Portuguese BANK Data Analysis

Siddharth, Nikhil, Sai Venkat

About the dataset

- **bank client data:**

1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

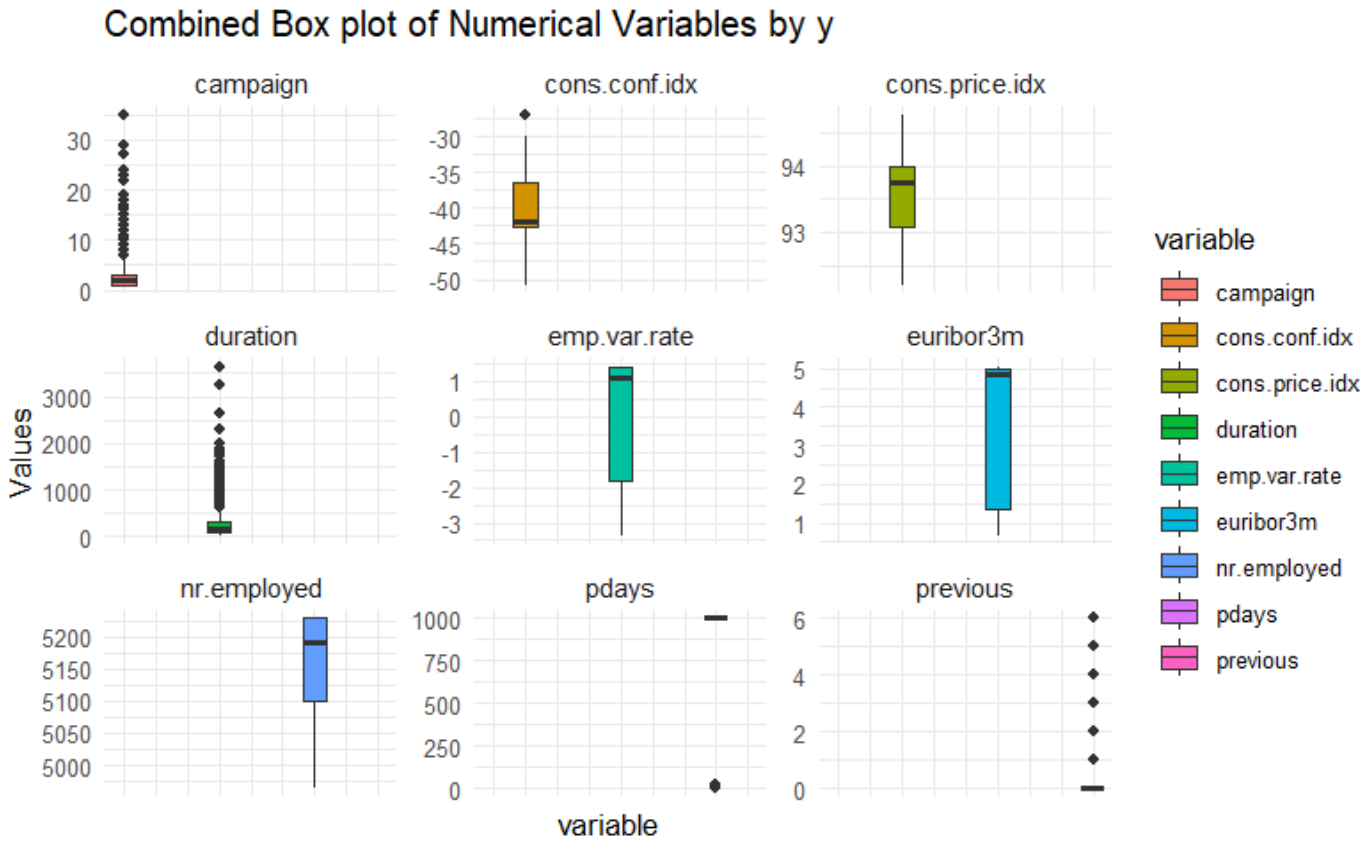
social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

