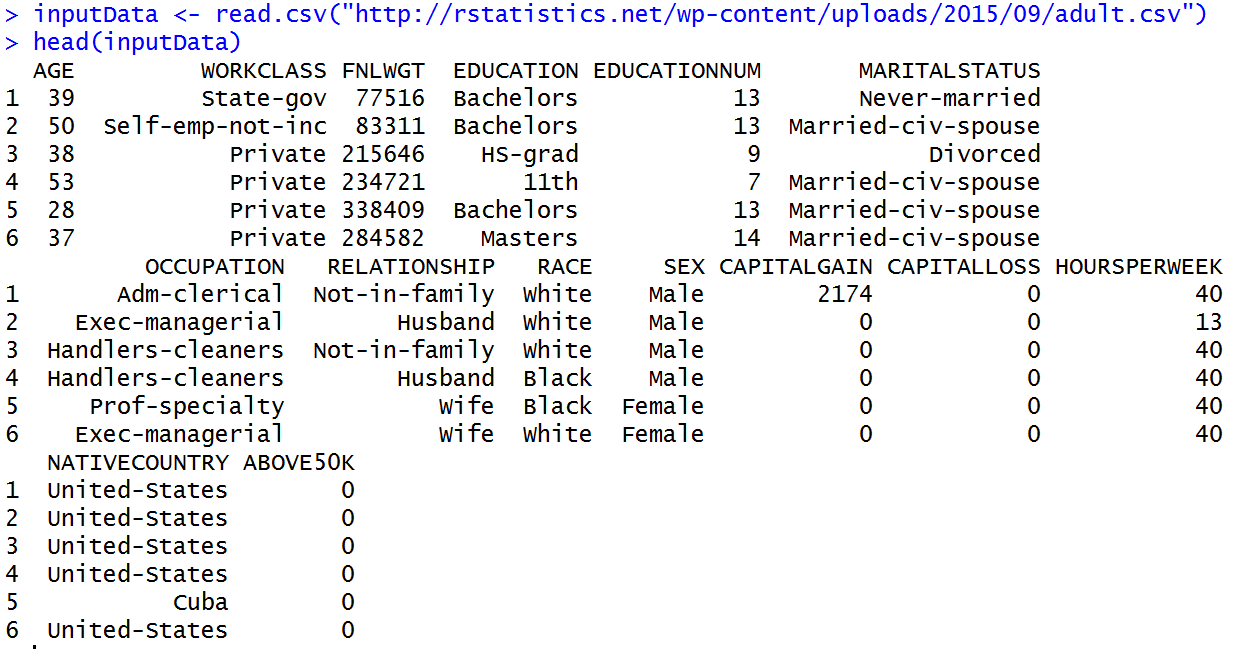
**Module 7: Additional Exercises with Answers**

