

A5-Regresión logística

Diego Alberto Baños Lopez
A01275100

2023-10-18

Encuentra un modelo logístico para encontrar el mejor conjunto de predictores que auxilien a clasificar la dirección de cada observación.

Análisis del dataset

```
# Librerías a usar
library(ISLR)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.3      v readr      2.1.4
## v forcats    1.0.0      v stringr    1.5.0
## v ggplot2    3.4.4      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.0
## v purrr      1.0.2
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Para llamar al dataset de manera mas sencilla usamos attach()
# Asi solo se requiere llamar las columnas
attach(Weekly)
```

```
# Vista previa del dataset
# Observamos estadísticas y graficas de como se comporta el dataset
head(Weekly)
```

```
##   Year  Lag1  Lag2  Lag3  Lag4  Lag5  Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514       Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712       Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178       Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

```
summary(Weekly)
```

```
##      Year      Lag1      Lag2      Lag3
## Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
## 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
## Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
## Mean   :2000   Mean    :  0.1506   Mean    :  0.1511   Mean    :  0.1472
```

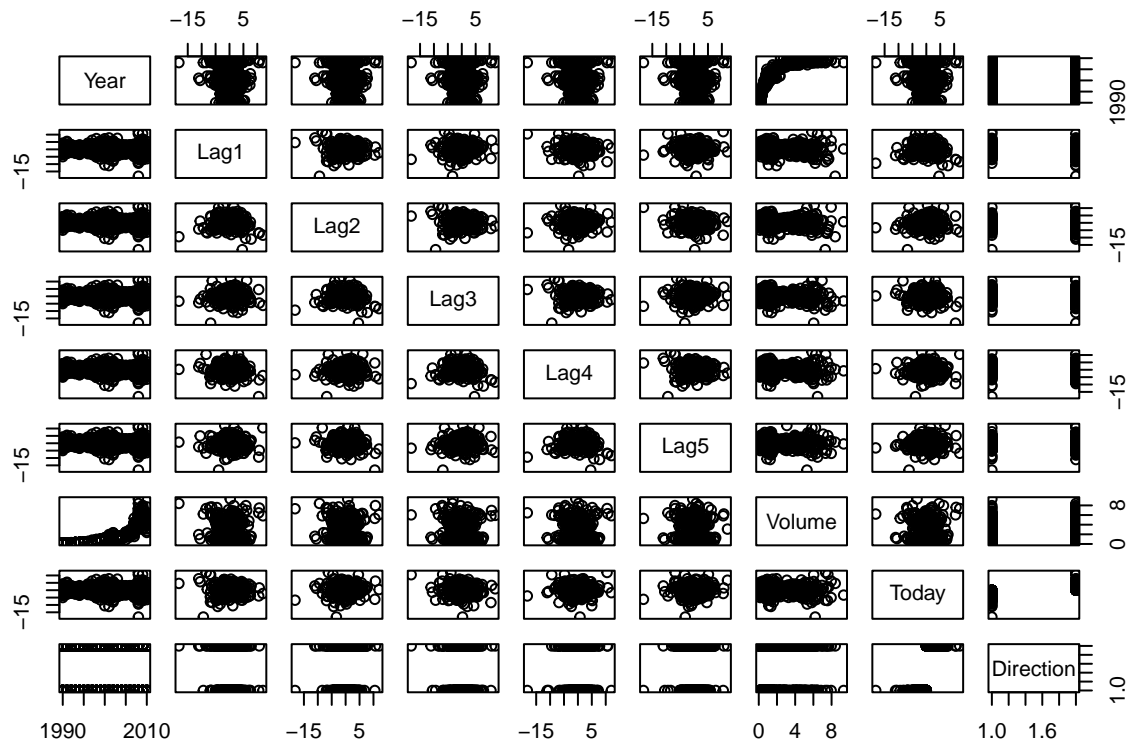
```
## 3rd Qu.:2005    3rd Qu.: 1.4050    3rd Qu.: 1.4090    3rd Qu.: 1.4090
## Max.    :2010    Max.    : 12.0260    Max.    : 12.0260    Max.    : 12.0260
##      Lag4          Lag5          Volume          Today
## Min.    :-18.1950    Min.    :-18.1950    Min.    :0.08747    Min.    :-18.1950
## 1st Qu.: -1.1580    1st Qu.: -1.1660    1st Qu.:0.33202    1st Qu.: -1.1540
## Median : 0.2380    Median : 0.2340    Median :1.00268    Median : 0.2410
## Mean   : 0.1458    Mean   : 0.1399    Mean   :1.57462    Mean   : 0.1499
## 3rd Qu.: 1.4090    3rd Qu.: 1.4050    3rd Qu.:2.05373    3rd Qu.: 1.4050
## Max.    : 12.0260    Max.    : 12.0260    Max.    :9.32821    Max.    : 12.0260
## Direction
## Down:484
## Up  :605
##
##
##
##
```

