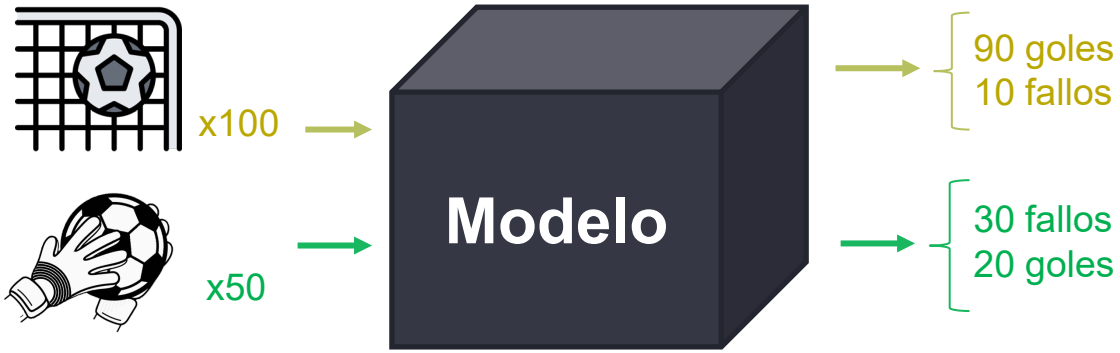
Métricas de desempeño de modelos de clasificación

Rafael Zambrano
rafazamb@gmail.com

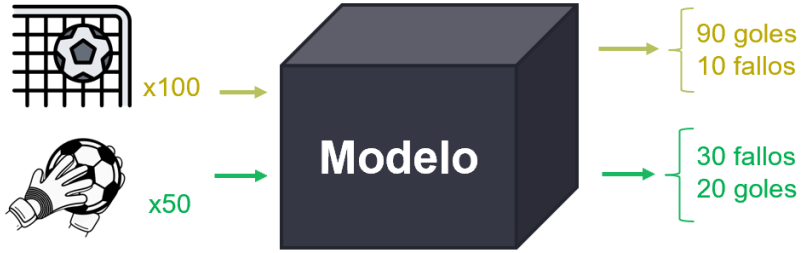
Evaluación de modelos de clasificación

- **Ejemplo**: Clasificador de goles



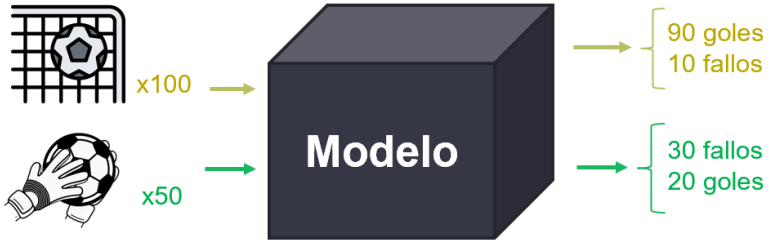
Evaluación de modelos de clasificación: Matriz de confusión

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)



		REALIDAD	
		GOLES	FALLOS
PREDICCIÓN	GOLES	TP 90	FP 20
	FALLOS	FN 10	TN 30

Evaluación de modelos de clasificación



		REALIDAD	
		GOLES	FALLOS
PREDICCIÓN	GOLES	TP 90	FP 20
	FALLOS	FN 10	TN 30

- **Accuracy**: En general, ¿cómo de bueno es el clasificador? (TP+TN)/Total = (90+30)/150= 80%
- **Precision**: Cuando predice "SÍ", ¿cuántas veces acierta? TP/(TP+FP) = 90/(90+20) = 82%
- **Recall**: En los casos reales de "SÍ", ¿cuánto predice correctamente? TP/(TP+FN) = 90/(90+10) = 90%
- **Specificity**: En los casos reales de "NO", ¿cuánto predice correctamente? TN/(TN+FP) = 30/(30+20)=60%
- **F1 Score:** Combina precision y recall en una sola métrica $\in[0,1]$: (2*Precision*Recall)/(Precision+Recall) = 0.81
- **Ratio de falsos positivos**: 1 – Specificity = FP/(TN+FP) = 1-0.6 = 40%
- **Mejora (Lift):** ¿Cómo mejora el modelo a una decisión aleatoria?

$$\text{Prior} = P(\text{moto}) = 100/150 = 66\%$$

$$\Longrightarrow \text{Mejora} = \text{Precision}/\text{Prior} = 82/66 = 1.24$$

Precision vs Recall

- **Detección de cáncer**: ¿precision o recall? **RECALL**(evitar falsos negativos)

Es preferible decir que una persona padece cáncer cuando no es así, que decirle a una persona que no padece cáncer cuando en realidad sí lo padece.

