A7-Regresión logística

Catherine Rojas

2024-11-05

Trabaja con el set de datos Weekly, que forma parte de la librería ISLR. Este set de datos contiene información sobre el rendimiento porcentual semanal del índice bursátil S&P 500 entre los años 1990 y 2010. **Se busca predecir el tendimiento (positivo o negativo)** dependiendo del comportamiento previo de diversas variables de la bolsa bursátil S&P 500.

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

Se cuenta con un set de datos con 9 variables (8 numéricas y 1 categórica que será nuestra variable respuesta: Direction). Las variables Lag son los valores de mercado en semanas anteriores y el valor del día actual (Today). La variable volumen (Volume) se refiere al volumen de acciones. Realiza:

1. El análisis de datos. Estadísticas descriptivas y coeficiente de correlación entre las variables.

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.3.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tibble' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Ver las primeras filas del conjunto de datos
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
```

```
# Ver una descripción rápida del conjunto de datos Weekly
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…
```

Observaciones * Aquí observamos que el conjunto de datos tiene 1,089 filas y 9 columnas. Direction es una variable categórica, mientras que las otras son numéricas. Esto ayuda a confirmar que Direction es la variable objetivo y que las demás pueden usarse como predictoras en el modelo.

```
# Estadísticas descriptivas
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
##
##
##
##
```

Observaciones * Observamos que los valores de Lag1 a Lag5 y Today tienen medias cercanas a cero, indicando una variabilidad alrededor del valor neutro. Volume muestra una tendencia creciente, con un rango amplio entre su mínimo (0.08747) y máximo (9.32821). Esto sugiere que el volumen de transacciones ha aumentado significativamente en el período de tiempo cubierto (1990-2010).

```
# Calcular la matriz de correlación para las variables
cor_matrix <- cor(Weekly[, -9])
cor_matrix
```

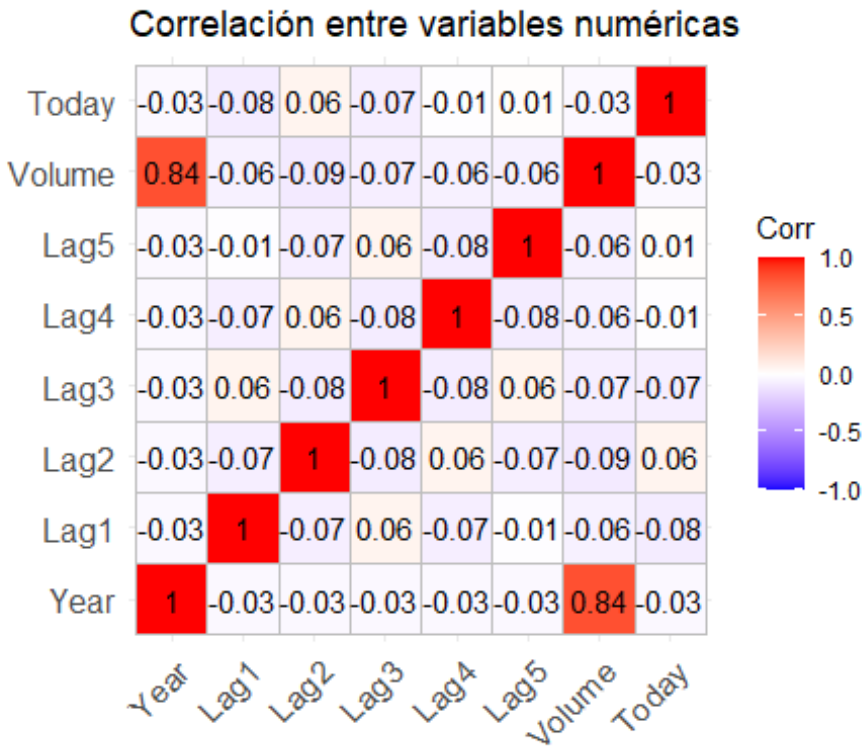
```
##              Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
```

