INFORME DE ANALÍTICA SOBRE EL RENDIMIENTO DEL SITIO WEB DE E-CORP

ESTEBAN CARDONA,

GILBERTO GIL,

LEIDYS GUERRERO,

MATEO CAICEDO

[2553984] ENFASIS PROFESIONAL II

MANUELA LONDOÑO OCAMPO

1 DE ABRIL DE 2024

INTRODUCCIÓN

En la actualidad, las empresas buscan constantemente formas innovadoras de expandir su presencia en línea y captar nuevos clientes a través de canales digitales. En este caso, la compañía E-Corp, especializada en la venta de productos de lujo, ha enfrentado el desafío de maximizar el rendimiento de su sitio web de comercio electrónico, lanzado con el propósito de ampliar su base de clientes y aumentar sus ventas en línea.

Aunque la estrategia de e-commerce de E-Corp ha estado en marcha durante un año, los informes recientes revelan que las ventas generadas a través de este canal no alcanzan el nivel esperado por la dirección de la empresa. Este bajo rendimiento ha llevado a cuestionamientos sobre la eficacia de las campañas de marketing digital y la optimización de la inversión en publicidad en línea.

Para abordar este desafío, la dirección de E-Corp ha decidido recurrir a modelos de aprendizaje automático (ML) como una solución potencial. En particular, desean utilizar el poder predictivo de estos modelos para anticiparse a la identificación de clientes potenciales y así optimizar su inversión en pauta digital. Esto implica una revisión exhaustiva de los datos disponibles y la implementación de estrategias inteligentes para dirigir las campañas de marketing digital hacia los segmentos de clientes más propensos a realizar compras en el sitio web.

En este contexto, este documento propone un enfoque integral para abordar el problema de E-Corp, que incluye análisis exploratorio de datos, selección de características, entrenamiento de modelos de ML (como Regresión Logística, Random Forest y Gradient Boosting), y finalmente, la implementación de estrategias de segmentación para optimizar el gasto en pauta digital. Este

enfoque tiene como objetivo principal ayudar a E-Corp a mejorar su rendimiento en línea y maximizar el retorno de la inversión en su estrategia de e-commerce de lujo.

DESARROLLO

LIMPIEZA Y TRANSFORMACIÓN DE LOS DATOS

Tomando entonces como base principal el dataset proporcionado por la empresa, ecommerce-data, se hace su respectiva importación de variables, contando entonces con:

1. **Reviews:** Este es el número de páginas de este tipo (Reviews) que visitó el usuario.
2. **Reviews_Duration:** Esta es la cantidad de tiempo dedicado a esta categoría de páginas.
3. **Informational:** Este es el número de páginas de este tipo (informativas) que visitó el usuario.
4. **Informational_Duration:** Esta es la cantidad de tiempo dedicado a esta categoría de páginas.
5. **ProductRelated:** Este es el número de páginas de este tipo (relacionadas con productos) que visitó el usuario.
6. **ProductRelated_Duration:** Esta es la cantidad de tiempo dedicado a esta categoría de páginas.
7. **BounceRates:** Métrica arrojada por Google Analytics que indica el porcentaje de visitantes que ingresan al sitio web a través de una página y salen sin realizar ninguna acción adicional durante la sesión.
8. **ExitRates:** Métrica arrojada por Google Analytics que indica el porcentaje de páginas vistas en el sitio web que terminan en esa página específica.

9. **PageValues:** Métrica arrojada por Google Analytics que representa el valor medio de una página web que un usuario visitó antes de completar una transacción de comercio electrónico.
10. **SpecialDay:** Este valor representa la proximidad de la fecha de navegación en el sitio a días especiales o festivos (por ejemplo, el Día de la Madre o San Valentín).
11. **Month:** Mes en el que se realizó la visita al sitio web.
12. **OperatingSystems:** Sistema operativo usado por el usuario para navegar en el sitio web.
13. **Browser:** Navegador usado por el usuario para navegar en el sitio web.
14. **Region:** Región (ubicación geográfica personalizada) desde la cual el usuario navega en el sitio web.
15. **TrafficType:** Variable que indica el tipo de tráfico al cual pertenece el usuario que navega en el sitio web (por ejemplo, si llegó al sitio desde un anuncio o a través de una búsqueda).
16. **VisitorType:** Tipo de usuario que ingresa al sitio web.
17. **Weekend:** Indica si la navegación se realizó en fin de semana.
18. **Purchase:** Indica si el usuario realizó una compra o no.

Sumando entonces un total de 12,330 registros en el último año. Además, se cuenta con un total de 18 variables asociadas al caso de estudio, la mayoría de las cuales son variables numéricas y el resto categóricas. La variable objetivo (y) es 'Purchase', la cual es de tipo 'bool' o también puede ser denominada una variable dicotómica.

Siguiendo con la estructura de esta base de datos, se comprende que se trabajarán con modelos de clasificación que, se espera, ayudarán a predecir mejor el comportamiento de los datos predictivos. Para ello, es necesario tener en cuenta los tipos de variables disponibles, por lo que

se realiza una separación de estas, agrupándolas en categóricas y numéricas, tal y como se muestra en la Tabla 1.

```
num_var = ['Reviews',      'Reviews_Duration', 'Informational',
       'Informational_Duration',  'ProductRelated',   'ProductRelated_Duration',
       'BounceRates',      'ExitRates', 'PageValues','SpecialDay']


cat_var   =   ['Month',   'OperatingSystems',   'Browser',   'Region',   'TrafficType',
'VisitorType', 'Weekend', 'Purchase']
```

Tabla 1. Separación de variables.

En este contexto, se decidió trabajar con la variable 'Month' como una variable categórica ordinal, mientras que las demás variables categóricas que no eran binarias fueron dummyficadas. La dummyficación es una técnica comúnmente empleada para codificar variables categóricas en variables binarias (0 o 1), lo que permite que sean utilizadas por algoritmos de aprendizaje automático.

La dummyficación implica la creación de variables ficticias para cada categoría presente en la variable original. Cada variable ficticia toma el valor de 1 si la observación pertenece a esa categoría y 0 si no. Este proceso se lleva a cabo con el fin de evitar que el algoritmo de aprendizaje automático asuma un orden o jerarquía entre las categorías.

ANÁLISIS EXPLORATORIO DE DATOS

Para el caso de estudio, resulta crucial comprender el comportamiento de cada una de las variables. Por este motivo, se generaron gráficos individuales para visualizar su distribución, tal y como se muestra en la Figura 1.

