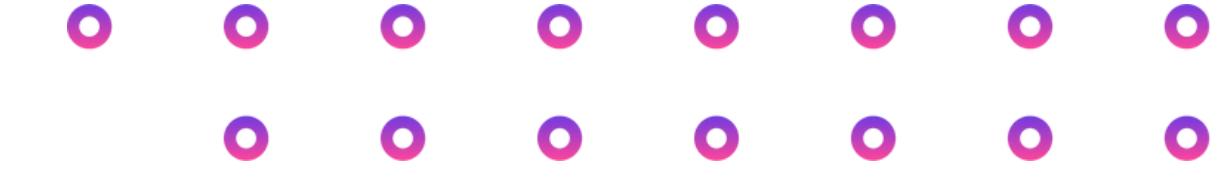


Machine Learning E-Commerce



Raquel Hernández Lozano
The Bridge 2025



Problema de Negocio

La mayoría del tráfico web no convierte.

Necesitamos identificar a los clientes reales antes de que se vayan





Problema de Negocio



APLICANDO MODELO MATCHINE LEARNING: CLASIFICACIÓN

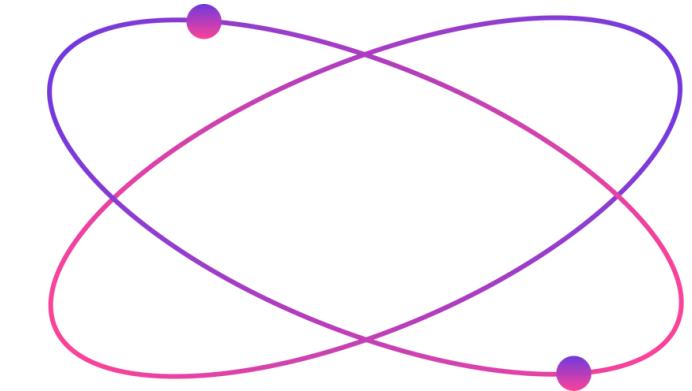
Predecir la variable target (compra/no compra) a nivel de sesión

**Mejorar previsiones
de producción para
evitar futuras
rotura de stock**

**Mejora en ventas
con estrategia de
marketing bien
enfocada**

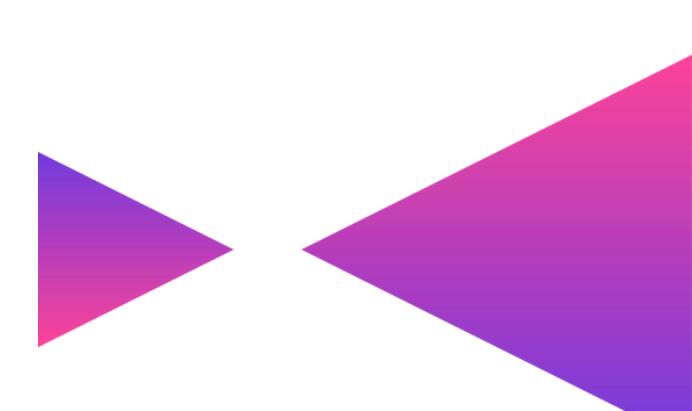


Fuente de datos disponible



Google Analytics

GA4 sera nuestro combustible para que el modelo de machine learning funcione correctamente para predecir si la sesion va a tener compra o no