```
glimpse(Weekly)
```

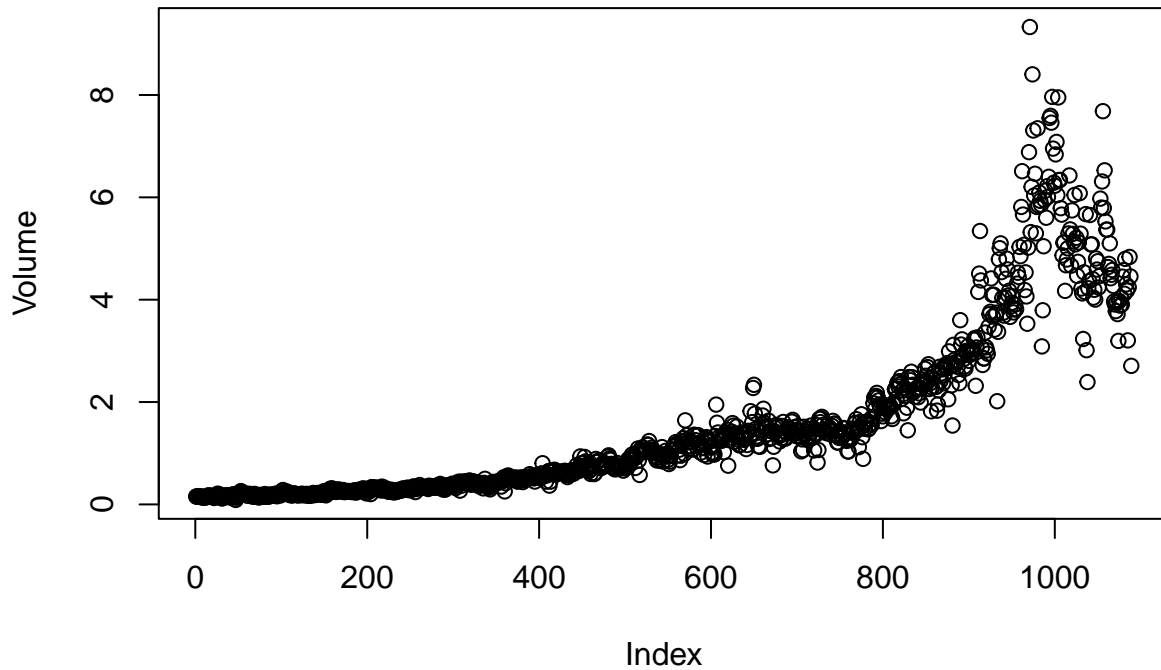
```
## Rows: 1,089
## Columns: 9
## $ Year      <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, ~
## $ Lag1      <dbl> 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0.807, 0~
## $ Lag2      <dbl> 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0~
## $ Lag3      <dbl> -3.936, 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, --
## $ Lag4      <dbl> -0.229, -3.936, 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, ~
## $ Lag5      <dbl> -3.484, -0.229, -3.936, 1.572, 0.816, -0.270, -2.576, 3.514,~
## $ Volume     <dbl> 0.1549760, 0.1485740, 0.1598375, 0.1616300, 0.1537280, 0.154~
## $ Today      <dbl> -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0.807, 0.041, 1~
## $ Direction <fct> Down, Down, Up, Up, Up, Down, Up, Up, Up, Down, Down, Up, Up~
```

```
plot(Weekly)
```

```
pairs(Weekly)
```



```
plot(Volume)
```



Coefficiente de correlación

```
cor(Weekly[, -9])
```

```
##           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
```

Formula un modelo logístico con todas las variables menos la variable “Today”. Calcula los intervalos de confianza para las β_i . Detecta variables que influyen y no influyen en el modelo. Interpreta el efecto de las variables en los odds (momios).

