Diplomatura en Ciencia de Datos e

Inteligencia Artificial y Sus

Aplicaciones en Economía y

Negocios

Módulo de Aplicación Integradora

Modelo Predictivo Orientado a Estrategias de Marketing Para Uso E-commerce En Páginas de Turismo

Tutor: Mgter. Giuliano Fassano

Integrantes:
Albornos, Juan Martin
Fogliatto, Robertino
Oviedo, Rocío Nazareth
Zurita, Emilia

Grupo 20
Albornos, Juan Martín; Fogliatto, Robertino; Oviedo, Rocío Nazareth; Zurita, Emilia

1. Introducción

Este informe tiene como objetivo presentar el desarrollo de las entregas del "Módulo de Aplicaciones Integradoras" con el propósito de aplicar de manera práctica los conocimientos adquiridos durante la Diplomatura en Ciencia de Datos e Inteligencia Artificial y Sus Aplicaciones en Economía y Negocios.

Este trabajo detalla el desarrollo completo de un modelo predictivo para estimar la propensión de los usuarios a comprar boletos. Dicho proceso abarca todas las etapas del proceso de ciencia de datos: desde la exploración y limpieza del dataset hasta la implementación y comparación de diversos modelos de clasificación. Entre los modelos evaluados se incluyen la Regresión Logística, Árboles de Decisión, Random Forest y XGBoost.

El análisis de comportamiento de usuarios en plataformas digitales es un campo clave en la ciencia de datos aplicada al marketing. Las empresas buscan comprender qué variables influyen en la conversión y cuáles son los patrones de comportamiento asociados a la compra. En este contexto, el proyecto tiene como objetivo construir un modelo de clasificación que prediga si un usuario comprará o no un boleto, a partir de un conjunto de datos provisto por la cátedra.

2. Descripción del dataset

El dataset original fue obtenido de una fuente en Google Drive y contiene información sobre usuarios de una plataforma de viajes. Incluye variables numéricas y categóricas relacionadas con la interacción del usuario con la página, como el número promedio de vistas, likes y check-ins, además de características demográficas y de comportamiento. El conjunto cuenta con 11,760 registros y 17 columnas que describen tanto la actividad del usuario como su contexto.

3. Procesamiento y limpieza de datos

La primera etapa consistió en identificar y corregir inconsistencias en los datos. Un análisis inicial con df.info() mostró la presencia de valores nulos en múltiples columnas y tipos de datos que, en principio, no correspondían con los verdaderos datos que la columna refleja. Luego se procedió a explorar los valores únicos de cada columna y se detectaron diversos errores tipográficos.

Para corregir esto, columna por columna, se realizaron las correspondientes transformaciones. Ejemplos de ello incluyen el reemplazo del símbolo '*' en la columna yearly_avg_Outstation_checkins por valores nulos y su cambio a tipo de dato Int64, el de 'Three' (reemplazado por 3) en member_in_family y también su cambio a tipo Int64 y el de 'Yeso' (reemplazado por 'Yes') en following_company_page.

En la columna Preferred_device, se detectó que se nombra de diferente manera a la misma categoría. Por lo que se decidió agrupar todos los dispositivos móviles bajo la categoría "Mobile" ya que la categoría "ios and android" es la más amplia y no se puede diferenciar el sistema operativo de cada usuario.

Para identificar si existían valores atípicos o outliers, se realizaron gráficos de BoxPlot en las variables numéricas. En dichos gráficos se pudo apreciar que en ciertas columnas existían muchos de estos valores. Para su tratamiento, se decidió trabajar con los percentiles 5% y 95% con el fin de hacerlo más exigente. De este modo, se eliminaron aquellos valores que superaban el percentil 95% más 1.5 veces el rango intercuartílico. Se decidió no establecer un límite inferior para la eliminación de outliers, ya que se consideró que los valores por debajo de ese umbral pueden presentarse de manera normal.

