Identifying Good Customers

Project Summary

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will purchase (yes/no) a term deposit (variable y). The data is a randomly selected subset of a much larger set of data.

Data understanding and preparation

- Numerical (5): age, duration, campaign, pdays, previous

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   32.00   38.00   40.11   47.00   88.00
> summary(duration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   103.0   181.0   256.8   317.0  3643.0
> summary(campaign)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.537   3.000  35.000
> summary(pdays)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   999.0   999.0   960.4   999.0   999.0
> summary(previous)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1903  0.0000  6.0000
```

Overall, the mean exceeds the median, indicating a right skewness in the distribution of the numerical data.

- Categorical (10): job (12), marital(4), education (8), default, housing, loan, contact (2), month (12), dayow (5), poutcome (3), y(output).

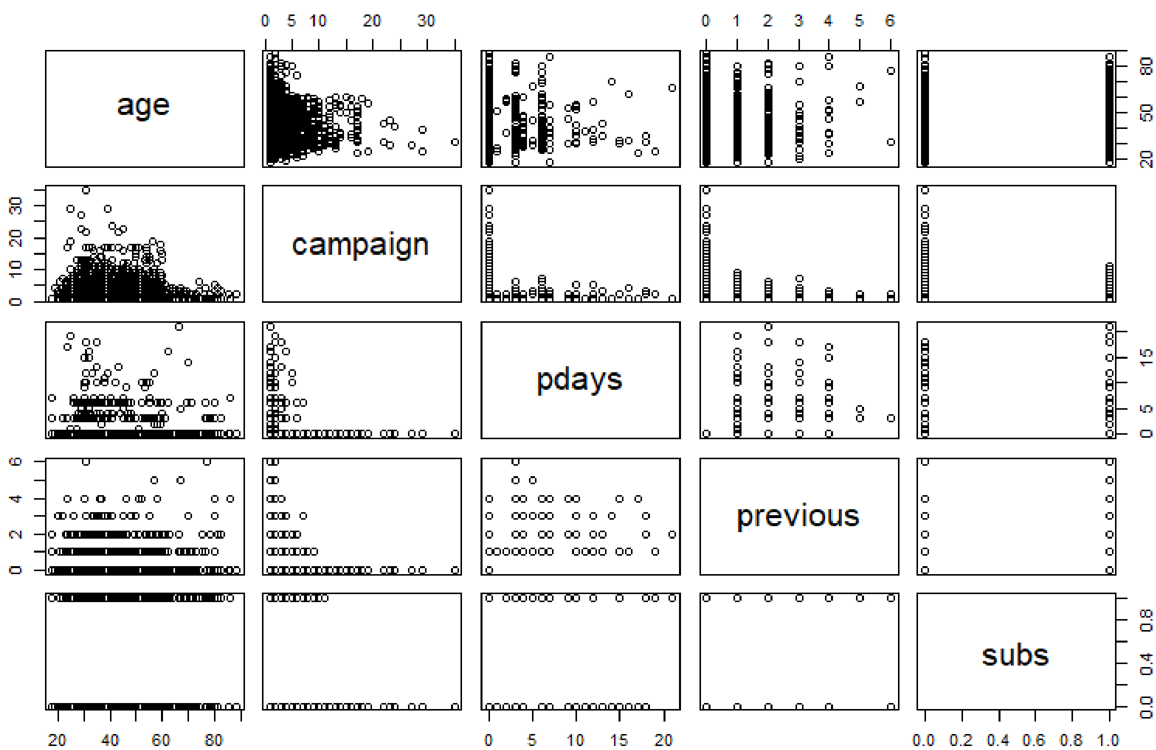
```
> table(job)
job
```

```
      admin.   blue-collar  entrepreneur     housemaid    management      retired
self-employed      services      student
       1012          884          148          110          324          166          159          393
82
   technician    unemployed      unknown
        691          111          39
> table(marital)
marital
divorced  married   single  unknown
    446     2509     1153       11
> table(education)
education
          basic.4y           basic.6y           basic.9y        high.school         illiterate
professional.course
              429                228                574                921                  1                535
  university.degree            unknown
             1264                167
> table(default)
default
     no unknown     yes
   3315     803       1
> table(housing)
housing
     no unknown     yes
   1839     105    2175
> table(loan)
loan
     no unknown     yes
   3349     105     665
> table(contact)
contact
 cellular telephone
    2652      1467
> table(month)
month
 apr  aug  dec  jul  jun  mar  may  nov  oct  sep
 215  636   22  711  530   48 1378  446   69   64
> table(dayow)
dayow
```

```
fri mon thu tue wed
768 855 860 841 795
> table(poutcome)
poutcome
    failure nonexistent     success
        454        3523         142
> table(y)
y
  no  yes
3668  451
```

The categorical data contain a number of unknown values. However, no observation or feature has too many missing values, so the data can remain the same without any deletion.

