Model Selection for bank telemarketing

Christoph Linne, Emil Brodersen
Xun Gong, Yufang Yan

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://case.hu-berlin.de

Outline

1. Dataset Introduction ✓
2. Data Pre-processing
3. Model Building and Prediction
4. Model Evaluation
5. Conclusion

Dataset Context (I)

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.

Dataset Context (II)

- Objective:
  - The classification goal is to predict if the client will subscribe a term deposit (variable y).
- Solution: predictive modeling
  - Predictive modeling helps in determining the main characteristics that affect success and selection of potential buying customers.
  - GLM, RPart Decision Tree, Random Forest Tree, Neural Network algorithms were used to build the model and the appropriate model is selected based ROC and AUC

Data Attributes

Demographic Information:	Marketing Information	Macro economy Information
age (numeric)	contact (categorical)	emp.var.rate (numeric)
job (categorical)	month(categorical)	cons.price.idx(numeric)
marital(categorical)	day_of_week (categorical)	cons.conf.idx (numeric)
education (categorical)	duration (numeric)	euribor3m (numeric)
Default (categorical)	campaign (numeric)	nr.employed: (numeric)
housing (categorical)	pdays (numeric)	
loan (categorical)	previous (numeric)	
	poutcome (categorical)	

Figure 1: Predictive Variables

Target Variable: y - has the client subscribed a term deposit? (binary: 'yes','no')

Outline

1. Dataset Introduction
2. Data Pre-processing ✓
3. Model Building and Prediction
4. Model Evaluation
5. Conclusion

Data cleaning (I)

Variable Selection:
Delete: Default and Duration

Demographic Information	age	job	marital	education	default	housing	loan
	0	330	80	1731	8597	990	990
	0	0.80%	0.19%	4.20%	20.87%	2.40%	2.40%

Figure 2: Missing Values

Data cleaning (II)

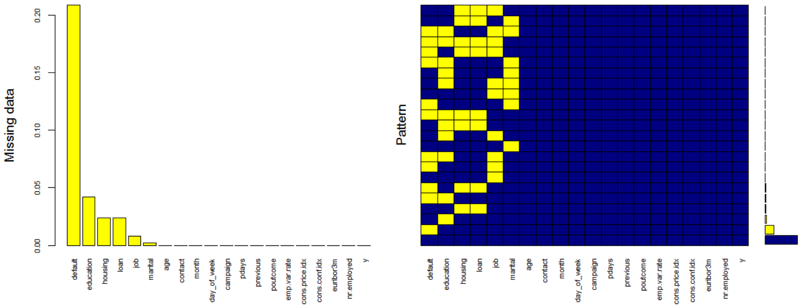


Figure 3: Missing Values Table

MICE Package

- Assumption:
  - the missing data are Missing at Random (MAR)
  - It imputes data on a variable by variable basis by specifying an imputation model per variable.
  - Linear regression is used to predict continuous missing values. Logistic regression is used for categorical missing values.
  - Suppose we have X1, X2?.Xk variables. If X1 has missing values, then it will be regressed on other variables X2 to Xk. The missing values in X1 will be then replaced by predictive values obtained.
- Methods:
  - polyreg (Bayesian polytomous regression) ? For Factor Variables (>= 2 levels)

Code

```r
library(mice)
library(VIM)
md.pattern(bank)
mice_plot <- aggr(bank, col=c('navyblue','yellow'),numbers=TRUE, sortVars=TRUE,
labels=names(bank), cex.axis=.7, gap=3,
ylab=c("Missing data","Pattern"))
tempData <- mice(bank,m=5,maxit=50,
meth='polyreg',seed=500)
summary(tempData)
bankfull_NonNA <- complete(tempData,1)
write.csv(bankfull_NonNA, "bankclean.csv")
```

Outline

1. Dataset Introduction
2. Data Pre-processing
3. Model Building and Prediction ✓
4. Model Evaluation
5. Conclusion

Logistic Regression Model

Logistic Regression

- Linear regression with a transformation such that the output is always between 0 and 1, and can thus be interpreted as a probability. It predicts the probability of occurrence of an event by fitting data to a logit function.
- To represent binary / categorical outcome, we use dummy variables.