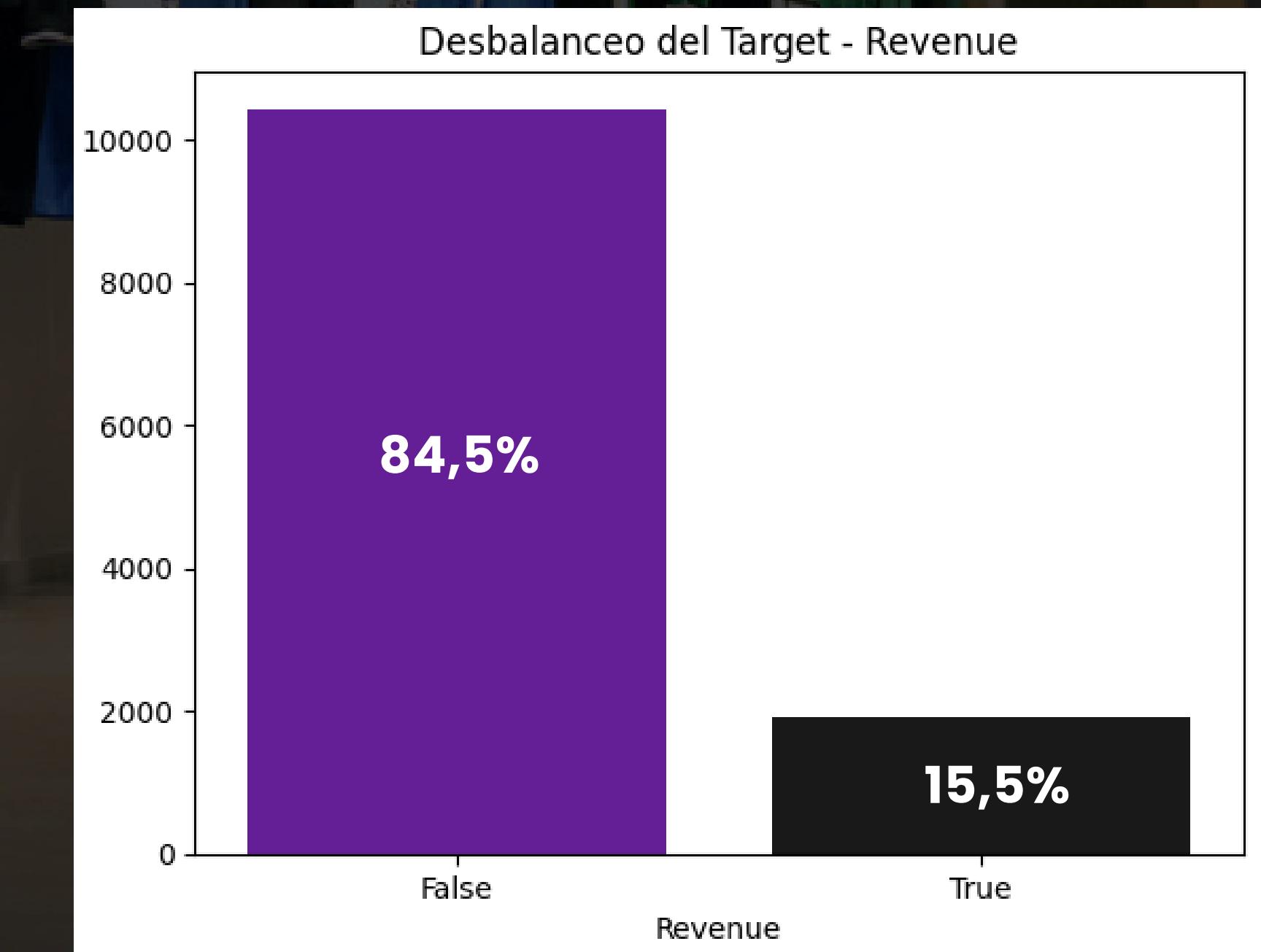


Reto E- Commerce:

La mayoría de sesiones no termina en compra

Desbalanceo de clases

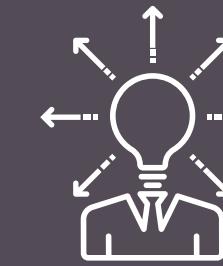
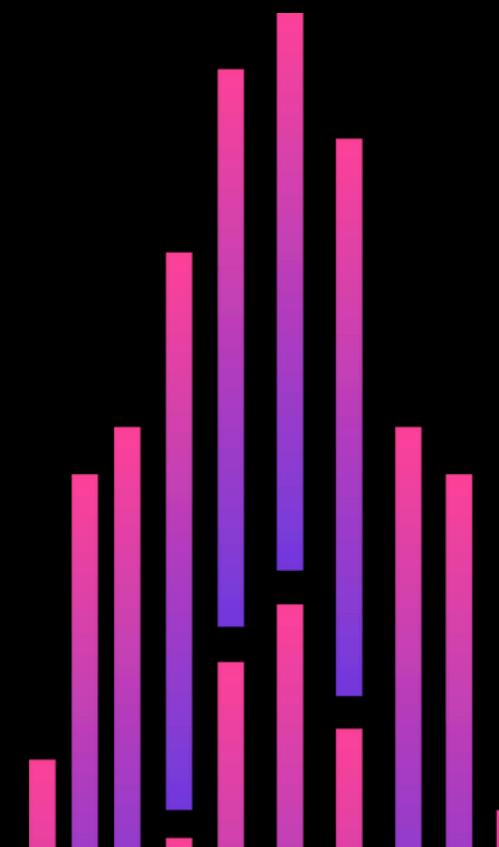
Total de sesiones: 12.330
18 variables
stratify=y



Huella digital de clientes



Google Analytics



Variables de Comportamiento Web

Administrative_Duration
Informational_Duration
ProductRelated_Duration

ExitRates
PageValues ***



Variables Contextuales

Browser
Region
TrafficType
VisitorType

Month
SpecialDay
Weekend
....



Variable Objetivo

Revenue YES/NO

PageValue info GA

Ingenieria de datos Ecommerce

Limpieza

01

En datos de GA no hay Nulos en principio por lo que este paso nos resulta fácil.

Transformacion. Feature Engineering

Aplicar logaritmo a las numericas para tratarlas, suavizar la influencia de los outliers. Binarizacion de variables como SpecialDay a 0 y 1 . Visitor Type 1 New Visitos 0 Returning Visitor. Standard Scaler

Codificacion y escalado

Traducir variables de texto a números. Estandarizar todas las variables numéricas para que todo tenga misma magnitud

02

03

04

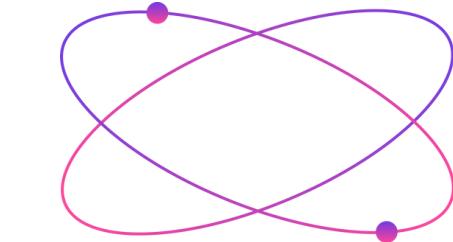
Division Train/Test

Separamos los datos en un conjunto de Entrenamiento (para que el modelo aprenda) y un conjunto de Prueba (datos nunca vistos) para validar su rendimiento. stratify=y





La "Magia" del Machine Learning



Modelado Predictivo

"Nos interesa que el modelo no se pierda ninguna posible compra. Si hay una compra, queremos que el modelo la detecte"

Métrica

Recall Alto



Buscamos crear un detector de clientes en nuestra web.

