
APRENDIZAJE AUTOMÁTICO (2025/26)

TRABAJO

Satisfacción de Usuarios en Compras Online

Motivación

En este trabajo utilizaremos un conjunto de datos disponible en UCI Machine Learning Repository sobre la intención de compra de los usuarios en plataformas de comercio electrónico. Este dataset fue recopilado para el análisis de comportamiento de compra de los usuarios en línea y es una herramienta valiosa para desarrollar modelos predictivos en el contexto del marketing y la experiencia del cliente.

La historia detrás de este conjunto de datos es la siguiente:

Imagina que trabajas como analista de datos en una tienda en línea internacional que ha notado un comportamiento de compra fluctuante en sus usuarios. Aunque muchos usuarios navegan por el sitio web y agregan productos al carrito, un porcentaje significativo de ellos no finaliza la compra, lo que genera una pérdida potencial de ventas. La empresa ha decidido recopilar datos detallados sobre el comportamiento de los usuarios en su plataforma con el fin de entender mejor las razones.

El reto para la tienda en línea es claro: el mercado de comercio electrónico es extremadamente competitivo, y la fidelización de los clientes está cada vez más complicada. Los compradores de hoy buscan una experiencia de usuario rápida, conveniente y sin fricciones, pero las razones por las cuales abandonan sus compras no son siempre obvias. ¿Es la velocidad del sitio web? ¿El diseño de la página? ¿La disponibilidad de productos? ¿Las recomendaciones de productos o las ofertas de descuento?

Para abordar este problema, la tienda ha recopilado una cantidad significativa de datos sobre el comportamiento de los usuarios en el sitio web, que incluyen información sobre las visitas, el tiempo de permanencia, la interacción con las páginas de productos y las características de los usuarios, como la duración de la sesión y la hora del día en que accedieron a la tienda. Sin embargo, el equipo de marketing y ventas no tiene claridad sobre qué factores están influyendo realmente en la intención de compra, y cómo pueden mejorar la experiencia del usuario para aumentar la tasa de conversión.

Tu misión como analista de datos es ayudar a la tienda a entender estos patrones. Deberás desarrollar un modelo predictivo que permita identificar si un usuario tiene alta o baja probabilidad de realizar una compra, basándose en su comportamiento y características demográficas. Con esta información, la tienda podrá ajustar su estrategia de marketing, optimizar la interfaz del sitio y mejorar la experiencia de usuario para incrementar las ventas y la fidelidad de los clientes.

Puedes consultar más información sobre este dataset y otros trabajos realizados en el siguiente enlace:

[Dataset de Compras Online](#)

Dataset

A continuación, se describen las características presentes en el dataset:

- **Administrative:** Número de páginas visitadas en la sección administrativa del sitio web.
- **Administrative Duration:** Duración total en segundos de la sesión en la sección administrativa.
- **Informational:** Número de páginas visitadas en la sección informativa del sitio web.
- **Informational Duration:** Duración total en segundos de la sesión en la sección informativa.
- **Product Related:** Número de páginas visitadas en la sección de productos del sitio web.
- **Product Related Duration:** Duración total en segundos de la sesión en la sección de productos.
- **Bounce Rates:** Tasa de rebote, que indica si el usuario abandona la página rápidamente o interactúa con el sitio (según Google Analytics) (Intervalo [0-1]).
- **Exit Rates:** Tasa de salida, que indica si el usuario abandona la página en determinado momento (según Google Analytics) (Intervalo [0-1]).
- **Page Value:** Valor total de una página, basado en el número de interacciones y conversiones realizadas en ella. (según Google Analytics).
- **Special Day:** Cercanía a un evento especial (por ejemplo, promociones, fechas clave) (normalizado entre 0-1).
- **Month:** Mes en que el usuario realizó la visita (Feb, May, Dec,...).
- **Operating Systems:** Tipo de sistema operativo usado por el usuario (Enum).
- **Browser:** Navegador web que el usuario usó para acceder al sitio (Enum).
- **Region:** Región geográfica desde la que el usuario accedió al sitio web (Enum).
- **Traffic Type:** Tipo de tráfico que ha dirigido al usuario al sitio (Enum).
- **Visitor Type:** Tipo de visitante (nuevo o recurrente).
- **Weekend:** Indicador de si la visita ocurrió en fin de semana (1: Sí, 0: No).
- **Revenue:** Si el usuario realizó una compra (1: Compra, 0: No compra).

