A7-Regresión logística

Luis Maximiliano López Ramírez

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 4.3.2
```

```
library(tidyverse)
```

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

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## Warning: package 'purrr' was built under R version 4.3.2
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

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

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
```

```
## ✓ 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
```

```
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
```

```
glimpse(Weekly)
```

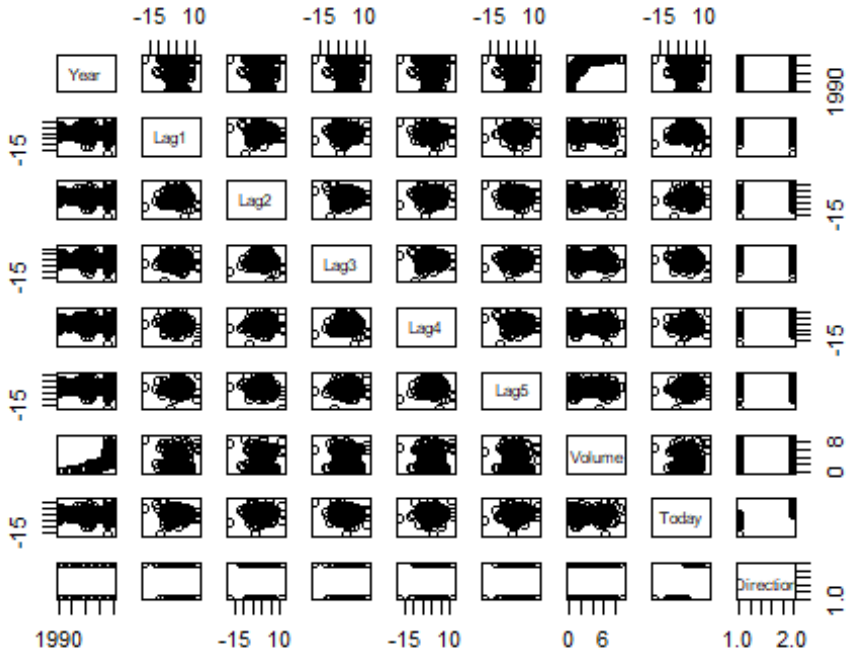
```
## Rows: 1,089
## Columns: 9
## $ Year      <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990,
1990, 1990, …
## $ Lag1      <dbl> 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372,
0.807, 0…
## $ Lag2      <dbl> 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -
1.372, 0…
## $ Lag3      <dbl> -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…
```

```
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
##
##
##
##
```

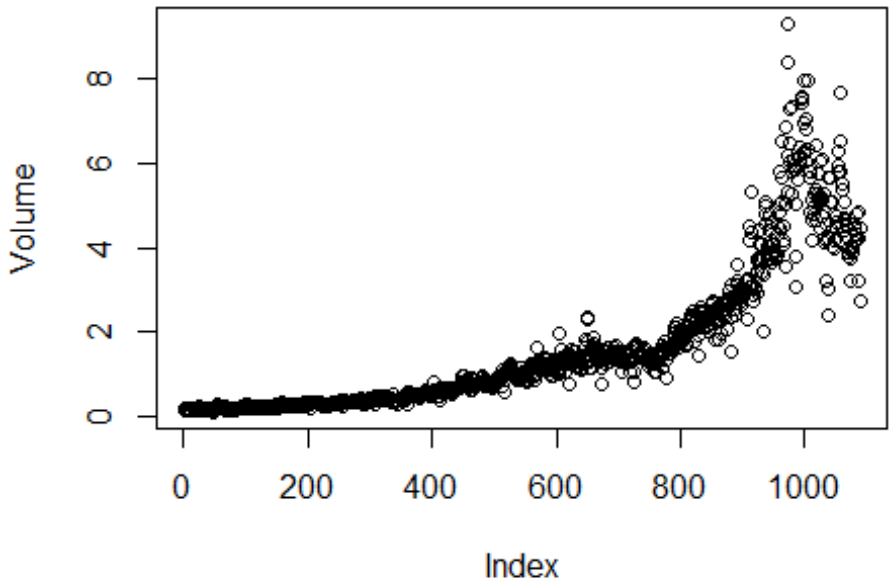
```
pairs(Weekly)
```




```
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

attach(Weekly)
plot(Volume)
```