Algoritmos Maching Learning
Supervisados probados:

- Regresión Logística
- KNeighbors
- Decision Tree
- Random Forest
- SVM
- XGBoost

Algorito Maching Learning **No Supervisado** probado.

- K-means



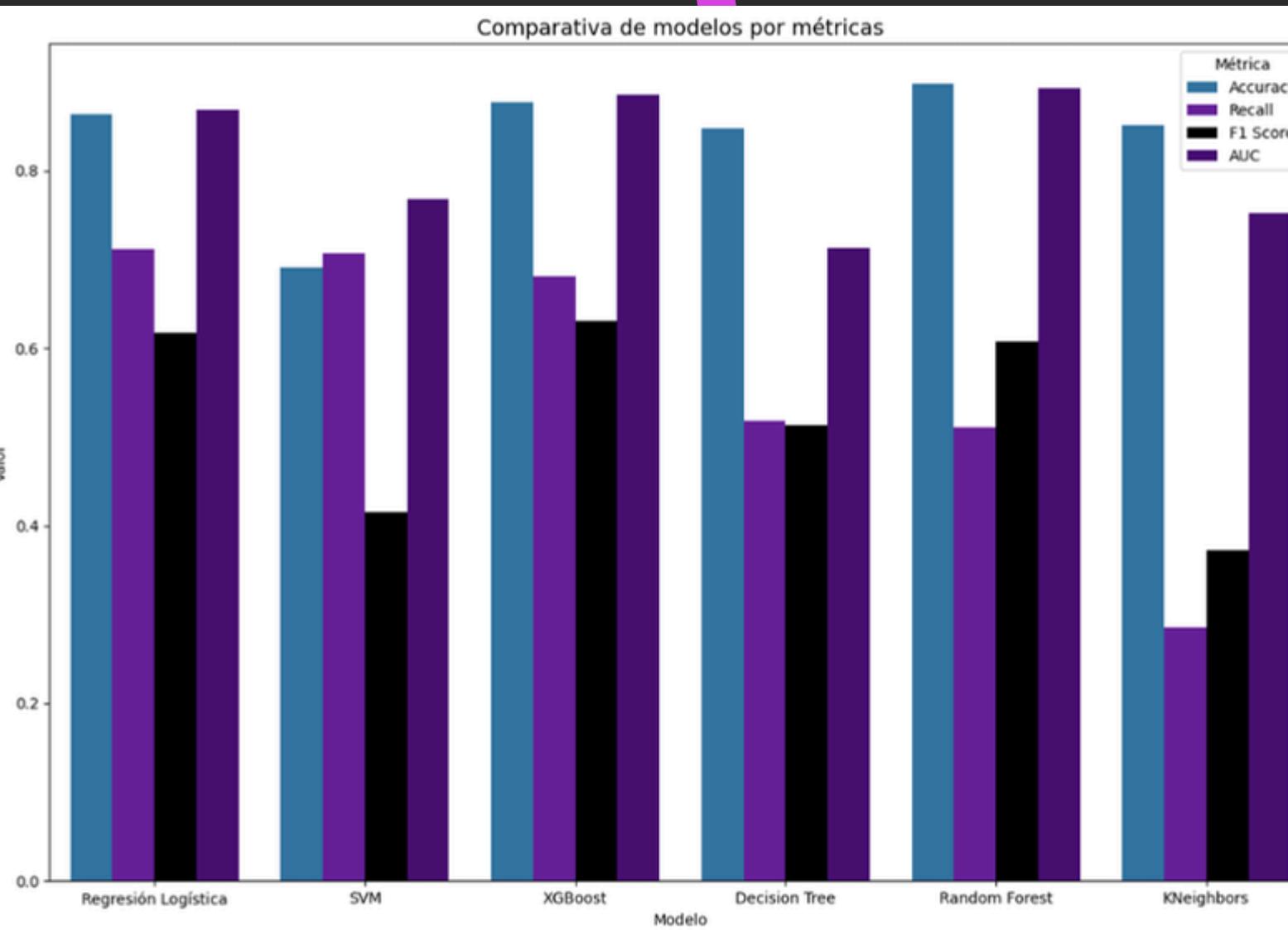
Baseline. Modelos ML Probados

`class_weight='balanced'`

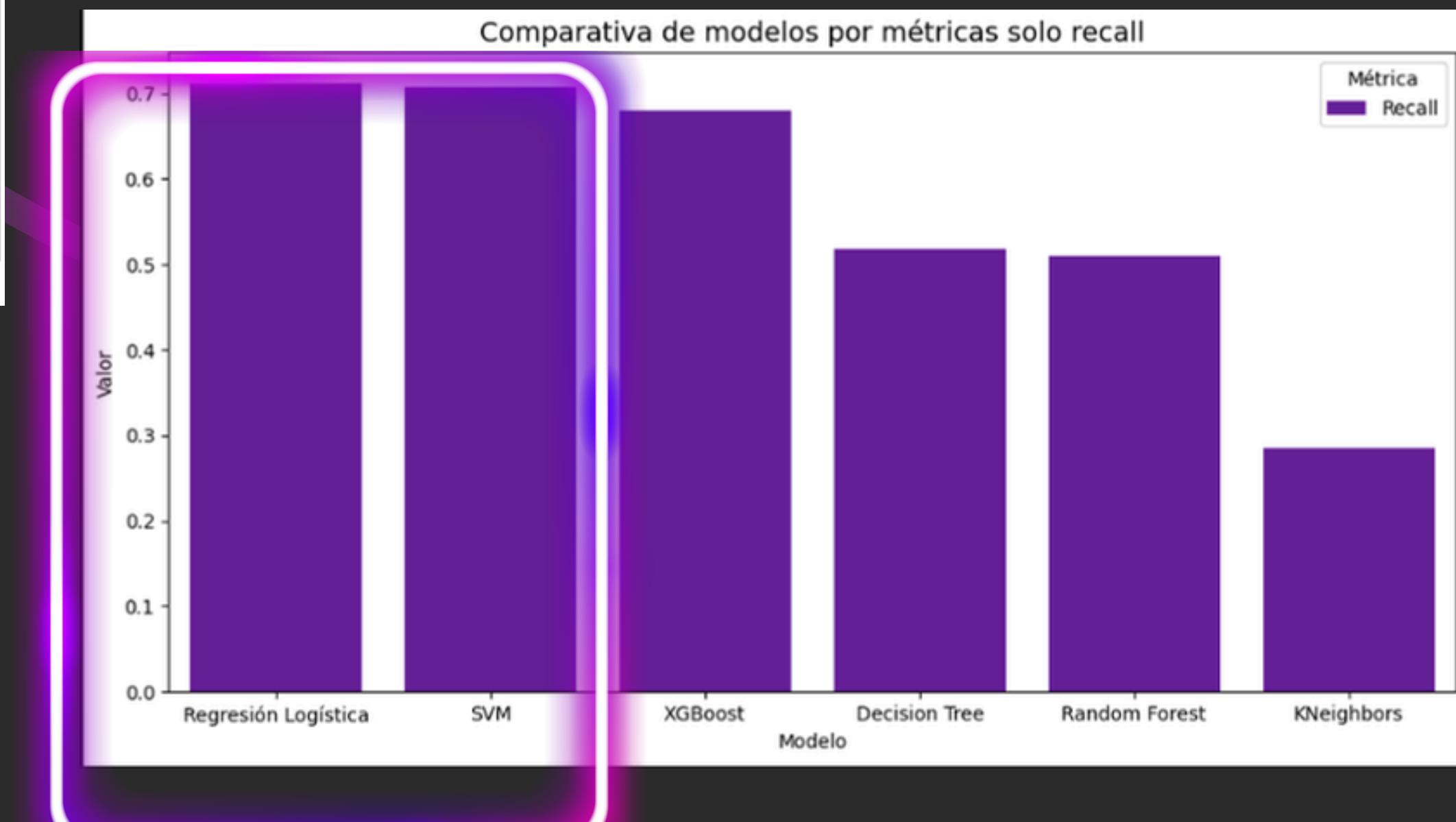
`XGBoost: scale_pos_weight=scale_pos_weight`

	Modelo	Accuracy	Recall	F1 Score	AUC
0	Regresión Logística	0.862936	0.712042	0.616780	0.868535
4	SVM	0.690998	0.706806	0.414747	0.767489
5	XGBoost	0.876318	0.680628	0.630303	0.885926
2	Decision Tree	0.847526	0.518325	0.512953	0.712288
3	Random Forest	0.897810	0.510471	0.607477	0.893171
1	KNeighbors	0.851176	0.285340	0.372650	0.752203

Modelos ML Supervisados Probados:



Nos interesa el recall alto.
Positivos(compradores)
correctamente
identificados



Baseline. Modelos ML Probados

Optimizado Vs Baseline

Etapa		Modelo	Accuracy	Recall	F1 Score	AUC	Umbral
2	Optimizado	Regresión Logística	0.871857	0.678010	0.621103	0.871009	0.55
0	Baseline	Regresión Logística	0.862936	0.712042	0.616780	0.868535	0.50
3	Optimizado	SVM	0.784266	0.782723	0.529204	0.824240	0.10
1	Baseline	SVM	0.690998	0.706806	0.414747	0.767489	0.50

SVM:

- El optimizado mejora en todo: Accuracy, Recall, F1 y AUC.
- El umbral baja a 0.10, lo que hace que el modelo sea más agresivo (detecta más positivos).
- El Recall sube bastante, lo que indica que detecta más compradores sin perder demasiado en precisión