Se observa que la mayoría de las variables exhiben asimetrías hacia la derecha, atribuibles a la variabilidad inherente en los datos. No obstante, más allá de esta asimetría, se destacan ciertos gráficos que revelan información relevante. Específicamente, se identifican diferencias significativas en las densidades entre las clases 'Compra' y 'No Compra'.

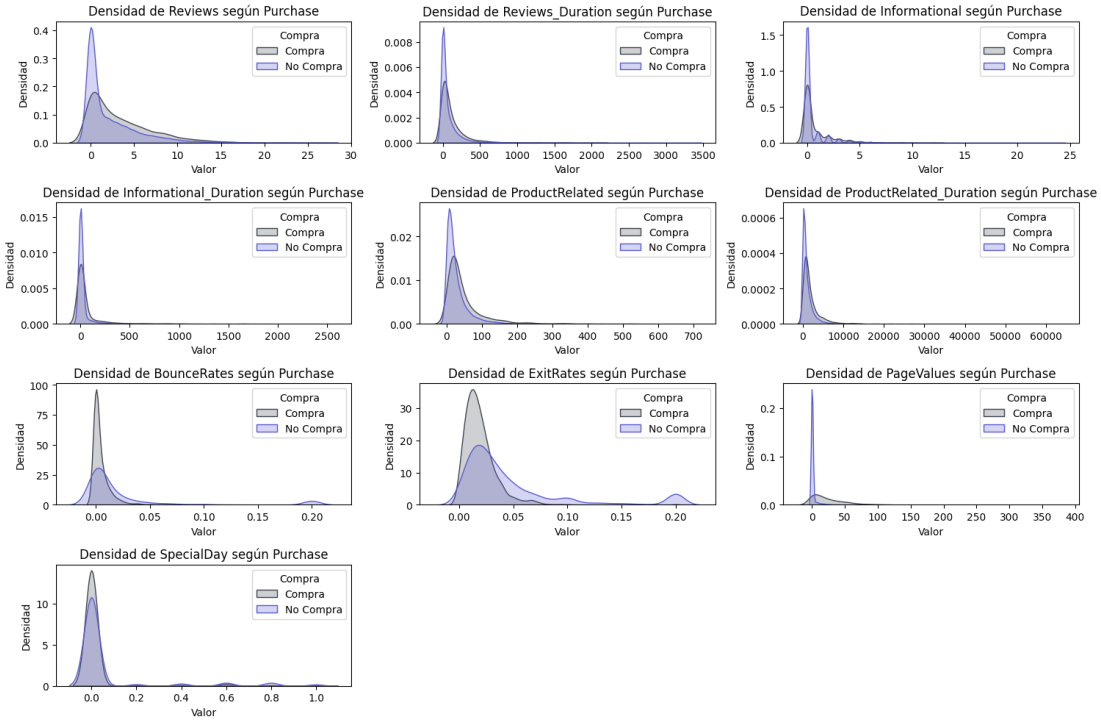

Fig.1 Gráficos individuales de distribución

- **Densidad de Reviews según Purchase:** Se observa una clara diferencia en las densidades entre las clases 'Compra' y 'No Compra', especialmente en valores bajos de 'Reviews'. La mayoría de las compras parecen ocurrir después de un número relativamente bajo de revisiones.
- **Densidad de ProductRelated según Purchase:** Similar a la variable 'Reviews', hay una clara diferencia en las densidades entre las clases 'Compra' y 'No Compra', con una mayor densidad para la clase 'Compra' en valores bajos de 'ProductRelated'. Esto sugiere que la mayoría de las compras se realizan después de interactuar con un número bajo de páginas relacionadas con productos.
- **Densidad de BounceRates según Purchase:** Aunque las diferencias no son tan marcadas como en las variables anteriores, se puede observar una ligera diferencia en las densidades entre las clases 'Compra' y 'No Compra', especialmente en valores más altos de 'BounceRates'. Esto sugiere que las tasas de rebote pueden influir en la probabilidad de compra de manera diferencial entre las dos clases.
- **Densidad de PageValues según Purchase:** Hay una clara diferencia en las densidades entre las clases 'Compra' y 'No Compra', especialmente en valores bajos de 'PageValues'. Esto indica que los valores de página más bajos están asociados con una mayor probabilidad de compra.
- **'SpecialDay':** Se puede evidenciar que las estrategias que esté usando la empresa para comercializar o promocionar en días festivos no está teniendo efecto, ya que la mayoría de los clientes compran en días ordinarios.

Estas gráficas proporcionan información relevante sobre las diferencias en el comportamiento entre las clases 'Compra' y 'No Compra', y pueden ser útiles para comprender mejor los factores que influyen en la realización de compras en el sitio web.

ANÁLISIS DE CORRELACIÓN

A continuación, en la Figura 2, se observa la matriz de correlación, la cual muestra las correlaciones lineales entre las variables numéricas del conjunto de datos. Su propósito es identificar correlaciones positivas o negativas entre las variables o directamente con la variable objetivo (y).

