A7-Regresión logística

Adrian Pineda Sanchez

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.

```
# Cargar librerías necesarias
library(ISLR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
# Cargar el set de datos Weekly
data("Weekly")

# Ver las primeras filas del set 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
```

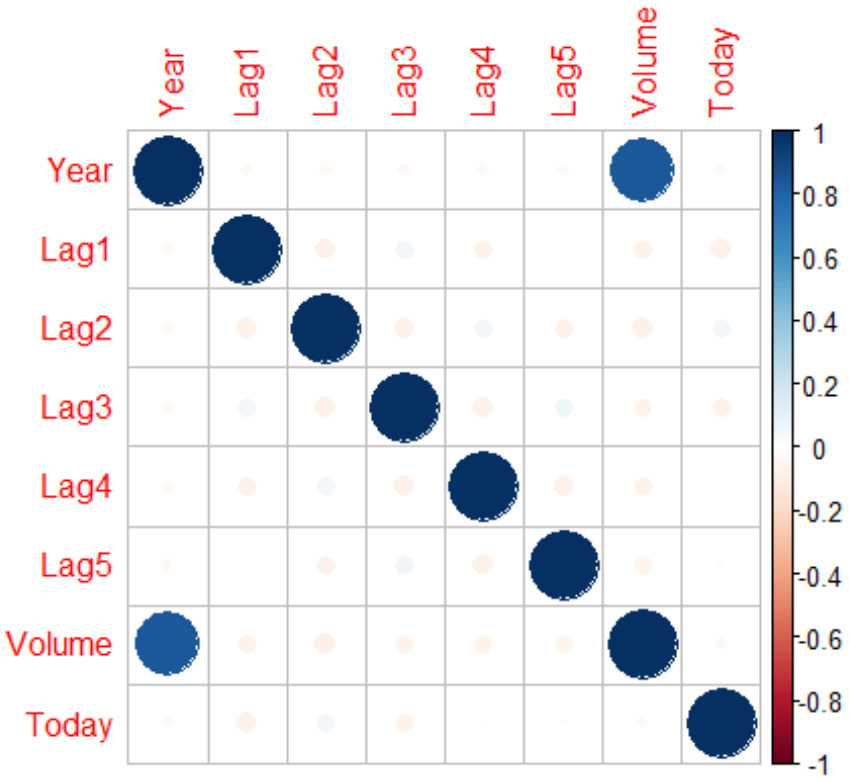
```
# Calcular estadísticas descriptivas
summary(Weekly)  # Resumen estadístico de todas las variables
```

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

```
# Calcular la matriz de correlación excluyendo la variable 'Direction'
correlation_matrix <- cor(Weekly %>% select(-Direction))
print(correlation_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

# Visualización de la matriz de correlación
corrplot(correlation_matrix, method = "circle")
```



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).

Para todos los coeficientes de las variables:

$$H_0: B_i = 0$$

$$H_1: B_i \neq 0$$

```
# Ajustar el modelo logístico con todas las variables menos 'Today'
model_full <- glm(Direction ~ Year + Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
Volume, data = Weekly, family = binomial)

# Resumen del modelo, incluyendo coeficientes y significancia
summary(model_full)
```

```
##
## Call:
## glm(formula = Direction ~ Year + Lag1 + Lag2 + Lag3 + Lag4 +
##     Lag5 + Volume, 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

# Calcular intervalos de confianza para los coeficientes
conf_intervals <- confint(model_full)

## Waiting for profiling to be done...

print(conf_intervals)

##                     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

# Interpretación de los efectos en los odds (momios)
odds_ratios <- exp(coef(model_full))  # Exponenciar los coeficientes
odds_ratios

##  (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
```

Descripción del Modelo

Este modelo de regresión logística predice la dirección (`Direction`) en función de varias variables predictoras: `Year`, `Lag1`, `Lag2`, `Lag3`, `Lag4`, `Lag5` y `Volume`. La variable dependiente es binaria, y el modelo se ajusta utilizando una familia binomial. El análisis se ha realizado sobre el conjunto de datos `Weekly`.