Con respecto al tratamiento de valores nulos, primero se analizó la cantidad de estos valores en cada columna. Luego se aplicaron estrategias de imputación basadas en la naturaleza de cada variable: la media para distribuciones normales y la mediana para las asimétricas. En los casos donde la proporción de nulos era baja, se optó por eliminar las filas afectadas.

Al finalizar todo este procesamiento, se obtuvo un dataset limpio sin errores tipográficos, sin valores nulos y sin outliers.

4. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos permitió obtener una visión integral del comportamiento de los usuarios. Primero se realizó un análisis gráfico general con histogramas para las variables continuas y gráficos de barras para las variables categóricas. De este primer análisis se pudo observar que:

- La mayoría de los usuarios no había adquirido boletos, lo que confirmó el desbalance de clases.
- Las vistas promedio anuales de la página de viajes presentan una distribución bastante simétrica y cercana a la normal, con valores concentrados entre 200 y 350. Esto indica un patrón de navegación regular y consistente.
- Las tablets y los dispositivos móviles concentran la mayoría de usuarios. Laptop es claramente minoritaria y “Other” casi residual.
- La distribución de likes otorgados por los usuarios es bastante plana/heterogénea (no claramente sesgada) con un pico visible alrededor de los 30 mil, y resulta en una gran dispersión general.
- La actividad de check-ins muestra una alta concentración entre 0 y 2 por usuario, lo que refleja una baja participación en esta funcionalidad, con muy pocos usuarios activos.
- Las familias pequeñas (de 1 a 4 miembros) son las que predominan, mientras que la proporción disminuye en grupos de 5 o más integrantes.
- Los destinos que concentran la mayor parte de las preferencias son Beach, Financial, Historical Site y Medical.

- Los comentarios anuales en la página de viajes se concentran principalmente entre 50 y 95, y muestran una participación moderada en las interacciones de este tipo. Es un posible indicador de interacción baja con la plataforma.
- La columna Total_likes_on_outstation_checkin_received presenta un sesgo fuerte a la derecha, con alta variabilidad.
- La mayoría de los usuarios realizó un check-in recientemente (entre 0 y 3 semanas).
- La cantidad de usuarios que no siguen la página es más del doble que los usuarios que sí la siguen.
- La actividad en comentarios dentro de la página de la empresa es reducida: la gran mayoría de los usuarios se concentra en valores bajos (0–50), con pocos casos de interacción elevada.
- En la columna Working_flag, la cantidad de "No" es casi 5 veces más grande que la cantidad de respuestas positivas.
- La columna Travelling_network_rating está concentrada entre 3 y 4.
- En Adult_flag la mayoría de valores son 0 y 1 con muchas diferencia del resto.
- La mayoría de los usuarios dedica entre 8 y 20 minutos diarios a la página y solo una minoría supera los 30–40 minutos.

Al segmentar los histogramas por compradores y no compradores, se identificaron diferencias relevantes: quienes compraron tienden a tener un mayor número promedio de vistas en la página de viajes, más check-ins anuales y un tiempo diario de navegación ligeramente superior (30–60 minutos), además de intervalos más cortos desde su último check-in, lo que sugiere un comportamiento más reciente y activo.

La matriz de correlación mostró asociaciones positivas entre el tiempo en la página y los likes recibidos, lo que evidenció que un mayor nivel de participación se traduce en más interacción dentro de la plataforma. También se aprecia una correlación moderada entre check-in recientes y nivel de actividad en la página (vistas y tiempo invertido).

Por otro lado, se realizaron tablas de contingencia para variables categóricas al segmentar por Taken_product. Estas revelaron que los usuarios que siguen la página de la compañía presentan una mayor propensión a la compra, que el uso de dispositivos móviles y laptops se asoció con una menor probabilidad de adquisición y también que la categoría de destino "OTT" se destaca del resto ya que su proporción de usuarios que compran es muy superior, mientras que otras categorías como "Tour and Travel" y "Hill Stations" también tienen proporciones altas de compradores. En cambio, "Other", "Medical" y "Game" presentan las proporciones más bajas de usuarios que adquieren el producto.

