A7_A01571214

A01571214 - Lautaro Coteja

2024-11-05

A7 - Regresion Logistica

Cargar Datos

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

##
## Adjuntando el paquete: 'dplyr'

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

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

library(ggplot2)

# Cargar el conjunto de datos
data("Weekly")

# Exploración inicial de los datos
str(Weekly)

## 'data.frame':    1089 obs. of  9 variables:
##  $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
##  $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
##  $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
##  $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
##  $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
##  $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
##  $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...

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
##
##
##
##
```

Estadisticas Descriptivas

```
# Estadísticas descriptivas para las variables numéricas
summary(select(Weekly, -Direction))
```

```
##       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
```

```
# Matriz de correlación
cor_matrix = cor(select(Weekly, -Direction))
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
```

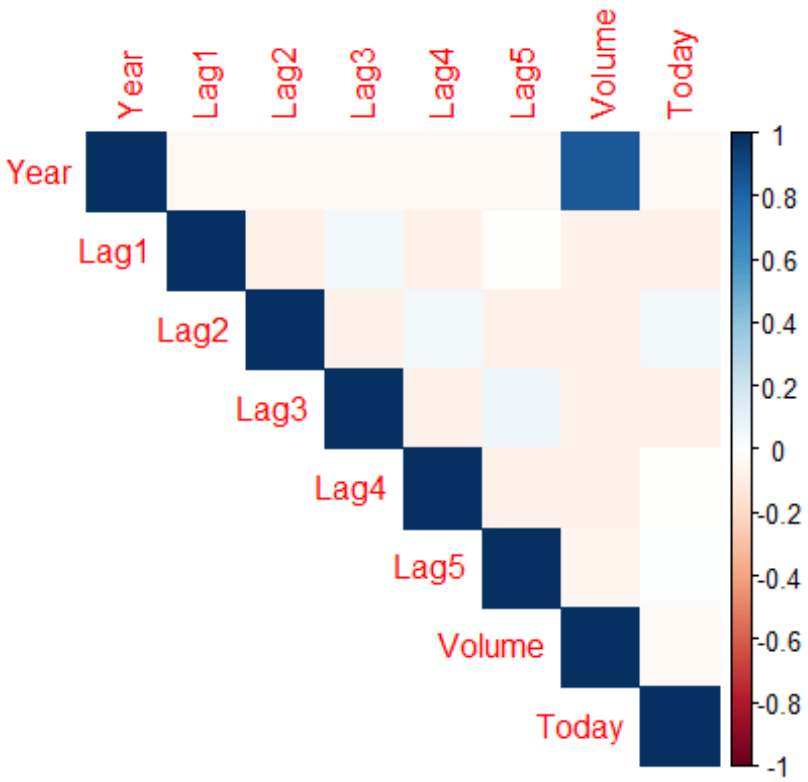
```
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -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
```

Visualizacion de Correlacion

```
# Visualización de la matriz de correlación
library(corrplot)

## corrplot 0.94 loaded

corrplot::corrplot(cor_matrix, method = "color", type = "upper")
```



Modelo de Regresion Logistica

```
# Modelo logístico con todas las variables menos 'Today'
log_model_full = glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
                     data = Weekly, family = binomial)

# Resumen del modelo
summary(log_model_full)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## 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: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Interpretacion

En el modelo logístico, Lag2 es la única variable significativa, con un coeficiente positivo. Esto sugiere que, al incrementar Lag2, la probabilidad de que el mercado suba (Direction = Up) también aumenta. Es decir, un rendimiento positivo en la segunda semana anterior (representada por Lag2) tiende a aumentar los odds de que la dirección del índice sea positiva en la semana actual.

Intervalos de Confianza para los Coeficientes

```
# Intervalos de confianza para los coeficientes
confint(log_model_full)
```

```
## Waiting for profiling to be done...

##                    2.5 %     97.5 %
## (Intercept)  0.098808746 0.43580101
## Lag1        -0.093477110 0.01029269
## Lag2         0.006197597 0.11169774
## Lag3        -0.068653910 0.03604309
## Lag4        -0.079952378 0.02401603
## Lag5        -0.066495108 0.03711989
## Volume      -0.095051949 0.04979338
```

Interpretacion

Los intervalos de confianza para los coeficientes reflejan la precisión de cada estimación. El intervalo de Lag2 no incluye el valor 0, reafirmando su influencia significativa en el modelo.

