SHOWDVANCED

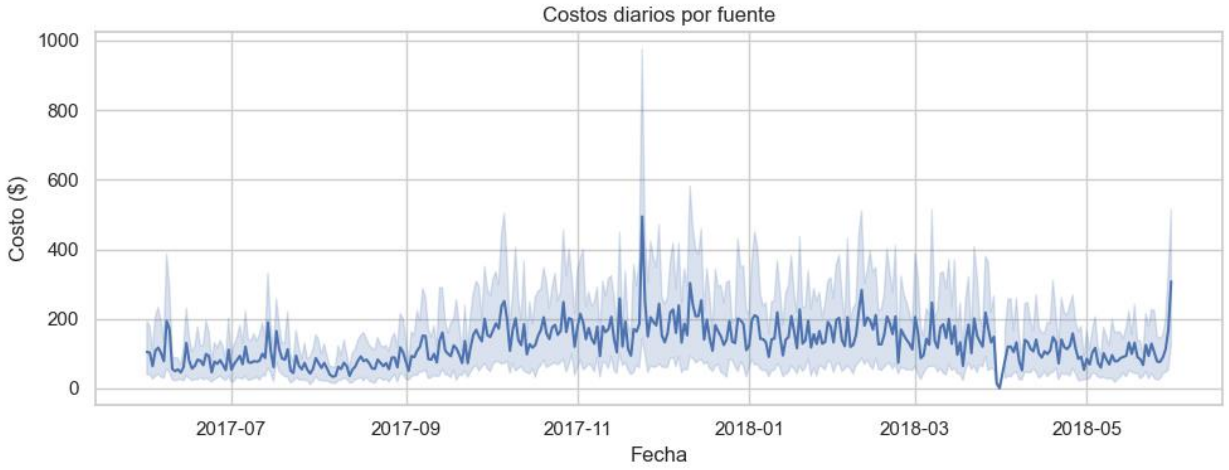
Guillermo Tapia

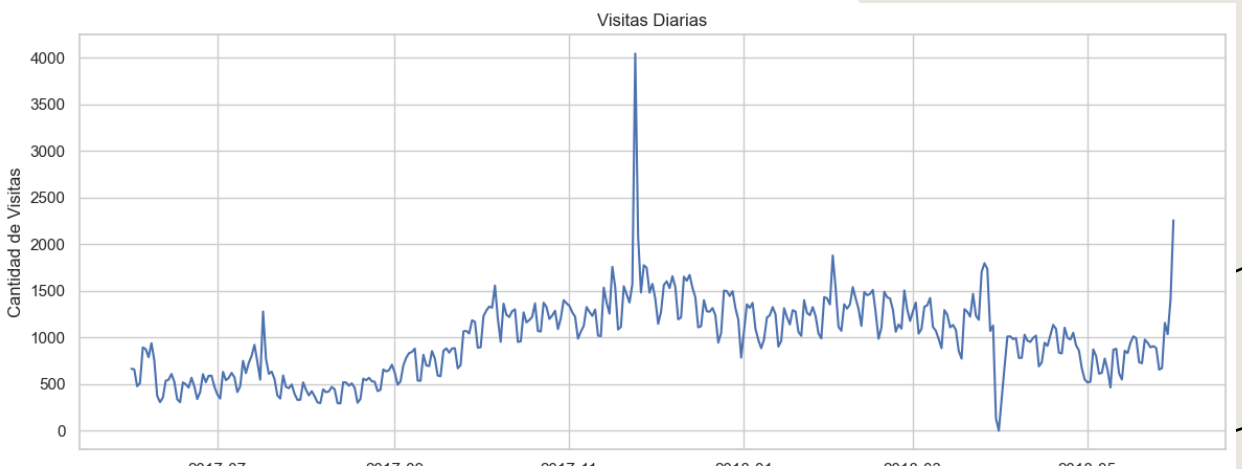
Proyecto Final de Minería de Datos

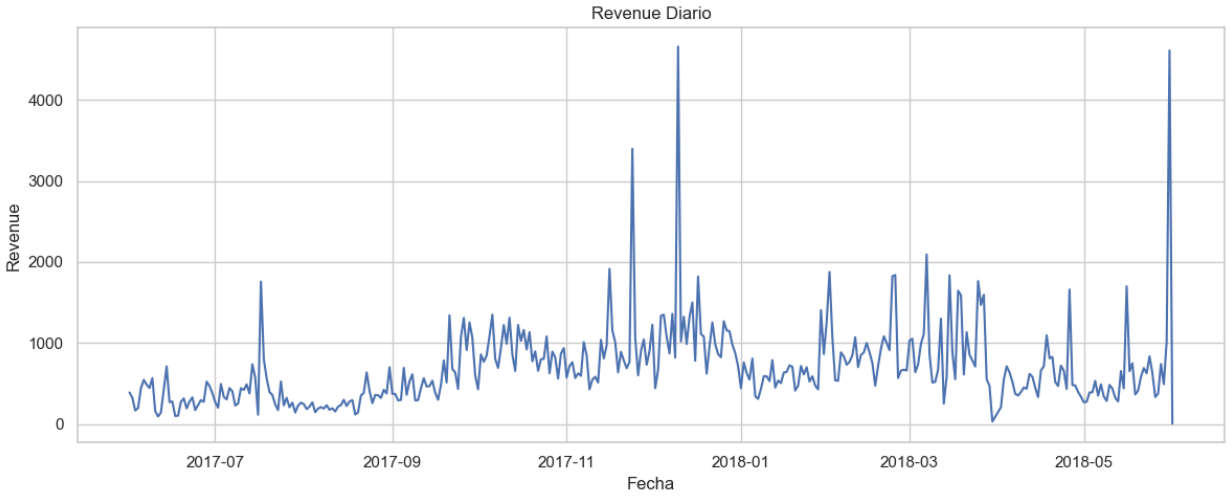
ANÁLISIS EXPLORATORIO DE DATOS

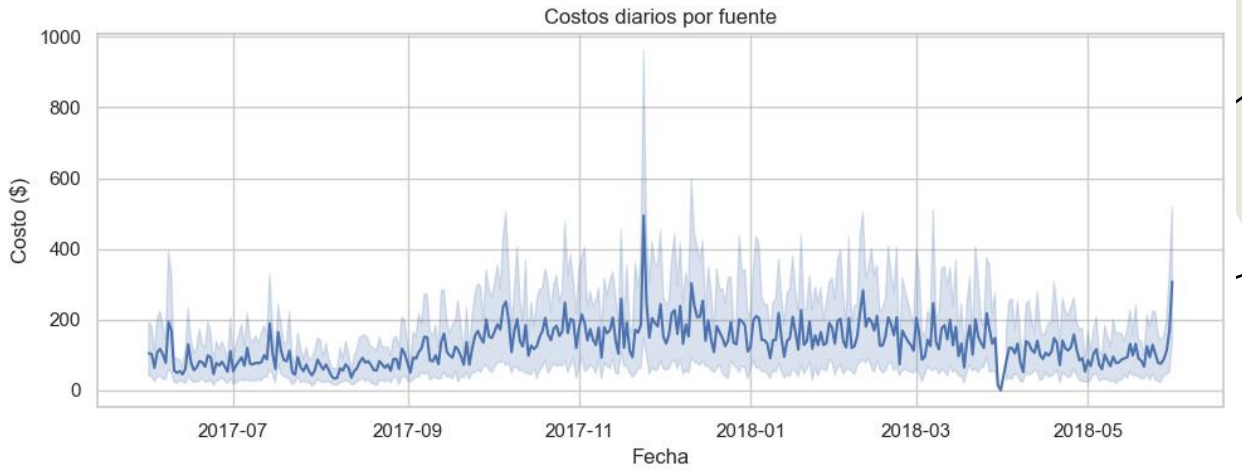
Archivos Usados

- orders.csv → (buy_ts, revenue, uid, buy_date)
- costs.csv → (source_id, dt, costs)
- visits.csv → (start_ts, end_ts, uid, source_id, etc.)