```
## 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
```

```
library(ggcorrplot)
```

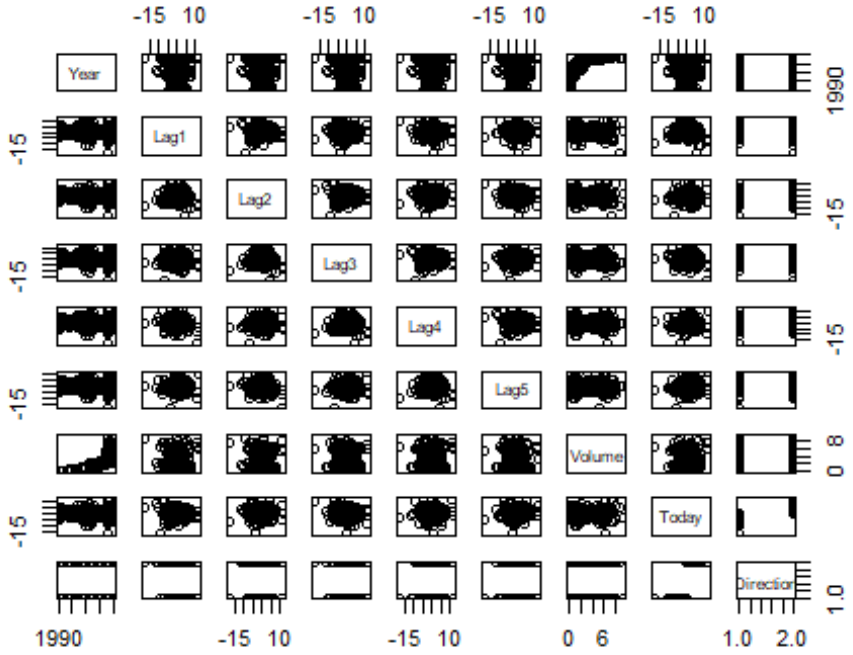
```
## Warning: package 'ggcorrplot' was built under R version 4.3.3
```

```
# Crear un gráfico de la matriz de correlación
ggcorrplot(cor_matrix, lab = TRUE, title = "Correlación entre variables numéricas")
```



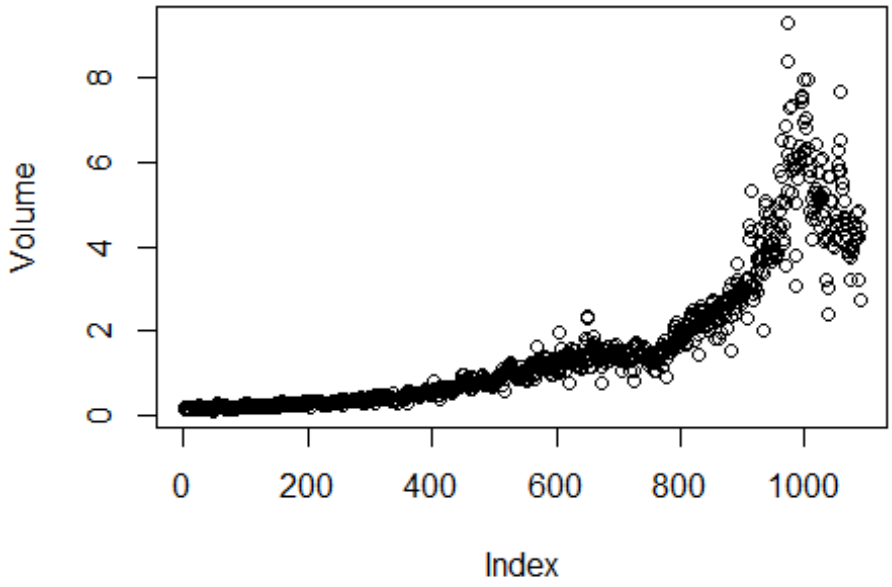
Observaciones * Observamos una fuerte correlación entre Year y Volume (0.84), lo que sugiere un aumento constante en el volumen a lo largo del tiempo. Las correlaciones entre los retrasos (Lag1 a Lag5) y Today son bajas, lo cual implica que el rendimiento de la semana actual no está fuertemente relacionado con los rendimientos de las semanas anteriores.

```r
# Crear una matriz de gráficos de dispersión entre todas las variables
pairs(Weekly)
```



Observaciones * Esta matriz muestra gráficos de dispersión entre todas las combinaciones de variables. La mayoría de los gráficos muestran una dispersión sin patrones lineales claros, excepto Year y Volume, donde se observa una tendencia creciente en Volume con el paso del tiempo. Esto refuerza la idea de que Volume ha crecido a lo largo de los años, mientras que los rendimientos pasados y actuales no presentan relaciones obvias.

