Predictive Modeling

Chapter 12: Discriminant Analysis and Other Linear Classification Models

STA 6543

The University of Texas at San Antonio

Overview

- Part I: General Strategies
- Part II: Regression Models
  - Chapter 6: Linear Regression and Its Cousins
  - Chapter 7: Nonlinear Regression Models
  - Chapter 8: Regression Trees and Rule-Based Models
- Part III: Classification Models
  - Chapter 11: Measuring Performance in Classification Models
  - Chapter 12: Discriminant Analysis and Other Linear Classification Models
  - Chapter 13: Nonlinear Classification Models
  - Chapter 14: Classification Trees and Rule-Based Models

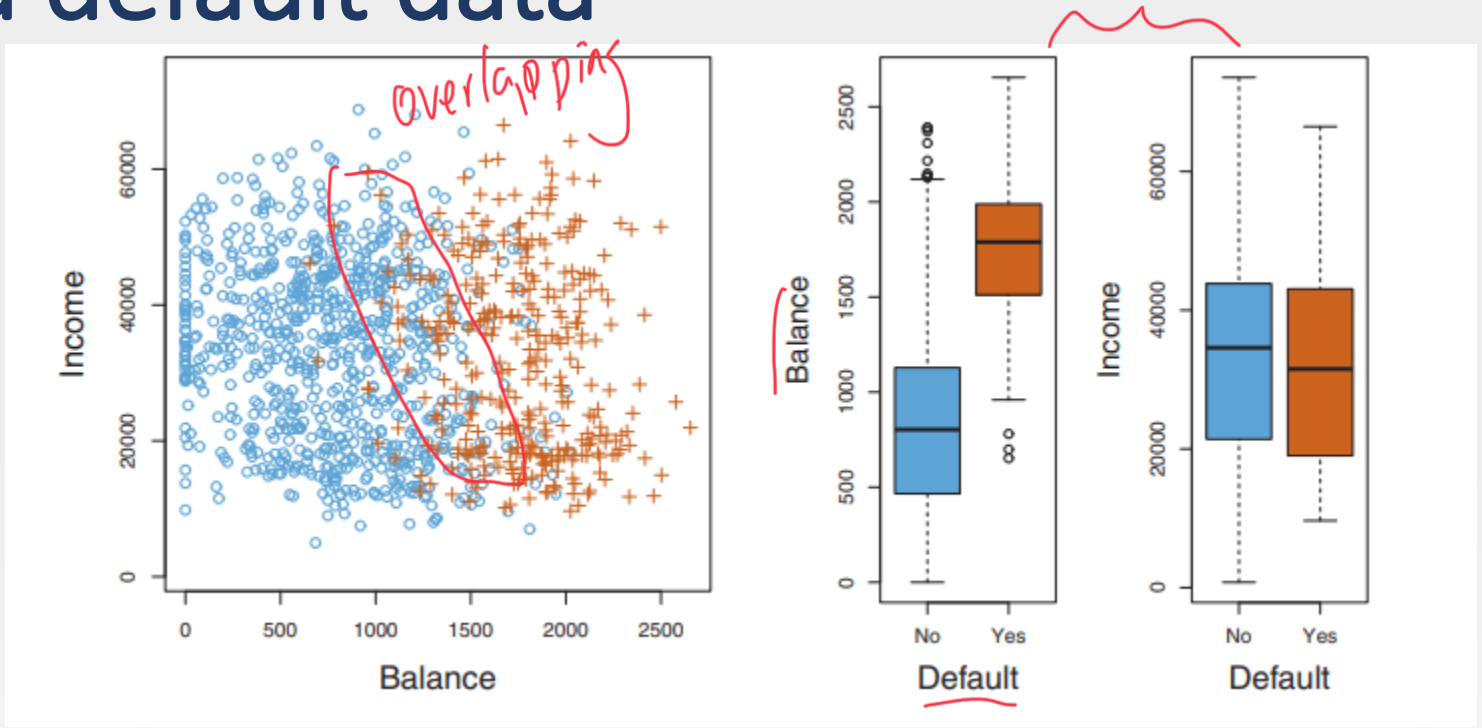
A further overview of classification

- Classification problems occur often, perhaps even more so than regression problems. Some examples include:
  - A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
  - An online banking service must be able to determine whether a transaction being performed on the site is fraudulent, based on the user's IP address, past transaction history, and so forth.
  - Based on DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing), and which are not.

predictors

features / predictors

Credit card default data



- Target: identify customers who are likely to default (Yes or No: binary response) given their monthly credit card balance (X_1) and annual incomes (X_2).