```
modelo_log_m <- glm(Direction ~ . - Today, data = Weekly, family = binomial)
summary(modelo_log_m)
```

```
##
## Call:
## glm(formula = Direction ~ . - Today, family = binomial, data = Weekly)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 17.225822  37.890522   0.455  0.6494
## Year        -0.008500   0.018991  -0.448  0.6545
## Lag1        -0.040688   0.026447  -1.538  0.1239
## Lag2         0.059449   0.026970   2.204  0.0275 *
## Lag3        -0.015478   0.026703  -0.580  0.5622
## Lag4        -0.027316   0.026485  -1.031  0.3024
## Lag5        -0.014022   0.026409  -0.531  0.5955
## Volume       0.003256   0.068836   0.047  0.9623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.2  on 1081  degrees of freedom
## AIC: 1502.2
##
## Number of Fisher Scoring iterations: 4
```

```
contrasts(Direction)
```

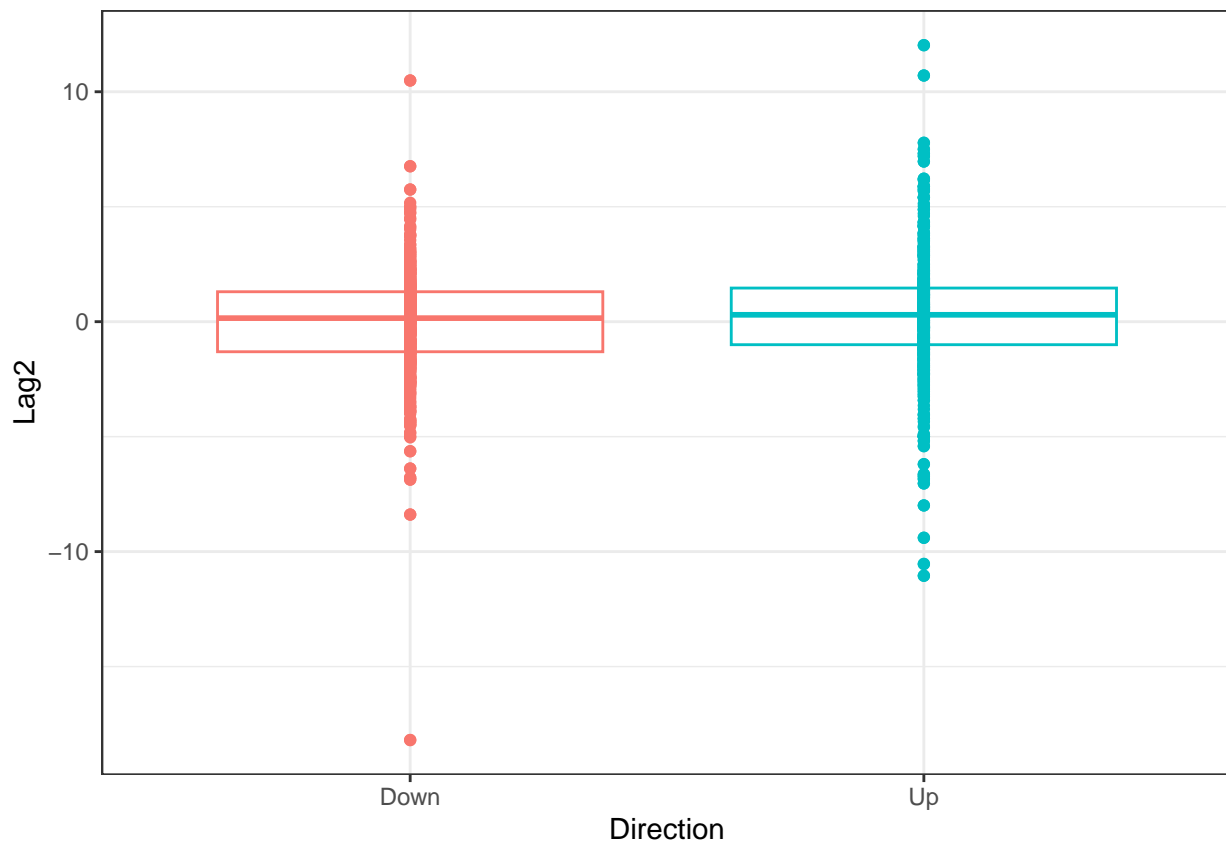
```
##      Up
## Down 0
## Up    1
```

