**REGRESIÓN LOGISTICA**

Es un modelo de para resolver problemas de clasificación no de regresión; predice probabilidades, por ejemplo: alguien compra: si o no, tiene una enfermedad 1 o 0.

La función logística (sigmoide) se usa para transformar valores en un rango de 0 a 1. Si el resultado es mas cercano a 1 es posible que sea de la clase positiva y su es cercano a 0 es más probable que sea de la clase negativa.

**NOTA1**: transforma cualquier número real en un valor entre 0 y 1, su gráfica tiene una forma de s

σ(z)=1/1+e−z​

Una vez que se obtiene la probabilidad, aplicamos un umbral (threshold) por defecto = 0.5, pero puede ajustarse según el problema.

**NOTA2**: el umbra o threshold es el valor que usamos para predecir a que clase pertenece una observación

**Interpretación de los coeficientes:**

**NOTA3:** como log-odds : los odds -> si una probabilidad de éxito es p, los odds se definen como:

Odds = p/(1-p), es decir la razón entre la probabilidad de éxito sobre probabilidad de fracaso. Visto de otra manera es la relación entre el éxito y fracaso

El log-odds es simplemente el logaritmo de los odds, es logaritmo convierte los odds en un rango qye va de -infinito hasta infinito.

En regresión logística esto se modela como: log(p/1−p​)=β0​+β1​X1​+β2​X2​+⋯+βn​Xn, esto no predice la probabilidad, predice los log-odds

Estos se interpretas como log-odds (logaritmo de la razón de probabilidades).

* Un coeficiente positivo -> aumenta la probabilidad de que Y = 1
* Un coeficiente negativo -> disminuye la probabilidad de que Y = 1

Ejemplo si B\_1 = 0.7 en un modelo de compra de un producto según el ingreso, cada unidad extra de ingreso (X) incrementa los log-odds de comprar en 0.7

**Métricas de evaluación:**

* Accuracy (exactitud): porcentaje de predicciones correctas. Proporción de aciertos totales del modelo, pero solo es confiables si las clases están balanceadas. La fórmula es:

Accuracy = # de predicciones correctas/ # total de predicciones. En términos de matriz de confusión:

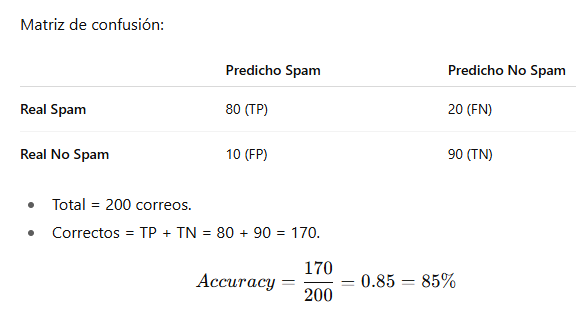
Acurracy = (TP+TN)/(TP+TN+FP+FN)

TP: +correctamente clasificados

TN: - correctamente clasificados

FP: - que el modelo clasificó como + (error tipo i)

FN: + que el modelo clasificó como – (error tipo II)



Fácil de interpretar y útil cuando las clases están equilibradas (=cantidad de + y -). Engañoso si las clases están desbalanceadas

* Precisión: de todos los positivos predichos, cuántos son realmente positivos; el % que son realmente +

Precisión = TP / (TP+FP).

En otras palabras “cuando el modelo dice que es + ¿Cuántas veces tienen la razón?

Precisión alta -> pocas falsas alarmas

Baja precisión -> el modelo etiqueta muchos – como +

La precisión es importante donde los falso + son muy costosos, ejemplo, fraude bancario, diagnóstico médico

* Recall (sensibilidad o TRP): de todos los positivos reales, cuántos fueron detectados. En otras palabras: de todos loe elementos que sí eran + en la realizada ¿cuántos fue capaz de detectar el modelo?

Recall = TP/(TP+FN)

Recall alto (cercano a 1) -> pocos falsos -, el modelo casi nunca deja escapar + reales

Recall bajo -> muchos falsos -, el modelo falla al detectar casos +.

El recall es crucial cuando es más costos pasar por alto un + que equivocarse clasificando – como +

* F1-score: balance entre precisión y reacall. Se utiliza cuando se quiere una medida única que combine ambas y evitar enfocarse solo en una; es la armonía entre la precisión y el recall

F1 = 2\* (precisión \* Recall)/(Precisión + Recall)

F1 = 1 -> precisión y recall perfectos

F1 = -> el modelo es muy malo en balancear la precisión y recall

Útil cuando se tiene clases desbalanceadas

Se usa cuando no se puede confiar solo en el accuracy ( por datos desbalanceados)

* Matriz de confusión: tabla que muestra TP, TN, FP, FN
* ROC curve y AUC: evalúa la capacidad del modelo para discriminar entre clases

El ROC es una gráfica que muestra la capacidad de un modelo de clasificación para separar las clases en distintos umbrales, la gráfica sería:

X (FPR -> Flase Positive Rate); FPR = FP/(FP+TN)

Y (TPR -> True Positive Rate = Recall); TPR = TP /(TP + FN)

La curva ROC se obtiene variando el umbral de decisión (por defecto 0.5)

El modelo perfecto se acerca a la esquina superior izquierda (TPR = 1, FPR=0)

AUC (área bajo la curva del ROC) mide en un solo número que tan buena es la curva ROC

AUC = 1.0 -> modelo perfecto

AUC = 0.5 -> modelo inútil (predice al azar

AUC < 0.5 -> modelo peor que el azar

AUC presenta la probabilidad que el modelo asigne una mayor puntuación a un + real que a un - real