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

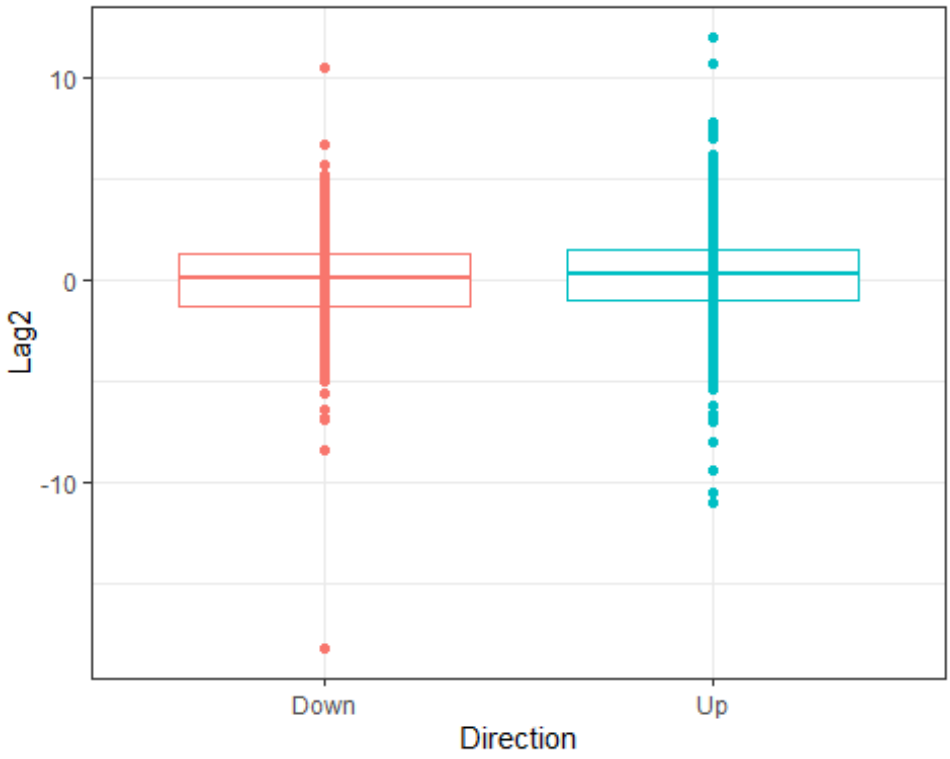
```
##      Up
## Down  0
## Up    1
```

```
confint(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
```

```
# Gráfico de las variables significativas (boxplot), ejemplo: Lag2):
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")
```



Variables que influyen en el modelo

Lag2: Es la única variable con un valor p menor a 0.05 (p = 0.0275). Esto indica que su influencia en el modelo es estadísticamente significativa, sugiriendo que el valor de Lag2 afecta los odds (momios) de Direction. Su coeficiente positivo (0.0594) indica que un incremento en Lag2 aumenta la probabilidad de que Direction sea "Up" (hacia arriba).

Variables que no influyen en el modelo Las siguientes variables no son estadísticamente significativas (p > 0.05):

Intercepto: p = 0.6494 Year: p = 0.6545 Lag1: p = 0.1239 Lag3: p = 0.5622 Lag4: p = 0.3024 Lag5: p = 0.5955 Volume: p = 0.9623 Estas variables tienen valores p mayores a 0.05, lo que sugiere que no contribuyen significativamente al modelo para predecir Direction.

Interpretación del efecto en los odds

Lag2: Con un coeficiente de 0.0594, cada incremento de una unidad en Lag2 incrementa los odds de Direction siendo "Up" en un factor de e^0.0594≈1.061. Esto implica un aumento moderado en la probabilidad de que Direction sea "Up" con mayores valores de Lag2.

Gráfica

La gráfica muestra que Lag2 tiene una distribución similar en ambas categorías de Direction, aunque con una ligera diferencia en la mediana y con varios valores

atípicos. Esta ligera diferencia en la mediana puede explicar por qué Lag2 es un predictor estadísticamente significativo en el modelo de regresión logística.

