

# Predicción de bajas de clientes

en empresa de Telecomunicaciones

XABI DEL REY  
Julio 2025 - Vitoria

PROYECTO MACHINE LEARNING  
BOOTCAMP DATA SCIENCE  
THE BRIDGE

# Dataset

publicado por IBM Cognos Analytics

en el que se organiza la información de **clientes** de una empresa de telecomunicaciones.

**7043 registros**  
**21 columnas**

## Información

2

**Servicios contratados** Detalles sobre los servicios a los que cada cliente está suscrito, como

- Teléfono
- Internet
- Almacenamiento en la nube
- Soporte técnico
- Servicios de streaming (TV y películas)...

### Información de la cuenta

- Antigüedad del cliente
- Tipo de contrato
- Método de pago
- Facturación total...

### Información demográfica

- Género
- Senior
- Si tienen pareja...

**Churn** / Bajas

**"Adquirir un nuevo cliente es hasta 25 veces más caro que retener uno existente."**

## TARGET: **CHURN**

Nos indica si el cliente ha causado baja en el último mes.

Tasa de Abandono

Rotación de clientes

Cancelaciones de servicio

Bajas de clientes

CAUSA BAJA

**27%**

Clase Minoritaria

NO CAUSA BAJA

**73%**

Métrica que indica el porcentaje de clientes que dejan de serlo para una empresa durante un período de tiempo determinado.

Especialmente importante para negocios basados en suscripciones.

### **OBJETIVO:**

Queremos detectar todos los Positivos.

Minimizar los Falsos Negativos.

**RECALL  $\geq$  0.85**

# Proceso

## Análisis Exploratorio



Correlaciones

Visualizaciones distribución de variables

**Entender el desequilibrio del Target : Habitual**

## Limpieza y Feature Engineering



Eliminar variables, nulos, vacíos, duplicados...

