datec

Anticipando Intenciones de Compra con Machine Learning

Caso de Éxito del Genio X

Powered by D4G

Resumen Ejecutivo

Se decidió emprender el proyecto de Anticipación de Intenciones de Compra para identificar los factores clave que influyen en la decisión de compra de los clientes en la plataforma de e-commerce del Genio X (EGX) y a su vez desarrollar un modelo predictivo de Machine Learning (ML) que pueda predecir la propensión de compra de los usuarios. El objetivo del proyecto es optimizar las estrategias de marketing, personalizar la experiencia de compra y mejorar la conversión de ventas de la plataforma.

Tras un análisis exhaustivo y la implementación de diversas técnicas de ML, se seleccionó el modelo con el mejor desempeño de todas las iteraciones. Este modelo predice correctamente el 97% de los usuarios que realizan compras en esta plataforma, con una precisión del 13%, y permite identificar al conjunto de usuarios que dan señales de querer comprar, pero que finalmente no lo hacen. Gracias a estos insights se puede dirigir las estrategias de marketing y esfuerzos de conversión de manera más efectiva hacia este grupo específico de usuarios.

1. Introducción

Contexto

En la actualidad para un negocio de e-commerce, la capacidad de entender y predecir el comportamiento de sus usuarios es fundamental para su éxito y sostenibilidad a largo plazo. EGX, a pesar de tener acceso a una rica fuente de datos a través de Google Analytics 4, no ha aprovechado plenamente este recurso para analizar y comprender mejor a sus usuarios. Por lo tanto este proyecto surge como una iniciativa estratégica para cerrar esta brecha, y así transformar datos brutos en insights accionables y ventajas competitivas.

Justificación

El proyecto de Propensión de Compra para EGX permitió:

- Mejorar la comprensión de los comportamientos de nuestros usuarios.
- Personalizar sus experiencias de compra.
- Mejorar la toma de decisiones basada en datos.
- Aprovechar al máximo los recursos de marketing para impulsar la rentabilidad.

Este proyecto representa un paso vital hacia la transformación de EGX en una organización data-driven, capaz de navegar con éxito el dinámico y competitivo entorno del retail online.

2. Objetivos

- Identificar los principales factores que afectan a la propensión de compra de los usuarios.
- Desarrollar un modelo predictivo para estimar la propensión de que un usuario que visite la página de EGX realice una compra.
- Proporcionar recomendaciones basadas en datos para mejorar la conversión de clientes.

3. Desarrollo

Se resume el desarrollo del proyecto en el siguiente esquema:

