STOR565 Final Project

Tao Bian

April 14, 2018

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Here is the Logistic Lasso regression: Step 1, read four dataset which is training dataset of basic and modified, test dataset of basic and modified:

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.4.3
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.4.3
```

```
## Loaded glmnet 2.0-13
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.4.4
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

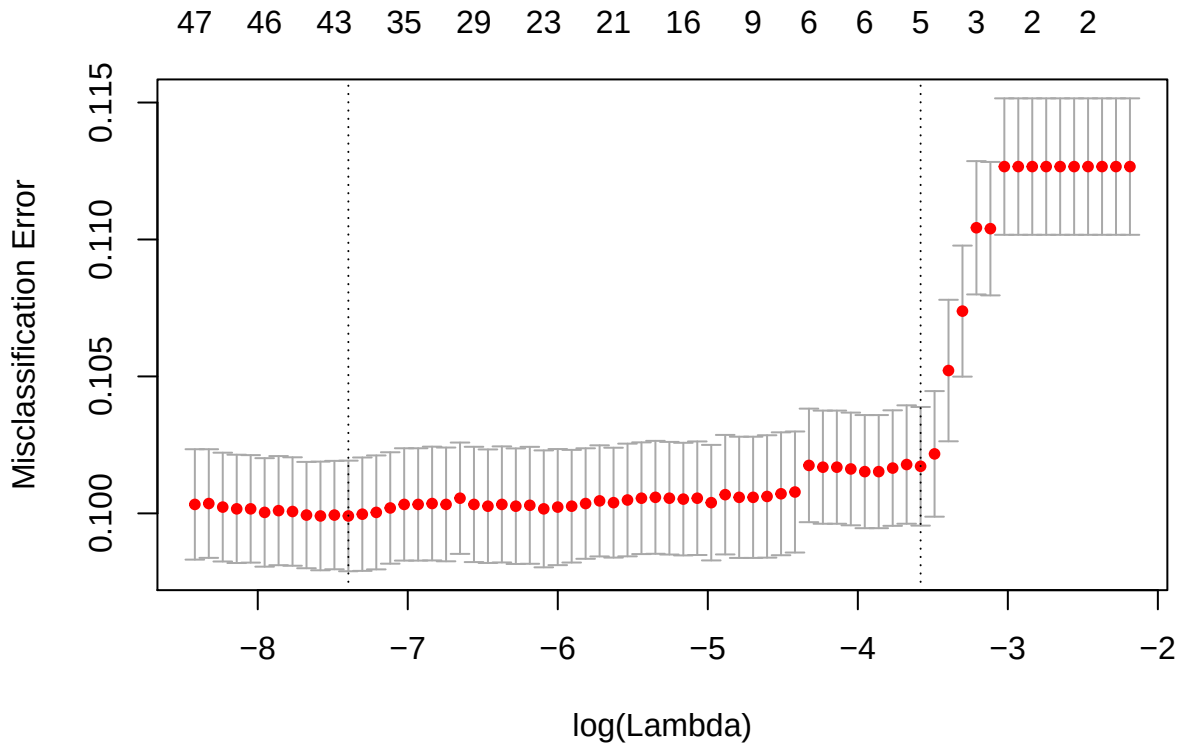
```
## The following object is masked from 'package:glmnet':
##
##     auc
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
bank.train.mod <- read.csv("C:\\Users\\tbian\\Documents\\GitHub\\565project\\data\\modified_train.csv")
bank.test.mod <- read.csv("C:\\Users\\tbian\\Documents\\GitHub\\565project\\data\\modified_test.csv")
bank.train <- read.csv("C:\\Users\\tbian\\Documents\\GitHub\\565project\\data\\basic_train.csv")
bank.test <- read.csv("C:\\Users\\tbian\\Documents\\GitHub\\565project\\data\\basic_test.csv")
```

Step2. do the lASSO logistic model based on the basic training dataset:

```
set.seed(1005)
x.matrix.b<-model.matrix(~.,bank.train[,-20])[,-1]
x.test.b<-model.matrix(~.,bank.test[,-20])[,-1]
y.test.b=bank.test$y
foldid=sample(1:4,size=length(bank.train$y),replace=TRUE)
bank.lasso.b<-cv.glmnet(x.matrix.b,bank.train$y, family="binomial", type.measure="class", alpha=1)
plot(bank.lasso.b)
```




```
min(bank.lasso.b$cvm)
```

```
## [1] 0.09990612
```

Step 3. Fit the model with tuning lamda and generate test error and ROC Curve.

