Pontificia Universidad JAVERIANA Bogotá

Regresión Logística

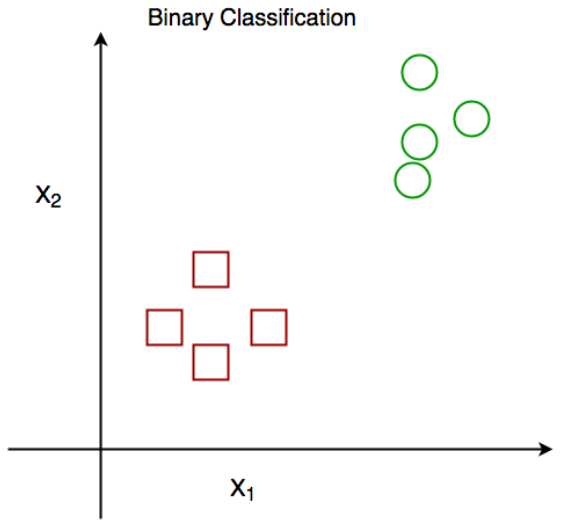
ING. HERIBERTO FELIZZOLA JIMENEZ

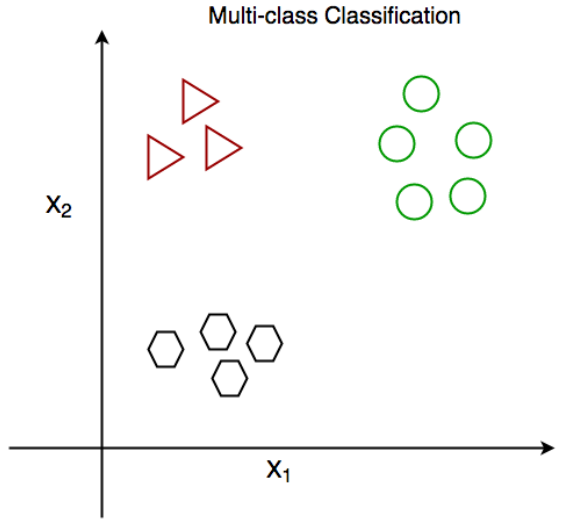
Educación **Continua**

Generamos experiencias educativas

Problema de Clasificación

- El problema de clasificación consiste en predecir una clase o categoría
- Se dividen en:
  - *Problema Clasificación binaria* -> la variable a predecir tiene dos posibles resultados.
  - *Problema Multi-Clase* -> la variable a predecir tiene más de dos posibles resultados.
- Los modelos y algoritmos para predecir son llamados clasificadores.
- Algunos ejemplos de problemas de clasificación:
  - Determinar si un proyecto tendrá retrasos o sobrecostos
  - Predecir del impago (default) de un crédito
  - Determinar si un correo es spam
  - Determinar la condición médica de una persona de acuerdo a sus síntomas
  - Detectar transacciones fraudulentas
  - Clasificar imágenes de objetos