Predicting Annual Income of Individuals

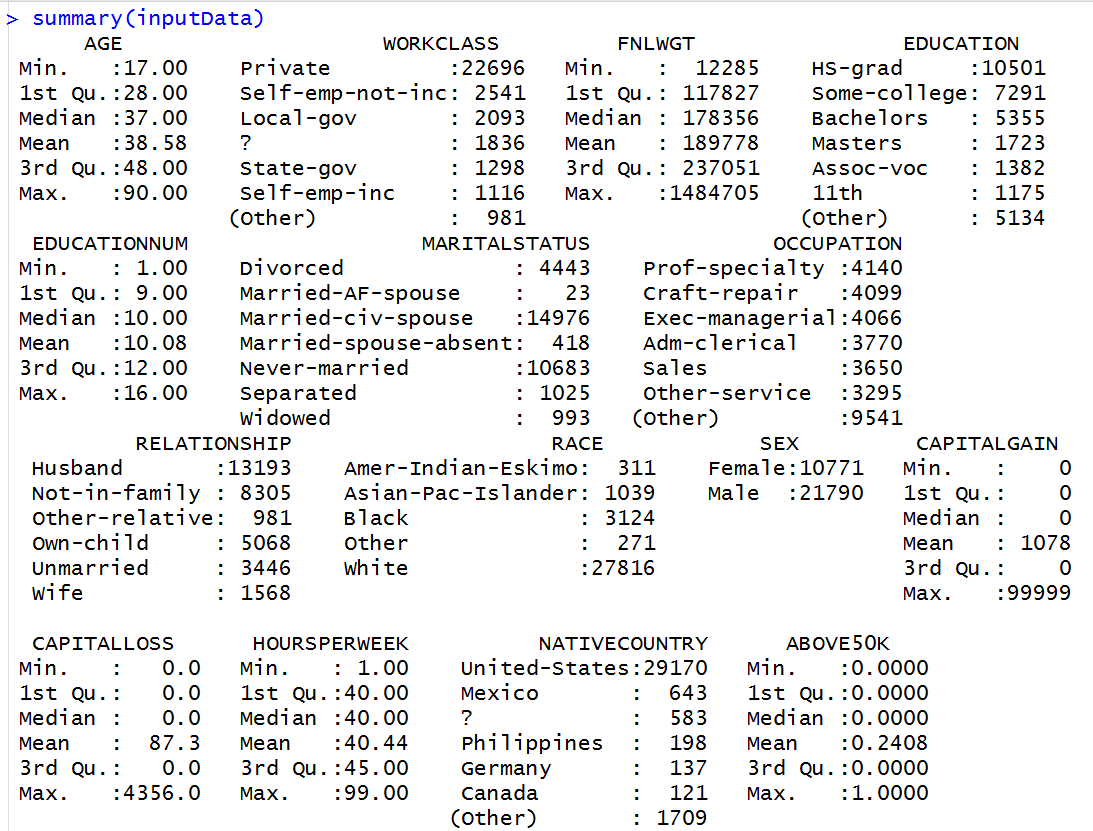
In this example, we will use logistic regression to predict if an individual earns more than $50k in a year or not. First read the ‘Adult’ CSV file using the following command (you need to be connected to the internet and it will take few moments for data to be loaded):

