Predicting the Odds of a Term Deposit Subscription

Logistic Regression

Li Sun

Introduction

The data set I will be investigating is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The objective of this project is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Objective

The original data was downloaded from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#. The dataset contains 21 variables, which includes 20 predictors and 1 response variable, and 45211 observations without missing values. Each observation is composed of client bank data, data related to the last contact of the current campaign, social and economic context attributes and desired target variable. The desired target is the output variable, which is binary 'yes' or 'no'. First a model will be determined to predict whether the client will subscribe to a term deposit. After confirming the model, a model will be constructed using all of the predictor variables and using a stepwise procedure considering the interactions of each predictor . The goodness of fit of that model will be implemented using a chi-square procedure. And McFadden Pseudo R^2 will be used to estimate the predictive power of our model. Then, the estimated effects of the predictor variables will be plotted to determine what they individually contribute to the model. Lastly, the accuracy of the model with regards to unknown variables will be measured and an ROC curve will be created to compare the rates of false positive predictions with false negative predictions.

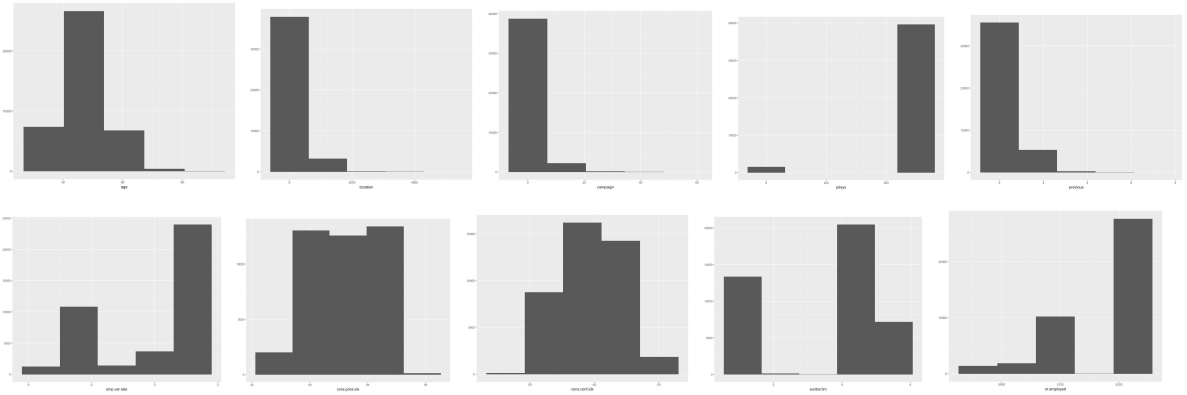
Data Description and Visualization

There are total 20 predictors. They are "age" ,"job", "marital" ,"education","default","housing", "loan","contact","month","day_of_week","duration","campaign","pdays","previous", "poutcome","emp.var.rate","cons.price.idx","cons.conf.idx","euribor3m","nr.employed". Among

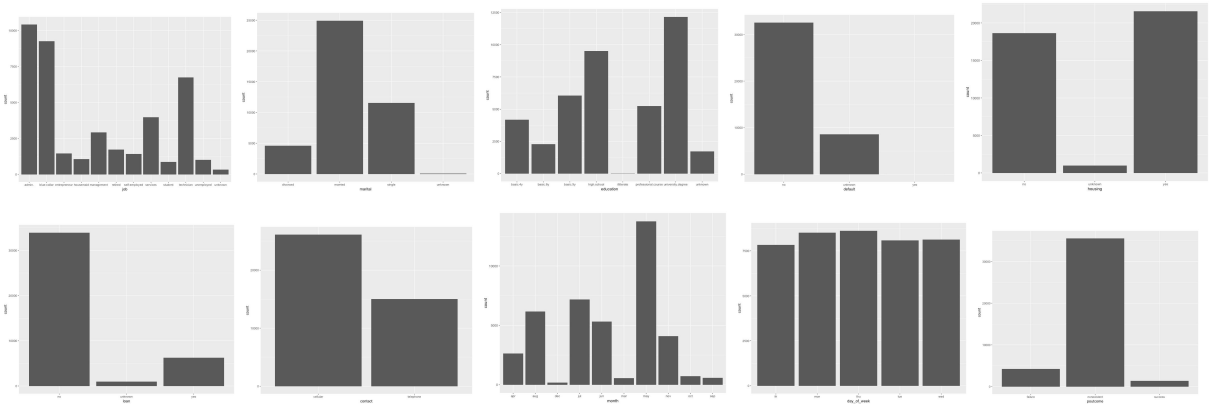
these 20 predictors, there are 10 each for continuous variable and categorical variables. There is 1 target variable "y". The positive class is ‘no’.

The histograms of each variable are plotted as follows. Most continuous variables do not follow normal distribution. The target feature contains unbalanced data with about 8 times more ‘no’ subscription than ‘yes’ subscription.

