**CMP 5002 Data Mining**

**Proyecto Final**

**Integrantes:** Melanie Álvarez, Marie Cucalón, Fabián De La Cruz, Roberth Lara, Ma. Emilia Rivadeneira

**Fecha:** 07-05-2025

**NRC:** 2106

**Link repositorio:** https://github.com/maemiliarv/DataMiningProyectoFinal/tree/main

**Link presentación:**
https://www.canva.com/design/DAGmtYOlWjw/4G93mFqBu4pi8DqqgBwGBA/edit?utm_content=DAGmtYOlWjw&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

## Resumen ejecutivo

Showz es una empresa dedicada a la venta de entradas para eventos, que recientemente ha fortalecido su área de Analítica Avanzada con el objetivo de tomar decisiones más informadas y estratégicas basadas en datos. En la dinámica industria de los eventos, vender entradas va más allá de una simple transacción: es el arte de generar expectativa e inspirar a las personas a vivir experiencias inolvidables. La venta de entradas no solo representa una fuente clave de ingresos para todo tipo de eventos, sino que también actúa como un indicador directo del atractivo y el éxito de cada uno. En un mercado caracterizado por la transformación digital y por clientes cada vez más exigentes, una gestión eficaz del ticketing se ha convertido en el núcleo de la planificación exitosa de eventos.

A partir de análisis previos sobre el comportamiento de los usuarios, sus patrones de compra y la inversión en marketing, este trabajo tiene como objetivo construir modelos predictivos que permitan:

- Estimar el Lifetime Value (LTV) esperado de cada cliente en un horizonte de 6 a 12 meses.
- Predecir el Customer Acquisition Cost (CAC) por fuente y cohorte para el siguiente trimestre.
- Identificar los factores clave que impulsan el LTV y CAC, utilizando técnicas de explicabilidad.
- Simular escenarios futuros para optimizar la asignación del presupuesto de marketing, maximizando el ROMI (Return on Marketing Investment).

**Metodología CRISP-DM**

***Entendimiento del negocio***

En el competitivo mundo de la venta de entradas para eventos, comprender el comportamiento del cliente no es solo una ventaja, es una necesidad estratégica. Showz opera en un entorno donde el éxito de cada campaña publicitaria se mide no solo en conversiones inmediatas, sino en la capacidad de generar valor sostenido a través del tiempo.

En este contexto, el crecimiento sostenible de la empresa depende de conocer no solo cuántos usuarios compran, sino cuánto valor generan a lo largo del tiempo (LTV) y cuánto cuesta adquirirlos (CAC). Las fuentes de adquisición, como campañas digitales o canales orgánicos, presentan niveles variables de eficiencia y fidelización.

Frente a estos desafíos, se propone una solución basada en inteligencia artificial capaz de anticipar el retorno esperado por cada cliente y permitir una asignación proactiva y optimizada del presupuesto de marketing, maximizando así el ROMI. Esto permitirá a Showz priorizar estratégicamente aquellas fuentes y cohortes con mayor potencial de rentabilidad, al tiempo que reduce inversiones poco efectivas y mejora la fidelización a largo plazo.

***Entendimiento de los datos***

El análisis se basó en tres fuentes principales de datos que capturan diferentes dimensiones del comportamiento del cliente y la inversión en marketing:

- **visits_log_us.csv**: contiene más de 350,000 registros de sesiones de usuarios. Se identificó que los dispositivos tipo *desktop* son los más utilizados, seguidos por *touch*. También se observaron 10 fuentes de adquisición con distribución heterogénea, destacando algunas con mayor volumen de sesiones. A partir de las marcas de tiempo, se calculó la duración de las sesiones, revelando 24,658 valores atípicos que fueron conservados por su potencial valor informativo (como indicadores de muy alto o bajo compromiso, campañas especiales o errores de navegación).

- **orders_log_us.csv**: registra los detalles de las compras realizadas. Se evidenció una clara estacionalidad positiva en los pedidos hacia finales del año, especialmente entre octubre y diciembre. El análisis de la variable de ingresos mostró una distribución asimétrica con 3,990 outliers que representan a usuarios de alto valor, los cuales fueron retenidos para reflejar el comportamiento real del negocio y enriquecer el modelado del Lifetime Value.

- **costs_us.csv**: contiene los gastos en marketing desagregados por fuente y fecha. Se observó una tendencia creciente de inversión hacia fin de año, coincidiendo con el patrón estacional de las compras. La distribución de costos mostró una fuerte concentración en valores bajos, con 189 outliers correspondientes a picos de inversión

que podrían reflejar campañas agresivas o promociones intensivas. Estos registros se los conservó por su relevancia potencial en el cálculo del CAC.