```r
# Usar attach() para acceder directamente a las columnas del conjunto de datos
attach(Weekly)

# Crear un gráfico simple de la variable Volume
plot(Volume)
```



Observaciones * El gráfico de Volume muestra cómo ha evolucionado el volumen de transacciones a lo largo de las observaciones, donde existe un aumento constante en el volumen a lo largo del tiempo, con un crecimiento más pronunciado hacia el final del período. Esto puede ser un reflejo de cambios en el mercado financiero o en la popularidad de las inversiones durante el período cubierto (1990-2010).

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

```
# Modelo con todos los predictores, excluyendo "Today"
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
```

Observaciones * Lag2 es la única variable con un valor p menor a 0.05, lo cual indica que es estadísticamente significativa en el modelo (con un nivel de confianza del 95%). Esto sugiere que el rendimiento de dos semanas anteriores tiene un efecto en la probabilidad de que Direction sea Up.

- Las otras variables (Year, Lag1, Lag3, Lag4, Lag5, Volume) tienen valores p altos, lo cual sugiere que no son significativamente influyentes para predecir Direction.
- El coeficiente positivo para Lag2 (0.059449) indica que un rendimiento positivo en esta variable aumenta la probabilidad de que el rendimiento de la semana actual (Direction) sea Up.

```
# Codificación de la variable dependiente
contrasts(Weekly$Direction)
```

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

```
# Intervalos de confianza al 95% para los coeficientes
confint(object = 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
```

Observaciones * Aquellos intervalos con valores p menores a 0.05 son considerados estadísticamente significativos y se considera que influyen en el modelo.

- El intervalo de confianza de Lag2 no incluye el valor 0 (va de aproximadamente 0.007 a 0.113), lo que refuerza su significancia estadística. Para las otras variables, los intervalos de confianza incluyen el valor 0, lo cual confirma que no son significativamente diferentes de cero, y por lo tanto, no influyen en la variable objetivo.

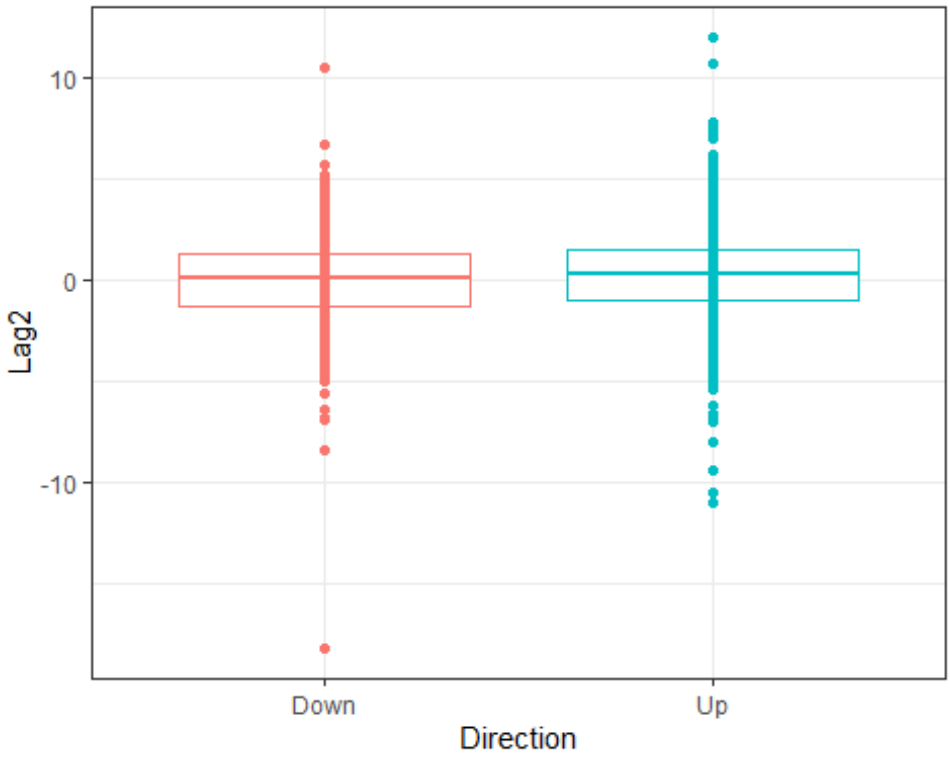
```
# Análisis del efecto de cada variable en los odds
# Para interpretar el efecto en los odds, exponenciamos los coeficientes:
exp(coef(modelo.log.m))
```