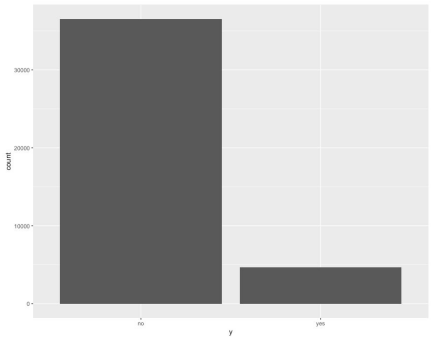
1. Histogram of continuous variable



2. Histogram of categorical variables

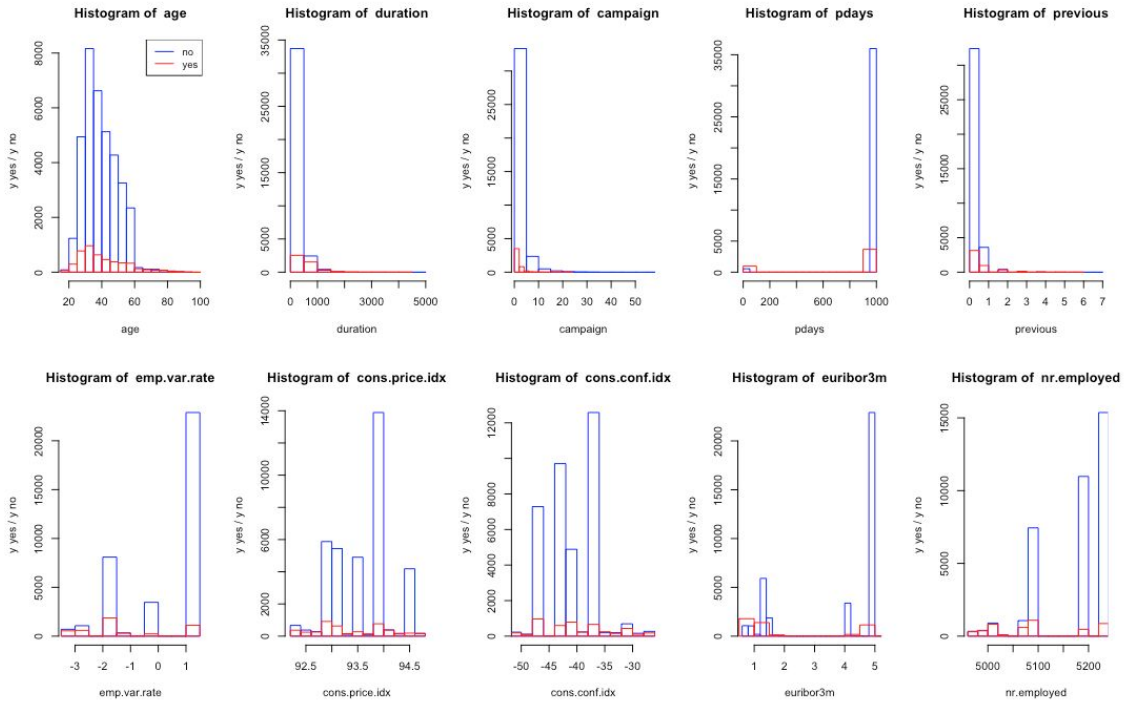


3. Histogram of target variable

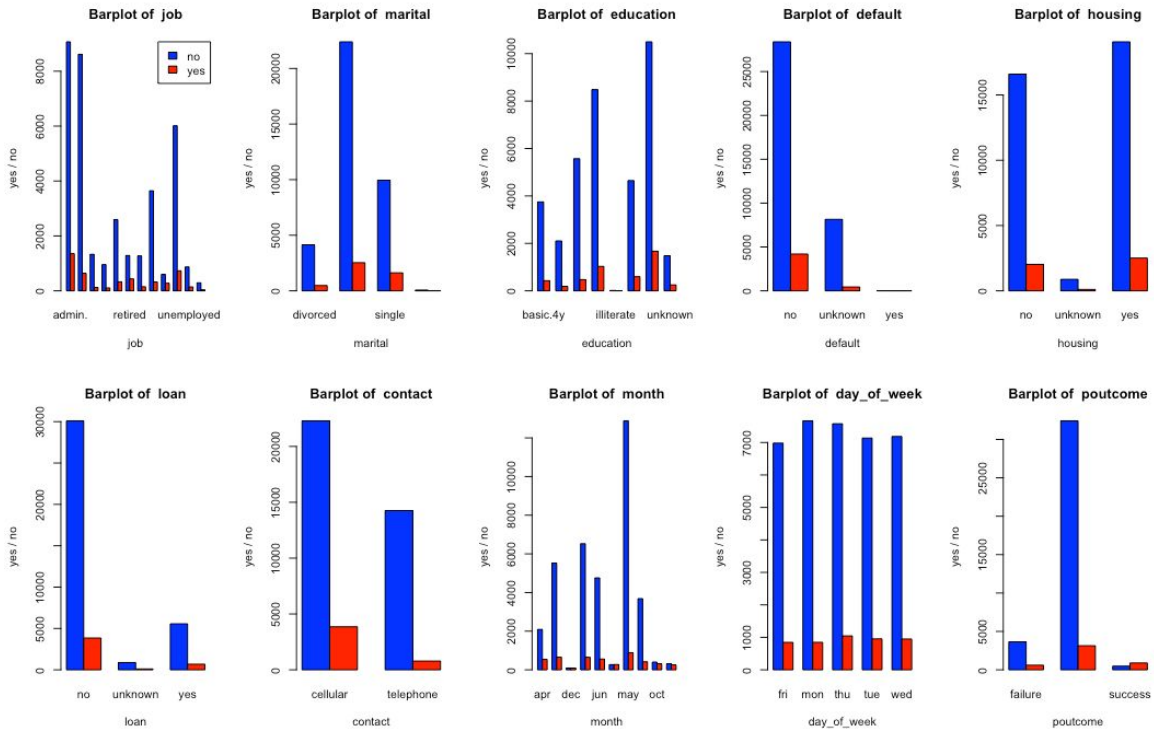


The histogram of subsamples are plotted as follows.

1. Histogram of subsamples for continuous variables



2. Histogram of subsamples for categorical variables



Among continuous variables, predictors odd (answer ‘no’/ answer ‘yes’) of "age" and odd of "previous" do not change obviously while age and previous increase. Among categorical variables, predictors odd between each level in "marital", "education","housing" and "loan" are similar. So variables "age" , "previous", "marital", "education","housing" and "loan" are potential non-significant factors in the logistic model.

Maximum Likelihood Estimation

First of all, since the target feature is binary, the logistic regression algorithm is selected. So a logistic regression model is going to be trained.

To obtain the model of interest, we need to find the values of the coefficients that solve:

$$\prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_i x_i}}{1+e^{\beta_0+\beta_i x_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_i x_i}}\right)^{1-y_i}$$

We cannot solve this equation by hand. As a result, we use statistical software to obtain the values of the coefficients.

Model Selection

1. Using all the predictor variables to determine the model, we come up with the following model:

```
 Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.366e+02  3.831e+01  -6.176 6.56e-10 ***
age                          1.966e-04  2.434e-03   0.081 0.935624
jobblue-collar              -2.347e-01  7.988e-02  -2.939 0.003295 **
jobentrepreneur             -1.780e-01  1.260e-01  -1.413 0.157566
jobhousemaid                -2.432e-02  1.478e-01  -0.165 0.869320
jobmanagement               -5.614e-02  8.536e-02  -0.658 0.510710
jobretired                   2.858e-01  1.071e-01   2.669 0.007606 **
jobself-employed            -1.578e-01  1.178e-01  -1.340 0.180396
jobservices                 -1.399e-01  8.610e-02  -1.624 0.104286
---
---
---
emp.var.rate                -1.758e+00  1.420e-01 -12.380  < 2e-16 ***
cons.price.idx               2.190e+00  2.524e-01   8.679  < 2e-16 ***
cons.conf.idx                2.069e-02  7.768e-03   2.664 0.007733 **
euribor3m                    3.316e-01  1.300e-01   2.551 0.010737 *
nr.employed                  5.413e-03  3.115e-03   1.738 0.082275 .
```

The predictor variables with p-value less than 0.05 are significant in this model which tells us that there is evidence to suggest that their slopes are different from zero (contribute to the overall adequacy of the model).

