Proyecto Final Modelos y Simulación II

Camilo Benavides Ramirez, Nicolás Carmona Cardona, Brayan Santiago Ramirez Silva

I. Introducción

El crecimiento del comercio electrónico ha intensificado la necesidad de comprender el comportamiento del consumidor en línea para optimizar la experiencia de usuario y mejorar las tasas de conversión. Un desafío crítico en este dominio es la predicción de la intención de compra, fundamental para reducir el abandono de carritos. El Aprendizaje Automático (Machine Learning) ofrece herramientas robustas para abordar este problema, permitiendo la identificación de patrones complejos a partir de los datos de navegación del usuario (*clickstream*). Este proyecto se enfoca en aplicar y evaluar diversas técnicas de clasificación supervisada sobre el conjunto de datos "Online Shoppers Purchasing Intention" con el objetivo de desarrollar un modelo predictivo preciso y eficiente.

II. Descripción del problema

El comercio electrónico ha crecido exponencialmente en los últimos años, convirtiéndose en un canal esencial para las empresas, pero enfrentando el desafío del alto abandono de compra por parte de los usuarios. Comprender y anticipar la intención de compra es clave para mejorar la experiencia del cliente y aumentar las tasas de conversión. En este contexto, una solución basada en Machine Learning (ML) resulta altamente útil, ya que permite analizar grandes volúmenes de datos de navegación, identificar patrones complejos en el comportamiento del usuario y predecir en tiempo real la probabilidad de que una sesión culmine en una compra, optimizando así las estrategias comerciales y la toma de decisiones.

El análisis exploratorio del dataset utilizado mostró que no existen valores faltantes en ninguna de las 18 columnas (17 características + 1 objetivo), lo que elimina la necesidad de aplicar estrategias de imputación de datos y simplifica el preprocesamiento. La base de datos está compuesta por variables numéricas continuas y discretas, categóricas y binarias, las cuales representan diferentes aspectos del comportamiento de los usuarios dentro de un entorno de comercio electrónico. A continuación, se presenta un resumen de las variables que componen el conjunto de datos.

TABLE I
Resumen de las Variables del Dataset

Tipo de Variable	Nombre	Descripción
Numéricas Continuas	ProductRelated_Duration, BounceRates, ExitRates	Métricas de tiempo, tasas o valores promedio (ej. tiempo en páginas, tasas de rebote/salida, valor de página).
Numéricas Discretas	Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, PageValues, SpecialDay, OperatingSystems, Browser, Region, TrafficType	Conteo de interacciones o códigos identificadores (ej. número de páginas visitadas, tipo de OS, navegador, región, tráfico).
Categóricas	Month, VisitorType	Variables con categorías fijas sin orden (ej. mes de la visita, tipo de visitante).
Binarias	Weekend, Revenue	Variables con dos posibles estados (ej. fin de semana sí/no, compra realizada sí/no).

Para el procesamiento de las variables categóricas y numéricas, se definió la siguiente estrategia de codificación:

Variables Categóricas Ordinales (Month): Se aplicará *Label Encoding* o *One-Hot Encoding* dependiendo del modelo, dado que los meses presentan cierta ordenación temporal.

Variables Categóricas Nominales (VisitorType): Se empleará *One-Hot Encoding* para crear variables dummy, evitando imponer un orden artificial entre las categorías.

Variables Numéricas Categóricas (OperatingSystems, Browser, Region, TrafficType): Aunque se expresan como números, representan categorías. Se aplicará *One-Hot Encoding* o *Target Encoding* según el rendimiento del modelo, con el fin de capturar relaciones no lineales.

Variables Binarias (Weekend): Se convertirán de booleanas a formato numérico (0/1) para compatibilidad con los algoritmos de aprendizaje automático.

Variables Numéricas Continuas y Discretas: Se aplicará normalización o estandarización dependiendo de los requerimientos de los modelos seleccionados, especialmente para aquellos sensibles a las escalas de las variables, como SVM, KNN y redes neuronales.

El anáslisis del dataset muestra variables predictoras correctamente estructuradas y sin valores faltantes, lo que proporciona una base confiable para el modelado. Sin embargo, se identificó un desbalance significativo en la variable objetivo (Revenue), ya que solo el 15% de las sesiones culminan en una compra, esto exige la aplicación de técnicas de balanceo de clases, como **SMOTE**, y el uso de métricas de evaluación adecuadas como un **F1 score** o **Balanced Error Rate**, en lugar de la exactitud (accuracy), la cual puede resultar engañosa en escenarios desbalanceados.

