A7-Regresión logística-fer

Fernanda Pérez

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)
library(tidyverse)
```

```
## ── 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
```

```
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)
```

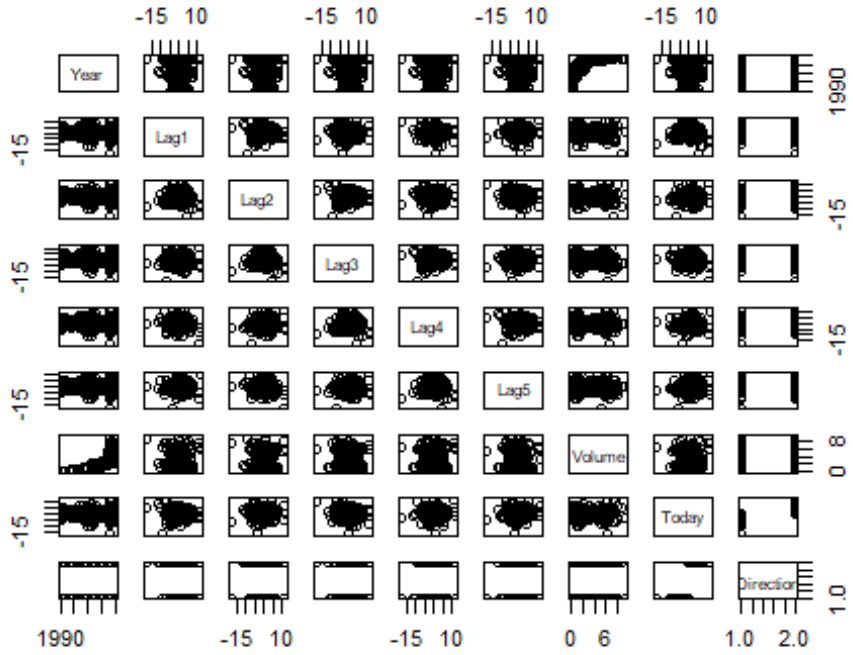
```
## Rows: 1,089
## Columns: 9
## $ Year      <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990,
1990, 1990, …
## $ Lag1      <dbl> 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372,
0.807, 0…
## $ Lag2      <dbl> 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -
1.372, 0…
## $ Lag3      <dbl>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

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

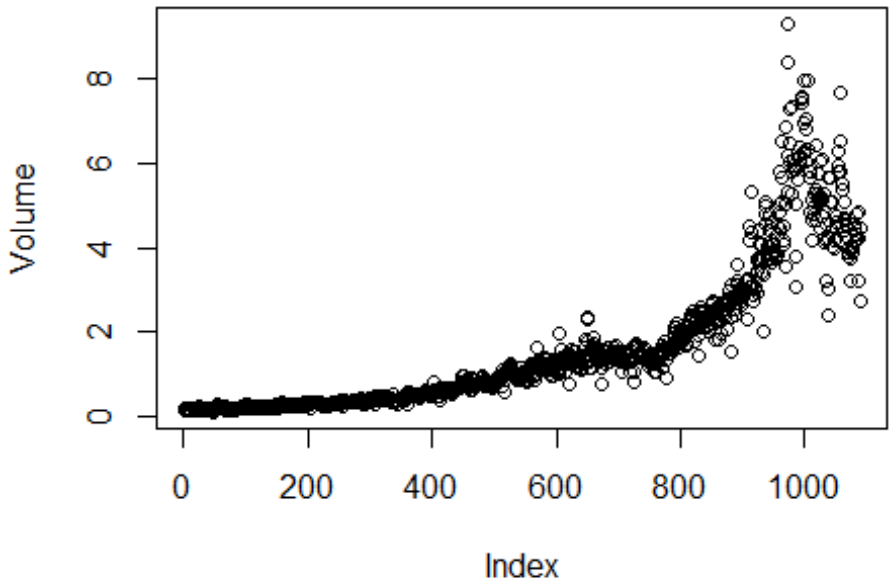
```
pairs(Weekly)
```




```
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)
```




```
detach(Weekly)
```

En el análisis exploratorio del database Weekly, se pueden observar algunas características importantes, como que Volume aumenta con el tiempo, lo cual indica un incremento en la actividad del mercado. No se ve una correlación fuerte en las variables Lag1 a Lag5 con la dirección del mercado (Direction), Lag1 vemos que es significativa en el modelo logístico. Esto nos sugiere que el rendimiento de la semana anterior (Lag1) tiene un impacto modesto en predecir si el mercado irá "Up" o "Down" en la semana siguiente.

