Chapter 10

Problem 10.3

**Sales of Riding Mowers.** A company that manufactures riding mowers wants to

identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer

is interested in classifying households as prospective owners or nonowners

on the basis of Income (in $1000s) and Lot Size (in 1000 ft2). The marketing expert

looked at a random sample of 24 households, given in the file *RidingMowers.csv*. Use

all the data to fit a logistic regression of ownership on the two predictors.

**a.** What percentage of households in the study were owners of a riding mower?

**A: 12 out of 24 are owners. 50% of households owned a riding mower.**

**b.** Create a scatter plot of Income vs. Lot Size using color or symbol to distinguish

owners from nonowners. From the scatter plot, which class seems to have a higher

average income, owners or nonowners?

**A:**

**From the scatterplot it is clear that owners(green) have more income compared to non-owners and since there are equal number of owners and non-owners it is safe to conclude that the average income of owners will be higher.**

**c.** Among nonowners, what is the percentage of households classified correctly?

At cut off = 0.5, the accuracy is 80 %.

The model predicted 5 non owners as non-owners correctly. But predicted an owner as a non-owner incorrectly.

**Among nonowners, the percentage of households classified correctly = 5/6\*100 = 83.33%**

The R code and confusion matrix is as follows.

> confusionMatrix(table(predict(log.reg,newdata = validds.df,type="response")>=0.5,validds.df$Ownership == 0))

Confusion Matrix and Statistics

Owner Nonowner

Owner 3 1

Nonowner 1 5

Accuracy : 0.8

95% CI : (0.4439, 0.9748)

No Information Rate : 0.6

P-Value [Acc > NIR] : 0.1673

Kappa : 0.5833

Mcnemar's Test P-Value : 1.0000

Sensitivity : 0.7500

Specificity : 0.8333

Pos Pred Value : 0.7500

Neg Pred Value : 0.8333

Prevalence : 0.4000

Detection Rate : 0.3000

Detection Prevalence : 0.4000

Balanced Accuracy : 0.7917

'Positive' Class : FALSE

**d.** To increase the percentage of correctly classified nonowners, should the cutoff

probability be increased or decreased?

While increasing the cut off:

**0.7**

> confusionMatrix(table(predict(log.reg,newdata = validds.df,type="response")>=0.7,validds.df$Ownership == 0))

Confusion Matrix and Statistics

FALSE TRUE

FALSE 3 1

TRUE 1 5

Accuracy : 0.8

95% CI : (0.4439, 0.9748)

No Information Rate : 0.6

P-Value [Acc > NIR] : 0.1673

Kappa : 0.5833

Mcnemar's Test P-Value : 1.0000

Sensitivity : 0.7500

Specificity : 0.8333

Pos Pred Value : 0.7500

Neg Pred Value : 0.8333

Prevalence : 0.4000

Detection Rate : 0.3000

Detection Prevalence : 0.4000

Balanced Accuracy : 0.7917

'Positive' Class : FALSE

**0.9**

> confusionMatrix(table(predict(log.reg,newdata = validds.df,type="response")>=0.9,validds.df$Ownership == 0))

Confusion Matrix and Statistics

FALSE TRUE

FALSE 4 2

TRUE 0 4

Accuracy : 0.8

95% CI : (0.4439, 0.9748)

No Information Rate : 0.6

P-Value [Acc > NIR] : 0.1673

Kappa : 0.6154

Mcnemar's Test P-Value : 0.4795

Sensitivity : 1.0000

Specificity : 0.6667

Pos Pred Value : 0.6667

Neg Pred Value : 1.0000

Prevalence : 0.4000

Detection Rate : 0.4000

Detection Prevalence : 0.6000

Balanced Accuracy : 0.8333

'Positive' Class : FALSE

While decreasing the cut off:

**0.4**

> confusionMatrix(table(predict(log.reg,newdata = validds.df,type="response")>=0.4,validds.df$Ownership == 0))

Confusion Matrix and Statistics

FALSE TRUE

FALSE 3 1

TRUE 1 5

Accuracy : 0.8

95% CI : (0.4439, 0.9748)