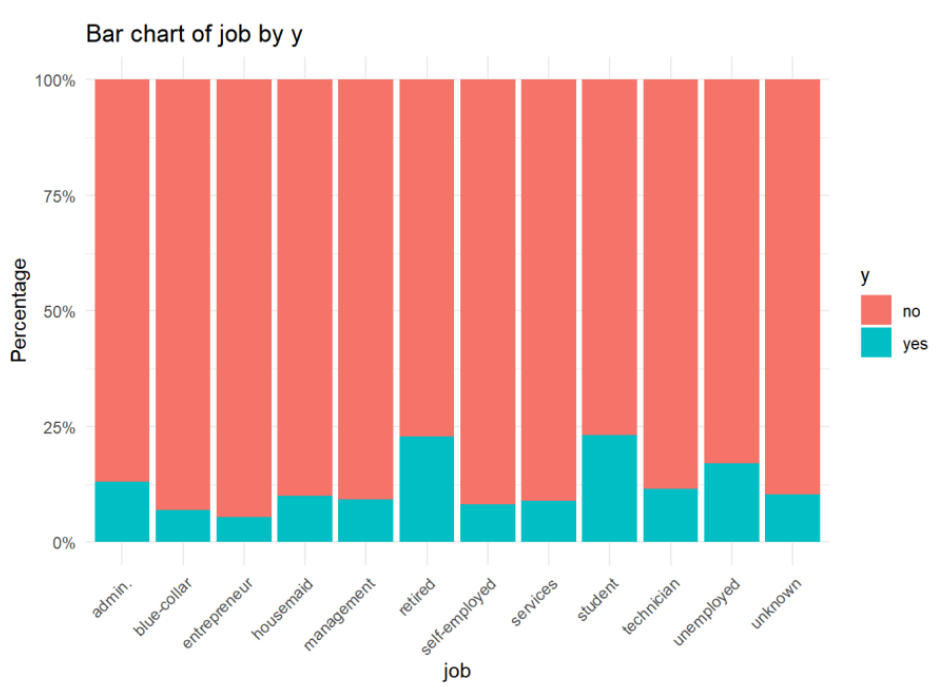
Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