SVM optimizado es claramente superior

MODELO SVM OPTIMIZADO

Ajuste de hiperparametros



El modelo detecta 78% de los compradores reales.
Recall



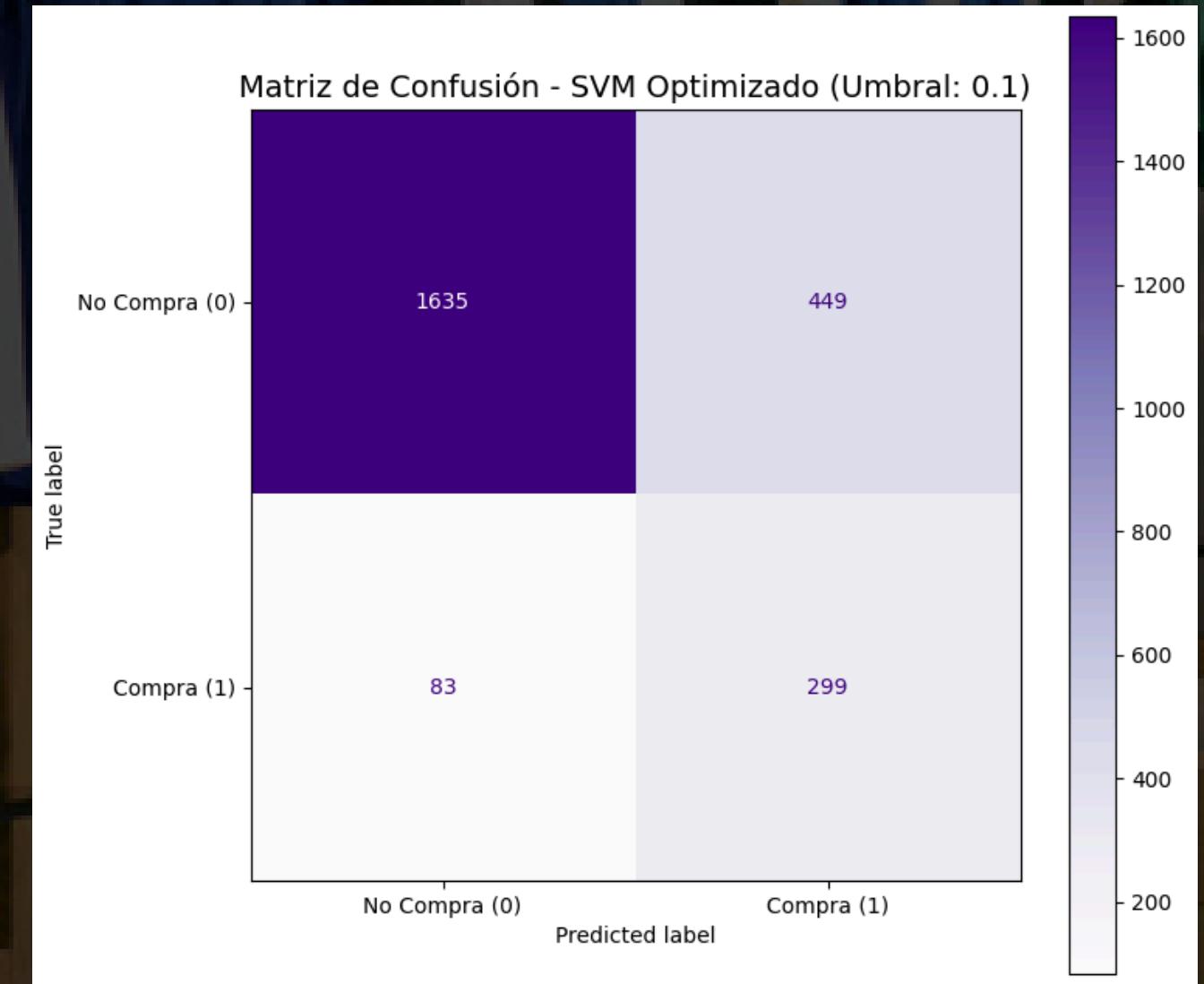
El modelo acierta en casi 8 de cada 10 casos.
Accuracy



El modelo tiene buena capacidad de ranking: distingue bien entre compradores y no compradores.
AUC

Matriz de confusión SVM

En nuestro caso un Falso Negativo (no identificar a un comprador) es más costoso que un Falso Positivo (dar un descuento innecesario).



TN (1635): Casos correctamente clasificados como "No Compra"

FP (449): Casos que no compraban, pero el modelo predijo que sí.

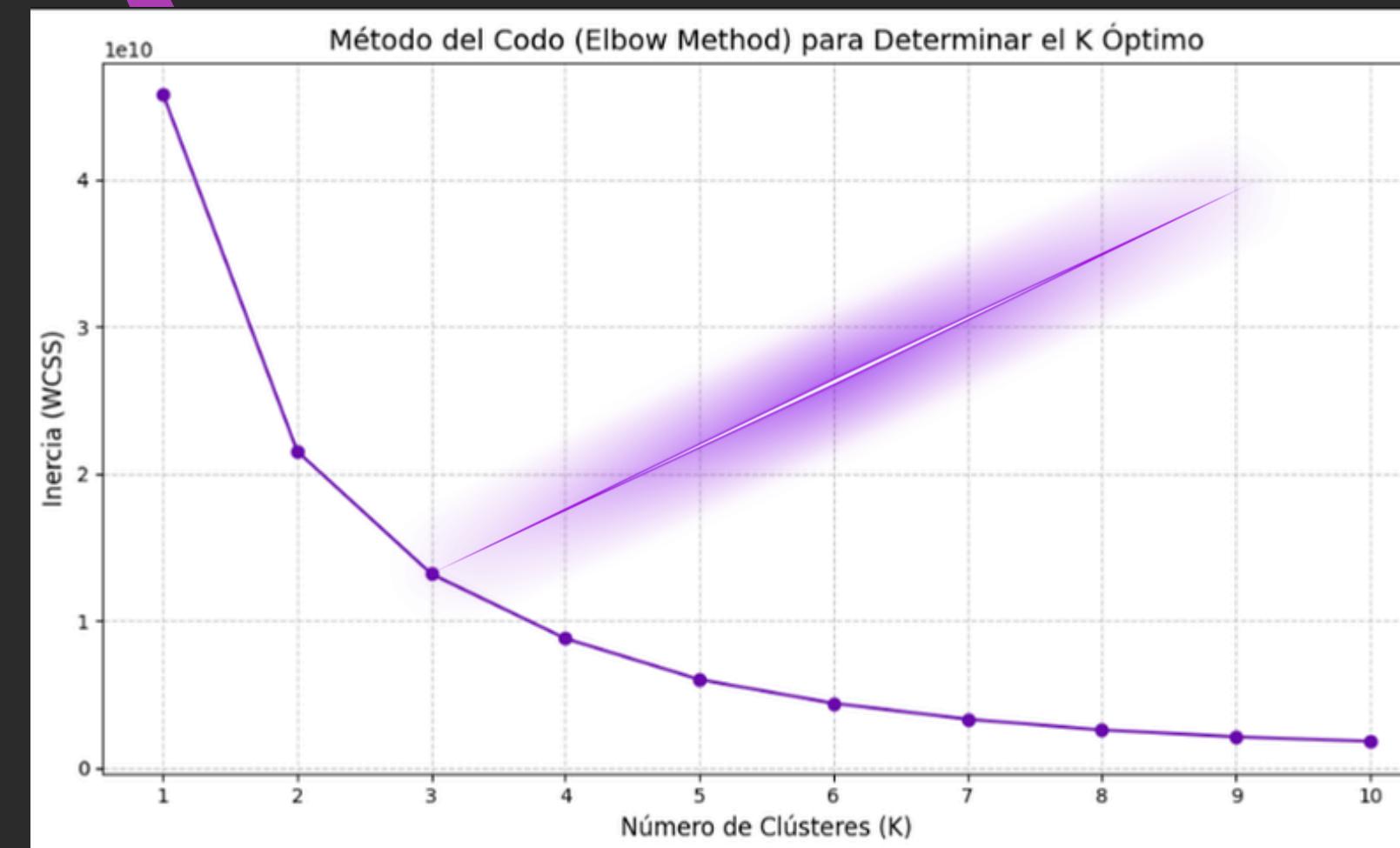
FN (83): Casos que sí compraban, pero el modelo no los detectó.



