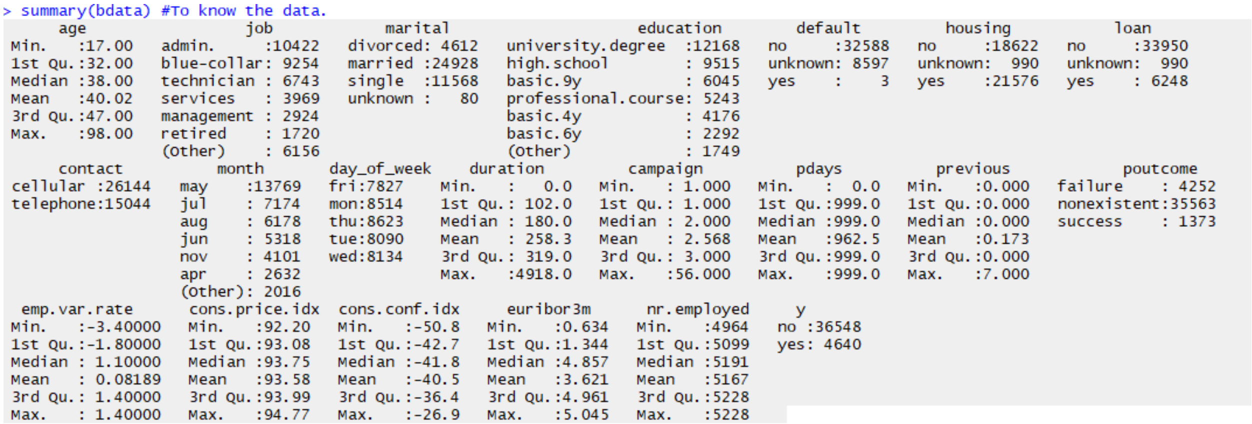
Assignment – Telemarketing to sell long term deposits

Mandeep Kumar 11/29/19 PGCBA -2

Contents

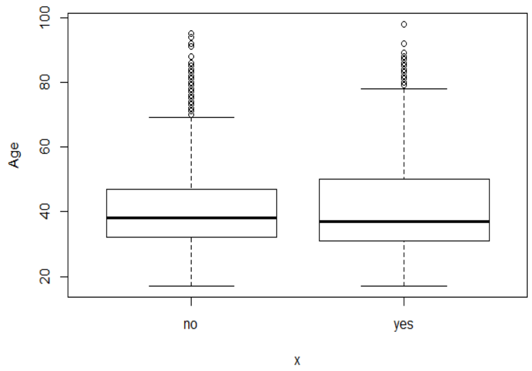
SECTION 1: Data Summary



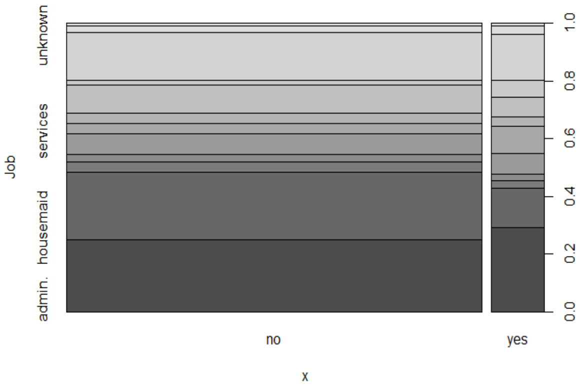
```
> summary(bdata) #To know the data.
      age                 job            marital                    education       default         housing          loan      
 Min.   :17.00   admin.     :10422   divorced: 4612   university.degree  :12168   no     :32588   no     :18622   no     :33950  
 1st Qu.:32.00   blue-collar: 9254   married :24928   high.school        : 9515   unknown: 8597   unknown:  990   unknown:  990  
 Median :38.00   technician : 6743   single  :11568   basic.9y           : 6045   yes    :    3   yes    :21576   yes    : 6248  
 Mean   :40.02   services   : 3969   unknown :   80   professional.course: 5243                                                  
 3rd Qu.:47.00   management : 2924                    basic.4y           : 4176                                                  
 Max.   :98.00   retired    : 1720                    basic.6y           : 2292                                                  
                 (Other)    : 6156                    (Other)            : 1749                                                  
      contact          month       day_of_week    duration         campaign          pdays          previous              poutcome    
 cellular :26144   may    :13769   fri:7827    Min.   :   0.0   Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252  
 telephone:15044   jul    : 7174   mon:8514    1st Qu.: 102.0   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563  
                   aug    : 6178   thu:8623    Median : 180.0   Median : 2.000   Median :999.0   Median :0.000   success    : 1373  
                   jun    : 5318   tue:8090    Mean   : 258.3   Mean   : 2.568   Mean   :962.5   Mean   :0.173                      
                   nov    : 4101   wed:8134    3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000                      
                   apr    : 2632               Max.   :4918.0   Max.   :56.000   Max.   :999.0   Max.   :7.000                      
                   (Other): 2016                                                                                                    
  emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m      nr.employed      y        
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964   no :36548  
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099   yes: 4640  
 Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857   Median :5191              
 Mean   : 0.08189   Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5167              
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228              
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228              
```

