Updated\_Assignment

Teja

September 15, 2016

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

Analysis of a Direct Marketing campaign of a European Banking institution. The marketing campaigns were conducted through phone calls. Often, more than one contact to the same client was required, in order to access if the product subscribed (bank term deposit) would be ('yes') or ('no').

library(data.table); library(ggplot2); library(plyr); library(rpart);library(rattle);

## Warning: package 'data.table' was built under R version 3.3.1

## Warning: package 'ggplot2' was built under R version 3.3.1

## Warning: package 'plyr' was built under R version 3.3.1

## Warning: package 'rpart' was built under R version 3.3.1

## Warning: package 'rattle' was built under R version 3.3.1

## Rattle: A free graphical interface for data mining with R.  
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

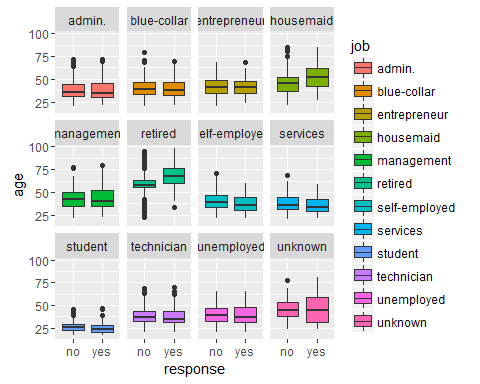
library(rpart.plot); library(RColorBrewer)

## Warning: package 'rpart.plot' was built under R version 3.3.1

myy <- read.table("New.txt", header = TRUE, sep = ";")  
  
m1 <- data.frame(myy)  
  
m1$ID<-seq.int(nrow(m1)) ## Adding a new column ID  
  
m2 <- m1[c(22,1:21)] ## Re-positioning the column ID to the beginning

A brief plot shows the response variation when mapped to multivariable w.r.t distribution of Age and profession of the clients

ggplot(m2, aes(x = response, y = age , fill = job )) + geom\_boxplot() +  
 facet\_wrap(~ job, ncol = 4)



So as to analyse the profession and age group which has the highest no. of subscriptions.

**Logistic Regression**

#Independent variables  
age<- m2$age  
job<- m2$job  
martial<-m2$marital  
education<-m2$education  
default<-m2$default  
housing<-m2$housing  
loan <- m2$loan  
  
# Dependent variable  
response<-m2$response  
  
#Generating labels  
l\_job<-rbind(levels(job))  
l\_martial<-rbind(levels(martial))  
l\_education<-rbind(levels(education))  
l\_default<-rbind(levels(default))  
l\_housing<-rbind(levels(housing))  
l\_loan<-rbind(levels(loan))  
  
  
#Factoring variables  
f.job<-factor(job,labels=l\_job)  
f.martial<-factor(martial,labels=l\_martial)  
f.education<-factor(education,labels=l\_education)  
f.default<-factor(default,labels=l\_default)  
f.housing<-factor(housing,labels=l\_housing)  
f.loan<-factor(loan,labels=l\_loan)  
  
  
  
# Create Contras matrix#  
contrasts(f.job) = contr.treatment(length(levels(job)))  
contrasts(f.martial) = contr.treatment(length(levels(martial)))  
contrasts(f.education) = contr.treatment(length(levels(education)))  
contrasts(f.default) = contr.treatment(length(levels(default)))  
contrasts(f.housing) = contr.treatment(length(levels(housing)))  
contrasts(f.loan) = contr.treatment(length(levels(loan)))  
  
  
  
#Model\_1  
Model\_1 <- glm(response ~ age + f.job +f.martial +f.education +f.default+f.default  
 +f.housing+f.loan, data=m2 ,family = binomial("logit"))  
summary(Model\_1)

