SUPPORT VECTOR MACHINES

Assignment 1

BP: Prepare a classification model using SVM for salary data

**PROCEDURE:**

STEP 1: First we have to Exploratory Data Analysis which can be done by plotting scattered plot, box plots and summary.

age workclass education educationno

Min. :17.00 Length:30161 Length:30161 Min. : 1.00

1st Qu.:28.00 Class :character Class :character 1st Qu.: 9.00

Median :37.00 Mode :character Mode :character Median :10.00

Mean :38.44 Mean :10.12

3rd Qu.:47.00 3rd Qu.:13.00

Max. :90.00 Max. :16.00

maritalstatus occupation relationship race

Length:30161 Length:30161 Length:30161 Length:30161

Class :character Class :character Class :character Class :character

Mode :character Mode :character Mode :character Mode :character

sex capitalgain capitalloss hoursperweek

Length:30161 Min. : 0 Min. : 0.0 Min. : 1.00

Class :character 1st Qu.: 0 1st Qu.: 0.0 1st Qu.:40.00

Mode :character Median : 0 Median : 0.0 Median :40.00

Mean : 1092 Mean : 88.3 Mean :40.93

3rd Qu.: 0 3rd Qu.: 0.0 3rd Qu.:45.00

Max. :99999 Max. :4356.0 Max. :99.00

native Salary

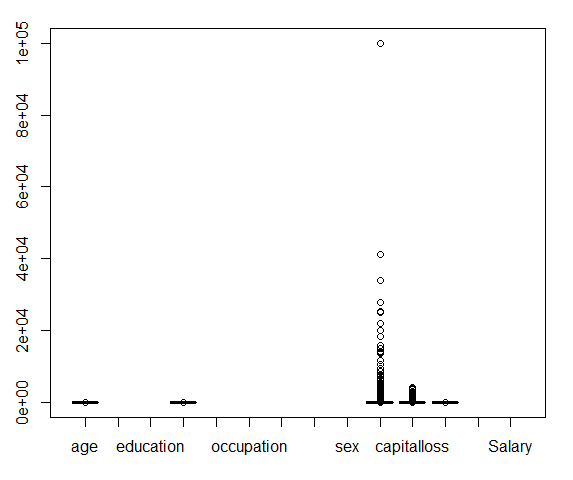
Length:30161 Length:30161

Class :character Class :character

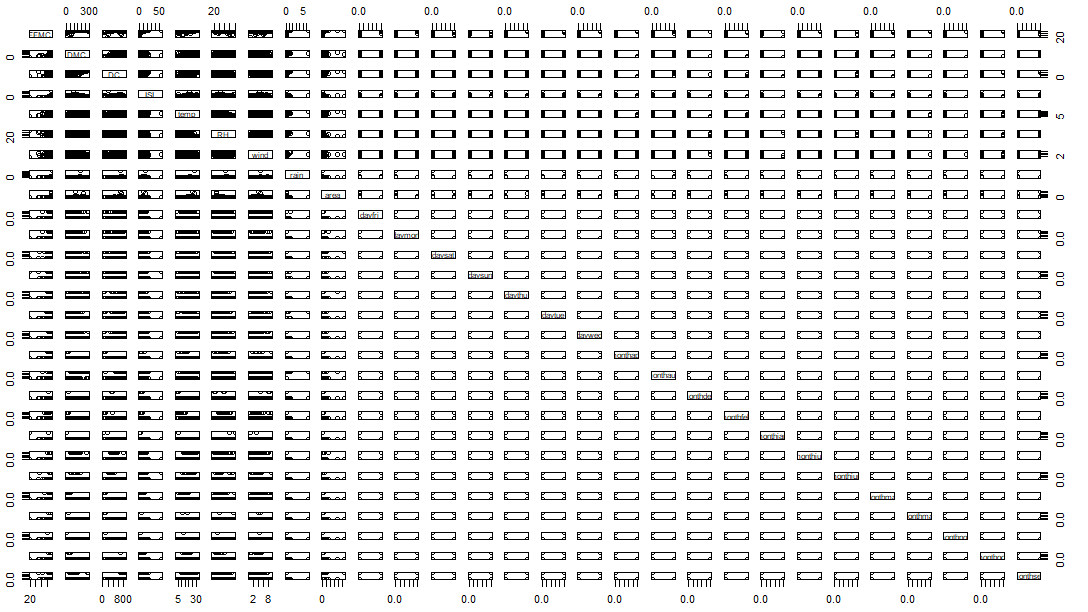
Mode :character Mode :character

From the summary we can see that in capital gain and capital loss the difference between mean and max is large so it may be right skewed it can my confirmed by using Box Plot .

It can be confirmed that capital gain and capital loss is right Skewed.



Scatter Plot for the Following data frame is:

  
STEP 2: Predicting the data using SVM techniques

|  |
| --- |
|  |
| |  | | --- | |  | |

|  |
| --- |
|  |
|  |
| |  | | --- | |  | |

Before building the model we need to install some packages required for SVM formation:

|  |
| --- |
| install.packages("kernlab") |
| library(kernlab) |

Now after splitting the data into test and train , we can Build model on training data which is given by ( here we use vanilla dot mode ):

salary\_model <- ksvm(Salary~ ., data=SalaryData\_Train\_1\_ , kernel="vanilladot")

Now predicting the model using some kernel techniques:

salary\_prediction <- predict(salary\_model ,SalaryData\_Test\_1\_)

>

>

> head(salary\_prediction)

[1] <=50K <=50K <=50K >50K <=50K >50K

Levels: <=50K >50K

table(salary\_prediction , SalaryData\_Test\_1\_$Salary)

salary\_prediction <=50K >50K

<=50K 10599 1554

>50K 761 2146

agreement <- salary\_prediction == SalaryData\_Test\_1\_$Salary

>

>

> table(agreement)

agreement

FALSE TRUE

2315 12745

> prop.table(table(agreement))

agreement

FALSE TRUE

0.1537185 0.8462815

The Above prediction shows 0.84 % of true results , we can check with other models to improve accuracy .  
  
polydot model :

salary\_model2 <- ksvm(Salary ~ ., data=SalaryData\_Train\_1\_ , kernel= "polydot" , parallel= TRUE)

Setting default kernel parameters

> salary\_prediction2 <- predict(salary\_model2 , SalaryData\_Test\_1\_)

>

>

>

> salary\_prediction2 <- predict(salary\_model2 , SalaryData\_Test\_1\_)

>

>

>

> head(salary\_prediction2)

[1] <=50K <=50K <=50K >50K <=50K >50K

Levels: <=50K >50K

>

>

> table(salary\_prediction2 , SalaryData\_Test\_1\_$Salary)

salary\_prediction2 <=50K >50K

<=50K 10598 1553

>50K 762 2147

> agreement <- salary\_prediction2 == SalaryData\_Test\_1\_$Salary

> table(agreement)

agreement

FALSE TRUE

2315 12745

> agreement2 <- salary\_prediction2 == SalaryData\_Test\_1\_$Salary

> table(agreement2)

agreement2

FALSE TRUE

2315 12745

> prop.table(table(agreement2))

agreement2

FALSE TRUE

0.1537185 0.8462815

Seems like there is no change in output so better to consider 1st model as we cannot afford the same efficiency with so much increase in time period.