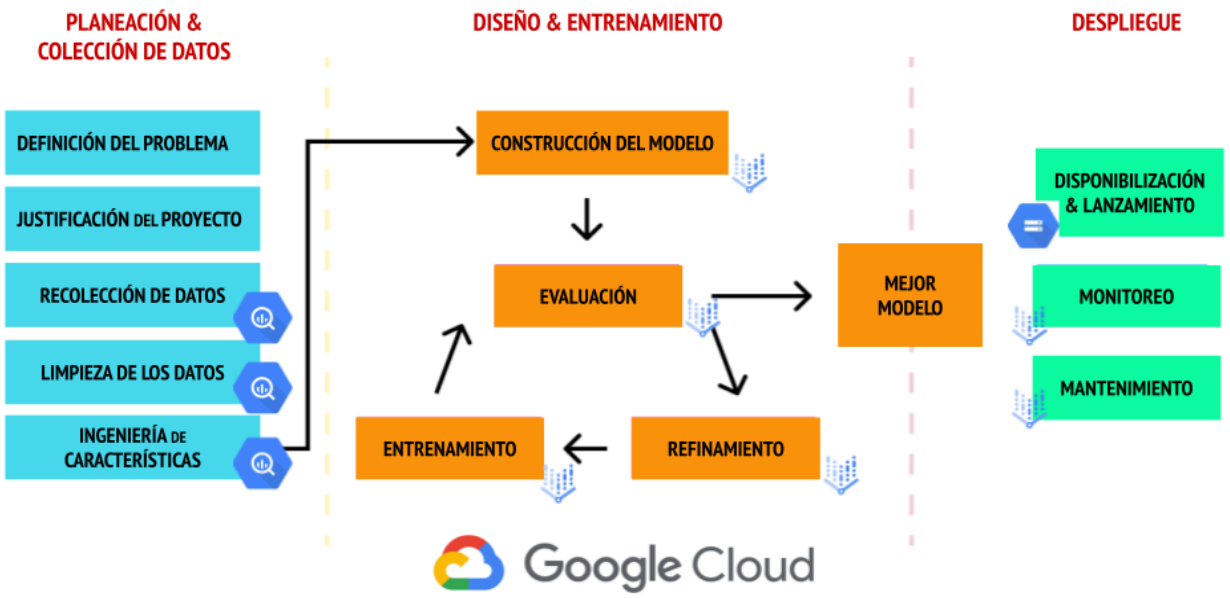

Imagen 1: Fases del Proyecto

Recolección y Preprocesamiento de Datos

Se utilizaron los datos históricos del comportamiento de los usuarios de EGX. Estos datos son de *Google Analytics 4*, de donde se extrajo información valiosa del comportamiento de cada usuario a través del tiempo. Estos datos fueron limpiados, normalizados y enriquecidos para prepararlos para el análisis y alimentar al modelo de propensión.

Procesamiento de Datos
Consideraciones a Priori

- La identificación de usuarios se realizó mediante la clave ***pseudo_user_id,*** un identificador único asignado a cada usuario que visita EGX, vinculado específicamente al dispositivo o navegador desde el cual se accede a EGX. Este identificador, de naturaleza anónima, preserva la privacidad del usuario mientras suministra información valiosa para análisis y personalización.

- Se segmentó la información por usuario en bloques semanales que contengan un resumen de las acciones que tenga el usuario en ese periodo de tiempo.

Resultados post procesamiento

- Resumen del comportamiento semanal de los usuarios, que contiene una serie de características, las mismas están especificadas en la Tabla 1.

Nombre de la variable	Descripción de la variable
SO_dispositivo	Sistema Operativo del dispositivo desde el que usuario accede a EGX
tiempo_de_sesion_mas_larga	Tiempo de sesión más larga que tuvo el usuario durante una semana
#sesiones	Cantidad de sesiones que tuvo el usuario durante una semana
#sesiones_1_periodo_atras	Cantidad de sesiones que tuvo el usuario durante una semana atrás
#sesiones_2_periodos_atras	Cantidad de sesiones que tuvo el usuario durante dos semanas atrás
#sesiones_3_periodos_atras	Cantidad de sesiones que tuvo el usuario durante tres semanas atrás
#scrolls	Cantidad de eventos scroll (cuantas veces el usuario llegó hasta el final de la página) que realizó el usuario durante la última semana
#page_view	Cantidad de páginas que visito el usuario durante la última semana
#add_to_cart	Cantidad de artículos agregados al carrito que hizo el usuario durante la última semana
#scrolls_1_periodo_atras	Cantidad de eventos scroll que realizó el usuario durante una semana atrás
#productos_comparacion_vistos	Cantidad de productos en que pertenecen a la categoría Productos de comparación
#productos_comparacion_en_carrito	Cantidad de productos en el carrito que pertenecen a la categoría Productos de comparación
#productos_especialidad_vistos	Cantidad de productos vistos que pertenecen a la categoría Productos de especialidad
horario_fav_ingreso	El horario favorito en el cual el usuario ingresa a EGX, estos horarios se dividen en: mañana, tarde y noche.
precio_promedio_productos_vistos	Precio promedio de los artículos que el usuario vio durante la última semana
categoria_favorita	Dado el conjunto de artículos que el usuario vio durante la última semana, se realizó una categorización del mismos y se obtuvo la moda de estas categorías, en caso de existir un empate se eligió una de las categorías al azar
propension_compra	Esta variable es un valor binario donde: 1 indica que

	el usuario es propenso a comprar (y realizo una compra en el periodo) y 0 indica que el usuario no es propenso (y efectivamente no realizó una compra en el periodo)