5. Análisis de componentes principales (PCA)

El PCA es útil para entender patrones globales y reducir dimensionalidad, pero en este caso la capacidad de separar las clases es limitada ya que, para poder explicar el 80% de la variabilidad original de los datos se deben tomar 8 componentes. Esto significa que la relación con la compra probablemente sea no lineal o requiera combinaciones más complejas

de variables. También nos indica que las variables cuantitativas no están muy correlacionadas entre sí.

Las diferencias en el comportamiento de compra no se explican de forma lineal y simple por las variables numéricas, por lo que fue necesario aplicar modelos de clasificación más complejos para capturar patrones relevantes.

6. Modelos predictivos

En la etapa de modelado se construyeron y evaluaron cuatro modelos de clasificación: Regresión Logística, Árbol de Decisión, Random Forest y XGBoost. Cada modelo tuvo como objetivo predecir la variable objetivo Taken_product (convertida a formato binario: No = 0, Yes = 1) a partir del conjunto de variables predictoras, con la exclusión de UserID.

En todos los casos se aplicó un proceso de pre procesamiento mediante ColumnTransformer. Este incluyó la estandarización de las variables numéricas con StandardScaler y la codificación de las variables categóricas mediante OneHotEncoder. Debido al desbalance de clases, se implementaron estrategias de ponderación (class_weight="balanced" en los modelos lineales y de árboles, y scale_pos_weight en XGBoost) con el propósito de asignar una penalización mayor a los errores en la clase minoritaria. La validación se llevó a cabo mediante un esquema de cross validation con cinco splits, y se calcularon las métricas en train y test de accuracy, precision, recall, F1-score y ROC-AUC para asegurar la solidez de los resultados. También se utilizó GridSearchCV para obtener los mejores parámetros para entrenar los modelos.

La **Regresión Logística**, utilizada como modelo base, obtuvo un F1-score de 0.464, un recall de 0.759 y un ROC-AUC de 0.804. El análisis de los coeficientes del modelo muestra que la variable más determinante es si el usuario sigue a la empresa, lo que incrementa las probabilidades de compra en 1.17 veces. A pesar de su simplicidad e interpretabilidad, este modelo presentó un número elevado de falsos positivos y no logró captar la complejidad de las relaciones entre variables. En conclusión, la Regresión Logística constituye una buena base inicial; sin embargo, al tratarse de un modelo lineal de clasificación, resulta insuficiente para realizar predicciones precisas y confiables. Por ello, fue necesario implementar modelos más complejos, como árboles de decisión o métodos de ensemble learning tales como Random Forest y Gradient Boosting.

El **Árbol de Decisión**, tras su ajuste con GridSearchCV, alcanzó métricas mucho más elevadas en train (F1-score de 0.92 y ROC-AUC de 0.953). Esto podía ser una señal de sobreajuste, sin embargo en test, los resultados fueron incluso más altos. Se realizó también un análisis de las curvas de F1 score para entrenamiento y validación ante distintos niveles de profundidad del árbol. Se pudo ver un leve sobreajuste a partir del max_depth 12 por lo que se decidió cortar ahí el árbol.

Se comprobó también que no existiera fuga de datos, es decir, si hay alguna feature que se derive directamente de la variable objetivo Taken_product. Por lo tanto el modelo funciona correctamente y clasifica bastante bien.

Teniendo en cuenta el nuevo árbol con max_depth = 12, se buscó el umbral que maximice el F1 score. El umbral óptimo es 0.813, por lo que diremos que el usuario va a comprar si su probabilidad es mayor a 0.813. Esto hace que el modelo sea más estricto y se mejore el balance entre precision y recall, es decir, entre acertar y detectar compradores.

En conclusión, cuando el modelo dice que el cliente comprará (clase 1), acierta el 89% de las veces, y además detecta el 89% de los compradores reales. En marketing, esto es muy bueno porque no se estarán dando promociones a usuarios que no comprarán (alta precision) y no se estarán perdiendo compradores potenciales (alta recall).

