# Bank Telemarketing (STAT 530)

Faiyaz Ahmad, Krutika Gopinathan, Rutuja Magdum

2023-04-21

# 1. Loading Data and Important Library

# 2. Some Important functions( For Model Evaluation)

### 2.1: Logistic Regression Model Evaluation (Precsion, Recall, and F1)

```
# Model evaluation for logistic regression
model_evaluation <- function(model,data,threshold){
  prob<-stats::predict(model, newdata=data, type="response")
  pred<- rep(0,dim(data)[1]) #createazerovector
  pred[prob>threshold]=1
  tab<-table(pred,data$y)
  tp<-tab[4]
  fp<-tab[2]
  tn<-tab[1]
  fn<-tab[3]
  Accuracy<-(tp+tn)/sum(tab)
  Precision<-(tp/(tp+fp))
  Recall<-(tp/(tp+fn))
  F_1<-2*Precision*Recall/(Precision+Recall)
  x<-c(Accuracy,Precision,Recall,F_1)
  return(x)
  }
```

### 2.2 Function for Drawing ROC__AUC__Curve for Logistic Regression

```
roc_auc_curve<-function(model,data){
  prob1<-stats::predict(model, newdata=data, type="response")
  rocobj <- roc(test$y, prob1)
  auc <- round(auc(test$y, prob1),4)
  #create ROC plot
  ggroc(rocobj, colour = 'steelblue', size = 2) +
    ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')'))
}
```

### 2.3. Model Evaluation Report for Logistic Regression

```
model_report<-function(model,train,test,threshold){
  test_set<-model_evaluation(model,test,threshold)
  test_set
  training_set<-model_evaluation(model,train,threshold)
  row.names<-c("Accuracy","Precision","Recall","F_1")
  df1<-data.frame(row.names,training_set,test_set)
  return(df1)
}
```

### 2.4. Decision Tree Model Evaluation function

```
decision_tree_eval <- function(model,data){
  y_pred<-predict(model,data,type='class')
  tab<-table(y_pred,data$y)
  tp<-tab[4]
  fp<-tab[2]
  tn<-tab[1]
  fn<-tab[3]
  Accuracy<-(tp+tn)/sum(tab)
  Precision<-(tp/(tp+fp))
  Recall<-(tp/(tp+fn))
  F_1<-2*Precision*Recall/(Precision+Recall)
  x<-c(Accuracy,Precision,Recall,F_1)
  return(x)
  }

decision_tree_report<-function(model,train,test){
  test_set<-decision_tree_eval(model,test)
  training_set<-decision_tree_eval(model,train)
  row.names<-c("Accuracy","Precision","Recall","F_1")
  df1<-data.frame(row.names,training_set,test_set)
  return(df1)
}
```

## 3. Data Preprocessing

### 3.1 Converting the data categorical data as factor

```
data$y<-as.factor(data$y)
data$job<-as.factor(data$job)
data$marital<-as.factor(data$marital)
data$education<-as.factor(data$education)
data$default<-as.factor(data$default)
data$housing<-as.factor(data$housing)
data$loan<-as.factor(data$loan)
data$month<-as.factor(data$month)
data$poutcome<-as.factor(data$poutcome)
```

### 3.2. Splitting the data into Training(80%) and Test(20%) set

```
#make this example reproducible
set.seed(1)
#use 80% of data set as training set and 20% as test set
sample <- sample(c(TRUE, FALSE), nrow(data),replace=TRUE,prob=c(0.8,0.2))

train <- data[sample, ]
test  <- data[!sample, ]
```

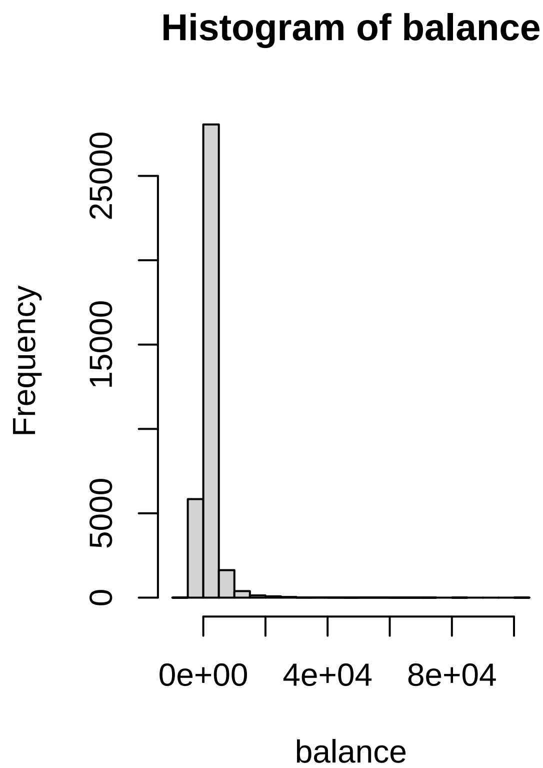
### 3.3. Treating outliers

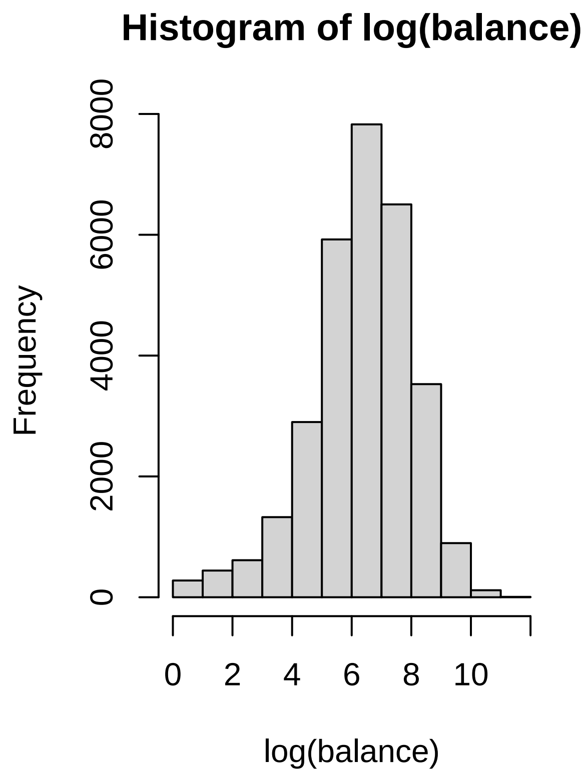
```
 1. For 'balance' there are lot of outliers.
 2. As data is divided into training and test.
 3. Process of treating outliers:
    IQR (Inter-quartile range): Q3-Q1
    If numerical feature is not in between Q1-1.5IQR to Q3+1.15IQR, then that particular data point wil
 4. We treating the outliers in training set, and in testing set, same value need to imputed as of trai
```

```
attach(train)
par(mfrow=c(1,2))
hist(balance)
hist(log(balance))
```