2) Formula un modelo logístico con todas las variables menos la variable "Today". Calcula los intervalos de confianza para las B_1. 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(Weekly$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
```

Se ha ajustado un modelo de regresión logística con todas las variables, pero se excluye la variable "Today". El resultado nos muestra que Lag2 es la única variable con un valor p significativo (p= 0.0275) indicandonos que Lag2 tiene un efecto estadísticamente significativo sobre la dirección del mercado (Direction).

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.

```
datos.entrenamiento <- Weekly$Year < 2009
datos.test <- Weekly[!datos.entrenamiento, ]
```

```r
modelo.log.m <- glm(Direction ~ . -Today, data = Weekly, family =
binomial, subset = datos.entrenamiento)
summary(modelo.log.m)
```

```
##
## Call:
## glm(formula = Direction ~ . - Today, family = binomial, data = Weekly,
##     subset = datos.entrenamiento)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.779438  42.446904   0.065   0.9478
## Year        -0.001227   0.021282  -0.058   0.9540
## Lag1        -0.062163   0.029466  -2.110   0.0349 *
## Lag2         0.044903   0.030066   1.493   0.1353
## Lag3        -0.015305   0.029595  -0.517   0.6050
## Lag4        -0.030967   0.029342  -1.055   0.2913
## Lag5        -0.037599   0.029353  -1.281   0.2002
## Volume      -0.085115   0.096432  -0.883   0.3774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1342.3  on 977  degrees of freedom
## AIC: 1358.3
##
## Number of Fisher Scoring iterations: 4
```

Vemos que el modelo está bien ajustado usando el conjunto de datos de entrenamiento de 1990 a 2008. De acuerdo con los resultados que obtuvimos , Lag1 es la única variable significativa p= 0.0349, indicandonos que el rendimiento de la semana anterior si tiene un efecto relevante en la dirección del mercado (Direction). Las demás variables (Year, Lag2, Lag3, Lag4, Lag5, y Volume) no tienen significancia estadística en este modelo, indicandonos que su influencia en la probabilidad de que el mercado suba o baje en la semana actual es muy pqueña o nula.

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

```r
modelo.log.s <- glm(Direction ~ Lag1, data = Weekly, family = binomial,
subset = datos.entrenamiento)
summary(modelo.log.s)
```

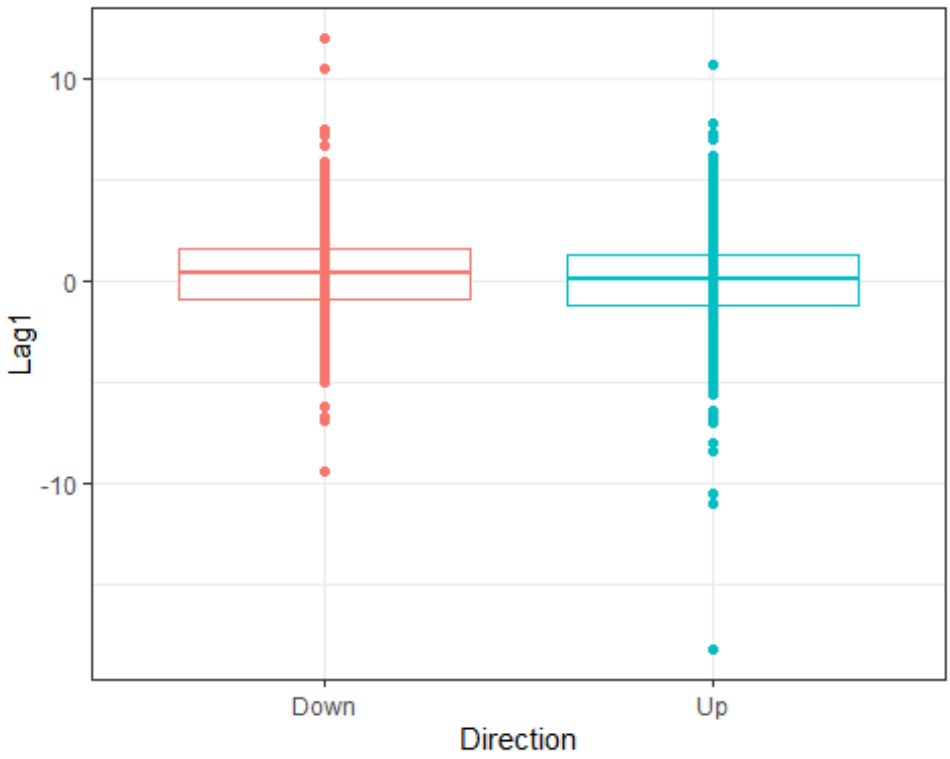
```
##
## Call:
## glm(formula = Direction ~ Lag1, family = binomial, data = Weekly,
##     subset = datos.entrenamiento)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.21829    0.06438   3.391 0.000697 ***
## Lag1        -0.05908    0.02892  -2.043 0.041059 *
## ---
## 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: 1350.4  on 983  degrees of freedom
## AIC: 1354.4
##
## Number of Fisher Scoring iterations: 4
```

Este modelo ya simplificado nos indica que Lag1 (que representa el rendimiento del mercado en la semana anterior) si tiene un impacto significativo en la dirección del mercado en la semana actual (Direction).Vemos que el coeficiente negativo de Lag1 (-0.05908) nos indica una relación inversa.

5) Representa gráficamente el modelo:

