Bank Marketing

Prediction Model for Bank Direct Marketing Campaign

This is a public dataset which was made usable for research by S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. The data was used for 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 (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

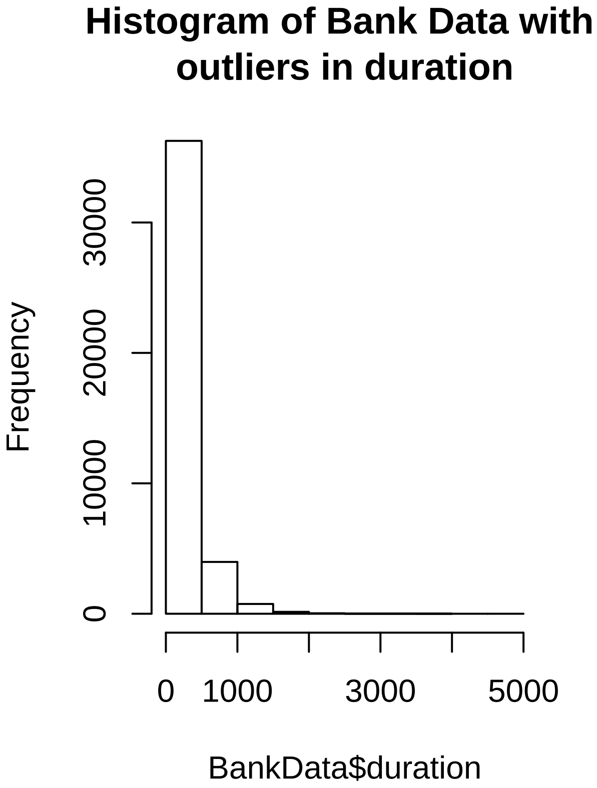
Library

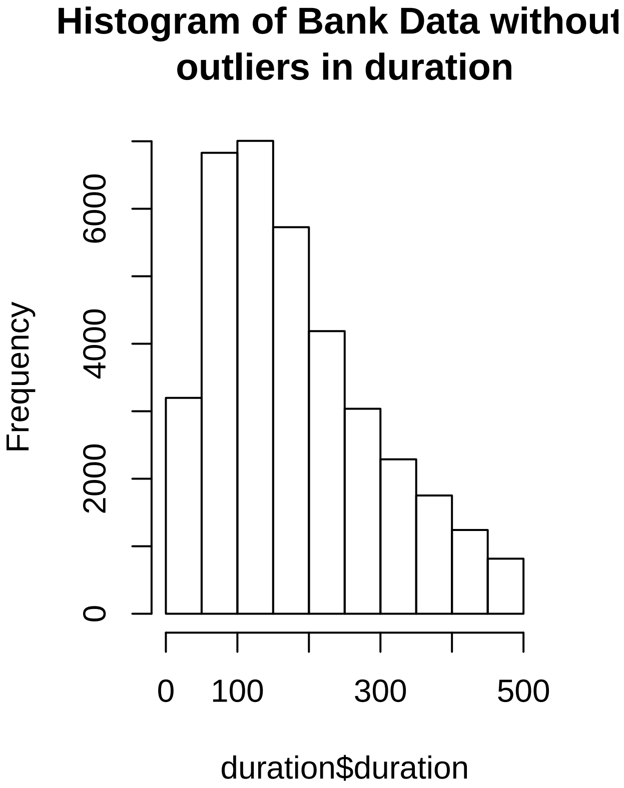
```
## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: lattice