Code (I)

```r
library(caret)
bank.train.dummy <- predict(dummyVars(y ~ . ,data=bank_train), newdata=bank_train)
bank.train.dummy <- data.frame(bank.train.dummy, y=factor(bank_train$y))
bank.test.dummy <- predict(dummyVars(y ~ . ,data=bank_test), newdata=bank_test)
bank.test.dummy <- data.frame(bank.test.dummy, y=factor(bank_test$y))
```

Code (II)

```
trControl=cvCtrl, metric="ROC", tuneLength=1)
predict.logit.train <- predict.train(logit)
predict.logit.test <- predict(logit, bank.test.dummy, type="prob")
(auc.logit <- auc(bank.test.dummy$y, predict.logit.test2$yes))
```

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0879  -0.3893  -0.3187  -0.2657   2.9410

Coefficients: (9 not defined because of singularities)
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.027e+12  7.251e+12   0.142 0.887353
age                           3.824e-04  2.358e-03   0.162 0.871179
job.admin.                    9.897e-03  1.221e-01   0.081 0.935405
job.blue.collar              -1.783e-01  1.284e-01  -1.388 0.165004
job.entrepreneur             -4.033e-03  1.598e-01  -0.025 0.979865
job.housemaid                -1.367e-01  1.796e-01  -0.761 0.446530
job.management               -1.219e-01  1.397e-01  -0.873 0.382836
job.retired                   2.266e-01  1.473e-01   1.538 0.124142
job.self.employed            -1.105e-01  1.580e-01  -0.699 0.484256
job.services                 -1.104e-01  1.365e-01  -0.809 0.418521
job.student                   2.742e-01  1.533e-01   1.789 0.073606 .
job.technician               -3.165e-02  1.280e-01  -0.247 0.804745
job.unemployed                       NA         NA      NA       NA
marital.divorced             -1.206e-01  7.706e-02  -1.564 0.117711
marital.married              -4.574e-02  5.043e-02  -0.907 0.364343
marital.single                       NA         NA      NA       NA
education.basic.4y           -1.741e-01  8.832e-02  -1.971 0.048746 *
education.basic.6y            5.133e-02  1.036e-01   0.495 0.620268
education.basic.9y           -1.354e-01  7.800e-02  -1.736 0.082606 .
education.high.school        -1.159e-01  5.807e-02  -1.995 0.046009 *
education.illiterate          3.343e-01  5.538e-01   0.604 0.546086
education.professional.course -8.343e-02  7.327e-02  -1.139 0.254851
education.university.degree          NA         NA      NA       NA
```