```
> pairs(data.frame(age,campaign,pdays,previous,subs))
```



```
> cor(age, campaign, method="pearson")
[1] -0.01416915
> cor(age, pdays, method="pearson")
[1] -0.04342483
> cor(age, previous, method="pearson")
[1] 0.05093104
> cor(pdays, campaign, method="pearson")
[1] 0.05874163
> cor(previous, campaign, method="pearson")
[1] -0.09148987
> cor(pdays, previous, method="pearson")
[1] -0.5879411
```

The scatterplot matrix shows no significant pattern or correlation between the variables. This is also supported by the correlation test above: no pair of variables is strongly correlated (>0.7), among which pdays and previous can be considered as most correlated. The significance of both the variables in terms of correlation can be checked further in the following classification method.

Logistic regression classification method

We will begin by regressing each of the independent variables with the response variables. Because the diagnosis variable is dichotomous, we use generalized linear model instead of the classical linear model.

```
> modelage<-glm(subs~age,family = binomial)
> summary(modelage)

Call:
glm(formula = subs ~ age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7089  -0.5029  -0.4621  -0.4391   2.2731

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.82698    0.19917 -14.194  < 2e-16 ***
```

```
age          0.01789    0.00463   3.863 0.000112 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2831.3  on 4117  degrees of freedom
AIC: 2835.3

Number of Fisher Scoring iterations: 4

> plot(campaign,subs)
> modelcampaign<-glm(subs~campaign,family = binomial)
> summary(modelcampaign)

Call:
glm(formula = subs ~ campaign, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5276  -0.5276  -0.4901  -0.4220   2.6481

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.74400    0.08162 -21.368  < 2e-16 ***
campaign    -0.15744    0.03203  -4.916 8.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2812.9  on 4117  degrees of freedom
AIC: 2816.9

Number of Fisher Scoring iterations: 5

> plot(pdays,subs)
```

```
> modelpdays<-glm(subs~pdays,family = binomial)
> summary(modelpdays)

Call:
glm(formula = subs ~ pdays, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.8871  -0.4498  -0.4498  -0.4498   2.1639  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.24015    0.05346  -41.90   <2e-16 ***
pdays        0.35512    0.03235   10.98   <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: 2845.8  on 4118  degrees of freedom
Residual deviance: 2685.8  on 4117  degrees of freedom
AIC: 2689.8

Number of Fisher Scoring iterations: 5

> plot(previous,subs)
> modelprevious<-glm(subs~previous,family = binomial)
> summary(modelprevious)

Call:
glm(formula = subs ~ previous, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-2.5960  -0.4209  -0.4209  -0.4209   2.2216  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.37923    0.05824  -40.86   <2e-16 ***
```

```
previous     0.95232    0.07238   13.16   <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: 2845.8  on 4118  degrees of freedom
Residual deviance: 2664.7  on 4117  degrees of freedom
AIC: 2668.7

Number of Fisher Scoring iterations: 5


> summary(glm(subs~job, family = binomial))

Call:
glm(formula = subs ~ job, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7261  -0.5309  -0.4408  -0.3782   2.4157

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.88844    0.09304 -20.297  < 2e-16 ***
jobblue-collar   -0.71365    0.16206  -4.403 1.07e-05 ***
jobentrepreneur  -0.97377    0.37522  -2.595  0.00945 **
jobhousemaid     -0.30879    0.33116  -0.932  0.35111
jobmanagement    -0.39395    0.21305  -1.849  0.06445 .
jobretired        0.67399    0.20684   3.258  0.00112 **
jobself-employed -0.53022    0.30402  -1.744  0.08115 .
jobservices      -0.43675    0.20005  -2.183  0.02902 *
jobstudent        0.68974    0.27778   2.483  0.01303 *
jobtechnician    -0.14463    0.15097  -0.958  0.33806
jobunemployed     0.31109    0.26862   1.158  0.24683
jobunknown       -0.28062    0.53594  -0.524  0.60055
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2781.3  on 4107  degrees of freedom
AIC: 2805.3

Number of Fisher Scoring iterations: 5

> summary(glm(subs~marital, family = binomial))

Call:
glm(formula = subs ~ marital, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.5373  -0.4601  -0.4601  -0.4601   2.1899  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -2.23774    0.16041 -13.950   <2e-16 ***
maritalmarried  0.04537    0.17361   0.261   0.7938    
maritalsingle   0.37541    0.18217   2.061   0.0393 *  
maritalunknown -0.06485    1.06079  -0.061   0.9513    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2835.9  on 4115  degrees of freedom
AIC: 2843.9

Number of Fisher Scoring iterations: 4

> summary(glm(subs~education, family = binomial))

Call:
glm(formula = subs ~ education, family = binomial)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5818  -0.5289  -0.4718  -0.3946   2.2786

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.3311     0.1699 -13.719   <2e-16 ***
educationbasic.6y             -0.1875     0.3040  -0.617   0.5374
educationbasic.9y             -0.1824     0.2324  -0.785   0.4324
educationhigh.school           0.1917     0.2010   0.954   0.3403
educationilliterate          -10.2349   324.7437  -0.032   0.9749
educationprofessional.course   0.3528     0.2154   1.638   0.1014
educationuniversity.degree     0.4349     0.1893   2.297   0.0216 *
educationunknown               0.6405     0.2728   2.348   0.0189 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2822.9  on 4111  degrees of freedom
AIC: 2838.9

Number of Fisher Scoring iterations: 11

> summary(glm(subs~default, family = binomial))

Call:
glm(formula = subs ~ default, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5085  -0.5085  -0.5085  -0.3549   2.3650