~~Tree-based~~ regression models

classification

- Various types of classification models
  - Logistic regression
  - Linear discriminant analysis (LDA)
  - Partial least squares discriminant analysis (PLSDA)
  - Penalized models
  - Nearest shrunken centroids
- R demonstrations for the stock market data

Why not linear regression?

- Suppose we code the response *default* as

$$Y = \begin{cases} 0 & \text{if no} \\ 1 & \text{if yes} \end{cases}$$

↑ which makes the response variable be numerical.

- Can we simply perform a linear regression

$$Y = X^T\beta + \epsilon$$

and classify as yes if $\hat{Y} > 0.5$?

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

X_1: balance

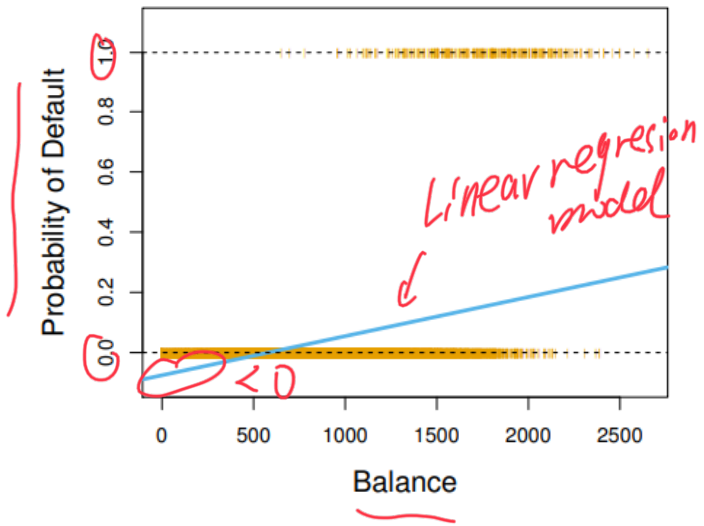
X_2: income

y: $\begin{cases} 0 \\ 1 \end{cases}$

Why not linear regression?

- For a binary response with a 0/1 coding, linear regression does make sense, and it is equivalent to linear discriminant analysis.
- Problems:
  - Some of our estimates might be outside the [0, 1] interval, making them hard to interpret as probabilities.
  - Cannot be easily extended to accommodate more than two levels.

Linear regression on default data



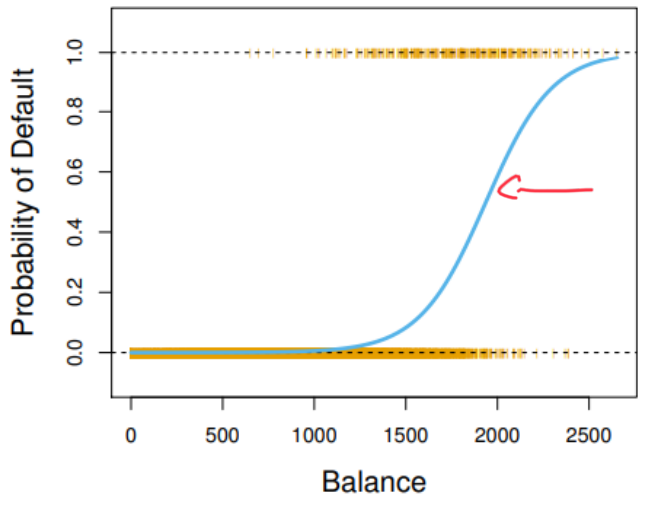


Figure: Left: *Estimated probability of default using linear regression.* Right: *Estimated probability of default using logistic regression.*

- Linear regression provides some negative estimated probabilities. Logistic regression seems well suited to the task.