Tabla 1: Tabla de descripción de variables

Diseño y Entrenamiento del Modelo

Diseño del Modelo

Considerando que el propósito fundamental es distinguir a los usuarios con mayor propensión de realizar una compra, se diseñó este proyecto para utilizar técnicas de aprendizaje automático supervisado, concretamente, una tarea de clasificación, dado que inicialmente conocemos las categorías que queremos prever.

Dicha clasificación es de tipo binaria, implicando dos categorías:

- La **clase 1**, que incluye a los ***usuarios propensos*** a realizar compras (y que efectivamente compraron).
- La **clase 0**, correspondiente a aquellos usuarios que no muestran dicha inclinación (y no efectuaron compra alguna).

Al analizar los datos, se constató una marcada disparidad entre las clases, siendo los usuarios propensos notablemente menos numerosos (Ver imagen 2).

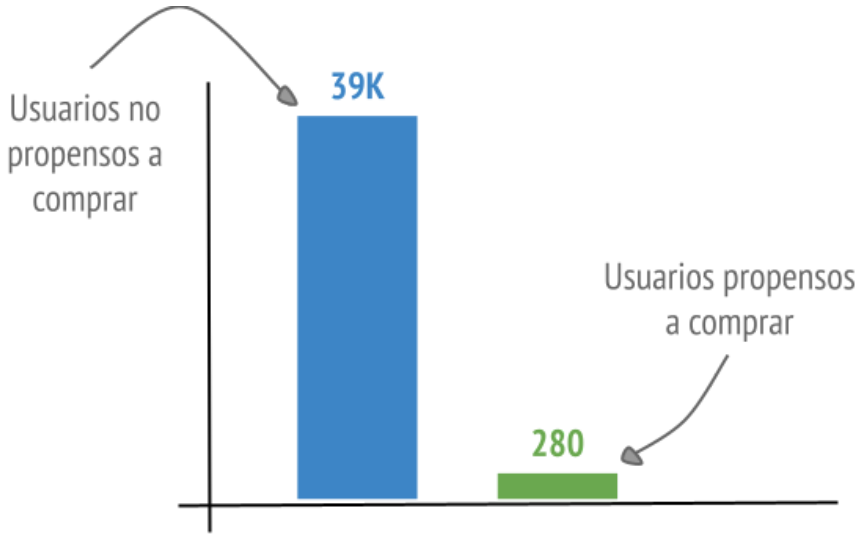

Imagen 2: Gráfico de distribución de clases

Ante el desequilibrio observado, se implementaron estrategias de balanceo con el fin de equiparar la representación de ambas clases.

Proceso de Entrenamiento

División del conjunto de datos

Se definió los conjuntos como:

- *Grupo de entrenamiento,* los datos del año 2023 y del primer mes del 2024.
- *Grupo de prueba,* los datos de las cuatro primeras semanas de febrero del 2024.

Definición de métrica
Dado el objetivo principal de este proyecto se decidió utilizar el ***recall*** como métrica de decisión, ya que se debe garantizar que se identifique la mayor cantidad posible de usuarios verdaderamente propensos a realizar una compra. Debido a que en este contexto, perder a un usuario potencialmente valioso (un falso negativo) puede significar perder una oportunidad de venta real. Esta métrica además, al enfocarse en la proporción de positivos reales que fueron correctamente identificados por el modelo, asegura que se minimice la pérdida de estos usuarios potenciales.

Experimentos
Durante la fase de entrenamiento, se realizaron una serie de experimentos para mejorar el desempeño de los modelos. Se evaluaron diferentes algoritmos de aprendizaje supervisado, tales como la regresión logística, árboles de decisión y modelos ensamblados. Además, se probó varias combinaciones de variables y se ajustaron los hiperparámetros con el objetivo de afinar el desempeño del modelo de clasificación.

