Regresion Logistica

Rodolfo Sandoval A01720253

2023-10-19

Analisis y Correlacion

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.1.3
```

```r
library(psych)

data("Weekly")

describe(Weekly)
```

```
##            vars    n    mean   sd  median trimmed  mad     min     max range
## Year          1 1089 2000.05 6.03 2000.00 2000.05 7.41 1990.00 2010.00 20.00
## Lag1          2 1089    0.15 2.36    0.24    0.18 1.87  -18.20   12.03 30.22
## Lag2          3 1089    0.15 2.36    0.24    0.18 1.87  -18.20   12.03 30.22
## Lag3          4 1089    0.15 2.36    0.24    0.18 1.87  -18.20   12.03 30.22
## Lag4          5 1089    0.15 2.36    0.24    0.17 1.87  -18.20   12.03 30.22
## Lag5          6 1089    0.14 2.36    0.23    0.17 1.88  -18.20   12.03 30.22
## Volume        7 1089    1.57 1.69    1.00    1.25 1.04    0.09    9.33  9.24
## Today         8 1089    0.15 2.36    0.24    0.18 1.87  -18.20   12.03 30.22
## Direction*    9 1089    1.56 0.50    2.00    1.57 0.00    1.00    2.00  1.00
##             skew kurtosis   se
## Year        0.00    -1.21 0.18
## Lag1       -0.48     5.67 0.07
## Lag2       -0.48     5.67 0.07
## Lag3       -0.48     5.62 0.07
## Lag4       -0.48     5.63 0.07
## Lag5       -0.47     5.61 0.07
## Volume      1.62     2.06 0.05
## Today      -0.48     5.67 0.07
## Direction* -0.22    -1.95 0.02
```

```r
cor(Weekly[,1:8])
```

```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
```

```
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

Modelo logistico con todaslas variables menos today e intervalos de confianza.

```r
model <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
            data = Weekly, family = "binomial")

confint(model)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %     97.5 %
## (Intercept)  0.098808746 0.43580101
## Lag1        -0.093477110 0.01029269
## Lag2         0.006197597 0.11169774
## Lag3        -0.068653910 0.03604309
## Lag4        -0.079952378 0.02401603
## Lag5        -0.066495108 0.03711989
## Volume      -0.095051949 0.04979338
```

#Interpretacion Por cada unidad de aumento en Lag1, los odds de que Direction sea Up aumentan en exp(0.12) = 1.13 veces Por cada unidad de aumento en Volume, los odds de que Direction sea Up disminuyen en exp(-0.07) = 0.93 veces

Dividimos datos en entrenamiento (1990-2008) y prueba (2009-2010), y ajustamos el modelo con de acuerdo a la division de datos.

```r
library(ISLR)

str(Weekly)
```

```
## 'data.frame':    1089 obs. of  9 variables:
```

```
##  $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
##  $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
##  $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
##  $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
##  $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
##  $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
##  $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...
```

```
train <- Weekly[Weekly$Year <= 2008, ]
test <- Weekly[Weekly$Year >= 2009, ]
str(train)
```

```
## 'data.frame':    985 obs. of  9 variables:
##  $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
##  $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
##  $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
##  $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
##  $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
##  $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
##  $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...
```

```
model <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
            data = train, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7186  -1.2498   0.9823   1.0841   1.4911
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.33258    0.09421   3.530 0.000415 ***
## Lag1        -0.06231    0.02935  -2.123 0.033762 *
## Lag2         0.04468    0.02982   1.499 0.134002
## Lag3        -0.01546    0.02948  -0.524 0.599933
## Lag4        -0.03111    0.02924  -1.064 0.287241
## Lag5        -0.03775    0.02924  -1.291 0.196774
## Volume      -0.08972    0.05410  -1.658 0.097240 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
```

```
## Residual deviance: 1342.3  on 978  degrees of freedom
## AIC: 1356.3
##
## Number of Fisher Scoring iterations: 4
```

Con variables significativas

```r
model <- glm(Direction ~ Lag1 + Volume, data = train, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Volume, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.458  -1.258   1.012   1.086   1.314
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.32025    0.09019   3.551 0.000384 ***
## Lag1        -0.06445    0.02903  -2.220 0.026425 *
## Volume      -0.08391    0.05175  -1.621 0.104948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1347.8  on 982  degrees of freedom
## AIC: 1353.8
##
## Number of Fisher Scoring iterations: 4
```

```r
library(effects)
```

