2025

Fundamentos de la Ciencia de Datos

Trabajo Práctico Especial

Grupo 04

Gabriel Cristian González

Torres Lagar Estanislao

Turri Pablo

1. Introducción

En este trabajo se analiza el dataset **Online Shoppers Purchasing Intention**, presentado por Sakar et al. (2018), que contiene información de sesiones de usuarios en un sitio de comercio electrónico. Cada fila representa una sesión, y la variable objetivo **Revenue** indica si la visita terminó o no en una compra (True/False).

El conjunto de datos combina:

- Métricas de navegación: cantidad de páginas visitadas por tipo (Administrative, Informational, ProductRelated) y sus respectivas duraciones.
- Métricas de comportamiento: BounceRates, ExitRates y PageValues, inspiradas en definiciones de Google Analytics.
- Variables de contexto: tipo de visitante (VisitorType), fuente de tráfico (TrafficType), sistema operativo, navegador, región, mes (Month) y si la sesión ocurrió en fin de semana (Weekend).
- Variable de salida: Revenue (compra/no compra).

El objetivo principal es entender **qué factores explican o se asocian con la probabilidad**

de compra, siguiendo el proceso pedido por la cátedra: análisis exploratorio, limpieza, planteo de hipótesis y su contraste con pruebas estadísticas apropiadas.

Enunciado Práctico Especial 2025

El informe se organiza de la siguiente manera: primero se detalla la limpieza y preparación de datos; luego se presenta el análisis exploratorio (univariado y bivariado). A partir de esas observaciones se plantean seis hipótesis (univariadas, bivariadas y multivariadas), se describen las pruebas utilizadas para validarlas y finalmente se discuten las conclusiones generales.

2. Limpieza y preparación de datos

Se comenzó explorando la estructura original del dataset:

- 12330 registros y 18 columnas.
- Sin valores nulos declarados por el archivo original.
- Tipos de datos ya compatibles con su significado (enteros, flotantes, booleanos y categóricos codificados como texto o identificadores).

Sobre esta base se realizaron las siguientes acciones:

1. **Revisión de tipos y códigos categóricos**
    - Se verificó que variables como **Month**, **VisitorType**, **Weekend** y **Revenue** estuvieran correctamente interpretadas (por ejemplo, Revenue como booleano/0–1).
    - No fue necesario recodificar categorías manualmente, ya que los valores coincidían con la documentación del paper de Sakar et al. (2018).
2. **Detección de outliers**
    - Se aplicó describe() y se generaron boxplots para variables de duración y PageValues, observando colas largas y valores extremos. Estos picos son esperables en datos de navegación web (sesiones muy largas, valores de página muy altos).
    - Se decidió **no eliminar outliers**, porque representan comportamiento real de usuarios y son relevantes para analizar compras “atípicas” pero posibles. TPE
3. **Construcción del dataset preprocesado**
    - Se creó un dataframe de trabajo preprocessed_dataset con las mismas 18

variables originales, garantizando que Revenue fuera numérica (0/1) para los modelos.

○ Se añadió una variable derivada **Informational_Total = Administrative + Informational**, utilizada luego en el contraste de la hipótesis H4.

○ Fueron detectadas 125 filas duplicadas las cuales fueron borradas. No fue necesario imputar valores faltantes, ya que el dataset no presentaba nulos reportados en las columnas utilizadas.

En síntesis, la limpieza fue principalmente de **verificación y preparación**: revisar estructura, tipos y valores extremos, y crear variables derivadas útiles para el análisis posterior, sin aplicar recortes agresivos que pudieran distorsionar la distribución real de las sesiones.

3. Análisis exploratorio de los datos

El análisis exploratorio se dividió en dos etapas: **univariado** (cada variable por separado) y **bivariado** (relación de las variables con Revenue).

3.1. Análisis univariado

Para cada variable del dataset se revisó:

- Tipo de dato y significado en el contexto de e-commerce.
- Estadísticos básicos (describe()): mínimo, máximo, media, mediana, desviación estándar.
- Distribuciones y outliers mediante histogramas, boxplots y gráficos de barras.