Entre las opciones evaluadas, el algoritmo *XGBoost* demostró ser superior, destacando por su excelente desempeño y su capacidad para identificar con mayor exactitud a los usuarios con mayor propensión a comprar, razón por la cual fue el modelo elegido.

Adicionalmente se ajustó el umbral de decisión utilizando la curva ROC (Ver imagen 3) con el fin de encontrar un balance adecuado entre la tasa de falsos positivos y falsos negativos. Este paso se llevó a cabo con el objetivo de identificar de manera precisa a los usuarios con alta propensión de compra, minimizando así el riesgo de perder potenciales oportunidades de recompra o de convertir usuarios en clientes.

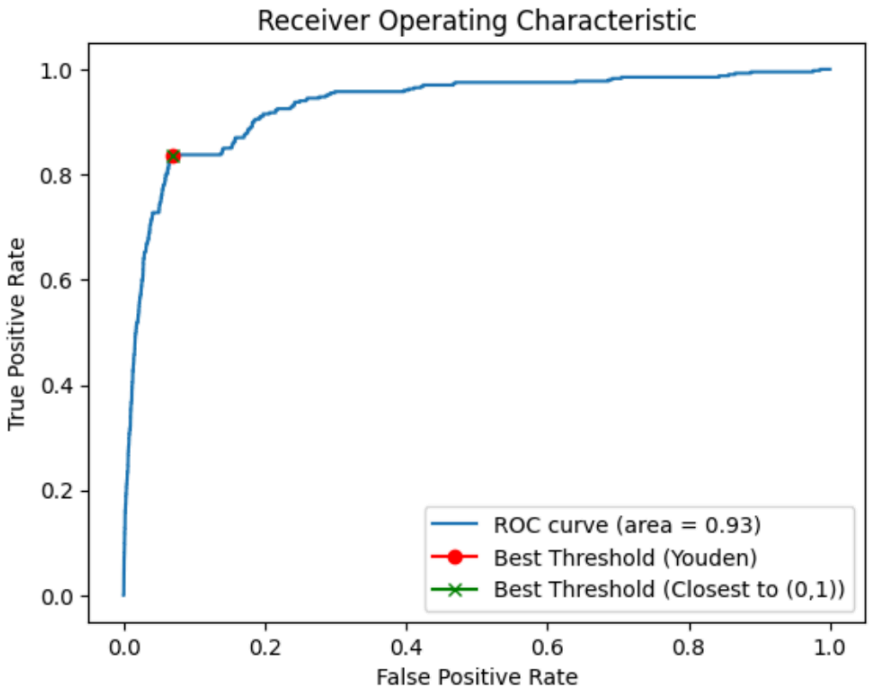

Imagen 3: Curva de ROC para optimizar el umbral de decisión

Evaluación del Modelo

Los resultados arrojados por el modelo elegido se ilustran en la Imagen 4. Aunque observamos una elevada incidencia de falsos positivos, este fenómeno no representa una desventaja en el contexto específico de nuestro proyecto. El propósito principal es identificar segmentos de usuarios que exhiben patrones de comportamiento similares, permitiéndonos así canalizar de manera efectiva nuestros esfuerzos de conversión hacia estos grupos.