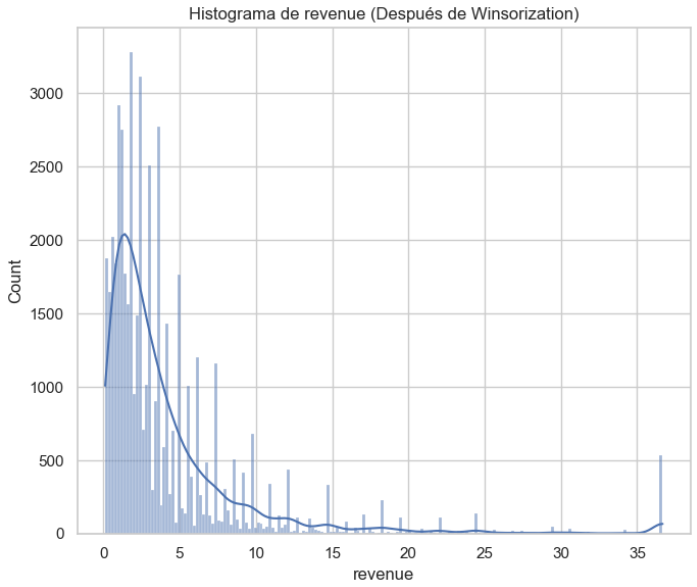


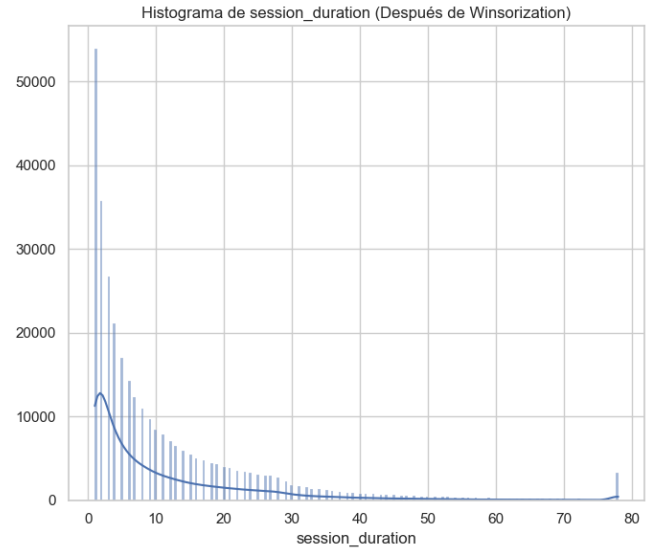


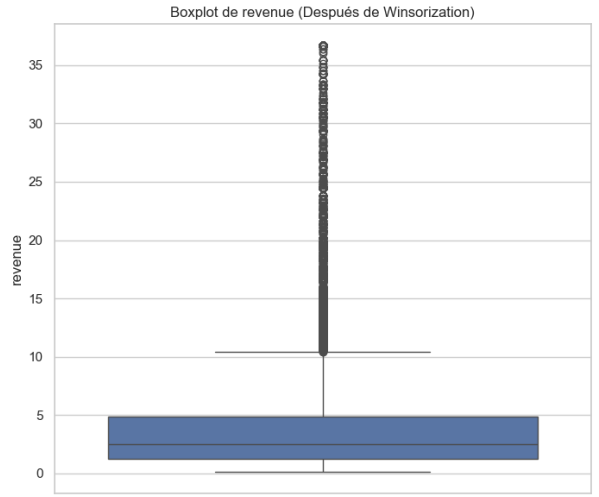


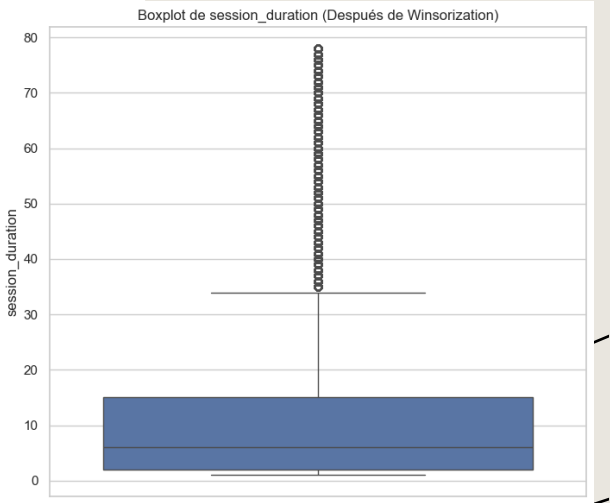
ANÁLISIS EXPLORATORIO DE DATOS

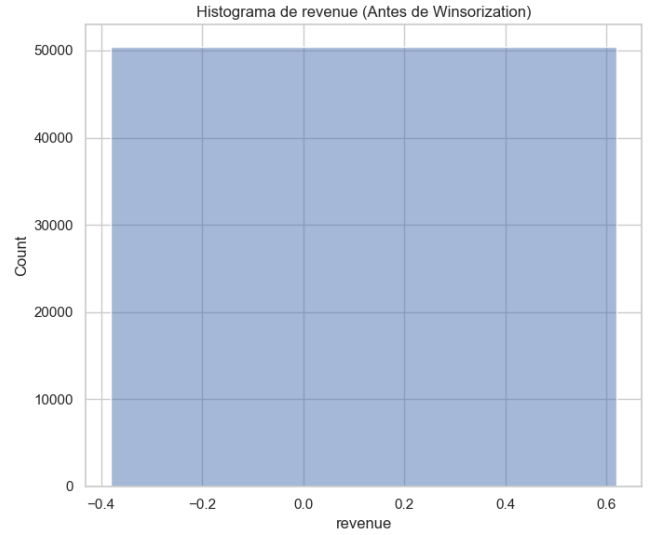
Sesiones negativas -> Borradas Winsorización **suave** (recorte por percentiles: 1% y 99%)

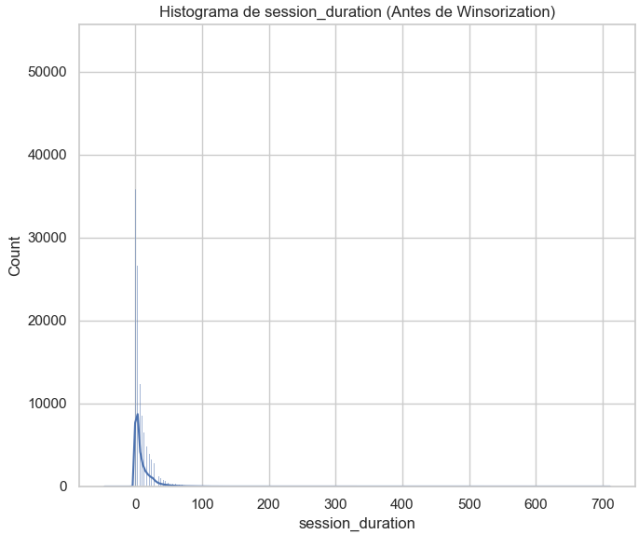


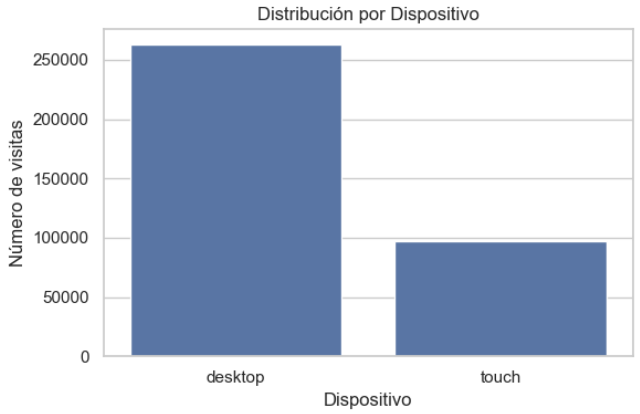


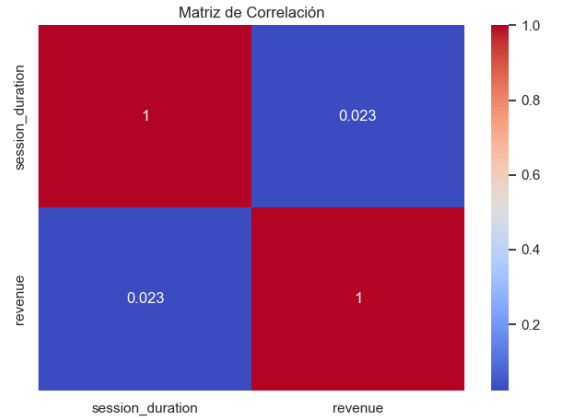












RESUMEN

- **Datos analizados:**
- 359,400 visitas
- 50,415 órdenes
- 2,542 registros de costos
- Columnas: uid, device, source_id, start_ts, end_ts, revenue, costs
- **Usuarios y visitas:**
- Dispositivos más usados: desktop (73%) y touch (27%)
- Duración media de sesión: **10.7 minutos**
- Valores atípicos detectados (duraciones negativas y muy largas)
- **Órdenes:**
- Ingresos con alta variabilidad, hasta **$2,633**
- Mediana de ingresos: **$2.50**, muchos valores bajos
- **Costos de campañas:**
- Costo medio diario por fuente: **$129.48**
- Valores extremos detectados (hasta **$1,788**)
- Fuente más común: source_id 4
- **Rango temporal:**
- Todos los datasets cubren de **junio 2017 a junio 2018**

Variable	Descripción
num_sessions	Total sesiones por usuario
avg_duration	Promedio por sesión del usuario
max_duration	Duración máxima de una sesión
first_session	Primera sesión del usuario
last_session	Última sesión del usuario

Variable	Descripción
CAC_source_30	Costo de adquisición por usuario, calculado como el costo de su fuente en los 30 días antes de su primera sesión
marketing_channel	Canal de marketing asociado al usuario (numérico o categórico codificado)