La curva Precision–Recall es bastante alta por lo que también demuestra que muchos compradores reales están efectivamente bien rankeados por el modelo. La curva ROC es excelente por lo que el modelo distingue compradores de no compradores muy claramente.
Se identificaron también las variables más importantes para la separación de clases. Las primeras 3 son total_likes_on_outofstation_checkin_received, total_likes_on_outstation_checkin_given y Yearly_avg_view_on_travel_page.

El **Random Forest** mejoró la estabilidad del modelo y presentó un mejor rendimiento en test, con un F1-score de 0.956 y ROC-AUC de 0.999. Tenía un leve desequilibrio entre precision y recall, la primera con un valor de 0.991, acertando casi siempre que dice que el cliente compra, y la segunda con un valor de 0.92.

El umbral que maximiza el F1 score en train es de 0.385, esto significa que el modelo tiende a asignar probabilidades moderadas (no muy altas) a la clase positiva. Si se usara el umbral por defecto de 0.5 se perderían positivos. Al bajar el umbral, baja la precision a 0.945 (sigue siendo muy buena) y aumenta el recall a 0.965, logrando así un mejor equilibrio entre ambas.
Las 3 variables más importantes para la separación de clases son las mismas que en el árbol de decisión.

En conclusión, el recall mejora con respecto al árbol de decisión. Ahora su valor es de 0.965. Esto es excelente ya que identifica casi todos los usuarios que van a comprar. La precision también mejora y el f1 score aumenta a 0.955, logrando un muy buen equilibrio entre no perder compradores y no contactar demás.

Analizando las curvas ROC y Precision-Recall, podemos afirmar que el modelo separa extremadamente bien ambas clases.

Finalmente, el modelo **XGBoost** superó a los anteriores al combinar el poder de los árboles de decisión con regularización avanzada. Su capacidad de combinar múltiples árboles de decisión y ajustar automáticamente la penalización por clases desbalanceadas resultó en un

modelo más robusto y generalizable. El modelo funciona de forma excelente, no sobreajusta y las métricas mejoraron con respecto al Random Forest. Recall (0.976), Precision (0.989) y F1 (0.982) son muy altos. Solo hay 4 falsos positivos y 9 falsos negativos por lo que las clases están muy bien aprendidas.

El umbral óptimo que maximiza F1 es de 0.595, se subió con respecto al 0.5 estándar para ser más estricto al clasificar los positivos. Sin embargo, podemos ver que la precisión sube muy poquito (a 0.992) y los demás scores se mantienen prácticamente igual. La performance en test (F1 ≈ 0.984) es ligeramente superior a la anterior, lo cual sugiere nada de sobreajuste.

Por lo tanto, podemos decir que el modelo XGBoost ya está tan bien calibrado y generaliza tan bien, que el umbral 0.5 está prácticamente en el punto de equilibrio entre precisión y recall, ajustarlo no casi no cambia el rendimiento.

Las curvas ROC y PR dieron 1 por lo que confirman que el modelo tiene un poder discriminativo excelente, separa casi perfectamente las clases.

En este modelo, las 3 variables más importantes o determinantes a nivel individual son: que siga la página de la empresa, Adult_flag y “Otro” como destino preferido. A nivel agregado son Adult_flag, total_likes_on_outofstation_checkin_received y travelling_network_rating.

7. Evaluación y comparación de modelos

La comparación de modelos muestra una evolución clara en términos de desempeño. La regresión logística ofrece una base interpretativa sólida, mientras que los modelos de ensamble, Random Forest y XGBoost, presentan una mejora significativa en la capacidad predictiva. La introducción de técnicas de validación cruzada y ajuste de hiperparámetros permitió reducir el riesgo de sobreajuste y garantizar una mejor generalización del modelo.

En términos de interpretación, las variables más determinantes en los modelos finales fueron aquellas relacionadas con la interacción digital y la afinidad con la empresa: seguir la página, la categoría de destino preferido y el tiempo de uso diario. Estas conclusiones son coherentes con el comportamiento esperado en contextos de marketing digital, donde la participación activa y la conexión con la marca son indicadores de mayor probabilidad de compra.