**inputData <- read.csv("http://rstatistics.net/wp-content/uploads/2015/09/adult.csv")**

Let’s have a look at the first 6 records and also a summary of the dataset



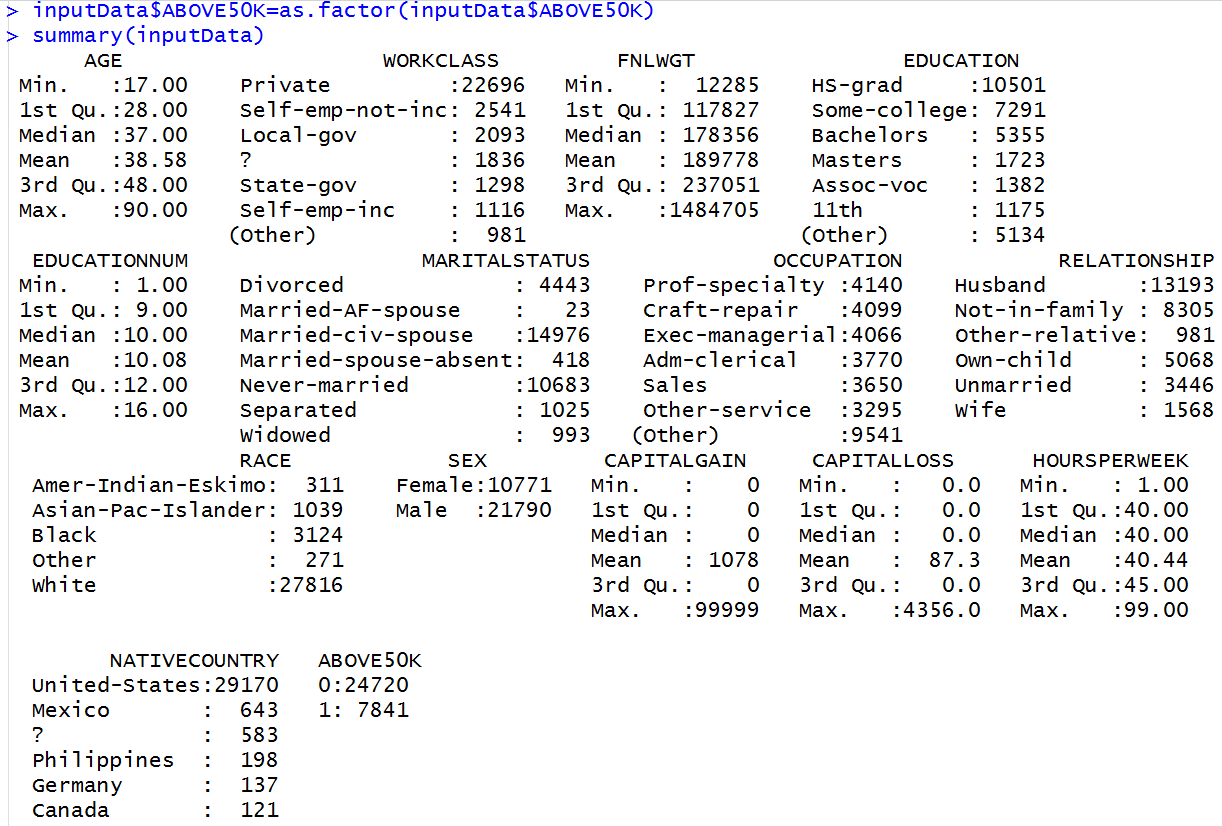
And



The variable ‘ABOVE50K’ is the variable that we are trying to predict. Currently, the variable is coded as a numeric variable that takes 0 and 1s. To use logistic regression, that is a classification method, we need to convert this variable to a factor (i.e. categorical variable):

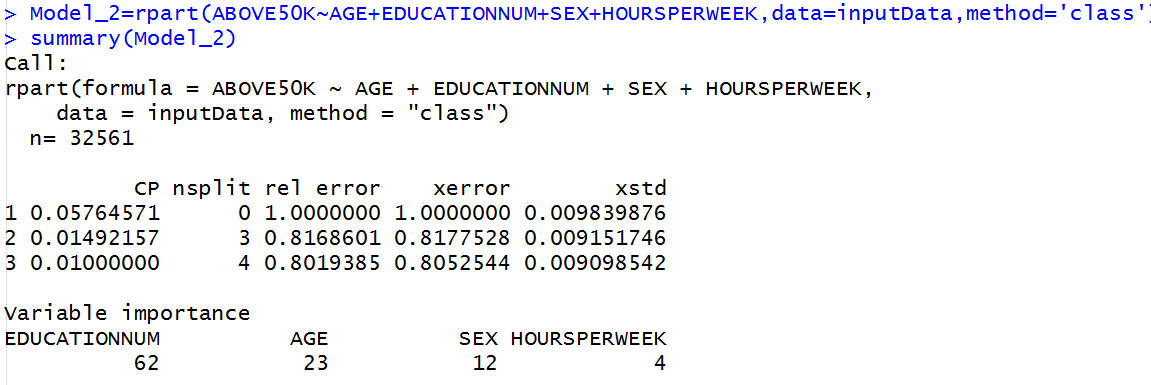
inputData$ABOVE50K=as.factor(inputData$ABOVE50K)

let’s look at the summary again:



**Questions**

**Q1. Build a decision trees model based on AGE, EDUCATIONNUM (Number of years of Education), SEX and HOURSPERWEEK (number of hours worked per week) to predict ABOVE50K.**



**Q2. How do we judge the statistical significance and the importance of variables in decision tree models?**

In decision tree models, we do not have coefficients for variables so we cannot use z-test or t-test to check the importance of variables. The variable importance field gives a measure of importance and significance of variables at the same time. The variable importance values are usually normalized so that the sum of all variable importance to be 100. In this example, EDUCATIONNUM, that is the number of years of education, is by far the most important variable followed by AGE and SEX and finally by the HOURSPERWEEK, which represents the number of hours worked per week. The same order of variable importance was suggested by the logistic regression model above.

