**Naïve Bayes**

**Classification Model For Salary Train and Test Dataset**

**Structure of data**

data.frame': 30161 obs. of 14 variables:

$ age : int 39 50 38 53 28 37 49 52 31 42 ...

$ workclass : Factor w/ 7 levels " Federal-gov",..: 6 5 3 3 3 3 3 5 3 3 ...

$ education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...

$ educationno : int 13 13 9 7 13 14 5 9 14 13 ...

$ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...

$ occupation : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...

$ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...

$ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...

$ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...

$ capitalgain : int 2174 0 0 0 0 0 0 0 14084 5178 ...

$ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...

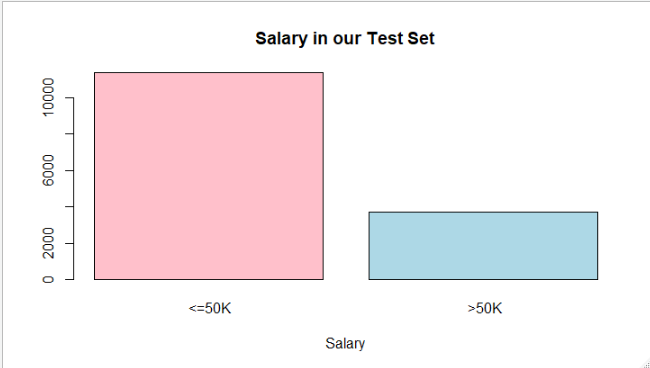
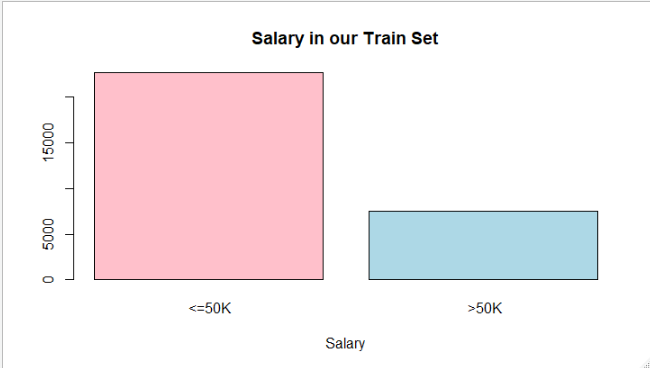
$ hoursperweek : int 40 13 40 40 40 40 16 45 50 40 ...

$ native : Factor w/ 40 levels " Cambodia"," Canada",..: 38 38 38 38 5 38 22 38 38 38 ...

$ Salary : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ..

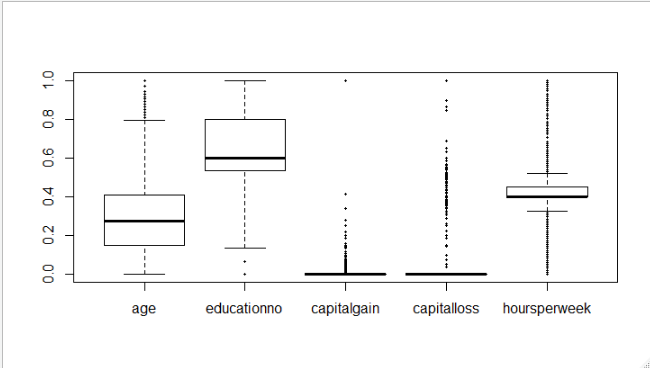
**From the above structure it is seen that, data contains 9 factor and 5 numeric**

**variables, so we have to create dummy variable to normalize the data.**



**From the above boxplot, data is imbalanced.**

**Boxplot of Numerical Variables In Test Dataset After Normalization**



From the above boxplot, lots of outlier are in the data but we are not going to remove them as we may lose a lot of information.

So we will create model with normalized dummy data.

**Model without Laplace smoothing 🡺**

**summary(model\_1)**

**Length Class Mode**

**apriori 2 table numeric**

**tables 102 -none- list**

**levels 2 -none- character**

**isnumeric 102 -none- logical**

**call 3 -none- call**

**Accuracy 🡺 0.7837317**

**Confusion Matrix**

**Predicted**

**Actual <=50K >50K**

**<=50K 10753 607**

**>50K 2650 1050**

**Model with Laplace smoothing 🡺**

**summary(model\_2)**

**Length Class Mode**

**apriori 2 table numeric**

**tables 102 -none- list**

**levels 2 -none- character**

**isnumeric 102 -none- logical**

**call 4 -none- call**

**Accuracy 🡺 0.7837317**

**Confusion Matrix**

**Predicted**

**Actual <=50K >50K**

**<=50K 10753 607**

**>50K 2650 1050**

**Model without laplace and with laplace smoothing giving same results.**