```
month.aug                  2.847e-01  1.426e-01   1.996 0.045955 *
month.dec                  3.289e-02  2.237e-01   0.147 0.883102
month.jul                 -1.424e-01  1.745e-01  -0.816 0.414664
month.jun                 -8.640e-01  2.349e-01  -3.678 0.000235 ***
month.mar                  1.349e+00  1.579e-01   8.540  < 2e-16 ***
month.may                 -5.649e-01  1.516e-01  -3.726 0.000194 ***
month.nov                 -6.370e-01  1.536e-01  -4.146 3.39e-05 ***
month.oct                 -1.151e-01  1.463e-01  -0.787 0.431438
month.sep                         NA         NA      NA       NA
day_of_week.fri           -1.718e-01  6.362e-02  -2.700 0.006928 **
day_of_week.mon           -3.925e-01  6.368e-02  -6.163 7.14e-10 ***
day_of_week.thu           -1.093e-01  6.120e-02  -1.786 0.074139 .
day_of_week.tue           -1.201e-01  6.247e-02  -1.923 0.054539 .
day_of_week.wed                   NA         NA      NA       NA
campaign                  -3.739e-02  1.018e-02  -3.674 0.000239 ***
pdays                      1.121e+00  2.299e-01   4.873 1.10e-06 ***
previous                  -6.913e-02  6.292e-02  -1.099 0.271876
poutcome.failure          -8.085e-01  2.268e-01  -3.565 0.000365 ***
poutcome.nonexistent      -3.264e-01  2.286e-01  -1.428 0.153363
poutcome.success                  NA         NA      NA       NA
emp.var.rate              -1.549e+00  1.390e-01 -11.140  < 2e-16 ***
cons.price.idx             2.117e+00  2.468e-01   8.580  < 2e-16 ***
cons.conf.idx              2.895e-02  7.842e-03   3.691 0.000223 ***
euribor3m                  2.479e-01  1.278e-01   1.940 0.052327 .
nr.employed                6.639e-03  3.034e-03   2.188 0.028636 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23071  on 32950  degrees of freedom
Residual deviance: 18118  on 32903  degrees of freedom
AIC: 18214
```

Decision Tree - General Introduction

Decision tree is a set of (splitting) rules to recursively partition a data set. The decision tree model is one of the most commonly used predictive models in statistics, data mining and machine learning.

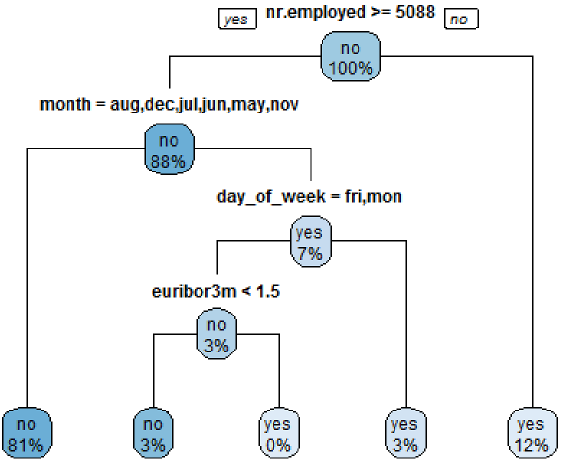
- Classification Tree ✓
  - The predicted outcome is the class to which the data belongs.
  - The Spilting Rule is to minimize mixture of classes (impurity) within nodes
- Regression Tree
  - The predicted outcome can be considered a real number.
  - The Spilting Rule is to minimize the variance of the response variable within nodes

Classification Tree - Splitting Criteria

Different decision tree algorithms use different splitting criteria for measuring the node impurity. Here listed two main splitting criterias. Let I(N) denote the impurity of some node N.

- Gini impurity
  - Gini Index I(N) = 1- $\sum_j p(c_j|N)^2$
  - Favors larger partitions.
  - Perfectly classified, Gini Index would be zero when perfectly classified. So a low Gini Index is preferred.
- Information gain and Entropy
  - Entropy I(N) = - $\sum_j p(c_j|N) * log_2(p(c_j|N))$
  - Favors splits with small counts but many unique values.
  - Information Gain = Entropy(parent) - Weighted Sum of Entropy(Children)

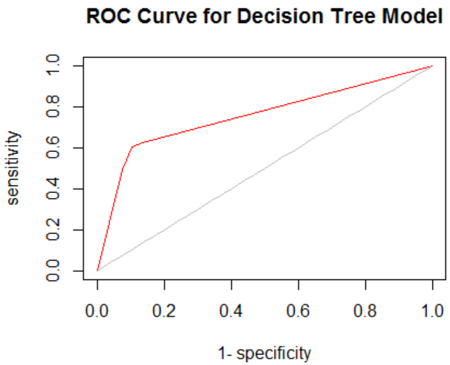
Classification Tree built with training data



Code