```
confint(modelo_log_m, level = 0.95)
```

```
## Waiting for profiling to be done...  
##              2.5 %      97.5 %  
## (Intercept) -56.985558236 91.66680901  
## Year        -0.045809580  0.02869546  
## Lag1        -0.092972584  0.01093101  
## Lag2         0.007001418  0.11291264  
## Lag3        -0.068140141  0.03671410  
## Lag4        -0.079519582  0.02453326  
## Lag5        -0.066090145  0.03762099  
## Volume      -0.131576309  0.13884038
```

Variables significativas

```
ggplot(data = Weekly, mapping = aes(x = Direction, y = Lag2)) +  
  geom_boxplot(aes(color = Direction)) +  
  geom_point(aes(color = Direction)) +  
  theme_bw() +  
  theme(legend.position = "null")
```



Analisis

- Solo Lag2 es estadísticamente significativa con un valor p de 0.0275.
 - En la salida de `summary(modelo_log_m)` esta marcado con un *
- Year, Lag1, Lag3, Lag4, Lag5, y Volume no son significativas en el modelo.

- El intervalo de confianza del 95% para Lag2 es [0.007001418, 0.11291264], lo que sugiere que es un predictor significativo y positivo.
- Un aumento de una unidad en Lag2 aumenta las odds de que la dirección sea “Up” en aproximadamente 1.061 veces.

Divide la base de datos en un conjunto de entrenamiento (datos desde 1990 hasta 2008) y de prueba (2009 y 2010). Ajusta el modelo encontrado.

Division de la base de datos

```
datos_entrenamiento <- (Year < 2009)
datos_test <- Weekly[!datos_entrenamiento, ]
```

Formula el modelo logístico sólo con las variables significativas en la base de entrenamiento.

Ajuste del modelo con variables significativas

```
modelo_log_s <- glm(Direction ~ Lag2,
  data = Weekly,
  family = binomial, subset = datos_entrenamiento
)

summary(modelo_log_s)

##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
##      subset = datos_entrenamiento)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## 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: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