No Information Rate : 0.6

P-Value [Acc > NIR] : 0.1673

Kappa : 0.5833

Mcnemar's Test P-Value : 1.0000

Sensitivity : 0.7500

Specificity : 0.8333

Pos Pred Value : 0.7500

Neg Pred Value : 0.8333

Prevalence : 0.4000

Detection Rate : 0.3000

Detection Prevalence : 0.4000

Balanced Accuracy : 0.7917

'Positive' Class : FALSE

**0.25**

> confusionMatrix(table(predict(log.reg,newdata = validds.df,type="response")>=0.25,validds.df$Ownership == 0))

Confusion Matrix and Statistics

FALSE TRUE

FALSE 2 0

TRUE 2 6

Accuracy : 0.8

95% CI : (0.4439, 0.9748)

No Information Rate : 0.6

P-Value [Acc > NIR] : 0.1673

Kappa : 0.5455

Mcnemar's Test P-Value : 0.4795

Sensitivity : 0.50

Specificity : 1.00

Pos Pred Value : 1.00

Neg Pred Value : 0.75

Prevalence : 0.40

Detection Rate : 0.20

Detection Prevalence : 0.20

Balanced Accuracy : 0.75

'Positive' Class : FALSE

**It can be concluded that when the cut off decreases, the percentage of correctly classified nonowners increases. At 0.25 cut off , the number of correctly classified non owners is maximum.**

**e.** What are the odds that a household with a $60K income and a lot size of 20,000

ft2 is an owner?

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 28.14728 14.93615 1.885 0.0595 .

Income -0.06299 0.05289 -1.191 0.2337

Lot\_Size -1.24586 0.67680 -1.841 0.0656 .

> odds <- exp(-28.14728+0.06299\*60.0+1.24586\*20)

> print(odds)

1.732075

> exp(coef(log.reg))

(Intercept) Income Lot\_Size

1675748654001.5273437 0.9389568 0.2876937

Every unit increase in Income, the odds of owning a riding mower is

0.9389568

For a household with income $60,000, the odds that they own a riding mower is 60\*0.9389568 = 56.3374

Every unit increase in Lot size, the odds of owning a riding mower is

0.2876937

For a household with lot size 20000, the odds that they own a riding mower is 20\*0.2876937 = 5.7538

> exp(cbind(OR = coef(log.reg), confint(log.reg)))

Waiting for profiling to be done...

OR 2.5 % 97.5 %

(Intercept) 1675748654001.5273437 1259.75221603 4117395881178132129280402240682.0000000

Income 0.9389568 0.81516630 1.0214018

Lot\_Size 0.2876937 0.04201136 0.7715462

Odds(Owner =1 |Income = 60k$ | Lot Size = 20000ft2) = e^bo+e^b1\*60+e^b2\*20

**f.** What is the classification of a household with a $60K income and a lot size of

20,000 ft2? Use cutoff = 0.5.

df <- data.frame("Income"=60,"Lot\_Size"=20)

> predict(log.reg,newdata = df,type="response")>=0.5

1

FALSE

The classification is Non owner.

P(Y=1|Income = *60&lotsize=20*) =1/1 + e^-(b0+b1\*income+b2\*lotsize)

= 1/1+e^-1.1042

=1/1+0.3314759

=0.751046

75.1046 %

**g.** What is the minimum income that a household with 16,000 ft2 lot size should have

before it is classified as an owner?

> confusionMatrix(table(predict(log.reg,newdata = validds.df,type="response")>=0.30,validds.df$Ownership == 1))

Confusion Matrix and Statistics

FALSE TRUE

FALSE 1 2

TRUE 5 2

A cut off of 0.3 or more the model can predict 2 owners correctly out of 4.

From the confusion matrix, p = 0.3

Odds = p/1-p = 0.3/0.7 = 0.4285

Log(Odds) = b0+b1\*X1+b2\*X2

Log(0.4285) = 28.14728-0.06299\*X-1.24586\*16

-0.3680 = 28.14728 -0.6299\*Income-19.9337

Income = (28.14728 -19.9337+0.3680)/0.6299

= 13.6237

The minimum income should be greater than $13k

Call:

glm(formula = Ownership ~ ., family = "binomial", data = trainds.df)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8200 -0.5018 -0.2828 0.2580 1.6186

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 28.14728 14.93615 1.885 0.0595 .