```
##  (Intercept)         Year         Lag1         Lag2         Lag3
Lag4
## 3.027468e+07 9.915361e-01 9.601291e-01 1.061251e+00 9.846412e-01
9.730534e-01
##         Lag5       Volume
## 9.860757e-01 1.003262e+00
```

Observaciones * exp(coef) para Lag2 es aproximadamente 1.061. Esto significa que por cada unidad adicional en Lag2, los odds de que Direction sea Up aumentan en un 6.1%.

- Los valores de exp(coef) para las demás variables están muy cerca de 1, lo cual indica que no tienen un efecto relevante en los odds de Direction = Up

```
# Gráfico de las variables significativas (boxplot):
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")
```



Observaciones * El boxplot muestra la distribución de Lag2 en función de Direction (Down vs. Up). Se observa que Lag2 tiene una mediana más alta cuando Direction es Up, lo cual apoya el hallazgo de que Lag2 es una variable significativa en el modelo. Esto sugiere que un rendimiento positivo en Lag2 está asociado con una mayor probabilidad de que el rendimiento de la semana actual (Direction) también sea positivo.

3. 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.

```
# Training: observaciones desde 1990 hasta 2008
datos.entrenamiento <- (Year < 2009)

# Test: observaciones de 2009 y 2010
datos.test <- Weekly[!datos.entrenamiento, ]

# Verifica:
sum(datos.entrenamiento) + nrow(datos.test) # nrow(datos.entrenamiento)
no funcionará ya que datos.entrenamiento es un vector
lógico, no un data.frame. En su lugar, se usa sum(datos.entrenamiento)
para                                                contar las
observaciones en el conjunto de entrenamiento.

## [1] 1089
```

- La suma de las observaciones en los conjuntos de entrenamiento y prueba es 1089, que es el total de filas en Weekly, lo que confirma que la división de datos se realizó correctamente.

```
# Ajuste del modelo logístico con variables significativas

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

Observaciones

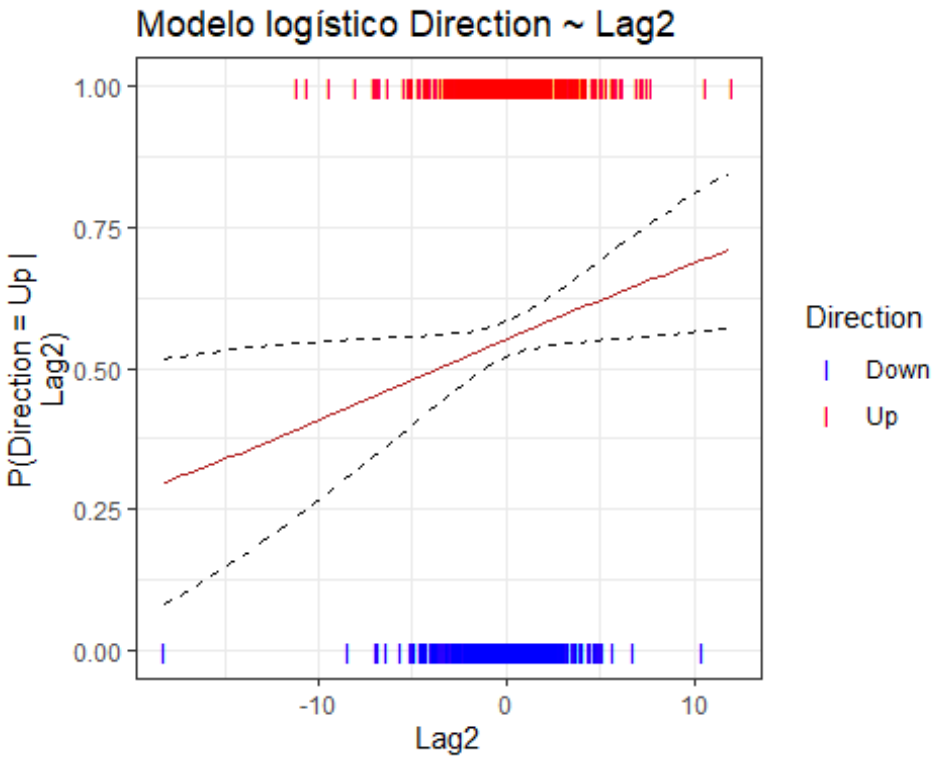
- Se ajusta un modelo de regresión logística usando Lag2 como predictor, ya que fue la única variable significativa encontrada en el análisis previo. El coeficiente estimado para Lag2 es 0.05810 con un valor p de 0.04298, lo cual indica que Lag2 es estadísticamente significativo al nivel de 5%. Esto significa que Lag2 tiene un efecto significativo en la probabilidad de que Direction sea Up.

```
# Predicciones en el conjunto de prueba
predictions <- predict(modelo.log.s, datos.test, type = "response")

# Convertir probabilidades a clases (Up o Down) usando un umbral de 0.5
predicted_class <- ifelse(predictions > 0.5, "Up", "Down")

# Calcular la precisión en el conjunto de prueba
actual_class <- datos.test$Direction
accuracy <- mean(predicted_class == actual_class)
accuracy

## [1] 0.625
```

