

# TRABAJO FINAL

## BIG DATA



# OBJETIVO

Predecir qué clientes tienen más probabilidad de responder a la próxima campaña.

## ¿Por qué es importante?

Apuntar a las personas correctas nos permite:

- ✓ Optimizar recursos de marketing
- ✓ Mejorar el resultado de futuras campañas
- ✓ Ahorrar dinero



# TIPOS DE VARIABLES

Demográficas	Comportamiento de compra	Canal de compra	Respuestas a Campañas	Relación con la empresa
Year_Birth Education Marital_Status Income	MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds	NumDealsPurchases NumCatalogPurchases NumWebPurchases NumStorePurchases NumWebVisitsMonth	AcceptedCmp1 AcceptedCmp2 AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 Response, Recency	Complain, Dt_Customer

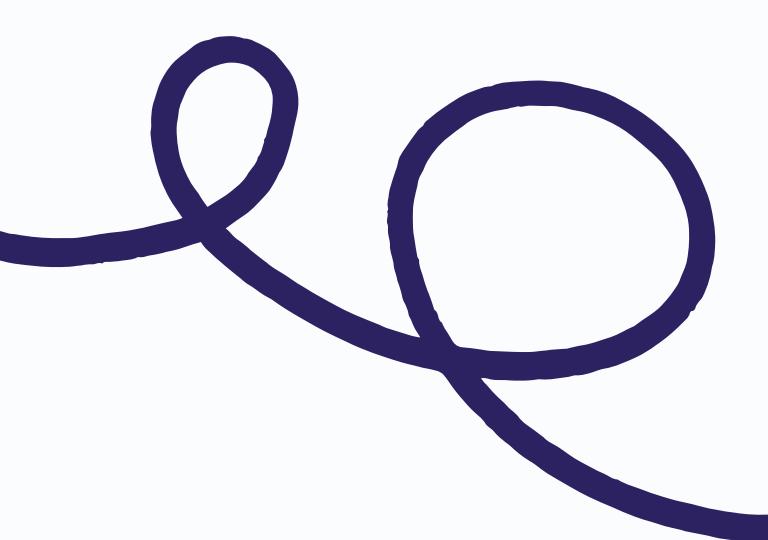
# ¿QUÉ VARIABLE QUEREMOS PREDICIR?

Variable target  
**RESPONSE**

# HIPÓTESIS

“Los clientes con mayor frecuencia de compras, mayor respuesta a campañas e ingreso tienen más probabilidad de responder a la próxima campaña.”





# ANÁLISIS EXPLORATORIO DE DATOS

## EDA



01

### Feature Statistics

Nos permitió ver estadísticas básicas de cada variable (promedios, mínimos, máximos y variación). Esto sirve para entender cómo se comportan los datos y detectar posibles valores raros o fuera de lo común.

02

### Distribution/Box plot/Violin plot

Estos gráficos nos permitieron analizar cómo se distribuyen las variables y comparar grupos. Con el Box Plot y el Violin Plot vimos, por ejemplo, cómo varía la respuesta a la última campaña según el nivel de ingresos, compras y frecuencia de la misma.

03

### Correlations

Este análisis mostró qué variables están más relacionadas con la respuesta a la campaña. Nos ayudó a identificar cuáles características de los clientes podrían influir más en la probabilidad de responder.

# PRINCIPALES VARIABLES



01

## Frecuencia de compra

La frecuencia de compra funciona como un indicador de fidelidad. Los clientes fieles suelen tener mayor tasa de respuesta frente a acciones comerciales.

02

## Nivel de ingresos

A mayor nivel de ingresos, mayor probabilidad de que el cliente responda a la campaña ya que tiene mayor capacidad de compra.

