Un Análisis Profundo del Conjunto de Datos 'Online Shoppers Purchasing Intention': Desde su Origen hasta la Formulación de Hipótesis

Sección 1: Génesis y Contexto del Conjunto de Datos 'Online Shoppers Purchasing Intention'

Para llevar a cabo un análisis de datos riguroso y significativo, es imperativo comprender no solo el contenido de un conjunto de datos, sino también su origen, el propósito para el cual fue creado y las decisiones metodológicas que guiaron su construcción. Esta sección explora el contexto académico y la estrategia de muestreo detrás del conjunto de datos 'Online Shoppers Purchasing Intention', proporcionando una base fundamental para su correcta interpretación y uso.

1.1. Orígenes Académicos y Objetivos de la Investigación

El conjunto de datos 'Online Shoppers Purchasing Intention' no es una simple recopilación de métricas de comercio electrónico; es el subproducto de una investigación académica específica con objetivos bien definidos. Fue presentado por primera vez en el artículo de 2018 titulado "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks", cuyos autores son C. Okan Sakar, S. Olcay Polat, Mete Katircioglu y Yomi Kastro.[1]

El objetivo principal de esta investigación era desarrollar un sistema capaz de predecir, en tiempo real, si un usuario que navega por un sitio de comercio electrónico completaría finalmente una compra.[3] El énfasis en el "tiempo real" es un aspecto fundamental que dictó la estructura del conjunto de datos. Explica por qué los datos están organizados por sesiones y por qué incluye características que pueden actualizarse dinámicamente con cada acción del usuario, como el número de páginas vistas y el tiempo de permanencia en ellas (por ejemplo, `Administrative_Duration`, `ProductRelated_Duration`).[2] En esencia, el conjunto de datos fue concebido como una herramienta para resolver un problema empresarial crítico: identificar a los usuarios con alta intención de compra para poder aplicar intervenciones dirigidas (como ofertas de descuento o la activación de un chat de soporte) con el fin de prevenir el abandono del sitio y aumentar las tasas de conversión.[3]

Este propósito original revela una capa más profunda de interpretación. El objetivo de los investigadores era, en sus propias palabras, "imitar el comportamiento de un vendedor en un entorno de compra virtual".[3] En una tienda física, un vendedor experimentado observa el comportamiento de un cliente —cuánto tiempo examina un producto, si mira la etiqueta del

precio, si busca ayuda— para evaluar su intención de compra. Las variables de este conjunto de datos, como ProductRelated_Duration, PageValues y BounceRates, son los equivalentes digitales de estas observaciones físicas. Actúan como indicadores del lenguaje corporal digital del usuario. Por lo tanto, estas variables no deben ser vistas como meras cifras abstractas. Cada una representa una pieza de comportamiento observable que refleja el nivel de interés y compromiso del usuario. Comprender esto transforma el análisis de un ejercicio puramente estadístico a uno de comportamiento, lo cual es esencial para formular hipótesis significativas y relevantes para el negocio.

1.2. Curación de Datos, Fuente y Estrategia de Muestreo

El conjunto de datos se compone de 12,330 vectores de características, donde cada vector representa una sesión de navegación única, recopilada a lo largo de un período de un año en un sitio web de comercio electrónico turco.[1] Una fuente sugiere que el sitio podría pertenecer a la marca de ropa Columbia.[11]

La decisión metodológica más crucial tomada por los autores fue garantizar que **cada sesión perteneciera a un usuario diferente**.[1] Esta elección fue deliberada y se hizo "para evitar cualquier tendencia hacia una campaña específica, un día especial, un perfil de usuario o un período".[1] Esta estrategia de muestreo mejora significativamente la capacidad de generalización de los hallazgos derivados del conjunto de datos. Implica que los patrones observados tienen menos probabilidades de ser artefactos de un pequeño grupo de "superusuarios" o del impacto de una única campaña de marketing muy exitosa.