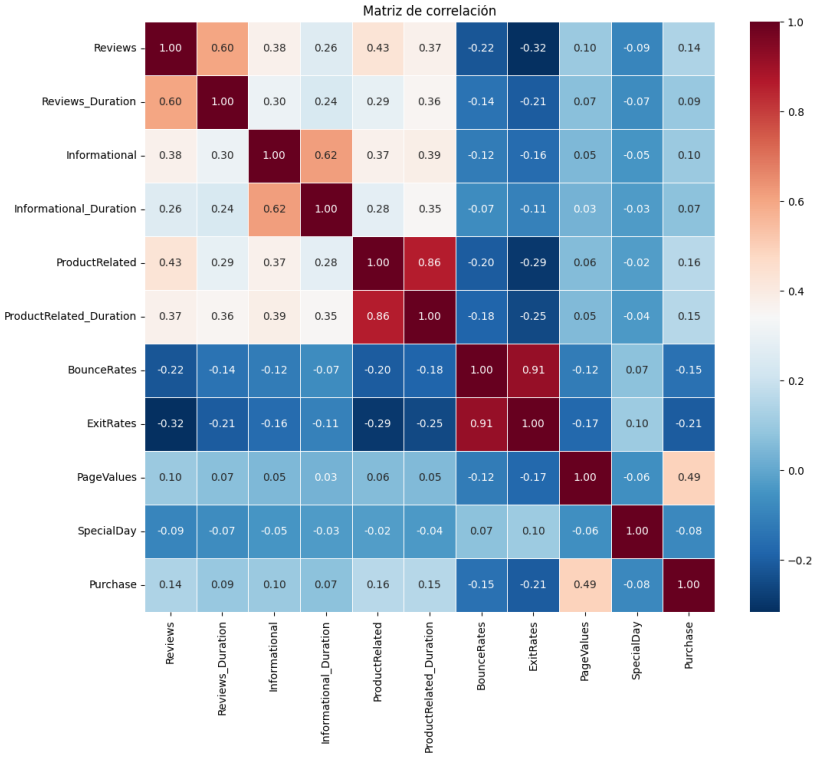

Fig.2 Matriz de correlación

- Las variables que llevan asociadas una duración tienen una correlación considerable entre sí, por lo cual se podrían suprimir estas variables para evitar problemas de multicolinealidad.
- La variable "PageValues" tiene una correlación de 0.49 con "Purchase", lo que sugiere una correlación moderadamente fuerte.
- 'BounceRates' y 'ExitRates' tienen una correlación bastante alta (0.91). Podríamos considerar eliminar una de ellas para evitar la multicolinealidad.

Se realizó una prueba chi-cuadrado (χ^2) para determinar la significancia de las variables categóricas, detallando los resultados en la Tabla 2.

Variable	Chi2	P-value	Significativo
Month	384934762	2,24E-71	Sí
OperatingSystems	75027056	1,42E-13	Sí
Browser	27715299	6,09E-03	Sí
Región	9252751	3,21E-01	No
Traffic Type	373145565	1,65E-67	Sí
Visitor Type	135251923	4,27E-30	Sí
Weekend	10390978	1,27E-03	Sí

Tabla 2. Prueba chi-cuadrado (χ^2)

De acuerdo con los resultados, la única variable que no se relaciona puede ser la variable asociada a la región, por lo que se puede considerar eliminar dicha variable.

Otra de las observaciones que se tienen en cuenta es que, la variable 'VisitorType' proporciona información valiosa que puede ser relevante para entender el comportamiento del cliente y optimizar las estrategias de marketing y la experiencia del usuario en el sitio web.

A continuación, en la Figura 3, se observa que los visitantes que regresan ('Returning_Visitor') tienen una proporción mucho más alta de compras en comparación con los visitantes nuevos ('New_Visitor') y otros tipos de visitantes ('Other'). Esta información es valiosa para comprender el comportamiento de compra de los diferentes tipos de visitantes. Puede ser útil para adaptar estrategias de marketing y experiencia del usuario para maximizar las conversiones, especialmente entre los visitantes que regresan y debido a que los visitantes que regresan muestran una mayor propensión a comprar, se pueden desarrollar estrategias específicas para fomentar la lealtad y la retención de estos clientes.

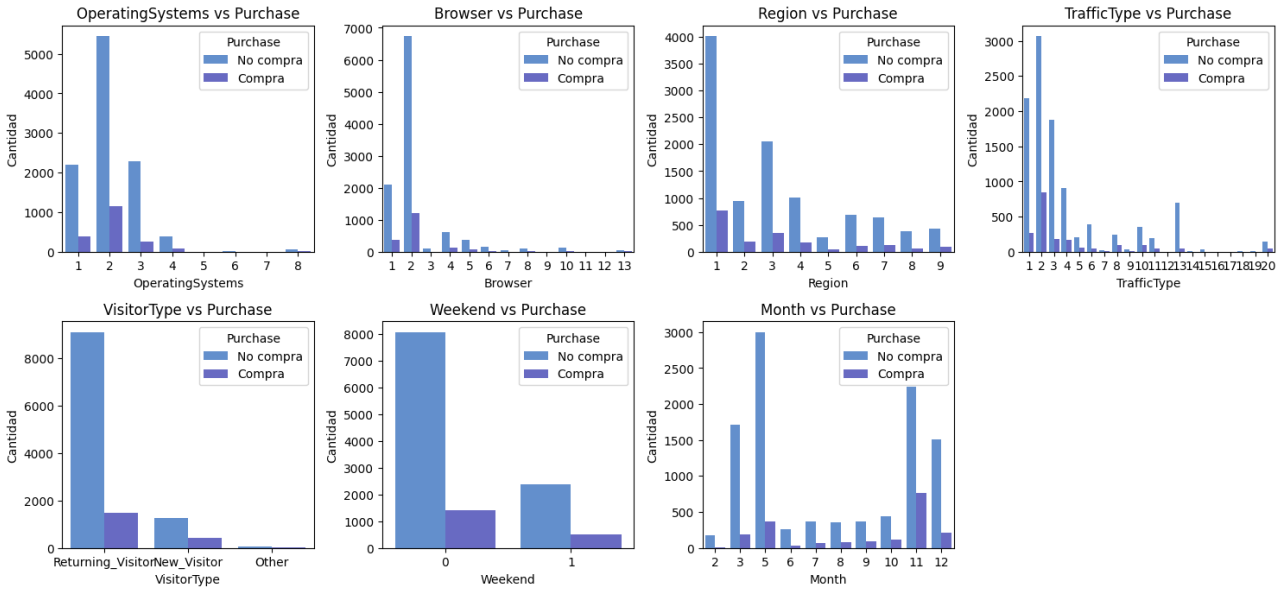

Fig.3 Barras Agrupadas Variables categóricas

DUMMIZAR LAS VARIABLES

Teniendo en cuenta esta información, y antes de eliminar alguna de las variables, se procede a hacer un reajuste en la base de datos con el fin de tener todas las variables en valores numéricos para facilitar la elaboración de modelos predictivos. Para este paso se utilizó la función get_dummies.

PREPARACIÓN DE LOS DATOS:

Para la preparación de los datos, entonces, se utilizó una proporción 20%-80% para los datos de prueba y entrenamiento, respectivamente, contando con un total de:

- Tamaño del conjunto de entrenamiento: (9864, 65).
- Tamaño del conjunto de validación (2466, 65).

SELECCIÓN DE VARIABLES

Para la selección de variables se tuvieron en cuenta algunos métodos, incluidos el Select K Best, RFE y Select From Model, tal y como se observa en la Tabla 3, los cuales dieron los siguientes resultados, ayudando a identificar algunas variables que ayudan a describir mejor el caso de estudio, entre ellas:

Variables Principales	SKB	RFE	SFM
Reviews	x	x	x
Reviews_Duration	x	x	x
Informational	x	x	x
ProductRelated	x	x	x
ProductRelated_Duration	x	x	x
BounceRates	x	x	x
ExitRates	x	x	x
PageValues	x	x	x
SpecialDay	x		
Month	x	x	x
OperatingSystems_3	x		
TrafficType_2	x	x	
TrafficType_3	x		
VisitorType_New_Visitor	x		
VisitorType_Returning_Visitor	x	x	
Informational_Duration		x	x
OperatingSystems_2		x	
Browser_2		x	
Region_1		x	

Tabla 3. Selección de variables

Teniendo en cuenta dichas variables, se procede a crear una nueva base de datos, seleccionando algunas variables, basado en los resultados de los métodos de filtrado y el análisis exploratorio de datos.

