Model Selection for Bank Telemarketing

Emil Brodersen, Xun Gong, Christoph Linne and Yufang Yan

Ladislaus von Bortkiewicz Chair of Statistics

Humboldt–Universität zu Berlin

http://lvb.wiwi.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 dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns base on phone calls, in order to access if the product (bank term deposit) is ("yes") or is not ("no") subscribed.
- The data is taken from *UCI Machine Learning Repository*.

Dataset Context (II)

- Objective:
  - The classification goal predicts 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, Decision Tree, Random Forest, and Neural Network algorithms were used to build models and the appropriate model is selected based on ROC and AUC.

Data Attributes

- Respondents:
  - 41, 188 observations, 20 variables.
- Target variable:
  - Has the client subscribed a term deposit? (“yes”/“no”).

Demographic Info	Marketing Info	Macroeconomy Info
age (numeric)	contact(categorical)	emp.yar.rate(numeric)
job(categrical)	month(categorical)	cons.price.idx(numeric)
marital(categorical)	day.of.week(categorical)	cons.conf.idx(numeric)
education(categorical)	duration(numeric)	euribor3m(numeric)
default(categrical)	campain(numeric)	nr.employed(numeric)
housing(categorical)	pdays(numeric)	
loan(categorical)	previous(numeric)	
	poutcome(categorical)	

Table 1: Predictor Variables

Outline

1. Dataset introduction
2. Data pre-processing ✓
3. Model building and prediction
4. Model evaluation
5. Conclusion

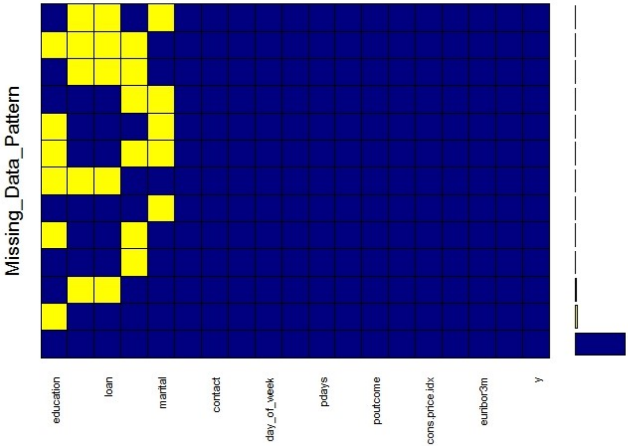
Data Cleaning (I)

- Variable selection: delete “default” and “duration”

Job	Marital	Education	Housing	Loan
330	80	1731	990	990
0.80%	0.19%	4.20%	2.40%	2.40%

Table 2: Demographic information.

Data Cleaning (II)



Multivariate Imputation via Chained Equations (MICE)

- Assumption:
  - The missing data are Missing at Random (MAR).
  - Linear regression is used to predict continuous missing values. Logistic regression is used for categorical missing values.
  - It imputes data on a variable by variable basis by specifying an imputation model per variable.
  - Suppose we have $x_1, x_2, ..., x_k$ variables. If x_1 has missing values, then it will be regressed on other variables x_2 to x_k. The missing values in x_1 will be then replaced by predictive values obtained.
- Methods:
  - Polyreg (Bayesian polytomous regression) - for factor variables (≥ 2 levels)

Code

```
library(VIM)
mice_plot<-aggr(bank,col=c('navyblue','yellow'),
  numbers=TRUE,sortVars=TRUE,labels=names(bank),
  cex.axis=.7,gap=3,ylab=c("Missing_Data_Ratio",
  "Missing_Data_Pattern"))
```

```
n <- nrow(bank)
sample.size <- ceiling(n*0.8)
idx.train <- sample(n, sample.size)
bank_train <- bank[idx.train, ]
bank_test <-  bank[-idx.train, ]
```

Code

```
library(mice)
# Data Imputing for Train Dataset
tempData1 <- mice(bank_train,m=5,maxit=10,meth="
  polyreg",seed=500,diagnostics=True)
bankclean_train <- complete(tempData1,1)
```

```
# Data Imputing for Test Dataset
tempData2 <- mice(bank_test,m=5,maxit=10,meth="
  polyreg",seed=500,diagnostics=True)
pred <- tempData2$predictorMatrix
pred[,"y"] <- 0
tempData3 <- mice(bank_test, pred=pred, pri=F)
bankclean_test <- complete(tempData3,1)
```