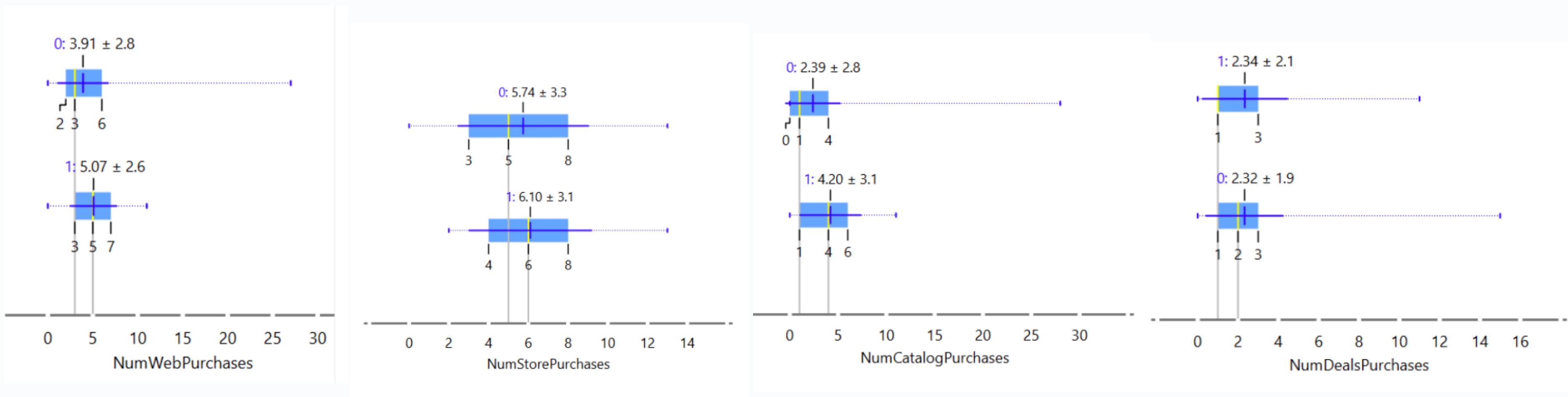
03

## Respuesta en campañas anteriores

Los clientes que ya respondieron en el pasado tienen una mayor probabilidad de volver a hacerlo, porque ya conocen la marca y mostraron interés real.

# ANÁLISIS EXPLORATORIO DE DATOS

## Boxplot



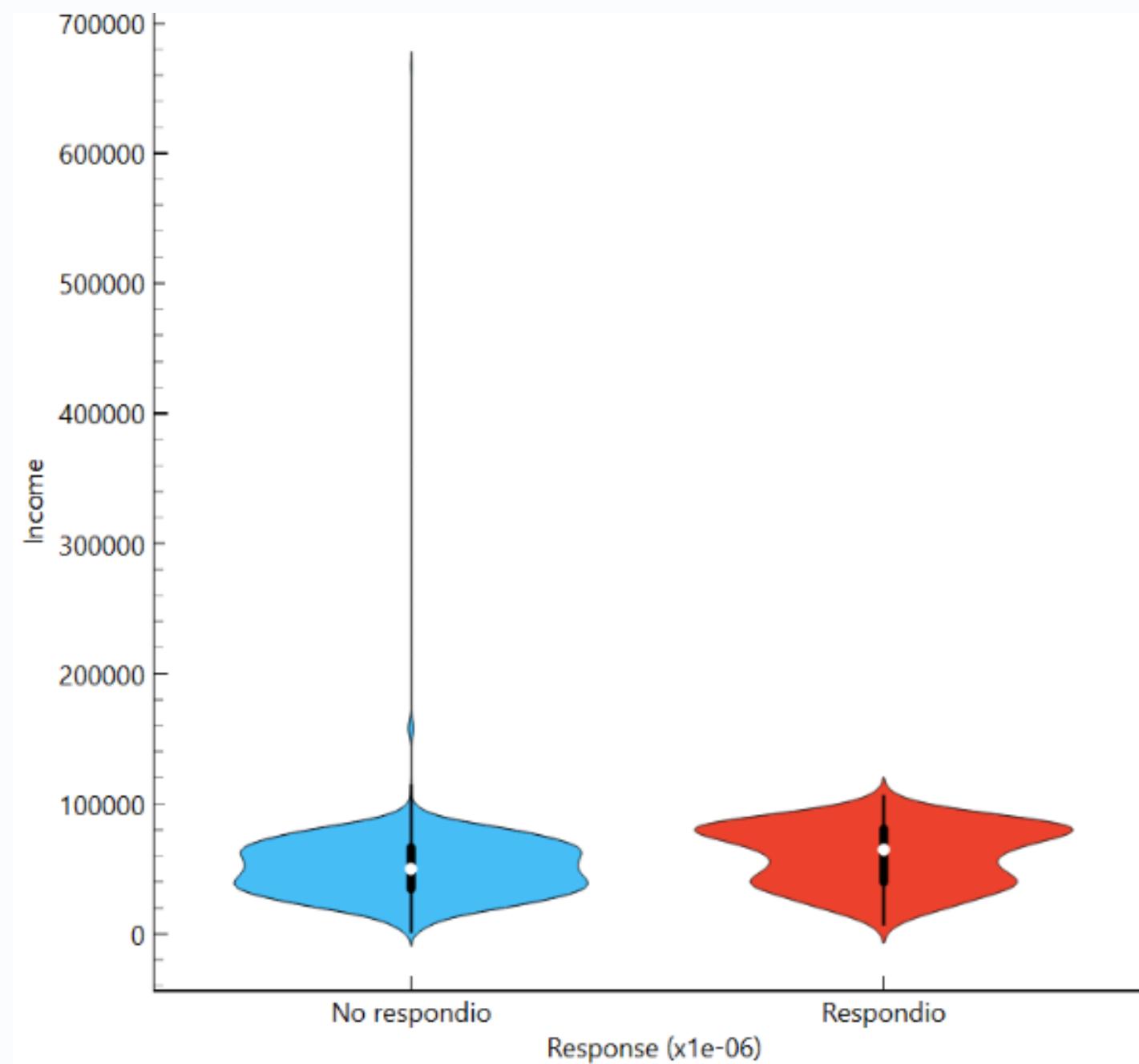
Los clientes que respondieron tuvieron un promedio más alto de compras que los que no respondieron. Esto muestra que quienes tienen mayor actividad de compra o mayor relación con la marca son los más propensos a responder.