- Estas predicciones son probabilidades de que Direction sea Up.
- Las probabilidades se convierten en predicciones de clase (Up o Down) utilizando un umbral de 0.5, donde la precisión se calcula como el porcentaje de observaciones en el conjunto de prueba que fueron clasificadas correctamente. La precisión obtenida es de 0.625 (62.5%).

5. 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() se calcula la probabilidad de que la variable respuesta
pertenezca al nivel de referencia (en este caso "Up")
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()
```



Interpretación * Este gráfico muestra que Lag2 tiene un efecto positivo en la probabilidad de que Direction sea "Up". Sin embargo, el modelo logístico basado únicamente en Lag2 no logra capturar toda la variabilidad de Direction, especialmente en valores cercanos a 0. Esto sugiere que Lag2 es un predictor moderadamente útil, pero pueden ser necesarias más variables para mejorar la precisión del modelo en la predicción de la dirección del rendimiento semanal.

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

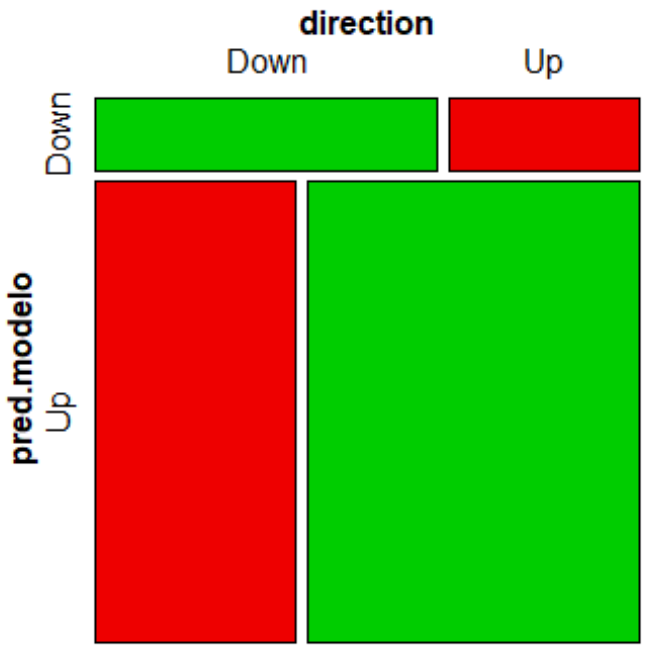
```
# Chi cuadrada: Se evalúa la significancia del modelo con predictores con respecto al modelo_nulo ("Residual deviance" vs "Null deviance").
anova(modelo.log.s, test = "Chisq") # Si valor p es menor que alfa será significativo.
```

```
## 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
```

- Como el valor p es menor que 0.05, concluimos que el modelo con Lag2 es significativamente mejor que el modelo nulo. Esto indica que Lag2 aporta información útil para predecir la dirección (Direction) del rendimiento del mercado.

```
# Cálculo de las predicciones correctas así como de los falsos negativos
y positivos.
# Normalmente se usa un límite de 0.5.