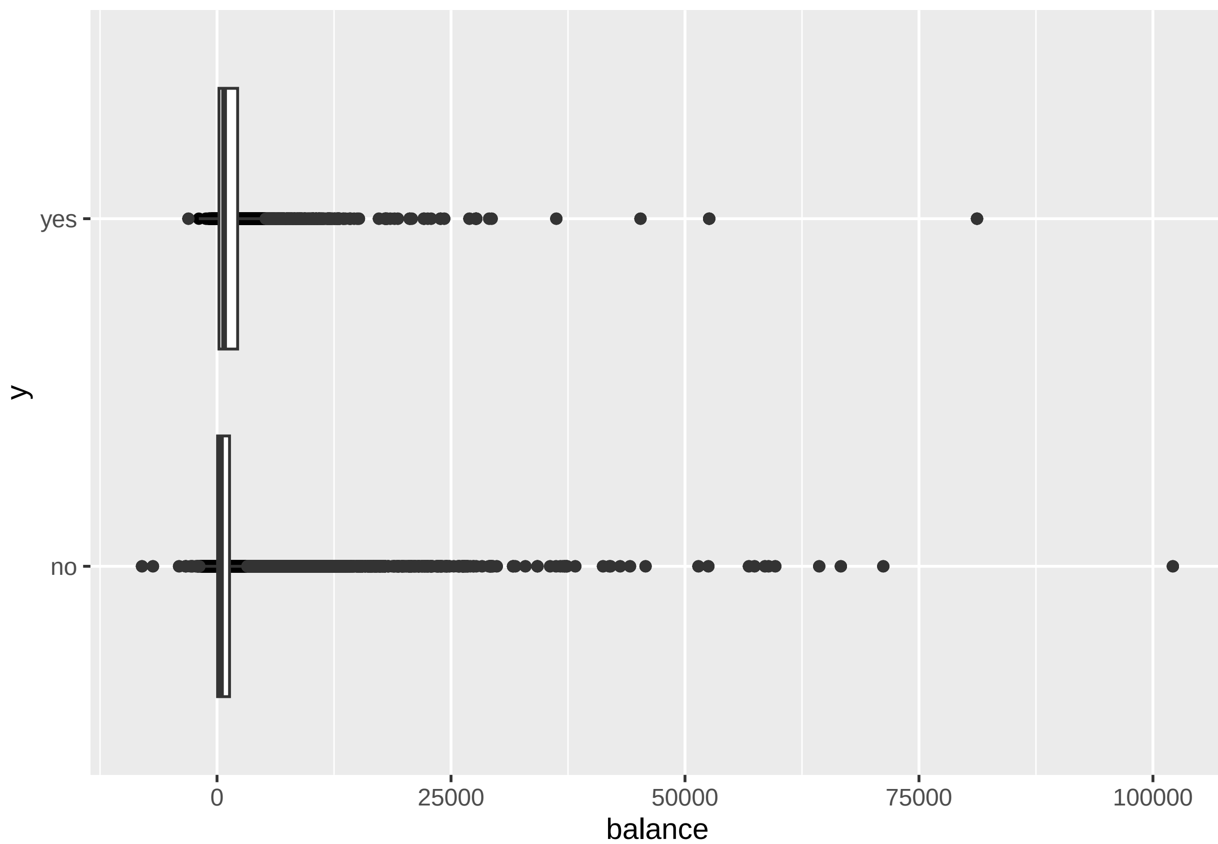
### 3.3.1 Let treat the outliers related to 'balance' feature

```
## Warning in log(balance): NaNs produced
```




```
ggplot(train, aes(balance,y))+
  geom_point()+
  geom_boxplot()
```




```
median<-median(train$balance)
q1<-quantile(train$balance,probs = c(.25, .5, .75))
IQR<-q1[3]-q1[1]
Max<-q1[3]+1.5*IQR
Min<-q1[1]-1.5*IQR
train$balance[train$balance>Max]<-median
train$balance[train$balance<Min]<-median
test$balance[test$balance>Max]<-median
test$balance[test$balance<Min]<-median
```

### 3.3.2. Treating the outliers for 'Duration' feature

```
median<-median(train$duration)
q1<-quantile(train$duration,probs = c(.25, .5, .75))
IQR<-q1[3]-q1[1]
Max<-q1[3]+1.5*IQR
Min<-q1[1]-1.5*IQR
train$duration[train$duration>Max]<-median
train$duration[train$duration<Min]<-median
test$duration[test$duration>Max]<-median
test$duration[test$duration<Min]<-median
```

### 3.3.3 Treating the outliers for 'Campaign' feature

```
median<-median(train$campaign)
q1<-quantile(train$campaign,probs = c(.25, .5, .75))
IQR<-q1[3]-q1[1]
Max<-q1[3]+1.5*IQR
Min<-q1[1]-1.5*IQR
train$campaign[train$campaign>Max]<-median
train$campaign[train$campaign<Min]<-median
test$campaign[test$campaign>Max]<-median
test$campaign[test$campaign<Min]<-median
```

# 4. Feature Selection:

### 4.1 Categorical Features

#### Test of Independence: For categorical Features

```
Null Hypothesis: There is no association between two categorical features.
Alternate Hypothesis: There is assocation between two features.
```

```
tab<-with(train,table(default,y))
addmargins(prop.table(tab))
```