Representa gráficamente el modelo

```
# Vector con nuevos valores interpolados en el rango del predictor Lag2:
nuevos_puntos <- seq(
  from = min(Weekly$Lag2), to = max(Weekly$Lag2),
  by = 0.5
)
```

```

# Predicción de los nuevos puntos según el modelo con el comando predict()

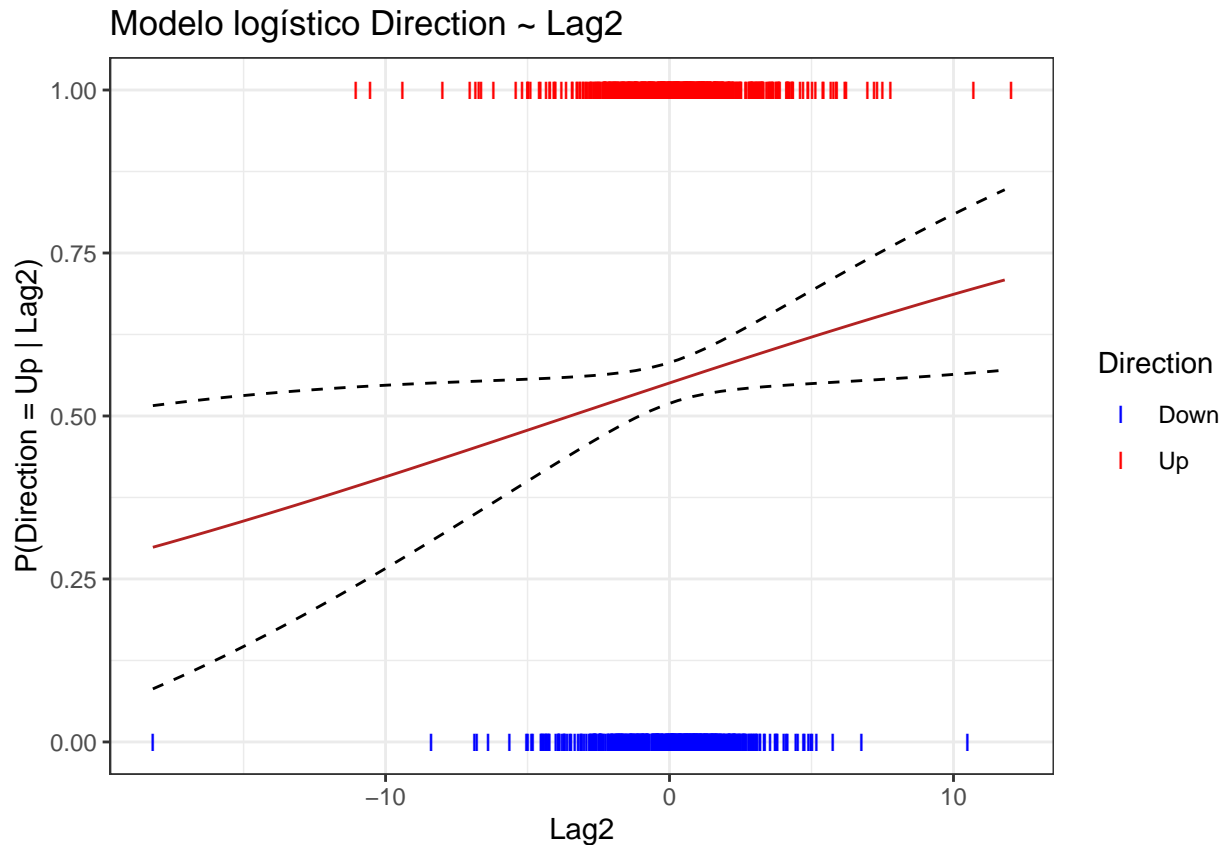
predicciones <- predict(modelo_log_s, newdata = data.frame(
  Lag2 =
    nuevos_puntos
), se.fit = TRUE, type = "response")

# Límites del intervalo de confianza (95%) de las predicciones
CI_inferior <- predicciones$fit - 1.96 * predicciones$se.fit
CI_superior <- predicciones$fit + 1.96 * predicciones$se.fit
# Matriz de datos con los nuevos puntos y sus predicciones
datos_curva <- data.frame(
  Lag2 = nuevos_puntos, probabilidad =
    predicciones$fit, CI.inferior = CI_inferior, CI.superior = CI_superior
)

# Codificación 0,1 de la variable respuesta Direction
Weekly$Direction <- ifelse(Weekly$Direction == "Down", yes = 0, no = 1)

ggplot(Weekly, aes(x = Lag2, y = Direction)) +
  geom_point(aes(color = as.factor(Direction)), shape = "I", size = 3) +
  geom_line(data = datos_curva, aes(y = probabilidad), color = "firebrick") +
  geom_line(data = datos_curva, aes(y = CI.superior), linetype = "dashed") +
  geom_line(data = datos_curva, aes(y = CI.inferior), linetype = "dashed") +
  labs(
    title = "Modelo logístico Direction ~ Lag2",
    y = "P(Direction = Up | Lag2)", x = "Lag2"
  ) +
  scale_color_manual(labels = c("Down", "Up"), values = c("blue", "red")) +
  guides(color = guide_legend("Direction")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw()