Una ventaja clave fue la alta calidad de los datos desde el inicio: no se encontraron valores nulos ni registros duplicados en ninguno de los archivos. Esto permitió enfocar los esfuerzos en transformar las variables temporales al formato datetime y calcular métricas derivadas como duración de sesión, mes de compra o gasto mensual, sin necesidad de aplicar estrategias de limpieza o imputación. Los datasets procesados se los almacenó en la carpeta data/processed/, listos para su uso en la ingeniería de características y modelado predictivo.

***Preparación de datos***

La etapa de ingeniería de características se diseñó con el objetivo de enriquecer los datos brutos provenientes de las tablas visits, orders y costs, generando un conjunto de variables explicativas que mejoren la capacidad predictiva de los modelos de LTV (Lifetime Value) y CAC (Customer Acquisition Cost). Este proceso se enfocó en tres grandes dimensiones: comportamiento del usuario, temporalidad y marketing, además de la construcción de las etiquetas objetivo.

1. Variables de Comportamiento

Se calcularon indicadores clave del comportamiento de compra de cada usuario:

- **n_compras:** número total de compras por usuario.
- **revenue_total:** suma total del ingreso generado.
- **aov:** valor promedio por orden (Average Order Value).
- **dias_entre_compras:** diferencia en días entre la primera y última compra.
- **is_retained:** indicador binario de retención si el usuario realizó una segunda compra al menos 30 días después de la primera.
- **gasto_mensual_promedio** y **frecuencia_compras**: promedios mensuales y de frecuencia de compra.
- **dias_primera_sesion_a_primera_compra:** tiempo (en días) desde la primera visita hasta la primera conversión.
- **trend_slope:** pendiente de la regresión lineal del gasto a lo largo del tiempo, como aproximación de la tendencia de gasto del usuario.
- **has_slope:** indicador binario que señala si el usuario tiene una tendencia significativa.

Estas variables permiten capturar patrones de lealtad, valor e intensidad de compra, fundamentales para predecir el LTV.

2. Variables Temporales

Para modelar efectos estacionales y contextuales del comportamiento, se extrajeron:

- **mes_primera_sesion** y **dia_semana_primera_sesion:** mes y día de la semana de la primera sesión.

- **es_fin_de_semana_primera_sesion:** indicador si la primera visita fue en fin de semana.
- **estacion_primera_sesion:** estación del año (codificada de 0 a 3).
- **dias_activo:** duración total de la actividad del usuario (de primera a última sesión).
- **cohort_year y cohort_month:** cohorte mensual de adquisición.

Estas variables ayudan a capturar efectos de tiempo que puedan influir en el valor o costo de adquisición del cliente.

3. Variables de Marketing

Se generaron variables que reflejan la interacción inicial del usuario con los canales de marketing:

- **primer_source** y **primer_dispositivo:** canal de adquisición y tipo de dispositivo de la primera sesión (codificado como 0=desktop, 1=touch).
- **n_sesiones:** número total de sesiones por usuario como proxy de engagement.
- **n_dispositivos_distintos:** diversidad de dispositivos utilizados por el usuario (omnicanalidad).
- **duracion_promedio_sesion:** duración media de cada sesión en segundos.

Estas variables fueron claves para modelar la conversión de marketing en valor (ROMI) y los costos asociados.

4. Construcción de Etiquetas (Targets)

Se definieron dos variables objetivo:

- **LTV_180:** suma del ingreso generado por un usuario en los 180 días posteriores a su primera sesión.
- **CAC_source_30:** costo promedio de adquisición por fuente, calculado como el gasto en marketing en los 30 días posteriores a la primera conversión dividido para el número de usuarios adquiridos por dicha fuente.

5. Dataset Final

Todas las variables generadas se integraron en un solo dataset (modeling_dataset.csv) que sirvió como entrada para los modelos predictivos. Este dataset cuenta con:

- 29 variables explicativas numéricas o categóricas.
- 2 etiquetas objetivo (LTV_180 y CAC_source_30).
- Unidades de análisis a nivel de usuario (uid).

Este conjunto de datos permite capturar múltiples dimensiones del comportamiento y adquisición del usuario, asegurando que los modelos predictivos sean informativos, robustos y explicables.