Esta regla de "una sesión por usuario único" tiene implicaciones importantes para el tipo de análisis que se puede realizar. El diseño del estudio proporciona una visión amplia pero superficial de la base de usuarios: observamos a 12,330 usuarios diferentes en una única ocasión. Esto constituye un estudio *transversal* de sesiones de usuario, no un estudio *longitudinal* del comportamiento del usuario a lo largo del tiempo. Si bien este enfoque es excelente para construir un modelo general de intención de compra basado en una interacción única, también impone una limitación clave. El conjunto de datos, por su diseño, no puede responder preguntas sobre el viaje del cliente a largo plazo, el valor de vida del cliente (customer lifetime value) o cómo evoluciona el comportamiento de un usuario a lo largo de múltiples visitas. Por lo tanto, cualquier hipótesis formulada debe estar estrictamente limitada al contexto de una única sesión. Por ejemplo, se puede plantear una hipótesis sobre el comportamiento de un 'Returning_Visitor' (visitante recurrente) dentro de su sesión *actual*, pero es imposible analizar los datos de sus sesiones anteriores, ya que no están incluidos en el conjunto de datos.

Sección 2: Un Léxico Exhaustivo de los Atributos del Conjunto de Datos

Para formular hipótesis válidas y realizar un análisis exploratorio de datos (EDA) significativo, es indispensable un dominio completo de cada una de las 18 variables del

conjunto de datos. Esta sección ofrece una deconstrucción detallada de cada atributo, agrupándolos lógicamente según su función en el contexto del comercio electrónico.

La siguiente tabla sirve como una guía de referencia rápida y esencial para todo el proyecto, centralizando la información clave de cada variable.

Tabla 1: Resumen de Atributos del Conjunto de Datos

Nombre de la Variable	Tipo de Dato Técnico	Tipo Conceptual	Descripción Detallada	Rango / Valores Únicos
Administrative	int64	Numérico-Discreto	Número de páginas de tipo administrativo visitadas en la sesión.	0 - 27
Administrative_Duration	float64	Numérico-Continuo	Tiempo total (en segundos) que el usuario pasó en páginas administrativas.	0 - 3398.75
Informational	int64	Numérico-Discreto	Número de páginas de tipo informativo visitadas en la sesión.	0 - 24
Informational_Duration	float64	Numérico-Continuo	Tiempo total (en segundos) que el usuario pasó en páginas informativas.	0 - 2549.375

ProductRelated	int64	Numérico-Discreto	Número de páginas relacionadas con productos visitadas en la sesión.	0 - 705
ProductRelated_Duration	float64	Numérico-Continuo	Tiempo total (en segundos) que el usuario pasó en páginas de productos.	0 - 63973.52
BounceRates	float64	Numérico-Continuo	Tasa de rebote promedio de las páginas visitadas en la sesión.	0.0 - 0.2
ExitRates	float64	Numérico-Continuo	Tasa de salida promedio de las páginas visitadas en la sesión.	0.0 - 0.2
PageValues	float64	Numérico-Continuo	Valor promedio de las páginas visitadas por el usuario en la sesión.	0.0 - 361.76
SpecialDay	float64	Numérico-Continuo	Proximidad de la visita a un día especial (ej. San Valentín). Valor de 0 a 1.	0.0 - 1.0

Month	object	Categórico-Nominal	Mes del año en que ocurrió la sesión.	10 valores únicos
OperatingSystems	int64	Categórico-Nominal	Código numérico que representa el sistema operativo del usuario.	8 valores únicos
Browser	int64	Categórico-Nominal	Código numérico que representa el navegador del usuario.	13 valores únicos
Region	int64	Categórico-Nominal	Código numérico que representa la región geográfica del usuario.	9 valores únicos
TrafficType	int64	Categórico-Nominal	Código numérico que representa la fuente de tráfico de la sesión.	20 valores únicos
VisitorType	object	Categórico-Nominal	Tipo de visitante ('New_Visitor', 'Returning_Visitor', 'Other').	3 valores únicos
Weekend	bool	Booleano	Indica si la sesión ocurrió durante el fin de semana.	True / False