2. Using stepwise procedure with interactions and then removing the insignificant terms (see note at the end). Because there are 20 predictors in the dataset, this stage contains 2 steps: the first stepwise procedure is used on 20 main factors without interaction and significant terms were selected. Then stepwise procedure is applied with interaction only on selected main factors.

Step 1: model selection without interaction

After the first procedure, the selected model without interaction is obtained as follows:

```
Step:  AIC=17170.27
y ~ duration + nr.employed + month + poutcome + emp.var.rate +
  cons.price.idx + job + contact + euribor3m + default + day_of_week +
  pdays + campaign + cons.conf.idx
```

```
Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.323e+02  3.822e+01  -6.078 1.22e-09 ***
duration             4.702e-03  7.450e-05  63.116  < 2e-16 ***
nr.employed          5.116e-03  3.108e-03   1.646 0.099749 .
monthaug             8.674e-01  1.202e-01   7.217 5.32e-13 ***
monthdec             3.019e-01  2.088e-01   1.446 0.148150
monthjul             1.361e-01  9.598e-02   1.418 0.156227
monthjun            -5.115e-01  1.258e-01  -4.068 4.75e-05 ***
monthmar             2.019e+00  1.441e-01  14.007  < 2e-16 ***
monthmay            -4.553e-01  8.233e-02  -5.530 3.20e-08 ***
monthnov            -4.253e-01  1.208e-01  -3.522 0.000429 ***
monthoct             1.803e-01  1.535e-01   1.175 0.240163
monthsep             3.607e-01  1.793e-01   2.012 0.044220 *
poutcomenonexistent  5.026e-01  6.411e-02   7.840 4.52e-15 ***
poutcomesuccess      1.036e+00  2.040e-01   5.081 3.76e-07 ***
emp.var.rate        -1.752e+00  1.419e-01 -12.350  < 2e-16 ***
cons.price.idx       2.160e+00  2.516e-01   8.586  < 2e-16 ***
jobblue-collar      -3.329e-01  6.586e-02  -5.055 4.31e-07 ***
jobentrepreneur     -2.029e-01  1.244e-01  -1.631 0.102947
jobhousemaid        -1.118e-01  1.409e-01  -0.793 0.427695
jobmanagement       -4.317e-02  8.341e-02  -0.518 0.604744
jobretired           2.047e-01  8.381e-02   2.442 0.014610 *
jobself-employed    -1.509e-01  1.169e-01  -1.291 0.196739
jobservices         -2.128e-01  8.168e-02  -2.605 0.009185 **
jobstudent           1.770e-01  1.018e-01   1.739 0.081968 .
jobtechnician       -2.728e-02  6.348e-02  -0.430 0.667323
jobunemployed       -3.686e-02  1.261e-01  -0.292 0.769998
jobunknown          -9.286e-02  2.344e-01  -0.396 0.691981
contacttelephone    -6.421e-01  7.674e-02  -8.368  < 2e-16 ***
```

```
euribor3m            3.422e-01  1.297e-01   2.638 0.008333 **
defaultunknown      -3.106e-01  6.636e-02  -4.681 2.86e-06 ***
defaultyes          -7.326e+00  1.134e+02  -0.065 0.948510
day_of_weekmon      -1.172e-01  6.604e-02  -1.775 0.075831 .
day_of_weekthu       5.858e-02  6.401e-02   0.915 0.360104
day_of_weektue       9.500e-02  6.575e-02   1.445 0.148527
day_of_weekwed       1.727e-01  6.562e-02   2.632 0.008488 **
pdays               -8.445e-04  2.036e-04  -4.149 3.34e-05 ***
campaign            -3.981e-02  1.155e-02  -3.448 0.000564 ***
cons.conf.idx        2.039e-02  7.734e-03   2.637 0.008377 **
  ---
Residual deviance: 17094  on 41150  degrees of freedom
AIC: 17170
```

The ANOVA with chi-squared test is implemented to test the goodness of fit on the model with main factors. The ANOVA table is shown as below:

```
Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           41187      28999
duration        1   4892.6     41186      24106 < 2.2e-16 ***
  nr.employed     1   5150.7     41185      18956 < 2.2e-16 ***
  month           9    923.4     41176      18032 < 2.2e-16 ***
  poutcome        2    497.2     41174      17535 < 2.2e-16 ***
  emp.var.rate    1    140.0     41173      17395 < 2.2e-16 ***
  cons.price.idx  1     60.3     41172      17335 7.983e-15 ***
  job            11     67.9     41161      17267 3.114e-10 ***
  contact         1     43.0     41160      17224 5.385e-11 ***
  euribor3m       1     44.7     41159      17179 2.302e-11 ***
  default         2     23.3     41157      17156 8.884e-06 ***
  day_of_week     4     24.7     41153      17131 5.679e-05 ***
  pdays           1     17.1     41152      17114 3.585e-05 ***
  campaign        1     12.7     41151      17101  0.000372 ***
  cons.conf.idx   1      7.0     41150      17094  0.008373 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It shows that the fitted model with selected main factors is a good fit.

Step 2: model selection with interaction on selected main factors

The stepwise procedures were performed again with interaction, which took hours for implementation. The selected model with interaction is displayed as follows:

```
# Step:  AIC=16333.39
# y ~ duration + nr.employed + month + poutcome + emp.var.rate +
#   contact + job + day_of_week + campaign + pdays + default +
#   euribor3m + nr.employed:month + duration:month + month:emp.var.rate +
#   duration:emp.var.rate + month:day_of_week + duration:contact +
#   nr.employed:contact + month:campaign + month:default + poutcome:emp.var.rate +
#   campaign:pdays + duration:poutcome + emp.var.rate:default +
```

```
#   day_of_week:pdays + contact:pdays + emp.var.rate:day_of_week +
#   month:euribor3m + duration:euribor3m + poutcome:euribor3m +
#   nr.employed:euribor3m + emp.var.rate:contact + contact:day_of_week +
#   job:euribor3m + contact:campaign
```