Variable	Descripción
LTV_180	Lifetime Value acumulado por el usuario en los 180 días posteriores a su primera compra

INGENIERÍA DE CARACTERÍSTICAS AVANZADA

ACTIVIDADES

Transformaciones Aplicadas

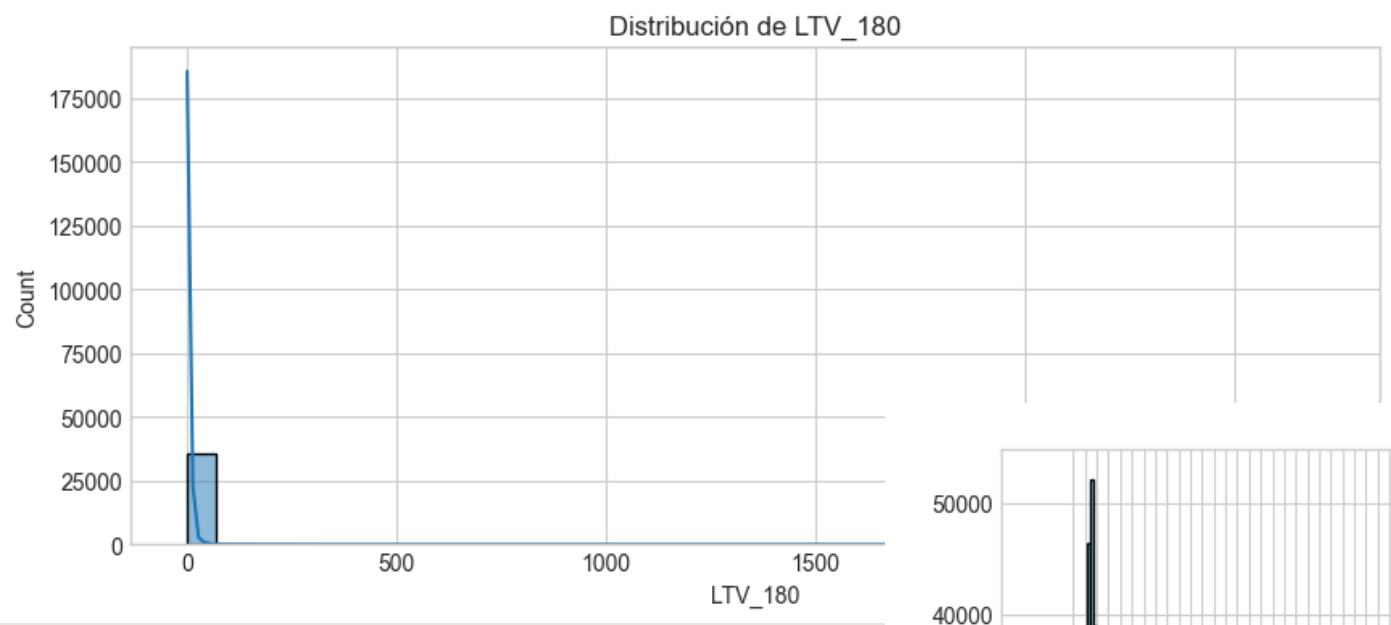
Creación de Nuevas Variables

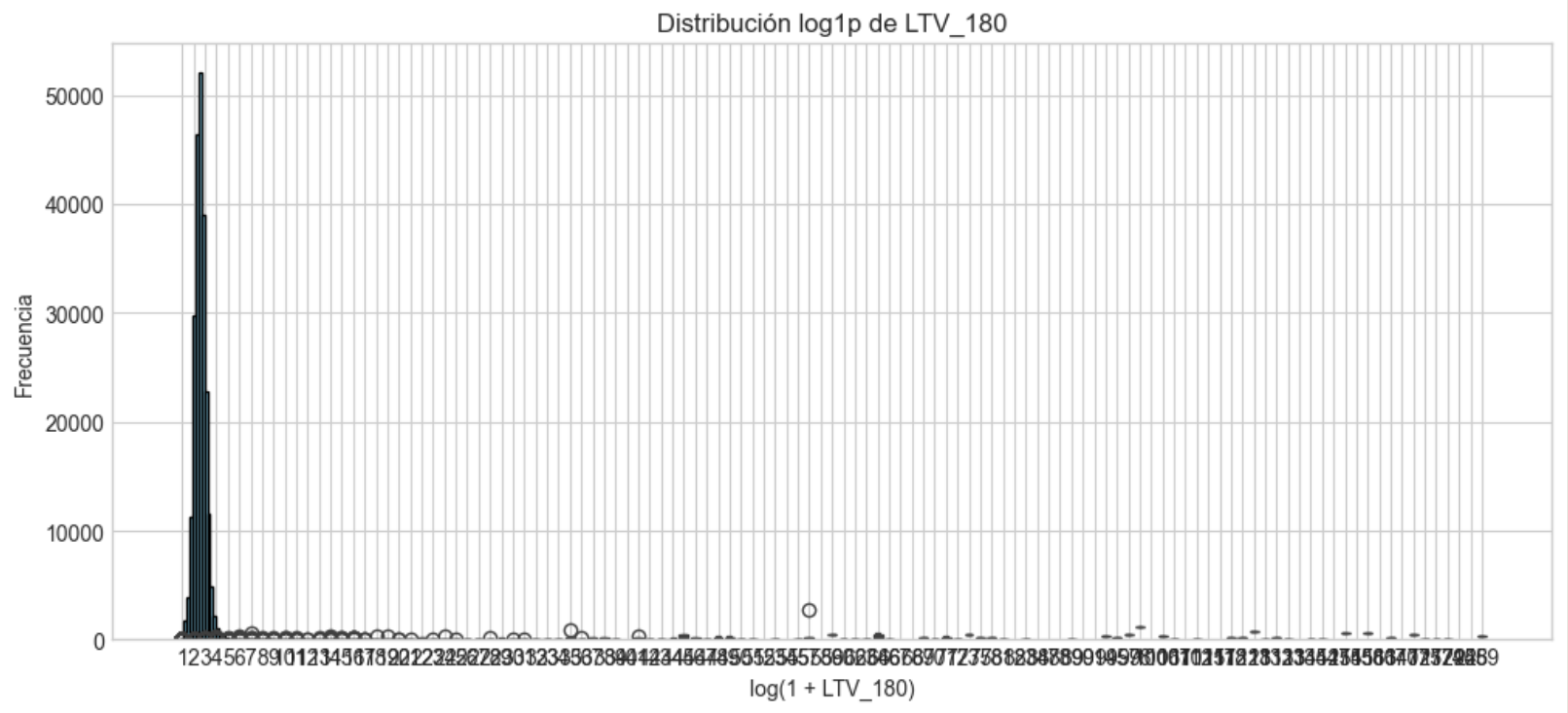