Revenue	bool	Booleano (Objetivo)	Indica si la sesión finalizó con una compra.	True / False

(Fuente de los datos de la tabla: [13])

2.1. Métricas de Interacción del Usuario (Datos de Clickstream)

Estas seis características son las medidas más directas de la actividad del usuario, derivadas de la información de las URL de las páginas que visita.[2]

- **Administrative** y **Administrative_Duration**: Se refieren a las páginas relacionadas con la gestión de la cuenta, como el inicio de sesión, la visualización de pedidos anteriores o la configuración del perfil. Valores altos en estas métricas pueden indicar que un cliente existente está gestionando su cuenta, lo que podría ser un precursor de una nueva compra o una señal de un cliente leal.
- **Informational** y **Informational_Duration**: Corresponden a páginas que proporcionan información general, como secciones de "Preguntas Frecuentes" (FAQ), políticas de envío o la página "Sobre nosotros". Una alta interacción con estas páginas sugiere que el usuario se encuentra en una fase de investigación o de construcción de confianza, evaluando la credibilidad y las condiciones del sitio antes de comprometerse a una compra.
- **ProductRelated** y **ProductRelated_Duration**: Representan el núcleo de la experiencia de compra. Incluyen páginas de categorías, páginas de detalles de productos y resultados de búsqueda. Se espera que esta sea la categoría más influyente en la intención de compra. La distinción entre el número de páginas (ProductRelated) y la duración (ProductRelated_Duration) es crucial; un usuario que visita 50 páginas de productos en 60 segundos está probablemente "saltando" de un producto a otro sin mucho interés, mientras que un usuario que visita 10 páginas en 600 segundos está probablemente considerando cada producto con mayor detenimiento.

2.2. Indicadores de Rendimiento de Google Analytics

Estas métricas no son datos brutos de clickstream, sino indicadores sofisticados calculados por plataformas de análisis web como Google Analytics.[2]

- **BounceRates (Tasa de Rebote)**: Se define como el porcentaje de visitantes que entran al sitio en una página y luego se van ("rebotan") *sin activar ninguna otra solicitud al servidor de análisis* durante esa sesión.[7] Un rebote es, por definición, una sesión con una sola página vista. Es una señal inequívoca de falta de interés o de que la página de destino no cumplió con las expectativas del usuario.
- **ExitRates (Tasa de Salida)**: Para una página específica, es el porcentaje de vistas de esa página que fueron las *últimas de la sesión*.[7] A diferencia de la tasa de

rebote, la tasa de salida no es inherentemente negativa. Todas las sesiones tienen una página de salida; por ejemplo, la página de "confirmación de pedido" tendrá una tasa de salida del 100%, lo cual es un resultado deseable.

- **PageValues (Valor de Página)**: Esta es la métrica más compleja y potencialmente la más predictiva. Representa el valor promedio de una página que un usuario visitó antes de completar una transacción. Google Analytics lo calcula dividiendo los ingresos de la transacción más el valor total de los objetivos entre las vistas de página únicas para esa página en particular.[7] En esencia, esta métrica distribuye el valor económico de una conversión entre todas las páginas que contribuyeron a ella. Un valor de PageValues distinto de cero es un fuerte indicador de que el usuario está en un camino que históricamente conduce a una compra.

La relación entre PageValues y la variable objetivo Revenue merece una consideración especial. La definición de PageValues implica que se calcula en función de las páginas visitadas *antes* de una transacción. Esto significa que una sesión que no genera ingresos (Revenue=False) tendrá, por definición, un PageValues de 0. Por lo tanto, no es que un PageValues alto *cause* una compra. Más bien, el acto de navegar a través de una secuencia de páginas que históricamente está correlacionada con las conversiones es lo que *genera* una puntuación alta de PageValues para la sesión. Un PageValues elevado es un síntoma, no una causa, de una sesión con alta intención de compra. Indica que el usuario está siguiendo un "camino dorado" hacia la conversión. Las hipótesis sobre esta variable deben reflejar esta relación, preguntando no si PageValues conduce a la compra, sino si las sesiones que resultan en compra se caracterizan por tener un PageValues significativamente mayor.