##   
## Call:  
## glm(formula = response ~ age + f.job + f.martial + f.education +   
## f.default + f.default + f.housing + f.loan, family = binomial("logit"),   
## data = m2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1088 -0.5252 -0.4523 -0.3302 2.7355   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.711864 0.127037 -21.347 < 2e-16 \*\*\*  
## age 0.014319 0.001925 7.439 1.02e-13 \*\*\*  
## f.job2 -0.380171 0.062567 -6.076 1.23e-09 \*\*\*  
## f.job3 -0.378287 0.099795 -3.791 0.00015 \*\*\*  
## f.job4 -0.128138 0.112300 -1.141 0.25386   
## f.job5 -0.185278 0.067068 -2.763 0.00574 \*\*   
## f.job6 0.714086 0.079687 8.961 < 2e-16 \*\*\*  
## f.job7 -0.192854 0.092621 -2.082 0.03733 \*   
## f.job8 -0.304591 0.069006 -4.414 1.01e-05 \*\*\*  
## f.job9 1.180574 0.085929 13.739 < 2e-16 \*\*\*  
## f.job10 -0.165880 0.055155 -3.008 0.00263 \*\*   
## f.job11 0.261646 0.096673 2.706 0.00680 \*\*   
## f.job12 0.002440 0.183210 0.013 0.98938   
## f.martial2 0.141883 0.054080 2.624 0.00870 \*\*   
## f.martial3 0.422868 0.060882 6.946 3.77e-12 \*\*\*  
## f.martial4 0.429426 0.322763 1.330 0.18336   
## f.education2 0.005580 0.095429 0.058 0.95337   
## f.education3 -0.154831 0.074816 -2.069 0.03850 \*   
## f.education4 -0.043379 0.071794 -0.604 0.54571   
## f.education5 1.018564 0.583695 1.745 0.08098 .   
## f.education6 0.045028 0.079045 0.570 0.56891   
## f.education7 0.176785 0.071216 2.482 0.01305 \*   
## f.education8 0.229515 0.092575 2.479 0.01317 \*   
## f.default2 -0.943032 0.053228 -17.717 < 2e-16 \*\*\*  
## f.default3 -8.613761 68.964693 -0.125 0.90060   
## f.housing2 -0.005653 0.107315 -0.053 0.95799   
## f.housing3 0.056639 0.032340 1.751 0.07989 .   
## f.loan2 NA NA NA NA   
## f.loan3 -0.056417 0.044879 -1.257 0.20872   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28999 on 41187 degrees of freedom  
## Residual deviance: 27659 on 41160 degrees of freedom  
## AIC: 27715  
##   
## Number of Fisher Scoring iterations: 9

After generating a linear model of the dependent variable w.r.t the variables. The accuracy of the model generated is accessed. So as to evaluate the model.

#Accuracy of the Model\_1 after prediction  
in\_d <- data.frame(age,job,martial,education,default,housing,loan, response)  
predicted\_response<-round(predict(Model\_1,in\_d,type="response"))

## Warning: contrasts dropped from factor f.job

## Warning: contrasts dropped from factor f.martial

## Warning: contrasts dropped from factor f.education

## Warning: contrasts dropped from factor f.default

## Warning: contrasts dropped from factor f.housing

## Warning: contrasts dropped from factor f.loan

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

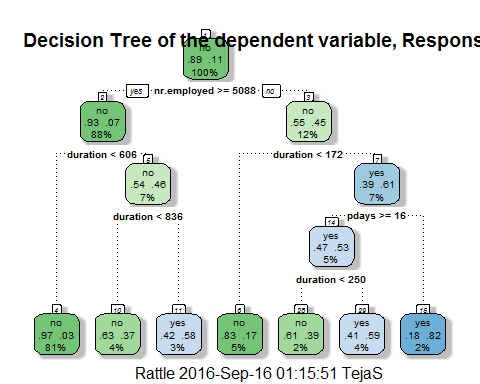
confusion\_matrix <- ftable(response, predicted\_response)  
accuracy <- sum(diag(confusion\_matrix))/41188\*100  
accuracy

## [1] 88.73458

**Accuracy** is found to be 88.7345829.'. Techinically the model can be used to predict the response with that accuracy.

**Using Rpart to make a decision tree**

fit = rpart(response ~ ., data = myy, method = "class")  
  
fancyRpartPlot(fit, main = "Decision Tree of the dependent variable, Response")



The fancy plot shows the decision tree of the response w.r.t other attributes like duration, No. of employees,

**Lets evaluate the accuracy of the decision tree**

pp = predict(fit, myy, type = "class")  
  
st <- data.frame(age = myy$age, response = pp)  
  
ct <- ftable(response, pp)  
ay <- sum(diag(ct))/41188\*100  
ay

## [1] 91.28387

**Accuracy** is found to be 91.2838691.'.

The accuracy is even more than the accuracy of linear model generated. Hence, it is highly recommeded to make sure the attributes lie within the positive decision tree so as to get a positive response.

For previous insights please visit [Gramener.Assigment](https://github.com/STrikEr-TejaS/Gramener_Assignment)