```
ggplot(data = Weekly, mapping = aes(x = Direction, y = Lag1)) +
  geom_boxplot(aes(color = Direction)) +
  geom_point(aes(color = Direction)) +
  theme_bw() +
  theme(legend.position = "none")
```



Graficamente usamos un diagrama de cajas para visualizar más sencillamente la relación entre Lag1 y Direction, nos muestra cómo los valores de Lag1 (rendimiento de la semana anterior) se distribuyen según la dirección del mercado (Up o Down) en la semana actual. La visualización nos indica que los valores de Lag1 tienden a ser ligeramente un poco mayores cuando Direction es "Down" en comparación con cuando Direction es "Up" y por otro lado las diferencias en la mediana y la dispersión de Lag1 entre "Up" y "Down" no llama mucho la atención, indicandonos que esta variable tiene un efecto predictivo modesto.

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

```
predicciones <- predict(modelo.log.s, newdata = datos.test, type =
"response")
predicciones_clase <- ifelse(predicciones > 0.5, "Up", "Down")

table(Predicción = predicciones_clase, Real = datos.test$Direction)
```

```
##           Real
## Predicción Down Up
##       Down    4  6
##       Up     39 55
```

Aquí se hizo una matriz de confusión, donde la tabla nos va a mostrar el desempeño del modelo en el conjunto de prueba. Esta nos indica que el modelo predijo correctamente 55 observaciones como "Up" y 4 como "Down". Sin embargo, también cometió 39 errores al predecir "Up" cuando la dirección real era "Down" y 6 errores al predecir "Down" cuando la dirección real era "Up". Vemos que apesar de que Lag1 nos mostró una relación significativa en el conjunto de entrenamiento, su poder predictivo en el conjunto de prueba es limitado, tiene sentido ya que es complejo predecir la dirección del mercado con un solo predictor.

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)

```
intercepto <- 0.21829
coef_lag1 <- -0.05908

ecuacion <- paste("logit(p) =", round(intercepto, 5), "+",
round(coef_lag1, 5), "* Lag1")

