# Análisis del Comportamiento de Navegación en un E-Commerce

Sebastián Flórez Jaramillo – Ingeniería de Sistemas – Universidad de Antioquia
sebastian.florezj@udea.edu.co

***Resumen*— Este artículo presenta el desarrollo de un proyecto de analítica de datos, en el cual se seleccionó un conjunto de datos con el objetivo de aplicar diferentes técnicas de análisis, descripción y visualización para comprender sus características y patrones relevantes y así obtener información útil para la toma de decisiones. Inicialmente, se llevó a cabo una exploración del contexto de los datos y el planteamiento de una pregunta problematizadora que enmarcó el flujo de estudio, seguida de un análisis descriptivo univariado, bivariado y multivariado que permitió conocer la distribución de las variables y las relaciones entre ellas. Posteriormente, se realizó el ejercicio de detectar – y eliminar – outliers. Para finalizar, se escalaron y codificaron los datos según su naturaleza.**

**Palabras clave:** e-commerce; comportamiento del usuario; análisis exploratorio; navegación; preprocesamiento de datos; revenue.

## I. Introducción

Al navegar en la red, un usuario interactúa continuamente y de distintas formas con la página, con el navegador y con el ordenador. Estas interacciones se pueden capturar, medir y transformar en información útil que apoye la toma de decisiones. Para eso se requiere un conocimiento previo y un análisis profundo de los datos. En un e-commerce, como en cualquier otro negocio, se requiere prestar atención a usuarios y clientes para garantizar una mejora continua de la experiencia y un crecimiento constante del negocio, por lo que es un territorio perfecto para aplicar la analítica de datos.

El presente estudio tiene como propósito realizar un análisis sistemático de un dataset real que contiene distintas métricas aplicadas a las sesiones de distintos usuarios al navegar en un e-commerce y comprender la relación entre su conducta y la probabilidad de realizar una compra. Esta información resulta muy valiosa al momento de implementar actualizaciones para incentivar la venta.

## II. Metodología

### *A. Selección del dataset*

Se revisaron diversos conjuntos de datos disponibles en el repositorio público UCI Machine Learning Repository, considerado una de las fuentes más utilizadas para el desarrollo y evaluación de modelos en ciencia de datos. Tras explorar varias alternativas, se seleccionó el dataset Online Shoppers Purchasing Intention Dataset, debido a que presenta un volumen adecuado de información, compuesto por 12 330 filas y 18 columnas, lo que permite realizar análisis estadísticos con suficiente representatividad.

Además, este conjunto de datos incluye una variedad equilibrada de tipos de variables, entre ellas variables continuas, discretas, categóricas y binarias, lo que lo convierte en un caso idóneo para aplicar múltiples técnicas estadísticas y de preprocesamiento de interés, tales como análisis exploratorio, imputación, detección de valores atípicos y escalamiento. Su estructura y diversidad permiten no solo la comprensión del comportamiento de los usuarios en un entorno de comercio electrónico, sino también la evaluación de patrones de navegación asociados con la intención de compra.

### *B. Técnicas empleadas en el análisis exploratorio, detección de datos atípicos e imputación.*

En el análisis univariado se utilizaron medidas de tendencia central y dispersión, como la media y la desviación estándar, junto con el cálculo de la asimetría y la curtosis. Además, se emplearon herramientas de visualización como histogramas, gráficos de barras y diagramas de caja.

Para el análisis bivariado se calcularon coeficientes de correlación de Pearson y Spearman entre todas las variables numéricas, así como porcentajes para describir el comportamiento de una variable en función de otra. Las técnicas de visualización incluyeron diagramas de dispersión y tablas cruzadas.

Finalmente, en el análisis multivariado se generó un mapa de calor de correlaciones y se realizó un ejercicio exploratorio de reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA); este procedimiento se llevó a cabo únicamente con fines analíticos y no se utilizó para modificar el conjunto de datos original.

Para la detección de valores atípicos se evaluaron inicialmente las variables mediante los métodos tradicionales de rango intercuartílico (IQR) y puntuaciones estándar (*z-scores*). Sin embargo, debido a la naturaleza heterogénea del conjunto de datos, estos enfoques no ofrecieron una discriminación adecuada. Por ello, se optó por utilizar un método basado en aprendizaje no supervisado, específicamente el algoritmo Isolation Forest, con el cual se identificaron y eliminaron aproximadamente el 5 % de los registros considerados anómalos. En cuanto al tratamiento de valores faltantes, el conjunto de datos original no presentaba ausencias, por lo que no fue necesario aplicar técnicas de imputación. Finalmente, se realizó un preprocesamiento adicional consistente en la estandarización de las variables numéricas mediante *StandardScaler* y la codificación de una variable categórica empleando el método *one-hot encoding*.

