

# Contexto del caso

**Escenario de negocio.** Un equipo de growth quiere estimar cuántas **conversiones** (compras/registros/leads) obtendrá una campaña digital a partir de tres factores operativos: **inversión** publicitaria, **clics** obtenidos y **segmentación** aplicada. El objetivo es **predecir conversiones** para planificar presupuestos semanales y elegir la segmentación más eficaz bajo restricciones de costo.

**Granularidad supuesta.** Cada fila representa el desempeño de una **campaña en un periodo** (p. ej., campaña-día o campaña-semana) con su inversión, clics, segmentación aplicada y conversiones logradas. Si tu granularidad real difiere (anuncio, ad-set, canal), solo hay que ajustar la interpretación; el modelo y el diccionario permanecen válidos.

**Objetivo del modelo.** Predecir **conversiones** a partir de **inversion**, **clics** y **segmentacion** usando una **regresión lineal múltiple** como línea base explicable. Con ello, el equipo puede:

- Estimar conversiones esperadas al variar el presupuesto.
- Comparar escenarios de segmentación manteniendo inversión/clics constantes.
- Detectar rendimientos decrecientes cuando aumentar inversión ya no sube conversiones al mismo ritmo.

**Preguntas que el caso permite responder.**

- ¿Cuántas conversiones espero si invierto X y obtengo ~Y clics con la segmentación Z?
- Dado un objetivo de conversiones, ¿qué combinación inversión-clics es razonable?
- A igual inversión, ¿qué segmentación históricamente rinde mejor?

**Decisiones operativas habilitadas.**

- Planificación semanal de presupuesto por segmento.
- Priorización de segmentos con mayor “pendiente” (impacto marginal) sobre conversiones.
- Señales tempranas de saturación (si los coeficientes y residuos muestran límites).

**Suposiciones explícitas (para alinear expectativas).**

- La relación entre variables y **conversiones** es **aproximadamente lineal** en el rango observado.
- **clicks** ya incorpora la mezcla de canales/creativos de ese periodo; si no, conviene añadirlos como variables futuras.
- **segmentacion** es una **categoría** (segmento demográfico/comportamental/look-alike, etc.), no un texto libre ruidoso.

## Diccionario de datos

Fuente: archivo proporcionado por ti con columnas **inversion**, **clicks**, **segmentacion**, **conversiones**. Tipos sugeridos y notas preparadas para modelado.

Campo	Tipo lógico	Unidad / dominio	Descripción funcional	Ejemplos válidos	Reglas de calidad / Notas para ML
<b>inversion</b>	Numérico continuo (float)	Moneda (p. ej., USD)	Monto invertido en el periodo y campaña correspondiente. Incluye pauta en plataformas digitales.	1 500; 25 000; 75 000	Debe ser $\geq 0$ . Conviene revisar outliers extremos. Estandarizar para la regresión. Si hay diferentes monedas, unificarlas previamente.
<b>clicks</b>	Numérico entero/continuo (int/float)	Cuento de clics	Número de clics registrados en el periodo.	430; 510; 1 250	Debe ser $\geq 0$ . Captar outliers percentil 1–99 si hay valores atípicos que distorsionen. Si existen métricas derivadas (CTR, CPC), se pueden añadir en iteraciones futuras.

<b>segmentacion</b>	Categorico (string → dummies)	Conjunto finito de etiquetas	Segmento de audiencia aplicado (demográfico, interés, look-alike, retargeting, etc.).	“Prospectos_AltaIntencion”, “Lookalike_1%”, “Intereses_Tecno”	Normalizar etiquetas (trim, casefold). One-Hot Encoding con <code>handle_unknown="ignore"</code> para permitir nuevos segmentos en predicción. Vigilar categorías con baja frecuencia.
<b>conversiones</b>	Numérico entero (int)	Conteo	Resultado objetivo del periodo: leads, registros o compras, según tu definición.	200; 1 350; 3 900	Debe ser $\geq 0$ . Sin decimales. Es la <b>variable objetivo</b> del modelo. Asegurar consistencia de la definición de “conversión” en todas las filas.

#### Derivadas recomendadas (opcional, futura iteración)

- `CPC = inversion / max(clics, 1)` para capturar eficiencia de clic.
- `log_inversion`, `log_clics` para suavizar asimetrías si hay colas largas.
- Indicadores por canal/creativo si están disponibles en otra tabla.

#### Preprocesamiento mínimo sugerido

- Conversión de tipos: `inversion`, `clics`, `conversiones` → numéricos; `segmentacion` → categórico.
- Tratamiento de outliers suaves: capado percentilar en `clics` (y, si procede, en `inversion`).
- División **train/test** estratificada por `segmentacion` si la distribución entre segmentos es muy desigual.

#### Validación y uso

- Métricas: RMSE, MAE,  $R^2$  en train y test.
- Interpretación: revisar coeficientes de la regresión (tras estandarizar numéricas y dummificar **segmentacion**) para entender impacto marginal.
- Predicción operacional: preparar una tabla “what-if” donde el usuario ingrese **inversion**, **clics**, **segmentacion** y el modelo devuelva **conversiones** estimadas.