Interpretación de los Coeficientes

Cada coeficiente representa el efecto estimado de la variable independiente correspondiente sobre la probabilidad de un cambio en la dirección. A continuación, se detallan las principales observaciones:

- **Intercepto**: El coeficiente del intercepto es 17.2258. Aunque positivo, su valor p es de 0.6494, lo que indica que no es estadísticamente significativo.
- **Year**: La variable `Year` tiene un coeficiente de -0.0085 con un valor p de 0.6545, indicando que no tiene un efecto significativo en el modelo a nivel del 5%.
- **Lag1, Lag3, Lag4, Lag5 y Volume**: Estas variables tienen coeficientes con valores p superiores a 0.05, lo que sugiere que no son estadísticamente significativas en el modelo. Esto implica que su influencia en `Direction` no es concluyente en este análisis.
- **Lag2**: La variable `Lag2` tiene un coeficiente de 0.0595 y un valor p de 0.0275, lo cual la hace la única variable con significancia estadística en el modelo (valor $p < 0.05$). Este resultado sugiere que `Lag2` tiene un impacto significativo en la dirección, y que su aumento está asociado a un cambio positivo en la probabilidad de `Direction`.

Métricas del Modelo

- **Deviance Nula**: La deviance nula es 1496.2 con 1088 grados de libertad, que representa la variabilidad total en el modelo sin considerar ninguna variable independiente.
- **Deviance Residual**: La deviance residual es 1486.2 con 1081 grados de libertad, que representa la variabilidad restante después de ajustar el modelo con las variables independientes. La diferencia entre ambas sugiere que el modelo ha capturado una pequeña porción de la variabilidad total.
- **AIC**: El criterio de información de Akaike (AIC) es de 1502.2, una métrica que evalúa la calidad del modelo considerando tanto su ajuste como su complejidad. Un AIC más bajo generalmente indica un mejor equilibrio entre ajuste y parsimoniocidad del modelo.

Conclusión

En conclusión, el análisis indica que la variable `Lag2` es la única predictora con un efecto estadísticamente significativo en la dirección (`Direction`). Las otras variables (`Year`, `Lag1`, `Lag3`, `Lag4`, `Lag5` y `Volume`) no presentan evidencia suficiente para ser consideradas significativas en este modelo. Esto sugiere que `Lag2` podría ser una variable clave para predecir la dirección, mientras que las demás podrían no ser relevantes y podrían ser candidatas para eliminación en futuros ajustes o simplificaciones del modelo.

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.

```
# Dividir el set de datos en entrenamiento y prueba
train <- Weekly %>% filter(Year < 2009)
test <- Weekly %>% filter(Year >= 2009)

# Ajustar el modelo en el conjunto de entrenamiento
model_train <- glm(Direction ~ Lag2, data = train, family = binomial)

# Predicciones en el conjunto de prueba
predictions <- predict(model_train, newdata = test, type = "response")
pred_class <- ifelse(predictions > 0.5, "Up", "Down")  # Clasificar según
probabilidad

head(train)
```

```
##   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
```

```
head(test)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5   Volume  Today Direction
## 1 2009  6.760 -1.698  0.926  0.418 -2.251 3.793110 -4.448      Down
## 2 2009 -4.448  6.760 -1.698  0.926  0.418 5.043904 -4.518      Down
## 3 2009 -4.518 -4.448  6.760 -1.698  0.926 5.948758 -2.137      Down
## 4 2009 -2.137 -4.518 -4.448  6.760 -1.698 6.129763 -0.730      Down
## 5 2009 -0.730 -2.137 -4.518 -4.448  6.760 5.602004  5.173        Up
## 6 2009  5.173 -0.730 -2.137 -4.518 -4.448 6.217632 -4.808      Down
```

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

```
# Ajustar el modelo solo con las variables significativas
model_significant <- glm(Direction ~ Lag2, data = train, family = binomial)
summary(model_significant)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = train)
##
## 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
```

Descripción del Modelo

Se ha ajustado un modelo de regresión logística utilizando la variable `Lag2` como predictor de la variable dependiente `Direction`. Este modelo busca predecir la dirección (ya sea "Up" o "Down") basándose en el valor de `Lag2`. Los datos utilizados para ajustar el modelo pertenecen al conjunto de entrenamiento (`train`).