Transformaciones de binarias y categóricas

Normalización de columnas numéricas

## Visualizaciones



Correlaciones

Distribución

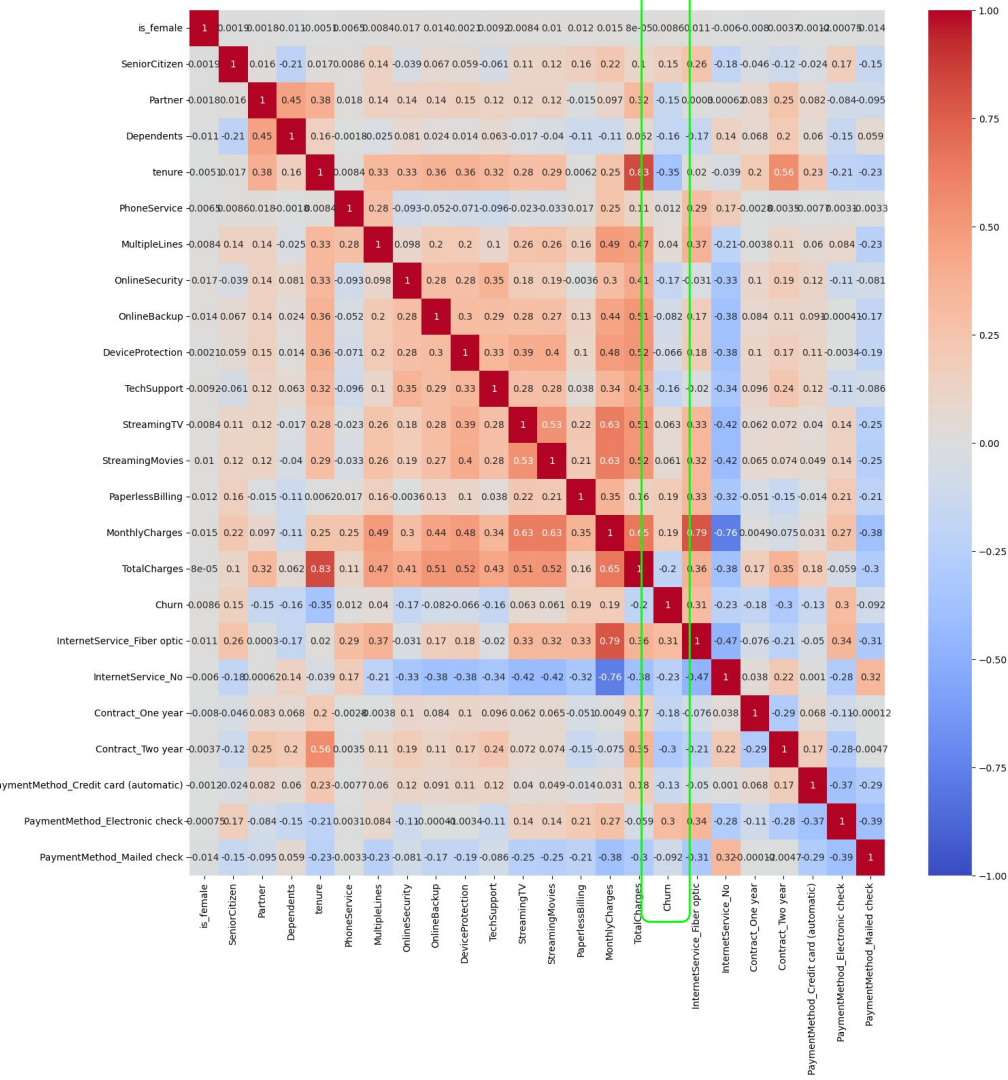
## Modelos



Probar diferentes modelos

Evaluarlos y ajustarlos

Objetivo: Recall



## Correlaciones

### Entre variables:

- Tipo de conexión a Internet (DSL, fibra o sin internet) determina en gran medida la factura mensual del cliente.
- 'TotalCharges' con 'tenure' con 0.83

### Churn:

- **No hay ninguna variable que destaque con una alta correlación con el 'Churn'.**
- Parece que si el cliente tiene Servicio de Internet contratado tendrá mayor probabilidad de producir baja.
- El método de pago ('PaymentMethod') también tiene una ligera correlación con el abandono del cliente.
- 'tenure' (con -0.35) y 'Contract' tienen una correlación negativa, lo que sugiere que los clientes antiguos o con contratos largos tienen un menor 'Churn'.