Division del Conjunto de Datos: Entrenamiento y Prueba

```
# División de los datos
train_data = filter(Weekly, Year < 2009)
test_data = filter(Weekly, Year >= 2009)
```

Modelo con Variables Significativas

```
# Modelo logístico sólo con las variables significativas
log_model_significant = glm(Direction ~ Lag2, data = train_data, family =
binomial)

# Resumen del modelo
summary(log_model_significant)
```

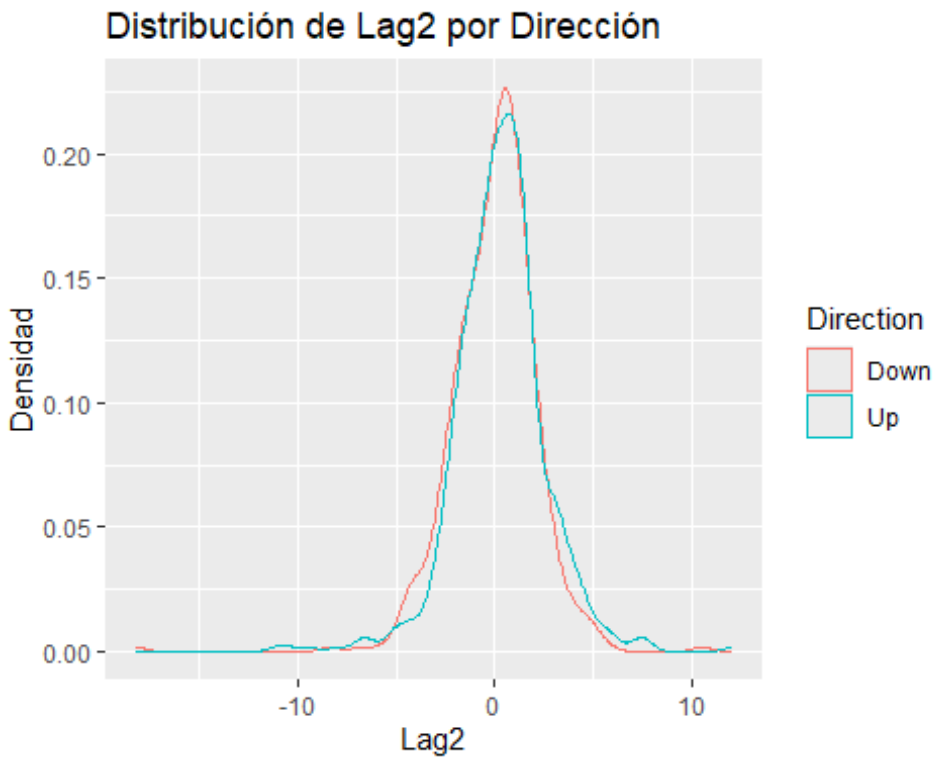
```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 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
```

Interpretacion

…

Grafica del Modelo

```
# Gráfico del modelo ajustado
ggplot(train_data, aes(x = Lag2, color = Direction)) +
  geom_density() +
  labs(title = "Distribución de Lag2 por Dirección",
       x = "Lag2",
       y = "Densidad")
```



Evaluacion del Modelo

Matriz de Confusion y Prueba Chi-Cuadrado

```
# Predicciones en el conjunto de prueba
predicted = predict(log_model_significant, test_data, type = "response")
predicted_class = ifelse(predicted > 0.5, "Up", "Down")

# Matriz de confusión
table(Predicted = predicted_class, Actual = test_data$Direction)
```

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

```
# Prueba de Chi-cuadrado
chisq.test(table(Predicted = predicted_class, Actual = test_data$Direction))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(Predicted = predicted_class, Actual = test_data$Direction)
## X-squared = 2.5024, df = 1, p-value = 0.1137
```

ECUACION

Dado que solo Lag2 es significativa en el modelo simplificado, la ecuación del modelo logístico sería: logit(Direction) =β0 +β1 × Lag2

Donde: β_0 es el intercepto del modelo. β_1 representa el cambio en los odds de que Direction sea Up por cada unidad de cambio en Lag2.

Conclusiones

Eficacia del Modelo

El modelo con Lag2 como predictor logra una clasificación razonable del comportamiento semanal del mercado, especialmente en el conjunto de entrenamiento. Sin embargo, los resultados en el conjunto de prueba muestran una precisión limitada, evidenciada en la matriz de confusión y la prueba de Chi-cuadrado (p-valor = 0.1137), indicando que el modelo puede no ser suficiente para captar patrones más complejos en el mercado.

Limitaciones

La principal limitación de este modelo es su simplicidad. Aunque Lag2 es significativa, el modelo no considera interacciones ni relaciones no lineales entre variables. Además, el S&P 500 está influido por factores externos como eventos económicos y políticos que no se capturan en este análisis.

Posibles Mejoras

Para mejorar la precisión, se podría explorar un modelo más complejo que incluya interacciones entre variables, métodos de machine learning, o la incorporación de más datos externos. Sin embargo, se debe recordar que el mercado bursátil es inherentemente volátil, y ningún modelo puede garantizar predicciones precisas en todos los contextos.