 # 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)

## [1] 1089

# Modelo utilizando solo la variable significativa Lag2
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
```

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

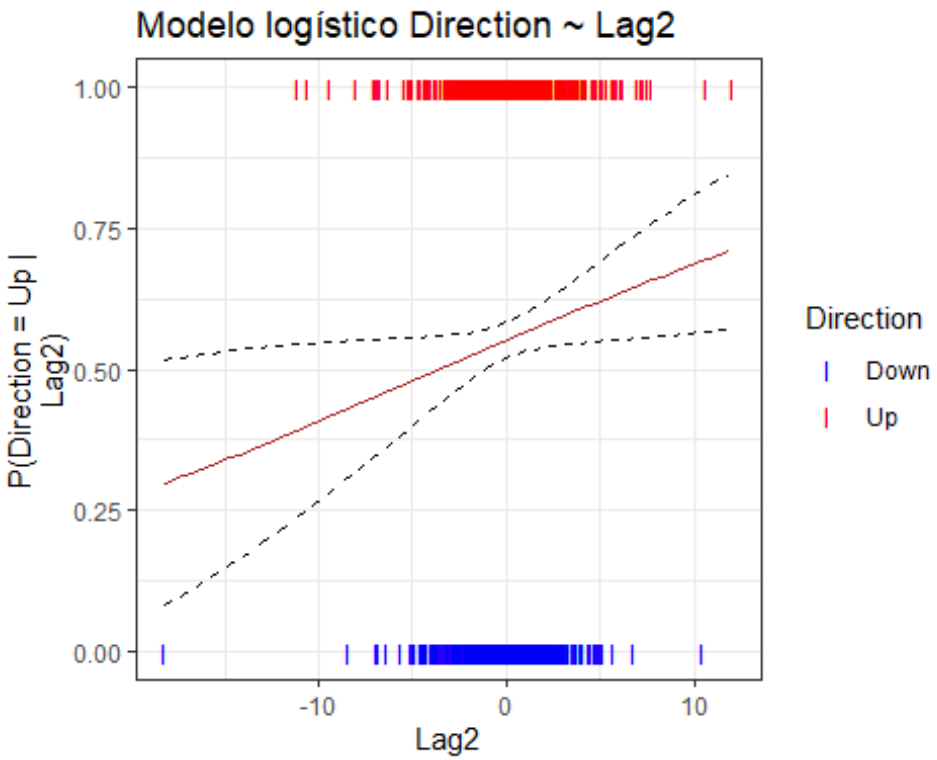
```
# Modelo utilizando solo la variable significativa Lag2
modelo.log.s <- glm(Direction ~ Lag2, data = Weekly,
                    family = binomial, subset = datos.entrenamiento)
summary(modelo.log.s)
```

```
##
## 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
```

 # 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()
```



Puntos de datos:

Cada punto representa una observación en los datos de Weekly, donde el eje x muestra los valores de Lag2 y el eje y muestra la variable binaria Direction, codificada como 0 para “Down” y 1 para “Up”.

Los puntos están coloreados según el valor de Direction: azul para “Down” y rojo para “Up”.

Curva de probabilidad (línea roja):

La línea roja representa la probabilidad predicha de que Direction sea “Up” (1) para diferentes valores de Lag2.

La pendiente de la línea indica que, a medida que Lag2 aumenta, la probabilidad de que Direction sea “Up” también aumenta. Esto es consistente con el signo positivo del coeficiente de Lag2 en el modelo, que indica una relación positiva entre Lag2 y la probabilidad de “Up”.

Intervalos de confianza (líneas punteadas):

Las líneas punteadas representan el intervalo de confianza del 95% para la probabilidad predicha. Estas líneas muestran el rango dentro del cual es probable que se encuentre la verdadera probabilidad para un valor dado de Lag2.

A medida que Lag2 se aleja de cero (en ambas direcciones), el intervalo de confianza se ensancha ligeramente, indicando mayor incertidumbre en la predicción para valores extremos de Lag2.

Interpretación del modelo:

La gráfica sugiere que Lag2 tiene un efecto positivo en la probabilidad de que Direction sea "Up": a medida que Lag2 aumenta, también aumenta la probabilidad de un cambio positivo en Direction.

La transición de "Down" a "Up" no es brusca, sino gradual, reflejando el carácter probabilístico de la regresión logística.

Conclusión

La gráfica confirma la significancia de Lag2 en el modelo y visualiza su efecto positivo sobre la probabilidad de Direction = Up. El intervalo de confianza sugiere que el modelo es más confiable cerca del centro de la distribución de Lag2 y menos confiable en los extremos.

