Proyecto Final de Data Mining

Predicción Avanzada de *LTV*, *CAC* y *ROMI* para *Showz*

28 de abril de 2025

Contexto

En el **PSet #4** analizaste las métricas históricas de visitas, ventas y marketing de *Showz*. El **Proyecto Final de Data Mining** amplía ese trabajo: construirás **modelos predictivos** que anticipen el *Lifetime Value* (LTV), el *Customer Acquisition Cost* (CAC) y la rentabilidad futura (*ROMI*) por usuario / fuente de adquisición. Integrarás las técnicas de **Machine Learning** vistas en la materia (regresión, clasificación, árboles, boosting, ensambladores, validación, explicabilidad).

1. Objetivos

1. Predecir el **LTV a 6–12 meses** de cada cliente nuevo.
2. Estimar el **CAC esperado** por fuente y cohort para el próximo trimestre.
3. Identificar los **drivers clave** que impulsan LTV y CAC (explicabilidad).
4. Generar una **estrategia de asignación de presupuesto** basada en predicciones y simulaciones de ROMI.

2. Datos disponibles

Se usarán los mismos `visits`, `orders` y `costs` de PSet #4. Puedes añadir **features externas** (festivos, clima, redes sociales, etc.) si lo documentas y lo incluyes en `data/external/`.

3. Actividades obligatorias

Paso 1. Ingeniería de características avanzada

- Variables de comportamiento (nº de sesiones, frecuencia, retencion/abandono, AOV-average order value, tendencia de compra/gasto).
- Variables temporales (efecto estacional, mes del evento, etc).
- Variables de marketing (fuente, dispositivo, interacción con campañas).
- Etiquetas ("target"):
  - *LTV_180*: ingreso acumulado de un usuario en los 180 días posteriores a su primera sesión.
  - *CAC_source_30*: costo medio de adquisición de su fuente en los 30 días posteriores a la conversion.

Paso 2. Modelado predictivo

Paso 2.a) Baseline: Regresiónes: Lineal, Estocastica, Ridge, etc.

Paso 2.b) Modelos avanzados: Random Forest, Gradient Boosting (XGBoost, LightGBM o CatBoost).

Paso 2.c) **Ensamblador** (stacking y blending).

Paso 2.d) Enfoque temporal: train 2017, validation 1^{er} sem 2018, test 2^{do} sem 2018.

Paso 2.e) Métricas: MAE, RMSE, MAPE para LTV; MAE y MAPE para CAC.

Paso 3. Validación y selección

- Validación cruzada con *TimeSeriesSplit*.
- Ajuste de hiperparámetros (`GridSearchCV`).
- Selección final basada en desempeño y simplicidad.

Paso 4. Explicabilidad y diagnóstico

- Importancia de variables (gain, permutation).
- SHAP o PDP para los features top 5.
- Análisis de errores sistemáticos (segmentos donde el modelo falla).

Paso 5. Estrategia de marketing basada en simulación

- Usar predicciones de LTV y CAC para estimar ROMI por fuente.
- Simular escenarios: +10 % de presupuesto en fuente A vs. redistribución proporcional.
- Recomendar la asignación óptima y cuantificar el beneficio esperado.

4. Entregables

1. **Informe PDF**: resumen ejecutivo + detalles técnicos con enfoque CRISP-DM.
2. **Presentación (10–15 diapositivas)**.
3. **Repositorio GitHub**:

.

```
data/
   raw/           CSV originales
   processed/     datasets con features y targets
   external/      fuentes adicionales (opcional)

notebooks/
   01_EDA.ipynb
   02_FeatureEngineering.ipynb
   03_ModelTraining.ipynb
   Final_Project_Showz_LTV_CAC.ipynb

models/            artefactos .pkl o .joblib
src/
   features.py
```

```
    train.py
    evaluation.py
    utils.py
 reports/
    figures/
    executive_summary.pdf
    slides/
        presentation.pdf
 requirements.txt
 README.md
 .gitignore
```

5. Criterios de evaluación

- **Calidad del dataset y features** – ingeniería creativa y bien documentada.
- **Rigor del modelado**: pipeline reproducible, validación temporal, tuning justificado.
- **Desempeño**: mejora significativa sobre el baseline.
- **Explicabilidad**: interpretación clara de variables y riesgos.
- **Recomendaciones de negocio**: simulaciones coherentes, ROI cuantificado.
- **Presentación profesional** y repositorio limpio/reproducible.

¡Éxitos desarrollando el sistema de predicción de LTV y CAC para Showz!