SECTION 2: Data Visualization

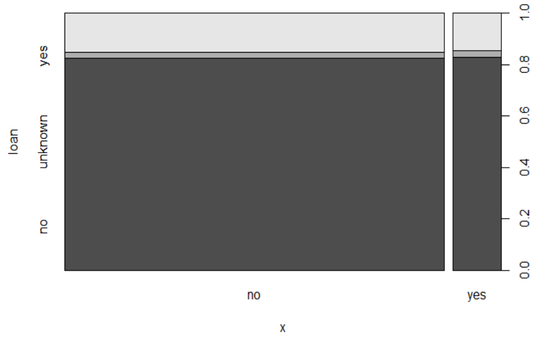
```
plot(bdata$y,bdata$age, ylab="Age")
```



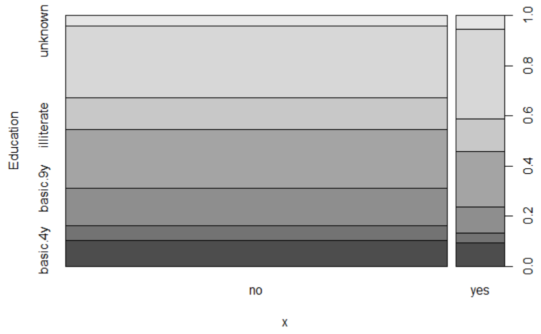
```
plot(bdata$y,bdata$job, ylab="Job")
```



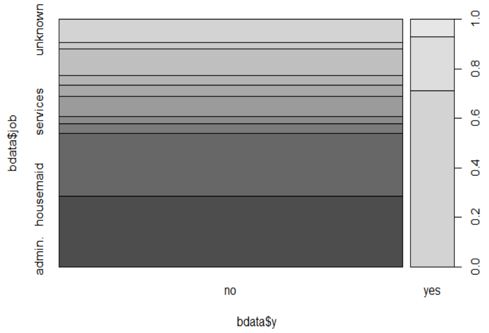
```
plot(bdata$y,bdata$loan, ylab="loan")
```



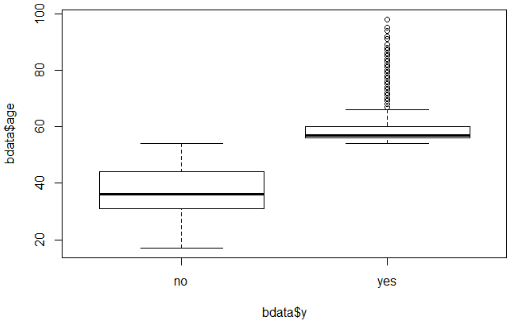
```
plot(bdata$y,bdata$education,ylab="Education" )
```



```
qqplot(bdata$y,bdata$job)
```



```
qqplot(bdata$y,bdata$age)
```



SECTION 3: Logistics Regression

Check for Multi-collinearity

Step 1:

```
model_LR<-glm(y~., data = bdata_trng_final, family = "binomial")

vif(model_LR) # throws error "there are aliased coefficients in t
he model"
Error in vif.default(model_LR) :
  there are aliased coefficients in the model
```