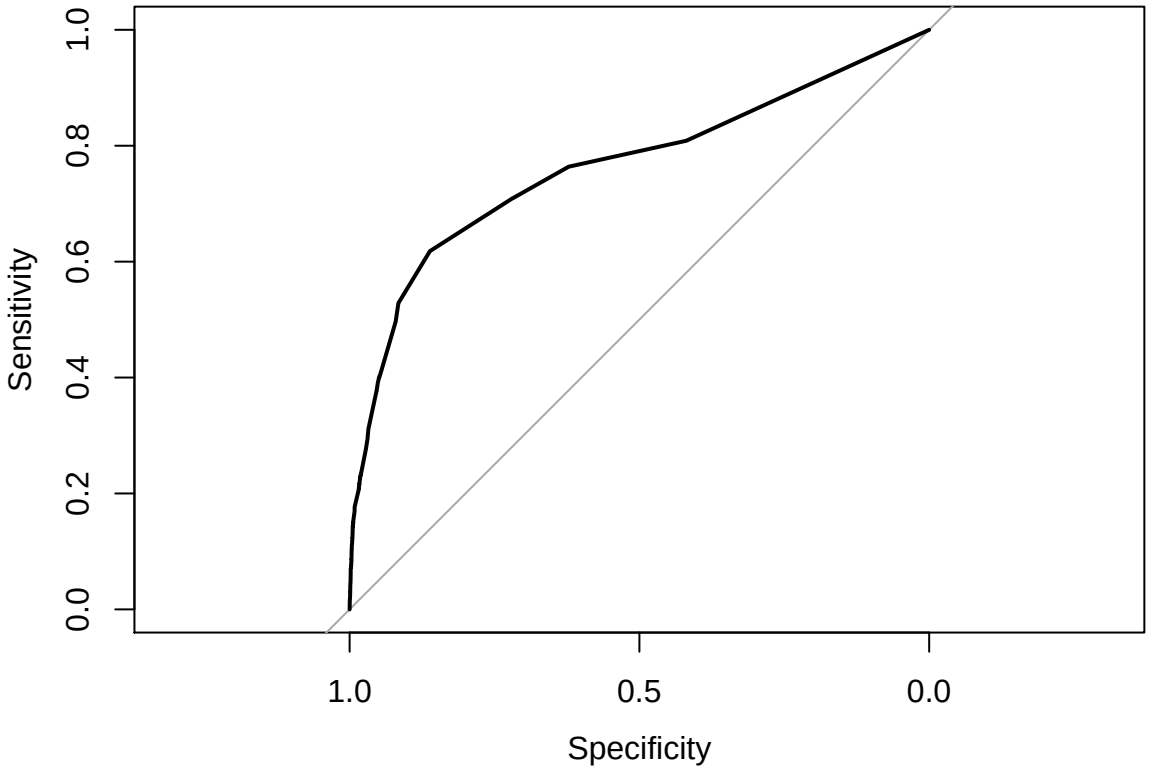
```
set.seed(1005)
fit.b<-glmnet(x.matrix.b,bank.train$y, family="binomial", alpha=1,lambda = bank.lasso.b$lambda.1se)
logistic.predict.b<-predict (fit.b, newx = x.test.b , type="response")
log.pre.b<-ifelse(logistic.predict.b<0.5,0,1)
y.test.b<-ifelse(y.test.b=='no',0,1)
table(y.test.b, log.pre.b)
```

```
##          log.pre.b
## y.test.b    0    1
##        0 9086   50
##        1  994  166
```

```
1-mean(y.test.b==log.pre.b)  #####[1] 0.1012044
```

```
## [1] 0.1013986
```

```
log.roc.b <- roc(y.test.b, as.numeric(logistic.predict.b))
plot(log.roc.b)
```