#### 4.1.1. Credit Default vs y

```
##        y
## default          no         yes         Sum
##     no  0.865998177 0.115853153 0.981851330
##     yes 0.016933234 0.001215436 0.018148670
##     Sum 0.882931411 0.117068589 1.000000000
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 15.758, df = 1, p-value = 7.2e-05
```

#### Comment:

1. P-value is less than the significance level of 0.05, we can reject the null hypothesis.
2. Credit default is associated with y

```
tab<-with(train,table(loan,y))
addmargins(prop.table(tab))
```

#### 4.1.2. personal loan vs y

```
##       y
## loan          no        yes        Sum
##   no  0.73398525 0.10640590 0.84039115
##   yes 0.14894616 0.01066269 0.15960885
##   Sum 0.88293141 0.11706859 1.00000000
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 167.47, df = 1, p-value < 2.2e-16
```

**Comment:**

1. P-value is less than the significance level of 0.05, we can reject the null hypothesis.
2. Personal loan is associated with y

```
tab<-with(train,table(housing,y))
addmargins(prop.table(tab))
```

#### 4.1.3. housing loan vs y

```
##        y
## housing         no        yes        Sum
##     no  0.36910583 0.07436259 0.44346841
##     yes 0.51382558 0.04270600 0.55653159
##     Sum 0.88293141 0.11706859 1.00000000
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 714.1, df = 1, p-value < 2.2e-16
```

**Comment:**

1. P-value is less than the significance level of 0.05, we can reject the null hypothesis.
2. Housing loan is associated with y

```
tab<-with(train,table(education,y))
addmargins(prop.table(tab))
```

#### 4.1.5. education vs y

```
##            y
## education            no         yes         Sum
##   primary   0.138476838 0.013369796 0.151846634
##   secondary 0.458247010 0.053865915 0.512112925
##   tertiary  0.250352200 0.044363415 0.294715616
##   unknown   0.035855363 0.005469462 0.041324825
##   Sum       0.882931411 0.117068589 1.000000000
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 189.07, df = 3, p-value < 2.2e-16
```

```
tab<-with(train,table(job,y))
addmargins(prop.table(tab))
```

#### 4.1.5. job vs y

```
##                y
## job                        no          yes          Sum
##   admin.        0.1002734731 0.0137841496 0.1140576227
##   blue-collar   0.2006574404 0.0156625508 0.2163199912
##   entrepreneur  0.0301096655 0.0028452253 0.0329548907
##   housemaid     0.0248611917 0.0024032485 0.0272644402
##   management    0.1798016629 0.0293914533 0.2091931162
##   retired       0.0386453413 0.0116295130 0.0502748543
##   self-employed 0.0310488661 0.0041435319 0.0351923980
##   services      0.0833678628 0.0079555813 0.0913234441
##   student       0.0150272092 0.0059943095 0.0210215187
##   technician    0.1490014088 0.0177066932 0.1667081020
##   unemployed    0.0243363443 0.0047512500 0.0290875943
##   unknown       0.0058009447 0.0008010828 0.0066020276
##   Sum           0.8829314107 0.1170685893 1.0000000000
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 706.09, df = 11, p-value < 2.2e-16
```

```r
tab<-with(train,table(month,y))
addmargins(prop.table(tab))
```

#### 4.1.6. month vs y

```
##      y
## month          no         yes         Sum
##   apr 0.052070385 0.012596337 0.064666722
##   aug 0.122289440 0.015441562 0.137731002
##   dec 0.002651860 0.002237507 0.004889368
##   feb 0.049169912 0.009751112 0.058921024
##   jan 0.028479876 0.003176708 0.031656584
##   jul 0.138918814 0.013618408 0.152537223
##   jun 0.105604818 0.012375349 0.117980166
##   mar 0.005082733 0.005248474 0.010331206
##   may 0.283970056 0.020330930 0.304300986
##   nov 0.078865225 0.009171017 0.088036242
##   oct 0.008756664 0.006878263 0.015634927
##   sep 0.007071628 0.006242921 0.013314549
##   Sum 0.882931411 0.117068589 1.000000000
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 2410.8, df = 11, p-value < 2.2e-16
```

```r
tab<-with(train,table(default,loan))
addmargins(prop.table(tab))
```

#### 4.1.7. Credit Default vs Personal Loan

```
##        loan
## default          no         yes         Sum
##     no  0.828844507 0.153006823 0.981851330
##     yes 0.011546642 0.006602028 0.018148670
##     Sum 0.840391149 0.159608851 1.000000000
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 206.4, df = 1, p-value < 2.2e-16
```

**Comment:**

1. P-value is less than the significance level of 0.05, we can reject the null hypothesis.
2. Credit default dependent on Loan.

```r
tab<-with(data,table(housing,loan))
addmargins(prop.table(tab))
```

#### 4.4.8. Housing Loan vs Personal Loan

```
##        loan
## housing         no        yes        Sum
##     no  0.38052686 0.06363496 0.44416182
##     yes 0.45924664 0.09659154 0.55583818
##     Sum 0.83977351 0.16022649 1.00000000
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 76.975, df = 1, p-value < 2.2e-16
```

**Comment:**

1. Null Hypothesis: Housing Loan not associated with Personal Loan .
   Alternate Hypothesis: Housing Loan is associated with  personal loan.

2. P-value is less than the significance level of 0.05, we can reject the null hypothesis

2. Housing loan is associated with personal loan.

```
tab<-with(train,table(education,job))
addmargins(prop.table(tab))
```

#### 4.1.9. Education vs Job

```
##            job
## education         admin.  blue-collar entrepreneur    housemaid   management
##   primary   0.0046407558 0.0831468744 0.0039225436 0.0140051380 0.0067677688
##   secondary 0.0932570923 0.1199138145 0.0120162426 0.0083423110 0.0240877324
##   tertiary  0.0122924781 0.0031767078 0.0154139388 0.0038949200 0.1730615176
##   unknown   0.0038672965 0.0100825944 0.0016021657 0.0010220712 0.0052760973
##   Sum       0.1140576227 0.2163199912 0.0329548907 0.0272644402 0.2091931162
##            job
## education        retired self-employed     services      student   technician
##   primary   0.0176790696  0.0029280959 0.0075964752 0.0009944477 0.0033700726
##   secondary 0.0218502251  0.0127897019 0.0759095053 0.0112704069 0.1147758349
##   tertiary  0.0080384520  0.0185630231 0.0045302616 0.0051103561 0.0433137206
##   unknown   0.0027071075  0.0009115770 0.0032872020 0.0036463081 0.0052484738
##   Sum       0.0502748543  0.0351923980 0.0913234441 0.0210215187 0.1667081020
##            job
## education     unemployed      unknown          Sum
##   primary   0.0056628270 0.0011325654 0.1518466341
##   secondary 0.0162426452 0.0016574128 0.5121129251
##   tertiary  0.0065467805 0.0007734593 0.2947156156
##   unknown   0.0006353416 0.0030385901 0.0413248253
##   Sum       0.0290875943 0.0066020276 1.0000000000
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 23045, df = 33, p-value < 2.2e-16
```