Algunos hallazgos relevantes:

- Las variables de conteo y duración (**Administrative, Informational, ProductRelated y sus duraciones**) presentan distribuciones muy asimétricas, con muchos ceros y pocas sesiones con valores muy altos. Esto es típico de navegación web: la mayoría de usuarios visita pocas páginas, y sólo una minoría navega mucho tiempo contenido específico.
- Las métricas **BounceRates** y **ExitRates** tienden a concentrarse en valores medios–altos, indicando que muchas sesiones terminan rápidamente en la misma

página o abandonan el sitio sin avanzar demasiado.

- **PageValues** está fuertemente cero-inflada: la mayoría de sesiones tiene valor 0, y los valores altos se observan en un subconjunto pequeño de visitas, lo que es consistente con su definición como valor económico esperado asociado a rutas de conversión.

- En las variables categóricas, **Returning_Visitor** es claramente la categoría dominante en VisitorType, y ciertos meses y tipos de tráfico concentran la mayor parte de las sesiones.

Este análisis permitió identificar patrones de comportamiento (por ejemplo, alta proporción de sesiones sin PageValue positivo ni compra) y sirvió como base para el diseño de las hipótesis.

3.2. Análisis bivariado

Dado que **Revenue** es la variable objetivo, el análisis bivariado se centró en estudiar cómo se comportan el resto de las variables respecto a la compra/no compra:

- **VisitorType vs Revenue**: los gráficos de barras mostraron que, tanto en términos absolutos como en tasas de conversión, los visitantes recurrentes (**Returning_Visitor**) compran más que los nuevos (**New_Visitor**).

- **Weekend vs Revenue y Month vs Revenue**: se observó que la distribución de compras varía según el mes y si la sesión ocurrió o no en fin de semana, sugiriendo cierta estacionalidad y diferencias de comportamiento entre días hábiles y fines de semana.

- **TrafficType, Region, Browser y OperatingSystems vs Revenue**: se identificaron tipos de tráfico y determinadas configuraciones de entorno donde la conversión es más alta.

- **ProductRelatedDuration y PageValues vs Revenue (boxplots)**: las sesiones con compra presentan medianas y rangos superiores en ambas variables, reforzando la idea de que el compromiso con las páginas de producto y el valor de página están asociados a la conversión.

- **Navegación Informativa (Administrative + Informational) vs Revenue**: se contrastó si las sesiones sin compra navegan más contenido informativo.

- **Matriz de correlación numérica**: se calculó la correlación entre variables

numéricas, observándose correlaciones positivas moderadas entre PageValues, ProductRelatedDuration y Revenue, y correlaciones negativas con ExitRates.

Estos resultados orientaron la elección final de las seis hipótesis que se presentan a continuación.

4. Hipótesis planteadas y resolución

A partir del análisis exploratorio y de la literatura del dominio, se formularon seis hipótesis: dos univariadas, dos bivariadas y dos multivariadas.

Para cada una se detalla la definición, la estrategia de abordaje (prueba estadística o modelo) y la discusión de resultados.

4.1. Hipótesis 1: ProductRelatedDuration y probabilidad de compra

3.1.1. Definición de la hipótesis

H1: La duración en páginas de productos (**ProductRelated_Duration**) influye positivamente en la probabilidad de compra (**Revenue**).

Intuición: los usuarios que pasan más tiempo viendo páginas de producto suelen estar más cerca de tomar una decisión de compra.

3.1.2. Estrategia de abordaje

- Variable explicativa: ProductRelated_Duration (continua, asimétrica). ●

Variable respuesta: Revenue (binaria: 0 = sin compra, 1 = con compra).

- Dadas la fuerte asimetría y presencia de outliers, se utilizó el **test de Mann-Whitney U** para comparar la distribución de ProductRelated_Duration entre sesiones con y sin compra, sin asumir normalidad.

3.1.3. Resultados obtenidos y discusión

El test arrojó:

- p-value ≈ 2.09×10^{-122} (mucho menor que 0.05).

Esto indica que la duración en páginas de producto es **significativamente mayor** en sesiones que terminan en compra que en sesiones sin compra. Combinando este resultado con los boxplots, se concluye que:

H1 queda aceptada: un mayor tiempo en páginas de producto se asocia fuertemente con una mayor probabilidad de compra.

4.2. Hipótesis 2: Tipo de visitante y compra

3.2.1. Definición de la hipótesis

H2: Los visitantes recurrentes (**Returning_Visitor**) compran más que los visitantes nuevos (**New_Visitor**).