print(ecuacion)
```

```
## [1] "logit(p) = 0.21829 + -0.05908 * Lag1"
```

La ecuación del modelo logístico nos muestra que la probabilidad de que Direction sea "Up" depende del valor de Lag1.

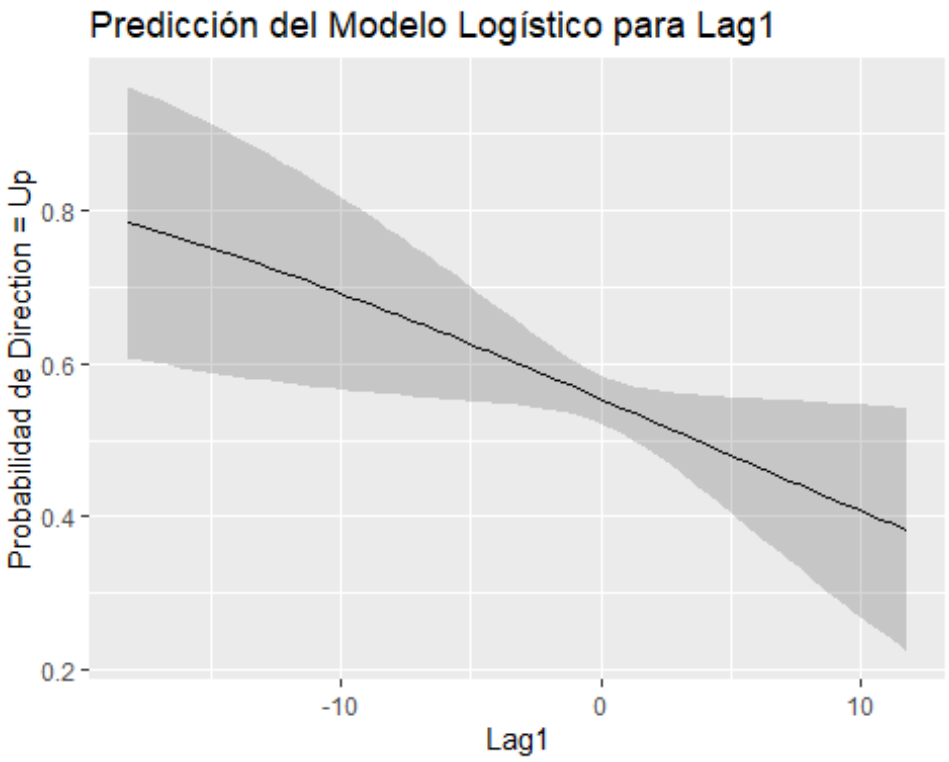
```r
nuevos_puntos <- seq(from = min(Weekly$Lag1), to = max(Weekly$Lag1), by =
0.5)

predicciones <- predict(modelo.log.s, newdata = data.frame(Lag1 =
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

datos_curva <- data.frame(Lag1 = nuevos_puntos, probabilidad =
predicciones$fit,
                          CI_inferior = CI_inferior, CI_superior =
CI_superior)

ggplot(datos_curva, aes(x = Lag1, y = probabilidad)) +
  geom_line() +
  geom_ribbon(aes(ymin = CI_inferior, ymax = CI_superior), alpha = 0.2) +
  labs(title = "Predicción del Modelo Logístico para Lag1", y =
"Probabilidad de Direction = Up", x = "Lag1")
```




```r
confusion_matrix <- table(Predicción = predicciones_clase, Real =
datos.test$Direction)
precision <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
precision_percentage <- precision * 100
cat("La precisión del modelo en el conjunto de prueba es:",
round(precision_percentage, 2), "%\n")
```

```
## La precisión del modelo en el conjunto de prueba es: 56.73 %
```

En la grafica vemos la relación entre Lag1 y la probabilidad de que Direction sea “Up”. Vemos que conforme Lag1 aumenta, la probabilidad de que el mercado suba (Direction = “Up”) disminuye. Esto es consistente con la interpretación de una relación inversa entre el rendimiento de la semana pasada y la dirección del mercado esta semana, lo cual sugiere una tendencia de reversión.

En esta actividad de análisis, se desarrolló un modelo de regresión logística para predecir la dirección semanal del mercado (Direction) usando el rendimiento de la semana anterior (Lag1) como predictor. Los resultados mostraron que Lag1 tiene una relación inversa con la probabilidad de que el mercado suba,indicando una posible tendencia de reversión en el mercado. Sin embargo, apesar de que Lag1 fue estadísticamente significativa, la precisión predictiva del modelo en el conjunto de prueba fue limitada, indicando que predecir la dirección del mercado con una sola variable tal vez no suficiente , obtuvimos una precisión del casi 57% lo cual es aceptable ya que no cae en la aleatoriedad, sin embargo mejorar el modelo, podriamos considerar variables adicionales y factores externos que afecten el mercado. En resumen, esta actividad destaca la importancia de múltiples factores en la predicción de mercados financieros y nos muestra las limitaciones de los modelos simplistas en entornos complejos.