Selección de Variables

Importancia de Variables y Explicación del Modelo

TRANSFORMACIONES APLICADAS

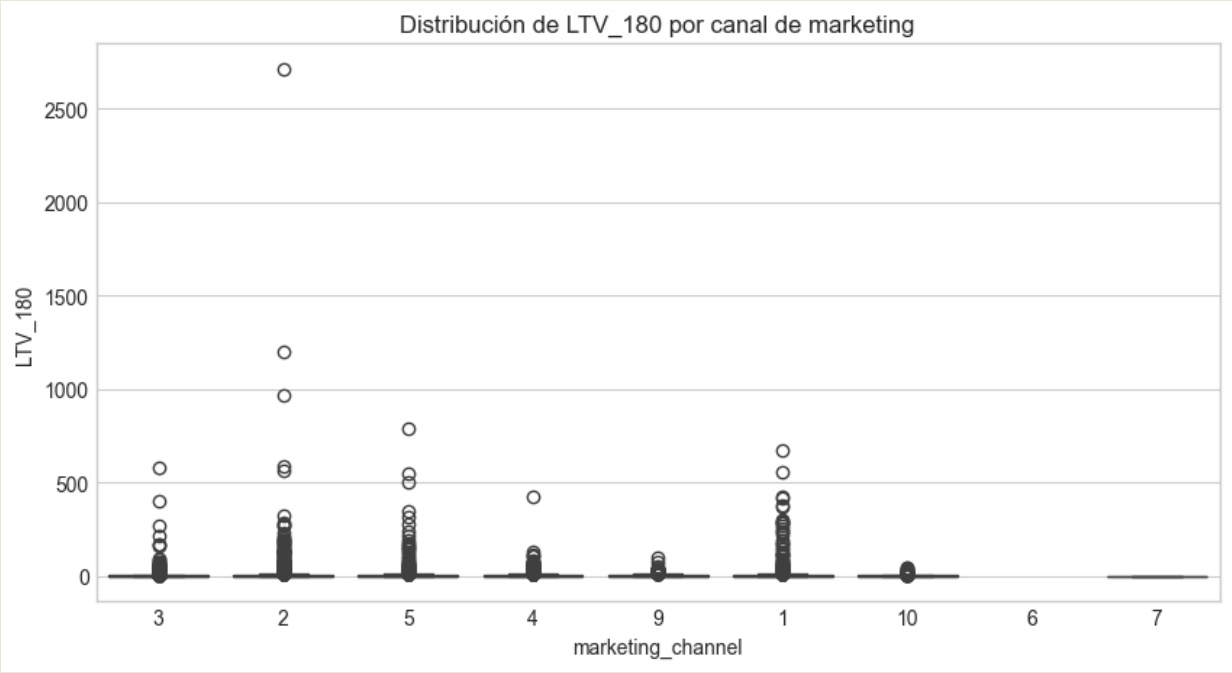
Algunas variables presentaban mucha asimetría y valores atípicos. Se aplicaron transformaciones logarítmicas para estabilizar la data.