## III. Resultados y Análisis

El análisis exploratorio inicial univariado, permitió depurar 9 de las 17 columnas predictoras originales del dataset. Muchas de estas columnas fueron eliminadas después del análisis de contexto, dado que, por su naturaleza, su comportamiento no presentaba una relación significativa con la variable objetivo.

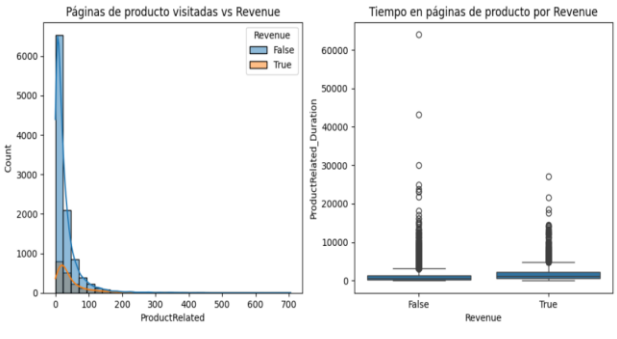


**Ilustración 1**

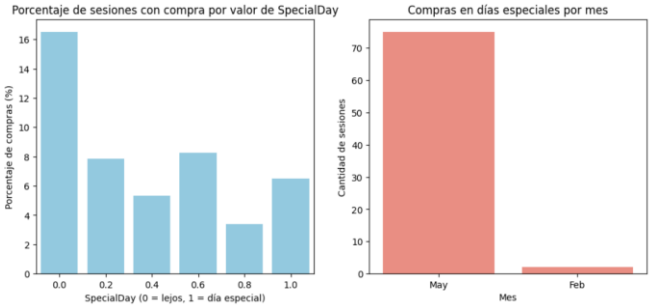


**Ilustración 2**

Las ilustraciones 1 y 2 muestran un ejemplo del análisis realizado de forma individual a cada columna, mostrando su relación con la variable predictora con el objetivo de deducir su relevancia e influencia sobre esta.

Variables - Tabla descriptiva

| Nombre | Tipo | Rango | Definición | Fuente |
|---|---|---|---|---|
| Administrative | Discreta | [0,27] | Páginas de tipo administrativo visitadas por el usuario en la sesión analizada. | Información tomada de la URL durante la navegación |
| Informational | Discreta | [0,24] | Páginas de tipo informativo visitadas por el usuario en la sesión analizada. | Información tomada de la URL durante la navegación |
| ProductRelated | Discreta | [0,705] | Páginas relacionadas con productos visitadas por el usuario en la sesión analizada. | Información tomada de la URL durante la navegación |
| Administrative_Duration | Continua | [0, 3398.75] | Tiempo que ha pasado el usuario visitando páginas de tipo administrativo | Información tomada de la URL durante la navegación |
| Informational_Duration | Continua | [0, 2549.375] | Tiempo que ha pasado el usuario visitando páginas de tipo informativo | Información tomada de la URL durante la navegación |
| ProductRelated_Duration | Continua | [0, 63973.52223] | Tiempo que ha pasado el usuario visitando páginas relacionadas a productos | Información tomada de la URL durante la navegación |
| Weekend | Binaria | {0, 1} | Toma un valor de 1 si es fin de semana y 0 en caso contrario | Toma de muestras |
| VisitorType | Categórica | "Returning_Visitor", "New_Visitor", "Other" | Describe si el visitante es nuevo, antiguo, o no registra información | Toma de muestras |

**Tabla 1**

La tabla 1 muestra una descripción de las columnas que permanecieron en el dataset luego de la depuración.

Luego de depurar las columnas necesarias para el análisis, el siguiente paso consiste es encontrar las correlaciones que, de no manejarse bien, pueden afectar considerablemente la utilidad del dataset.