**Comment:**

```
 1. Null Hypothesis: Education not associated with job type.
    Alternate Hypothesis: Education is associated with  job type.

 2. P-value is less than the significance level of 0.05, we can reject the null hypothesis

 3. Education is associated with the Job type
```

### 4.1.10 Does adding education along with job as feature in model will result in better model than alone with job type?

```
  i) We will first fit the model with job and do anova test on it.
  ii) Then add education varible in first model to check the impact of new variable on model. To perfor
```

### 4.1.4 Does adding education along with housing along with loan as feature in model will result in better model than alone with housing type?

```
mod1<-glm(y~housing,family='binomial',data=train)
mod2<-glm(y~housing+loan,family='binomial',data=train)
anova(mod1,mod2,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ housing
## Model 2: y ~ housing + loan
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     36199      25428
## 2     36198      25262  1   165.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Comment:

```
1.From chisquare test,it is evident that housing and loan is associated with each other.
2. From above anova test, it suggest that adding the housing along with loan will give significant impa
3. So, we can use housing along with the loan.
```

```
mod1<-glm(y~loan,family='binomial',data=train)
mod2<-glm(y~loan+default,family='binomial',data=train)
anova(mod1,mod2,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ loan
## Model 2: y ~ loan + default
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     36199      25950
## 2     36198      25938  1   11.666 0.0006366 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Comment:

```
1. From above anova test, it suggest adding the loan along with the credit default have significant eff
```

```
mod1<-glm(y~loan,family='binomial',data=train)
mod2<-glm(y~loan+housing,family='binomial',data=train)
anova(mod1,mod2,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ loan
## Model 2: y ~ loan + housing
```

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     36199      25950
## 2     36198      25262  1   688.08 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1. From above, it is clear, that housing and loan are completely associated with each other. Can act as
```

```r
mod1<-glm(y~month,family='binomial',data=train)
anova(mod1,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  36200      26140
## month 11   1741.2     36189      24399 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**let's check the distributin of 'y' in training and testing set.**

```r
# Distribution of 'y' in training set
tab<-table(train$y)
prop.table(tab)
```

```
##
##        no       yes
## 0.8829314 0.1170686
```

```r
# Distribution of 'y' in test set
tab<-table(test$y)
prop.table(tab)
```

```
##
##        no       yes
## 0.8833518 0.1166482
```

## Train the model.

**1. Demography: Age,job,marital**

```r
dem_mod<-glm(y~age+job+marital,family='binomial',data=train)
summary(dem_mod)
```

```
##
## Call:
## glm(formula = y ~ age + job + marital, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9516  -0.5238  -0.4542  -0.3847   2.4183
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.512657   0.108645 -23.127  < 2e-16 ***
## age               0.010919   0.001953   5.591 2.26e-08 ***
## jobblue-collar   -0.507696   0.065134  -7.795 6.46e-15 ***
## jobentrepreneur  -0.324338   0.114046  -2.844 0.004456 **
## jobhousemaid     -0.322807   0.122984  -2.625 0.008670 **
## jobmanagement     0.176137   0.058258   3.023 0.002500 **
## jobretired        0.676357   0.083843   8.067 7.21e-16 ***
## jobself-employed -0.011202   0.099448  -0.113 0.910315
## jobservices      -0.335990   0.078174  -4.298 1.72e-05 ***
## jobstudent        0.956036   0.097046   9.851  < 2e-16 ***
## jobtechnician    -0.148539   0.063577  -2.336 0.019473 *
## jobunemployed     0.352020   0.096342   3.654 0.000258 ***
## jobunknown       -0.030076   0.204776  -0.147 0.883232
## maritalmarried   -0.077945   0.053752  -1.450 0.147036
## maritalsingle     0.372490   0.060799   6.127 8.98e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26140  on 36200  degrees of freedom
## Residual deviance: 25390  on 36186  degrees of freedom
## AIC: 25420
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(dem_mod,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    36200      26140
```

```
## age       1    32.13     36199      26108 1.441e-08 ***
## job      11   601.94     36188      25506 < 2.2e-16 ***
## marital  2   116.18     36186      25390 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
threshold=0.1
model_report(dem_mod,train,test,threshold)
```

```
##   row.names training_set  test_set
## 1  Accuracy    0.4838264 0.4834628
## 2 Precision    0.1460905 0.1443238
## 3    Recall    0.7036338 0.6955281
## 4       F_1    0.2419473 0.2390451
```