Income -0.06299 0.05289 -1.191 0.2337

Lot\_Size -1.24586 0.67680 -1.841 0.0656 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19.1214 on 13 degrees of freedom

Residual deviance: 9.7639 on 11 degrees of freedom

AIC: 15.764

Number of Fisher Scoring iterations: 6

Valid 10

4 owners

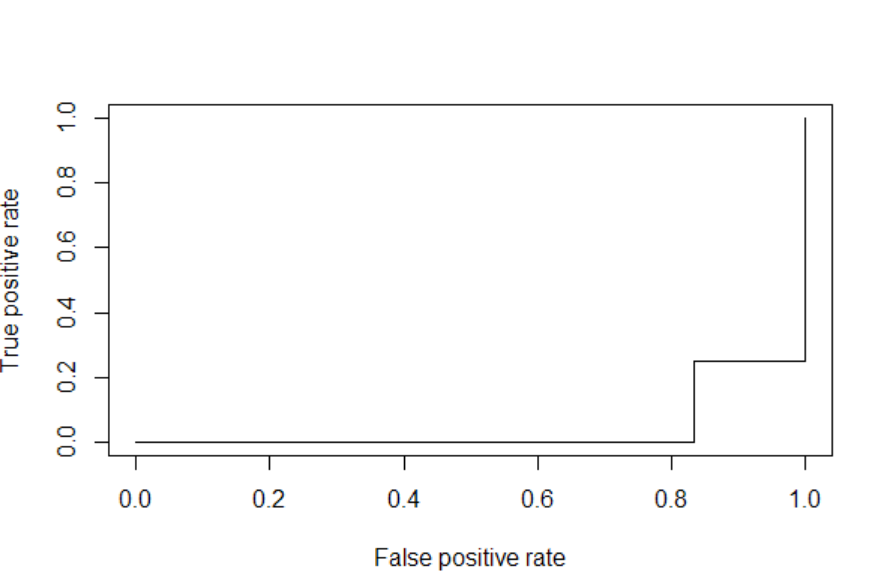
6 non owners

|  |  |  |
| --- | --- | --- |
| **2** | 1 | 0.861948383 |
| **5** | 1 | 0.001187459 |
| **7** | 1 | 0.358403668 |
| **8** | 1 | 0.006859890 |
| **14** | 0 | 0.251198066 |
| **15** | 0 | 0.933207067 |
| **20** | 0 | 0.743914787 |
| **22** | 0 | 0.933801336 |
| **23** | 0 | 0.999443330 |
| **24** | 0 | 0.996797020 |

Use of options(scipen=999)?

Logical meaning of cut off at 0.3, 5 non owners were predicted right and 2 owners were predicted as non owners.

At 0.3 actual\non owner count becomes 7 which is not the case. How?



Call:

glm(formula = Ownership ~ ., family = "binomial", data = trainds.df)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.46031 -0.39382 -0.07795 0.24078 1.70730

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 27.52640 16.13508 1.706 0.088 .

Income -0.11295 0.06907 -1.635 0.102

Lot\_Size -1.09268 0.68325 -1.599 0.110

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 19.1214 on 13 degrees of freedom

Residual deviance: 7.8235 on 11 degrees of freedom

AIC: 13.824

Number of Fisher Scoring iterations: 6

Confusion Matrix and Statistics

Predicted

Actual FALSE TRUE

FALSE 0 6

TRUE 4 0

Predicted True- 6 times and False-4 times. Actually, there were 4 True and 6 false. The classifier predicted wrongly.

Accuracy : 0

95% CI : (0, 0.3085)

No Information Rate : 0.6

P-Value [Acc > NIR] : 1.0000

Kappa : -0.9231

Mcnemar's Test P-Value : 0.7518

Sensitivity : 0.0

Specificity : 0.0

Pos Pred Value : 0.0

Neg Pred Value : 0.0

Prevalence : 0.4

Detection Rate : 0.0

Detection Prevalence : 0.6

Balanced Accuracy : 0.0

'Positive' Class : FALSE