*No hay ninguna variable que por sí sola explique el abandono de los clientes.*

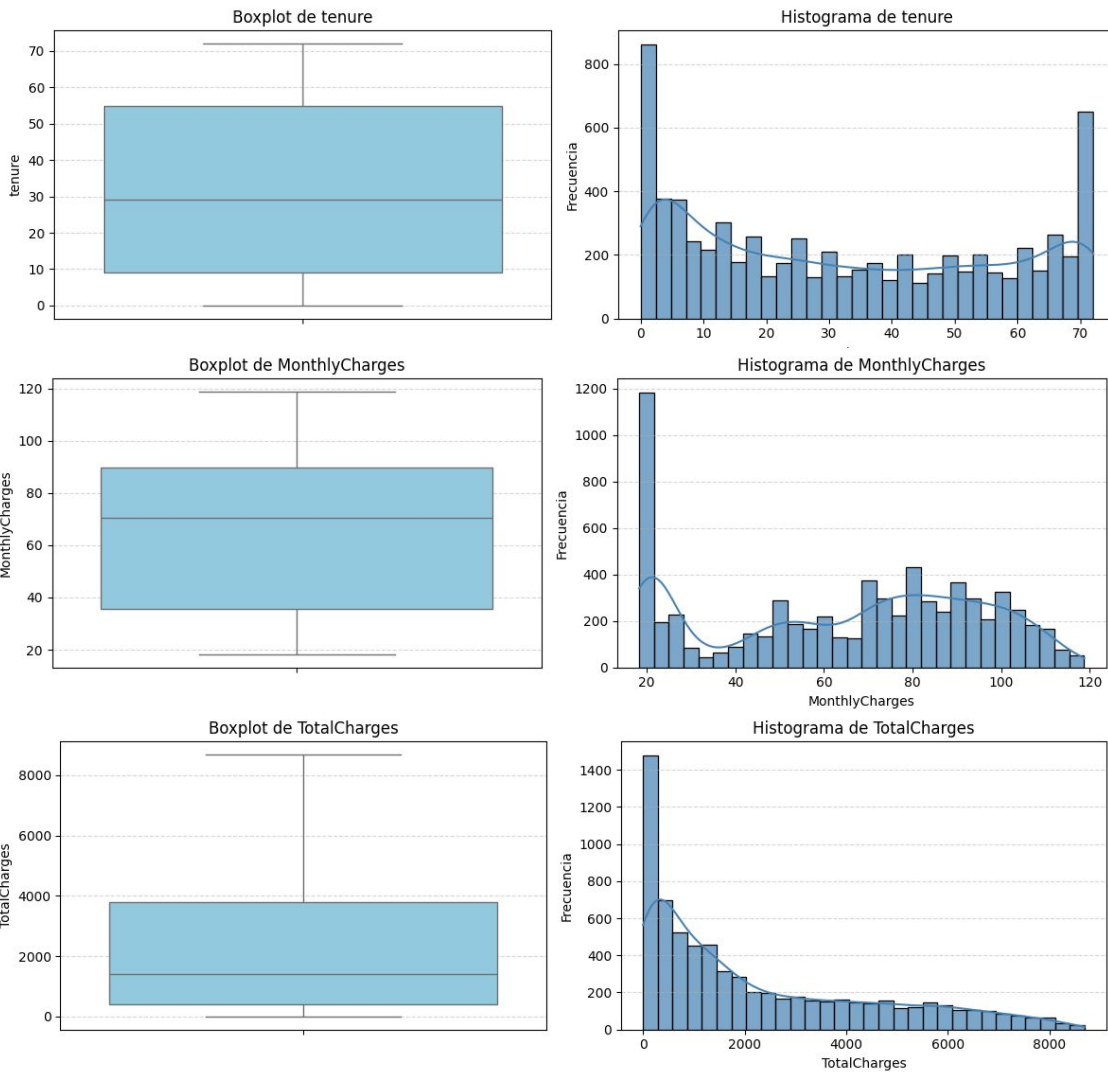
## Variables numéricas:

Distribuciones normales. Se explican con facilidad.

Aparentemente no hay valores fuera de lo que se puede considerar natural. Visualmente no se aprecian *outliers*.

Sin embargo, aunque no parece necesario eliminar registros, convendría transformar estos valores **escalándolos**.

De esta manera, minimizaremos las dificultades en trabajarlos que podrían encontrar algunos modelos.