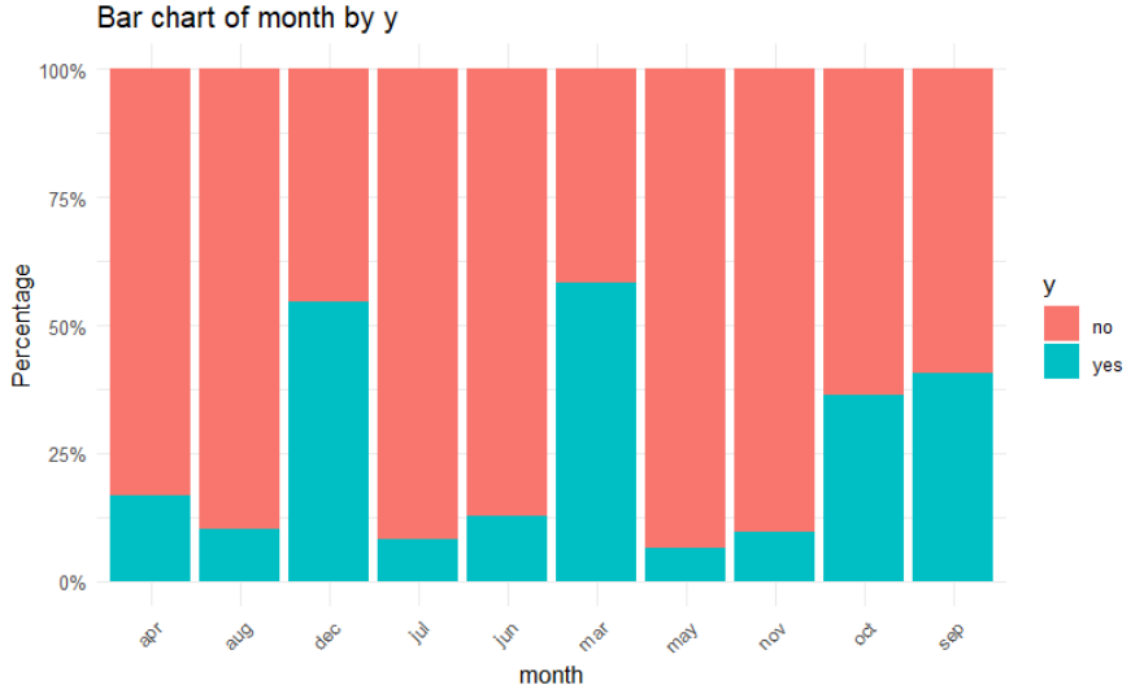
Exploratory Data Analysis

No Missing values in the dataset



Categorical Data distributions





Lasso

- Performed Lasso to remove which have low significance on the target variable y.

- Address multicollinearity

```
Warning: variables are collinearSetting levels: control = no, case = yes
Setting direction: controls < cases
Setting levels: control = no, case = yes
Setting direction: controls < cases
AUC for Logistic Regression: 0.9211915
AUC for LDA: 0.9230863
```

```
55 x 1 sparse Matrix of class "dgCMatrix"
                                        s1
(Intercept)                     50.6585245160
age                              .
jobadmin.                        .
jobblue-collar                   .
jobentrepreneur                 -0.4556030044
jobhousemaid                     .
jobmanagement                   -0.1147423311
jobretired                      -0.0492980488
jobself-employed                -0.0858386403
jobservices                      .
jobstudent                       .
jobtechnician                    0.2158854419
jobunemployed                    .
jobunknown                       .
maritalmarried                  -0.0134033271
maritalsingle                    0.0413690231
maritalunknown                   .
educationbasic.6y                .
educationbasic.9y                .
educationhigh.school             .
educationilliterate              .
educationprofessional.course     .
educationuniversity.degree       .
educationunknown                 .
defaultunknown                   .
defaultyes                       .
housingunknown                  -0.0248545509
housingyes                       .
loanunknown                      .
loanyes                          .
contacttelephone                -0.2487895857
monthaug                         0.0467547721
monthdec                         1.2354963752
monthjul                         .
monthjun                         0.5032635694
monthmar                         1.5631203142
monthmay                        -0.5560849939
monthnov                        -0.1836799590
monthoct                         .
monthsep                        -0.2552679016
day_of_weekmon                   .
day_of_weekthu                   .
day_of_weektue                  -0.0258929094
day_of_weekwed                   .
duration                         0.0047792850
campaign                        -0.0203305097
pdays                           -0.0008660701
previous                         .
poutcomenonexistent              0.2282509012
poutcomesuccess                  0.9514258360
emp.var.rate                    -0.1554985363
cons.price.idx                   .
cons.conf.idx                    0.0228007359
euribor3m                        .
nr.employed                     -0.0102910962
```