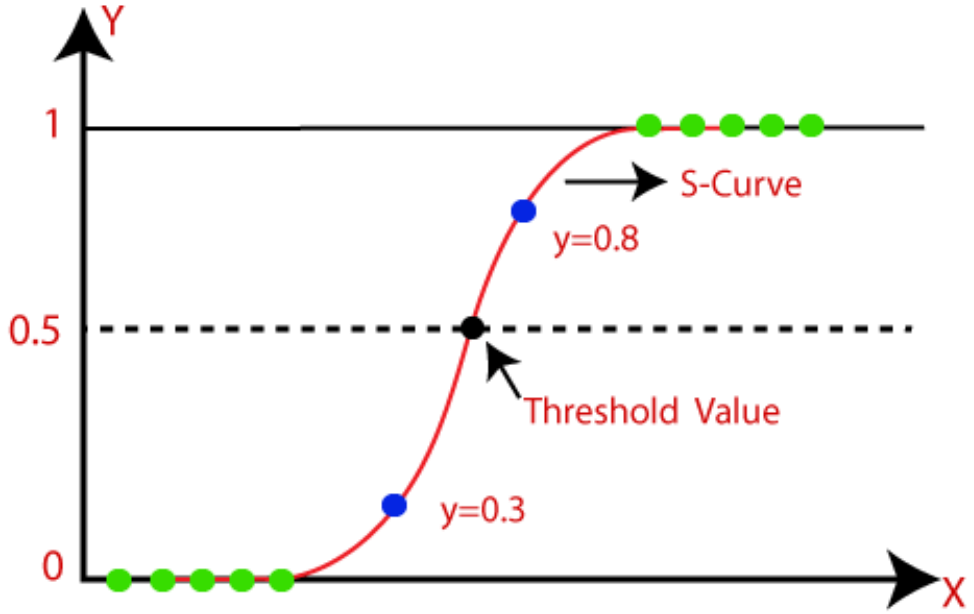




Regresión Logística - RL

- Se utiliza para modelar y predecir una variable discreta o binaria, esto es $\mathrm{Y} = \{0, 1\}$
- Los valores 0 y 1 representan categorías binarias, ejemplo:
  - Compra (1) vs No Compra (0)
  - Fraude (1) vs No Fraude (0)
  - Cumple (1) vs No Cumple (0)
  - Enfermo (1) vs Sano (0)
  - Spam (1) vs No Spam (0)
- Usualmente la categoría que interesa predecir se codifica con 1.
- La RL predice la probabilidad de *Y = 1* a partir de la siguiente ecuación:

$$E(Y) = P(Y = 1|X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_pX_p}}$$



Esto se denomina *función logística*. Es la que da la forma de S a la gráfica.

Estrategia de Promóción

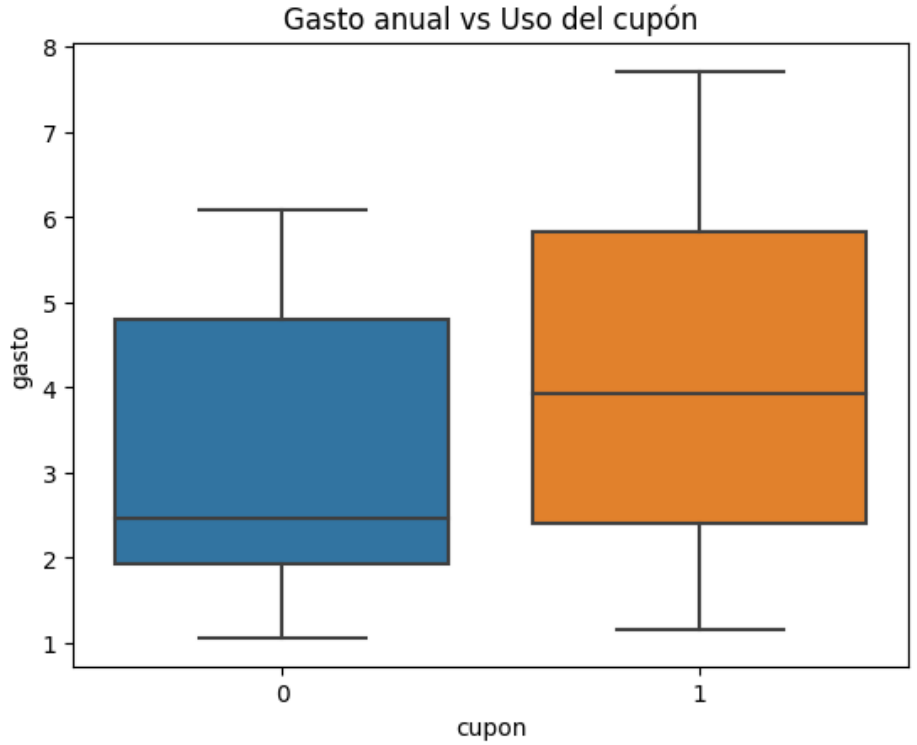
- Una cadena nacional de ropa realizará una promoción por correo.
- Ha ordenado imprimir copias de su catálogo y en cada uno incluye un cupón de $50 de descuento en compras por $200 o más.
- La cadena desea enviarlo sólo a aquellos clientes que tengan mayor probabilidad de usar el cupón.
- La gerencia considera que:
  1. La cantidad gastada anualmente por el cliente (X_1)
  2. Si posee o no una tarjeta de crédito de la tienda (X_2)
- Son dos variables útiles para predecir si ese cliente usará el cupón.
- Realiza un testeo con 32 clientes para obtener datos y construir un modelo de RL que le permita predecir si un cliente utilizará o no un cupón.