La motivación viene de la teoría de e-commerce: usuarios que ya conocen el sitio suelen tener mayor confianza y claridad en lo que buscan.

3.2.2. Estrategia de abordaje

- Variables categóricas: VisitorType (Returning_Visitor, New_Visitor, Other) y Revenue (True/False).

- Se utilizó una **prueba de Chi-cuadrado de independencia** sobre la tabla de contingencia VisitorType × Revenue, ya que se trata de evaluar la asociación entre dos variables categóricas.

3.2.3. Resultados obtenidos y discusión

El contraste dio:

- $\chi^2 \approx 130.67$

- p-value $\approx 4.22 \times 10^{-29}$

El p-value es muy inferior a 0.05, por lo que se rechaza la hipótesis nula de independencia. Esto significa que el tipo de visitante está **estadísticamente asociado** a la compra.

Al observar las proporciones, los **Returning_Visitor presentan una tasa de conversión claramente superior** a la de los New_Visitor.

H2 queda aceptada: ser visitante recurrente aumenta la probabilidad de compra.

4.3. Hipótesis 3: PageValues y probabilidad de compra

3.3.1. Definición de la hipótesis

H3: A mayor **PageValues**, mayor probabilidad de compra.

PageValues es una métrica que sintetiza el valor económico esperado de las páginas visitadas antes de la conversión, por lo que se espera que sea mayor en sesiones que terminan en compra.

3.3.2. Estrategia de abordaje

- Variable explicativa: PageValues (continua, fuertemente cero-inflada y asimétrica).
- Variable respuesta: Revenue (0/1).
- Se volvió a utilizar **Mann-Whitney U** con hipótesis alternativa “greater” (PageValues en compra > PageValues en no compra), por las mismas razones de asimetría y presencia de outliers.

3.3.3. Resultados obtenidos y discusión

El test devolvió:

- p-value ≈ 0.0 (numéricamente muy cercano a 0, menor que cualquier umbral usual).

Los boxplots de PageValues según Revenue muestran que las sesiones con compra concentran la mayoría de los valores altos, mientras que en sesiones sin compra la distribución se mantiene muy cerca de cero.

> **H3 queda aceptada**: PageValues es significativamente mayor en sesiones con compra. Cuanto mayor es el PageValue, mayor es la probabilidad de que la sesión termine en una transacción.

4.4. Hipótesis 4: Navegación informativa en sesiones sin compra

3.4.1. Definición de la hipótesis

> **H4:** Las sesiones sin compra tienen más páginas informativas (Administrative + Informational) que las sesiones que sí finalizan en compra.

La idea de dominio es que usuarios que sólo buscan información (ayuda, términos, información general) tienden a no concretar una compra.

3.4.2. Estrategia de abordaje

- Se construyó la variable **Informational_Total = Administrative + Informational**.
- Se aplicó un test de **Mann-Whitney U** comparando Informational_Total entre sesiones sin compra (Revenue = 0) y con compra (Revenue = 1), utilizando la hipótesis alternativa “greater” para reflejar que se esperaba más navegación informativa en sesiones sin compra.

3.4.3. Resultados obtenidos y discusión

El contraste produjo:

- p-value ≈ 1.0

Es decir, no hay evidencia estadística que apoye que las sesiones sin compra tengan más navegación informativa que las que sí compran; de hecho, el resultado sugiere que las distribuciones son muy similares.

> **H4 queda rechazada**: en este dataset, la navegación informativa no diferencia de manera significativa a las sesiones con y sin compra.

Este resultado es interesante porque contradice la intuición inicial y muestra que incluso usuarios que terminan comprando también consultan páginas administrativas o informativas.

4.5. Hipótesis 5: Modelo multivariado con ProductRelated_Duration, PageValues y ExitRates

3.5.1. Definición de la hipótesis

> **H5:** La combinación de **ProductRelated_Duration**, **PageValues** y **ExitRates** influye de forma conjunta en la probabilidad de compra.

Estas tres variables representan, respectivamente, compromiso con páginas de producto, valor económico de la ruta visitada y probabilidad de abandonar el sitio.

3.5.2. Estrategia de abordaje

Se ajustó un modelo de **regresión logística** con:

- Variable respuesta: Revenue (0/1).
- Predictores: ProductRelated_Duration, PageValues y ExitRates.
- Se utilizó *statsmodels* para obtener coeficientes, p-values y pseudo-R^2 del modelo.