TP (299): Casos correctamente clasificados como "Compra"

Modelos ML No Supervisado Probado

K-Means Segmentación de Clientes.
Clustering realizado en 3 segmentos



Segmento 1: clientes con alta probabilidad de compra.
Segmento 2: clientes indecisos o de bajo gasto.
Segmento 3: visitantes que rara vez compran.

¿Por qué es útil?

Marketing personalizado: diseñar campañas distintas para cada segmento.

Optimización de recursos: invertir más en los segmentos con mayor retorno.

Estrategia de negocio: entender qué tipo de clientes tienes y cómo se comportan.

REAL TIME: PROBEMOS EL MODELO

Streamlit

Modelo de Predicción de Intención de Compra (E-commerce)

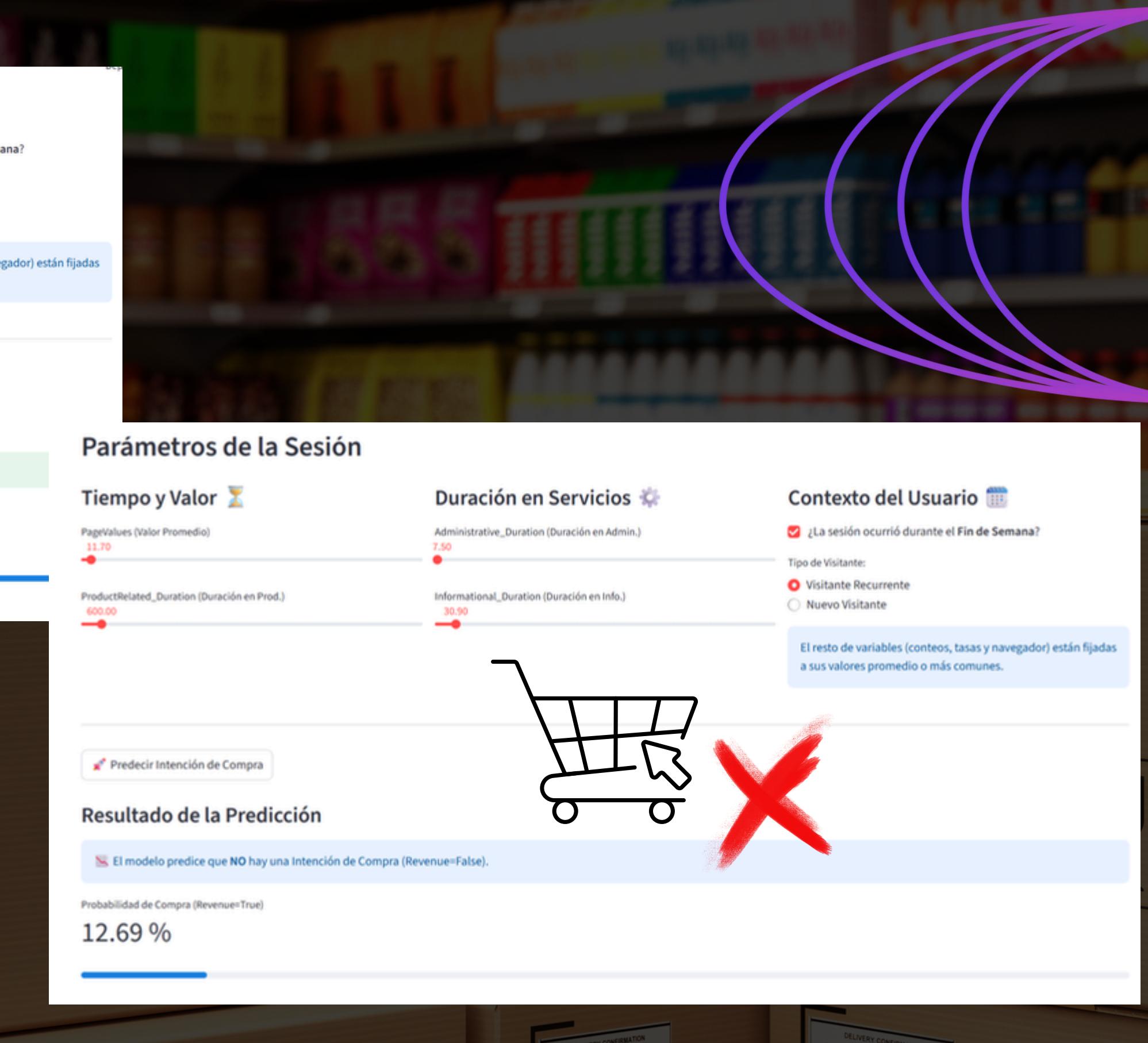
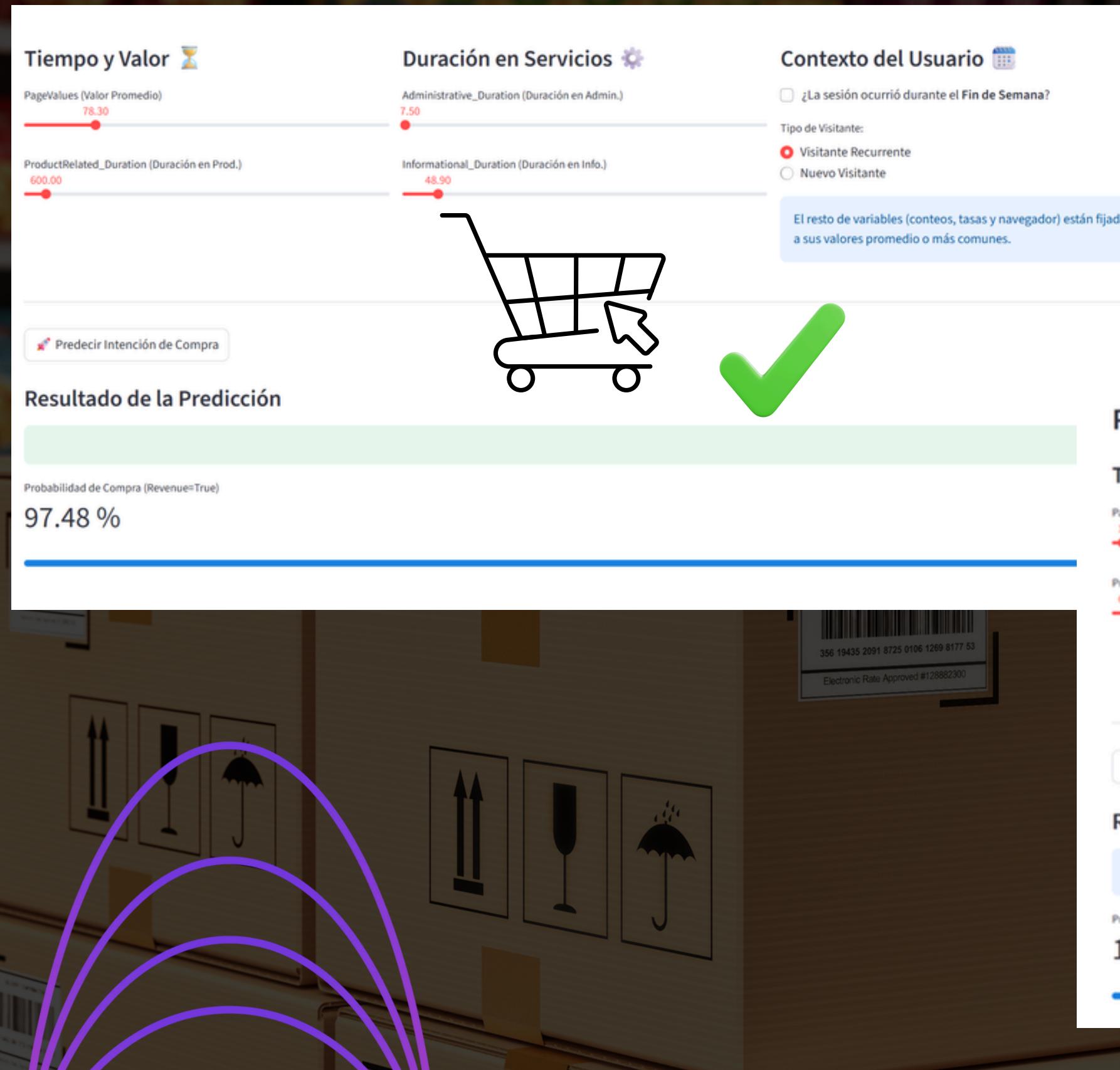
Ajusta las variables clave para predecir la intención de compra. Ahora con manejo robusto de tipos de datos para evitar errores.