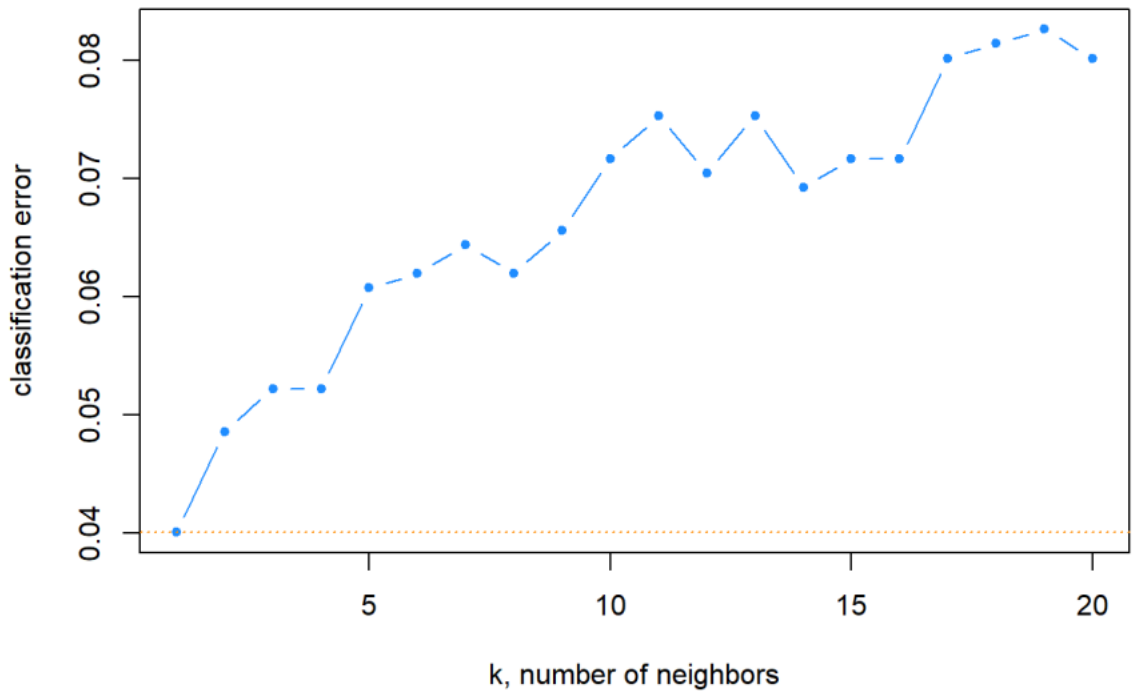
Logistic Regression

```
#testing error
print(paste("Logistic testing error", 1-cm$overall[1]))
```

```
[1] "Logistic testing error 0.108140947752126"
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 717  73
         1  16  17

               Accuracy : 0.8919
                 95% CI : (0.8686, 0.9123)
    No Information Rate : 0.8906
    P-Value [Acc > NIR] : 0.4835

                  Kappa : 0.2313

 Mcnemar's Test P-Value : 2.921e-09

            Sensitivity : 0.9782
            Specificity : 0.1889
         Pos Pred Value : 0.9076
         Neg Pred Value : 0.5152
             Prevalence : 0.8906
         Detection Rate : 0.8712
   Detection Prevalence : 0.9599
      Balanced Accuracy : 0.5835

       'Positive' Class : 0
```