Linear regression for more than two levels

- Suppose we want to predict the medical condition of a patient in the emergency room on the basis of her symptoms. There are three possible values on the response: *stroke, drug overdose, and epileptic seizure*. Consider the following coding:

$$Y = \begin{cases} 1 \text{ if } stroke \\ 2 \text{ if } drug\ overdose \\ 3 \text{ if } epileptic\ seizure \end{cases}$$

Linear regression for more than two levels

- Unfortunately, this coding implies an *ordering*, insisting that the difference between *stroke* and *drug overdose* is the same as between *drug overdose* and *epileptic seizure*.
- A different coding could end up with a totally different relationship among the three conditions, leading to fundamentally different linear models.
- Therefore, linear regression is not appropriate here.

Logistic regression

Chapter 12: Discriminant Analysis and Other Linear Classification Models

Logistic regression

- Suppose we want to predict *default = yes* using balance (X_1), then we can model the probability $p(X_1) = Pr(Y = 1 | X_1)$

$$P(X_1) = \frac{e^{\beta_0+\beta_1 X_1}}{1 + e^{\beta_0+\beta_1 X_1}}$$

where e ≈ 2.71828 is a mathematical constant, and $p(X_1)$ is called the *logistic function* having values between 0 and 1.

- A little algebra yields

$$\log\left(\frac{P(X_1)}{1 - P(X_1)}\right) = \beta_0 + \beta_1 X_1$$

The left-hand side is called the log odds or logit. The quantity $p(X_1)/[1 - p(X_1)]$ is called the odds.

Interpretation of β_1

$$\log\left(\frac{p(X_1)}{1-p(X_1)}\right) = \beta_0 + \beta_1 X_1$$

β_1: 0.0055

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513 ($\hat{\beta}_0$)	0.3612	−29.5	<0.0001
balance (X_1)	0.0055 ($\hat{\beta}_1$)	0.0002	24.9	<0.0001

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

- For every one unit change in X_1, the log odds of Y = 1 changes by β_1, or the odds changes by a factor $\exp(\beta_1)$.
- If $\beta_1 = 0$, then there is no relationship between Y and X_1.
- If $\beta_1 > 0$, then increasing X_1 will be associated with increasing $p(X_1)$ (i.e., probability of Y = 1).
- If $\beta_1 < 0$, then increasing X_1 will be associated with decreasing $p(X_1)$.
- Maximum likelihood estimation is often used to estimate X_1 .

Are the coefficients significant?

- We can perform a hypothesis test to see whether β_1 is significantly different from zero.
- We use a *Z* test instead of a *t* test, but of course that doesn't change the way we interpret the p-value.
- The p-value associated with balance is less than 0.05, we can reject H_0 : $\beta_1 = 0$ and conclude that there is indeed an association between balance (X_1) and probability of default.

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$< 0.0001 < 0.05 = \alpha$

positive.

Interpretation of β_1

- The $\widehat{\beta_1} = 0.0055$ indicates that a one dollar increase in balance is associated with an increase in the log odds of default by 0.0055 units.

$$\log \frac{p(X_1)}{1-p(X_1)} = \beta_0 + \beta_1 X_1$$

$$\frac{p(X_1)}{1-p(X_1)} = e^{\beta_0 + \beta_1 X_1} = e^{\beta_0} \cdot e^{\beta_1 X_1}$$

0.0055

Making predictions

- What is the default probability for an individual with a balance of $1000?

replace X_1 by 1000

$$\hat{p}(X_1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

$\hat{\beta}_0$ $\hat{\beta}_1$

$p(Y=1|X_1) = 0.006$

- In contrast, the predictive probability of default for an individual with a balance of $2000 is much higher and equals 0.586.

$\hat{p}(X_1) = 0.586 \Rightarrow p(Y=1|X_1) = 0.586 > p(Y=0|X_1)$

$p(Y=0|X_1) = 1 - 0.586 = 0.416$

Other considerations in logistic regression

- Analyzing rare events with logistic regression (link).
- Diagnostics and model checking for logistic regression (link).
- We use maximum likelihood estimation to train logistic regression. In machine learning, stochastic gradient ascent is often used (link)

Logistic regression for > 2 response classes

- Sometimes we wish to classify a response variable that has more than two classes. In this case, one multiple-class extension is

k = 1, 2, 3 if you have 3 classes

p predictors

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}.$$

- This form is used in the R package *glmnet*.
- When response classes are ordered, e.g., “poor”, “fair”, “good”, the ordinal logistic regression can be used.