***Modelado***

Para esta sección decidimos hacer una función para cada modelo. También decidimos utilizar una función de entrenamiento y evaluación global. Esta función implementa GridSearchCV y TimeSeriesSplit. También incluye un segundo entrenamiento con el dataset de evaluación antes del test. Esta función de entrenamiento y evaluación nos devuelve el mejor modelo con un diccionario de las métricas MAE, RMSE Y MAPE. Este es el caso para todos los modelos que no eran ensambladores. Para los modelos ensambladores se hizo una función de entrenamiento y evaluación específica que sigue la misma idea. Esta función entrena cada modelo dentro del ensamblador y devuelve lo mismo que la función global que utilizan el resto de modelos.

Hay una función que divide el dataset en la siguiente manera: train 2017, validation 1er trimestre 2018, test 2do trimestre 2018. Una vez dividido hace drop a la columna de las fechas y procede a dividir el dataset en train, validation y test.

Hay una 'función maestra' que como parámetros tiene solo el df, la columna target y la columna que contiene los valores datetime. Aquí se ejecutan todas las funciones. Primero dividiendo el dataset, y de ahí yendo en el siguiente orden: Regresiones (Lineal, Estocástica y Ridge), Modelos avanzados (Random Forest y Gradient Boosting (XGBoost, LightGBM y CatBoost) y Ensamblador (stacking y blending).

***Evaluación***

La evaluación del desempeño de los modelos se realizó con base en tres métricas estándares de regresión: MAE, RMSE y MAPE, tanto en validación como en test. Estas métricas permiten cuantificar qué tan precisas y útiles son las predicciones para su aplicación práctica dentro de la empresa Showz.

1. Significado de las Métricas

- MAE (Mean Absolute Error): representa el error promedio absoluto entre las predicciones y los valores reales, expresado en las mismas unidades que el target. Por ejemplo: un MAE de 1.36 en LTV_180 indica que, en promedio, el modelo se equivoca por $1.36 en la estimación del valor futuro de cada cliente.
- RMSE (Root Mean Squared Error): penaliza errores grandes con más fuerza que el MAE. Aunque menos interpretable directamente en negocio, es útil para detectar modelos que cometen errores extremos con mayor frecuencia.
- MAPE (Mean Absolute Percentage Error): expresa el error en términos relativos (%) y permite comparar errores en distintos rangos de valores. Un MAPE de 2.17% implica que, en promedio, la predicción del LTV está muy cerca del valor real (97.83% de precisión). En cambio, un MAPE de 45% indica que los errores de predicción son muy relevantes en proporción al valor real.

## 2. Umbrales de Evaluación

Para determinar si un modelo es "bueno", se definieron umbrales de referencia derivados de los datos reales:

- LTV_180: se fijaron como aceptables MAE ≤ 20% de la media histórica ($1.28), RMSE ≤ 25% de la media ($1.60) y MAPE < 30%.
- CAC_source_30: los valores ideales serían MAE ≤ $0.07 y MAPE < 30%, debido a que los CAC suelen ser bajos y cualquier error relativo impacta fuertemente en decisiones de inversión.

Estos umbrales permiten comparar modelos no solo entre sí, sino también con el desempeño esperado a nivel de negocio.

## 3. Análisis de Resultados para LTV_180

- El Random Forest logró un MAE de $1.36 y un MAPE de 2.17%, lo que significa que el modelo predice el ingreso futuro de cada cliente con precisión casi exacta en términos porcentuales, y con un error absoluto promedio tolerable y útil para la toma de decisiones.
- Otros modelos como XGBoost y Blending mostraron un rendimiento competitivo, aunque con ligeras desviaciones en test.
- En comparación, el baseline por media tuvo un MAE de $5.56 y un MAPE superior al 430%, lo que significa que usar el promedio histórico para estimar el valor de un nuevo cliente es completamente ineficaz para toma de decisiones individualizadas.

Por lo tanto, si la empresa Showz usa este modelo para predecir que un cliente generará $28 de ingreso, el modelo acertará en promedio con un margen de error de ±$1.36. Esta precisión permite asignar presupuestos de retención o adquisición alineados con el valor real esperado del cliente.

## 4. Análisis de Resultados para CAC_source_30

- Todos los modelos regresivos lograron MAE de 0.13 y MAPE entre 42% y 45%, muy por encima del umbral ideal.
- Incluso el baseline por mediana logró un mejor MAPE (28.29%), lo cual indica que los modelos actuales no son significativamente más útiles que una estimación estadística simple.
- Esto puede explicarse por:
    - La agregación del CAC a nivel fuente (no a nivel usuario).
    - La escasez de variables específicas de marketing (campañas, presupuesto diario, impresiones, etc.).
    - La presencia de valores pequeños en el target (ej. CAC de $0.10), que generan grandes errores relativos con ligeras desviaciones absolutas.

