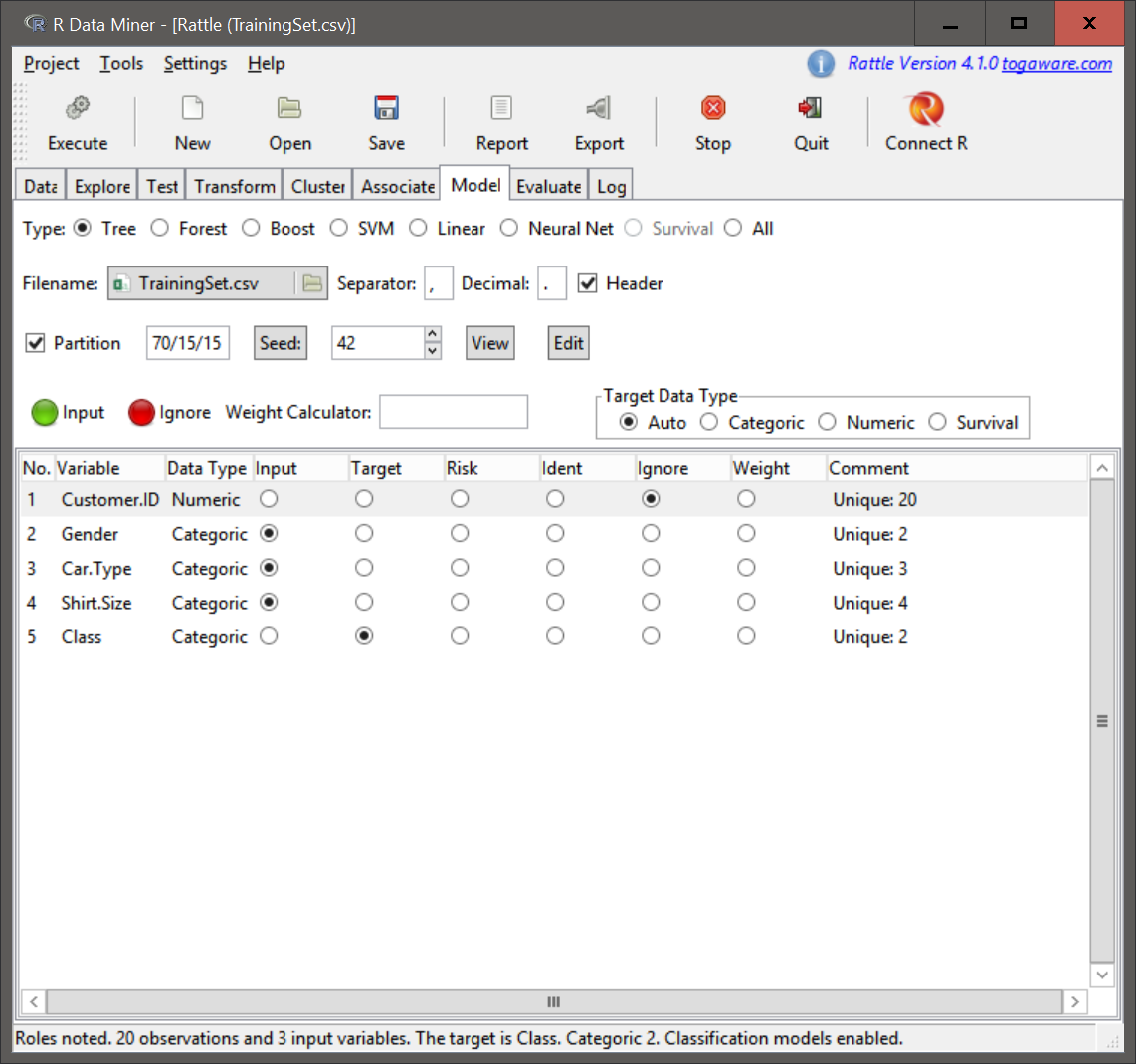
We now construct the decision tree as follows:



In the above decision tree, there were two records which led to an inability to further split, which is marked by a ‘?’ in the circle (or node), as both records had luxury cars with female applicants who wear Large shirts. However, we can still classify the test record or new record. Based on the Decision Tree above the new record can be classified as ‘C0’.

One may also ask as to why further splitting was not done using the gender attribute (which has higher information gain), to which the answer is that for further splitting the gender attribute of all remaining records of ‘Family’ car type was male while those of ‘Luxury’ car type was female.

We now create a decision tree using Rattle GUI through R studio as follows:



As can be seen in the above screenshots, a Decision tree with only one node was created by the traditional algorithm (also using conditional algorithm which is not shown to save space), such a decision tree would immediately classify the new record as ‘C1’ which would be most likely wrong.

**Hence, we conclude that out initial decision tree is sufficiently accurate, and the classification of the new record is ‘C0’.**

# References:

[1] “Decision Tree Induction” Lecture Slides, Majid Shaalan, PhD., Harrisburg University of Science and Technology, CISC-520, Fall 2016

[2] “Decision Trees & Rattle” Lecture Slides, Stephen Penn, DM, PMP, Harrisburg University of Science and Technology, ANLY-510, Summer 2016