# 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 = Direction[!datos.entrenamiento]

library(dplyr)
library(vcd)

## Warning: package 'vcd' was built under R version 4.3.3

## Loading required package: grid

##
## Attaching package: 'vcd'

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

# Matriz de confusión
matriz.confusion <- table(pred.modelo, direction)
matriz.confusion

##            direction
## pred.modelo Down Up
##        Down    9  5
##        Up     34 56

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




```
mean(pred.modelo == direction)

## [1] 0.625
```

Interpretación

- Verdaderos Positivos (Up predicho y Up real): 56 casos fueron correctamente clasificados como “Up”.
- Verdaderos Negativos (Down predicho y Down real): 9 casos fueron correctamente clasificados como “Down”.
- Falsos Positivos (Up predicho pero Down real): 34 casos fueron incorrectamente clasificados como “Up” cuando realmente eran “Down”.
- Falsos Negativos (Down predicho pero Up real): 5 casos fueron incorrectamente clasificados como “Down” cuando realmente eran “Up”.
- La precisión del modelo es de 0.625, o 62.5%. Esto significa que el modelo predice correctamente el 62.5% de los casos en el conjunto de prueba.
- El gráfico de mosaico proporciona una visualización de la matriz de confusión, donde:
- Las celdas verdes representan las clasificaciones correctas (verdaderos positivos y verdaderos negativos).

- Las celdas rojas representan las clasificaciones incorrectas (falsos positivos y falsos negativos).
- Dado que el modelo tiene un número considerable de falsos positivos, es posible que se beneficie de incluir más predictores o probar modelos más complejos para capturar mejor la variabilidad en la dirección del mercado.

7. Escribe (ecuación), grafica el modelo significativo e interprétalo en el contexto del problema. Añade posibles errores e indica si es buen modelo, en qué no lo es, cuánto cambia)

Escribir la Ecuación del Modelo Significativo

- Dado que el modelo logístico incluye solo `Lag2` como predictor significativo para la variable de respuesta `Direction`, la ecuación del modelo se expresa como:

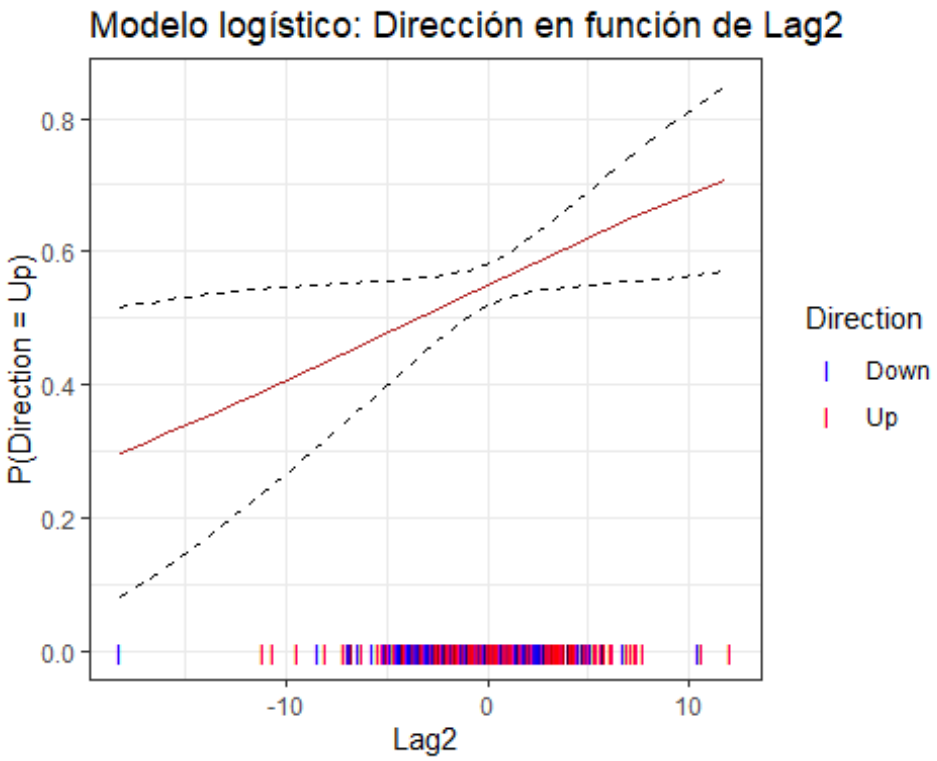
$$\log\left(\frac{P(\text{Direction} = \text{Up})}{1 - P(\text{Direction} = \text{Up})}\right) = \beta_0 + \beta_1 \times \text{Lag2}$$