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.98049    0.05321 -37.223  < 2e-16 ***
defaultunknown  -0.75309    0.15673  -4.805 1.55e-06 ***
defaultyes     -10.58558  324.74370  -0.033    0.974
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2818.4  on 4116  degrees of freedom
AIC: 2824.4

Number of Fisher Scoring iterations: 11

> summary(glm(subs~housing, family = binomial))

Call:
glm(formula = subs ~ housing, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4836  -0.4836  -0.4824  -0.4824   2.2166

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.092353   0.074572 -28.058   <2e-16 ***
housingunknown -0.274770   0.356381  -0.771    0.441
housingyes      0.005129   0.101214   0.051    0.960
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2845.2  on 4116  degrees of freedom
AIC: 2851.2

Number of Fisher Scoring iterations: 4

> summary(glm(subs~loan, family = binomial))

Call:
```

```
glm(formula = subs ~ loan, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4867  -0.4867  -0.4867  -0.4645   2.2166

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.07374    0.05486 -37.800   <2e-16 ***
loanunknown -0.29338    0.35278  -0.832    0.406
loanyes     -0.09867    0.13924  -0.709    0.479
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2844.7  on 4116  degrees of freedom
AIC: 2850.7

Number of Fisher Scoring iterations: 4

> summary(glm(subs~contact, family = binomial))

Call:
glm(formula = subs ~ contact, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5522  -0.5522  -0.5522  -0.3262   2.4332

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.80369    0.05573 -32.365   <2e-16 ***
contacttelephone -1.10336    0.13031  -8.467   <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: 2845.8  on 4118  degrees of freedom
Residual deviance: 2759.3  on 4117  degrees of freedom
AIC: 2763.3

Number of Fisher Scoring iterations: 5

> summary(glm(subs~month, family = binomial))

Call:
glm(formula = subs ~ month, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.3232  -0.4606  -0.4162  -0.3675   2.3361  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.6039     0.1827  -8.781  < 2e-16 ***
monthaug     -0.5864     0.2253  -2.603 0.009234 ** 
monthdec      1.7862     0.4655   3.837 0.000125 ***
monthjul     -0.7986     0.2277  -3.507 0.000452 ***
monthjun     -0.3122     0.2241  -1.393 0.163653    
monthmar      1.9403     0.3451   5.623 1.88e-08 ***
monthmay     -1.0572     0.2127  -4.970 6.71e-07 ***
monthnov     -0.6339     0.2431  -2.607 0.009124 ** 
monthoct      1.0386     0.3100   3.350 0.000807 ***
monthsep      1.2244     0.3133   3.908 9.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2642.7  on 4109  degrees of freedom
AIC: 2662.7

Number of Fisher Scoring iterations: 5
```

```
> summary(glm(subs~dayow, family = binomial))

Call:
glm(formula = subs ~ dayow, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4934  -0.4865  -0.4786  -0.4696   2.1258

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.110578   0.116220 -18.160   <2e-16 ***
dayowmon     0.066182   0.158215   0.418    0.676
dayowthu     0.036359   0.158846   0.229    0.819
dayowtue     0.001365   0.160713   0.008    0.993
dayowwed    -0.038659   0.164190  -0.235    0.814
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2845.8  on 4118  degrees of freedom
Residual deviance: 2845.3  on 4114  degrees of freedom
AIC: 2855.3

Number of Fisher Scoring iterations: 4

> summary(glm(subs~poutcome, family = binomial))

Call:
glm(formula = subs ~ poutcome, family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.445  -0.416  -0.416  -0.416   2.232

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.7537     0.1323  -13.25  < 2e-16 ***
```

```
poutcomenonexistent  -0.6501     0.1458   -4.46 8.19e-06 ***
poutcomesuccess       2.3635     0.2200   10.75  < 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: 2845.8  on 4118  degrees of freedom
Residual deviance: 2577.7  on 4116  degrees of freedom
AIC: 2583.7

Number of Fisher Scoring iterations: 5
```

As can be seen, age, campaign, pdays, previous, contact, month, and poutcome are highly significant with almost every value of p smaller than 0.05. The job variable is moderately significant with half of the features showing significance. We continue generating the generalized linear model with all the independent variables based on the training data set.

```
> bank.additional$subs=ifelse(y=="yes",1,0)

> trainsample <- sample(4119,3119)
> trainsubs <- bank.additional[trainsample,]
> testsubs <- bank.additional[-trainsample,]

> mod1 = glm(subs ~
age+campaign+pdays+previous+job+marital+education+default+housing+loan+contact
+month+dayow+poutcome, data=trainsubs, family=binomial)
> summary(mod1)