## Rattle: A free graphical interface for data mining with R.
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

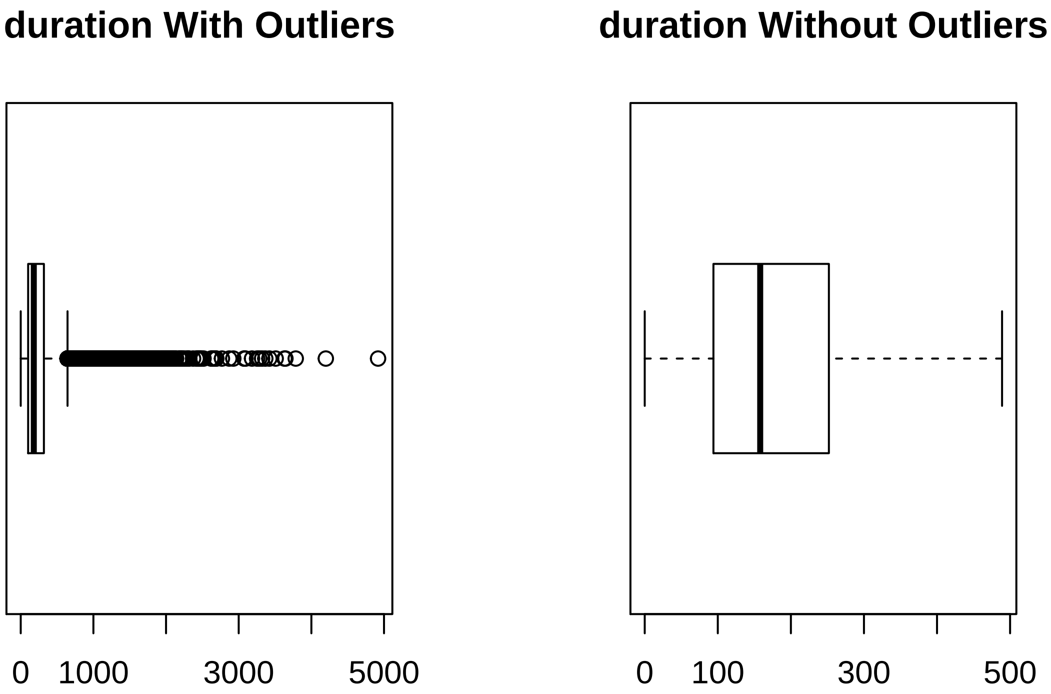
Loading Dataset

Data Cleaning

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

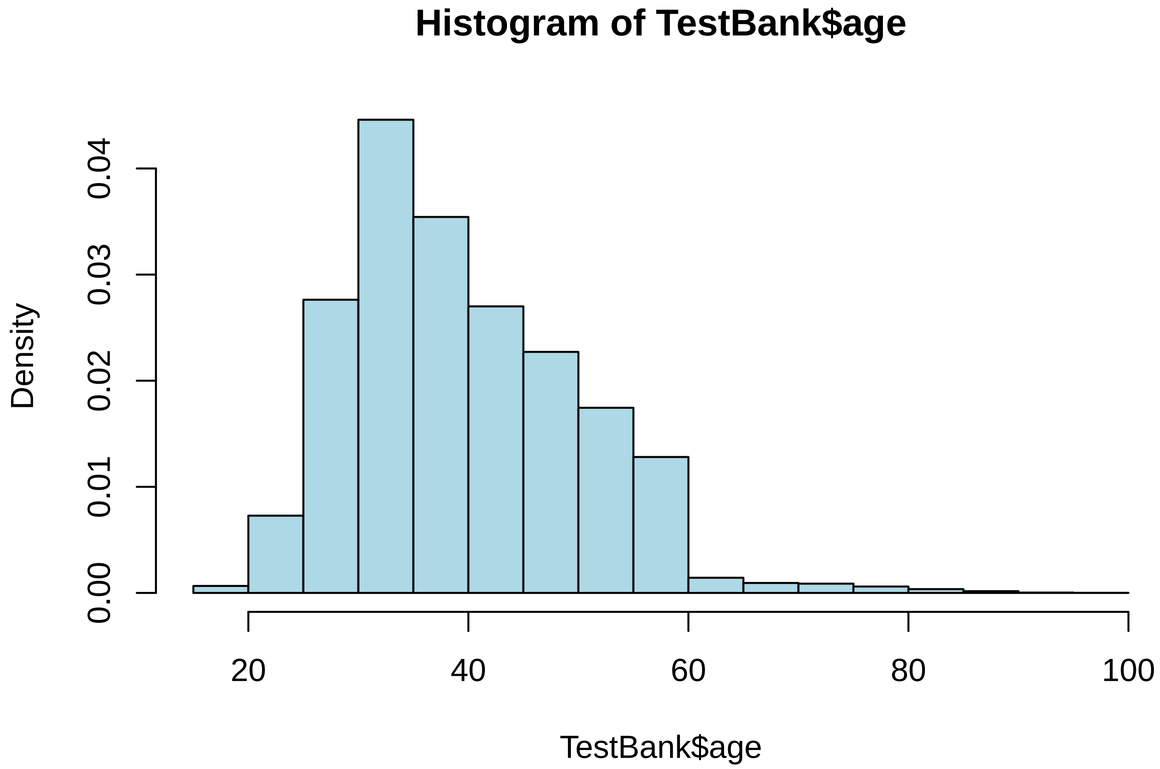

Histogram of Bank Data with
outliers in duration
Frequency
0
10000
20000
30000
0
1000
3000
5000
BankData$duration


Histogram of Bank Data withou
outliers in duration
Frequency
0
2000
4000
6000
0
100
300
500
duration$duration

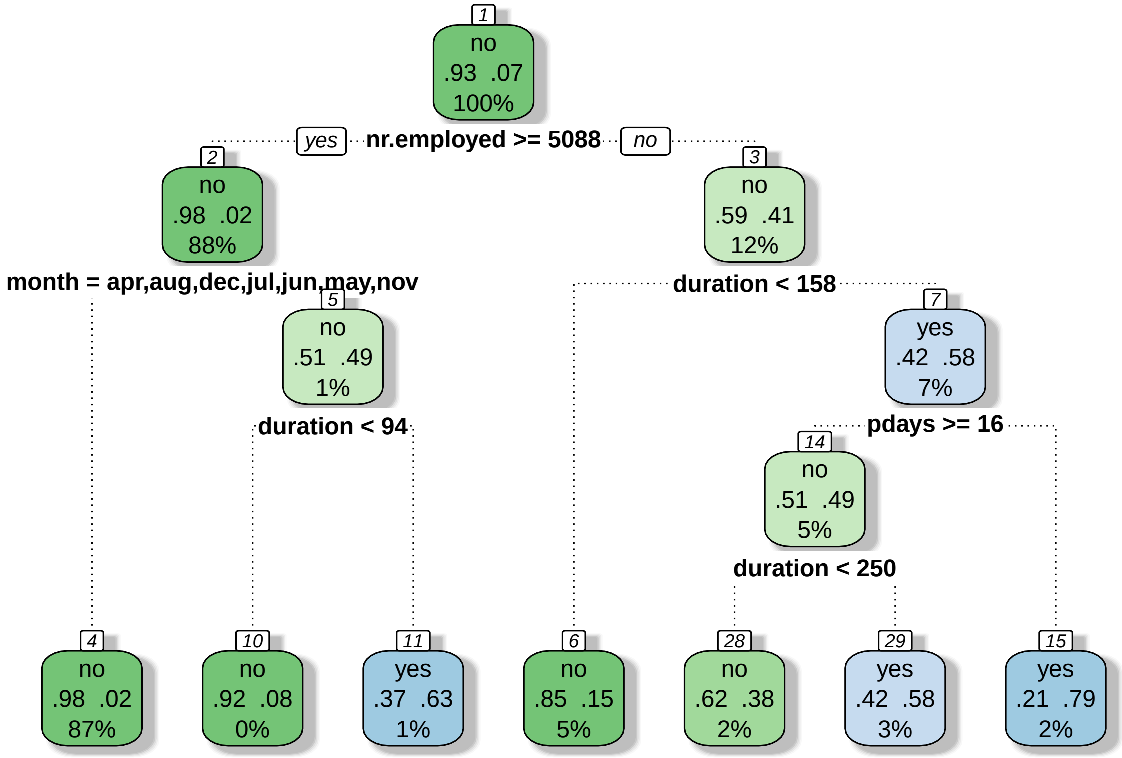


Training and Testing datasets