En consecuencia, se concluye que XGBoost es el modelo seleccionado para predecir la propensión de compra, dado que ofrece el mejor desempeño general, una excelente capacidad de generalización y una interpretación coherente con los patrones de comportamiento observados en los datos.

8. Análisis Final

Analizamos los falsos negativos, es decir, aquellos usuarios que el modelo no identificó como compradores, pero que finalmente realizaron una compra. En este caso, el modelo tiende a

subestimarlos. Observamos que todos estos usuarios acceden desde dispositivos móviles, y que 6 de los 9 comparten la misma cantidad en la variable *likes_on_outstation_checkin_given*, además de que el campo *working_flag* toma el valor "No".

Luego analizamos los falsos positivos, es decir, clientes que el modelo predijo como compradores, pero que finalmente no realizaron la compra. Este grupo es clave, ya que son usuarios sobre los que conviene aplicar acciones de marketing específicas para intentar revertir su decisión. En este caso, todos también acceden desde dispositivos móviles, y 4 de los 5 presentan la misma cantidad de *likes_on_outstation_checkin_given* (coincidente con los valores observados en los falsos negativos), mientras que *working_flag* también toma el valor "No".

A partir de estas observaciones, no se puede afirmar que exista un grupo de usuarios con características particulares que el modelo clasifique sistemáticamente de forma errónea.

Posteriormente, dividimos el conjunto de datos en deciles para analizar la probabilidad de compra (y_proba) en cada uno.

Se observa que:

- El 10% de clientes con menor probabilidad (decil 1) tiene valores inferiores a 0.0007, es decir, prácticamente nulos.
- El 10% superior (decil 10) presenta probabilidades mayores a 0.98, lo que indica una altísima propensión a comprar.
- La distribución es muy concentrada: la mayoría de los clientes muestra una probabilidad baja, mientras que un grupo reducido presenta una probabilidad muy alta.

Esto sugiere que el modelo posee un excelente poder discriminativo, ya que separa claramente a los compradores de los no compradores, algo consistente con las métricas ROC-AUC y PR-AUC, ambas con valor 1.

En base a esta información, se puede concluir que:

- Del decil 1 al 5, la probabilidad de compra es prácticamente nula, por lo que cualquier esfuerzo de marketing sería ineficiente.
- Del decil 6 al 8, la probabilidad aumenta levemente, aunque sigue siendo baja, por lo que tampoco se recomienda invertir recursos en esos segmentos.
- Al decil 10 tampoco es necesario destinar esfuerzos, dado que la probabilidad de compra es casi total.

De esta manera, el valor principal que aporta el modelo es permitir identificar de manera precisa al 80% de los usuarios que no comprarán y al 10% que con seguridad lo harán.

El siguiente paso consiste en analizar en detalle el decil 9, ya que allí se concentran los usuarios con una probabilidad intermedia de compra. Dentro de este grupo, los subdeciles 4 y 5 resultan los más adecuados para implementar acciones de marketing que estimulen la conversión.

9. Aplicación de los resultados del modelo en estrategias de marketing

A partir de los resultados obtenidos en el modelo predictivo, se propone diseñar una campaña de marketing orientada principalmente a los usuarios indecisos, con el objetivo de maximizar la conversión dentro de este segmento de potenciales clientes. El análisis evidenció que la propensión a la compra se relaciona principalmente con el nivel de interacción digital, expresado en el seguimiento de la página de la empresa, la cantidad de *likes* y *check-ins,* el tiempo promedio de navegación y el tipo de ubicación preferida registrado en la plataforma.

En función de estos hallazgos, se recomienda implementar estrategias de segmentación y personalización de contenido dirigidas a los usuarios más activos, reforzando su vínculo con la marca mediante campañas de remarketing, notificaciones o promociones exclusivas según su patrón de comportamiento. Asimismo, se sugiere emplear el modelo predictivo de forma continua para actualizar los segmentos y monitorear el impacto de las acciones implementadas mediante métricas de conversión y retención. De esta manera, la integración entre el análisis de datos y el marketing personalizado permitirá optimizar los recursos disponibles y potenciar la captación de nuevos clientes.