Resumen de los Resultados

Coeficientes del Modelo

- **Intercepto**: El coeficiente del intercepto es 0.20326, con un valor p de 0.00157. Esto indica que, cuando `Lag2` es cero, el log-odds de que la dirección sea "Up" es positivo y significativo.

- **Lag2**: El coeficiente de `Lag2` es 0.05810, con un valor p de 0.04298, lo cual es menor al nivel de significancia típico de 0.05. Esto sugiere que `Lag2` tiene un efecto estadísticamente significativo en la probabilidad de que `Direction` sea "Up". Un coeficiente positivo indica que, a medida que `Lag2` aumenta, la probabilidad de que `Direction` sea "Up" también aumenta.

Métricas del Modelo

- **Null Deviance**: La devianza nula es 1354.7 con 984 grados de libertad, lo que representa la variabilidad total de la variable dependiente sin ajustar por ningún predictor.

- **Residual Deviance**: La devianza residual es 1350.5 con 983 grados de libertad, lo que indica la variabilidad no explicada por el modelo después de incluir `Lag2` como predictor. La reducción en la devianza, aunque pequeña, es significativa dado el valor p de `Lag2`.

- **AIC (Criterio de Información de Akaike)**: El valor del AIC es 1354.5, lo que sugiere una medida de la calidad del modelo en términos de ajuste y complejidad. Un AIC más bajo generalmente indica un mejor equilibrio entre ajuste y parsimoniocidad del modelo.

Interpretación

Este modelo sugiere que la variable `Lag2` es un predictor significativo de la dirección (`Direction`). A medida que `Lag2` aumenta, la probabilidad de que `Direction` sea "Up" también aumenta. Este resultado implica que los cambios en `Lag2` tienen una relación importante con la dirección del mercado en el conjunto de entrenamiento.

En general, el modelo con `Lag2` como único predictor es una versión simplificada que aún proporciona información valiosa sobre el comportamiento de la variable dependiente.

5. Representa gráficamente el modelo:

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

$$H_0: B_i = 0$$

$$H_1: B_i \neq 0$$

```
# Prueba de Chi-cuadrado
anova(model_significant, 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
```

Descripción de la Prueba

La prueba de chi-cuadrado se utiliza para evaluar la significancia del modelo logístico ajustado con la variable `Lag2` como único predictor (`model_significant`). La prueba compara la devianza del modelo nulo (sin predictores) con la devianza residual del modelo que incluye `Lag2`, para determinar si `Lag2` aporta una mejora significativa en la predicción de `Direction`.

Resultados de la Prueba

- **Deviance del Modelo Nulo**: 1354.7 con 984 grados de libertad. Este valor representa la variabilidad total en la variable de respuesta sin incluir ningún predictor.
- **Deviance Residual del Modelo con `Lag2`**: 1350.5 con 983 grados de libertad. La inclusión de `Lag2` en el modelo reduce la devianza en 4.1666.
- **Valor p**: 0.04123. Este valor p es menor que el nivel de significancia típico de 0.05, lo que indica que la inclusión de `Lag2` en el modelo mejora significativamente la predicción de `Direction` en comparación con el modelo nulo.

Interpretación

Dado que el valor p es menor que 0.05, podemos concluir que el predictor `Lag2` es significativo en el modelo y aporta una mejora estadísticamente significativa en la predicción de `Direction`. Esto respalda la importancia de `Lag2` como predictor de la dirección del mercado en este contexto.