```
Coefficients: (13 not defined because of singularities)
Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -7.516e+02  1.202e+02  -6.253 4.02e-10 ***
  duration                         8.503e-03  7.577e-04  11.221  < 2e-16 ***
  nr.employed                      1.519e-01  2.420e-02   6.275 3.49e-10 ***
  monthaug                         2.693e+02  2.557e+02   1.053 0.292364
  monthdec                        -1.289e+03  1.452e+03  -0.888 0.374639
  monthjul                         5.669e+02  2.338e+02   2.425 0.015294 *
  monthjun                         6.911e+02  1.575e+02   4.388 1.14e-05 ***
  monthmar                         8.052e+02  1.286e+02   6.263 3.77e-10 ***
  monthmay                         6.563e+02  1.176e+02   5.581 2.39e-08 ***
  monthnov                         9.413e+02  1.227e+02   7.670 1.73e-14 ***
  monthoct                         1.587e+02  2.714e+02   0.585 0.558781
  monthsep                         2.580e+02  3.223e+02   0.800 0.423491
  poutcomenonexistent              1.844e+00  3.391e-01   5.437 5.42e-08 ***
  poutcomesuccess                  2.558e+00  4.535e-01   5.641 1.69e-08 ***
  emp.var.rate                     1.016e+00  1.396e+00   0.728 0.466868
  ---
  ---
  jobtechnician:euribor3m         -7.471e-03  3.697e-02  -0.202 0.839857
  jobunemployed:euribor3m          2.481e-03  8.129e-02   0.031 0.975658
  jobunknown:euribor3m             2.285e-01  1.230e-01   1.858 0.063197 .
  Contacttelephone:campaign        4.628e-02  2.901e-02   1.595 0.110660
```

By looking up the p-value from above summary table, the p-value of variable 'emp.var.rate' is greater than 0.05, which indicates it is not significant. The variable 'emp.var.rate' is removed from the model.

Goodness of Fit

The ANOVA with chi-squared is tested by adding the terms of the model sequentially.

```
Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
  NULL                              41187      28999
  duration               1   4892.6     41186      24106 < 2.2e-16 ***
  nr.employed            1   5150.7     41185      18956 < 2.2e-16 ***
  month                  9    923.4     41176      18032 < 2.2e-16 ***
  poutcome               2    497.2     41174      17535 < 2.2e-16 ***
  contact                1     61.2     41173      17474 5.129e-15 ***
  job                   11     72.9     41162      17401 3.351e-11 ***
  day_of_week            4     26.0     41158      17375 3.216e-05 ***
  campaign               1     19.7     41157      17355 9.275e-06 ***
  pdays                  1     16.6     41156      17339 4.674e-05 ***
  default                2     28.6     41154      17310 6.176e-07 ***
  euribor3m              1     27.3     41153      17283 1.715e-07 ***
  nr.employed:month      9    270.7     41144      17012 < 2.2e-16 ***
  duration:month         9    241.6     41135      16770 < 2.2e-16 ***
  month:day_of_week     36    189.9     41099      16580 < 2.2e-16 ***
```

```
duration:contact        1      0.7     41098      16580 0.3930268
nr.employed:contact     1     37.2     41097      16543 1.074e-09 ***
month:campaign          9     36.2     41088      16506 3.717e-05 ***
month:default          10     32.7     41078      16474 0.0003083 ***
campaign:pdays          1      8.8     41077      16465 0.0030674 **
duration:poutcome       2      4.0     41075      16461 0.1374390
day_of_week:pdays       4     11.9     41071      16449 0.0181411 *
contact:pdays           1      3.1     41070      16446 0.0767107 .
month:euribor3m         9    201.6     41061      16244 < 2.2e-16 ***
duration:euribor3m      1     24.1     41060      16220 9.134e-07 ***
poutcome:euribor3m      2      0.8     41058      16220 0.6715170
nr.employed:euribor3m   1      2.9     41057      16216 0.0863545 .
contact:day_of_week     4      8.2     41053      16208 0.0836825 .
job:euribor3m          11     24.3     41042      16184 0.0114478 *
contact:campaign        1      1.1     41041      16183 0.2991032
```

ANOVA test is iteratively executed to remove nonsignificant variables in the ANOVA test one by one. In this project, the iteration time is 6. Since there are too many terms in the final model, only the variables in the final model is listed as follows:

y ~ duration + nr.employed + month + poutcome +contact + job + day_of_week + campaign + pdays + default + euribor3m + nr.employed:month + duration:month + month:day_of_week +nr.employed:contact + month:campaign + month:default +campaign:pdays + day_of_week:pdays + month:euribor3m + duration:euribor3m + job:euribor3m

The coefficients of significant terms are listed in the following summary table:

```
Coefficients: (8 not defined because of singularities)
Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -6.306e+02  1.039e+02  -6.068 1.29e-09 ***
  duration                      3.032e-03  2.320e-04  13.073  < 2e-16 ***
  nr.employed                   1.280e-01  2.107e-02   6.075 1.24e-09 ***
  monthaug                      6.711e+02  1.044e+02   6.426 1.31e-10 ***
  monthdec                     -1.238e+03  1.335e+03  -0.927 0.353756
  monthjul                      6.714e+02  1.050e+02   6.396 1.60e-10 ***
  monthjun                      5.945e+02  1.051e+02   5.658 1.53e-08 ***
  monthmar                      8.035e+02  1.286e+02   6.249 4.13e-10 ***
  monthmay                      7.788e+02  1.048e+02   7.431 1.08e-13 ***
  monthnov                      6.412e+02  1.052e+02   6.098 1.08e-09 ***
  ---
  ---
  jobstudent:euribor3m          9.943e-02  9.557e-02   1.040 0.298138
  jobtechnician:euribor3m      -9.405e-03  3.666e-02  -0.257 0.797508
  jobunemployed:euribor3m      -1.103e-02  8.115e-02  -0.136 0.891921
  jobunknown:euribor3m          2.380e-01  1.229e-01   1.938 0.052681 .
```

Interpreting the Model

Since there are too many variables selected in the final model, the model is interpreted in general.

Holding the other variables constant:

A. For a main continuous variable
   1. If it's coefficient is positive, for each unit it grows, the odds of having 'no' answer increase by exp(coefficient).
   2. If it's coefficient is negative, for each unit it grows, the odds of having 'no' answer decrease by exp(coefficient).

B. For each dummy variable
   1. If it's coefficient is positive, for each unit it grows, compared with reference level, the odds of having 'no' answer increase by exp(coefficient).
   2. If it's coefficient is negative, for each unit it grows, compared with reference level, the odds of having 'no' answer decrease by exp(coefficient).

Goodness of Fit

The ANOVA table is created by adding the terms of the model sequentially.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			41187	28999		
duration	1	4892.6	41186	24106	< 2.2e-16	***
nr.employed	1	5150.7	41185	18956	< 2.2e-16	***
month	9	923.4	41176	18032	< 2.2e-16	***
poutcome	2	497.2	41174	17535	< 2.2e-16	***
contact	1	61.2	41173	17474	5.129e-15	***
job	11	72.9	41162	17401	3.351e-11	***
day_of_week	4	26.0	41158	17375	3.216e-05	***
campaign	1	19.7	41157	17355	9.275e-06	***
pdays	1	16.6	41156	17339	4.674e-05	***
default	2	28.6	41154	17310	6.176e-07	***
euribor3m	1	27.3	41153	17283	1.715e-07	***
nr.employed:month	9	270.7	41144	17012	< 2.2e-16	***
duration:month	9	241.6	41135	16770	< 2.2e-16	***
month:day_of_week	36	189.9	41099	16580	< 2.2e-16	***
nr.employed:contact	1	37.4	41098	16543	9.643e-10	***
month:campaign	9	36.1	41089	16507	3.740e-05	***
month:default	10	32.6	41079	16474	0.0003124	***
campaign:pdays	1	8.7	41078	16466	0.0030970	**
day_of_week:pdays	4	11.8	41074	16454	0.0187934	*
month:euribor3m	9	195.8	41065	16258	< 2.2e-16	***
duration:euribor3m	1	7.7	41064	16250	0.0053819	**
job:euribor3m	11	24.2	41053	16226	0.0117907	*

Since the residual deviance of the model decreases with each added predictor variable along with the fact that the p-values are significant, there is evidence that our fitted model is a good fit.

Cook's distances for the data are created, yet none of them are significantly large. This indicates that there are no influential points.