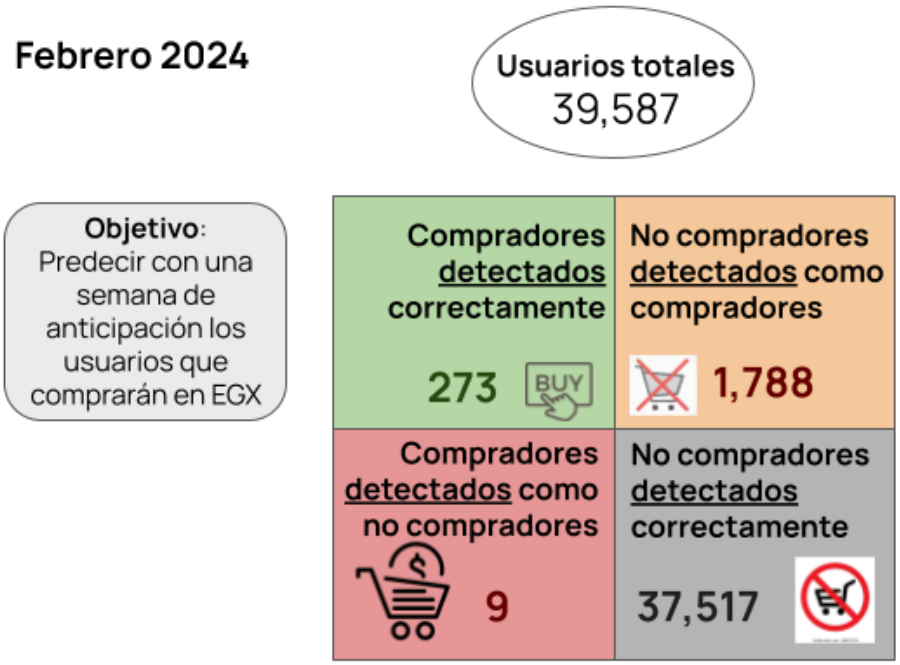

Imagen 4: Matriz de confusión del modelo seleccionado

	Precision	Recall	F1-Score	Total
0 (no propensa)	1.00	0.95	0.98	39,305
1 (propensa)	0.13	0.97	0.23	282
accuracy			0.95	39,587
macro avg	0.57	0.96	0.60	39,587
weighted avg	0.99	0.95	0.97	39,587

Tabla 2: Resultados de la evaluación del modelo

Interpretación

De la tabla de resultados anterior se pueden extraer las siguientes interpretaciones:

- En Febrero 2024 el modelo predijo que 2,061 usuarios comprarían en EGX.
- 273 compraron y 1,788 no compraron (precisión del 13%).
- 1,7880 usuarios demostraron un comportamiento propenso a comprar pero no compraron.

Además, se identificaron las características clave que influyen en la propensión de un usuario. En el gráfico de valores SHAP adjunto, se detalla cómo varían los resultados al ajustar cada característica específica, por ejemplo se puede notar que la característica de tiempo en sesión más larga de la semana es una de las más influyentes, conjuntamente con la cantidad de eventos scroll que realiza el usuario.