2.3. Variables Contextuales y Temporales

Estos atributos proporcionan contexto sobre el momento y las circunstancias de la sesión.[7]

- **SpecialDay**: Es una característica de ingeniería numérica que indica la proximidad de la fecha de la visita a un día especial de compras (por ejemplo, San Valentín, Día de la Madre).[7] Es un valor de punto flotante entre 0 y 1, donde 0 indica ninguna proximidad y 1 indica la máxima proximidad. Su cálculo es sofisticado, ya que tiene en cuenta dinámicas del comercio electrónico como el tiempo entre la fecha del pedido y la entrega. Por ejemplo, para el día de San Valentín, el valor es distinto de cero entre el 2 y el 12 de febrero, y alcanza su máximo el 8 de febrero, no el 14, para dar tiempo al envío.[7]
- **Month**: El mes en que ocurrió la visita. Esta variable captura la estacionalidad (por ejemplo, compras navideñas en noviembre y diciembre). Es importante señalar que algunos análisis sugieren que faltan datos de dos meses en el conjunto de datos, lo que podría introducir un sesgo en los análisis basados en esta variable.[16]
- **Weekend**: Una variable booleana que indica si la visita tuvo lugar en fin de semana.

2.4. Demografía del Usuario y de la Sesión

Estas variables categóricas describen al usuario y los aspectos técnicos de su sesión.[2]

- **VisitorType**: Una variable crucial con tres niveles: 'Returning_Visitor' (visitante recurrente), 'New_Visitor' (nuevo visitante) y 'Other' (otro).[13] Los visitantes recurrentes suelen ser el segmento más valioso para un negocio de comercio electrónico.
- **OperatingSystems**, **Browser**, **Region**, **TrafficType**: Aunque están codificadas como enteros, estas variables son de naturaleza categórica.[13] Son útiles para identificar problemas técnicos (por ejemplo, una baja tasa de conversión en un navegador específico) o para evaluar el rendimiento de los canales de marketing (a través de TrafficType).

2.5. La Variable Objetivo (Revenue)

Esta es la variable dependiente que se busca predecir.[13]

- **Revenue**: Una variable booleana (True/False) que indica si la sesión concluyó con una transacción económica.

Un aspecto fundamental de esta variable es que el conjunto de datos está muy desbalanceado. Aproximadamente el 84.5% de las sesiones no resultaron en una compra (Revenue=False), mientras que solo el 15.5% sí lo hicieron (Revenue=True).[2] Este desequilibrio debe ser considerado en cualquier modelo predictivo.

Sin embargo, este desbalance debe ser interpretado con conocimiento del dominio. Una tasa de conversión del 15.5% por sesión puede parecer baja desde una perspectiva de balance de clases en ciencia de datos, pero en el contexto del comercio electrónico, es excepcionalmente alta. Las tasas de conversión típicas en la industria suelen oscilar entre el 1% y el 3%. Esta alta tasa de conversión de referencia en el conjunto de datos sugiere que la población de usuarios representada puede estar prefiltrada o ser más propensa a la compra que el tráfico promedio de un sitio web. Este contexto es vital. Significa que un factor que aumenta *ligeramente* la probabilidad de compra en este conjunto de datos podría ser un impulsor *muy significativo* en un sitio con una tasa de conversión más típica del 2%.

Sección 3: Formulación de Hipótesis Univariadas

Con un conocimiento profundo de cada variable, ahora es posible formular hipótesis comprobables sobre la relación entre las características individuales y la variable objetivo, Revenue. Cada hipótesis se presenta formalmente con una hipótesis nula (H_0) y una alternativa (H_1), seguidas de una justificación basada en el conocimiento del dominio.