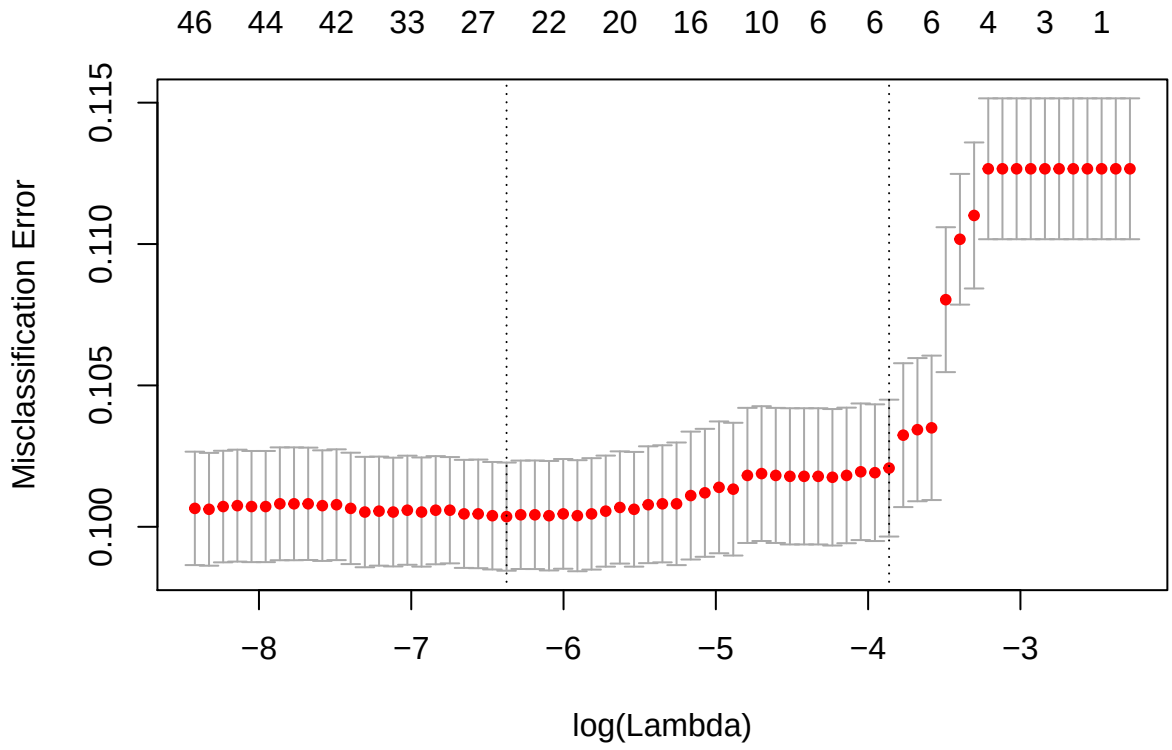



```
log.roc.b$auc
```

```
## Area under the curve: 0.764
```

Now, we will do the lasso logistic regression on the modified train dataset and test dataset:

```
set.seed(1005)
x.matrix.m<-model.matrix(~.,bank.train.mod[,-14])[,-1]
x.test.m<-model.matrix(~.,bank.test.mod[,-14])[,-1]
y.test.m=bank.test.mod$y
foldid.m=sample(1:4,size=length(bank.train.mod$y),replace=TRUE)
bank.lasso.m<-cv.glmnet(x.matrix.m,bank.train.mod$y, family="binomial", type.measure="class", alpha=1)
plot(bank.lasso.m)
```




```
min(bank.lasso.m$cvm)
```

```
## [1] 0.1003594
```

