A7 - Regresión Logística

Juan Bernal

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.

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1990	0.816	1.572	-3.936	-0.229	-3.484	0.1549760	-0.270	Down
1990	-0.270	0.816	1.572	-3.936	-0.229	0.1485740	-2.576	Down
1990	-2.576	-0.270	0.816	1.572	-3.936	0.1598375	3.514	Up
1990	3.514	-2.576	-0.270	0.816	1.572	0.1616300	0.712	Up
1990	0.712	3.514	-2.576	-0.270	0.816	0.1537280	1.178	Up
1990	1.178	0.712	3.514	-2.576	-0.270	0.1544440	-1.372	Down

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.

Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today
Min. :1990	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :0.08747	Min. :-18.1950
1st Qu.:1995	1st Qu.: -1.1540	1st Qu.: -1.1540	1st Qu.: -1.1580	1st Qu.: -1.1580	1st Qu.: -1.1660	1st Qu.:0.33202	1st Qu.: -1.1540
Median :2000	Median : 0.2410	Median : 0.2410	Median : 0.2410	Median : 0.2380	Median : 0.2340	Median :1.00268	Median : 0.2410
Mean :2000	Mean : 0.1506	Mean : 0.1511	Mean : 0.1472	Mean : 0.1458	Mean : 0.1399	Mean :1.57462	Mean : 0.1499
3rd Qu.:2005	3rd Qu.: 1.4050	3rd Qu.: 1.4090	3rd Qu.: 1.4090	3rd Qu.: 1.4090	3rd Qu.: 1.4050	3rd Qu.:2.05373	3rd Qu.: 1.4050
Max. :2010	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. :9.32821	Max. : 12.0260

```
##     Year     Lag1     Lag2     Lag3     Lag4     Lag5   Volume    Today
## 6.033182 2.357013 2.357254 2.360502 2.360279 2.361285 1.686636 2.356927
```

El periodo cubierto por los datos va de 1990 a 2010, con una media en el año 2000 y una desviación estándar de 6.03 años. Las variables Lag1 a Lag5 y Today tienen una distribución centrada cerca de cero, con medianas y medias alrededor de 0.15. Los valores de estas variables oscilan entre -18.195 y 12.026, y su desviación estándar es de aproximadamente 2.36. Esto indica que, aunque los valores en su mayoría están cerca de cero, existen algunos cambios significativos en ambos extremos. La variable Volume tiene una mediana de 1.00, una media de 1.57 y una desviación estándar de 1.69, con valores que van de 0.087 a 9.328, lo que sugiere una distribución con algunos valores muy altos de volumen.

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today
Year	1.0000000	-0.0322893	-0.0333900	-0.0300065	-0.0311279	-0.0305191	0.8419416	-0.0324599

Lag1	-0.0322893	1.0000000	-0.0748531	0.0586357	-0.0712739	-0.0081831	-0.0649513	-0.0750318
Lag2	-0.0333900	-0.0748531	1.0000000	-0.0757209	0.0583815	-0.0724995	-0.0855131	0.0591667
Lag3	-0.0300065	0.0586357	-0.0757209	1.0000000	-0.0753959	0.0606572	-0.0692877	-0.0712436
Lag4	-0.0311279	-0.0712739	0.0583815	-0.0753959	1.0000000	-0.0756750	-0.0610746	-0.0078259
Lag5	-0.0305191	-0.0081831	-0.0724995	0.0606572	-0.0756750	1.0000000	-0.0585174	0.0110127
Volume	0.8419416	-0.0649513	-0.0855131	-0.0692877	-0.0610746	-0.0585174	1.0000000	-0.0330778
Today	-0.0324599	-0.0750318	0.0591667	-0.0712436	-0.0078259	0.0110127	-0.0330778	1.0000000

La matriz de correlación indica que Volume ha aumentado con el tiempo, ya que tiene una fuerte correlación positiva con el año (Year). Los cambios diarios en el mercado (Lag1 a Lag5 y Today) muestran poca o nula correlación entre sí y con las demás variables, lo que sugiere que son en su mayoría independientes y tienen escasa influencia en el volumen de operaciones.