**Q3. How does the probability of earning above $50k changes with these variables?**

We can answer this question by ploting the decision tree model. You can use the plot() function from R-base (no need for additional library), or use fancyRpartPlot() from the ‘rattle’ library which has a nicer presentation.

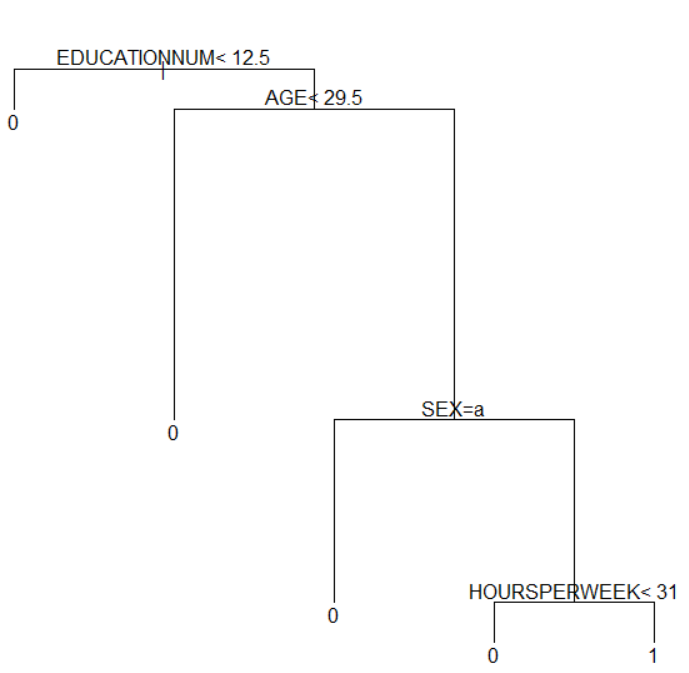
Using the plot() function:

plot(Model\_2) #to plot the tree   
text(Model\_2) #to add labels

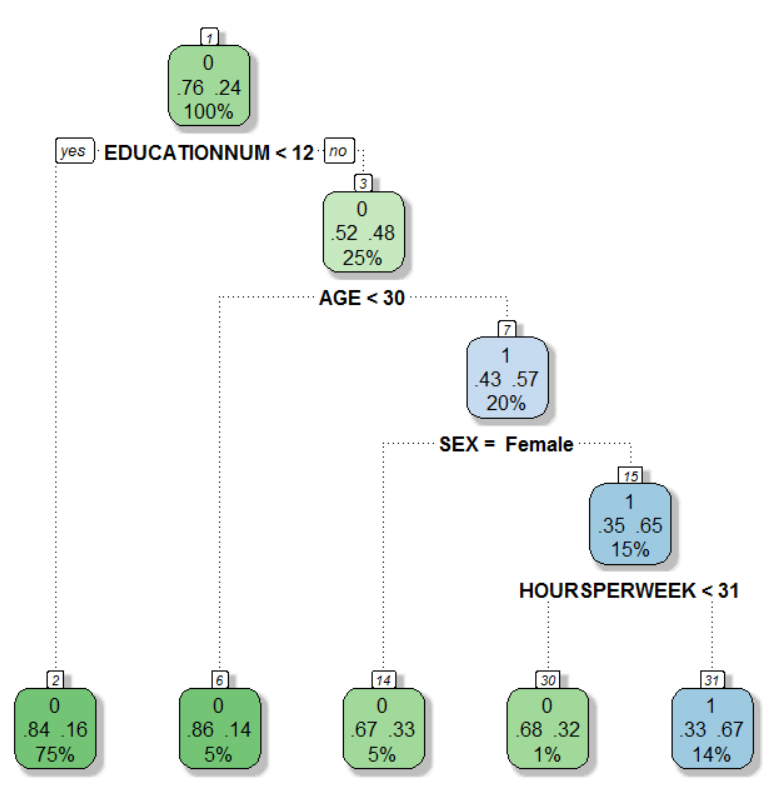
Using the fancyRpartPlot()

library(rattle)  
fancyRpartPlot(Model\_2)