# ANÁLISIS EXPLORATORIO DE DATOS

## Violin Plot

*Los clientes que respondieron a la campaña tienen, en promedio, un nivel de ingreso mayor que quienes no respondieron.*

*Esto sugiere que el ingreso influye en la probabilidad de respuesta, ya que quienes tienen mayor capacidad económica están más dispuestos a realizar la compra.*



# PREPARACIÓN DE DATOS

## EDIT DOMAIN

Permite definir qué tipo de dato es cada variable (numérica, categórica, fecha, etc.) y cuál es la variable objetivo del modelo.

*Transformamos response de numérica a categorica.*

## SELECT COLUMNS

Permite elegir qué columnas queremos usar y eliminar las que no aportan información.

*Nos quedamos solo con las relevantes para el análisis.*

## PREPROCESS

Se encargan de preparar los datos para que los modelos puedan trabajar correctamente.

*Tratamos valores faltantes.*

## DATA SAMPLER

Se encarga de dividir los datos entre aquellos que van a ir a training y los que van a ir al test.

*Utilizamos la proporcion de 80% para el training y el 20% para el test*

# TRAINING

## Modelos de aprendizaje

### LOGISTIC REGRESION

Predice la probabilidad de respuesta (sí/no) en función de las variables del cliente.  
Es un modelo estadístico simple y fácil de interpretar.

### NAIVE BAYES

Calcula la probabilidad de que un cliente responda, suponiendo independencia entre las variables. Es rápido y funciona bien con muchos datos.

### RANDOM FOREST

Utiliza muchos árboles de decisión y cada uno “vota” la respuesta. La predicción final es la combinación de todas esas votaciones

# TESTING

## TEST & SCORE

Evalúa y compara los modelos que entrenamos. Muestra métricas como accuracy, precisión, recall y AUC para ver cuál funcionó mejor.

## ROC ANALYSIS

Grafica la curva ROC y muestra qué tan bien cada modelo distingue entre clientes que responden y los que no.  
*Cuanto más alta el área (AUC), mejor es el modelo.*

## CONFUSION MATRIX

Permite ver los aciertos y errores del modelo. Muestra cuántos clientes fueron clasificados correctamente y dónde se equivocó (*falsos positivos y falsos negativos*).

# TEST AND SCORE

¿A qué modelo le fue mejor?

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.811	0.869	0.845	0.851	0.869	0.378
Naive Bayes	0.747	0.849	0.838	0.832	0.849	0.349
Random Forest	0.742	0.847	0.822	0.816	0.847	0.268

**Logistic Regression** fue el mejor porque tuvo los mayores valores en cada métrica, lo que significa que distingue mejor entre quienes responden y quienes no, y tiene menos errores.