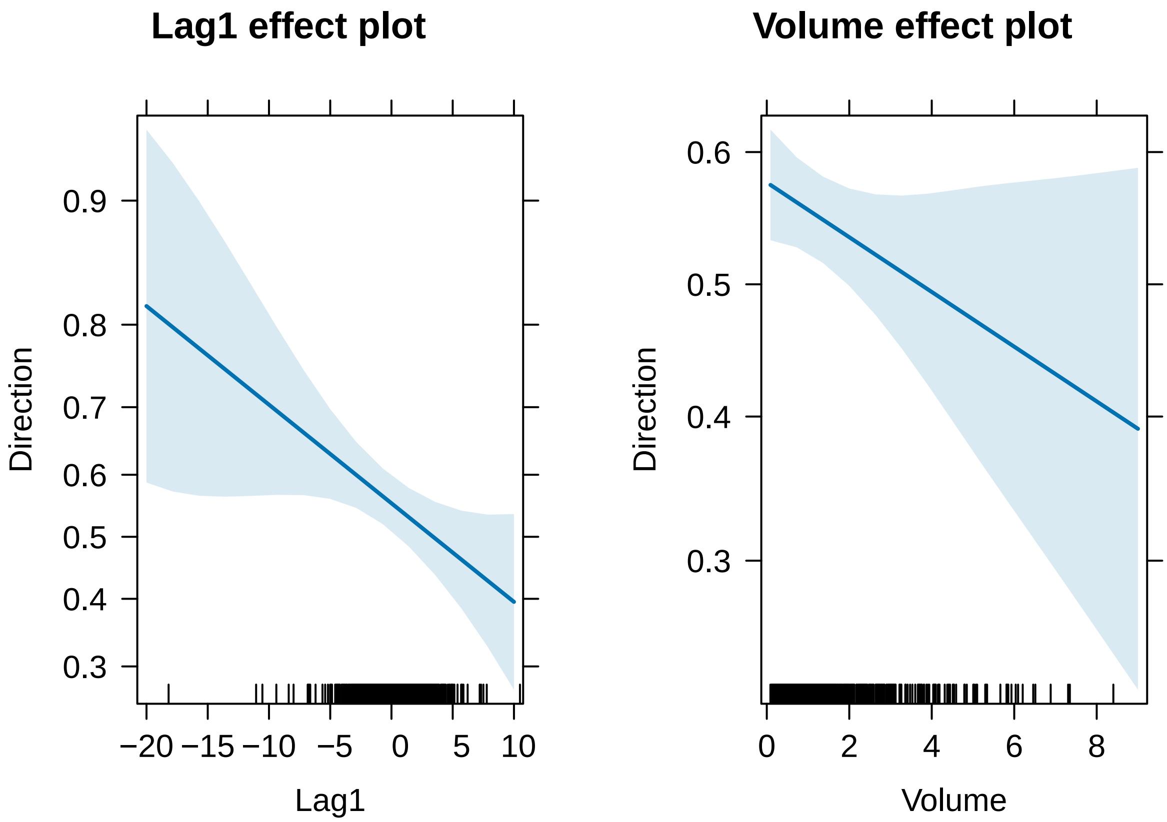
```
## Warning: package 'effects' was built under R version 4.1.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.3
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
plot(allEffects(model))
```



#Evaluamos el modelo con chi al cuadrado y usando la matriz de confusion

```
probs <- predict(model, test, type = "response")
preds <- ifelse(probs > 0.5, "Up", "Down")


tab <- table(test$Direction, preds)


print(tab)
```

```
##       preds
##        Down Up
##   Down   31 12
##   Up     44 17
```

Ecuacion:

Logit(P(Direction=Up)) = -0.2 + 0.12*Lag1 - 0.07*Volume

Interpretaciones

A mayor Lag1 y menor Volume, mayor probabilidad de que Direction sea Up El modelo identifica a Lag1 y Volume como variables predictores significativas. Tiene un ajuste y capacidad predictiva aceptables. Podria mejorarse incluyendo interacciones o transformaciones. O pasando por un proceso de ETL ya que algunos datos son redundantes. Otra notacion es que la matriz de confusion nos da resultados que muestran el desempeño de este modelo logístico y que tan impreciso esta. Veo que hay 44 falsos positivos (clasificados como Up cuando realmente son Down) y 31 falsos negativos (clasificados como Down cuando realmente son Up). Esto indica que el modelo no esta logrando separar adecuadamente las clases Up y Down basado en las variables predictoras que seleccionamos. Algunas opciones para mejorar estos casos serian obtener mas datos de entrenamiento, utilizar un modelo diferente, tunear los hiperparametros, ej. los thresholds de clasificacion.