Se os facilitará un dataset diferente al que aparece en la página de UCI Machine Learning para que entrenéis vuestros modelos y nos reservaremos un dataset para evaluar el modelo final que habéis seleccionado. Deberéis dividir vosotros mismos el dataset proporcionado en train/test para poder validar los modelos que estáis entrenando. **NO ESTÁ PERMITIDO ENTRENAR LOS MODELOS CON UN DATASET DIFERENTE AL PROPORCIONADO EN EL MOODLE.**

Código

Deberéis entregar un fichero en Python llamado APS_Solver.py en el que se encuentre una clase denominada APS_Solver que deberá contener los siguientes métodos (y todos los demás que necesites):

- `load_model(self, model_path)` -> Cargará un modelo ya entrenado de una ruta dada.
- `save_model(self, model_path)` -> Guardará el modelo entrenado en la ruta especificada.
- `train_model(self, file_path)` -> Realizará todo el proceso de limpieza, preprocesado de datos y entrenamiento del fichero de entrenamiento ubicado en `file_path`.
- `test_model(self, file_path)` -> Realizará el proceso de limpieza, preprocesado y predicción del fichero de test ubicado en `file_path`.

A mayores, el modelo podrá tener todos los métodos extras que consideréis oportunos para el correcto funcionamiento del modelo (para la limpieza, preprocesado, normalización, etc)

Prueba del código

Para probar el código vamos a ejecutar el siguiente bloque de código que DEBE funcionar con vuestra clase APS_SOLVER:

```
from GrupoA.APS_Solver import APS_Solver
model = APS_Solver()
model.train_model("online_shoppers_train.csv")
model.test_model("online_shoppers_test.csv")
```

cuyo resultado debe ser que se muestre por terminal lo siguiente:

Tasa de error: 0.16564417177914115

Precisión: 0.8365231259968102

Recall: 0.8221003134796239

F1-score: 0.8292490118577075

Informe

Deberéis entregar un informe con una estructura como la que se indica a continuación, donde indicaréis lo que se ha hecho en cada apartado.

0. Portada indicando autores del trabajo

1. Análisis exploratorio de datos (ADE):

- Explorar y analizar los datos para comprender su estructura y características.

2. Limpieza de datos:

- Indicar si se han tratado valores faltantes o nulos y de qué manera.

3. Ingeniería de características:

- Indicar si se han implementado métodos para generar nuevas características a partir de otra característica del conjunto de datos.
- Buscar información relevante que se pueda obtener combinando características

4. Selección de características:

- Indicar si se ha realizado selección de las características (variables) relevantes para el problema.
- Indicar si se ha realizado eliminación de características irrelevantes o altamente correlacionadas.

5. Transformación de datos:

- Escalar o normalizar las características para que tengan una escala similar.
- Codificar variables categóricas en un formato numérico si es necesario.

6. División de datos:

- Dividir el conjunto de datos en conjuntos de entrenamiento y prueba.
- Los datos de entrenamiento se utilizan para entrenar el modelo y los de prueba para evaluar el rendimiento final.

7. Selección de modelo:

- Elegir el tipo de modelo de aprendizaje automático que se utilizará (regresión, clasificación, agrupamiento, etc.).
- Seleccionar algoritmos y arquitecturas específicas (seleccionar dos modelos).

8. Entrenamiento del modelo:

- Alimentar los datos de entrenamiento al modelo para que aprenda de los patrones presentes en los datos.

9. Validación y ajuste de hiperparámetros:

- Evaluar el rendimiento del modelo en el conjunto de prueba.
- Ajustar hiperparámetros para optimizar el rendimiento.

10. Evaluación de los modelos:

- Evaluar el rendimiento final del modelo en el conjunto de prueba para estimar su capacidad de generalización. Emplear como métricas tasa de error, precisión, recall y F-score.

11. Elección del modelo final

- Elegir el mejor modelo para este problema.

El trabajo para la asignatura APRENDIZAJE AUTOMÁTICO se realizará en grupos de 3 personas.

La puntuación asignada al trabajo es 15 puntos (sobre un total de 100 en la asignatura).

Se realizará una entrega en Agora del análisis de datos (EDA) con fecha límite el viernes 21 de noviembre a las 23:00.

Se realizará la entrega en Ágora (por uno de los miembros del grupo) del código, el modelo guardado como.pkl, y del informe final con fecha límite jueves 11 diciembre 23:00h.

Se realizará la entrega en Ágora de las diapositivas empleadas para la presentación con fecha límite viernes 12 diciembre 23:00h.

La presentación del trabajo se hará la semana del 16 de diciembre.

*Por cada día o fracción de día de retraso en algunas de las entregas, se aplicará una penalización del **10 %** sobre la nota final del trabajo. Por ejemplo, entre 0-24h de retraso en alguna de las entregas, 10% menos en la nota final. Entre 24-48h de retraso en alguna de las de las entregas, 20% menos en la nota final.*