```



Evalúa el modelo con las pruebas de verificación correspondientes (Prueba de chi cuadrada, matriz de confusión)

```
# Usando Chi Cuadrada
anova(modelo_log_s, test = "Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Direction
##
## Terms added sequentially (first to last)
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                984      1354.7
## Lag2    1      4.1666    983      1350.5 0.04123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Usando Matriz de confusión

# Cálculo de la probabilidad predicha por el modelo con los datos de test
prob_modelo <- predict(modelo_log_s, newdata = datos_test, type = "response")
# Vector de elementos "Down"
pred_modelo <- rep("Down", length(prob_modelo))
```



```
# Sustitución de "Down" por "Up" si la p > 0.5
pred_modelo[prob_modelo > 0.5] <- "Up"
Direction_0910 <- Direction[!datos_entrenamiento]
# Matriz de confusión
matriz_confusion <- table(pred_modelo, Direction_0910)
matriz_confusion
```

```
##           Direction_0910
## pred_modelo Down Up
##           Down    9  5
##           Up     34 56
```

```
library(vcd)
```

```
## Loading required package: grid
```

```
##
```

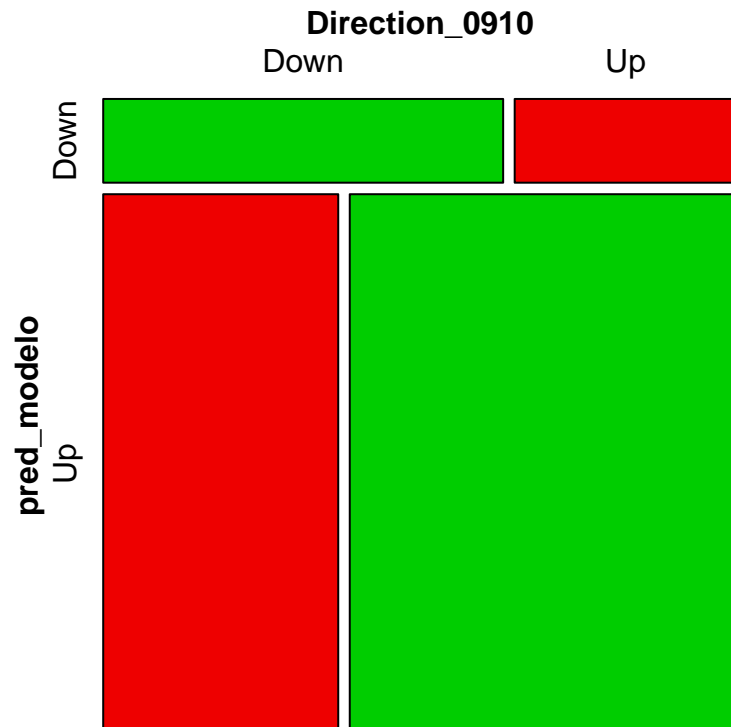
```
## Attaching package: 'vcd'
```

```
## The following object is masked from 'package:ISLR':
```

```
##
```

```
## Hitters
```

```
mosaic(matriz_confusion,
  shade = T, colorize = T,
  gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2))
)
```



```
mean(pred_modelo == Direction_0910)
```

```
## [1] 0.625
```

```
# Coeficientes del modelo
```

```
coeficientes <- coef(modelo_log_s)
```

Ecuación de modelo

```
cat(  
  "Ecuación del modelo: logit(P(Up)) =",  
  coeficientes[1], "+", coeficientes[2], "* Lag2", "\n"  
)
```

```
## Ecuación del modelo: logit(P(Up)) = 0.2032574 + 0.05809527 * Lag2
```

De igual forma podría ser escrito como: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{Lag2}$

En resumen del conjunto de datos se puede observar que entre el rendimiento de las semanas (Lag1 a Lag6) así como el año (Year) y el volumen (Volume) Lag2 es el que demuestra ser un valor significativo para las predicciones del mercado, en que dirección del se moverán acorde al rendimiento de Lag2.

Esto quiere decir que si Lag2 aumenta, la dirección del mercado tendrá la tendencia de estar hacia arriba (Up).

Este modelo puede ser de gran ayuda para en primera instancia saber cómo se comportará el S&P500, no obstante se recomienda no confiar del 100% del modelo ya que este puede no tener previsto otros factores que puedan afectar la bolsa, así como se recomienda hacer otros análisis y verificaciones.