Outline

1. Dataset introduction
2. Data pre-processing
3. Model building and prediction ✓
4. Model evaluation
5. Conclusion

Logistic Regression Model

- 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

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

```
cv<-trainControl(method="cv",classProbs=TRUE,summaryFunction=twoClassSummary)
logit <- train(y ~ ., data=bank.train.dummy, method="glm", family = binomial("logit"), preProc=c("center","scale"),metric="ROC", tuneLength=1,trControl=cv,summaryFunction=twoClassSummary, verboseIter=TRUE)
```

$$\begin{aligned}y =&0.21 - 0.06 * age - 0.10 * job.bluecollar + 0.05 * job.retired+\\&0.07 * job.student - 0.16 * marital.married - 0.04*\\&education.basic4y - 0.03 * education.basic9y - 0.04*\\&education.highschool - 0.05 * education.professionalcours+\\&0.08 * housing.no - 0.52 * loan.no - 0.10 * contact.cellular+\\&0.10 * month.aug + 0.18 * month.jul + 0.21 * month.mar-\\&0.31 * month.may - 0.09 * month.nov + 0.06 * month.oct-\\&0.05 * day\_of\_weekfri - 0.12 * day\_of\_weekmon-\\&0.04 * day\_of\_weekthu - 0.04 * day_o f_w eektue-\\&0.10 * campaign + 0.34 * pdays - 0.16 * poutcome.nonexistent-\\&1.36 * emp.var.rate + 0.48 * cons.price.idx+\\&0.07 * cons.conf.idx + 0.70 * euribor3m + -0.65 * nr.employed\end{aligned} \tag{1}$$

Decision Tree

Decision tree is a set of rules (splitting) 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 minimizes 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, two main splitting criterias are listed. Let $I(N)$ denote the impurity of some node N.

- Gini impurity
  - Gini index $I_G(N) = 1 - \sum_j p(c_j|N)^2$:
  - 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 $I_{IG}(N) = -\sum_j p(c_j|N) * \log_2(p(c_j|N))$
  - Favors splits with small counts but many unique values.
  - Information gain = entropy(parent) - weighted sum of entropy(children)

Classification Tree Built With Training Data

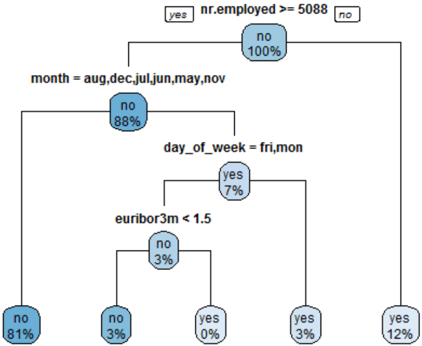
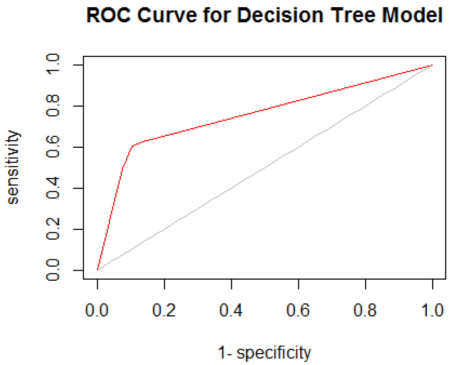


Figure 1: Decision tree.

Code

```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))
logit <- glm (y~.,data = bank.train.dummy, family =
  binomial(link="logit"))
summary(logit)
predict.logit.test <- predict(logit, newdata = bank.
  test.dummy, type="response")
```

Prediction Result - ROC Curve



Area under the ROC Curve(AUC) $=$ 0.756398

Random Forest Output

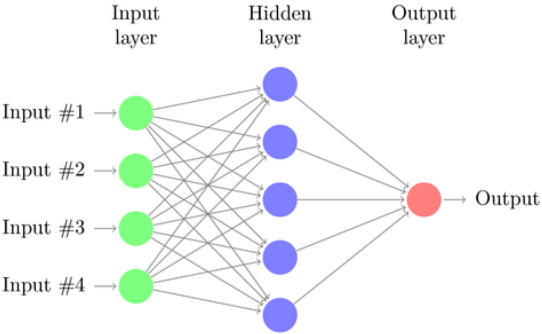
```
randomForest(x = bank.train.1, y = bank.label,
  ntree = 1000, importance = TRUE)
Type of random forest: classification
Number of trees: 1000
No. of variables tried at each split: 4

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.
- Receives input signals (variable values).
- Aggregates input signals (weighted sum).
- Non-linear transformation (logistic, hyperbolic).
- Sends output signal (result).

Neural Networks



Codes

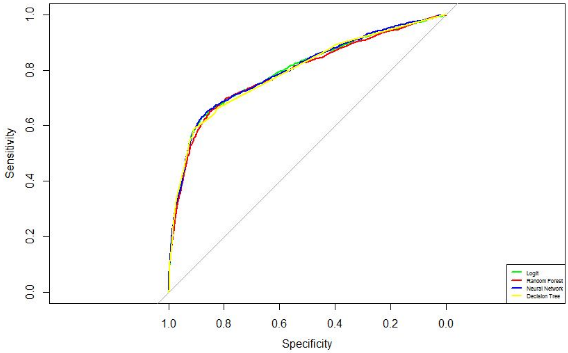
```
model.control<- trainControl(method = "cv", number = 5, classProbs = TRUE, summaryFunction = twoClassSummary,returnData = FALSE )
nn.parms <- expand.grid(decay = c(0, 10^seq(-3, 0, 1)), size = seq(3,15,2))
```

```
nn <- train(y~., data = dataclean_train, method = "nnet",maxit = 200,trace = FALSE,tuneGrid = nn.parms, 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.
- Better succession rate of the campaign during March and August and worse during May, June, 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.