3.1. Hipótesis sobre el Tipo de Visitante

- **H_0**: La proporción de sesiones que resultan en una compra (Revenue=True) es la misma para 'New_Visitor' y 'Returning_Visitor'.

- **H_1**: La proporción de sesiones que resultan en una compra es significativamente mayor para 'Returning_Visitor' que para 'New_Visitor'.
- **Justificación**: Los visitantes recurrentes ya han superado la barrera inicial de confianza y han demostrado un interés previo en los productos del sitio. Esta familiaridad y precalificación deberían traducirse en una tasa de conversión más alta, ya que es más probable que regresen con una intención de compra más definida.[21]

3.2. Hipótesis sobre la Interacción del Usuario (Duración)

- **H_0**: La duración promedio en páginas de productos (ProductRelated_Duration) para las sesiones que resultan en una compra es igual a la duración promedio para las sesiones que no resultan en una compra.
- **H_1**: La duración promedio en páginas de productos (ProductRelated_Duration) es significativamente mayor para las sesiones que resultan en una compra.
- **Justificación**: El tiempo dedicado a las páginas de productos es un indicador directo del interés y la consideración. Se espera que los usuarios que están contemplando seriamente una compra pasen más tiempo investigando detalles, especificaciones e imágenes de los productos, en comparación con aquellos que solo están navegando casualmente.

3.3. Hipótesis sobre la Adherencia al Sitio (Tasa de Rebote)

- **H_0**: La tasa de rebote promedio (BounceRates) para las sesiones que resultan en una compra es igual a la tasa de rebote promedio para las sesiones que no resultan en una compra.
- **H_1**: La tasa de rebote promedio (BounceRates) es significativamente menor para las sesiones que resultan en una compra.
- **Justificación**: Un rebote representa una falta total de interacción. Es conceptualmente imposible que una sesión de rebote (una sesión con una sola página vista) resulte en una compra, ya que esta requiere navegar al menos a una página de producto, al carrito y a la página de pago. Por lo tanto, las sesiones que generan ingresos deben tener una tasa de rebote de 0, y el promedio para este grupo debería ser drásticamente más bajo.[22]

3.4. Hipótesis sobre la Señal de Valor Económico (Valor de Página)

- **H_0**: El valor de página promedio (PageValues) para las sesiones que resultan en una compra es igual a 0.
- **H_1**: El valor de página promedio (PageValues) para las sesiones que resultan en una compra es significativamente mayor que 0.
- **Justificación**: Como se estableció en la sección 2.2, la métrica PageValues está intrínsecamente ligada a las rutas de navegación que generan ingresos. Por definición, una sesión con Revenue=True debe tener un PageValues mayor que cero, mientras que una sesión sin ingresos tendrá un PageValues de 0. Esta hipótesis busca verificar esta propiedad fundamental de la métrica dentro del conjunto de datos.

Sección 4: Formulación de Hipótesis Bivariadas

Esta sección explora la interacción entre pares de variables para descubrir relaciones más matizadas en el comportamiento del usuario. Estas hipótesis investigan efectos de interacción que una análisis univariado no podría capturar.

4.1. Hipótesis sobre el Comportamiento en Fin de Semana y el Tipo de Visitante

- **H_0**: El efecto de ser un 'Returning_Visitor' en la probabilidad de compra es independiente de si la sesión ocurre en fin de semana (`Weekend`).
- **H_1**: El aumento en la probabilidad de compra para un 'Returning_Visitor' (en comparación con un 'New_Visitor') es aún mayor durante el fin de semana (`Weekend`).
- **Justificación**: Los fines de semana pueden representar un período de tiempo en el que los usuarios realizan compras más consideradas y menos apresuradas. Un visitante recurrente, que ya está familiarizado con el sitio, podría aprovechar este tiempo para finalizar una compra que estuvo investigando durante la semana laboral. Esta hipótesis explora la interacción entre la familiaridad del usuario y el momento de la compra.