The stock market data

- The stock market data is from the ISLR library.
- This data set consists of percentage returns for the S&P 500 stock index over 1250 days, from the beginning of 2001 until the end of 2005.
- For each date, we have recorded the percentage returns for each of the five previous trading days, Lag1 through Lag5.
- We have also recorded Volume (the number of shares traded on the previous day, in billions), Today (the percentage return on the date in question) and Direction (whether the market was Up or Down on this date: Response is binary).

The stock market data

```
# required packages
library(AppliedPredictiveModeling)
library(caret)
library(ISLR) #the stock market data
library(pROC) #roc
library(MASS) #lda
library(glmnet)

### The Stock Market Data ####
head(Smarket)
names(Smarket)
dim(Smarket)
summary(Smarket);
cor(Smarket); # pairwise correlations for numerical predictors
cor(Smarket[,-9]); ## The only substantial correlation is between Year and Volume.
attach(Smarket)
plot(Volume);
```

Data information

```
> head(Smarket)
  Year   Lag1   Lag2   Lag3   Lag4   Lag5 Volume  Today Direction
1 2001  0.381 -0.192 -2.624 -1.055  5.010 1.1913  0.959        Up
2 2001  0.959  0.381 -0.192 -2.624 -1.055 1.2965  1.032        Up
3 2001  1.032  0.959  0.381 -0.192 -2.624 1.4112 -0.623      Down
4 2001 -0.623  1.032  0.959  0.381 -0.192 1.2760  0.614        Up
5 2001  0.614 -0.623  1.032  0.959  0.381 1.2057  0.213        Up
6 2001  0.213  0.614 -0.623  1.032  0.959 1.3491  1.392        Up
> names(Smarket)
[1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"      "Volume"    "Today"     "Direction"
> dim(Smarket)
[1] 1250    9
> summary(Smarket);
      Year           Lag1                Lag2                Lag3                Lag4         
 Min.   :2001   Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.922000  
 1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500   1st Qu.:-0.640000   1st Qu.:-0.640000  
 Median :2003   Median : 0.039000   Median : 0.039000   Median : 0.038500   Median : 0.038500  
 Mean   :2003   Mean   : 0.003834   Mean   : 0.003919   Mean   : 0.001716   Mean   : 0.001636  
 3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.596750  
 Max.   :2005   Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.733000  
      Lag5              Volume           Today          Direction 
 Min.   :-4.92200   Min.   :0.3561   Min.   :-4.922000   Down:602  
 1st Qu.:-0.64000   1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648  
 Median : 0.03850   Median :1.4229   Median : 0.038500             
 Mean   : 0.00561   Mean   :1.4783   Mean   : 0.003138             
 3rd Qu.: 0.59700   3rd Qu.:1.6417   3rd Qu.: 0.596750             
 Max.   : 5.73300   Max.   :3.1525   Max.   : 5.733000             
```

