Логистическая регрессия

Возможно потребуется установить пакет

```
library(ISLR)
```

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

```
attach(Smarket)
help("Smarket")
```

```
## starting httpd help server ... done
```

Разделим данные на два подмножества: для обучения и для проверки Выберем для обучения данные ранее 2005г, остальные используем для проверочного подмножества

Векторы данных, содержат флажки для каждой далее отбираемой строки

```
train <- Year < 2005
test <- !train
```

Выберем данные из 8го столбца (Today)

```
training_data = Smarket[train, -8]
testing_data = Smarket[test, -8]
```

для тестов отберём значения колонки Direction которые будем стараться предсказать

```
testing_y = Direction[test]
```

Выполним подгонку логистической модели. Знак . означает выбор всех предикторов (влияющих переменных)

```
fit <- glm(Direction ~ ., data = training_data, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = Direction ~ ., family = "binomial", data = training_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.382  -1.184   1.030   1.146   1.451
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.990e+02  1.185e+02  -2.523   0.0116 *
## Year         1.495e-01  5.922e-02   2.525   0.0116 *
## Lag1        -5.824e-02  5.200e-02  -1.120   0.2627
## Lag2        -5.378e-02  5.210e-02  -1.032   0.3019
## Lag3        -1.059e-03  5.190e-02  -0.020   0.9837
## Lag4        -2.359e-03  5.199e-02  -0.045   0.9638
## Lag5        -1.074e-02  5.139e-02  -0.209   0.8344
## Volume      -2.665e-01  2.481e-01  -1.074   0.2828
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1374.7  on 990  degrees of freedom
## AIC: 1390.7
##
## Number of Fisher Scoring iterations: 3
```

Дефолты

```
attach(Default)
help("Default")

summary(Default)
```

```
##  default    student       balance           income
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
##                        Median : 823.6   Median :34553
##                        Mean   : 835.4   Mean   :33517
##                        3rd Qu.:1166.3   3rd Qu.:43808
##                        Max.   :2654.3   Max.   :73554
```

```
train_d <- income < mean(income)
test_d <- !train_d
```

```
training_data_d <- Default[train_d,]
testing_data_d <- Default[test_d, -1]
```

```
testing_y_d <- default[test_d]
fit_d <- glm(default ~ ., data = training_data_d, family = "binomial")
summary(fit_d)
```

```
##
## Call:
## glm(formula = default ~ ., family = "binomial", data = training_data_d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1866  -0.1537  -0.0615  -0.0233   3.6746
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.020e+01  7.567e-01 -13.478  < 2e-16 ***
## studentYes  -7.733e-01  2.990e-01  -2.586  0.00971 **
## balance      5.603e-03  3.174e-04  17.651  < 2e-16 ***
## income      -1.467e-05  2.108e-05  -0.696  0.48651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1497.81  on 4735  degrees of freedom
## Residual deviance:  816.56  on 4732  degrees of freedom
## AIC: 824.56
##
## Number of Fisher Scoring iterations: 8
```

предскажем вероятность дефолта

```
logistic_probabs_d <- predict(fit_d, testing_data_d, type = "response")
head(logistic_probabs_d)
```

```
##            1            4            5            9           14
## 1.155872e-03 4.275550e-04 1.725127e-03 1.416637e-02 5.758818e-04
##           16
## 9.555899e-05
```

просматривать вероятности не очень удобно, категоризируем вероятность приняв за дефолт вероятность большую чем 50%

Подготовим вектор с длиной равной вектору проверочных данных

```
logistic_pred_y_d <- rep('No',  length(testing_y_d))
#как мы поминм из лаборторных! rep повторяет занчение указанное чилсо раз
```

Изменим флаг дефолта для тех у кого вероятность этого больше 50%

```
logistic_pred_y_d[logistic_probabs_d > 0.5] = 'Yes'
```

Покажем таблицу истинности

```
table(logistic_pred_y_d, testing_y_d)
```

```
##                  testing_y_d
## logistic_pred_y_d   No  Yes
##               No  5097  117
##               Yes    9   41
```

процент ошибок классифицирования

```
mean(logistic_pred_y_d != testing_y_d)
```

```
## [1] 0.02393617
```