Cliente	Gasto Anual Millones	Posee Tarjeta de Crédito	Cupón
1	2,291	1	1
2	3,215	1	0
3	2,135	1	1
4	3,924	0	1
5	2,528	1	1
6	2,473	0	0
7	2,384	0	0
8	7,708	0	1
9	1,182	1	1
10	3,345	0	1
11	5,547	0	0
12	1,917	0	0
…	…	…	…

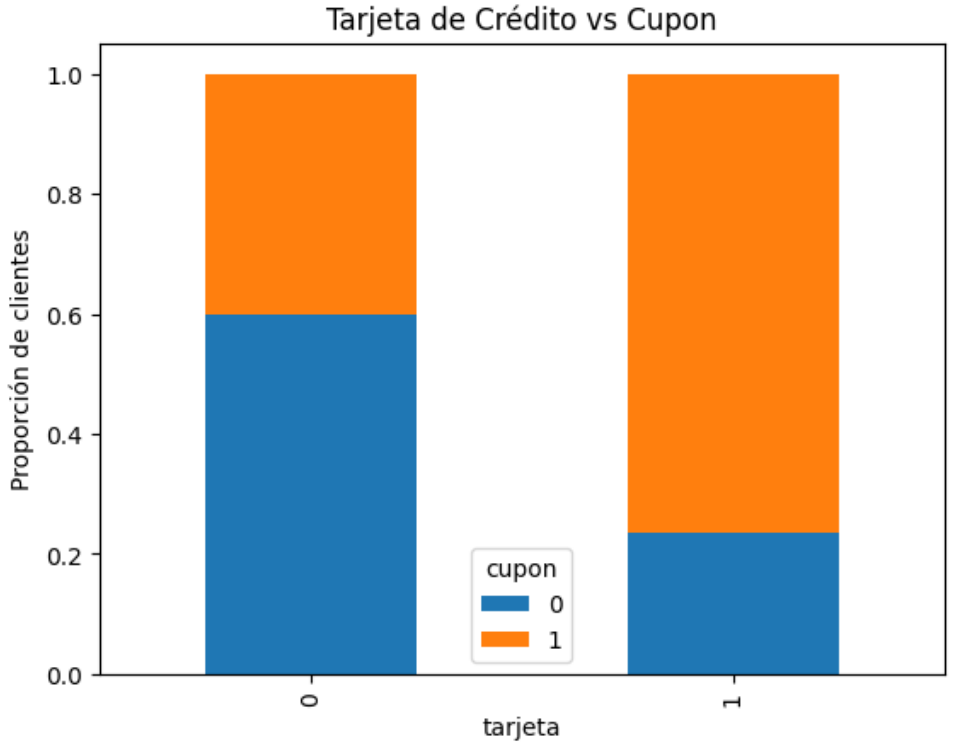
$$Y = 0 \rightarrow No\ utiliza\ el\ cupon$$

$$Y = 1 \rightarrow Utiliza\ el\ cupon$$

Análisis exploratorio



Los clientes que utilizan el cupón (1) tienen un gasto medio superior respecto a los que no lo utilizan (0)



Clientes que tienen tienen tarjeta son más propensos a utilizar el cupón en comparación a los que no tienen tarjeta

Estimación de Parámetros

- Para entrenar el modelo utilizamos el 70% de los datos (22 datos), y el resto (10 datos) lo vamos a utilizar para evaluar.
- Con estos datos la ecuación de regresión logística tiene la siguiente forma:

$$P(Y = 1|X_1, X_2) = \frac{e^{-4.2587+0.6863X_1+3.4334X_2}}{1 + e^{-4.2587+0.6863X_1+3.4334X_2}}$$

- Si se quiere predecir la probabilidad de que un cliente use el cupón, entonces tendremos que reemplazar en la ecuación el gasto anual y si tiene o no tarjeta de crédito.
- Considere un cliente con gasto anual X_1 = 2.64 millones y Si tiene tarjeta de crédito X_2 = 1. La probabilidad de que use el cupón es:

$$P(Y = 1|X_1, X_2) = \frac{e^{-4.2587+0.6863(2.64)+3.4334(1)}}{1 + e^{-4.2587+0.6863(2.64)+3.4334(1)}} = 0.7284$$

- Esto indica que el cliente tiene una probabilidad de 0.7284 de utilizar el cupón.
- Cuando la función da como resultado una probabilidad igual o superior a 0.5, entonces se asigna la categoría y = 1.

Esta ecuación se obtiene con el *método de máxima verosimilitud*.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

Las estimaciones β_0 y β_1 se eligen para maximizar esta función de verosimilitud.

Cociente de Posibilidad (Odds Ratio)

- Otra forma de interpretar el resultado es haciendo uso del cociente de posibilidades (odds ratio):

$$odds = \frac{P(Y=1)}{1-P(Y=1)}$$

- Para la predicción que se realizó anteriormente el odds:

$$odds = \frac{0.7284}{1-0.7284} = 2.68$$

- Esto indica que el cliente tiene 2.68 veces más de posibilidades de usar el cupón que de no usarlo.
- Esto también se puede calcular como:

$$odds = e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_pX_p} = e^{-4.2587+0.6863(2.64)+3.4334(1)} = 2.68$$

- Al aplicar logaritmo natural a la ecuación para el odds, obtenemos:

$$\log(odds) = \beta_0 + \beta_1X_1 + \beta_2X_2 = -2.5813 + 0.4288X_1 + 2.6628X_2$$

- Por tanto, los parámetros de la regresión logística β_1 y β_2 se interpretan como el cambio que se produce en el log(odds) cuando X_1 o X_2 cambian en una unidad, siendo que las demás variables permanecen constantes.
- Observe que el uso de la tarjeta de crédito tiene un mayor efecto en el log(odds) para uso del cupón, en comparación a un incremento en 1 millón de pesos en el gasto anual.