```
##       age                   job           marital     
##  Min.   :17.00   admin.     :9168   divorced: 4062  
##  1st Qu.:32.00   blue-collar:8078   married :21863  
##  Median :38.00   technician :5947   single  :10094  
##  Mean   :40.02   services   :3468   unknown :   62  
##  3rd Qu.:47.00   management :2568                   
##  Max.   :98.00   retired    :1474                   
##                  (Other)    :5378                   
##                education        default         housing     
##  university.degree  :10708   no     :28514   no     :16247  
##  high.school        : 8316   unknown: 7564   unknown:  875  
##  basic.9y           : 5271   yes    :    3   yes    :18959  
##  professional.course: 4621                                  
##  basic.4y           : 3632                                  
##  basic.6y           : 2012                                  
##  (Other)            : 1521                                  
##       loan            contact          month       day_of_week
##  no     :29723   cellular :22717   may    :12102   fri:6865   
##  unknown:  875   telephone:13364   jul    : 6094   mon:7553   
##  yes    : 5483                     aug    : 5572   thu:7487   
##                                    jun    : 4724   tue:7129   
##                                    nov    : 3612   wed:7047   
##                                    apr    : 2230              
##                                    (Other): 1747
##     duration      campaign         pdays          previous
##  Min.   :  0   Min.   : 1.00   Min.   : 0.00   Min.   :0.0000
##  1st Qu.: 94   1st Qu.: 1.00   1st Qu.:25.00   1st Qu.:0.0000
##  Median :158   Median : 2.00   Median :25.00   Median :0.0000
##  Mean   :182   Mean   : 2.59   Mean   :24.33   Mean   :0.1718
##  3rd Qu.:252   3rd Qu.: 3.00   3rd Qu.:25.00   3rd Qu.:0.0000
##  Max.   :489   Max.   :56.00   Max.   :27.00   Max.   :7.0000
##
##         poutcome      emp.var.rate      cons.price.idx  cons.conf.idx
##  failure    : 3784   Min.   :-3.40000   Min.   :92.20   Min.   :-50.80
##  nonexistent:31146   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70
##  success    : 1151   Median : 1.10000   Median :93.44   Median :-41.80
##                      Mean   : 0.08674   Mean   :93.57   Mean   :-40.46
##                      3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40
##                      Max.   : 1.40000   Max.   :94.77   Max.   :-26.90
##
##    euribor3m      nr.employed     y
##  Min.   :0.634   Min.   :4964   no :33576
##  1st Qu.:1.344   1st Qu.:5099   yes: 2505
##  Median :4.857   Median :5191
##  Mean   :3.630   Mean   :5167
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
##
```



Building Models



Rattle 2017−Jan−27 19:43:12 Vijay