Call:
glm(formula = subs ~ age + campaign + pdays + previous + job +
    marital + education + default + housing + loan + contact +
    month + dayow + poutcome, family = binomial, data = trainsubs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3528  -0.4344  -0.3334  -0.2301   2.9266
```

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.1959535	0.9553880	-3.345	0.000822	***
age	0.0237305	0.0079161	2.998	0.002720	**
campaign	-0.0960820	0.0393691	-2.441	0.014665	*
pdays	-0.0002703	0.0006544	-0.413	0.679542	
previous	0.5105013	0.1796730	2.841	0.004493	**
jobblue-collar	-0.4848043	0.2617205	-1.852	0.063972	.
jobentrepreneur	-0.6549200	0.4569203	-1.433	0.151762	
jobhousemaid	-0.0480069	0.4491438	-0.107	0.914880	
jobmanagement	-0.6211810	0.2886910	-2.152	0.031420	*
jobretired	-0.0746932	0.3454415	-0.216	0.828812	
jobself-employed	-0.6746112	0.3836856	-1.758	0.078707	.
jobservices	-0.5521881	0.2910989	-1.897	0.057840	.
jobstudent	0.7289630	0.3830790	1.903	0.057053	.
jobtechnician	-0.0194521	0.2100236	-0.093	0.926207	
jobunemployed	-0.4716835	0.4218446	-1.118	0.263505	
jobunknown	-0.6835914	0.7755126	-0.881	0.378063	
maritalmarried	0.3970838	0.2433142	1.632	0.102684	
maritalsingle	0.4492338	0.2737224	1.641	0.100755	
maritalunknown	0.9448227	1.1628643	0.812	0.416507	
educationbasic.6y	0.0472589	0.4382035	0.108	0.914117	
educationbasic.9y	0.1443019	0.3372827	0.428	0.668770	
educationhigh.school	0.3684444	0.3151013	1.169	0.242287	
educationprofessional.course	0.4428013	0.3338174	1.326	0.184681	
educationuniversity.degree	0.4419253	0.3148392	1.404	0.160422	
educationunknown	0.3754754	0.4080422	0.920	0.357475	
defaultunknown	-0.2587865	0.2080926	-1.244	0.213642	
housingunknown	-0.3134144	0.4786517	-0.655	0.512606	
housingyes	-0.1368157	0.1352401	-1.012	0.311705	
loanunknown	NA	NA	NA	NA	
loanyes	-0.1921157	0.1889134	-1.017	0.309177	
contacttelephone	-1.3311016	0.2039582	-6.526	6.74e-11	***
monthaug	-0.9648931	0.2884859	-3.345	0.000824	***
monthdec	2.1746254	0.6976054	3.117	0.001825	**
monthjul	-0.7536737	0.2880400	-2.617	0.008882	**
monthjun	0.6516773	0.3073693	2.120	0.033991	*
monthmar	1.6847434	0.4246654	3.967	7.27e-05	***
monthmay	-0.4265109	0.2677842	-1.593	0.111218	

```
monthnov                 -0.9118617  0.3138547  -2.905 0.003668 **
monthoct                  0.6831510  0.3955340   1.727 0.084139 .
monthsep                  0.8251139  0.4161026   1.983 0.047372 *
dayowmon                  0.0102198  0.2011053   0.051 0.959470
dayowthu                 -0.0974596  0.2130142  -0.458 0.647293
dayowtue                 -0.1209930  0.2142669  -0.565 0.572289
dayowwed                  0.0801171  0.2121565   0.378 0.705704
poutcomenonexistent       0.7371133  0.3067481   2.403 0.016262 *
poutcomesuccess           2.1200435  0.6517640   3.253 0.001143 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2152.8  on 3118  degrees of freedom
Residual deviance: 1673.0  on 3074  degrees of freedom
AIC: 1763

Number of Fisher Scoring iterations: 6

> anova(mod1,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: subs

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       3118     2152.8
age        1   11.593      3117     2141.2 0.0006619 ***
campaign   1   25.895      3116     2115.3 3.605e-07 ***
pdays      1  217.090      3115     1898.2 < 2.2e-16 ***
previous   1   16.208      3114     1882.0 5.675e-05 ***
job       11   38.596      3103     1843.4 6.202e-05 ***
marital    3    5.719      3100     1837.7 0.1261335
education  6    7.946      3094     1829.8 0.2420815
```

```
default    1    6.625      3093     1823.1 0.0100567 *
housing    2    0.749      3091     1822.4 0.6877968
loan       1    0.604      3090     1821.8 0.4371156
contact    1   23.816      3089     1798.0 1.060e-06 ***
month      9  109.824      3080     1688.1 < 2.2e-16 ***
dayow      4    1.073      3076     1687.1 0.8985358
poutcome   2   14.116      3074     1673.0 0.0008605 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this first model containing every variables, the AIC value is 1763 with a residual deviance of 1673, which is relatively high. The anova test shows the significance of the majority of variables at 0.05, except for marital, education, housing, loan, and dayow. We generate a second model without the insignificant variables below.

```
> mod2 = glm(subs ~ age+campaign+pdays+previous+contact+month+poutcome,
data=trainsubs, family=binomial)
> summary(mod2)

Call:
glm(formula = subs ~ age + campaign + pdays + previous + contact +
    month + poutcome, family = binomial, data = trainsubs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4743  -0.4335  -0.3799  -0.2357   2.7886

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.1137162  0.7866936  -2.687 0.007213 **
age                0.0130200  0.0057349   2.270 0.023188 *
campaign          -0.0991824  0.0389533  -2.546 0.010891 *
pdays             -0.0005501  0.0006355  -0.866 0.386688
previous           0.4772463  0.1731684   2.756 0.005852 **
contacttelephone  -1.3999576  0.2029369  -6.898 5.26e-12 ***
monthaug          -0.7842393  0.2809443  -2.791 0.005247 **
monthdec           2.2593489  0.6770499   3.337 0.000847 ***
```

```
monthjul           -0.7512743  0.2812757  -2.671 0.007564 **
monthjun            0.6918997  0.3020849   2.290 0.021997 *
monthmar            1.8226295  0.4157253   4.384 1.16e-05 ***
monthmay           -0.4747184  0.2610953  -1.818 0.069037 .
monthnov           -0.8896256  0.3078530  -2.890 0.003855 **
monthoct            0.9741479  0.3881704   2.510 0.012087 *
monthsep            0.9471287  0.4058589   2.334 0.019615 *
poutcomenonexistent  0.6508384  0.2989105   2.177 0.029453 *
poutcomesuccess      1.9230276  0.6304635   3.050 0.002287 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2152.8  on 3118  degrees of freedom
Residual deviance: 1715.5  on 3102  degrees of freedom
AIC: 1749.5

Number of Fisher Scoring iterations: 6

> anova(mod2,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: subs

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      3118     2152.8
age       1   11.593      3117     2141.2 0.0006619 ***
campaign  1   25.895      3116     2115.3 3.605e-07 ***
pdays     1  217.090      3115     1898.2 < 2.2e-16 ***
previous  1   16.208      3114     1882.0 5.675e-05 ***
contact   1   33.527      3113     1848.5 7.027e-09 ***
month     9  120.777      3104     1727.7 < 2.2e-16 ***
poutcome  2   12.202      3102     1715.5 0.0022409 **
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Comparing to mod1, mod2 shows a decrease in AIC (from 1763 to 1749.5) with a much better anova test as every variable appears to be significant. This shows that mod2's prediction is more accurate than mod1.