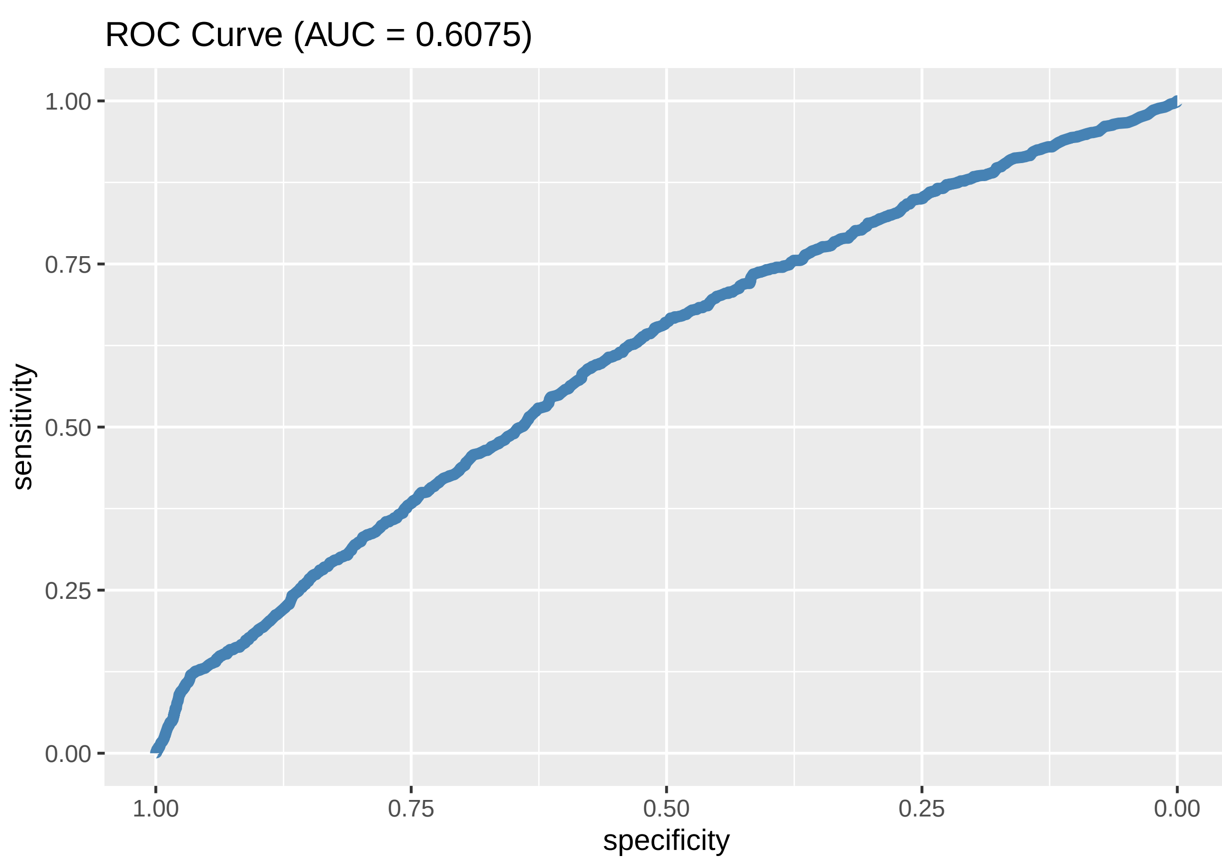
```
roc_auc_curve(dem_mod,test)
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```



**2.Financial Characteristics: Balance, housing,loan**

```
fin_mod<-glm(y~balance+housing+loan,family='binomial',data=train)
summary(fin_mod)
```

```
##
## Call:
## glm(formula = y ~ balance + housing + loan, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8512  -0.5763  -0.4149  -0.3836   2.6209
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.719e+00  2.670e-02  -64.39   <2e-16 ***
## balance      2.596e-04  1.893e-05   13.72   <2e-16 ***
## housingyes  -8.532e-01  3.405e-02  -25.06   <2e-16 ***
## loanyes     -6.113e-01  5.603e-02  -10.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26140  on 36200  degrees of freedom
## Residual deviance: 25084  on 36197  degrees of freedom
## AIC: 25092
##
## Number of Fisher Scoring iterations: 5
```

```
anova(fin_mod,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    36200      26140
## balance  1   245.39     36199      25895 < 2.2e-16 ***
## housing  1   675.17     36198      25220 < 2.2e-16 ***
## loan     1   135.90     36197      25084 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
threshold=0.1
model_report(fin_mod,train,test,threshold)
```

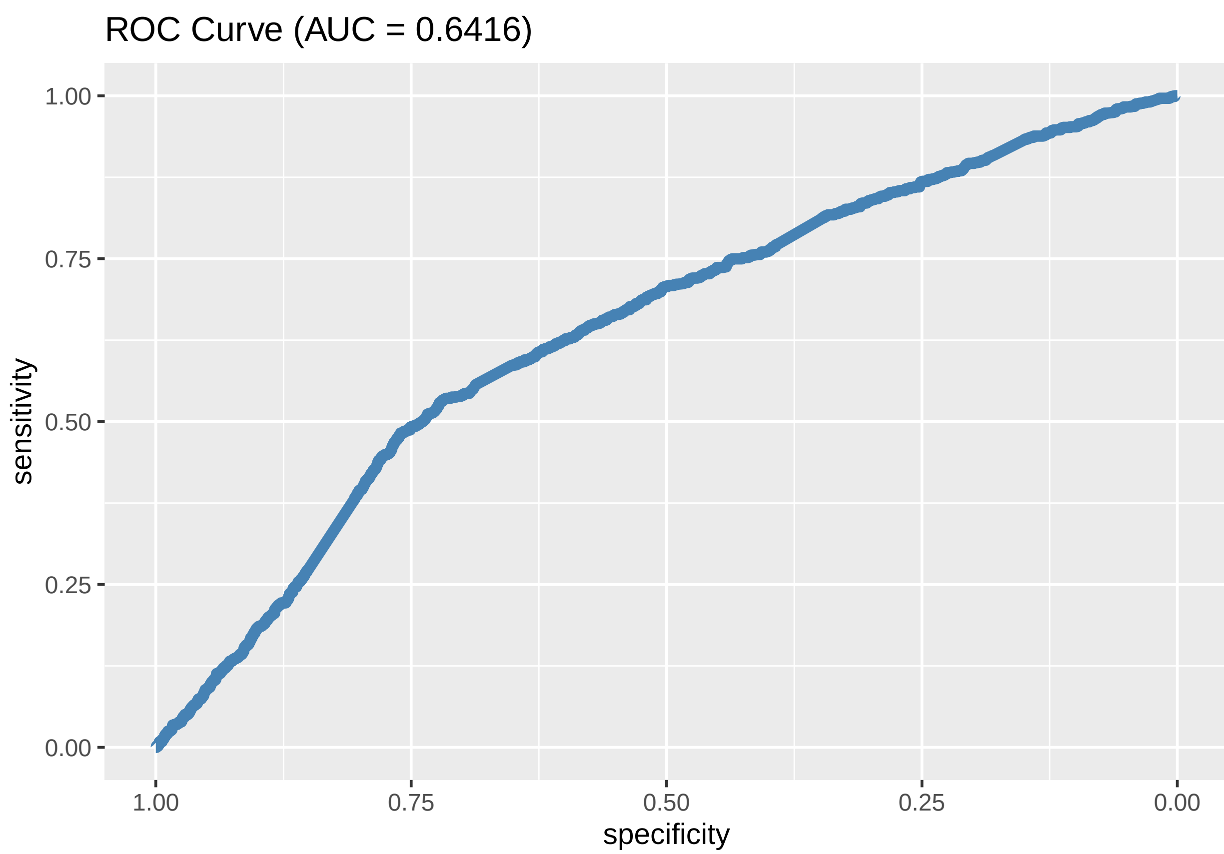
```
##   row.names training_set  test_set
## 1  Accuracy    0.5779122 0.5731410
## 2 Precision    0.1679697 0.1651090
## 3    Recall    0.6590373 0.6555661
## 4       F_1    0.2677082 0.2637825
```

### Comment:

1. Precision and Recall is very less on positive cases (i.e Response variable: 'y')
2. From here, It can be drawin, that Financial characteristics are not good predictor of subscriptio
3. Precision and Recall for positive class is zero for test set.