Con esta nueva base de datos se logra reducir las características a: Purchase Reviews Informational ProductRelated BounceRates ExitRates PageValues Month Weekend VisitorType_New_Visitor VisitorType_Other VisitorType_Returning_Visitor.

MODELOS PROPUESTOS

1. REGRESIÓN LOGÍSTICA

La regresión logística es una elección adecuada debido a su capacidad para abordar problemas de clasificación binaria. Este modelo es especialmente útil cuando se trabaja con conjuntos de datos con **pocas características**, ya que es fácil de interpretar y entender. Además, la regresión logística tiende a ser **menos propensa al sobreajuste**, lo que la hace útil cuando se dispone de datos limitados. Sin embargo, puede no funcionar tan bien en casos donde las relaciones entre las características y el resultado son altamente no lineales. Para trabajar con este modelo, entonces, se realizó la estandarización de las variables y se eliminaron las características que estaban correlacionadas, para así continuar con la evaluación del modelo.

a. Todas las características (Modelo Base)

Métricas de desempeño sobre el conjunto de entrenamiento:

Accuracy Score: 82.55%

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.84	0.89	8319
1	0.47	0.76	0.58	1545
accuracy			0.83	9864
macro avg	0.71	0.80	0.73	9864
weighted avg	0.87	0.83	0.84	9864

El análisis de los resultados revela un buen desempeño general del modelo en la predicción de compras en el sitio web de E-Corp. Con un Accuracy Score del 82.55%, el modelo logra predecir correctamente la mayoría de las instancias. Sin embargo, se observa una disparidad en la precisión entre las clases "Compra" y "No Compra", con una precisión del 95% para la clase "No Compra" y del 47% para la clase "Compra". A pesar de esto, el modelo logra una recuperación (recall) aceptable para ambas clases, con valores del 84% y 76% respectivamente. Esto sugiere que el modelo es eficaz para identificar correctamente las compras, aunque hay margen para mejorar la precisión en la predicción de esta clase.

```
Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 81.79%
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.83      0.89      2103
           1       0.43      0.75      0.55       363
    accuracy                           0.82      2466
   macro avg       0.69      0.79      0.72      2466
weighted avg       0.87      0.82      0.84      2466
```

El análisis de los resultados sobre el conjunto de validación muestra un desempeño consistente del modelo, con un Accuracy Score del 81.79%. Aunque la precisión para la clase "No Compra" sigue siendo alta (95%), se observa una mejora en la precisión para la clase "Compra" en comparación con el conjunto de entrenamiento, alcanzando el 43%. Sin embargo, la recuperación para esta clase es del 75%, lo que sugiere que el modelo sigue siendo efectivo para identificar la mayoría de las compras, aunque con cierto margen de mejora en la precisión.

b. Características Filtradas (Modelo Ajustado)

```
Métricas de desempeño sobre el conjunto de entrenamiento:
Accuracy Score: 86.92%
Classification Report:
```

	precision	recall	f1-score	support
0	0.95	0.89	0.92	8367
1	0.55	0.74	0.63	1497
accuracy			0.87	9864
macro avg	0.75	0.82	0.78	9864
weighted avg	0.89	0.87	0.88	9864

El análisis de los resultados sobre el conjunto de entrenamiento revela un desempeño sólido del modelo, con un Accuracy Score del 86.92%. Se observa una alta precisión del 95% para la clase "No Compra", indicando que el modelo es efectivo en la predicción de esta clase. Sin embargo, la precisión para la clase "Compra" es del 55%, lo que sugiere que hay margen para mejorar la precisión en la identificación de compras. La recuperación (recall) para la clase "Compra" es del 74%, lo que indica que el modelo es capaz de identificar la mayoría de las compras, aunque con cierta tasa de falsos positivos.

El análisis de los resultados sobre el conjunto de validación muestra un sólido desempeño del modelo, con un Accuracy Score del 86.86%. Se destaca una alta precisión del 94% para la clase "No Compra", lo que indica que el modelo es efectivo en la predicción de esta clase. Además, la precisión para la clase "Compra" es del 58%, lo que sugiere una mejora en la precisión en la identificación de compras en comparación con el conjunto de entrenamiento. La recuperación (recall) para la clase "Compra" es del 74%, lo que indica que el modelo es capaz de identificar la mayoría de las compras, aunque con cierta tasa de falsos positivos.

Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 86.86%
Classification Report:

	precision	recall	f1-score	support
0	0.94	0.89	0.92	2055
1	0.58	0.74	0.65	411
accuracy			0.87	2466
macro avg	0.76	0.82	0.79	2466
weighted avg	0.88	0.87	0.87	2466

2. RANDOM FOREST

Los bosques aleatorios son modelos de ensamble que combinan múltiples árboles de decisión. Son una opción sólida para **problemas de clasificación binaria** debido a su capacidad para manejar una variedad de conjuntos de datos y tipos de características. Los bosques aleatorios tienden a tener un **rendimiento robusto** y son menos propensos al sobreajuste que algunos otros modelos, gracias a la diversidad de los árboles individuales en el ensamble.

El análisis de los resultados sobre el conjunto de entrenamiento revela un rendimiento perfecto del modelo, con un Accuracy Score del 100%. Esto indica que el modelo fue capaz de clasificar correctamente todas las instancias del conjunto de entrenamiento. Tanto la precisión como la recuperación (recall) para ambas clases son del 100%, lo que sugiere una clasificación impecable de las clases "No Compra" y "Compra".

a. Todas las características (Modelo Base)

```
Métricas de desempeño sobre el conjunto de entrenamiento:
Accuracy Score: 100.00%
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      8319
           1       1.00      1.00      1.00      1545

    accuracy                           1.00      9864
   macro avg       1.00      1.00      1.00      9864
weighted avg       1.00      1.00      1.00      9864
```

```
Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 90.11%
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.97      0.94      2103
           1       0.75      0.49      0.59       363

    accuracy                           0.90      2466
   macro avg       0.83      0.73      0.77      2466
weighted avg       0.89      0.90      0.89      2466
```