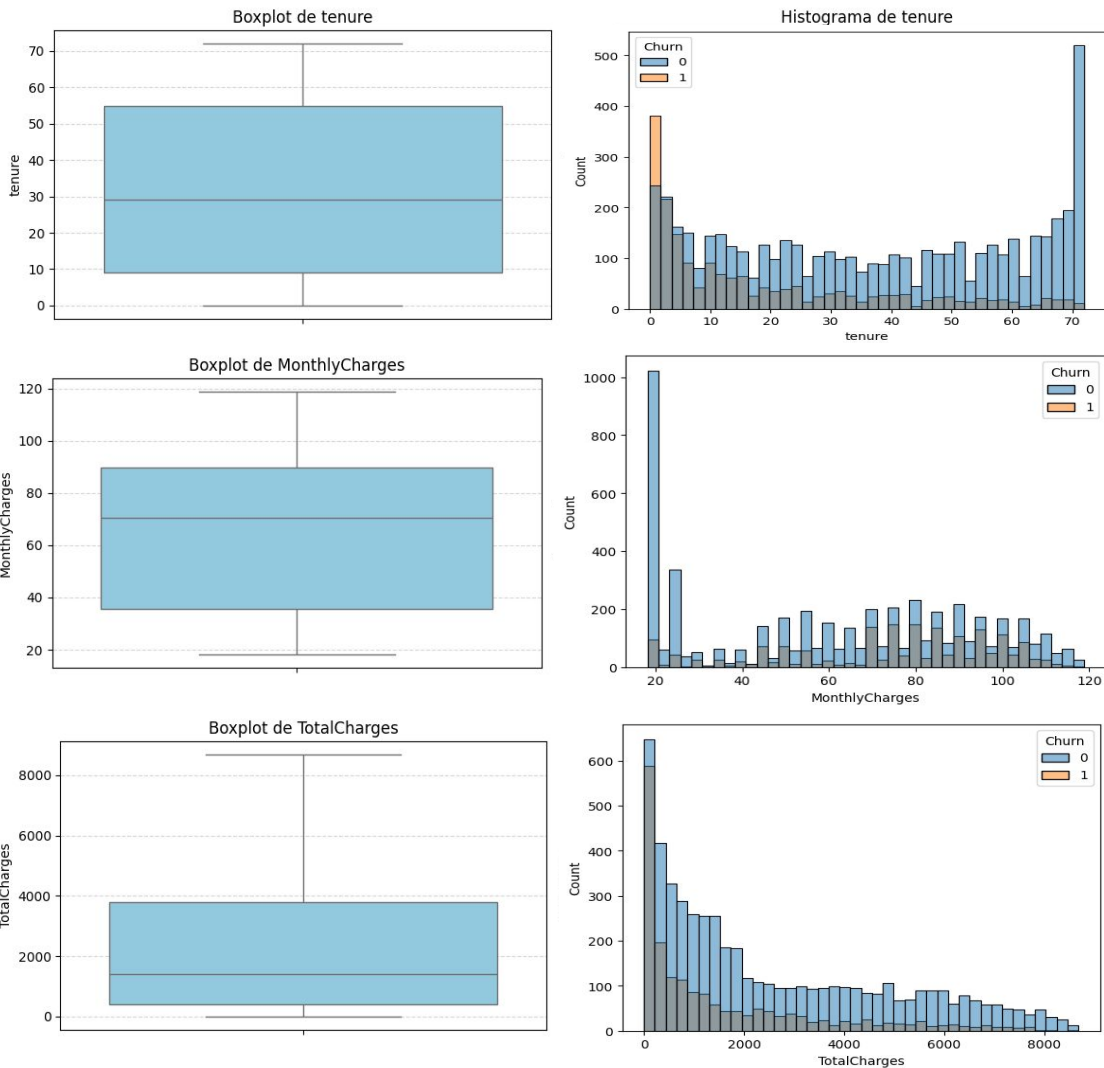
## Variables numéricas:

Distribuciones normales. Se explican con facilidad.

Aparentemente no hay valores fuera de lo que se puede considerar natural. Visualmente no se aprecian *outliers*.