Parámetros de la Sesión

Tiempo y Valor	Duración en Servicios	Contexto del Usuario
PageValues (Valor Promedio) 18.00 <input type="range" value="18.00"/>	Administrative_Duration (Duración en Admin.) 7.50 <input type="range" value="7.50"/>	<input type="checkbox"/> ¿La sesión ocurrió durante el Fin de Semana?
ProductRelated_Duration (Duración en Prod.) 600.00 <input type="range" value="600.00"/>	Informational_Duration (Duración en Info.) 48.90 <input type="range" value="48.90"/>	Tipo de Visitante: <input checked="" type="radio"/> Visitante Recurrente <input type="radio"/> Nuevo Visitante

El resto de variables (conteos, tasas y navegador) están fijadas a sus valores promedio o más comunes.

PRECICCIONES MODELO SVM OPTIMIZADO



Limitaciones

Datos solo de 1 año

Falta automatización. Futuro csv para subida de datos

Seguir alimentando el modelo, para mejorar futuras clasificaciones

Escalabilidad

Posibilidad de reentrenar el modelo con más datos

Si el tráfico se duplica, duplicamos aprendizaje

No afecta en nada a la velocidad de carga de web

Posibilidad de enriquecer el modelo incrementando la precisión

Incluir datos offline. Clientes registrados

Futuros beneficios de usar Maching Learning para predecir clientes

01

Aumento de la tasa de
conversión. Código dto/chatbot

02

Maximización ROI en Marketing.
Publicidad dirigida y bien
enfocada

03

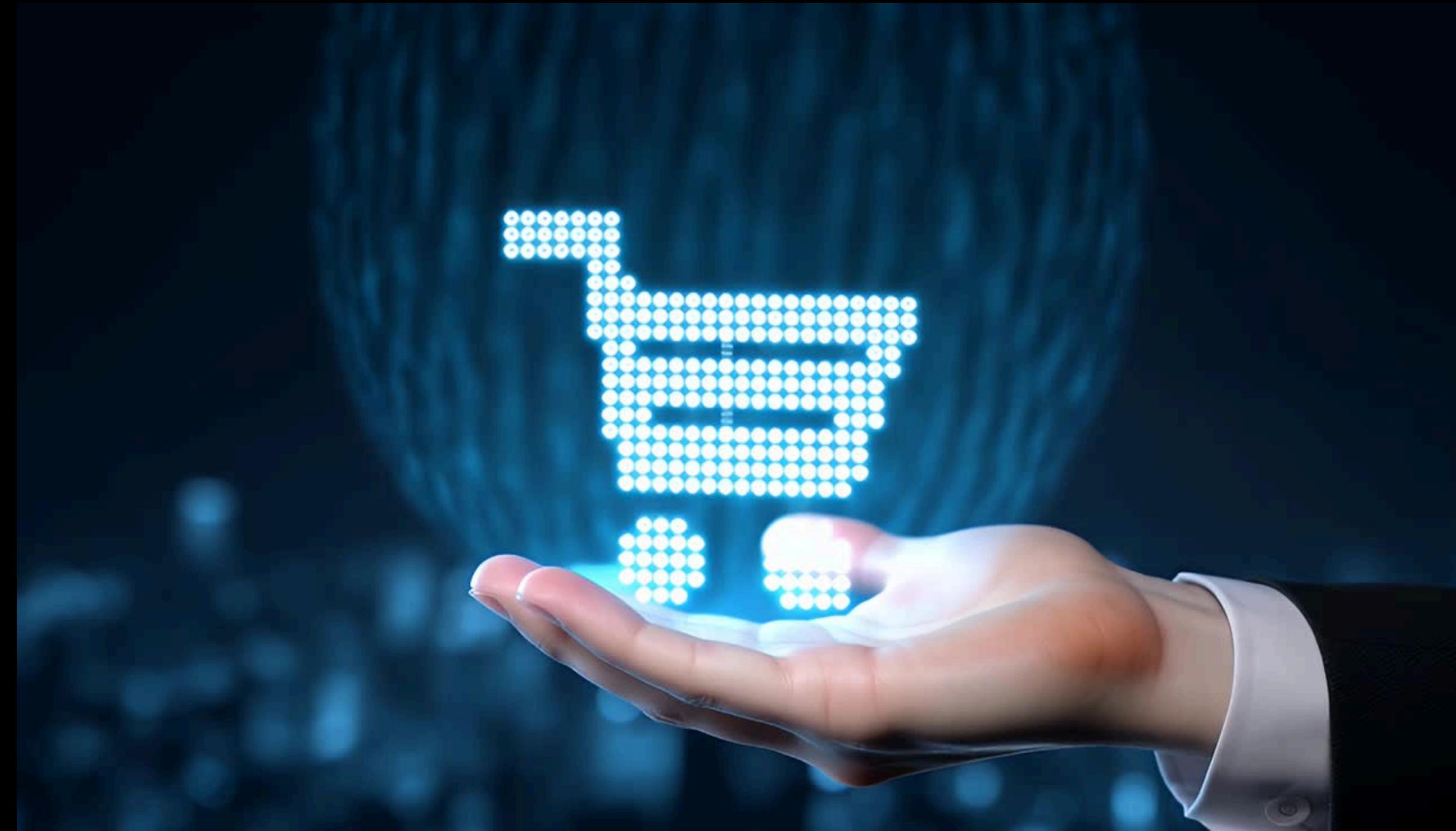
Optimización de customer
Journey

04

Mejorar prevision de stock con
las señales de intención de
compra

05

El modelo de ML aumenta la
inteligencia del negocio.
Prediccion del valor de vida de cliente



¡Gracias por vuestra atención!