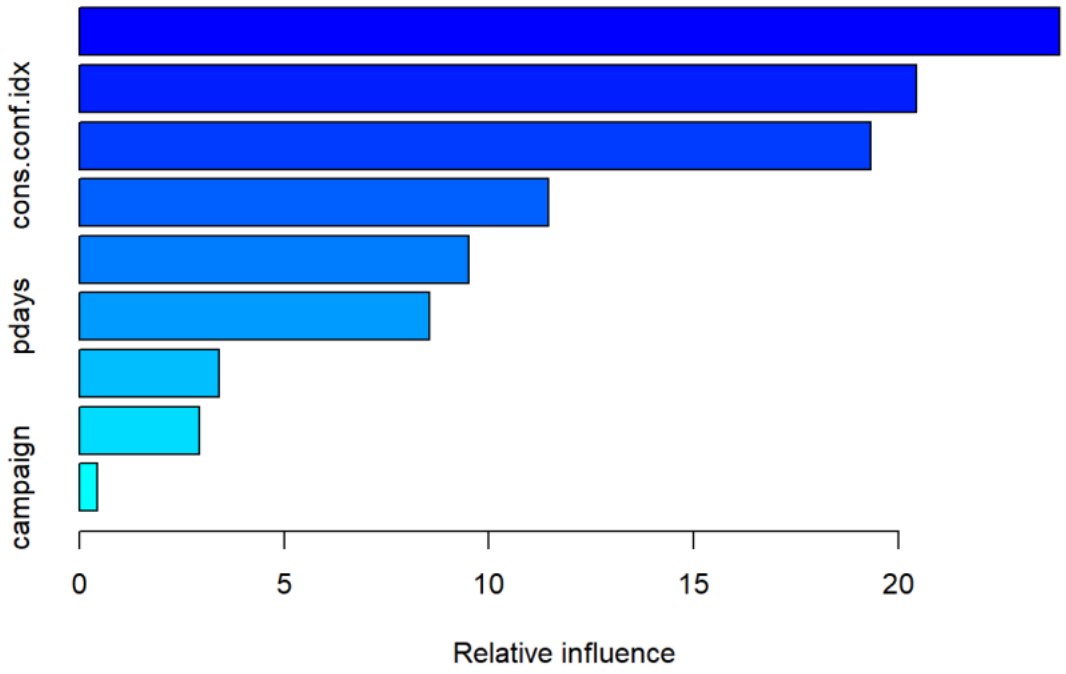
KNN



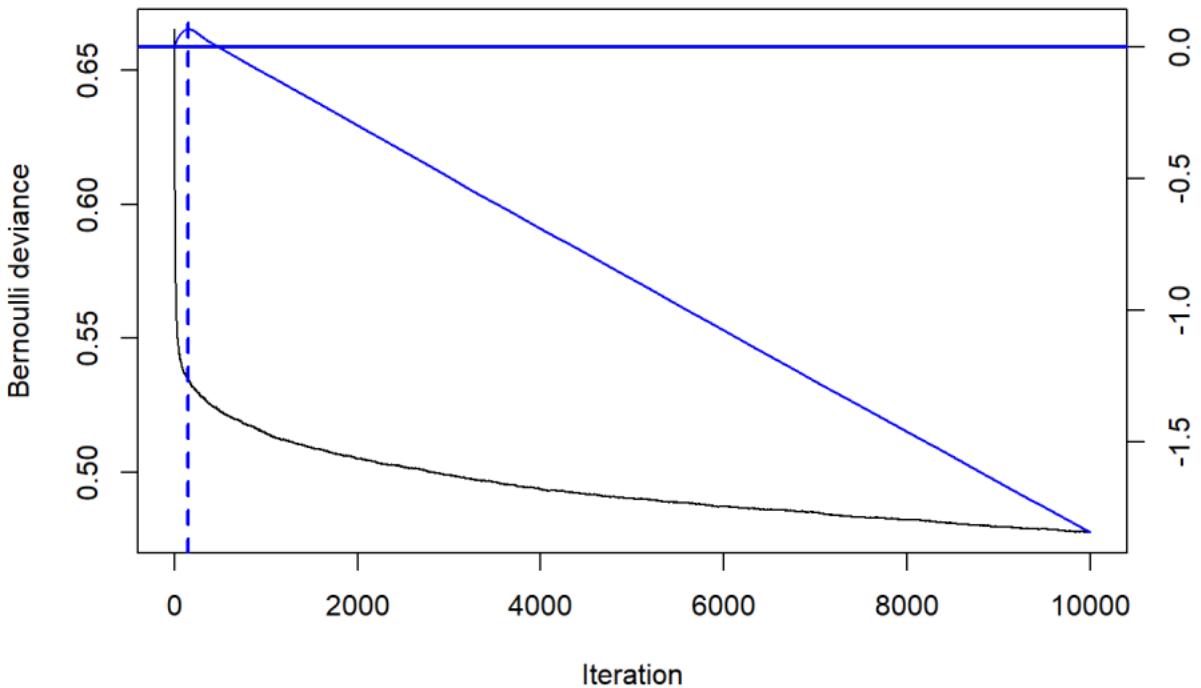

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 710  47
         1  23  43

               Accuracy : 0.9149
                 95% CI : (0.8938, 0.9331)
    No Information Rate : 0.8906
    P-Value [Acc > NIR] : 0.012575