La regresión logística es adecuada porque:

- Revenue es binaria.
- Permite interpretar signos y significancia de los coeficientes.
- Es un modelo estándar y fácilmente explicable en un contexto académico.

3.5.3. Resultados obtenidos y discusión

El modelo ajustado mostró:

- Pseudo $R^2 \approx 0.296$ (modelo con capacidad explicativa moderada). ● Los tres predictores resultaron altamente significativos (p-value < 0.001). ● Signos coherentes con el dominio:
    - Coeficiente positivo para **ProductRelated_Duration**: más tiempo en productos → mayor probabilidad de compra.
    - Coeficiente positivo para **PageValues**: mayor valor de las páginas visitadas → mayor probabilidad de compra.
    - Coeficiente negativo fuerte para **ExitRates**: cuanto mayor es la tasa de salida, menor es la probabilidad de compra.

H5 queda aceptada: el conjunto ProductRelated_Duration, PageValues y ExitRates constituye un modelo multivariado estadísticamente significativo para explicar la probabilidad de compra.

4.6. Hipótesis 6: Influencia de BounceRates y ExitRates

3.6.1. Definición de la hipótesis

H6: Las tasas de abandono del sitio (**BounceRates** y **ExitRates**) influyen conjuntamente en la probabilidad de compra.

Ambas variables miden la propensión de los usuarios a abandonar el sitio, por lo que se espera que se relacionen negativamente con la conversión.

3.6.2. Estrategia de abordaje

Se ajustó otra **regresión logística** con:

- Variable respuesta: Revenue.
- Predictores: BounceRates y ExitRates.

De nuevo, la regresión logística permite cuantificar el efecto de cada tasa de abandono sobre la probabilidad de compra y evaluar su significancia conjunta.

3.6.3. Resultados obtenidos y discusión

El modelo presentó:

- Pseudo $R^2 \approx 0.087$ (menor capacidad explicativa que el modelo anterior, pero distinto de cero).

- **ExitRates** resultó altamente significativa (p-value < 0.001) con coeficiente negativo pronunciado.

- **BounceRates** no fue estadísticamente significativa (p-value ≈ 0.27).

Esto indica que:

- **ExitRates** es un predictor importante y negativo de la probabilidad de compra: sesiones donde la página actual concentra muchas salidas tienen menor probabilidad de terminar en una compra.

- BounceRates, por sí sola, no añade información estadísticamente significativa una vez que ExitRates está incluida en el modelo.
  H6 queda parcialmente aceptada: las tasas de abandono influyen en la compra, pero el efecto relevante se concentra en ExitRates. BounceRates no aporta evidencia adicional en este modelo.

5. Conclusiones

El análisis completo del dataset Online Shoppers Purchasing Intention permite extraer varias conclusiones relevantes para un contexto de comercio electrónico:

- Las métricas relacionadas con la **interacción con productos** (ProductRelated, ProductRelated_Duration y PageValues) son los indicadores más fuertemente asociados a la compra. Las sesiones con compra pasan más tiempo en páginas de producto, visitan más productos y presentan valores de página más altos.

- El **tipo de visitante** importa: los visitantes recurrentes muestran una tasa de conversión significativamente superior a la de los nuevos, confirmando que la familiaridad con el sitio incrementa la probabilidad de compra.

- Las variables de **contexto** (tipo de tráfico y mes del año) también influyen en la conversión, aunque su efecto es más heterogéneo y depende de la combinación de canales y periodos.

- Las pruebas estadísticas y los modelos multivariados reforzaron varias de las hipótesis planteadas, destacando el papel de ProductRelated_Duration, PageValues y ExitRates como predictores clave del comportamiento de compra.

En conjunto, los resultados son coherentes con modelos clásicos de analítica web: el compromiso profundo con contenido de producto, unido a una estructura de navegación eficiente que minimice tasas de salida, es fundamental para fomentar la conversión en sitios de e-commerce.

6. Referencias

1. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention based on web analytics data. *Neural Computing and Applications*.

2. Google Analytics Documentation. Definiciones de PageValue, BounceRate y ExitRate.

3. Material teórico de la cátedra de Fundamentos de la Ciencia de Datos – UNICEN.