4.2. Hipótesis sobre Días Especiales e Interacción

- **H_0**: No existe correlación entre la métrica `SpecialDay` y `ProductRelated_Duration` para las sesiones que resultan en una compra.
- **H_1**: Existe una correlación positiva entre la métrica `SpecialDay` y `ProductRelated_Duration` para las sesiones que resultan en una compra.
- **Justificación**: Las visitas cercanas a una festividad comercial importante (indicadas por un valor alto de `SpecialDay`) suelen estar asociadas con una mayor urgencia de compra y una intención de hacer un regalo. Esto podría llevar a los usuarios a pasar más tiempo evaluando cuidadosamente los productos para asegurarse de que están tomando la decisión correcta, lo que a su vez aumentaría su tiempo de permanencia en las páginas relacionadas con productos.

4.3. Hipótesis sobre la Fuente de Tráfico y la Adherencia al Sitio

- **H_0**: La tasa de rebote promedio (`BounceRates`) es la misma en todas las principales categorías de `TrafficType`.
- **H_1**: La tasa de rebote promedio (`BounceRates`) es significativamente diferente para sesiones que se originan en distintas categorías de `TrafficType` (por ejemplo, tráfico directo frente a un anuncio de banner).
- **Justificación**: La fuente del tráfico a menudo predetermina la intención del usuario. Un usuario que escribe la URL directamente (tráfico directo) probablemente tiene un objetivo claro y, por lo tanto, una menor probabilidad de rebotar. En cambio, un usuario que hace clic en un anuncio de banner genérico puede tener una intención más baja y ser más propenso a rebotar si la página de destino no cumple inmediatamente sus expectativas. Este análisis es crucial para evaluar la efectividad de los diferentes canales de marketing.

4.4. Hipótesis sobre el Valor de Página y la Profundidad de la Interacción

- **H_0**: Entre las sesiones que resultaron en una compra, no hay correlación entre PageValues y el número total de páginas de productos (ProductRelated) visitadas.
- **H_1**: Entre las sesiones que resultaron en una compra, existe una correlación positiva entre PageValues y el número total de páginas de productos (ProductRelated) visitadas.
- **Justificación**: Esta hipótesis busca comprender la naturaleza de las sesiones de compra "valiosas". ¿Una transacción de alto valor proviene de un usuario que encuentra rápidamente lo que busca (bajo recuento de ProductRelated, alto PageValues), o de un usuario que navega extensamente, posiblemente añadiendo múltiples artículos a su carrito (alto recuento de ProductRelated, alto PageValues)? La respuesta puede informar estrategias para optimizar la navegación del sitio y los sistemas de recomendación de productos.

Conclusión

El conjunto de datos 'Online Shoppers Purchasing Intention' es una herramienta rica y bien estructurada para el análisis del comportamiento del consumidor en el comercio electrónico. Su origen académico, centrado en la predicción en tiempo real, dota a sus variables de un propósito claro y orientado a la acción. La comprensión profunda de su metodología de muestreo (una sesión por usuario único) y de las definiciones precisas de sus 18 atributos, especialmente las métricas de Google Analytics como PageValues, es fundamental para cualquier análisis válido.

El análisis de su contexto revela que no es simplemente un registro de clics, sino una representación de la "intención" del usuario, capturada a través de su lenguaje corporal digital. La tasa de conversión anormalmente alta del 15.5% sugiere que se trata de un entorno de alta intención, un factor que debe tenerse en cuenta al generalizar los hallazgos.

Las hipótesis univariadas y bivariadas propuestas en este informe sirven como punto de partida para un análisis exploratorio de datos riguroso. Al probar estas suposiciones, un analista puede descubrir los impulsores clave de la conversión dentro de este conjunto de datos, sentando las bases para análisis predictivos más complejos y, en última instancia, para la toma de decisiones empresariales informadas. Este documento proporciona el conocimiento de dominio necesario para pasar de ser un mero usuario de los datos a un experto capaz de interrogarlo de manera significativa.