                  Kappa : 0.5055

 Mcnemar's Test P-Value : 0.005977

            Sensitivity : 0.9686
            Specificity : 0.4778
         Pos Pred Value : 0.9379
         Neg Pred Value : 0.6515
             Prevalence : 0.8906
         Detection Rate : 0.8627
   Detection Prevalence : 0.9198
      Balanced Accuracy : 0.7232

       'Positive' Class : 0
```

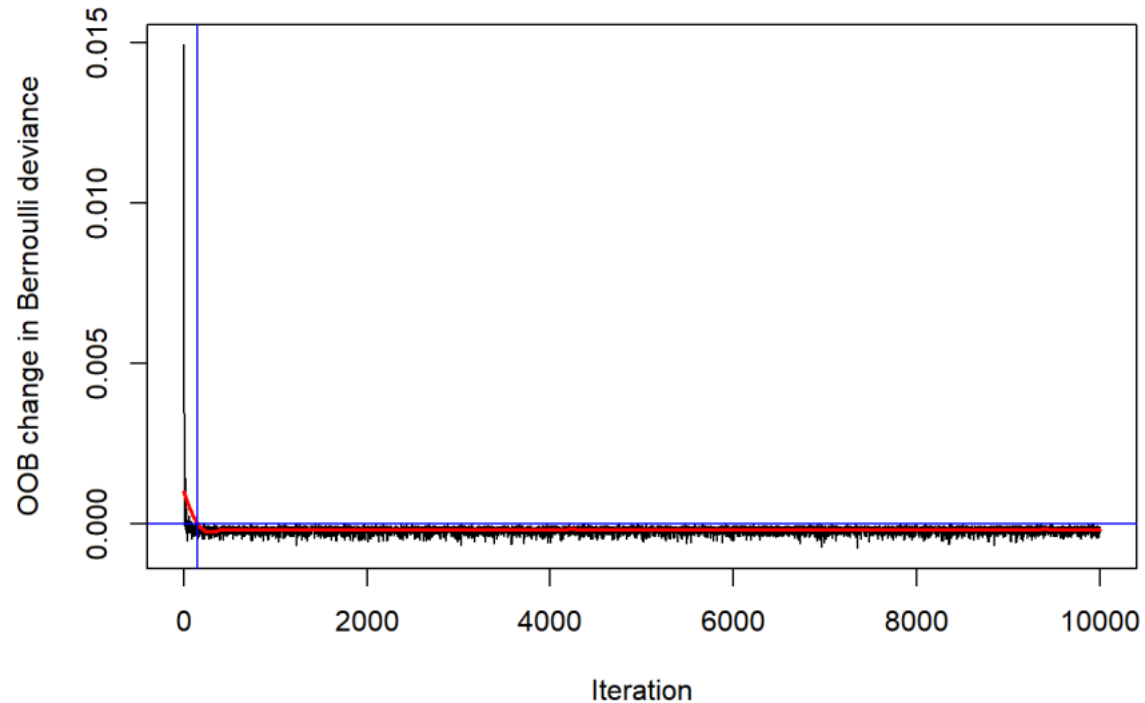
```
[1] "KNN testing error 0.0850546780072904"
```

Tree Based Model




```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 727  18
         1  64  21

               Accuracy : 0.9012
                 95% CI : (0.8789, 0.9207)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2932

 Mcnemar's Test P-Value : 6.715e-07

            Sensitivity : 0.9191
            Specificity : 0.5385
         Pos Pred Value : 0.9758
         Neg Pred Value : 0.2471
             Prevalence : 0.9530
         Detection Rate : 0.8759
   Detection Prevalence : 0.8976
      Balanced Accuracy : 0.7288

[1] "Gradient Descent Boosting testing error 0.0987951807228916"
```


Bernoulli deviance
0.50
0.55
0.60
0.65
-1.5
-1.0
-0.5
0.0
0
2000
4000
6000
8000
10000
Iteration


OOB change in Bernoulli deviance
0.000
0.005
0.010
0.015
0
2000
4000
6000
8000
10000
Iteration

SVM