Asimismo, las variables numéricas presentan distribuciones sesgadas y valores atípicos, especialmente en las métricas de duración, mientras que la matriz de correlación sugiere posibles problemas de multicolinealidad entre variables relacionadas. Con estos hallazgos se tiene la necesidad de aplicar transformaciones y normalizaciones adecuadas, manejar los valores atípicos y realizar una selección cuidadosa de características, priorizando aquellas con mayor capacidad explicativa para optimizar la precisión. En consecuencia, la aproximación de ML propuesta se basará en un paradigma de aprendizaje supervisado, dado que la variable objetivo (*Revenue*) es binaria y representa la conversión o no del usuario. Específicamente, se trata de un problema de clasificación binaria. Este enfoque permitirá entrenar modelos capaces de predecir la probabilidad de compra a partir de los patrones de navegación observados.

III. Estado del arte

La predicción de la intención de compra en línea ha sido abordada desde diversas perspectivas dentro del aprendizaje supervisado. A continuación, se revisan cinco estudios relevantes que emplean diferentes técnicas y metodologías.

A. Sakar et al. (2018) [1]

Este trabajo propone un sistema dual para predecir en tiempo real tanto la intención de compra como el abandono del sitio. Para la intención de compra, se enmarcan en una **clasificación binaria supervisada**, comparando modelos como *Random Forest (RF)*, *Support Vector Machines (SVM)* y *Multilayer Perceptron (MLP)*. Para el abandono, utilizan una red recurrente *Long Short-Term Memory (LSTM)*. La validación se realizó mediante una división *hold-out* (70/30) repetida 100 veces, evaluando con métricas como *Accuracy* y *F1-Score*. El mejor resultado fue obtenido por un **MLP** con sobremuestreo y selección de características, alcanzando un ***Accuracy* del 87.94%** y un **F1-Score de 0.87**.

B. Requena et al. (2020) [2]

Este estudio compara dos estrategias para la predicción: una basada en características diseñadas manualmente (*hand-crafted*) y otra en aprendizaje profundo. En el primer enfoque, se utilizaron clasificadores como *XGBoost* y redes neuronales (*NN*) sobre características como k-grams y motivos de grafos de visibilidad horizontal (HVGm). El segundo empleó arquitecturas como *LSTM* (S2L). Utilizando una validación *hold-out* (80/20), el modelo **S2L** destacó en datos balanceados con un **F1-Score del 91.03%**. Sin embargo, en un escenario desbalanceado, una **NN** con características manuales fue superior, logrando un **AUC de 94.75%**.

C. Tanvir et al. (2023) [3]

Este trabajo se centra en optimizar la predicción mediante *Gradient Boosting*. Se compararon modelos como *SVM*, *RF*, *MLP*, *Decision Tree (DT)* y *XGBoost*, manejando el desbalance de clases con *SMOTE*. La validación se realizó con una división *hold-out* (70/30). El rendimiento se midió con un conjunto robusto de métricas, incluyendo el **Coeficiente de Correlación de Matthews (MCC)**, definido como:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

El modelo que combina **selección de características, SMOTE y XGBoost** obtuvo los mejores resultados, alcanzando una ***accuracy* del 90.65%** y un **auROC de 0.937**.

References

[1] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6893–6908, 2018. [Online]. Available: https://2024.sci-hub.cat/6856/e5cfb30c80dd18fbc513644a1aa089ea/10.1007@s00521-018-3523-0.pdf

[2] B. Requena, G. Cassani, J. Tagliabue, C. Greco, and L. Lacasa, "Shopper intent prediction from clickstream e-commerce data with minimal browsing information," *Scientific Reports*, vol. 10, no. 1, p. 16983, 2020. Available: https://www.nature.com/articles/s41598-020-73622-y

[3] A.-A. Tanvir, I. A. Khandokar, A. M. Muzahidul Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, p. e15163, 2023. [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(23)02370-8

[4] M. A. Gomes, R. Meyes, P. Meisen, and T. Meisen, "Will This Online Shopping Session Succeed? Predicting Customer's Purchase Intention Using Embeddings," in *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22)*, 2022, pp. 2873–2882. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3511808.3557127

[5] M. S. Satu and S. F. Islam, "Modeling online customer purchase intention behavior applying different feature engineering and classification techniques," *Discover Artificial Intelligence*, vol. 3, no. 1, p. 36, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s44163-023-00086-0