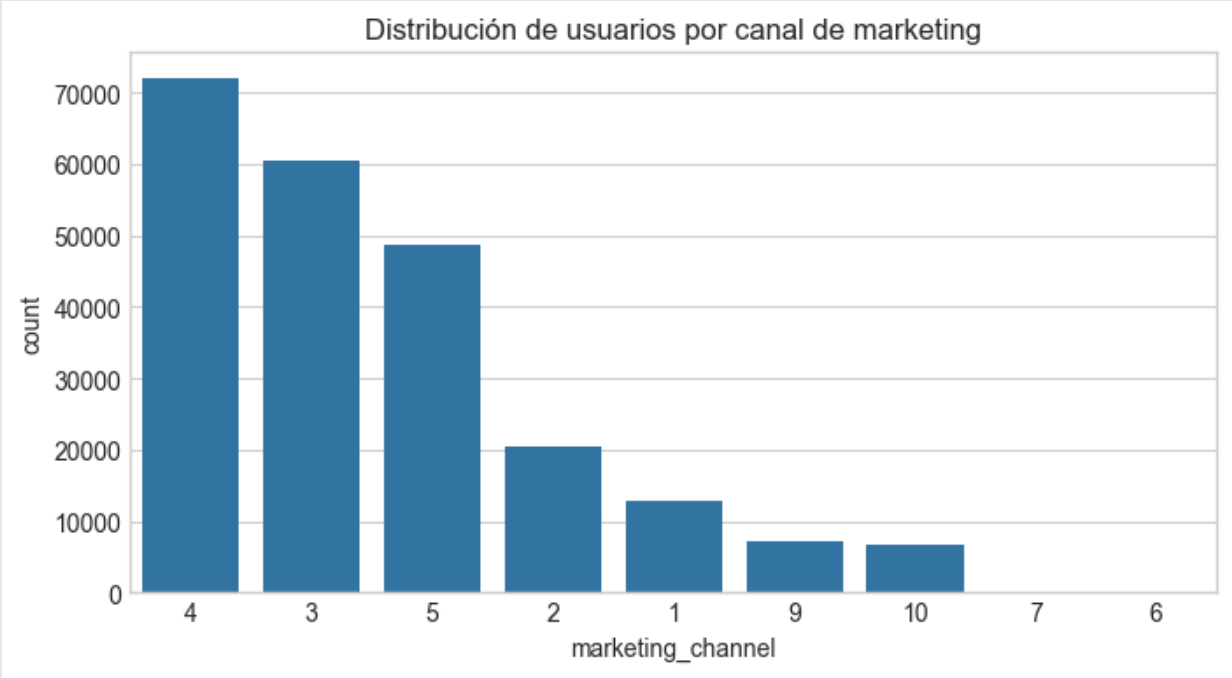




NUEVAS VARIABLES

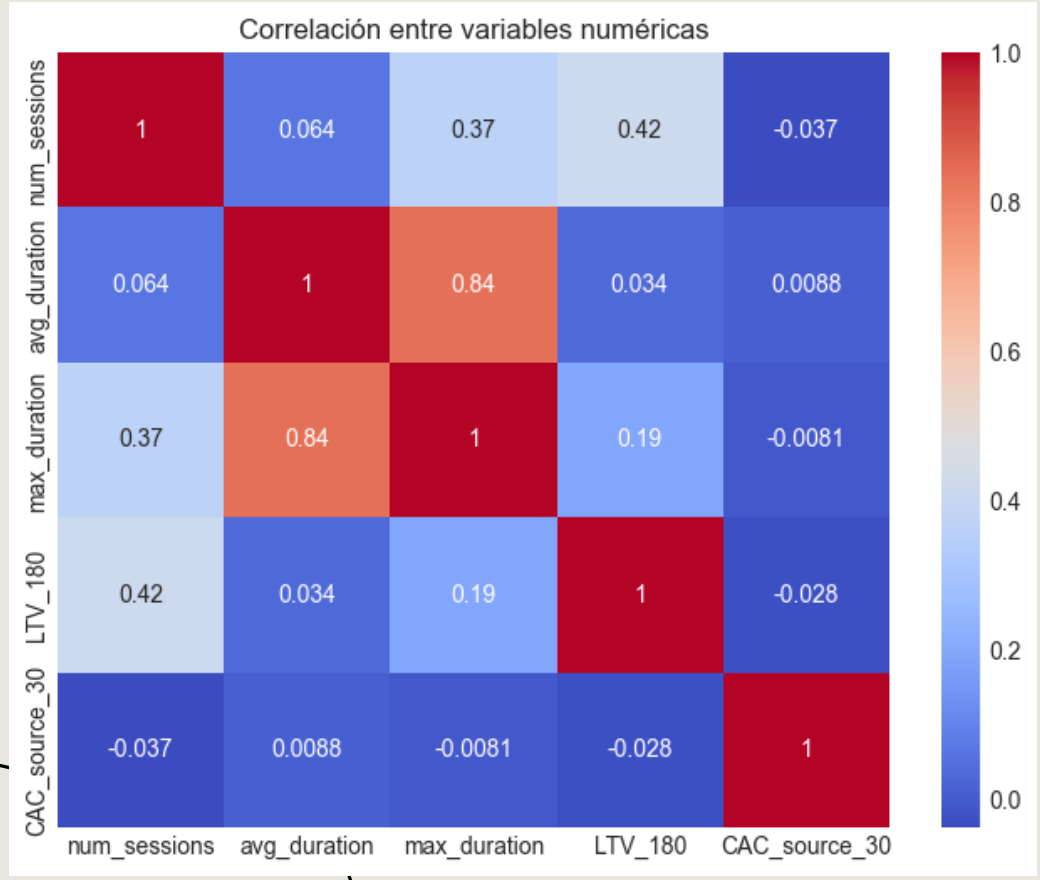
Variables generadas: sesiones por usuario, canales de adquisición codificados, tiempos de conversión y frecuencia de retorno, todas fundamentales para estimar el LTV.

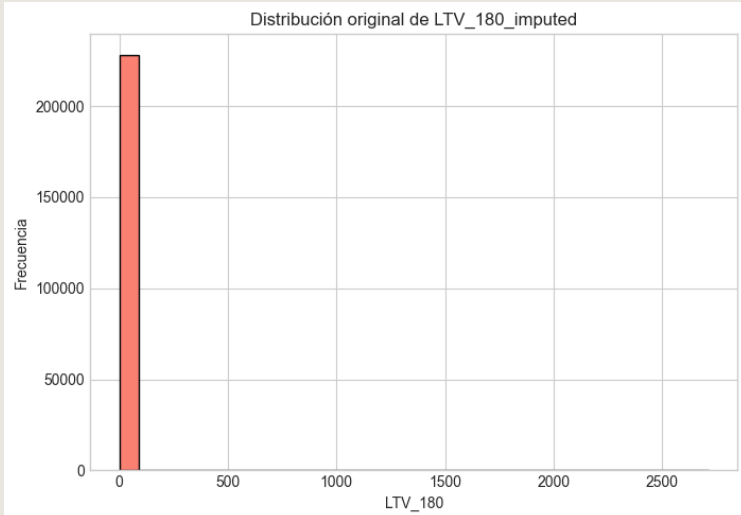


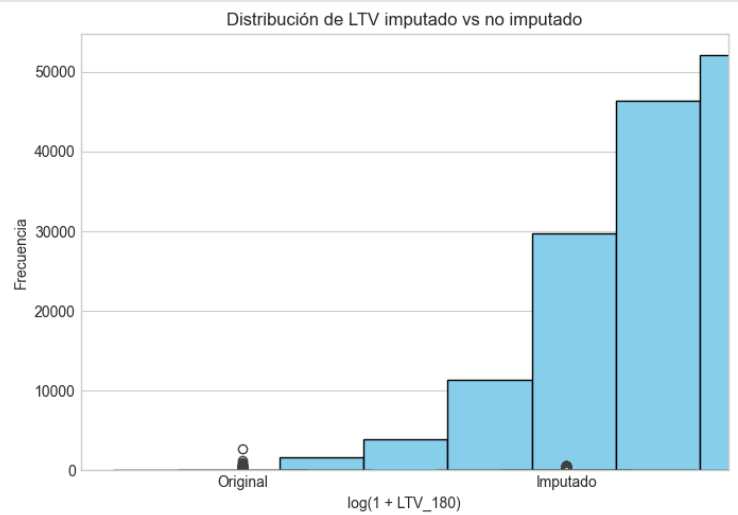


SELECCIONADO DE VARIABLES

Mediante el análisis de correlaciones y lógica de negocio, se imputó con transformación logarítmica.