Step 2:

Find the problematic variable. Below line gives the name of the attribute with aliased coefficients

```
> ld_vars <- attributes(alias(model_LR)$Complete)$dimnames[[1]]
> ld_vars
[1] "loanunknown"
```

Step 3:

Remove the predictor variable “loan” and build the model again.

```
> model_LR<-glm(y~.-loan, data = bdata_trng_final, family = "bino
mial")
> vif(model_LR)
                     GVIF Df GVIF^(1/(2*Df))
age              2.097427  1        1.448250
job              6.009406 11        1.084929
marital          1.488378  3        1.068527
education        3.496928  7        1.093540
default          1.161552  2        1.038149
housing          1.026318  2        1.006515
contact          2.523594  1        1.588582
month           74.878885  9        1.270958
day_of_week      1.077809  4        1.009410
duration         1.411048  1        1.187875
campaign         1.064896  1        1.031938
pdays            9.123700  1        3.020546
previous         5.460787  1        2.336833
poutcome        26.479101  2        2.268432
emp.var.rate   135.674988  1       11.647961
cons.price.idx  54.897808  1        7.409306
cons.conf.idx    5.234861  1        2.287982
euribor3m      142.832425  1       11.951252
nr.employed    145.223657  1       12.050878
```

Step 4:

VIF result shows presence of multicollinearity. Remove variables that has vif > 5 (Yellow lines above) and build the model again.

```
> model_LR<-glm(y~age+marital+education+default+housing+contact+d
ay_of_week+duration+campaign, data = bdata_trng_final, family = "
binomial")
> vif(model_LR)
                 GVIF Df GVIF^(1/(2*Df))
age          1.463590  1        1.209789
marital      1.339616  3        1.049937
education    1.243936  7        1.015714
default      1.112016  2        1.026899
housing      1.010428  2        1.002597
contact      1.062167  1        1.030615
day_of_week  1.016600  4        1.002060
duration     1.134090  1        1.064937
campaign     1.018416  1        1.009166
```

Step 5:

VIF result shows multicollinearity have been rectified. Above model looks fine.

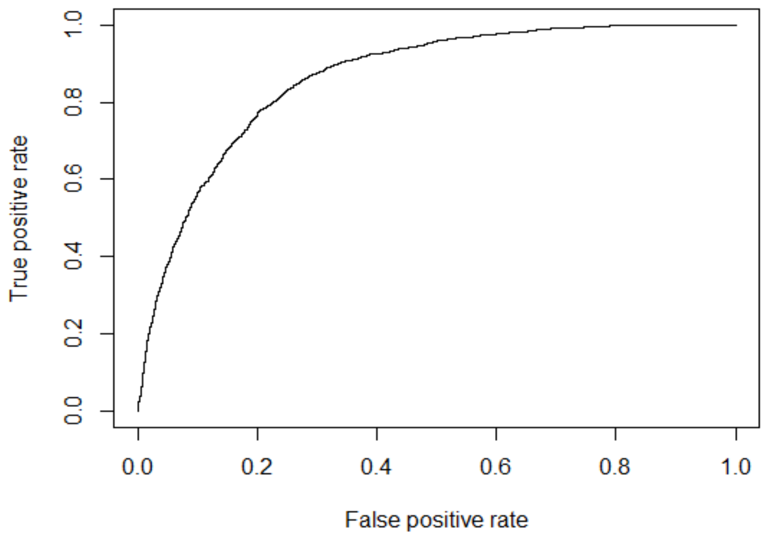
Accuracy of the model:

Finding accuracy of the above Logistic Regression Model using R and AUC (Area Under the Curve)

R code snippet:

```
> pred_y<-predict(model_LR,bdata_tst, type="response")
> pred<-prediction(pred_y,bdata_tst$y)
> rocc<-performance(pred,"tpr","fpr")
> plot(rocc)
> aucrp<-performance(pred,"auc")
> aucrp
```