Lo primero que se hizo, fue calcular los coeficientes de correlación de Pearson y Spearman para detectar relaciones lineales y/o monótonas. De este ejercicio surgió la conclusión de que en el dataset se tienen 3 parejas de variables que están fuertemente relacionadas entre sí y débilmente relacionadas con las demás. Esto constituye uno de los hallazgos mas importantes del ejercicio. En las siguientes figuras se pueden observar estas relaciones.

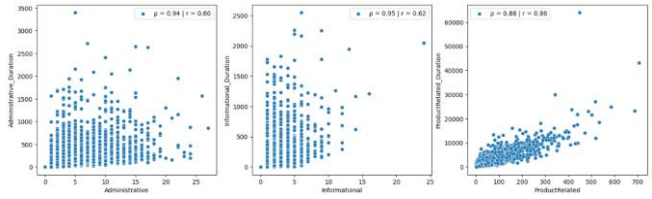

**Ilustración 3**

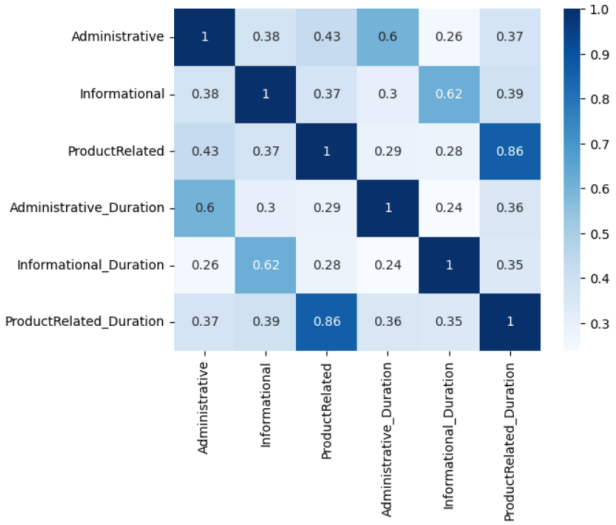


**Ilustración 4**

Se tiene entonces que las parejas de predictoras que consisten en la cantidad de páginas del tipo X (administrativas, informativas o relacionadas a productos) visitadas y el tiempo total en dicho tipo de páginas guardan una relación no lineal fuerte.

Las otras variables se analizaron principalmente en función de la variable objetivo. A continuación, repasamos los hallazgos.

```
Porcentaje de compras vs no compras por Weekend:
Revenue      False      True
Weekend
False     85.108856  14.891144
True      82.601116  17.398884
```


**Ilustración 5**

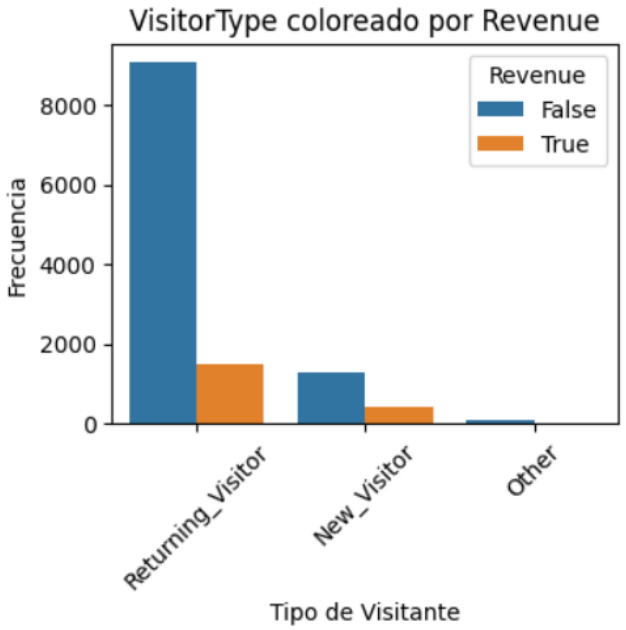


**Ilustración 6**

Las gráficas anteriores también brindan información importante para el análisis, se puede evidenciar que hay una mayor tendencia a realizar compras en aquellas sesiones durante los fines de semana y aquellas sesiones pertenecientes a usuarios nuevos.

Por ejemplo, con esta información un administrador podría decidir concentrar sus esfuerzos de marketing los fines de semana y con un enfoque mas agresivo a usuarios nuevos.