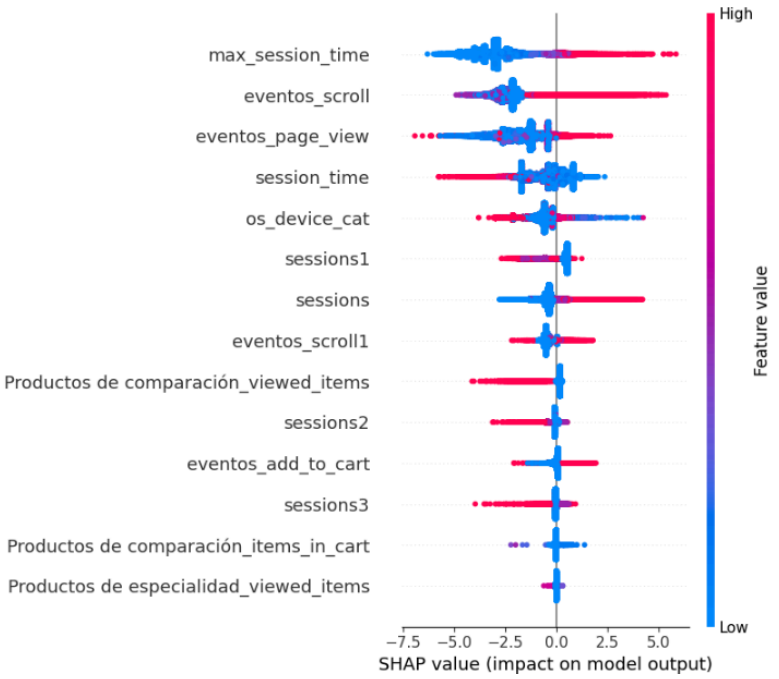

Imagen 5: Gráfico de valores SHAP - Importancia de las características

Despliegue del Modelo

La implementación se llevó a cabo mediante la automatización de la extracción y procesamiento semanal de datos desde Google Analytics 4, seguida de un proceso de entrenamiento también automatizado que se realiza cada semana. El modelo realiza predicciones de propensión de compra de los usuarios de EGX, y los datos resultantes se almacenan en una base de datos.

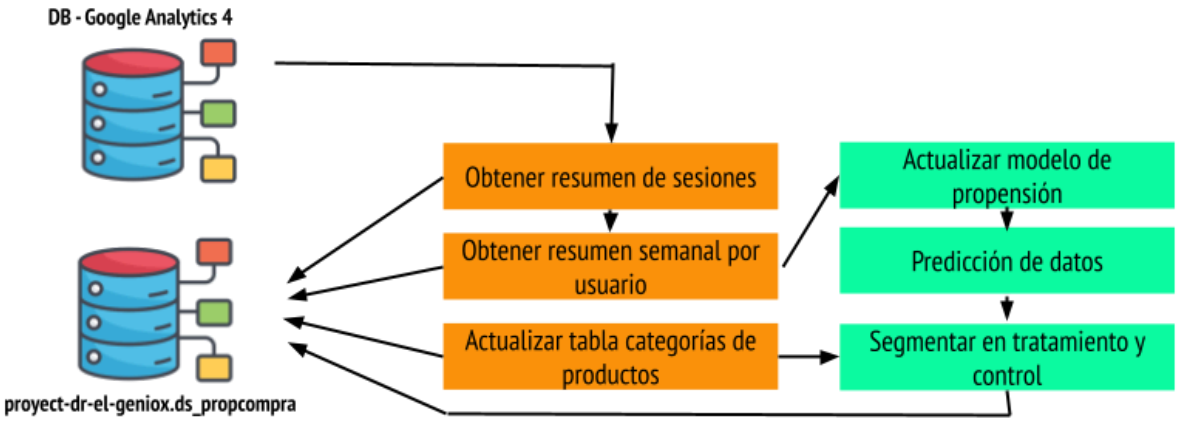

Imagen 6: Esquema conceptual despliegue

4. Recomendaciones y Acciones de Negocio

Teniendo en cuenta los resultados obtenidos y el objetivo propuesto, esta solución nos brinda la capacidad de influir en el comportamiento de compra de los usuarios propensos que no realizaron una adquisición. Esto podría aumentar significativamente la base de compradores, al implementar acciones específicas dirigidas hacia estos usuarios propensos.

Integración con Emarsys

Como primera medida estratégica para alcanzar los objetivos propuestos, se decidió emplear los resultados obtenidos para iniciar una campaña de envío de correos electrónicos dirigidos a aquellos usuarios identificados como propensos a realizar una compra. Esta decisión se fundamentó en una detallada segmentación de estos usuarios, considerando dos criterios clave: su categoría de producto preferida y su franja horaria de actividad predominante en la plataforma EGX.

Tomando como referencia estas directrices de segmentación, se procedió a organizar los usuarios propensos en grupos distintos según sus preferencias. Por ejemplo en la categoría de Electrónica, se formaron tres subgrupos basados en la preferencia de horario de los usuarios: mañana, tarde y noche. Este enfoque se replicó para cada una de las categorías de producto disponibles en EGX, asegurando así una personalización y precisión máxima en nuestra comunicación por correo electrónico.