 # 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"). Si valor p es menor que alfa será significativo.
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
```

```
# 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"
```

```r
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
```

```r
library(vcd)
```

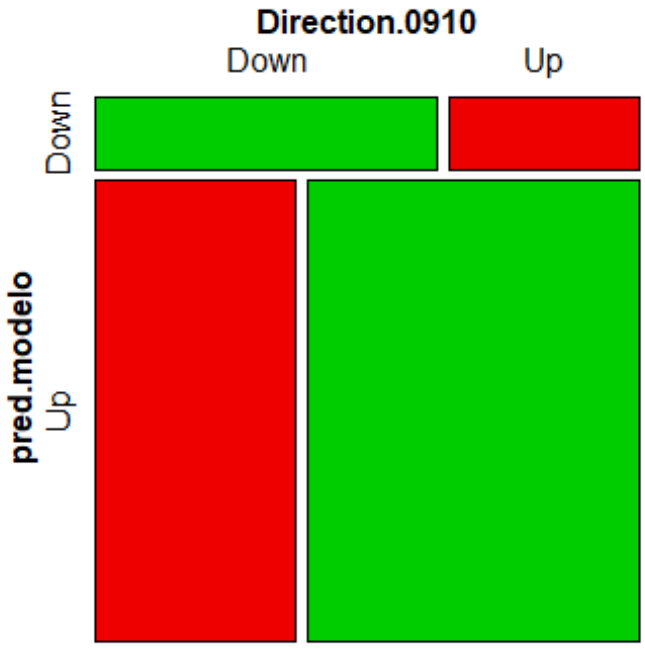
```
## Warning: package 'vcd' was built under R version 4.3.3
```

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

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

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

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



```r
mean(pred.modelo == Direction.0910)
```

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

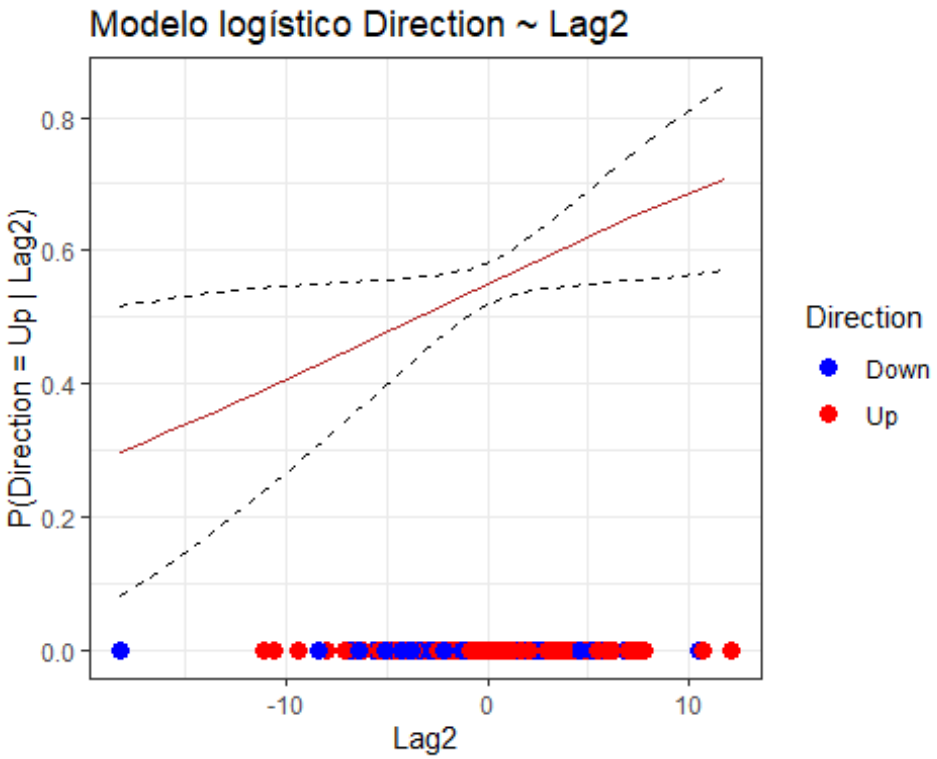
 # 7. Escribe (ecuación), grafica el modelo significativo e interprétalo en el contexto del problema. Añade posibles errores, si es buen modelo, en qué no lo es, cuánto cambia)

```r
# 1. Crear el modelo logístico con Lag2 como predictor
modelo.log.s <- glm(Direction ~ Lag2, family = binomial, data = Weekly,
subset = datos.entrenamiento)