```
library(rpart)
rpart.control= rpart.control(minsplit=20, minbucket = round(20/3), cp=0.005)
dtm<-rpart(y~.,data_train,
              parms=list(loss=matrix(c(0,1,5,0),byrow=TRUE,nrow=2)),
              control=rpart.control)
printcp(dtm)
bestcp <- dtm$cptable[which.min(dtm$cptable[,"xerror"]),"CP"]
rpart.control= rpart.control(minsplit=20, minbucket = round(20/3), cp=bestcp)
dtm<-rpart(y~.,data_train,method= control=rpart.control)
summary(dtm)
rpart.plot(dtm,type=1,extra=100)
```

Prediction Result - ROC Curve



Area under the ROC Curve(AUC)=0.756398

Random Forest Output

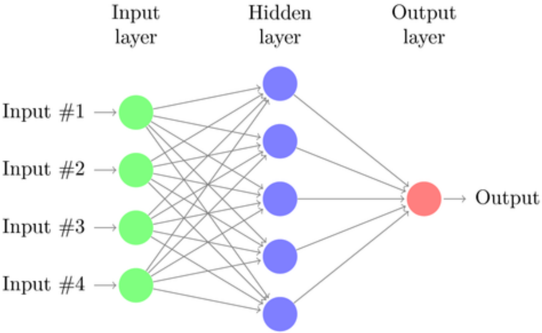
```
randomForest(x = bank.train.1, y = bank.label,
  ntree = 1000, importance = TRUE)
Type of random forest: classification
Number of trees: 1000
No. of variables tried at each split: 4

OOB estimate of  error rate: 10.01%
Confusion matrix:
no   yes  class.error
no   28581  689  0.02353946
yes   2610 1071  0.70904645
```

Neural Networks

- Neural networks is a computational approach that is modeled on the way a biological brain solves problems.
- Recieves input signals (variable values)
- Aggregates input signals (a weighted sum)
- Non-linear transformation (logistic, hyperbolic)
- Sends output signal (result)

Neural Networks



Code for Neural Network

```
model.control<- trainControl(
method = "cv", # 'cv' for cross validation
number = 5, # number of folds in cross validation
classProbs = TRUE,
summaryFunction = twoClassSummary,
returnData = FALSE # The training data will not be included in the ouput training object)

nn.parms <- expand.grid(decay = c(0, 10^seq(-3, 0, 1)),
size = seq(3,15,2))
```

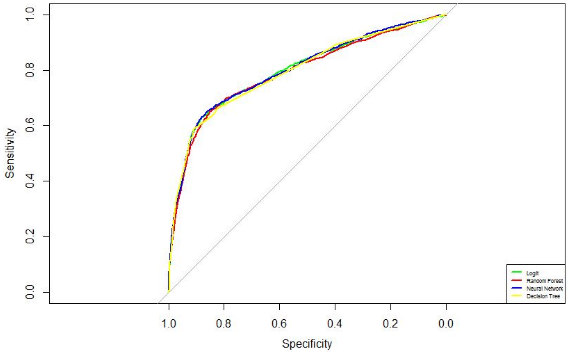
Code for Neural Network

```
    nn <- train(y~., data = train,
    method = "nnet",
    maxit = 200,
    trace = FALSE, # options for nnet function
    tuneGrid = nn.parms, # parameters to be tested
    metric = "ROC", trControl = model.control)
```

Outline

1. Dataset Introduction
2. Data Pre-processing
3. Model Building and Prediction
4. Model Evaluation ✓
5. Conclusion

Model Selection



Outline

1. Dataset Introduction
2. Data Pre-processing
3. Model Building and Prediction
4. Model Evaluation
5. Conclusion ✓

Conclusion (I)

- A client with an education of 4 years basic and high school are less likely.
- A client contacted by bank via cellular are significantly more likely.
- Campaign is more likely to be successful during March and August, and less likely during May, Jun, and November.
- Campaign conducted on Friday and Monday are less likely.

Conclusion (II)

- A client who used to be contacted are more likely to deposit their money in the bank.
- A client who used to give negative reply are more likely to reject again.
- When macro economy statistics, such as cons.price.idx (consumer price index), cons.conf.idx (consumer confidence index) and nr.employed (number of employees increases), increases, the more likely clients sign for term deposit.
- When macro economy statistics emp.var.rate (employment variation rate) increase, the less likely.