El análisis de los resultados sobre el conjunto de validación muestra un buen rendimiento del modelo, con un Accuracy Score del 90.11%. La precisión para la clase "No Compra" es del 92%, lo que indica que el 92% de las predicciones de "No Compra" fueron correctas, mientras que para la clase "Compra" es del 75%. El recall para la clase "No Compra" es del 97%, lo que sugiere que el 97% de las instancias de "No Compra" fueron correctamente identificadas, mientras que para la clase "Compra" es del 49%, indicando que el modelo tiene dificultades para identificar todas las instancias de "Compra". El valor F1-score, que combina precisión y recall, es del 94% para la clase "No Compra" y del 59% para la clase "Compra".

b. Características Filtradas (Modelo Ajustado)

```
Métricas de desempeño sobre el conjunto de entrenamiento:
Accuracy Score: 93.63%
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.94      0.96      8319
           1       0.74      0.93      0.82      1545

    accuracy                           0.94      9864
   macro avg       0.86      0.93      0.89      9864
weighted avg       0.95      0.94      0.94      9864
```

El análisis de los resultados sobre el conjunto de entrenamiento revela un buen rendimiento del modelo, con un Accuracy Score del 93.63%. La precisión para la clase "No Compra" es del 99%, lo que indica que el 99% de las predicciones de "No Compra" fueron correctas, mientras que para la clase "Compra" es del 74%. El recall para la clase "No Compra" es del 94%, lo que sugiere que el 94% de las instancias de "No Compra" fueron correctamente identificadas, mientras que para la clase "Compra" es del 93%, indicando que el modelo tiene un buen rendimiento en la identificación de la clase "Compra". El valor F1-score, que combina precisión y recall, es del 96% para la clase "No Compra" y del 82% para la clase "Compra".

```
Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 88.65%
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.91      0.93      2103
           1       0.59      0.75      0.66       363
    accuracy                           0.89      2466
   macro avg       0.77      0.83      0.80      2466
weighted avg       0.90      0.89      0.89      2466
```

El análisis de los resultados sobre el conjunto de validación muestra un buen rendimiento del modelo, con un Accuracy Score del 88.65%. La precisión para la clase "No Compra" es del 96%, lo que indica que el 96% de las predicciones de "No Compra" fueron correctas, mientras que para la clase "Compra" es del 59%. El recall para la clase "No Compra" es del 91%, lo que sugiere que el 91% de las instancias de "No Compra" fueron correctamente identificadas, mientras que para la clase "Compra" es del 75%, indicando que el modelo tiene un buen rendimiento en la identificación de la clase "Compra". El valor F1-score, que combina precisión y recall, es del 93% para la clase "No Compra" y del 66% para la clase "Compra".

3. GRADIENT BOOSTING

El aumento de gradiente es otro modelo de ensamble que combina múltiples árboles de decisión, pero de forma secuencial. Es conocido por su capacidad para mejorar el rendimiento del modelo con respecto a los árboles de decisión individuales. Gradient Boosting es especialmente útil en conjuntos de datos grandes y complejos, donde puede capturar relaciones no lineales entre las características y el resultado. Sin embargo, es más propenso al sobreajuste que Random Forest y puede requerir una mayor atención a la regularización y ajuste de hiperparámetros.

a. Todas las características (Modelo Base)

```
Métricas de desempeño sobre el conjunto de entrenamiento:

Accuracy Score: 91.68%

Classification Report:

              precision    recall f1-score   support


           0       0.94      0.97      0.95      8319

           1       0.78      0.65      0.71      1545
    accuracy                           0.92      9864

   macro avg       0.86      0.81      0.83      9864

weighted avg       0.91      0.92      0.91      9864
```

El análisis de los resultados revela un desempeño satisfactorio del modelo en el conjunto de entrenamiento, con una precisión global del 91.68%. Se observa una alta precisión para la clase 0 y una sensibilidad aceptable para la clase 1, aunque podría mejorarse.

```
Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 90.27%
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.96      0.94      2103
           1       0.70      0.60      0.64       363

    accuracy                           0.90      2466
   macro avg       0.82      0.78      0.79      2466
weighted avg       0.90      0.90      0.90      2466
```

El análisis de los resultados en el conjunto de validación muestra un rendimiento sólido del modelo, con una precisión general del 90.27%. Se destaca una buena precisión para la clase 0, mientras que la sensibilidad para la clase 1 podría ser mejorada. Aunque el modelo muestra una capacidad general para predecir correctamente las clases en el conjunto de datos de validación, aún existen áreas que podrían ser refinadas para mejorar el rendimiento en la predicción de la clase minoritaria.

4. GRADIENT BOOSTING XTREMME

XGBoost es una variante mejorada del algoritmo de aumento de gradiente, diseñada para mejorar la eficiencia y el rendimiento. Ofrece muchas de las mismas ventajas que Gradient Boosting, como la capacidad para manejar relaciones complejas entre características y resultado, pero con una

mayor eficiencia computacional. XGBoost es particularmente útil en conjuntos de datos grandes y complejos, donde puede proporcionar resultados precisos con tiempos de entrenamiento más cortos. Sin embargo, al igual que otros modelos de ensamble, puede requerir ajuste de hiperparámetros para obtener el mejor rendimiento posible.

a. Todas las características (Modelo Base)

```
Métricas de desempeño sobre el conjunto de entrenamiento:
Accuracy Score: 98.60%
Classification Report:
              precision    recall f1-score   support

           0       0.99      1.00      0.99      8319
           1       0.99      0.92      0.95      1545

    accuracy                           0.99      9864
   macro avg       0.99      0.96      0.97      9864
weighted avg       0.99      0.99      0.99      9864
```

El modelo muestra un excelente rendimiento en el conjunto de entrenamiento, con una precisión general del 98.60%. Se destaca una alta precisión y sensibilidad para ambas clases, lo que indica una capacidad significativa para predecir correctamente tanto las instancias positivas como negativas.

```
Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 89.78%
Classification Report:
              precision    recall f1-score   support

           0       0.93      0.95      0.94      2103
           1       0.68      0.57      0.62       363

    accuracy                           0.90      2466
   macro avg       0.81      0.76      0.78      2466
weighted avg       0.89      0.90      0.89      2466
```

El modelo muestra un buen desempeño en el conjunto de validación, con una precisión general del 89.78%. Aunque la precisión para la clase negativa es alta, la precisión para la clase positiva es relativamente menor.