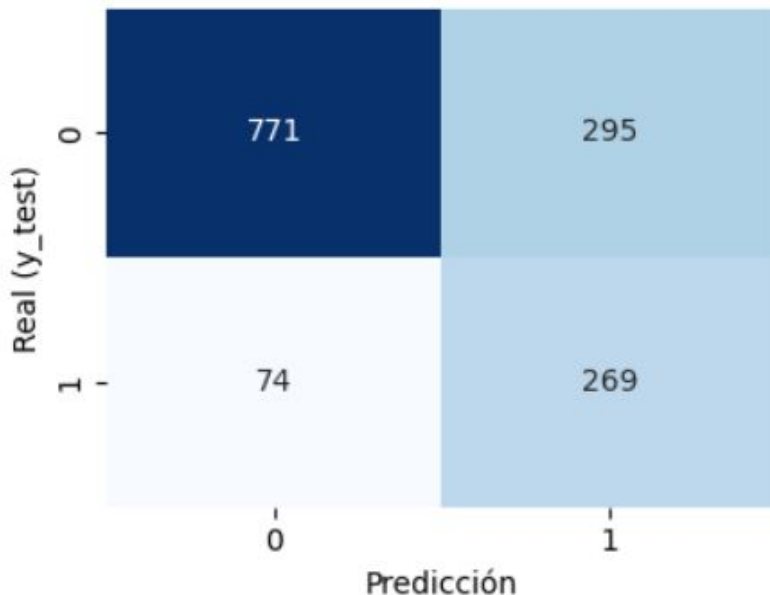
Sin embargo, aunque no parece necesario eliminar registros, convendría transformar estos valores **escalándolos**.

De esta manera, minimizaremos las dificultades en trabajarlos que podrían encontrar algunos modelos.



# MODELO 1: REGRESIÓN LOGÍSTICA

Matriz de Confusión  
Modelo 1: Regresión Logística



AUC Score: 0.843

Precisión (Churn): 0.477

Recall (Churn): 0.784

Exactitud (Accuracy): 0.738

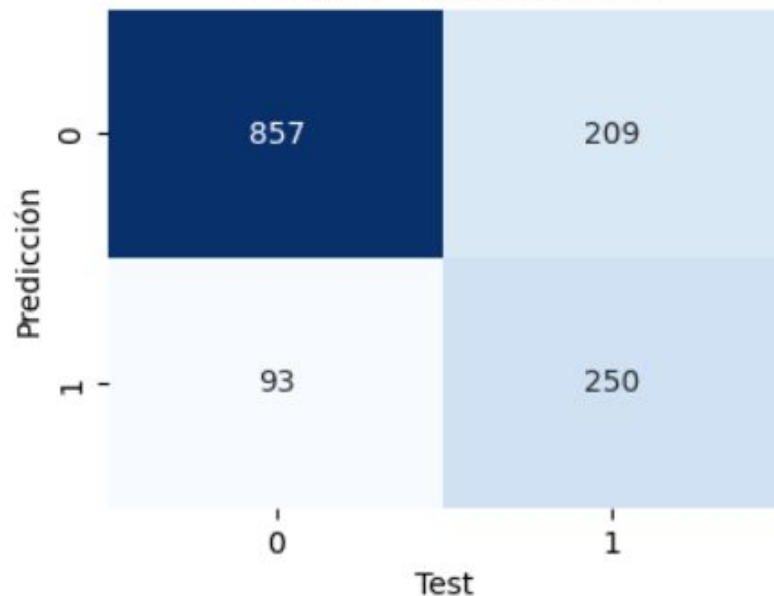
Reporte de Clasificación:

	precision	recall	f1-score	support
No Churn	0.91	0.72	0.81	1066
Churn	0.48	0.78	0.59	343
accuracy			0.74	1409
macro avg	0.69	0.75	0.70	1409
weighted avg	0.81	0.74	0.75	1409