- Donde:
  - $\beta_0 = 0.20326$ es el intercepto.
  - $\beta_1 = 0.05810$ es el coeficiente de `Lag2`.
- Esta ecuación se puede reescribir en términos de probabilidad de `Direction = Up`:

$$P(\text{Direction} = \text{Up}) = \frac{1}{1 + e^{-(0.20326 + 0.05810 \times \text{Lag2})}}$$

```
# 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
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)

# Graficar el modelo
```

```r
library(ggplot2)
ggplot(Weekly, aes(x = Lag2, y = as.numeric(Direction == "Up"))) +
  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: Dirección en función de Lag2", y =
"P(Direction = Up)", 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()
```




```r
# Cálculo de la probabilidad predicha por el modelo con los datos de
prueba
prob.modelo <- predict(modelo.log.s, newdata = datos.test, type =
"response")

# Vector de predicciones inicializado en "Down"
pred.modelo <- rep("Down", length(prob.modelo))

# Cambia a "Up" donde la probabilidad predicha es mayor a 0.5
pred.modelo[prob.modelo > 0.5] <- "Up"
```

```r
# Extrae la variable Direction del conjunto de prueba para compararla con las predicciones
direction <- datos.test$Direction

# Genera la matriz de confusión
matriz.confusion <- table(Predicted = pred.modelo, Actual = direction)
print(matriz.confusion)
```

```
##          Actual
## Predicted Down Up
##      Down    9  5
##      Up     34 56
```

- Recordando el objetivo de este análisis, predecir la dirección semanal (Direction) del índice bursátil S&P 500, es decir, si el rendimiento será positivo ("Up") o negativo ("Down"), utilizando como predictor el rendimiento de dos semanas previas (Lag2), se concluye lo siguiente:

**Relación entre Lag2 y Direction:*

- El modelo logístico sugiere que existe una relación positiva entre Lag2 y la probabilidad de que el rendimiento sea "Up". Esto significa que, cuando el rendimiento de dos semanas previas (Lag2) es alto, aumenta la probabilidad de que el rendimiento en la semana actual también sea positivo. Sin embargo, la magnitud del efecto es pequeña (coeficiente de Lag2 = 0.05810), lo que indica que esta variable por sí sola no es un fuerte predictor de Direction.

Predicción Limitada:

- El modelo tiene una precisión de 62.5% en el conjunto de prueba. Aunque esto es mejor que una predicción aleatoria, no es lo suficientemente alto para aplicaciones en las que se requiere una predicción confiable, como en la toma de decisiones de inversión. La predicción del modelo tiende a ser más precisa para identificar la clase "Up" que la clase "Down".

Posibles Errores del Modelo según la Matriz de Confusión

- **Falsos Positivos (34 casos):** El modelo predijo "Up" cuando en realidad fue "Down", lo que podría inducir a decisiones incorrectas en finanzas, como mantener o comprar activos ante una falsa expectativa de subida. Este es el error más frecuente y costoso.
- **Falsos Negativos (5 casos):** El modelo predijo "Down" cuando en realidad fue "Up", lo que podría llevar a una venta prematura o a perder oportunidades de inversión en semanas positivas.
- **Verdaderos Positivos y Negativos:** El modelo identificó correctamente 56 semanas con rendimiento "Up" y solo 9 con rendimiento "Down", lo que muestra dificultades para reconocer períodos de rendimiento negativo.

Evaluación del Modelo

- Los 34 falsos positivos sugieren que el modelo predice “Up” incorrectamente con frecuencia, lo que puede inducir a errores de inversión.
- El modelo esta excesivamente simplificado al solo usar Lag2, ignorando otros factores relevantes que podrían mejorar la predicción.

Recomendaciones:

- Incorporar otras variables del mercado podría mejorar la precisión.
- Usar modelos como árboles de decisión o redes neuronales para capturar relaciones no lineales.
- Aplicar Lasso para seleccionar predictores relevantes y evitar el sobreajuste.

Conclusion

- En el contexto de predicción del rendimiento semanal del S&P 500, este modelo basado únicamente en Lag2 ofrece una precisión limitada y sufre de una alta tasa de falsos positivos. Aunque el modelo proporciona información preliminar sobre la relación entre el rendimiento de dos semanas anteriores y la dirección actual, no es lo suficientemente preciso para aplicaciones prácticas en la toma de decisiones de inversión.
- No es un buen modelo para predecir la dirección del mercado en escenarios reales debido a su simplicidad y precisión limitada.
- El modelo debería considerar múltiples factores y utilizar técnicas de modelado más avanzadas para mejorar su precisión y reducir los errores de clasificación.
- El efecto de Lag2 en la probabilidad de “Up” es leve, lo que indica que Lag2 por sí solo no captura una relación fuerte con Direction.