b. Características Filtradas (Modelo Ajustado)

```
Métricas de desempeño sobre el conjunto de entrenamiento:
Accuracy Score: 91.59%
Classification Report:
              precision    recall f1-score   support

           0       0.94      0.97      0.95      8319
           1       0.78      0.65      0.71      1545

    accuracy                           0.92      9864
   macro avg       0.86      0.81      0.83      9864
weighted avg       0.91      0.92      0.91      9864
```

El modelo ajustado muestra un buen desempeño en el conjunto de entrenamiento, con una precisión general del 91.59%. Sin embargo, al igual que en el modelo original, la precisión para la clase positiva es relativamente menor, lo que sugiere que el modelo puede tener dificultades para identificar correctamente las instancias positivas.

```
Métricas de desempeño sobre el conjunto de validación:
Accuracy Score: 90.47%
Classification Report:
              precision    recall f1-score   support

           0       0.94      0.95      0.94      2103
           1       0.70      0.62      0.66       363

    accuracy                           0.90      2466
   macro avg       0.82      0.79      0.80      2466
weighted avg       0.90      0.90      0.90      2466
```

El modelo ajustado también muestra un buen desempeño en el conjunto de validación, con una precisión general del 90.47%. Sin embargo, al igual que en el conjunto de entrenamiento, la precisión para la clase positiva sigue siendo relativamente menor en comparación con la clase negativa.

Teniendo en cuenta los resultados de cada modelo, se podría trabajar con un modelo Radiant Forest o el XGBoost, ya que tienen el mejor comportamiento, sin embargo se preferiría el modelo XGBoost ya que asemeja más los comportamientos tanto de entrenamiento como de prueba, sin embargo a continuación se muestran las matrices y gráficas comparativas que ayudan a visualizar mejor los datos.

MODELO RANDOM FOREST

Según lo representado en la Figura 4, tanto en el conjunto de entrenamiento como en el de validación, se evidencia una tendencia consistente en la matriz de confusión. Ambos modelos exhiben un rendimiento satisfactorio en términos generales en la clasificación. Sin embargo, se resalta la importancia de mejorar la capacidad para identificar de manera correcta las instancias positivas, lo que podría contribuir a la disminución de la tasa de falsos negativos.

En última instancia, la empresa debe utilizar estos insights para tomar decisiones informadas que optimicen su estrategia de comercio electrónico y maximicen el retorno de inversión en marketing digital.

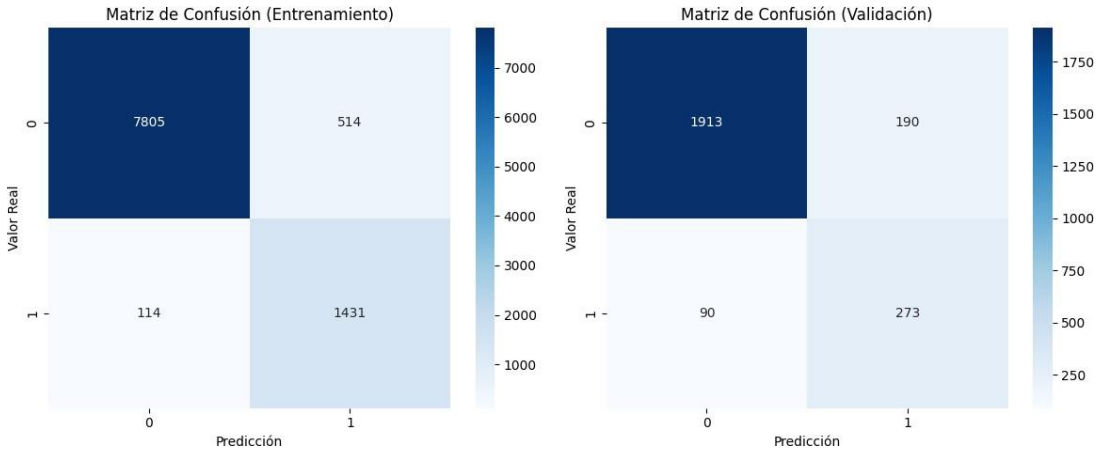

Fig.4 Matriz de confusión modelo Random Forest

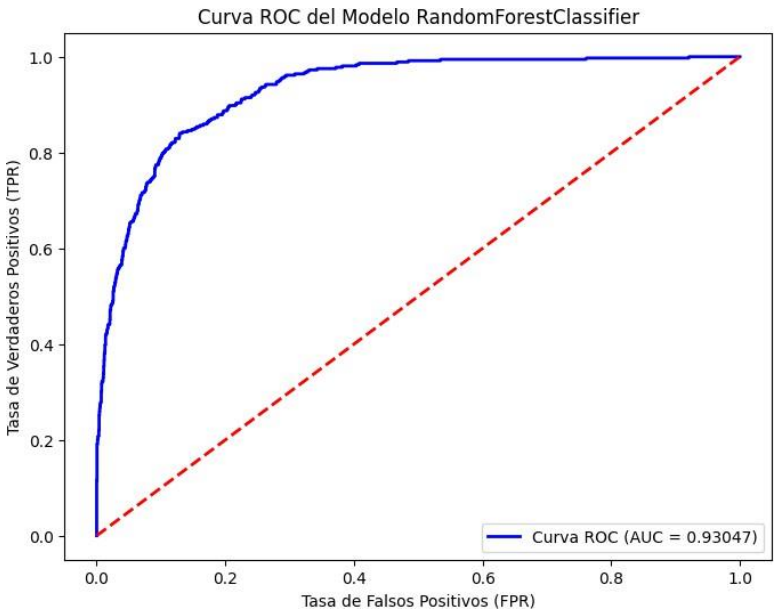

Fig.5 Curva ROC modelo Random Forest

Tal y como se observa en la Figura 5, un AUC de 0.93047, en el modelo Random Forest Classifier muestra una capacidad de discriminación bastante sólida entre las clases positivas y negativas. Lo que sugiere que el modelo es efectivo para clasificar las instancias correctamente en sus respectivas clases y que tiene un buen equilibrio entre la sensibilidad y la especificidad.

MODELO XGBOOST

Según lo representado en la gráfica 6, tanto el conjunto de entrenamiento y validación muestran una capacidad general para clasificar correctamente las instancias, aunque cuenta con un alto número de verdaderos negativos y positivos.