# 2. Imprimir la ecuación del modelo
intercepto <- coef(modelo.log.s)[1]
coef_lag2 <- coef(modelo.log.s)[2]
cat("Ecuación del modelo: log(p / (1 - p)) =", round(intercepto, 4), "+",
round(coef_lag2, 4), "* Lag2\n")
```

```
## Ecuación del modelo: log(p / (1 - p)) = 0.2033 + 0.0581 * Lag2
```

```r
# 3. Generar nuevos puntos para la gráfica de predicción
nuevos_puntos <- seq(from = min(Weekly$Lag2), to = max(Weekly$Lag2), by =
0.5)

# 4. Predicción de probabilidades y sus intervalos de confianza
predicciones <- predict(modelo.log.s, newdata = data.frame(Lag2 =
nuevos_puntos), se.fit = TRUE, type = "response")
CI_inferior <- predicciones$fit - 1.96 * predicciones$se.fit
CI_superior <- predicciones$fit + 1.96 * predicciones$se.fit

# Crear un DataFrame para la curva de predicción y sus intervalos
datos_curva <- data.frame(Lag2 = nuevos_puntos,
                          probabilidad = predicciones$fit,
                          CI.inferior = CI_inferior,
                          CI.superior = CI_superior)

# 5. Graficar la curva de probabilidad y los intervalos de confianza
library(ggplot2)
grafica_modelo <- ggplot(Weekly, aes(x = Lag2, y = as.numeric(Direction
== "Up"))) +
  geom_point(aes(color = as.factor(Direction)), shape = 16, 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"))
```

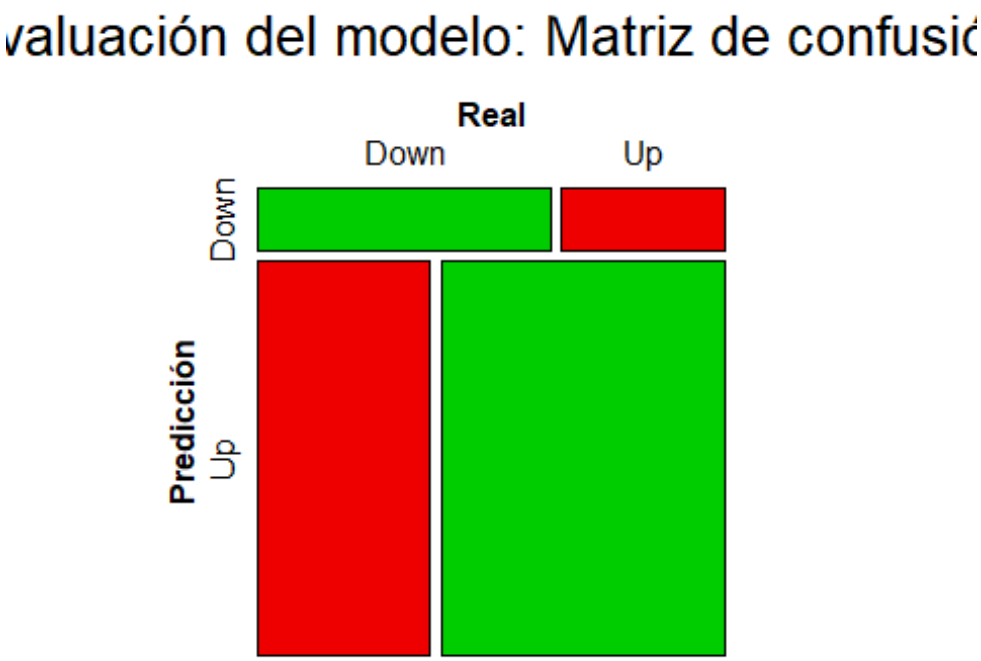
```
+
  guides(color=guide_legend("Direction")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw()

print(grafica_modelo)
```




```
# 6. Evaluación del modelo: Matriz de confusión y precisión
# Cálculo de la probabilidad predicha para los datos de prueba
prob.modelo <- predict(modelo.log.s, newdata = datos.test, type =
"response")

# Clasificación: "Up" si la probabilidad > 0.5, "Down" en caso contrario
pred.modelo <- rep("Down", length(prob.modelo))
pred.modelo[prob.modelo > 0.5] <- "Up"
Direction.0910 <- datos.test$Direction

# Matriz de confusión
matriz.confusion <- table(pred.modelo, Direction.0910)
print(matriz.confusion)
```

```
##             Direction.0910
## pred.modelo Down Up
##         Down    9  5
##         Up     34 56
```