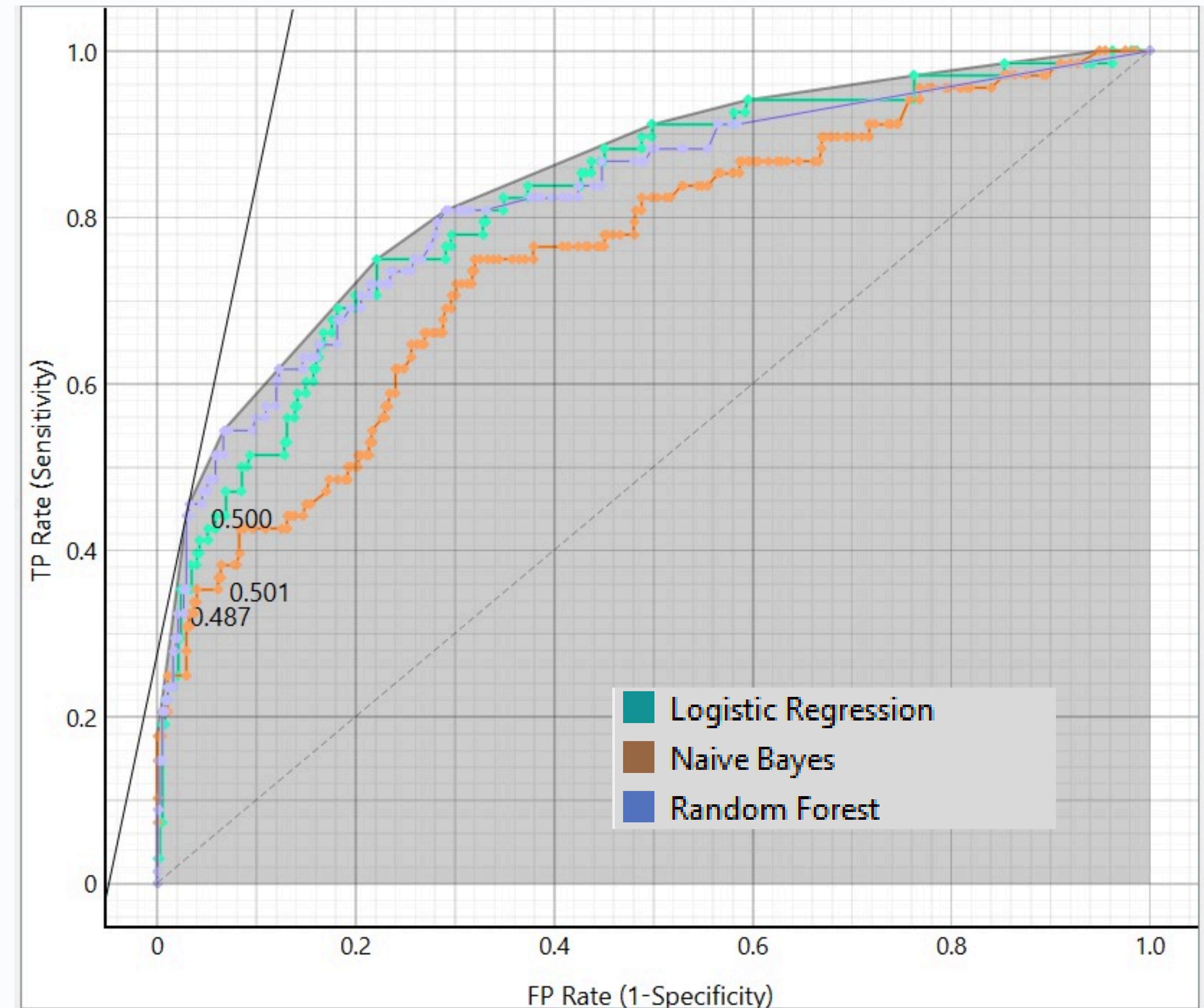
# ROC ANALYSIS

¿A qué modelo le  
fue mejor?

Eje Y (Vertical): Tasa de Aciertos (Verdaderos Positivos)

Eje X (Horizontal): Tasa de Errores (Falsos Positivos)

TARGET: RESPONDIO



# CONFUSION MATRIX

## LOGISTIC REGRESSION

		Predicted		$\Sigma$
		No respondio	Respondio	
Actual	No respondio	366	9	375
	Respondio	46	22	68
$\Sigma$		412	31	443

"El modelo se equivoca 9 veces prediciendo que una persona iba a responder cuando en realidad no lo hizo (falsos positivos), y 46 veces predijo que no iba a responder cuando sí lo hizo (falsos negativos)."



# CONFUSION MATRIX

## NAIVE BAYES

		Predicted		$\Sigma$
		No respondio	Respondio	
Actual	No respondio	351	24	375
	Respondio	42	26	68
$\Sigma$		393	50	443

*"El modelo se equivoca 24 veces prediciendo que una persona iba a responder cuando en realidad no lo hizo (falsos positivos), y 42 veces predijo que no iba a responder cuando sí lo hizo (falsos negativos)."*



# CONFUSION MATRIX

## RANDOM FOREST

		Predicted		$\Sigma$
		No respondio	Respondio	
Actual	No respondio	358	17	375
	Respondio	37	31	68
$\Sigma$		395	48	443

*“El modelo se equivoca 17 veces prediciendo que una persona iba a responder cuando en realidad no lo hizo (falsos positivos), y 37 veces predijo que no iba a responder cuando sí lo hizo (falsos negativos).”*



## CONCLUSIÓN

Gracias al modelo, la empresa puede identificar qué clientes tienen más probabilidad de responder y enfocar la campaña en ellos.

Esto permite ahorrar presupuesto, aumentar la tasa de respuesta y mejorar el retorno de inversión.

# MUCHAS GRACIAS

