**Assignment 05**

Hands-on: Feature Selection

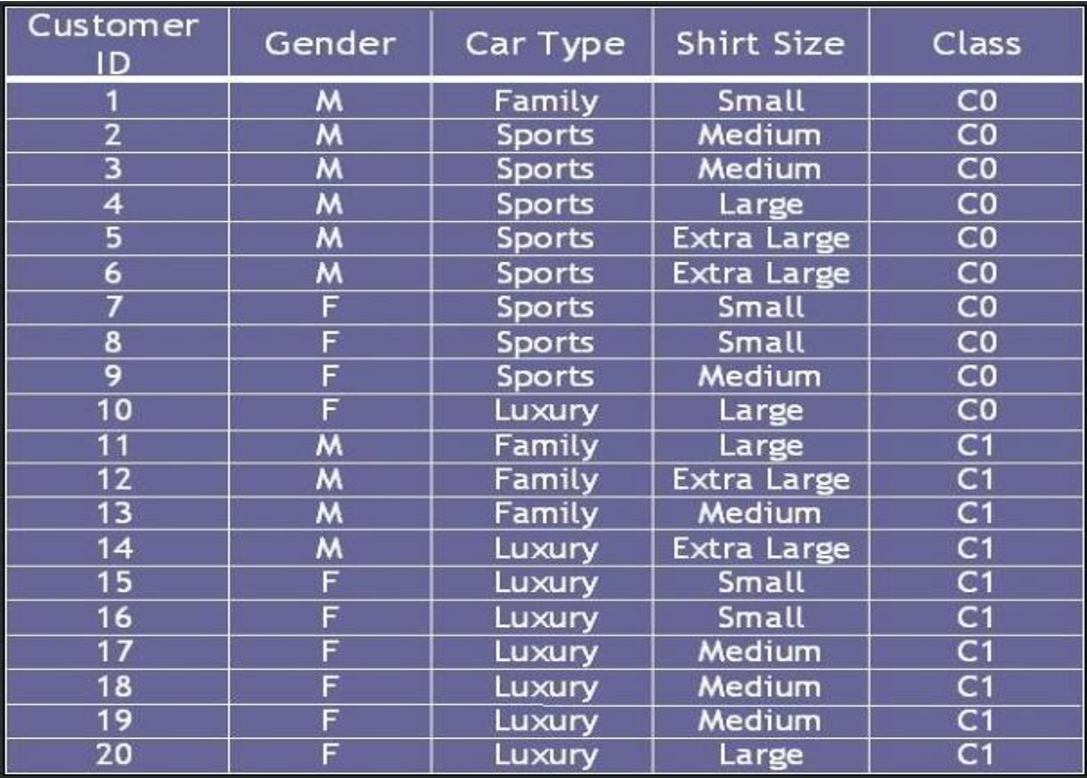
**Name: Dean D’souza**

**H.U. ID: 168424**

# Solutions:

For this Assignment, I chose to perform the Hands-on of Feature Selection in the lecture slides (page 92).

**Consider the training example shown in the table below:**



**A) Compute the Gini index for the overall collection of the training examples.**

For this question I assume we are calculating the Gini index of the system, so we will look at the estimated variable, which is “class”, and calculate it as follows:

Gini(Class) = 1 – [] = 1 – [] = = 0.5

**B) Compute the Gini index for the Customer ID attribute.**

We compute the Gini index of Customer ID as follows:

Gini(Customer ID:1) = 1 – [] = 0

This will be the case for each Customer ID, so the overall gini index for Customer ID is 0.

**C) Compute the Gini index for the Gender attribute.**

We compute the Gini Index of Gender as follows:

Gini(Gender:female) = 1- [] = 1 - [0.16+0.36] = 0.48

Gini(Gender:male) = 1- [] = 1 - [0.36+0.16] = 0.48

Gini(Gender) = [[] + []] = 0.48

The overall Gini index of Gender is 0.48.

**D) Compute the Gini index for the car type attribute using multiway split.**

Using multiway split, we compute the Gini Index for Car Type as follows:

Gini(Car Type: family) = 1 - [] = 1 – [0.0625+0.5625] = 0.375

Gini(Car Type: sports) = 1 – [] = 0

Gini(Car Type: luxury) = 1 – [] = 1 – [0.0156+0.7656] = 0.21875

Gini(Car Type) = [[]+ []+ []] = 0.163

The overall Gini index of Car Type is 0.163.

**E) Compute the Gini index for the Shirt Size attribute using multiway split.**

Using multiway split, we compute the Gini index for Shirt Size as follows:

Gini(Shirt Size: Small) = 1 - [] = 1 - [0.36+0.16] = 0.48

Gini(Shirt Size: Medium) = 1 - [] = 1 - [0.184+0.327] = 0.49

Gini(Shirt Size: Large) = 1 - [] = 1 – [0.25+0.25] = 0.5

Gini(Shirt Size: Extra Large) = 1 - [] = 1 – [0.25+0.25] = 0.5

Gini(Shirt Size) = [[]+ [] + [] + []] = 0.4915

The overall Gini index of Shirt size is 0.4915.

**F) Which attribute is better, Gender, Car Type, or Shirt Size?**

Since we usually consider an attribute with a lower Gini index to be better, Car Type has the lowest overall Gini index and is hence the best attribute or feature for classification in this case.

**G) Explain why customer ID should not be used as the attribute test condition even though it has the lowest Gini.**

Customer ID has a Gini index of 0 as it is unique for each record and hence does not contribute any information to the classifier. Hence it should not be used as the attribute test condition even though it has the “lowest” Gini index.