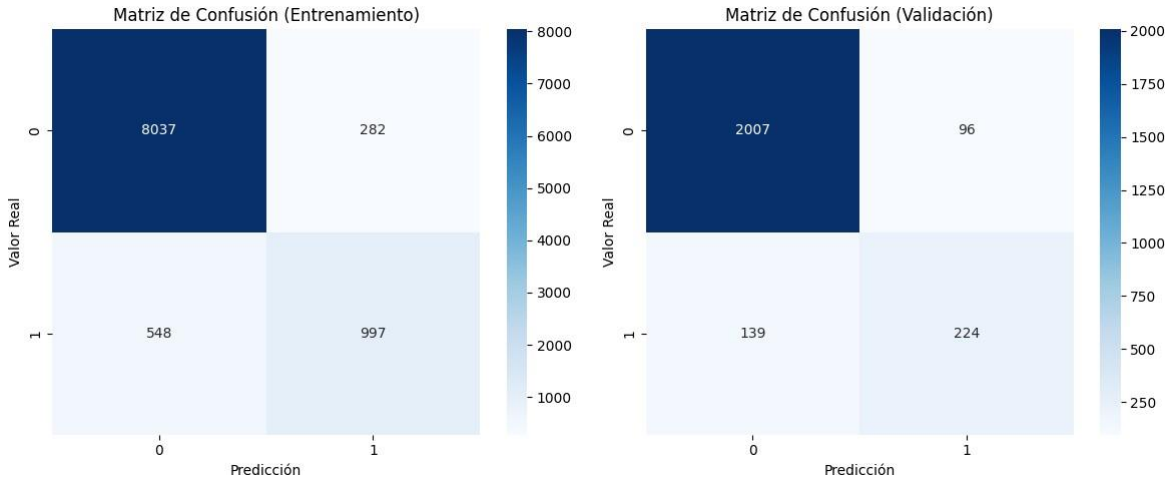

Fig 6. Matriz de confusión modelo XGBOOST

Con esto la empresa puede ajustar sus estrategias de marketing digital para dirigirse de manera más efectiva a los segmentos de clientes identificados como propensos a realizar compras en el sitio web. Ya que implica una optimización de la inversión en publicidad en línea y la personalización de mensajes y ofertas. Además, se puede mejorar la experiencia del usuario en el sitio web para aumentar la conversión de clientes, lo que incluye la optimización del diseño del sitio y la facilitación del proceso de compra.

La curva ROC en la Figura 7, muestra un área bajo la curva (AUC) de 0.93034 para el modelo XGBoost, lo que indica un buen rendimiento en la capacidad de clasificación del modelo. Dado que, cuanto más cercano esté el AUC a 1, mejor será el modelo en distinguir entre las clases positivas y negativas. En este caso, el AUC sugiere que el modelo XGBoost tiene una alta

capacidad para clasificar correctamente las instancias positivas y negativas en el conjunto de datos. Lo que puede ser útil para la empresa al identificar y dirigirse de manera efectiva a los clientes potenciales en sus campañas de marketing digital.

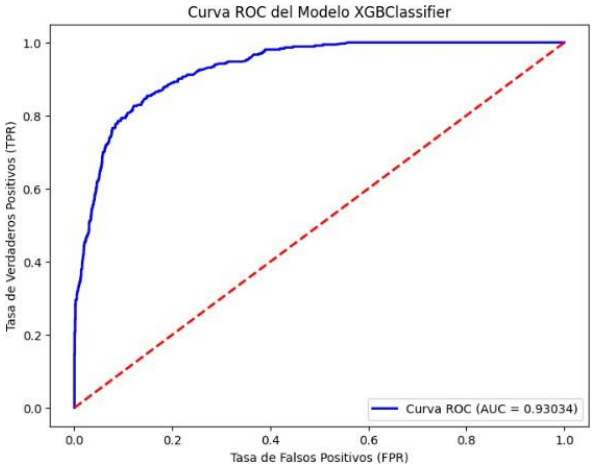

Fig. 7 Curva ROC modelo XGBOOST

AJUSTE DE SUBMUESTREO XGBOOST

```
Métricas de desempeño sobre el conjunto de entrenamiento:
Accuracy Score: 86.63%
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.88      0.87      1523
           1       0.87      0.86      0.87      1529

    accuracy                           0.87      3052
   macro avg       0.87      0.87      0.87      3052
weighted avg       0.87      0.87      0.87      3052
```

Las métricas de desempeño sobre el conjunto de entrenamiento muestran un puntaje de precisión global del 86.63%. Con una precisión media del 87% para ambas clases, el modelo demuestra un buen equilibrio entre la capacidad de predecir correctamente las instancias positivas y negativas.

```
Métricas de desempeño sobre el conjunto de validación:

Accuracy Score: 84.95%

Classification Report:

              precision    recall  f1-score   support

           0       0.87      0.82      0.85       385
           1       0.83      0.88      0.85       379

    accuracy                           0.85       764
   macro avg       0.85      0.85      0.85       764
weighted avg       0.85      0.85      0.85       764
```

Las métricas de desempeño sobre el conjunto de validación indican un puntaje de precisión global del 84.95%. El reporte de clasificación muestra que tanto la precisión, el recall y el f1-score son equilibrados para ambas clases, con valores cercanos al 85%. Esto sugiere que el modelo es capaz de generalizar bien a datos no vistos y mantener un rendimiento consistente en la predicción de instancias positivas y negativas.

Las matrices de confusión tanto en el conjunto de entrenamiento como en el de validación para el Ajuste de Submuestreo XGBoost, muestran un buen desempeño en la clasificación de verdaderos positivos y negativos, tal y como se observa en la Figura 8.