Matriz de Confusión

Pontificia Universidad JAVERIANA Bogotá

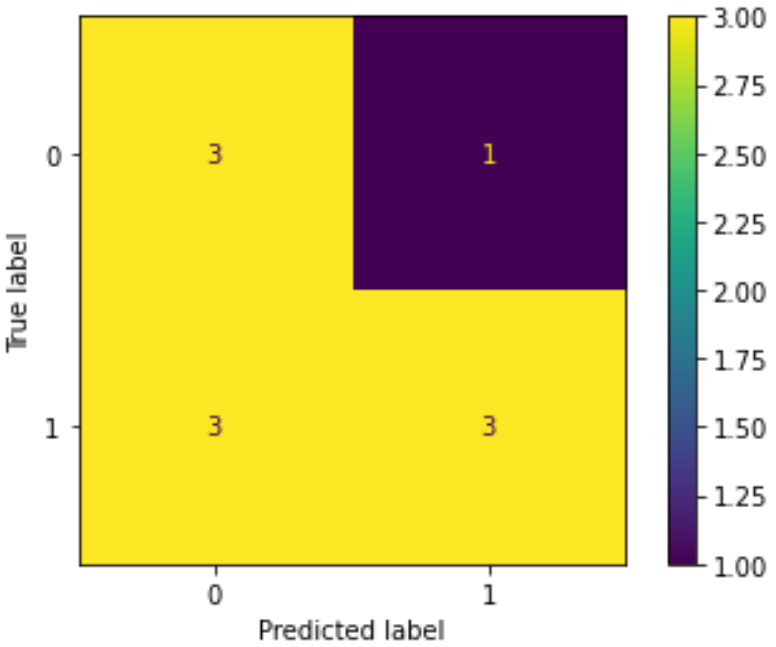
Al utilizar el modelo surge la pregunta:
¿Qué tanto acierta el clasificador en las predicciones?
Para responder a esta pregunta se utiliza la matriz de confusión, la cual resume los resultados de las predicciones.

		Predicción	
		1: Positivo	0: Negativo
Clase Real	1: Positivo	Verdadero Positivo VP	Falso Negativo FN
	0: Negativo	Falso Positivo FP	Verdadero Negativo VN

Medida de Desempeño	Formula
Exactitud (Accuracy): proporción de predicciones acertadas	$\text{Exactitud} = \frac{VP + VN}{VP + FP + FN + VN}$
Recuperación (Recall): proporción predicciones positivas correctas, con respecto al total de clases reales positivas.	$\text{Recuperación} = \frac{VP}{VP + FN}$
Precisión : proporción de predicciones positivas correctas, con respecto al total de predicciones.	$\text{Precisión} = \frac{VP}{VP + FP}$
F-1 Score: se calcula como la media harmónica entre el Recall y la Precisión.	$\text{F1-Score} = 2\frac{\text{Precisión} \times \text{Recuperación}}{\text{Precisión} + \text{Recuperación}}$

Predicciones para el Cojunto de Prueba

Cliente	Gasto Anual	Tarjeta	Utiliza Cupón	Probabilidad	Predicción
12	1.92	0	0	0.0502	0
11	5.55	0	0	0.3895	0
4	3.92	0	1	0.1725	0
9	1.18	1	1	0.4961	0
27	1.27	1	0	0.5116	1
25	2.64	1	1	0.7284	1
20	6.28	0	1	0.5129	1
29	4.80	0	0	0.2760	0
10	3.35	0	1	0.1235	0
31	5.53	1	1	0.9512	1



$$Accuracy = \frac{3+3}{3+1+3+3} = 0.6$$ En solo en 6 de las 10 predicciones el modelo acertó (60%)

$$Sensibilidad = \frac{3}{3+3} = 0.5$$ De los 6 clientes que usaron el cupón, el modelo detectó 3 (50%)

$$Precisión = \frac{3}{3+1} = 0.75$$ El modelo predijo que 4 clientes usarían el cupón, pero solo 3 lo hicieron (75%)