```
roc_auc_curve(fin_mod,test)
```

```
## Setting levels: control = no, case = yes

## Setting direction: controls < cases

## Setting levels: control = no, case = yes

## Setting direction: controls < cases
```



3. **Campaign: duration,campaign,day,month**

```
camp_mod<-glm(y~duration+campaign+day+month,'binomial',data=train)
summary(camp_mod)
```

```
##
## Call:
## glm(formula = y ~ duration + campaign + day + month, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0203  -0.4817  -0.3719  -0.2945   2.6395
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.2659330  0.0785817 -28.835  < 2e-16 ***
## duration     0.0042783  0.0001122  38.144  < 2e-16 ***
## campaign    -0.1298589  0.0156978  -8.272  < 2e-16 ***
## day         -0.0007434  0.0023657  -0.314    0.753
## monthaug    -0.3821556  0.0723401  -5.283 1.27e-07 ***
## monthdec     1.3250660  0.1663189   7.967 1.63e-15 ***
## monthfeb    -0.0390787  0.0853410  -0.458    0.647
## monthjan    -0.7650684  0.1169555  -6.542 6.09e-11 ***
## monthjul    -0.7842069  0.0730203 -10.740  < 2e-16 ***
## monthjun    -0.5638121  0.0762615  -7.393 1.43e-13 ***
## monthmar     1.6653224  0.1205127  13.819  < 2e-16 ***
## monthmay    -1.1623814  0.0669558 -17.360  < 2e-16 ***
## monthnov    -0.6303403  0.0806412  -7.817 5.43e-15 ***
## monthoct     1.2736502  0.1039319  12.255  < 2e-16 ***
## monthsep     1.4177111  0.1102343  12.861  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26140  on 36200  degrees of freedom
## Residual deviance: 22880  on 36186  degrees of freedom
## AIC: 22910
##
## Number of Fisher Scoring iterations: 5
```

```
anova(camp_mod,test='Chisq')
```

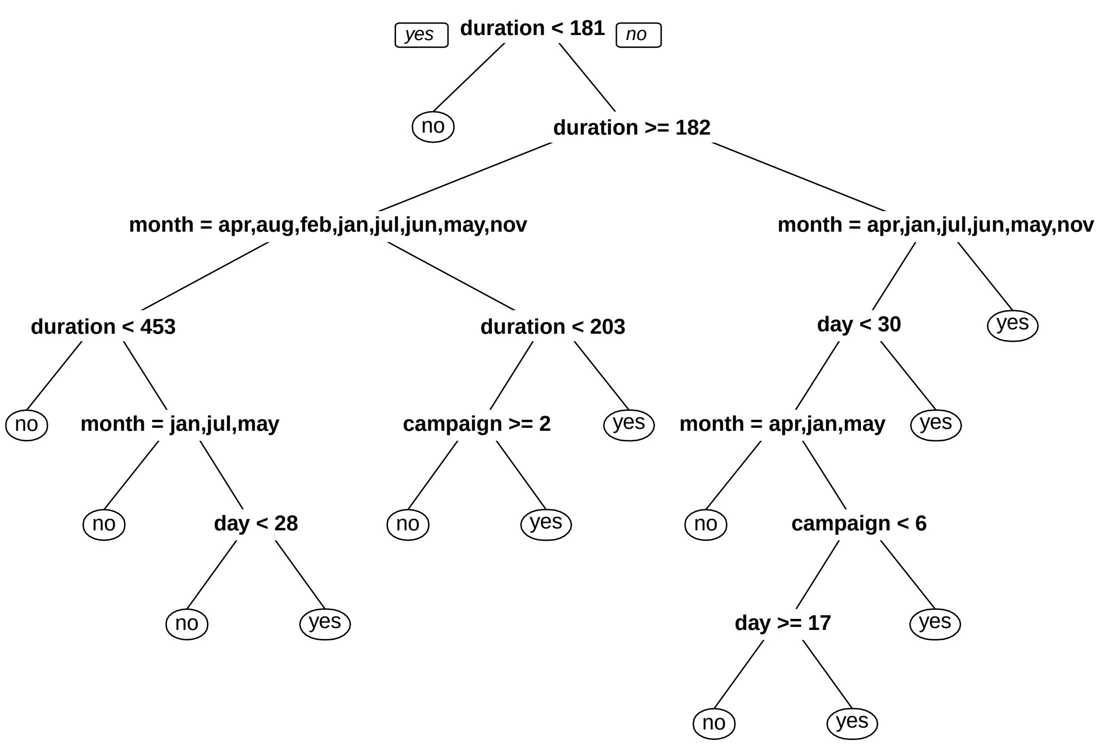
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
```

```
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      36200      26140
## duration  1  1433.91      36199      24706 < 2.2e-16 ***
## campaign  1   124.41      36198      24582 < 2.2e-16 ***
## day       1    12.58      36197      24569  0.000389 ***
## month    11  1688.99      36186      22880 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
threshold=0.1
model_report(camp_mod,train,test,threshold)
```

```
##   row.names training_set  test_set
## 1  Accuracy    0.6749261 0.6826859
## 2 Precision    0.2173848 0.2190802
## 3    Recall    0.6833412 0.6707897
## 4       F_1    0.3298405 0.3302881
```

```
options(repr.plot.width=6, repr.plot.height=5)

dt_camp<-rpart(y~duration+campaign+day+month,cp=0.001,maxdepth=7, minbucket=5,method='class',data=train
prp(dt_camp)  #,space=4,split.cex=1.5, nn.border.col=0)
```



```r
decision_tree_report(dt_camp,train,test)
```

```
##   row.names training_set  test_set
## 1  Accuracy    0.8936493 0.8906770
## 2 Precision    0.6008316 0.5660000
## 3    Recall    0.2727702 0.2692674
## 4       F_1    0.3752029 0.3649259
```