ROC Curve



AUC: 0.8636697 ~ 86.366%

SECTION 4: Model for Decision Tree classification

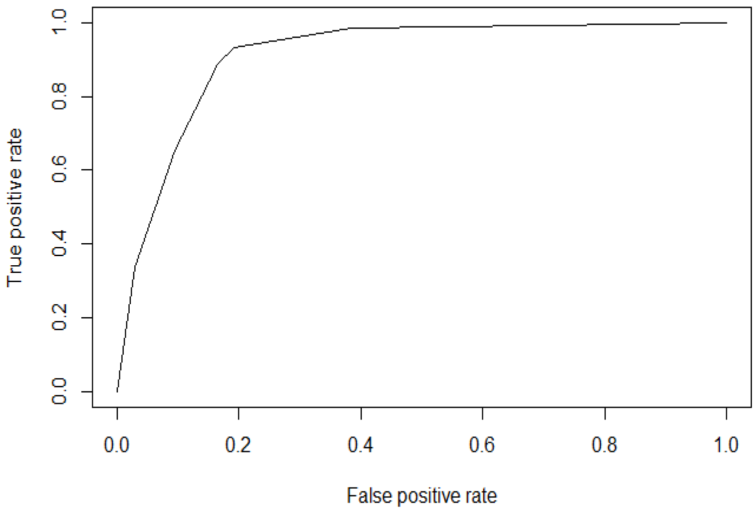
Model building using rpart:

```
model_DT<-rpart(y~.,data=bdata_trng_final, method="class")
```

Accuracy of the model:

```
pred_DT<-predict(model_DT,bdata_tst, type = "prob")
head(pred_DT)
prsc_DT<-pred_DT[,2]
pred<-prediction(prsc_DT,bdata_tst$y)
rocc<-performance(pred,"tpr","fpr")
plot(rocc)
aucrp<-performance(pred,"auc")
aucrp
```

ROC Curve



AUC: 0.911614 ~ 91.16%

SECTION 5: Model for Bayesian classification

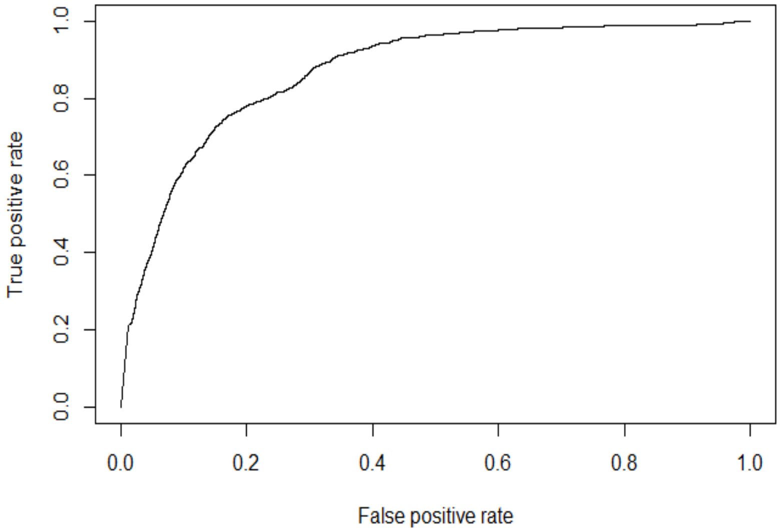
Model building using rpart:

```
> model_BA<-naiveBayes(y~.,data = bdata_trng_final)
```

Accuracy of the model:

```
> pred_BA<-predict(model_BA,bdata_tst,type = "raw")
> prsc_BA<-pred_BA[,2]
> pred<-prediction(prsc_BA,bdata_tst$y)
> rocc<-performance(pred,"tpr","fpr")
> plot(rocc)
> aucrp<-performance(pred,"auc")
> aucrp
```

ROC Curve:



AUC: 0.8697756 ~ 86.97%

SECTION 6: Comparison of the 2 Model

Comparing the AUC values for Decision Tree and Bayesian's Classification, Decision Tree accuracy is higher.

	Decision Tree	Bayesian Classification
AUC %	91.16	86.97

SECTION 7: Complete R File (code)

Assignment.R

SECTION 8: Data File

bankdataPWork.csv