## MODELO 2: **RANDOM FOREST**

Matriz de Confusión  
Modelo 2: Random Forest



AUC Score: 0.848

Precisión (Churn): 0.545

Recall (Churn): 0.729

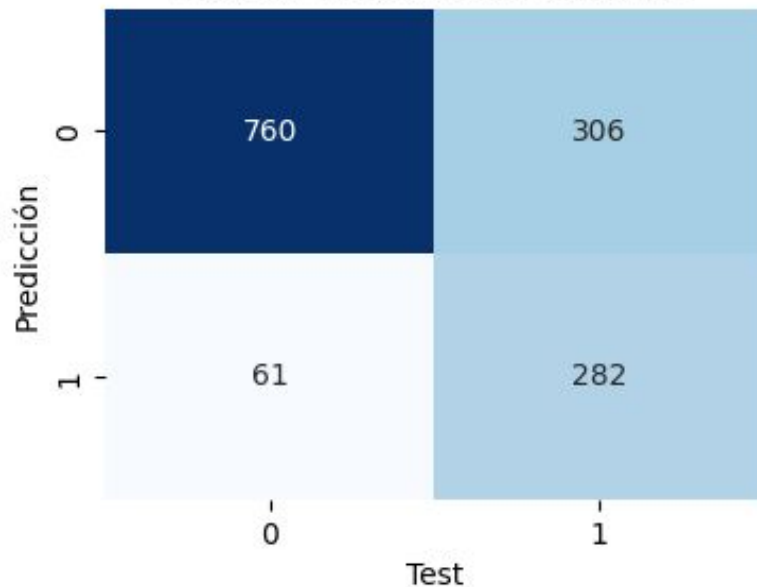
Exactitud (Accuracy): 0.786

Reporte de Clasificación:

	precision	recall	f1-score	support
No Churn	0.90	0.80	0.85	1066
Churn	0.54	0.73	0.62	343
accuracy			0.79	1409
macro avg	0.72	0.77	0.74	1409
weighted avg	0.82	0.79	0.79	1409

## MODELO 3: **XG BOOST**

Matriz de Confusión  
Modelo 3: XGBoost Classifier



AUC Score: 0.846

Precisión (Churn): 0.480

Recall (Churn): 0.822

Exactitud (Accuracy): 0.740

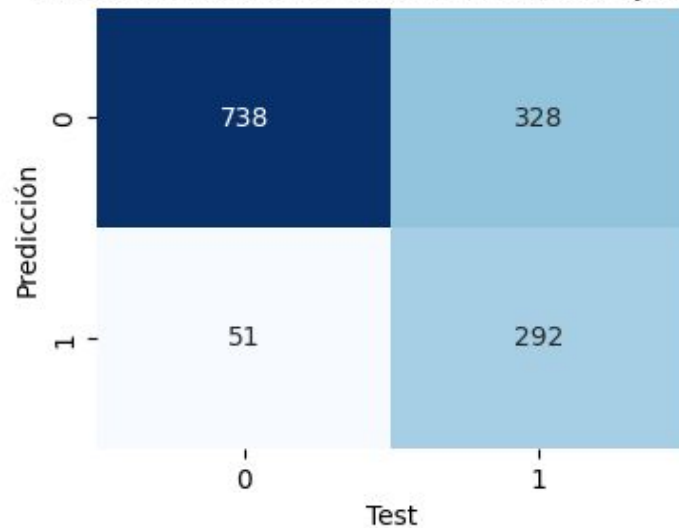
Reporte de Clasificación:

	precision	recall	f1-score	support
No Churn	0.93	0.71	0.81	1066
Churn	0.48	0.82	0.61	343
accuracy			0.74	1409
macro avg	0.70	0.77	0.71	1409
weighted avg	0.82	0.74	0.76	1409

## MODELO 4: **XG BOOST**, THRESHOLD AJUSTADO

Matriz de Confusión

Modelo Final: XGBoost con threshold ajustado



AUC Score: 0.846

Precisión (Churn): 0.471

Recall (Churn): 0.851

Exactitud (Accuracy): 0.731

Reporte de Clasificación:

	precision	recall	f1-score	support
No Churn	0.94	0.69	0.80	1066
Churn	0.47	0.85	0.61	343
accuracy			0.73	1409
macro avg	0.70	0.77	0.70	1409
weighted avg	0.82	0.73	0.75	1409