### 4. Mixed Model

```r
mix_mod<-glm(y~age+job+balance+marital+housing+loan+duration+campaign+day+month+previous,'binomial',data
summary(mix_mod)
```

```
##
## Call:
## glm(formula = y ~ age + job + balance + marital + housing + loan +
##     duration + campaign + day + month + previous, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7949  -0.4769  -0.3489  -0.2537   2.8926
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.972e+00  1.460e-01 -13.505  < 2e-16 ***
## age               9.963e-04  2.103e-03   0.474  0.63572
## jobblue-collar   -3.237e-01  6.942e-02  -4.662 3.13e-06 ***
## jobentrepreneur  -2.542e-01  1.206e-01  -2.107  0.03509 *
## jobhousemaid     -3.909e-01  1.312e-01  -2.979  0.00290 **
## jobmanagement     1.005e-01  6.328e-02   1.588  0.11222
## jobretired        2.913e-01  9.355e-02   3.113  0.00185 **
## jobself-employed -7.223e-02  1.068e-01  -0.676  0.49881
## jobservices      -1.865e-01  8.286e-02  -2.251  0.02438 *
## jobstudent        4.784e-01  1.082e-01   4.423 9.74e-06 ***
## jobtechnician    -9.811e-02  6.850e-02  -1.432  0.15206
## jobunemployed     1.027e-01  1.054e-01   0.975  0.32977
## jobunknown       -2.880e-01  2.178e-01  -1.322  0.18604
## balance           2.000e-04  2.093e-05   9.555  < 2e-16 ***
## maritalmarried   -1.707e-01  5.762e-02  -2.962  0.00306 **
## maritalsingle     1.250e-01  6.586e-02   1.898  0.05775 .
## housingyes       -6.393e-01  4.254e-02 -15.027  < 2e-16 ***
## loanyes          -4.614e-01  5.937e-02  -7.771 7.79e-15 ***
## duration          4.270e-03  1.150e-04  37.118  < 2e-16 ***
## campaign         -1.208e-01  1.585e-02  -7.622 2.50e-14 ***
## day              -2.316e-03  2.382e-03  -0.972  0.33096
## monthaug         -6.743e-01  7.762e-02  -8.688  < 2e-16 ***
## monthdec          8.064e-01  1.722e-01   4.684 2.82e-06 ***
## monthfeb         -2.857e-01  8.854e-02  -3.227  0.00125 **
## monthjan         -1.024e+00  1.199e-01  -8.538  < 2e-16 ***
## monthjul         -7.078e-01  7.629e-02  -9.278  < 2e-16 ***
```

```
## monthjun         -6.624e-01  7.930e-02  -8.352  < 2e-16 ***
## monthmar          1.201e+00  1.255e-01   9.566  < 2e-16 ***
## monthmay         -9.313e-01  6.891e-02 -13.514  < 2e-16 ***
## monthnov         -7.274e-01  8.311e-02  -8.752  < 2e-16 ***
## monthoct          8.210e-01  1.087e-01   7.555 4.19e-14 ***
## monthsep          8.655e-01  1.152e-01   7.510 5.93e-14 ***
## previous          8.967e-02  7.537e-03  11.897  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26140  on 36200  degrees of freedom
## Residual deviance: 22000  on 36168  degrees of freedom
## AIC: 22066
##
## Number of Fisher Scoring iterations: 5
```

```
anova(mix_mod,test='Chisq')
```