```
named integer(0)
```

We can also perform Wald Tests on each of the predictors to check and see if they are needed in the model.

```
[1] "duration"
F =  170.8934  on  1  and  41053  df: p= < 2.22e-16
[1] "nr.employed"
F =  36.90044  on  1  and  41053  df: p= 1.2541e-09
[1] "month"
F =  17.23507  on  9  and  41053  df: p= < 2.22e-16
---
[1] "duration:euribor3m"
F =  0.1217418  on  1  and  41053  df: p= 0.72715
[1] "job:euribor3m"
F =  18.08437  on  11  and  41053  df: p= < 2.22e-16
```

Like the results before, these p-values indicate that each of the predictor variables are significant in prediction, except term 'duration:euribor3m', which should be removed from our final model.

Collinearity

After assessing the goodness of fit of the logistic model, we will check to see if there is any collinearity between the predictor variables. We will check this using Variance Inflation Factors. If any is greater than 10, we will remove that variable from the model.

```
> vif(fit.interaction.6)
Error in vif.default(fit.interaction.6) :
  there are aliased coefficients in the model
```

```
alias(fit.interaction.6)
# ...#               day_of_weekthu day_of_weektue day_of_weekwed campaign   pdays defaultunknown
defaultyes
# monthdec:defaultyes  0              0              0              0        0     0              0
# monthjul:defaultyes  0              0              0              0        0     0              0
# monthjun:defaultyes  0              0              0              0        0     0              0
# monthmar:defaultyes  0              0              0              0        0     0              0
# monthmay:defaultyes  0              0              0              0        0     0              0
# monthnov:defaultyes  0              0              0              0        0     0              1
# monthoct:defaultyes  0              0              0              0        0     0              0
# ...
#               monthsep:defaultunknown monthaug:defaultyes campaign:pdays day_of_weekmon:pdays
# monthdec:defaultyes  0                       0                  0              0
```