```
> mod3 = glm(subs ~ age+campaign+pdays+previous+contact+month+poutcome+job, data=trainsubs, family=binomial)
> summary(mod3)
```

```
Call:
glm(formula = subs ~ age + campaign + pdays + previous + contact +
    month + poutcome + job, family = binomial, data = trainsubs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3843  -0.4351  -0.3399  -0.2339   2.9278

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.3487630  0.8292907  -2.832 0.004622 **
age                 0.0179503  0.0070162   2.558 0.010515 *
campaign           -0.0954660  0.0390608  -2.444 0.014524 *
pdays              -0.0003074  0.0006522  -0.471 0.637377
previous            0.5297899  0.1792688   2.955 0.003124 **
contacttelephone   -1.3723221  0.2021562  -6.788 1.13e-11 ***
monthaug           -0.9558745  0.2849464  -3.355 0.000795 ***
monthdec            2.3411177  0.6882744   3.401 0.000670 ***
monthjul           -0.7748563  0.2837698  -2.731 0.006322 **
monthjun            0.6754493  0.3041724   2.221 0.026377 *
monthmar            1.7232002  0.4213694   4.090 4.32e-05 ***
monthmay           -0.4181978  0.2640378  -1.584 0.113227
monthnov           -0.8754351  0.3108090  -2.817 0.004853 **
monthoct            0.7706597  0.3931318   1.960 0.049960 *
monthsep            0.8507661  0.4140448   2.055 0.039901 *
poutcomenonexistent 0.7551912  0.3055169   2.472 0.013442 *
poutcomesuccess     2.1218767  0.6482826   3.273 0.001064 **
jobblue-collar     -0.7394892  0.2142166  -3.452 0.000556 ***
```

```
jobentrepreneur     -0.7472999  0.4500934  -1.660 0.096850 .
jobhousemaid        -0.2493403  0.4253555  -0.586 0.557746
jobmanagement       -0.6323807  0.2846002  -2.222 0.026284 *
jobretired          -0.1349391  0.3354213  -0.402 0.687465
jobself-employed    -0.7006112  0.3809935  -1.839 0.065929 .
jobservices         -0.6326724  0.2747640  -2.303 0.021301 *
jobstudent           0.7045507  0.3613095   1.950 0.051177 .
jobtechnician       -0.0125873  0.1902879  -0.066 0.947260
jobunemployed       -0.4869152  0.4138495  -1.177 0.239375
jobunknown          -0.8101770  0.7539896  -1.075 0.282590
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2152.8  on 3118  degrees of freedom
Residual deviance: 1685.3  on 3091  degrees of freedom
AIC: 1741.3

Number of Fisher Scoring iterations: 6

> anova(mod3,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: subs

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      3118     2152.8
age       1   11.593      3117     2141.2 0.0006619 ***
campaign  1   25.895      3116     2115.3 3.605e-07 ***
pdays     1  217.090      3115     1898.2 < 2.2e-16 ***
previous  1   16.208      3114     1882.0 5.675e-05 ***
contact   1   33.527      3113     1848.5 7.027e-09 ***
month     9  120.777      3104     1727.7 < 2.2e-16 ***
```