Check the test error and ROC curve:

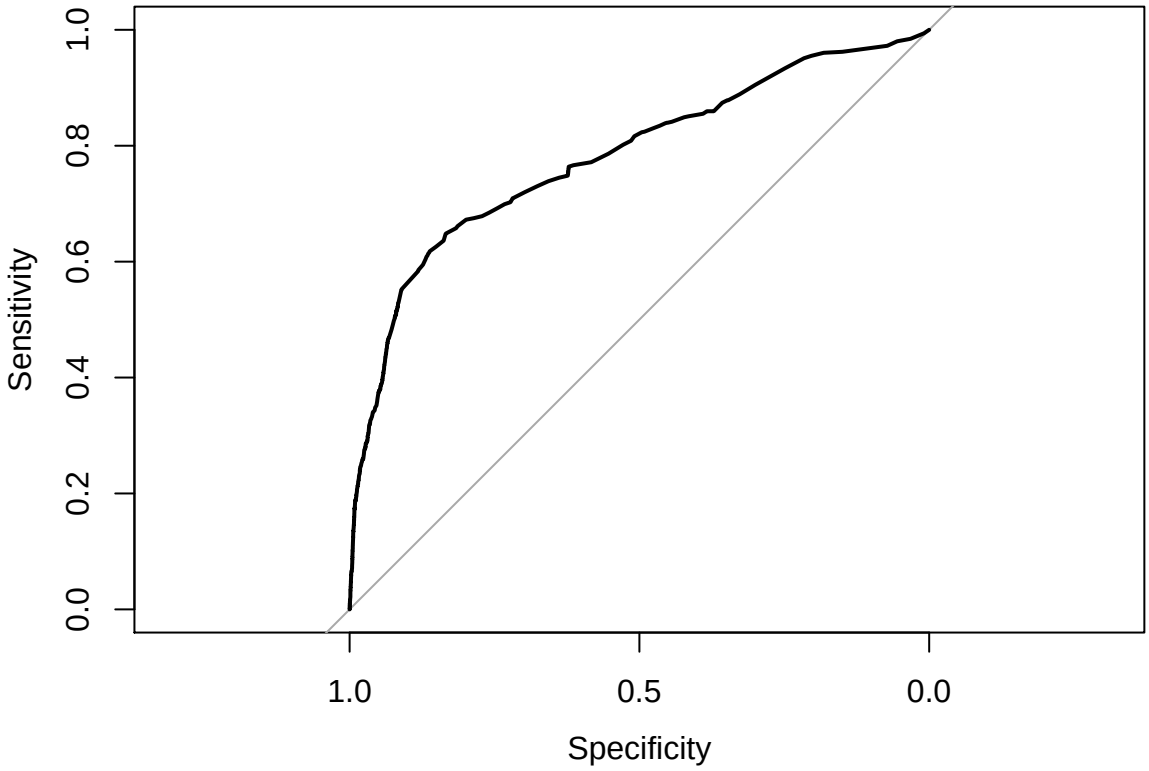
```
set.seed(1005)
fit.m<-glmnet(x.matrix.m,bank.train.mod$y, family="binomial", alpha=1,lambda = bank.lasso.m$lambda.1se)
logistic.predict.m<-predict (fit.m, newx = x.test.m , type="response")
log.pre.m<-ifelse(logistic.predict.m<0.5,0,1)
y.test.m<-ifelse(y.test.m=="no",0,1)
table(y.test.m, log.pre.m)
```

```
##          log.pre.m
## y.test.m    0    1
##        0 9062   74
##        1  973  187
```

```
1-mean(y.test.m==log.pre.m)   #####[1] 0.1007187
```

```
## [1] 0.10169
```

```
log.roc.m <- roc(y.test.m, as.numeric(logistic.predict.m))
plot(log.roc.m)
```




```
log.roc.m$auc
```

```
## Area under the curve: 0.7808
```