RESUMEN

Mediante el análisis de correlaciones y lógica de negocio, se imputó con transformación logarítmica.

Tipo de acción	Detalle
Transformaciones	log() a LTV para reducir asimetría.
Imputación inteligente	Medianas para Valores Faltantes
Variables derivadas	Número de sesiones, frecuencia de retorno, duración promedio, CAC.
Codificación categórica	Canales de marketing codificados y agrupados por performance.
Segmentación temporal	Cohortes y fechas relativas para captar evolución del comportamiento.
Filtrado y selección	Se eliminaron variables redundantes y se conservaron las más predictivas.

MODELADO PREDICTIVO

- **Entrenamiento y Resultados**
  - **Lifetime Value en 180 días.**
  - **Costo de Adquisición en 30 días según el canal.**
  - **Modelos Base**
    - **LightGBM (balance velocidad-precisión.**
    - **Regresiones (Rigde, Lasso, SGD), como Baseline.**
    - **Seleccionado de variables: numéricas y fechas convertidas a tipo timestamp**

VALORACIONES

Evaluación – LTV_180

Métrica	Entrenamiento	Validación	Test
MAE	0.0358	0.0060	0.0117
RMSE	3.92	0.1991	0.6691
MAPE	≈0%	≈0%	≈0%

Evaluación – CAC_source_30

Métrica	Entrenamiento	Validación	Test
MAE	38.17	97.86	81.88
RMSE	77.66	117.38	115.58
MAPE	38%	91%	92%

3. VALIDACIÓN SELECCIÓN.
4. EXPLICABILIDAD/DIAGNÓSTICO
5. ESTRATEGIA DE MARKETING BASADA DE SIMULACIÓN

VALIDACIÓN/SELECCIÓN

- Modelos Entrenados: Lineal, Ridge, Lasso, SGD y LightGBM.
- Evaluación con RMSE.
- Modelos elegidos: Ridge para LTV y LightGBM para CAV.

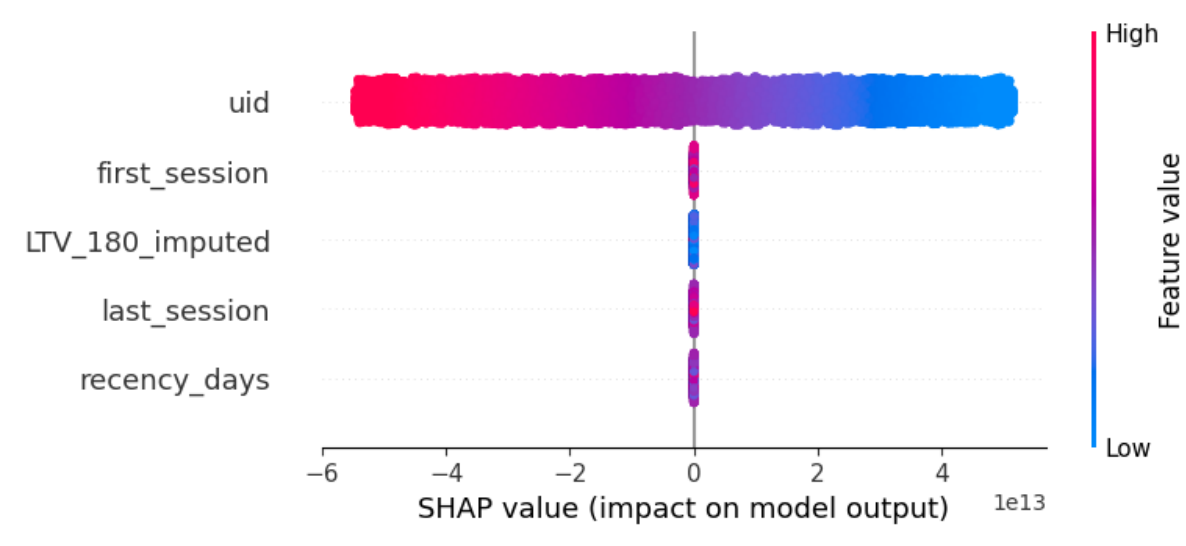
- TimeSeriesSplit no fue aplicado.
- GridSearchCV no se usó.

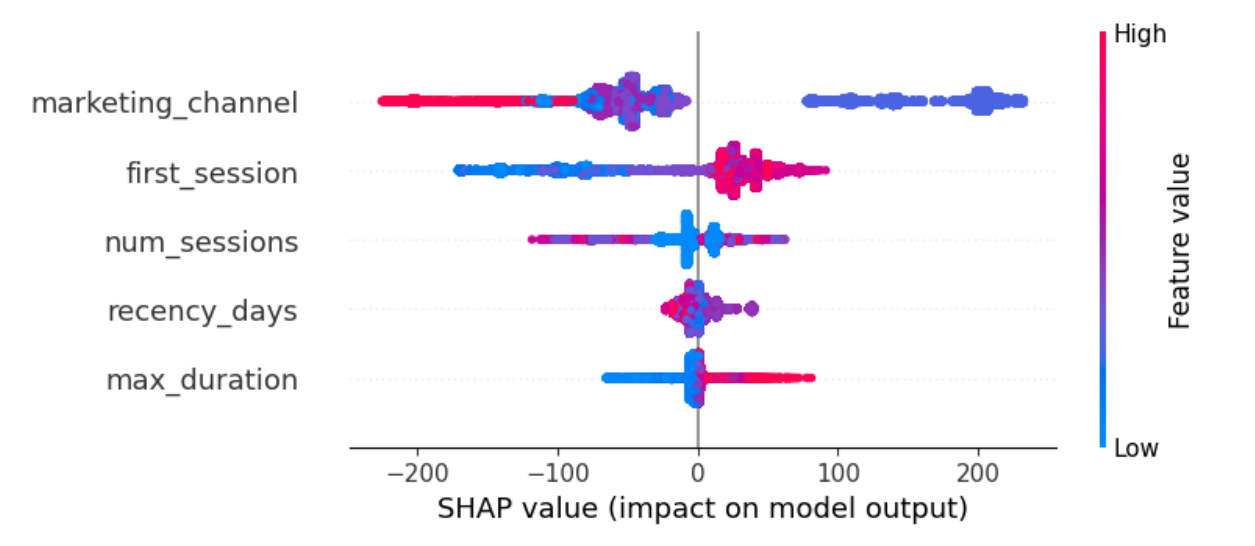
El trabajo individual y el tiempo limitado priorizaron una comparación directa basada en desempeño.