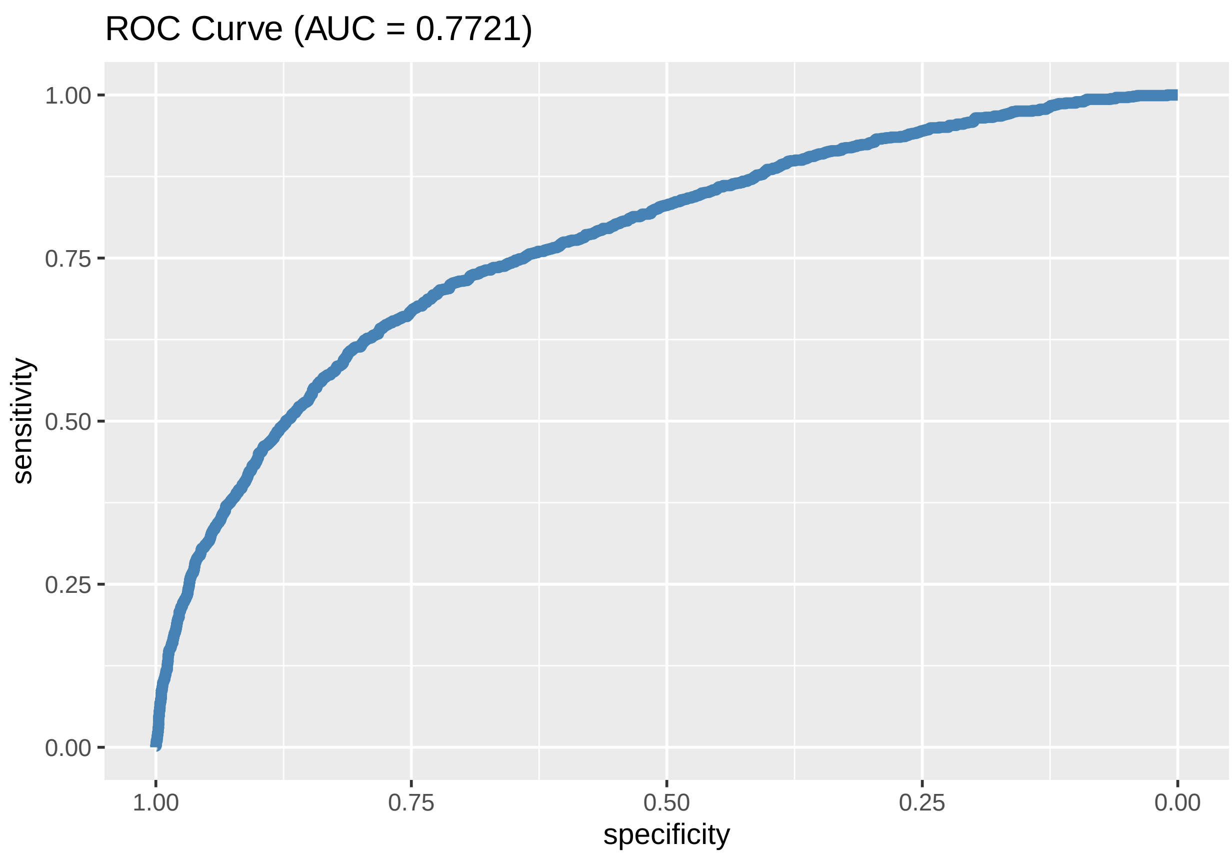
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     36200      26140
## age       1    32.13     36199      26108 1.441e-08 ***
## job      11   601.94     36188      25506 < 2.2e-16 ***
## balance   1   196.95     36187      25309 < 2.2e-16 ***
## marital   2   110.51     36185      25199 < 2.2e-16 ***
## housing   1   428.36     36184      24770 < 2.2e-16 ***
## loan      1   104.42     36183      24666 < 2.2e-16 ***
## duration  1  1419.02     36182      23247 < 2.2e-16 ***
## campaign  1   114.20     36181      23133 < 2.2e-16 ***
## day       1    15.59     36180      23117 7.861e-05 ***
## month    11   983.56     36169      22134 < 2.2e-16 ***
## previous  1   133.43     36168      22000 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
threshold=0.3
model_report(mix_mod,train,test,threshold)
```

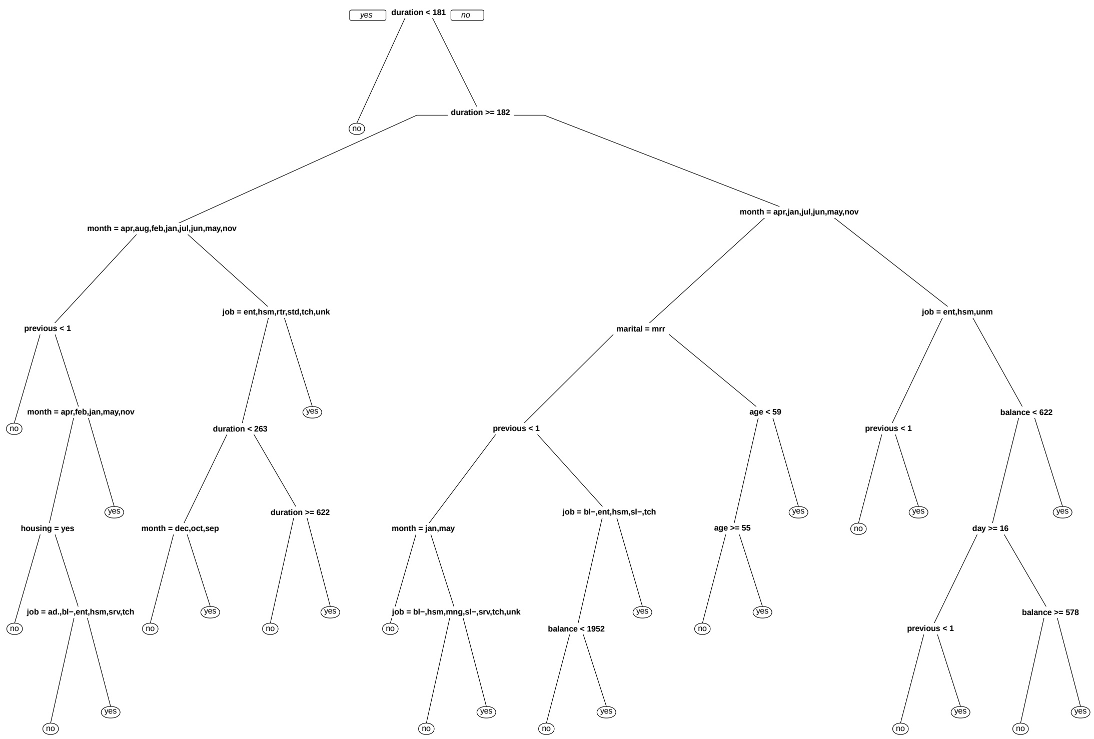
```
##   row.names training_set  test_set
## 1  Accuracy    0.8741747 0.8753607
## 2 Precision    0.4459229 0.4506849
## 3    Recall    0.3084002 0.3130352
## 4       F_1    0.3646255 0.3694554
```

```
roc_auc_curve(mix_mod,test)
```

```
## Setting levels: control = no, case = yes

## Setting direction: controls < cases

## Setting levels: control = no, case = yes

## Setting direction: controls < cases
```




```
#options(repr.plot.width=6, repr.plot.height=5)

dt_mix<-rpart(y~age+job+balance+marital+housing+loan+duration+campaign+day+month+previous,cp=0.001,maxd
prp(dt_mix) #,space=4,split.cex=1.5, nn.border.col=0)
```




```r
decision_tree_report(dt_mix,train,test)
```

```
##   row.names training_set  test_set
## 1  Accuracy    0.9014944 0.8952275
## 2 Precision    0.6119254 0.5710491
## 3    Recall    0.4334592 0.4091342
## 4       F_1    0.5074586 0.4767184
```