Experimento A/B

Para evaluar la efectividad de esta estrategia, se estableció un experimento A/B (Imagen 7), asignando a los participantes a los grupos de tratamiento y control con una proporción equitativa de 50/50. Estos datos y la subsegmentación mencionada previamente se registraron en una tabla específica en nuestra base de datos, permitiéndonos monitorear la evolución de los resultados entre los grupos de tratamiento y control, con el objetivo de evaluar el impacto significativo de la estrategia a lo largo del tiempo.

La métrica utilizada para evaluar el impacto será la ***Tasa de Conversión (CR)***, es decir, la proporción de usuarios que realizaron una compra dividido por el total de usuarios que iniciaron sesión en la plataforma EGX. Debido a que la estrategia implica el envío de correos electrónicos, solo el grupo de tratamiento recibirá estos mensajes personalizados, mientras que el grupo de control no será objeto de dichas acciones dirigidas.

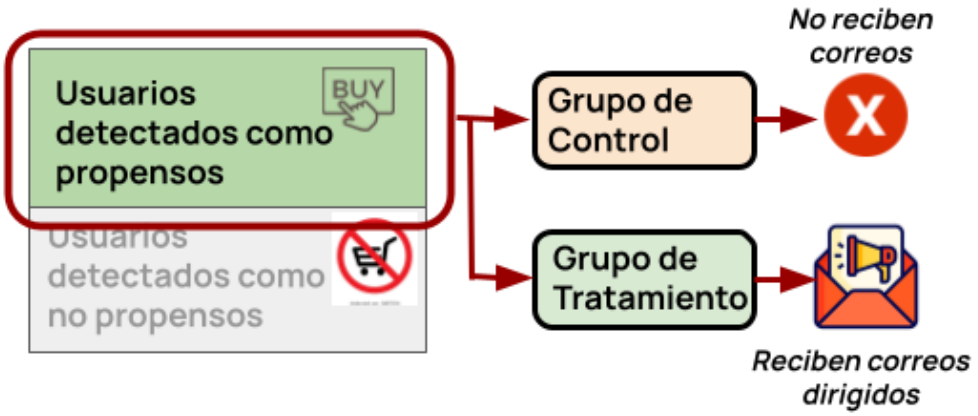

Imagen 7: Gráfico Experimento A/B

5. Conclusión

El proyecto de Anticipación de Intención de Compra en EGX se centró en identificar los determinantes cruciales que influyen en la decisión de compra de los usuarios, utilizando para ello un modelos predictivos de aprendizaje automático. A través de un análisis detallado y la aplicación de diversas técnicas, se logró desarrollar un modelo capaz de prever con cierta precisión la tendencia de compra de los usuarios. Este avance representa un paso significativo hacia la personalización de la experiencia de compra, la optimización de estrategias de marketing y, en última instancia, el incremento potencial en la conversión de ventas, subrayando la importancia de una metodología basada en datos para comprender mejor y actuar sobre los comportamientos de los usuarios en el ecosistema de e-commerce.

El empleo de datos históricos de Google Analytics 4 y su procesamiento mediante técnicas avanzadas de preprocesamiento y análisis permitieron segmentar eficazmente a los usuarios y predecir su comportamiento de compra. Esto no solo mejoró la comprensión de los patrones de comportamiento de los clientes sino que también facilitó la personalización de las experiencias de compra y la toma de decisiones basada en datos.

Finalmente, las acciones estratégicas y recomendaciones derivadas de los análisis subrayan el potencial de influir positivamente en el comportamiento de compra de los usuarios propensos que aún no han realizado una compra. A través de campañas de marketing dirigido y experimentos A/B, EGX está en camino de convertirse en una entidad impulsada por datos, capaz de navegar con éxito en el competitivo entorno del e-commerce. La implementación del modelo y las estrategias de seguimiento no solo auguran un aumento en la base de compradores sino también marcan un precedente en la utilización de analítica avanzada para la toma de decisiones estratégicas en el ámbito del marketing digital.

datec

Haz realidad proyectos de analítica avanzada en tu empresa. Comunícate con nosotros y descubre cómo podemos ayudarte.

Powered by D4G