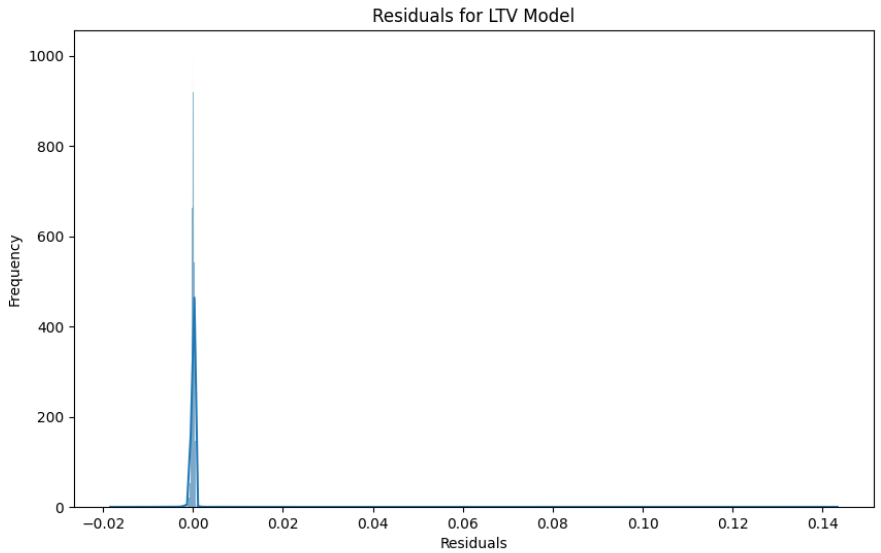
Se cumplió el objetivo de seleccionar los mejores modelos, aunque sin validación formal.

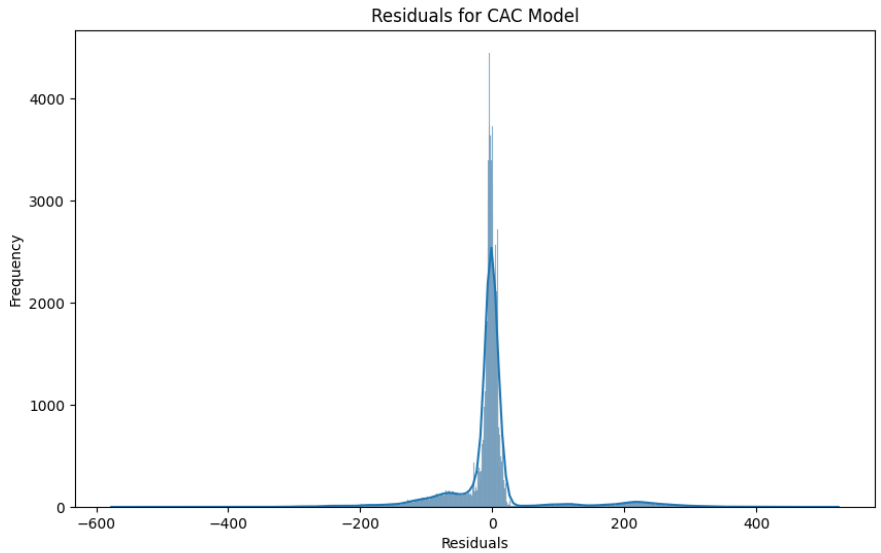
EXPLICABILIDAD DIAGNÓSTICO

- Uso de SHAP para interpretación del modelo de LTV.
- Análisis de residuos para tanto LTV como CAC.
- Permutation Importance y PDP no se aplicaron
- Sin diagnóstico de fallos por segmento (canales y clientes)









SIMULACIÓN ESTRATEGIA DE MARKETING

- Se implementó simulate_romi_from_real_values() para:
  - Calcular ROMI actual.
  - Simular +10% y redistribución proporcional.
  - Generar gráfico comparativo y recomendar.

La función clave fue implementada y se logró una simulación válida, aunque la simulación completa quedó pendiente por un error técnico.

- simulate_marketing_budget() no se ejecutó (faltó parámetro).

Categoría	Contenido
Logros Clave	- Se cumplió el objetivo central: estimar LTV y CAC con modelos predictivos. - Ridge (LTV) y LightGBM (CAC) ofrecieron buen desempeño. - Se logró interpretabilidad básica con gráficos SHAP. - Simulaciones ROMI reales generaron recomendaciones útiles.
Limitaciones	- No se aplicó validación cruzada formal ni ajuste de hiperparámetros. - No se finalizó la simulación basada en predicciones del modelo. - Faltó análisis de errores por segmento.
Recomendaciones	- Implementar TimeSeriesSplit y GridSearchCV en futuros ciclos. - Finalizar simulate_marketing_budget() con datos modelo. - Analizar desempeño por canal, tipo de cliente u otros segmentos.- Escalar metodología a otros productos o campañas.