La prueba de chi-cuadrado valida la elección de `Lag2` como variable significativa en el modelo, confirmando que su inclusión ayuda a reducir la devianza residual de manera significativa en comparación con un modelo sin predictores.

```
# Matriz de confusión en el conjunto de prueba con el modelo significativo
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

pred_significant <- predict(model_significant, newdata = test, type =
"response")
pred_class_significant <- ifelse(pred_significant > 0.5, "Up", "Down")
confusionMatrix(as.factor(pred_class_significant), test$Direction)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9  5
##       Up     34 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.2439
##
##                   Kappa : 0.1414
##
##  Mcnemar's Test P-Value : 7.34e-06
```

```
##
##             Sensitivity : 0.20930
##             Specificity : 0.91803
##          Pos Pred Value : 0.64286
##          Neg Pred Value : 0.62222
##              Prevalence : 0.41346
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.13462
##       Balanced Accuracy : 0.56367
##
##        'Positive' Class : Down
##
```

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

# Vector inicial de predicciones, asumiendo todos como "Down"
pred.modelo <- rep("Down", length(prob.modelo))

# Cambiar la predicción a "Up" para las probabilidades mayores a 0.5
pred.modelo[prob.modelo > 0.5] <- "Up"

# Crear una variable con las observaciones reales de `Direction` en el conjunto de prueba
Direction.test <- test$Direction