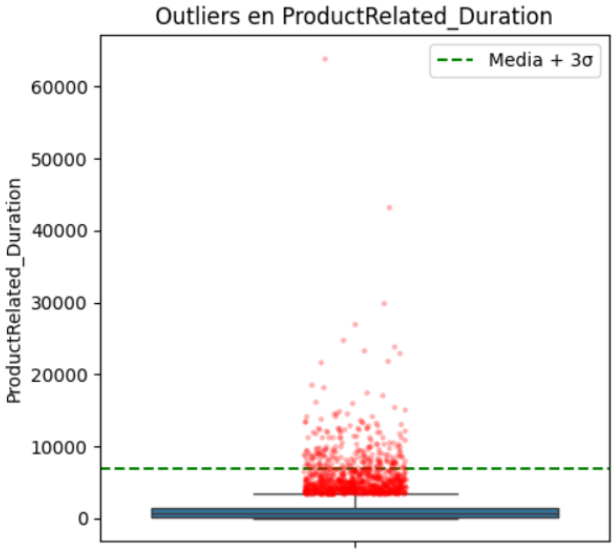


**Ilustración 7**

Pasando a la fase de detección de atípicos, en esta ilustración podemos observar tanto la técnica de z-score como IQR. Se puede observar una alta densidad de candidatos a outliers. Una densidad similar para cada variable derivó en la utilización de una herramienta un tanto mas robusta y efectiva para aplicar simultáneamente a distintas variables.

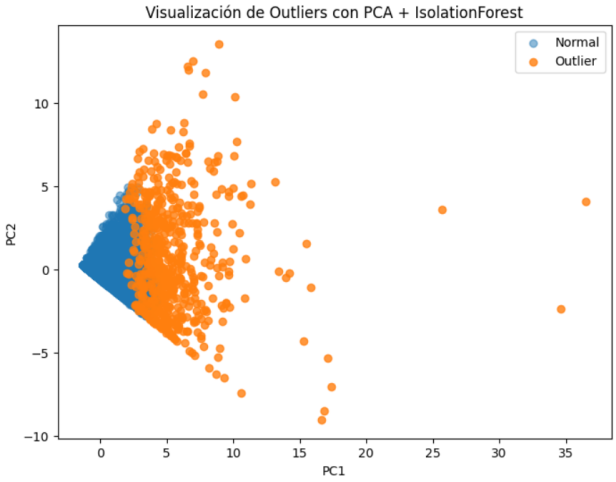


**Ilustración 8**

Así entonces, la técnica seleccionada para detección de atípicos fue isolation forest, que presenta ventajas para datasets grandes y de alto grado de dimensionalidad. Para observar gráficamente las poblaciones etiquetadas como outlier y no outlier por Isolation Forest, dado que el dataset tiene 8 variables predictoras, se usó reducción de dimensionalidad por PCA a 2 componentes y se grafico el diagrama de dispersión de PC1 vs PC2.

Una vez se eliminaron los outliers, el paso restante en el procesamiento del dataset sería escalonamiento de variables númericas y codificación de variables categóricas.

Estos procesos se realizaron por StandarScaler y one-hot encoding respectivamente

## IV. CONCLUSIONES

Finalizado el procesamiento del dataset original, se obtuvo un conjunto de datos compuesto por 8 columnas predictoras y 1 columna objetivo, a partir de un total inicial de 18 columnas. Asimismo, el número de registros pasó de 12330 a 11713, luego de descartar las muestras consideradas atípicas.

En cuanto al comportamiento de los usuarios en un e-commerce, comparado con el comportamiento habitual en comercios físicos, se encuentran algunas concordancias como una mayor probabilidad de conversión (venta) para usuarios o clientes nuevos, y aunque la diferencia de compras entre semana y fin de semana era poca, al igual que sucede en el mundo físico, hay una mayor tendencia a compras en findes de semana. Por otro lado, muy a diferencia de lo que se podría esperar, no hay evidencia que sostenga que hay una mayor tendencia de compras en fechas cercanas a un día feriado o especial (navidad, día de la madre, San valentin, etc).
Por último, los datos indican que no hay una relación evidente entre ver muchas o pocas páginas del tipo que sean, y la probabilidad de compra; sin embargo, si existe una tendencia al aumento de probabilidad de compra en sesiones en las cuales se pasa mas tiempo mirando las páginas relacionadas a productos.

En general, el proceso permitió preparar un dataset limpio, consistente y adecuado para la fase de modelado que no corresponde al alcance de este proyecto.