```
poutcome  2   12.202      3102     1715.5 0.0022409 **
job      11   30.225      3091     1685.3 0.0014605 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the job variable is moderately significant, we can try adding it to our generalized linear model. As elaborated above, mod3 with variable job shows a further decrease in both AIC and residual deviance. The anova test also indicates the significance of every variable.

```
> mod4 = glm(subs ~ age+campaign+previous+contact+month+poutcome+job,
data=trainsubs, family=binomial)
> summary(mod4)

Call:
glm(formula = subs ~ age + campaign + previous + contact + month +
    poutcome + job, family = binomial, data = trainsubs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3765  -0.4348  -0.3398  -0.2338   2.9284

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.676954   0.453083  -5.908 3.46e-09 ***
age                 0.017904   0.007017   2.551 0.010729 *
campaign           -0.095172   0.039029  -2.439 0.014747 *
previous            0.558983   0.169921   3.290 0.001003 **
contacttelephone   -1.372519   0.202194  -6.788 1.14e-11 ***
monthaug           -0.951862   0.284811  -3.342 0.000832 ***
monthdec            2.341492   0.687924   3.404 0.000665 ***
monthjul           -0.773110   0.283729  -2.725 0.006434 **
monthjun            0.678924   0.304009   2.233 0.025534 *
monthmar            1.732261   0.420327   4.121 3.77e-05 ***
monthmay           -0.416811   0.263933  -1.579 0.114284
monthnov           -0.872043   0.310658  -2.807 0.004999 **
monthoct            0.769004   0.392980   1.957 0.050365 .
monthsep            0.870080   0.411195   2.116 0.034347 *
poutcomenonexistent  0.773588   0.304286   2.542 0.011012 *
```

```
poutcomesuccess      2.397882   0.281197   8.527  < 2e-16 ***
jobblue-collar      -0.739797   0.214271  -3.453 0.000555 ***
jobentrepreneur     -0.750465   0.450487  -1.666 0.095734 .
jobhousemaid        -0.250138   0.425823  -0.587 0.556920
jobmanagement       -0.629643   0.284603  -2.212 0.026942 *
jobretired          -0.128629   0.334827  -0.384 0.700855
jobself-employed    -0.701116   0.381080  -1.840 0.065795 .
jobservices         -0.635291   0.274857  -2.311 0.020814 *
jobstudent           0.717537   0.359257   1.997 0.045795 *
jobtechnician       -0.009203   0.190063  -0.048 0.961382
jobunemployed       -0.489453   0.414143  -1.182 0.237267
jobunknown          -0.811327   0.753702  -1.076 0.281723
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2152.8  on 3118  degrees of freedom
Residual deviance: 1685.5  on 3092  degrees of freedom
AIC: 1739.5

Number of Fisher Scoring iterations: 6

> anova(mod4,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: subs

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      3118     2152.8
age       1   11.593      3117     2141.2 0.0006619 ***
campaign  1   25.895      3116     2115.3 3.605e-07 ***
previous  1  150.630      3115     1964.7 < 2.2e-16 ***
contact   1   33.321      3114     1931.4 7.814e-09 ***
```

```
month      9  133.699      3105      1797.7 < 2.2e-16 ***
poutcome  2   81.418      3103      1716.2 < 2.2e-16 ***
job       11   30.741      3092      1685.5 0.0012107 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As elaborated before, pdays and previous are moderately correlated. Although their correlation test does not go above 0.7, we can try omitting one of these two variables. Mod4 without the pdays variable shows a drop in AIC to 1739.5 with the residual deviance of 1685.5 and significant anova test.

```
>  mod5 = glm(subs ~ 
age+campaign+previous+contact+month+poutcome+job+campaign*previous,
data=trainsubs, family=binomial)
> summary(mod5)

Call:
glm(formula = subs ~ age + campaign + previous + contact + month + 
    poutcome + job + campaign * previous, family = binomial, 
    data = trainsubs)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3620  -0.4338  -0.3387  -0.2345   2.9267  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -2.718704   0.456654  -5.954 2.62e-09 ***
age                 0.018058   0.007028   2.570 0.010184 *  
campaign           -0.078395   0.039300  -1.995 0.046070 *  
previous            0.756689   0.225339   3.358 0.000785 ***
contacttelephone   -1.360859   0.202337  -6.726 1.75e-11 ***
monthaug           -0.970219   0.285639  -3.397 0.000682 ***
monthdec            2.397232   0.696002   3.444 0.000573 ***
monthjul           -0.772466   0.283631  -2.723 0.006460 ** 
monthjun            0.654939   0.305067   2.147 0.031804 *  
monthmar            1.757987   0.421901   4.167 3.09e-05 ***
monthmay           -0.424425   0.264582  -1.604 0.108684    
monthnov           -0.882211   0.311485  -2.832 0.004622 ** 
```

```
monthoct            0.758591   0.392615   1.932 0.053341 .
monthsep            0.869919   0.412316   2.110 0.034873 *
poutcomenonexistent  0.778850   0.307820   2.530 0.011399 *
poutcomesuccess      2.420428   0.283892   8.526  < 2e-16 ***
jobblue-collar      -0.743236   0.214475  -3.465 0.000529 ***
jobentrepreneur     -0.771218   0.452247  -1.705 0.088138 .
jobhousemaid        -0.246177   0.424716  -0.580 0.562166
jobmanagement       -0.637594   0.285714  -2.232 0.025642 *
jobretired          -0.133639   0.334932  -0.399 0.689890
jobself-employed    -0.690891   0.381899  -1.809 0.070437 .
jobservices         -0.643439   0.275269  -2.337 0.019414 *
jobstudent           0.753009   0.358213   2.102 0.035542 *
jobtechnician       -0.004490   0.190268  -0.024 0.981172
jobunemployed       -0.522625   0.419355  -1.246 0.212669
jobunknown          -0.840184   0.759304  -1.107 0.268502
campaign:previous   -0.107376   0.078395  -1.370 0.170788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2152.8  on 3118  degrees of freedom
Residual deviance: 1683.5  on 3091  degrees of freedom
AIC: 1739.5