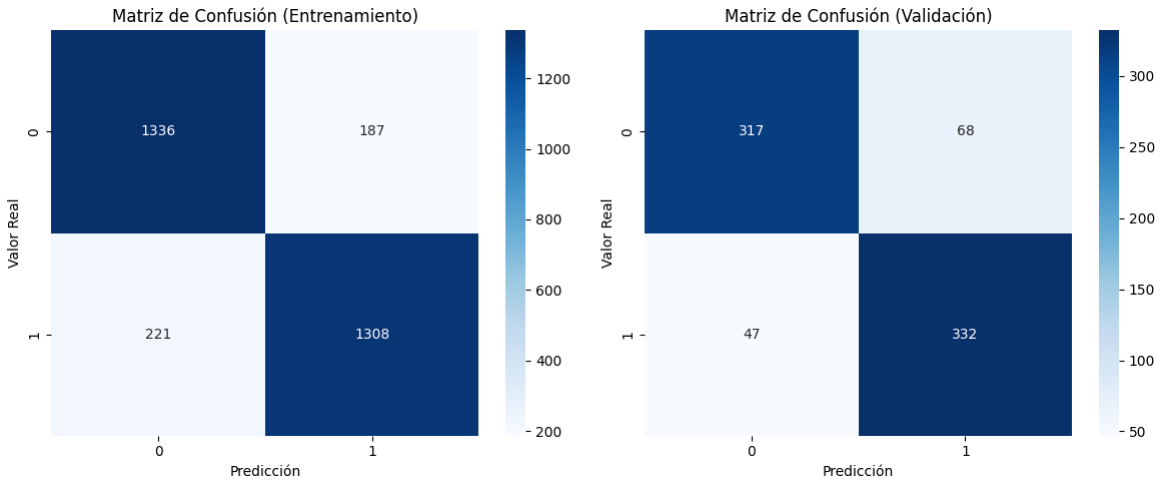

Fig.8 Matriz de confusión modelo XGBOOST

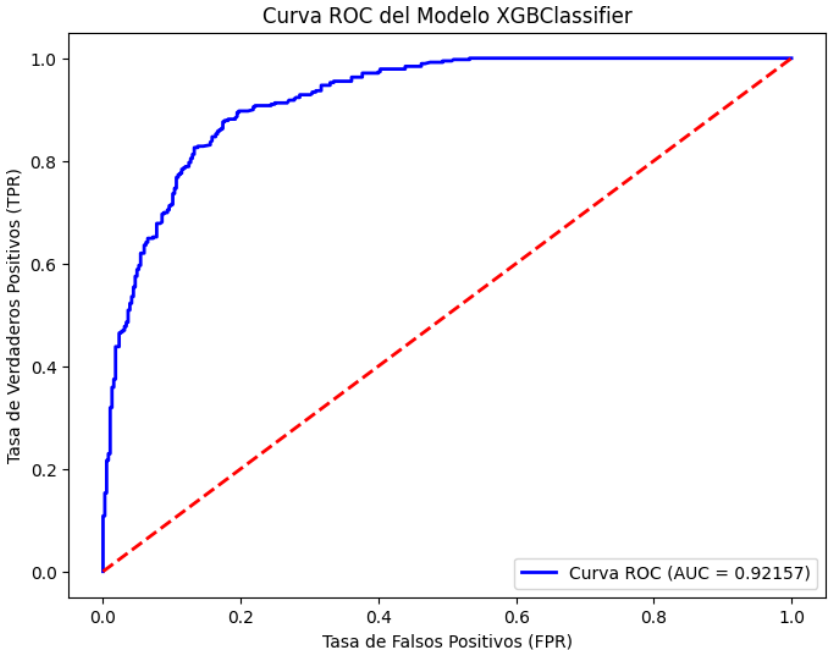

Fig.9 Curva ROC modelo XGBOOST

Como se observa en la Figura 9, una Curva ROC (AUC) de 0.92157, muestra un buen rendimiento en la capacidad de clasificación. Ya que indica que el modelo es capaz de distinguir entre las clases positivas y negativas con un alto grado de precisión. Un valor de AUC cercano a 1 sugiere que el modelo tiene una alta tasa de verdaderos positivos y una baja tasa de falsos positivos, lo cual es deseable para la mayoría de los casos de clasificación.

CONCLUSIONES

Tras un minucioso análisis, hemos llegado a la conclusión de que el modelo XGBoost exhibe un rendimiento excepcional en la clasificación de instancias tanto positivas como negativas, destacando por su capacidad notable para discriminar entre clases. Este rendimiento se ha mantenido constante tanto en los conjuntos de entrenamiento como en los de validación, lo que

sugiere su habilidad para generalizar efectivamente a datos no vistos y mantener un rendimiento robusto en la predicción de compras en el sitio web de E-Corp.

Durante la investigación, se exploró el impacto del ajuste de submuestreo en el rendimiento del modelo XGBoost, lo que demostró ser una solución efectiva para abordar el desequilibrio de clases en el conjunto de datos. Este enfoque mejoró significativamente la precisión en la clasificación de instancias minoritarias, señalando su potencial aplicabilidad en escenarios similares de desequilibrio de clases.

Los resultados obtenidos ofrecen valiosos insights para la toma de decisiones en el ámbito del comercio electrónico y el marketing digital. Dichos insights podrían ser empleados para optimizar estrategias de publicidad en línea, personalizar mensajes y ofertas, y mejorar la experiencia del usuario en el sitio web, todo ello con el objetivo de incrementar la conversión de clientes para E-Corp.

A pesar del sólido rendimiento demostrado por el modelo XGBoost, se sugiere la exploración de otras técnicas de modelado o ajustes de hiperparámetros con el fin de mejorar aún más la precisión y el recall. Además, es recomendable mantenerse alerta ante posibles cambios en el conjunto de datos o en el entorno empresarial que puedan requerir actualizaciones en el modelo en el futuro.

En resumen, este análisis proporciona una base sólida para la implementación de un modelo predictivo en el contexto del comercio electrónico de E-Corp. Estos hallazgos son esenciales para la toma de decisiones informadas en el mundo empresarial, con el potencial de mejorar la eficacia de las estrategias de marketing digital y aumentar el retorno de inversión en publicidad en línea para E-Corp.

RECOMENDACIONES

- Asegurarse de que las páginas de reseñas y páginas informativas proporcionen contenido valioso y relevante para los productos que ofrece E-Corp. Esto puede incluir opiniones de clientes, detalles sobre productos y guías de compra que ayuden a los usuarios a tomar decisiones informadas.
- Optimizar la navegación y la presentación de las páginas relacionadas con productos para facilitar la búsqueda y exploración de los productos por parte de los usuarios. Esto puede incluir filtros de búsqueda, recomendaciones personalizadas y una estructura de categorías intuitiva.
- Identificar las páginas con altas tasas de salida y realizar mejoras en el diseño y contenido de esas páginas para reducir la tasa de abandono. Esto puede implicar agregar llamadas a la acción claras, mejorar la usabilidad y optimizar el rendimiento de la página.
- Analizar las páginas con valores de página más altos y comprender qué características las hacen más valiosas para los usuarios. Esto puede ayudar a enfocar los esfuerzos de marketing en promover y destacar esas páginas para aumentar la conversión.
- Aprovechar las tendencias estacionales referente a los meses del año y de fin de semana para adaptar las estrategias de marketing y promociones. Por ejemplo, ofrecer ofertas especiales o promociones exclusivas durante períodos de alta demanda, como las vacaciones o los fines de semana.
- Personalizar la experiencia del usuario según el sistema operativo y el tipo de visitante. Esto puede incluir optimizar la compatibilidad del sitio web con diferentes sistemas operativos y ofrecer contenido específico para diferentes tipos de usuarios, como nuevos visitantes versus clientes recurrentes.