# Crear la matriz de confusión
matriz.confusion <- table(pred.modelo, Direction.test)
print(matriz.confusion)
```

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

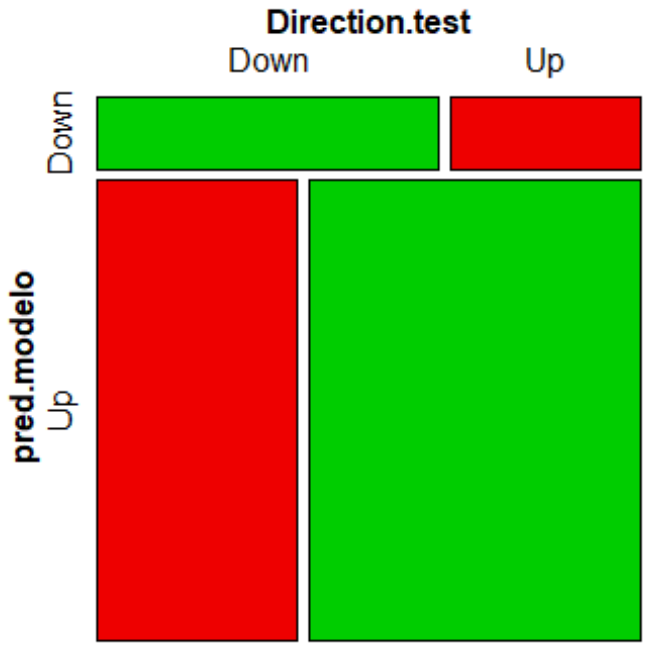
```r
# Visualización de la matriz de confusión utilizando el paquete vcd
library(vcd)
```

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

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

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

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



Descripción de la Matriz de Confusión

La matriz de confusión muestra los resultados de las predicciones del modelo significativo (`model_significant`) en el conjunto de datos de prueba. Los valores indican el número de observaciones clasificadas correctamente e incorrectamente en cada clase.

- **Clase Positiva (`Down`)**:
  - Predicciones Correctas (`Down`): 9
  - Falsos Negativos (`Up` predicho como `Down`): 5
- **Clase Negativa (`Up`)**:
  - Predicciones Correctas (`Up`): 56
  - Falsos Positivos (`Down` predicho como `Up`): 34

Métricas de Rendimiento

- **Exactitud (Accuracy)**: 0.625 (Intervalo de confianza del 95%: 0.5247 a 0.718). Esto significa que el modelo predijo correctamente el 62.5% de las observaciones en el conjunto de prueba.
- **Kappa**: 0.1414, una medida de la concordancia entre las predicciones del modelo y los valores reales, ajustada por el azar. Un valor bajo de Kappa indica que la concordancia entre predicciones y observaciones reales no es fuerte.

Métricas de Sensibilidad y Especificidad

- **Sensibilidad (Recall para `Down`)**: 0.2093. Indica que el modelo identificó correctamente el 20.93% de las observaciones `Down`.

- **Especificidad (Recall para `Up`)**: 0.9180. Indica que el modelo identificó correctamente el 91.80% de las observaciones `Up`.

Predicción Positiva y Negativa

- **Valor Predictivo Positivo (PPV)**: 0.6429. Esto significa que, de todas las observaciones clasificadas como `Down`, el 64.29% eran realmente `Down`.
- **Valor Predictivo Negativo (NPV)**: 0.6222. Esto significa que, de todas las observaciones clasificadas como `Up`, el 62.22% eran realmente `Up`.

Tasa de Falsos Negativos y Falsos Positivos