Number of Fisher Scoring iterations: 6

> anova(mod5,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: subs

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              3118     2152.8
```

```
age              1   11.593      3117     2141.2 0.0006619 ***
campaign          1   25.895      3116     2115.3 3.605e-07 ***
previous          1  150.630      3115     1964.7 < 2.2e-16 ***
contact           1   33.321      3114     1931.4 7.814e-09 ***
month             9  133.699      3105     1797.7 < 2.2e-16 ***
poutcome          2   81.418      3103     1716.2 < 2.2e-16 ***
job              11   30.741      3092     1685.5 0.0012107 **
campaign:previous  1    1.986      3091     1683.5 0.1587474
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now test the significance of the interaction terms in the model. Among all the interaction terms that can be generated from mod4, only campaign*previous shows a better result (mod5): AIC remaining the same, but the residual evidence is smaller. However, the anova test indicates the insignifcance of the variable campaign*previous among the other variables.

Therefore, we can choose mod4 to produce the confusion matrix below.

```
> install.packages("gmodels")
> library(gmodels)

> predmod <- predict(mod4, newdata = testsubs, type='response')
> pred = ifelse(predmod>=0.3,1,0)
> CrossTable(subs[-trainsample],pred[1:1000])
```

```
   Cell Contents
|-------------------------------|
|                             N |
| Chi-square contribution |
|                 N / Row Total |
|                 N / Col Total |
|               N / Table Total |
|-------------------------------|


Total Observations in Table:  1000
```

```
                     | pred[1:1000]
subs[-trainsample] |          0 |          1 | Row Total |
-------------------|------------|------------|-----------|
                 0 |        856 |         34 |       890 |
                   |      0.370 |      6.014 |           |
                   |      0.962 |      0.038 |     0.890 |
                   |      0.909 |      0.586 |           |
                   |      0.856 |      0.034 |           |
-------------------|------------|------------|-----------|
                 1 |         86 |         24 |       110 |
                   |      2.996 |     48.662 |           |
                   |      0.782 |      0.218 |     0.110 |
                   |      0.091 |      0.414 |           |
                   |      0.086 |      0.024 |           |
-------------------|------------|------------|-----------|
      Column Total |        942 |         58 |      1000 |
                   |      0.942 |      0.058 |           |
-------------------|------------|------------|-----------|
```

We classify 856 + 24 = 880 of the 1000 test cases correctly, for an accuracy rate of 88%, and consequently an error rate of 22%. To predict more easily, we change the lower cutoff points to 0.2, 0.15, and 0.1.

```
> pred = ifelse(predmod>=0.2,1,0)
> CrossTable(subs[-trainsample],pred[1:1000])


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1000
```

```
                         | pred[1:1000]
subs[-trainsample] |         0 |          1 | Row Total |
-------------------------|------------|------------|--------------|
                       0 |       848 |         42 |         890 |
                         |     0.799 |      9.719 |             |
                         |     0.953 |      0.047 |       0.890 |
                         |     0.918 |      0.553 |             |
                         |     0.848 |      0.042 |             |
-------------------------|------------|------------|--------------|
                       1 |        76 |         34 |         110 |
                         |     6.468 |     78.638 |             |
                         |     0.691 |      0.309 |       0.110 |
                         |     0.082 |      0.447 |             |
                         |     0.076 |      0.034 |             |
-------------------------|------------|------------|--------------|
            Column Total |       924 |         76 |        1000 |
                         |     0.924 |      0.076 |             |
-------------------------|------------|------------|--------------|
```

The 0.2 cutoff point increases the accuracy rate to 88.2% with a lower false negatives of 76 comparing to the 0.3 cutoff point.

```
> pred = ifelse(predmod>=0.1,1,0)
>
> CrossTable(subs[-trainsample],pred[1:1000])
```

```
Cell Contents
|-------------------------------|
|                             N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------------|
```

```
Total Observations in Table:  1000


                   | pred[1:1000] 
subs[-trainsample] |         0 |         1 | Row Total | 
-------------------|-----------|-----------|-----------|
                 0 |       680 |       210 |       890 | 
                   |     0.951 |     2.652 |           | 
                   |     0.764 |     0.236 |     0.890 | 
                   |     0.924 |     0.795 |           | 
                   |     0.680 |     0.210 |           | 
-------------------|-----------|-----------|-----------|
                 1 |        56 |        54 |       110 | 
                   |     7.695 |    21.453 |           | 
                   |     0.509 |     0.491 |     0.110 | 
                   |     0.076 |     0.205 |           | 
                   |     0.056 |     0.054 |           | 
-------------------|-----------|-----------|-----------|
      Column Total |       736 |       264 |      1000 | 
                   |     0.736 |     0.264 |           | 
-------------------|-----------|-----------|-----------|
```

The 0.1 cuttoff point, however, lower the accuracy rate to 73.4%, but with a much lower false negative of 56.

```
> pred = ifelse(predmod>=0.15,1,0)
> CrossTable(subs[-trainsample],pred[1:1000])


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