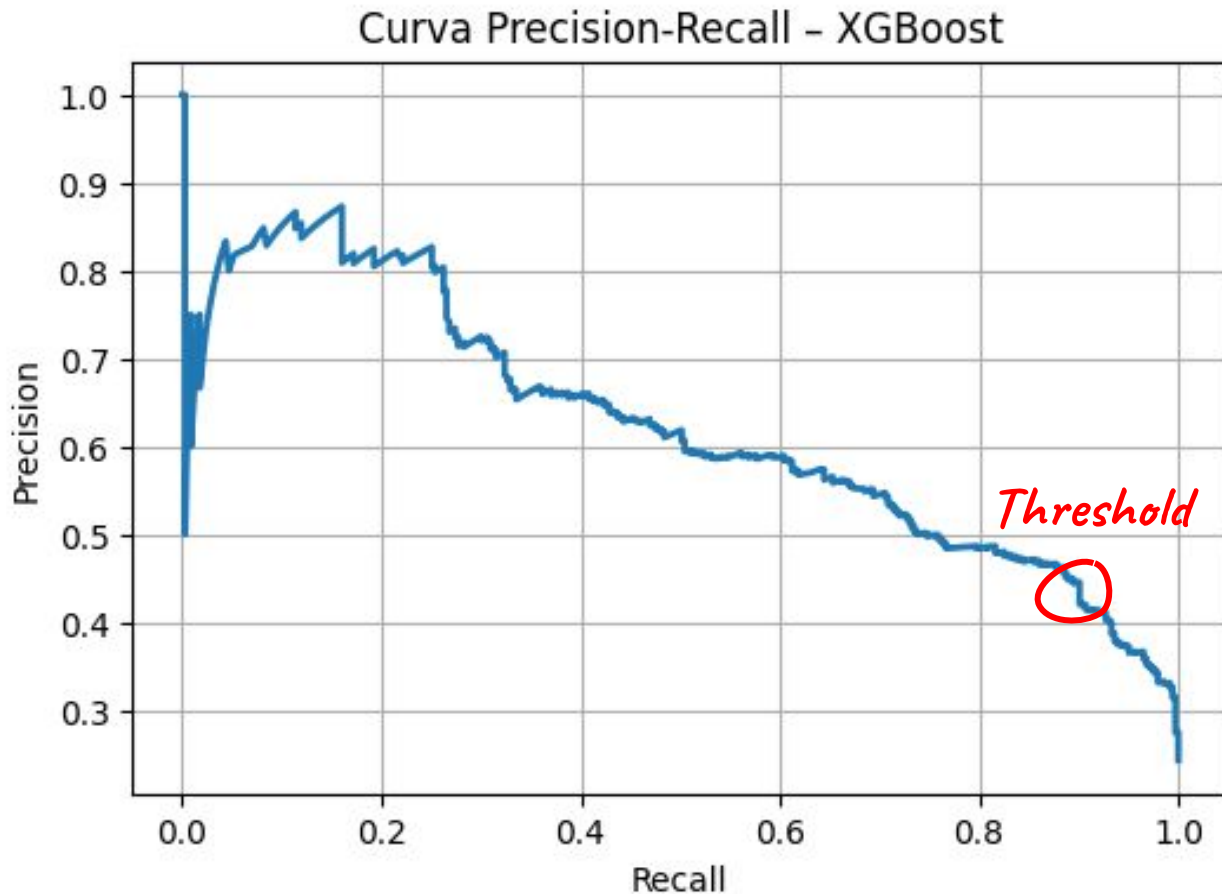
## MODELO 3: **XG BOOST**

### **RECALL $\geq 0.85$**

Para lograr el  $\text{Recall} \geq 0.85$  que nos interesa, vamos a tener que aceptar una `Precision` cerca de 0.45.

**Eso significa que aprox. el 55 % de las detecciones serán "falsas alarmas".**

Confirmaremos con "Negocio" si el coste de la gestión de esos clientes (potenciales Churn que no vamos a predecir satisfactoriamente) es asumible frente al beneficio de retener a los clientes que realmente iban a causar baja.



# Gracias!

XABI DEL REY  
Julio 2025 - Vitoria

PROYECTO MACHINE LEARNING  
BOOTCAMP DATA SCIENCE  
THE BRIDGE