```
# monthjul:defaultyes  0                    0                  0             0
# monthjun:defaultyes  0                    0                  0             0
# monthmar:defaultyes  0                    0                  0             0
# monthmay:defaultyes  0                    0                  0             0
# monthnov:defaultyes  0                   -1                  0             0
# monthoct:defaultyes  0                    0                  0             0
```

There are aliased coefficients in the final model. And the aliased coefficients are detected in the above table. The further method is needed to deal with collinearity.

Power

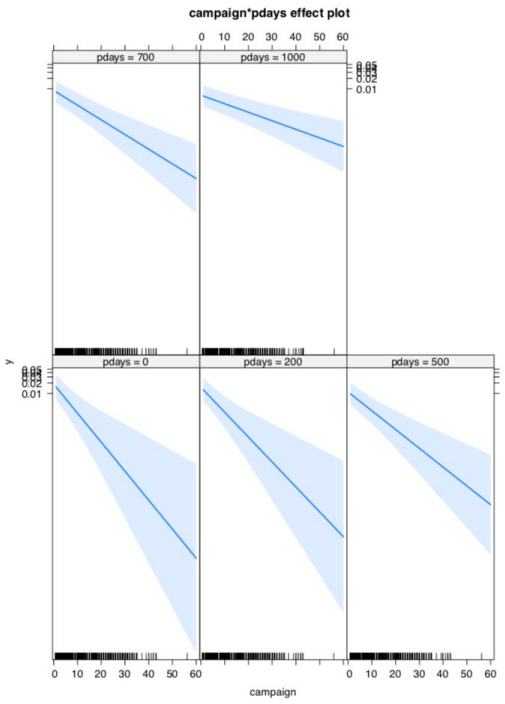
To assess the predictive power of the model, we use the McFadden R^2.

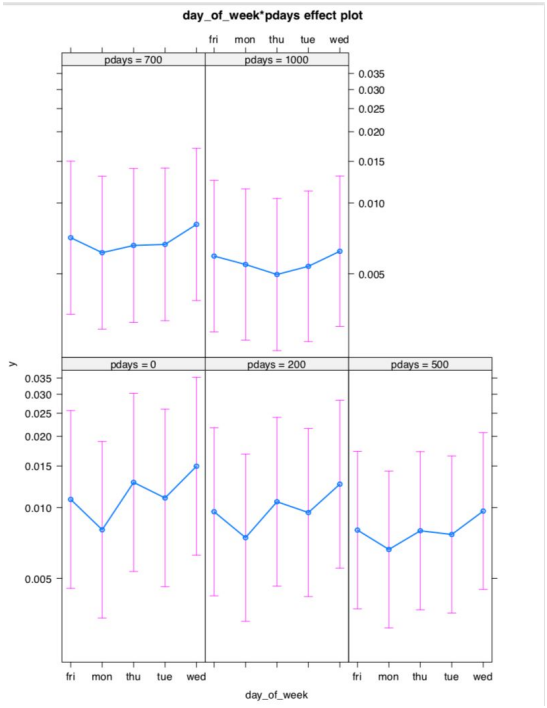
```
      llh        llhNull            G2      McFadden          r2ML          r2CU
-8.112964e+03 -1.449936e+04  1.277280e+04  4.404606e-01  2.666335e-01  5.275425e-01
```

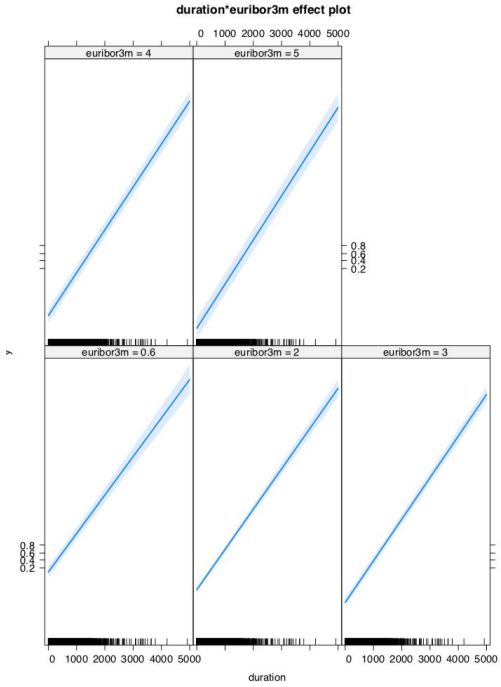
A McFadden R^2 value between 0.2 and 0.4 is considered good. Therefore, since our McFadden $R^2$is .4404 and the model contains lots of variables, we can say that the model selected is a good fit for predicting subscription .

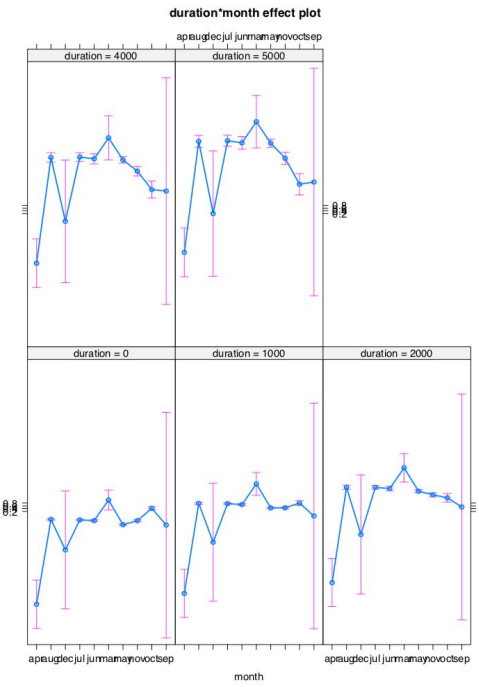
Effect Size

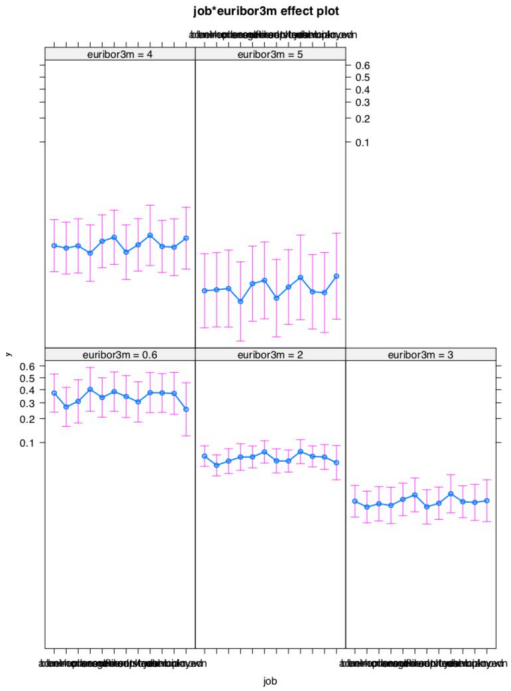
To determine the effect of the individual predictor variables on the chances of subscription, let's make a plot to determine the effect each one has, individually, on subscription:

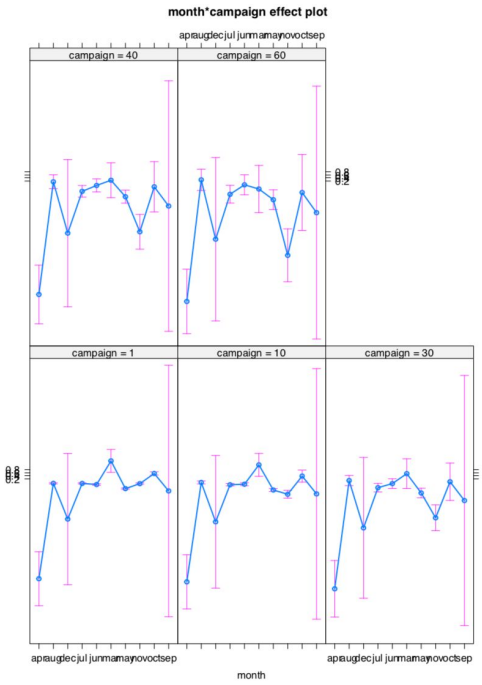

campaign*pdays effect plot
pdays = 700
pdays = 1000
pdays = 0
pdays = 200
pdays = 500
campaign
y

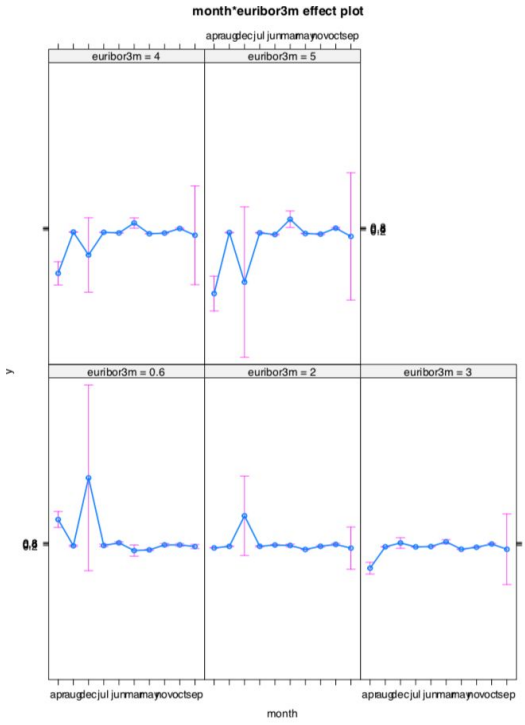

day_of_week*pdays effect plot
pdays = 700
pdays = 1000
pdays = 0
pdays = 200
pdays = 500
fri mon thu tue wed
day_of_week
y

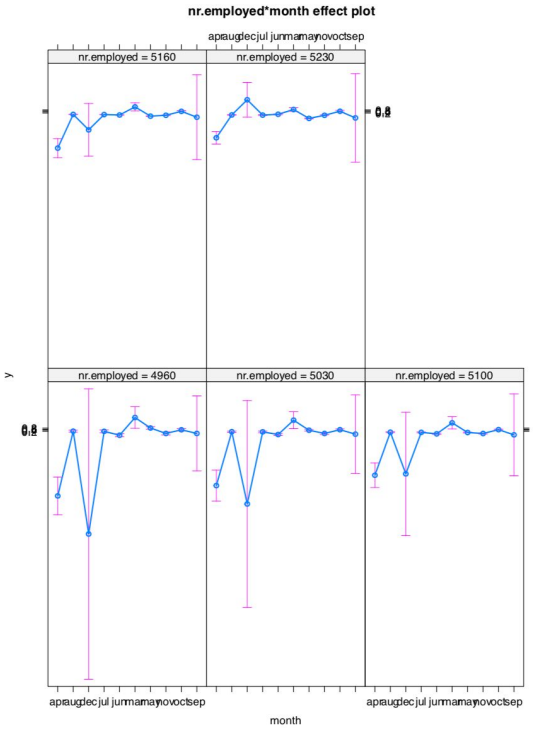

duration*euribor3m effect plot