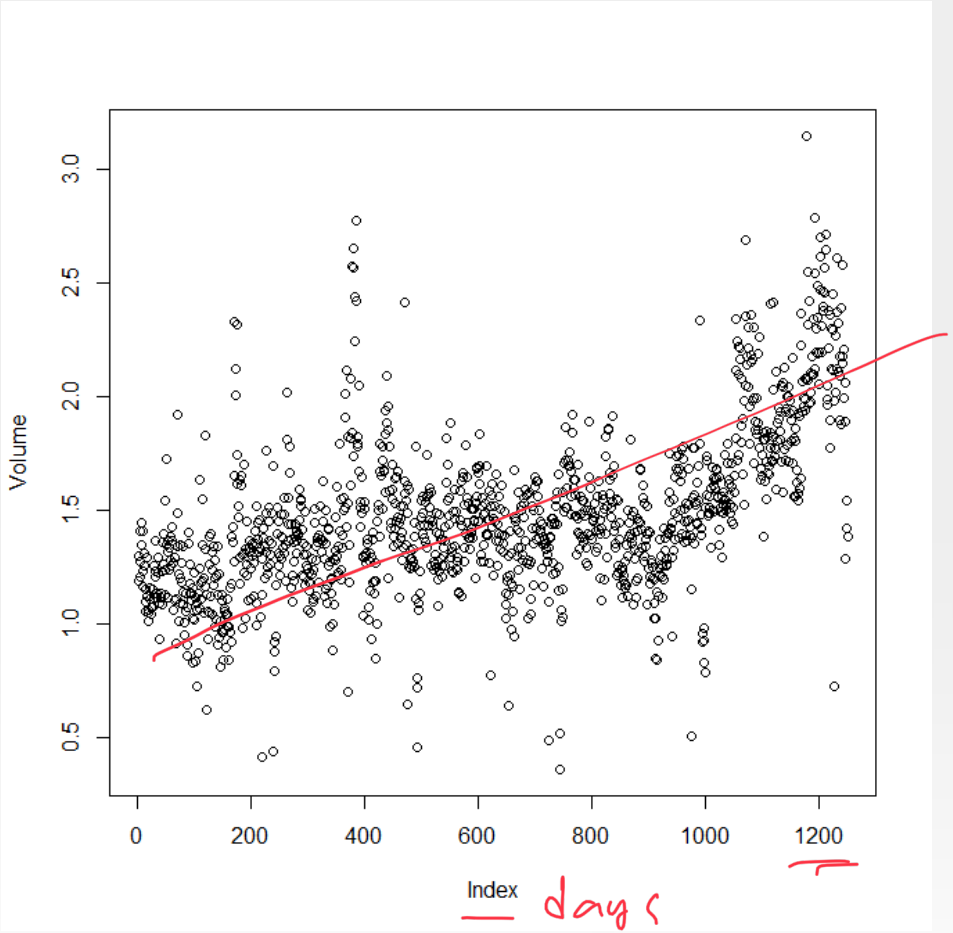
Data information

the response is binary "up" or "down"

```
> cor(Smarket); # pairwise correlations for numerical predictors
Error in cor(Smarket) : 'x' must be numeric
> cor(Smarket[,-9]); ## The only substantial correlation is between Year and Volume.
              Year         Lag1         Lag2         Lag3         Lag4         Lag5      Volume        Today
Year   1.00000000  0.029699649  0.030596422  0.033194581  0.035688718  0.029787995  0.53900647  0.030095229
Lag1   0.02969965  1.000000000 -0.026294328 -0.010803402 -0.002985911 -0.005674606  0.04090991 -0.026155045
Lag2   0.03059642 -0.026294328  1.000000000 -0.025896670 -0.010853533 -0.003557949 -0.04338321 -0.010250033
Lag3   0.03319458 -0.010803402 -0.025896670  1.000000000 -0.024051036 -0.018808338 -0.04182369 -0.002447647
Lag4   0.03568872 -0.002985911 -0.010853533 -0.024051036  1.000000000 -0.027083641 -0.04841425 -0.006899527
Lag5   0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641  1.000000000 -0.02200231 -0.034860083
Volume 0.53900647  0.040909908 -0.043383215 -0.041823686 -0.048414246 -0.022002315  1.00000000  0.014591823
Today  0.03009523 -0.026155045 -0.010250033 -0.002447647 -0.006899527 -0.034860083  0.01459182  1.000000000
> attach(Smarket)
The following objects are masked from Smarket (pos = 3):

    Direction, Lag1, Lag2, Lag3, Lag4, Lag5, Today, Volume, Year

The following objects are masked from Smarket (pos = 4):

    Direction, Lag1, Lag2, Lag3, Lag4, Lag5, Today, Volume, Year

The following objects are masked from Smarket (pos = 5):

    Direction, Lag1, Lag2, Lag3, Lag4, Lag5, Today, Volume, Year

> plot(Volume);
```

- As one would expect, the correlations between the lag variables and today's returns are close to zero. In other words, there appears to be little correlation between today's returns and previous days' returns.

Data information



- The only substantial correlation is between Year and Volume. We see from the above plot that Volume is increasing over time. In other words, the average number of shares traded daily increased from 2001 to 2005.