- **Detección de SPAM**: ¿precision o recall? **PRECISION**(los falsos negativos no son preocupantes)

No pasa nada si un correo spam no se detecta, pero si un email no es spam, no queremos que vaya a la carpeta de spam. Es decir, si el modelo predice SPAM, tiene que estar muy seguro de ello.

Evaluación de modelos de clasificación

Ejemplo: Detección de fraude en tarjetas. De 1000 clientes, solo 10 cometieron fraude. Se entrena un modelo y se prueba con el conjunto de test, obteniendo la siguiente matriz de confusión

		FRAUDE	NO FRAUDE
PREDICCIÓN	FRAUDE	0	0
	NO FRAUDE	10	990

- **Accuracy**: En general, ¿cómo de bueno es el clasificador? (TP+TN)/Total = (0+990)/1000= 99%
- **Precision**: Cuando predice "SÍ", ¿cuántas veces acierta? TP/(TP+FP) = 0/(0+0) = 0%
- **Recall**: En los casos reales de SÍ, ¿cuánto predice correctamente? TP/(TP+FN) = 0/(0+10) = 0%
- **Specificity**: En los casos reales de NO, ¿cuánto predice correctamente? TN/(TN+FP) = 990/(990+0)=100%
- **F1 Score:** Combina precision y recall en una sola métrica $\in[0,1]$: (2*Precision*Recall)/(Precision+Recall) = 0
- **Ratio de falsos positivos**: 1 – Specificity = FP/(TN+FP) = 1-0 = 100%
- **Mejora:** ¿Cómo mejora el modelo a una decisión aleatoria?

  Prior = P(fraude) = 10/1000 = 1%

  $\Longrightarrow$Mejora = Precision/Prior = 0/1 = 0

Otro ejemplo

Medición del desempeño de un clasificador

- La matriz de confusión es una de las herramientas básicas de medición del desempeño de un modelo de clasificación:

	Real:	
Predicción:	**Heavy User**	**No Heavy User**
Heavy User	4	1
No Heavy User	20	75

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

- ¿El modelo es bueno?

Precision — % de positivos del modelo que son correctos

Recall — % de positivos reales que detecta el modelo

Ejemplo (2)

- *Precision = 4/(4+1)=80%*
- *Recall = 4/(4+20)=17%*
- *...No está teniendo un buen desempeño. Sólo es capaz de detectar un 17% de los positivos*

Evaluación de modelos de clasificación: Curva ROC

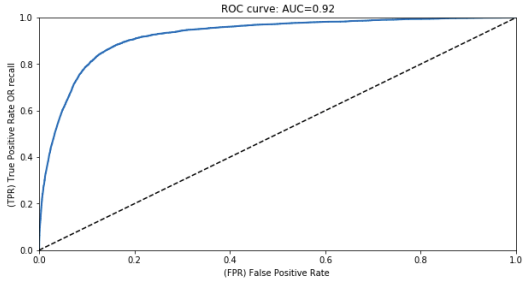
- En los modelos de clasificación binarios, existe un compromiso entre el error de falsos positivos y el de falsos negativos, pudiendo aumentar uno para disminuir el otro, y viceversa.

- Ejemplos:

    - Quiero que mi modelo detecte todos los fraudes de tarjetas: habrá muchos falsos positivos (baja precisión) y pocos falsos negativos (mayor recall)

    - Quiero que mi modelo detecte solo los casos reales de fraude de tarjetas: habrá muchos falsos negativos (bajo recall) y pocos falsos positivos (alta precisión)

- La **curva ROC** relaciona el recall con el ratio de falsos positivos

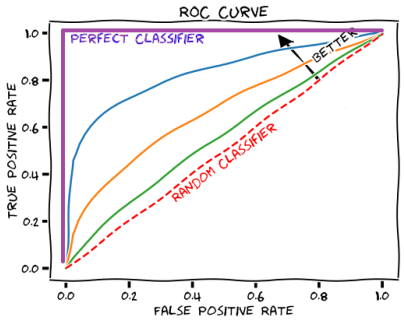


Evaluación de modelos de clasificación: Curva ROC

- En las curvas ROC, nos interesa que la curva se acerque lo máximo posible a la esquina superior izquierda de la gráfica, de manera que el hecho de aumentar el recall no haga que nuestro modelo introduzca más falsos positivos.

- En este caso también podemos calcular el ROC AUC (área bajo la curva), que también nos sirve como métrica para resumir la curva y poder comparar modelos.

  - AUC = 1: Clasificador perfecto
  - AUC = 0.5: Clasificador aleatorio



¡Gracias!

Contacto: Rafael Zambrano

rafazamb@gmail.com