```
## [1] 0.9520576
```

```
##   no  yes
## 6865  352
```

```
## Number of cases in table: 1
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.30746, df = 1, p-value = 0.5792
##  Chi-squared approximation may be incorrect
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  6655  242
##        yes  102  218
##
##                Accuracy : 0.9523
##                  95% CI : (0.9472, 0.9571)
##     No Information Rate : 0.9363
##     P-Value [Acc > NIR] : 3.273e-09
##
##                   Kappa : 0.5346
##  Mcnemar's Test P-Value : 6.661e-14
```

```
##
##             Sensitivity : 0.9849
##             Specificity : 0.4739
##          Pos Pred Value : 0.9649
##          Neg Pred Value : 0.6813
##              Prevalence : 0.9363
##          Detection Rate : 0.9221
##    Detection Prevalence : 0.9557
##       Balanced Accuracy : 0.7294
##
##        'Positive' Class : no
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  6695  282
##        yes   62  178
##
##                Accuracy : 0.9523
##                  95% CI : (0.9472, 0.9571)
##     No Information Rate : 0.9363
##     P-Value [Acc > NIR] : 3.273e-09
##
##                   Kappa : 0.4861
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9908
##             Specificity : 0.3870
##          Pos Pred Value : 0.9596
##          Neg Pred Value : 0.7417
##              Prevalence : 0.9363
##          Detection Rate : 0.9277
##    Detection Prevalence : 0.9667
##       Balanced Accuracy : 0.6889
##
##        'Positive' Class : no
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  6650  226
##        yes  107  234
##
##                Accuracy : 0.9539
##                  95% CI : (0.9488, 0.9586)
##     No Information Rate : 0.9363
##     P-Value [Acc > NIR] : 8.596e-11
##
##                   Kappa : 0.5604
##  Mcnemar's Test P-Value : 1.004e-10
##
```

```
##             Sensitivity : 0.9842
##             Specificity : 0.5087
##          Pos Pred Value : 0.9671
##          Neg Pred Value : 0.6862
##              Prevalence : 0.9363
##          Detection Rate : 0.9214
##    Detection Prevalence : 0.9528
##       Balanced Accuracy : 0.7464
##
##        'Positive' Class : no
##
```

Conclusion

1. Based on the prediction produced using rpart, SVM, Random Forest and KNN models the accuracy comes to 94%. It was a small validation datset (20%), but this result is within our expected margin of 97% +/-4%. This definitely suggest that we may have an accurate and a reliable accurate model.

The decision tree provides following facts about the data:- 1. The top node shows 93% of the customers are employed and 7% are unemployed. This shows that the top node represents 100% of the customer base.

2. On looking at the node for customers who are unemployed and the duration of the call was less than 162 seconds, the prediction model concludes that 84% of them did not subscribe to the term account. Only 16% of them subscribed to term deposit account.