Data splitting

```
###Data splitting
train = which(Year<2005)
Smarket.train= Smarket[train,]; # observations before 2005 are served as test data.
Smarket.test= Smarket[-train,]; # observations from 2005 are served as test data.
```

training

Train control function

```
### Create a control function that will be used across models.
set.seed(100)
ctrl <- trainControl(method = "LGOCV",
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE,
                     savePredictions = TRUE)

# LGOCV: Repeated Train/Test Splits Estimated (25 reps, 75%)
```

Logistic regression

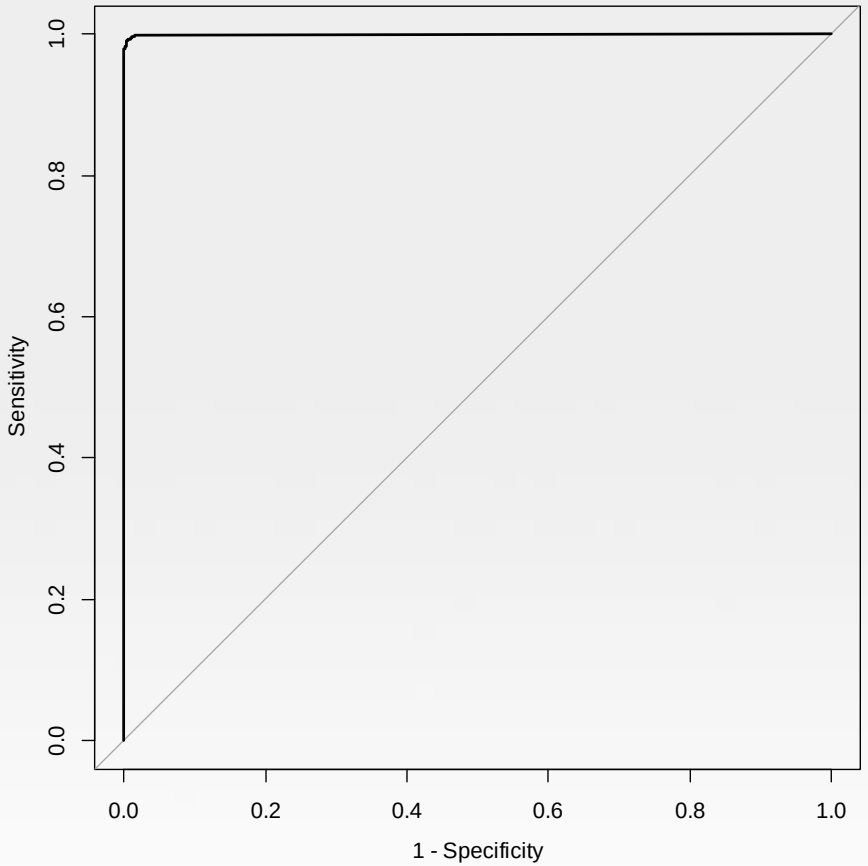
```
set.seed(476)
logisticTune <- train(x = as.matrix(Smarket.train[,1:8]),
y = Smarket.train$Direction,
method = "glm",
metric = "ROC",
trControl = ctrl)
logisticTune


### Save the test set results in a data frame
testResults <- data.frame(obs = Smarket.test$Direction,
                          logistic = predict(logisticTune, Smarket.test))
```

test predictors

Roc curve of the logistic regression

```
library(pROC)

### Predict the test set based the logistic regression
Smarket.test$logistic <- predict(logisticTune,Smarket.test, type = "prob")[,1]

#ROC for logistic model
logisticROC <- roc(Smarket.test$Direction, Smarket.test$logistic)
plot(logisticROC, col=1, lty=1, lwd=2)

#Confusion matrix of logistic model
confusionMatrix(data = predict(logisticTune, Smarket.test), reference =
Smarket.test$Direction)
```

ROC curve of the logistic regression



Confusion matrix

```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
      Down  110   0
      Up      1 141

               Accuracy : 0.996
                 95% CI : (0.9781, 0.9999)
    No Information Rate : 0.5595
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9919

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9910
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9930
             Prevalence : 0.4405
         Detection Rate : 0.4365
   Detection Prevalence : 0.4365
      Balanced Accuracy : 0.9955

       'Positive' Class : Down
```