- **Tasa de Falsos Negativos**: 0.7907. El modelo etiquetó incorrectamente el 79.07% de los casos reales `Down` como `Up`.
- **Tasa de Falsos Positivos**: 0.0825. El modelo etiquetó incorrectamente el 8.25% de los casos reales `Up` como `Down`.

Prueba de McNemar

- **Valor p de McNemar**: 7.34e-06. Esta prueba evalúa si hay una diferencia significativa entre las tasas de error para las clases positivas y negativas. Un valor p bajo (< 0.05) sugiere que existe una diferencia significativa en las tasas de error, lo cual podría indicar un sesgo en la predicción del modelo.

Conclusión

El modelo significativo (`model_significant`) muestra un rendimiento moderado en términos de exactitud (62.5%) y especificidad (91.80%), pero tiene una baja sensibilidad (20.93%), lo que indica que el modelo tiene dificultades para identificar correctamente las observaciones de la clase `Down`. La prueba de McNemar sugiere una diferencia significativa en las tasas de error, lo cual indica un sesgo en las predicciones hacia la clase `Up`. Este análisis sugiere que, aunque `Lag2` es un predictor significativo, el modelo puede no ser lo suficientemente robusto para predecir ambas clases de manera equilibrada.

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

El modelo logístico ajustado tiene la siguiente forma, donde `Direction` es la variable de respuesta y `Lag2` es el predictor significativo:

$$\log\left(\frac{P(\text{Direction} = \text{Up})}{1 - P(\text{Direction} = \text{Up})}\right) = 0.20326 + 0.05810 \times \text{Lag2}$$

Interpretación de los Coeficientes

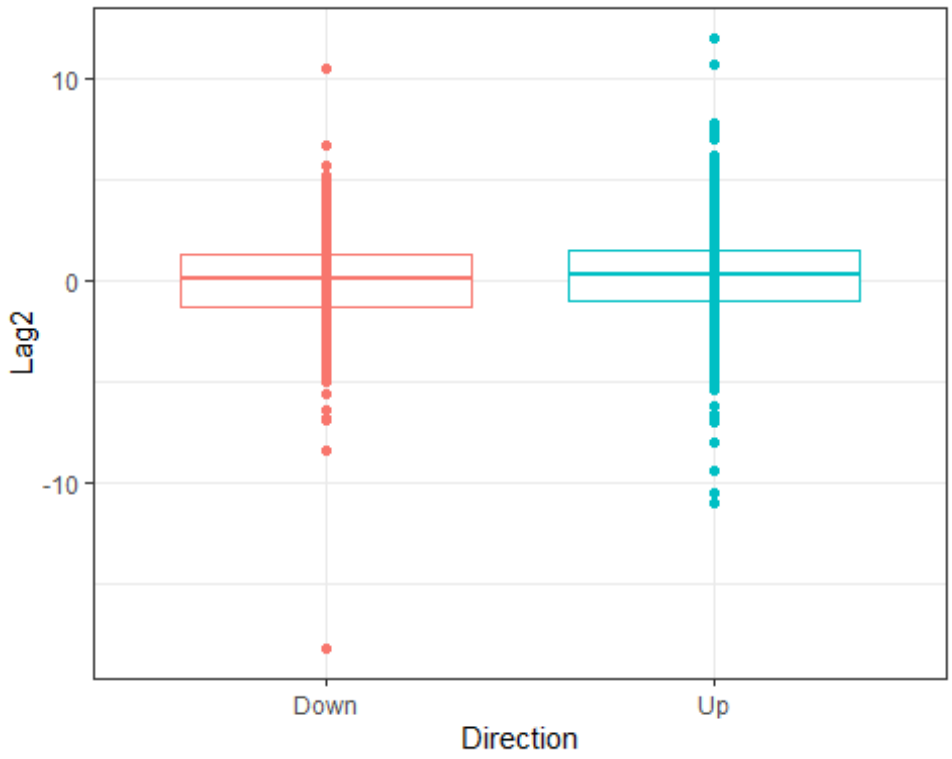
- **Intercepto (0.20326)**: Este valor indica el log-odds de que `Direction` sea “Up” cuando `Lag2` es cero.
- **Coeficiente de `Lag2` (0.05810)**: Un coeficiente positivo sugiere que un incremento en `Lag2` aumenta la probabilidad de que `Direction` sea “Up”. Específicamente, por

cada unidad de aumento en `Lag2`, el log-odds de `Direction = Up` se incrementa en 0.05810. En términos de odds, por cada unidad de aumento en `Lag2`, las probabilidades de que `Direction` sea "Up" se multiplican aproximadamente por

$$e^{0.05810} \approx 1.06$$

.