euribor3m = 4
euribor3m = 5
euribor3m = 0.6
euribor3m = 2
euribor3m = 3
duration
y

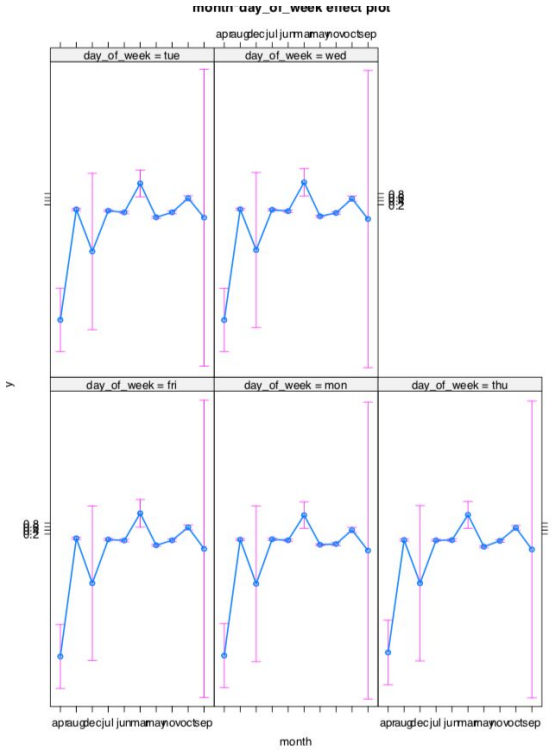

duration*month effect plot
duration = 4000
duration = 5000
duration = 0
duration = 1000
duration = 2000
apr aug dec jul jun mar may nov oct sep
month

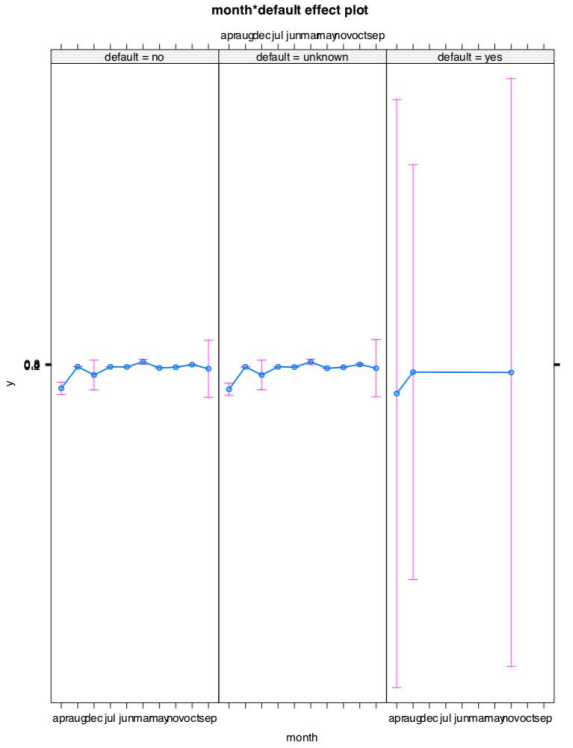

job*euribor3m effect plot
euribor3m = 4
euribor3m = 5
euribor3m = 0.6
euribor3m = 2
euribor3m = 3
y
job

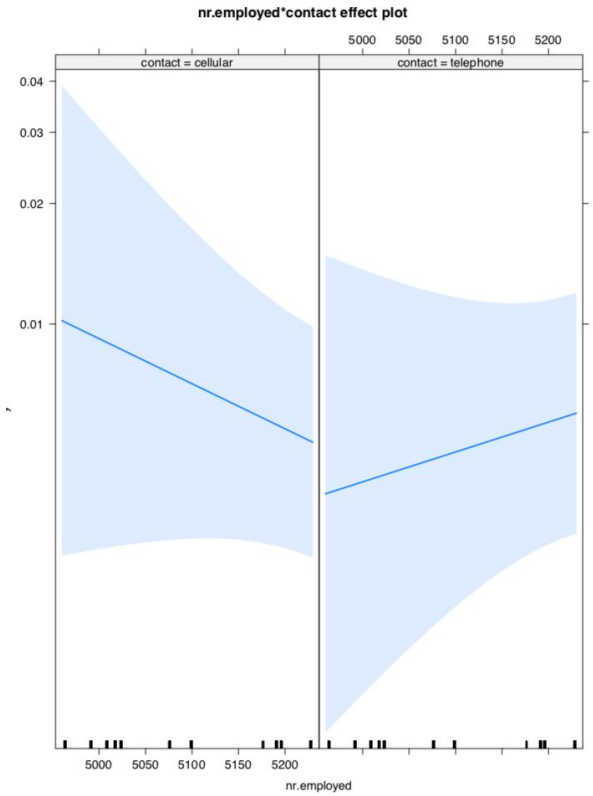

month*campaign effect plot
campaign = 40
campaign = 60
campaign = 1
campaign = 10
campaign = 30
aprauguecjul junmamaynovoctsep
month

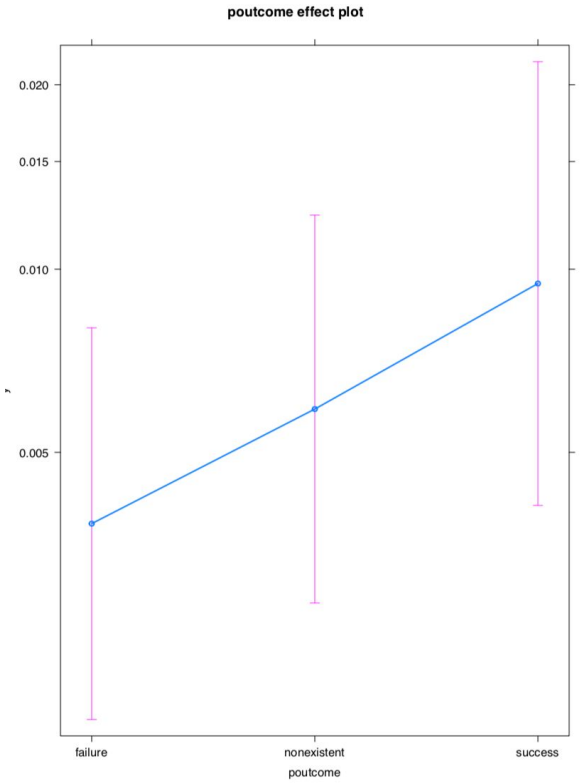

month*euribor3m effect plot
euribor3m = 4
euribor3m = 5
euribor3m = 0.6
euribor3m = 2
euribor3m = 3
y
aprauguecjul junmamaynovoctsep
month


nr.employed*month effect plot
nr.employed = 5160
nr.employed = 5230
nr.employed = 4960
nr.employed = 5030
nr.employed = 5100
y
aprauguecjul junmamaynovoctsep
month


month*day_of_week effect plot
day_of_week = tue
day_of_week = wed
day_of_week = fri
day_of_week = mon
day_of_week = thu
apraugdecjul junmamaynovoctsep
month
y


month*default effect plot
default = no
default = unknown
default = yes
apraugdecjul junmamaynovoctsep
month
y


nr.employed*contact effect plot
contact = cellular
contact = telephone
5000 5050 5100 5150 5200
0.04
0.03
0.02
0.01
nr.employed
y


poutcome effect plot
0.020
0.015
0.010
0.005
failure
nonexistent
success
poutcome
y

These plots are consistent with previous test conclusions. For example, the coefficients for ‘poutcomenonexistent’and ‘poutcomesuccess’ are .4368 and .9155 respectively. They represent that compared with reference level ‘failure’, the chance of not subscripted for ‘existent’ and ‘success’ is higher, and level ‘success’ has the highest chance.

Cross Validation

Using Cross Validation techniques on the model, we obtain the following results:

```
Confusion Matrix and Statistics
          Reference
Prediction   no  yes
       no  7139  511
      yes  170   417
Accuracy : 0.9173
95% CI : (0.9112, 0.9232)
No Information Rate : 0.8873
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.5075
Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9767
Specificity : 0.4494
Pos Pred Value : 0.9332
Neg Pred Value : 0.7104
Prevalence : 0.8873
Detection Rate : 0.8667
Detection Prevalence : 0.9287
Balanced Accuracy : 0.7130

'Positive' Class : no
```

The overall accuracy of the model to predict survival rate is .9173 with a sensitivity (the proportion who not subscripted who were predicted not to subscript based on the model) is .9767 yet the specificity (the proportion who subscribe a term who were predicted to subscribe a term based on the model) was .4494. This indicates that our model does a better job at correctly predicting the chances that someone not subscribed than predicting the chances that someone subscribed.

Variable of Importance

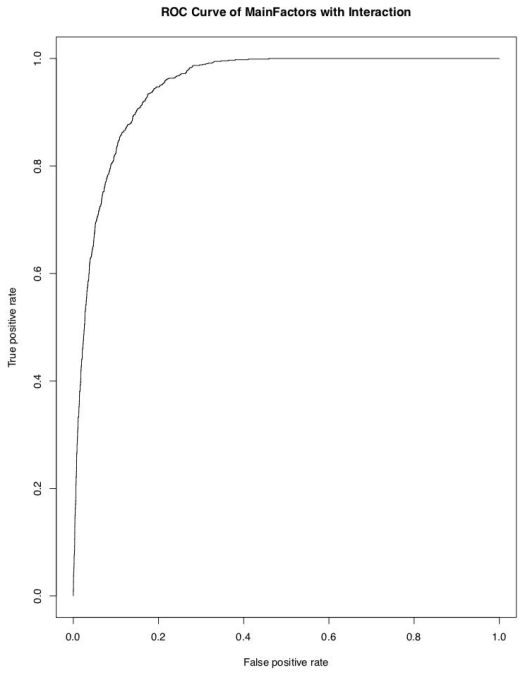
We can assess the importance of individual predictors in the model. Based on the sample of 45211 observations.

```
only 20 most important variables shown (out of 134)
```

```
Overall
duration                   100.00
`duration:monthmay`         63.83
`monthmay:day_of_weektue`   62.86
`monthjun:day_of_weektue`   59.32
`monthaug:day_of_weektue`   56.36
day_of_weekwed              56.22
monthmay                    55.67
`nr.employed:monthmay`      55.62
`monthaug:day_of_weekwed`   55.21
`monthmay:day_of_weekwed`   54.12
`monthnov:day_of_weektue`   52.83
`monthjul:day_of_weektue`   52.64
`monthnov:day_of_weekwed`   52.62
`monthmar:euribor3m`        52.45
poutcomenonexistent         51.68
`monthmay:euribor3m`        51.58
`monthoct:euribor3m`        51.28
euribor3m                   50.97
`duration:monthjul`         50.76
day_of_weektue              49.95
```

It appears that the ‘duration’ has the biggest impact on the probability of subscription. And ‘monthmay’, ‘ day_of_weekwed’ and interactions between some ‘month’ and’day_of_week’ are important factors too.

ROC Curve



The area underneath this ROC curve is .9470. The curve is close to the left-hand border and the top of the curve reaches the y-value of 1 pretty quickly. This indicates that the test is accurate. Since the area is .9470, the test does an excellent job of separating the client who makes the subscription or not and making predictions using the chosen model.

***Note:** I repeated the analysis with the two models that I produced: one with the main significant factors without interaction terms and one with interaction terms to compare which performed better at predicting the results. For the model without interaction, its prediction accuracy is 0.916, and its sensitivity and specificity are 0.9789 and 0.4203 respectively. Its AUC is 0.9413. Both models give similar predictions and ROC curves. Another big concern is time consuming on model selection for the model with interaction. As a result, I would recommend the model without interaction.