Case Study: Feature Selection Process in R

Introduction:

In this section, we illustrate the feature selection process in R Programming Language.

R is a free software programming language and a software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

Why R? Because it is an Open Source, it is an interpreted language; users typically access it through a command-line interpreter, fast execution and cost-effective.

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.

Method:

Regressions subset selection: In this considers all possible subsets of the pool of explanatory variables and finds the model that best fits the data according to some criteria (e.g. Adjusted R^2, AIC and BIC). These criteria assign scores to each model and allow us to choose the model with the best score.

Scenario:

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Input Data:

A data frame mtcars with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (lb/1000)
[, 7] qsec 1/4 mile time
[, 8] vs V/S

[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburettors

Execution:

```
>install.packages("leaps")
>leaps=regsubsets(am~mpg+cyl+disp+hp+drat +wt+qsec+vs+gear+carb,data=mtcars,
nbest=10)
#To view the ranked models according to the adjusted R-squared criteria and #BIC,
#respectively, type
> plot(leaps, scale="adjr2")
> plot(leaps, scale="bic")
#The R function step() can be used to perform variable selection. To perform forward
#selection we need to begin by specifying a starting model and the range of models which
#we want to examine in the search.
#Starting Model
>null=glm(am~1, family="binomial", data=mtcars)
#Range Model
>full=glm(am~mpg+cyl+disp+hp+drat +wt+qsec+vs+gear+carb,family="binomial"
,data=mtcars)
#We can perform forward selection using the command:
> step(null, scope=list(lower=null, upper=full), direction="forward")
```

Output:

```
Start:  AIC=45.23
am ~ 1
 Df Deviance    AIC
+ gear  1   15.276 19.276
+ wt    1   19.176 23.176
+ drat  1   21.650 25.650
+ mpg   1   29.675 33.675
+ disp  1   29.732 33.732
+ cyl   1   33.951 37.951
+ hp    1   41.228 45.228
<none>      43.230 45.230
+ qsec  1   41.465 45.465
+ vs    1   42.323 46.323
+ carb  1   43.124 47.124
Step:  AIC=19.28
am ~ gear
```

```
Df Deviance    AIC
+ wt    1   0.0000  6.000
+ disp  1   9.5137 15.514
+ drat  1  10.5354 16.535
+ mpg   1  11.6587 17.659
+ qsec  1  12.1454 18.145
<none>     15.2763 19.276
+ vs    1  13.4602 19.460
+ carb  1  13.5452 19.545
+ hp    1  13.9898 19.990
+ cyl   1  14.5425 20.543

Step:  AIC=6
am ~ gear + wt
 Df   Deviance AIC
<none>    5.7611e-09   6
+ qsec  1 1.8197e-09   8
+ hp    1 2.6852e-09   8
+ mpg   1 2.7646e-09   8
+ disp  1 2.8158e-09   8
+ cyl   1 2.9620e-09   8
+ vs    1 3.0594e-09   8
+ carb  1 3.3481e-09   8
+ drat  1 5.1960e-09   8

Call:  glm(formula = am ~ gear + wt, family = "binomial", data = mtcars)
Coefficients:
(Intercept)        gear          wt
     24.98      105.57     -148.47
Degrees of Freedom: 31 Total (i.e. Null);  29 Residual
Null Deviance:      43.23
Residual Deviance: 5.761e-09     AIC: 6
```

Analysis:

This tells R to start with the null model and search through models lying in the range between the null and full model using the forward selection algorithm .At last (Call :) containing some independent variable i.e. our selected variables while considering am as dependent variable.