```
Total Observations in Table:  1000


                   | pred[1:1000]
subs[-trainsample] |         0 |         1 | Row Total |
-------------------|-----------|-----------|-----------|
                 0 |       827 |        63 |       890 |
                   |     0.904 |     8.044 |           |
                   |     0.929 |     0.071 |     0.890 |
                   |     0.920 |     0.624 |           |
                   |     0.827 |     0.063 |           |
-------------------|-----------|-----------|-----------|
                 1 |        72 |        38 |       110 |
                   |     7.312 |    65.083 |           |
                   |     0.655 |     0.345 |     0.110 |
                   |     0.080 |     0.376 |           |
                   |     0.072 |     0.038 |           |
-------------------|-----------|-----------|-----------|
      Column Total |       899 |       101 |      1000 |
                   |     0.899 |     0.101 |           |
-------------------|-----------|-----------|-----------|
```

The 0.15 cutoff point indicates an accuracy rate of 86.5% (higher than 0.1 cutoff point) with 72 false negative (lower than 0.2 cutoff point).

Overall, the cutoff point of 0.2 shows the best prediction of the mod4 logistic regression (88.2% accuracy and 76 false negatives).

kNN classification

```
> install.packages("class")
> install.packages("rpart")
> library(class)

> normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x))) }
> bankdata <- as.data.frame(lapply(bankdata, normalize))

> job<-as.numeric(bank.additional$job)
```

```
> marital<-as.numeric(bank.additional$marital)
> education<-as.numeric(bank.additional$education)
> default<-as.numeric(bank.additional$default)
> housing<-as.numeric(bank.additional$housing)
> loan<-as.numeric(bank.additional$loan)
> contact<-as.numeric(bank.additional$contact)
> month<-as.numeric(bank.additional$month)
> dayow<-as.numeric(bank.additional$day_of_week)
> poutcome<-as.numeric(bank.additional$poutcome)

>bankdata=data.frame(age,job,marital,education,default,housing,loan,contact,month,da
yow,bank.additional$duration,campaign,pdays,previous,poutcome,subs)

> traink<-bankdata[1:3118,]
> trainlabel<-bankdata[1:3118,15]
> testk<-bankdata[3119:4119,]
> testlabel<-bankdata[3119:4119,15]
>  train<-traink[complete.cases(traink),]
> test<-testk[complete.cases(testk),]
> train_lable<-train$subs
```

First, we test the knn model with k=3 below.

```
> model <- knn(train=train, test=test, cl=train_lable, k=3)
> test_label<-test$subs
> library(gmodels)
> CrossTable(test_label, model, chisq=FALSE)
```

```
   Cell Contents
|-------------------------------|
|                             N |
| Chi-square contribution |
|               N / Row Total |
|                N / Col Total |
|             N / Table Total |
|--------------------------------|
```

```
Total Observations in Table:  1000


             | model
  test_label |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |       859 |        36 |       895 |
             |     0.236 |     3.978 |           |
             |     0.960 |     0.040 |     0.895 |
             |     0.910 |     0.643 |           |
             |     0.859 |     0.036 |           |
-------------|-----------|-----------|-----------|
           1 |        85 |        20 |       105 |
             |     2.011 |    33.907 |           |
             |     0.810 |     0.190 |     0.105 |
             |     0.090 |     0.357 |           |
             |     0.085 |     0.020 |           |
-------------|-----------|-----------|-----------|
Column Total |       944 |        56 |      1000 |
             |     0.944 |     0.056 |           |
-------------|-----------|-----------|-----------|
```

We classify 859 + 20 = 879 of the 1000 test cases correctly, for an accuracy rate of 87.9%, and consequently an error rate of 22.1%. To predict more easily, we change k with its accuracy rate and false negative as shown respectively below.

k=2: 85.7%, 80
k=3: 87.9%, 85
k=4: 87.7%, 85
k=7: 88.9%, 82
k=8: 88.8%, 83
k=9: 89.1%, 84
k=10: 89.3%, 82
k=11: 89.6%, 83

```
> model2 <- knn(train=train, test=test, cl=train_lable, k=10)
> CrossTable(test_label, model2, chisq=FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1000


             | model2
  test_label |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |       870 |        25 |       895 |
             |     0.379 |     7.508 |           |
             |     0.972 |     0.028 |     0.895 |
             |     0.914 |     0.521 |           |
             |     0.870 |     0.025 |           |
-------------|-----------|-----------|-----------|
           1 |        82 |        23 |       105 |
             |     3.227 |    64.000 |           |
             |     0.781 |     0.219 |     0.105 |
             |     0.086 |     0.479 |           |
             |     0.082 |     0.023 |           |
-------------|-----------|-----------|-----------|
Column Total |       952 |        48 |      1000 |
             |     0.952 |     0.048 |           |
-------------|-----------|-----------|-----------|
```

K of 10 indicates a significantly high accuracy rate of 89.3% with 82 false negative. Comparing to k=2 that has the lowest number of false negative, k=10 shows a much higher accuracy rate.

We can conclude that the best knn model has k=10.

In conclusion, the two classification method (logistic regression and knn) both show significant predictions based on the observations given. The difference in these two method is not huge as they both reach an accuracy rate of approximately 89% and false negatives of 80 out of 1000.