```r
# Calcular precisión
precision <- mean(pred.modelo == Direction.0910)
cat("Precisión del modelo en datos de prueba:", round(precision, 4),
"\n")

## Precisión del modelo en datos de prueba: 0.625

# 7. Graficar la matriz de confusión en un gráfico de mosaico
library(vcd)
mosaic(matriz.confusion, shade = TRUE, colorize = TRUE,
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2,
2)),
       main = "Evaluación del modelo: Matriz de confusión",
       labeling_args = list(set_varnames = c(pred.modelo = "Predicción",
Direction.0910 = "Real")))
```



Ecuación del modelo

Ecuación del modelo: log(p / (1 - p)) = 0.2033 + 0.0581 * Lag2

Interpretación matriz de confusión

La matriz de confusión se organiza de la siguiente manera:

Eje vertical (pred.modelo): Predicciones del modelo. Eje horizontal (Direction.0910): Valores reales.

Cada celda contiene el número de observaciones para cada combinación de predicción y valor real:

Celdas verdes representan las predicciones correctas: Down - Down (arriba a la izquierda): 9 observaciones correctamente predichas como “Down”. Up - Up (abajo a la derecha): 56 observaciones correctamente predichas como “Up”.

Celdas rojas representan las predicciones incorrectas: Down - Up (arriba a la derecha): 5 observaciones predichas como “Down” cuando eran “Up”. Up - Down (abajo a la izquierda): 34 observaciones predichas como “Up” cuando eran “Down”.

Precisión del modelo La precisión del modelo se calcula como el número de predicciones correctas dividido por el total de predicciones. Esto significa que el modelo tiene una precisión del 62.5% en los datos de prueba, indicando que el modelo predice correctamente la dirección en el 62.5% de las observaciones de 2009 y 2010.

Prueba de Chi cuadrada

Dado que el valor p (0.04123) es menor que 0.05, concluimos que el modelo con Lag2 es significativamente mejor que el modelo nulo. Esto respalda el uso de Lag2 como predictor significativo para Direction.

Posibles errores, si es un buen modelo y en qué no lo es

El modelo logístico con Lag2 como único predictor tiene algunas ventajas, como su simplicidad y la significancia estadística de Lag2, que demuestra tener un impacto real en la variable de respuesta Direction. La reducción en la devianza al incluir Lag2 indica que el modelo es mejor que uno sin predictores. Sin embargo, este modelo presenta varias limitaciones.

En términos de rendimiento, su precisión en los datos de prueba es moderada, con un 62.5%, lo cual indica que es propenso a cometer errores de clasificación. Este nivel de precisión sugiere que, aunque Lag2 es un predictor significativo, el modelo no es lo suficientemente robusto para predecir Direction de manera confiable, lo que lo hace limitado en un contexto en el que se necesita una predicción más precisa.

Además, al depender solo de Lag2, el modelo podría estar perdiendo otros factores importantes que influyen en Direction, como Volume o Year. Esto limita la complejidad del modelo y hace que sea incompleto para capturar toda la variabilidad en la dirección del mercado.

Para mejorar el modelo, se pueden considerar las siguientes estrategias:

Agregar más predictores: Incluir variables adicionales podría ayudar a capturar más factores que influyen en Direction, aumentando así la precisión del modelo.

Probar modelos no lineales: La regresión logística asume una relación lineal en los log-odds, lo cual podría no ser adecuado en un contexto financiero. Modelos como árboles de decisión o redes neuronales pueden captar relaciones no lineales y mejorar la predicción.

Ingeniería de características: Transformar Lag2 o crear variables derivadas de esta puede mejorar el rendimiento del modelo sin agregar demasiada complejidad.

Conclusión

El modelo logístico que usa Lag2 como predictor tiene una precisión del 62.5% en los datos de prueba y es significativamente mejor que un modelo sin predictores. Aunque tiene una tasa de error considerable (particularmente en predicciones de “Up” para observaciones que eran “Down”), el modelo muestra que Lag2 tiene un efecto estadísticamente significativo en la predicción de Direction. El modelo actual es un buen comienzo debido a su simplicidad y la significancia de Lag2, pero tiene una precisión moderada y carece de complejidad. Con ajustes adicionales, como agregar más predictores y probar modelos no lineales, se podría mejorar su capacidad predictiva.