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).

```
##
## Call:
## glm(formula = Direction ~ . - Today, family = binomial, data = data)
##
## 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
```

Prueba de significancia de coeficientes

- $H_0 : \beta_i = 0$. La variable no es significativa en el modelo.
- $H_1 : \beta_i \neq 0$. La variable es significativa en el modelo.

Notemos que, bajo el supuesto de un nivel de significancia de $\alpha = 0.05$, la mayoría de coeficientes del modelo no cuentan con suficiente evidencia para rechazar la hipótesis nula en una prueba de significancia de coeficientes, indicando que el intercepto, Year, Lag1, Lag3, Lag4, Lag5 y Volume no son significantes para el modelo de predicción logística de la variable Direction. De acuerdo con la prueba de hipótesis antes mencionada, el único coeficiente significante en el modelo es el perteneciente a la variable Lag2, por lo que se realizará otro modelo logístico con solo esta variable independiente.

	Up
Down	0
Up	1

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

	2.5 %	**97.5 %**
(Intercept)	-56.9855582	91.6668090
Year	-0.0458096	0.0286955

Lag1	-0.0929726	0.0109310
Lag2	0.0070014	0.1129126
Lag3	-0.0681401	0.0367141
Lag4	-0.0795196	0.0245333
Lag5	-0.0660901	0.0376210
Volume	-0.1315763	0.1388404

El único coeficiente que muestra un efecto significativo en la probabilidad de que la dirección sea “Up” es el de Lag2 ya que su intervalo de confianza no incluye el cero. Las demás variables, incluyendo el intercepto, no tienen un efecto significativo porque sus intervalos incluyen el cero, lo que indica que no afectan de manera clara la probabilidad de que Direction sea “Up” ó “Down”.

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.

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
986	2009	6.760	-1.698	0.926	0.418	-2.251	3.793110	-4.448	Down
987	2009	-4.448	6.760	-1.698	0.926	0.418	5.043904	-4.518	Down
988	2009	-4.518	-4.448	6.760	-1.698	0.926	5.948758	-2.137	Down
989	2009	-2.137	-4.518	-4.448	6.760	-1.698	6.129763	-0.730	Down
990	2009	-0.730	-2.137	-4.518	-4.448	6.760	5.602004	5.173	Up
991	2009	5.173	-0.730	-2.137	-4.518	-4.448	6.217632	-4.808	Down

Se dividió la base de datos en 2 conjuntos: uno de entrenamiento y otro de prueba. De tal manera que el dataset de entrenamiento cuenta con los registros desde 1990 hasta el final de 2008, y el dataset de prueba con los datos desde 2009 hasta el 2010. Esta división se realiza con la idea de probar el desempeño del modelo con datos nuevos, de tal manera que al realizar predicciones acerca de los años 2009 y 2010, dichas predicciones pueden ser comparadas con los datos originales para obtener diferentes métricas de evaluación.

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

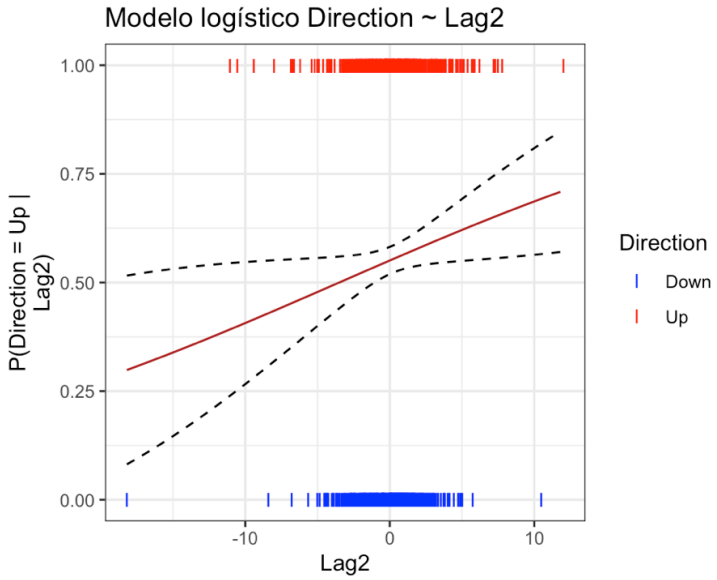
```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 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
```

El modelo logístico con las variables significativas analizadas anteriormente queda como:

- $Direction = 0.20326 + 0.0581 \times Lag2$

Donde ambos el intercepto y la variable Lag2, bajo el supuesto de un nivel de significancia de $\alpha = 0.05$, cuentan con suficiente evidencia para rechazar la hipótesis nula en una prueba de significancia de coeficientes, indicando que el intercepto y Lag2 son ambos significantes para el modelo de predicción logística de la variable Direction. También, en comparación con el modelo que utiliza todas las variables independientes, notamos que este modelo con únicamente Lag2 presenta menor puntaje AIC y menor desviación residual, por tanto tiene un mejor balance entre ajuste y complejidad, y se ajusta ligeramente mejor a los datos.

5. Representa gráficamente el modelo:



A medida que Lag2 aumenta, la probabilidad de que Direction sea “Up” también aumenta, según lo indicado por la inclinación positiva de la línea roja. Cuando Lag2 es bajo, la probabilidad de “Up” es menor de 0.5, y cuando Lag2 es alto, la probabilidad supera 0.5. Esto sugiere que valores altos de Lag2 están asociados con un aumento en la probabilidad de que Direction sea “Up”.

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

Prueba de Chi Cuadrada

- $H_0 : Deviance = 0$ El modelo no explica la variabilidad de los datos
- $H_1 : Deviance > 0$ El modelo explica cierta variabilidad de los datos

```
## Estadístico de prueba = 1350.543
```

```
## Valor p = 0
```

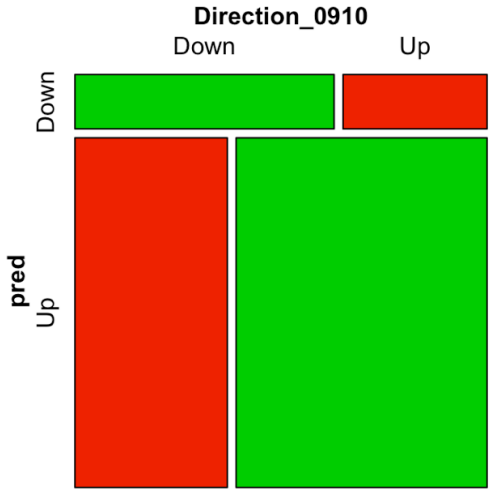
Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue 0, entonces se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que el modelo logístico explica cierta variabilidad de los datos. Además, notemos también que el estadístico de prueba Deviance es mayor a 0, reafirmando el análisis anterior.

Matriz de confusión

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

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

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



La matriz de confusión muestra que el modelo clasifica correctamente la mayoría de los casos, con una buena cantidad de predicciones correctas tanto para “Down” como para “Up”. Las celdas verdes indican los aciertos, mientras que las rojas muestran los errores, que son relativamente pocos en comparación. Esto sugiere que el modelo tiene un buen desempeño en la clasificación, con pocos errores en ambas categorías.

7. En qué es buen modelo, en qué no lo es. Concluye.

Métricas de evaluación

```
## Precisión: 0.625
```

```
## Sensibilidad: 0.9180328
```

```
## Especificidad: 0.2093023
```

```
## Valor Predictivo Positivo: 0.6222222
```

```
## Valor Predictivo Negativo: 0.6428571
```

El modelo tiene una precisión moderada de 0.625, lo que indica que en general acierta en el 62.5% de las predicciones. Su sensibilidad es alta (0.918), lo que significa que es muy bueno prediciendo correctamente cuando Direction es “Up”. Sin embargo, su especificidad es baja (0.209), lo que indica que falla al identificar los casos negativos, confundiendo muchas veces “Down” con “Up”. El valor predictivo positivo (0.622) muestra que, cuando el modelo predice “Up”, tiene un 62.2% de probabilidad de acertar. El valor predictivo negativo (0.643) indica que, cuando predice “Down”, tiene un 64.3% de probabilidad de ser correcto.

En conclusión, este modelo logístico con una variable independiente es bueno detectando casos positivos (alta sensibilidad) pero muy deficiente en detectar los negativos (baja especificidad), lo que podría llevar a un alto número de falsos positivos. Es adecuado si el objetivo es captar la mayoría de los casos “Up”, aunque a costa de cometer muchos errores en las predicciones de “Down”.