```r
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")
```



Descripción de la Gráfica

La gráfica es un diagrama de caja (boxplot) que muestra la distribución de la variable `Lag2` en función de la variable categórica `Direction` (con niveles "Down" y "Up"). Además, se han añadido puntos individuales para observar la dispersión de los valores de `Lag2` dentro de cada categoría de `Direction`.

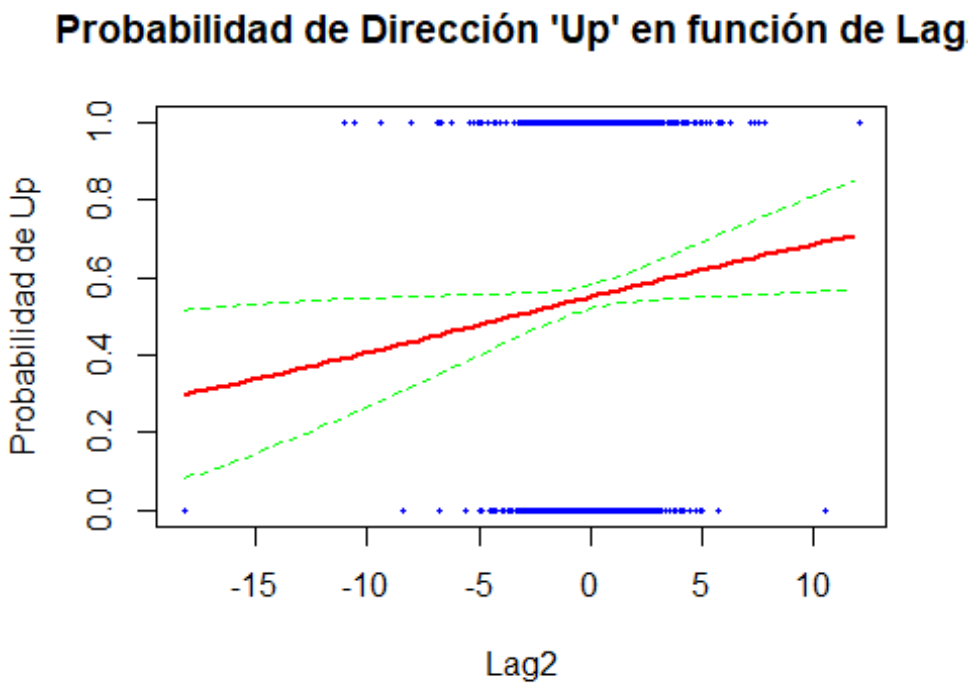
- **Eje X**: Representa las dos posibles direcciones: "Down" y "Up".
- **Eje Y**: Representa los valores de `Lag2`.
- **Diagrama de Caja**: El boxplot muestra la mediana, los cuartiles, y los valores atípicos de `Lag2` para cada categoría de `Direction`.

Interpretación

- La posición de las cajas y la dispersión de los puntos muestran cómo varían los valores de `Lag2` en función de `Direction`.
- En esta gráfica, podemos observar que la distribución de `Lag2` tiene un comportamiento diferente según la dirección:
  - **Direction = Down**: La mediana y el rango intercuartílico para `Lag2` están centrados en torno a un valor ligeramente inferior.
  - **Direction = Up**: La mediana de `Lag2` es un poco más alta en comparación con la de "Down", lo que sugiere una posible relación positiva entre `Lag2` y la probabilidad de que `Direction` sea "Up".

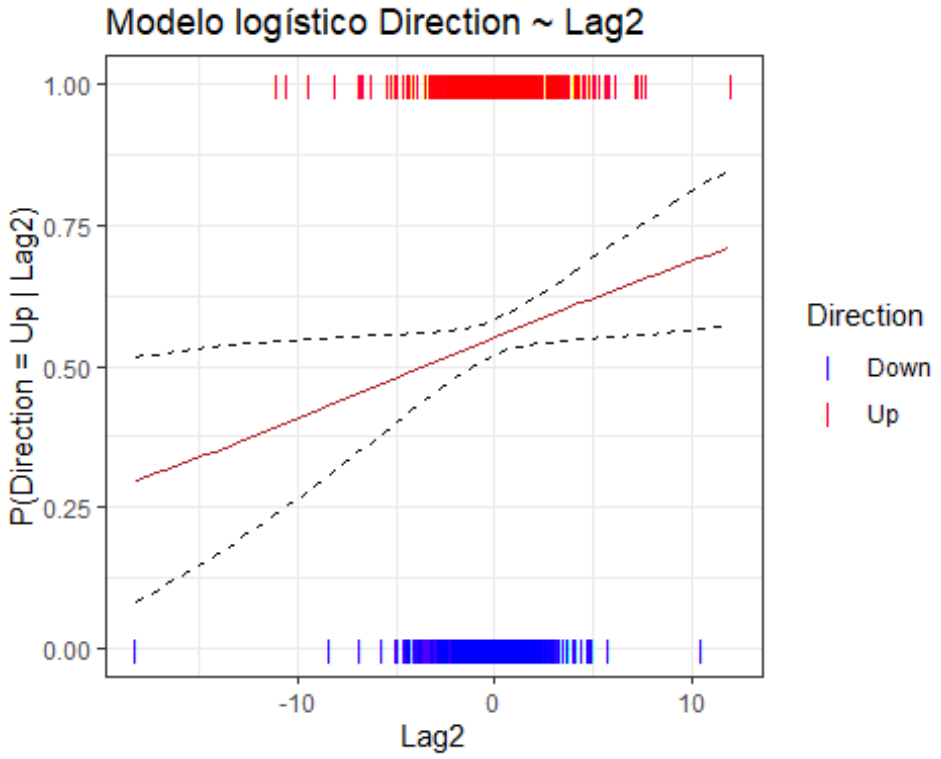
Este patrón es consistente con los resultados del modelo logístico, que indicaron que un incremento en `Lag2` se asocia con una mayor probabilidad de que `Direction` sea "Up". La gráfica ayuda a visualizar esta relación y refuerza la interpretación de `Lag2` como un predictor significativo en el modelo.