With plot() function



With fancyRpartPlot()



**Q4. James and Hannah are two individuals. Given the information below, what is the probability that each of them is earning more than $50k a year?**

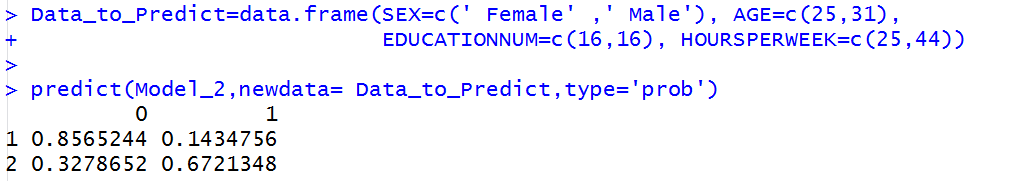
**Hannah: Female, 25 Year, 16 Years of Education, Working part time 25 hours a week   
James: Male, 31 Year, 16 Years of Education, Working 44 hours a week**

We create a new dataset, first record for Hannah and second record for James:

Data\_to\_Predict=data.frame(SEX=c(' Female' ,' Male'), AGE=c(25,31),

EDUCATIONNUM=c(16,16), HOURSPERWEEK=c(25,44))

predict(Model\_2,newdata= Data\_to\_Predict,type='prob')

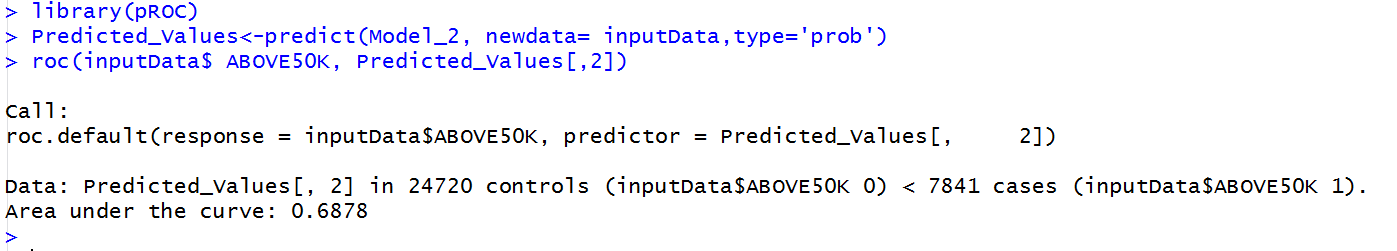


Unlike logistic regression model which provides the probability of the second level class (by default alphabetically), rpart decision tree models gives the probability of each class explicitly. In our example, the probability of 1 (i.e. earning above 50K) is 14.3% for the first observation (i.e. Hannah) and the probability of 1 (again earning abobr50K) for the second observation (i.e. James) is 67.2%. These values are somehow similar to what we got from the logistic regression: 19.4% for Hannah and 66.6% for James.

**Q5. What is the accuracy of this model in terms of Area Under Curve (AUC) of ROC ?**

library(pROC)

Predicted\_Values<-predict(Model\_2, newdata= inputData,type='prob')   
roc(inputData$ ABOVE50K, Predicted\_Values[,2])



We passed Predicted\_Values[,2] to the predict function because the second column of the vector Predicted\_Values contains probabilities for 1 (i.e. income above50K), the first column is the probability for 0.

Comparing the decision tree and the logistic regression model, it is apparent that the logistic regression model was more accurate where the AUC was 0.81.