```
# Definir un vector con nuevos valores en el rango de la variable Lag2
nuevos_puntos <- seq(from = min(train$Lag2), to = max(train$Lag2), by = 0.5)

# Realizar predicciones con el modelo significativo usando los nuevos puntos
de Lag2
# Convertimos las predicciones al tipo "response" para obtener probabilidades
predicciones <- predict(model_significant, newdata = data.frame(Lag2 =
nuevos_puntos),
                        se.fit = TRUE, type = "response")

# Extraer las predicciones y el intervalo de confianza
probabilidades <- predicciones$fit
se <- predicciones$se.fit
limite_inferior <- probabilidades - 1.96 * se
limite_superior <- probabilidades + 1.96 * se

# Graficar los resultados
plot(train$Lag2, as.numeric(train$Direction == "Up"),
     xlab = "Lag2", ylab = "Probabilidad de Up",
     main = "Probabilidad de Dirección 'Up' en función de Lag2",
     col = "blue", pch = 20, cex = 0.6)
lines(nuevos_puntos, probabilidades, col = "red", lwd = 2)
lines(nuevos_puntos, limite_inferior, col = "green", lty = 2)
lines(nuevos_puntos, limite_superior, col = "green", lty = 2)
```




```r
# Paso 1: Crear los límites de los intervalos de confianza al 95%
CI_inferior <- predicciones$fit - 1.96 * predicciones$se.fit
CI_superior <- predicciones$fit + 1.96 * predicciones$se.fit

# Paso 2: Crear un data frame con los nuevos puntos y las predicciones
datos_curva <- data.frame(
  Lag2 = nuevos_puntos,
  probabilidad = predicciones$fit,
  CI_inferior = CI_inferior,
  CI_superior = CI_superior
)

# Paso 3: Codificar la variable `Direction` como binaria (0 y 1) en los datos originales
train$Direction <- ifelse(train$Direction == "Down", 0, 1)

# Paso 4: Graficar el modelo utilizando ggplot2
library(ggplot2)

ggplot(train, aes(x = Lag2, y = Direction)) +
  geom_point(aes(color = as.factor(Direction)), shape = "|", 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()
```



Descripción de la Gráfica

Esta gráfica muestra la probabilidad predicha de que `Direction` sea "Up" en función de `Lag2` utilizando el modelo logístico ajustado solo con la variable `Lag2`. La gráfica incluye:

- **Línea Roja**: Representa la probabilidad predicha de `Direction = Up` en función de `Lag2`.
- **Líneas Negras Punteadas**: Estas líneas representan el intervalo de confianza al 95% para la probabilidad predicha.
- **Puntos Horizontales en la Parte Superior e Inferior**: Los puntos rojos en la parte superior representan observaciones donde `Direction` es "Up", mientras que los puntos azules en la parte inferior representan observaciones donde `Direction` es "Down".

Interpretación

- **Relación Positiva**: La gráfica muestra que a medida que `Lag2` aumenta, la probabilidad predicha de que `Direction` sea "Up" también aumenta. Esto es consistente con el coeficiente positivo de `Lag2` en el modelo logístico, lo que sugiere que `Lag2` tiene una influencia positiva sobre la probabilidad de `Direction = Up`.
- **Intervalos de Confianza**: Las líneas negras punteadas muestran los intervalos de confianza al 95% alrededor de la probabilidad predicha. A medida que nos alejamos de valores de `Lag2` cercanos a cero, el intervalo de confianza se ensancha ligeramente, indicando una mayor incertidumbre en la predicción de `Direction` para valores extremos de `Lag2`.
- **Distribución de Observaciones**: Los puntos horizontales en la parte superior e inferior de la gráfica muestran la distribución de las observaciones en función de `Lag2`. Se observa que las observaciones con `Direction = Up` tienden a tener valores de `Lag2` más altos, mientras que las observaciones con `Direction = Down` tienen valores de `Lag2` más bajos.

Conclusión

Esta gráfica confirma la relevancia de `Lag2` como predictor en el modelo logístico para `Direction`. La probabilidad de que `Direction` sea "Up" aumenta con `Lag2`, y el modelo captura esta relación de manera adecuada. El intervalo de confianza proporciona una idea de la precisión de estas predicciones en diferentes rangos de `Lag2`.