Es decir, aunque el error promedio es bajo ($0.13 por usuario adquirido), cuando el CAC real es pequeño (por ejemplo $0.20), este error representa una desviación del 65%, lo cual compromete la capacidad de usar esta predicción como base sólida para redistribuir presupuesto de adquisición entre canales.

5. Consideraciones Finales

- Los modelos para LTV superan ampliamente los criterios de precisión esperados y aportan un valor directo al negocio. La empresa puede identificar clientes de alto valor desde su primera sesión y optimizar sus estrategias de fidelización, retención y segmentación.
- Los modelos para CAC requieren mejoras, ya que sus errores relativos aún son altos y no permiten confiar plenamente en ellos para tomar decisiones presupuestarias automatizadas.
- MAPE debe considerarse siempre junto al MAE: un MAPE bajo indica precisión general, pero un MAE bajo garantiza utilidad financiera real para decisiones a nivel de unidad (cliente o fuente).

***Explicabilidad y Diagnóstico***

En esta etapa se profundiza en la interpretación de los dos modelos seleccionados: XGBoost para LTV_180 y regresión Ridge para CAC_source_30, con el propósito de entender qué variables impulsan sus predicciones y de identificar patrones de error que permitan mejorar su desempeño.

**Importancia de variables**

Para cada modelo se calculó la importancia de las características mediante feature importances (XGBoost) y análisis de coeficientes absolutos (Ridge):

- **LTV_180 (XGBoost)**

    1. *Average Order Value (AOV)*: el valor promedio por compra resulta determinante para estimar el valor futuro del cliente.
    2. *Revenue Total*: muestra la capacidad histórica de generación de ingresos.
    3. *Gasto Mensual Promedio*: refleja consistencia en el nivel de gasto.
    4. *Días desde primera sesión a primera compra*: captura la rapidez de conversión inicial.
    5. *Trend Slope*: pendiente de la tendencia de gasto a lo largo del tiempo.

Estas cinco variables concentran la mayor contribución a la reducción del error en la predicción de LTV, lo cual confirma que tanto la magnitud como la frecuencia de compra son factores clave.

- **CAC_source_30 (Ridge)**

El análisis de coeficientes del modelo Ridge para CAC_source_30 revela los siguientes impactos:

- Coeficientes positivos (a mayor valor de la variable, mayor CAC_predicho):

1. *LTV_180*: el coeficiente más alto, indica que clientes con mayor valor de vida requieren mayor inversión para ser adquiridos.
2. *Días entre compras*: su incremento se asocia con un CAC más elevado, sugiriendo que la baja frecuencia de compra exige mayores esfuerzos de adquisición.
3. *Frecuencia de compras*: a medida que aumenta la frecuencia, el modelo predice un CAC más alto, posiblemente por campañas específicas para compradores repetitivos.
4. *Número de dispositivos distintos*: mayor omnicanalidad implica mayor complejidad y coste en la adquisición.
5. *Cohort_year*: usuarios de cohortes más recientes o antiguas muestran diferencias en coste de adquisición.

- Coeficientes negativos (a mayor valor de la variable, menor CAC_predicho):

1. *Revenue Total*: clientes que han generado más ingresos históricamente costaron menos en marketing, reflejando eficiencia de adquisición.
2. *Días activos*: usuarios con mayor permanencia orgánica fueron más baratos de adquirir.
3. *Primer dispositivo*: ciertos canales de acceso inicial (p. ej. desktop vs. touch) se asocian a menor coste.
4. *Número de sesiones*: mayor engagement previo a la compra se traduce en menor CAC.
5. *Trend Slope*: usuarios cuya tendencia de gasto crece con el tiempo resultan más eficientes de captar.

Estas diez variables explican la mayor parte de la variación en la predicción de CAC_source_30, confirmando que tanto indicadores de valor del cliente (como LTV_180 y revenue_total) como métricas de engagement y temporalidad (frecuencia, sesiones, cohortes) influyen de manera opuesta en el costo de adquisición.

**Análisis de dependencia parcial (PDP)**

Se generaron gráficos de Partial Dependence para las cinco variables más importantes del modelo XGBoost (LTV_180) y del modelo Ridge (CAC_source_30), con los siguientes hallazgos:

- En **LTV_180**:

    - *revenue_total*: la curva de dependencia parcial es creciente y casi lineal, lo que confirma que a mayor ingreso histórico de un usuario corresponde un LTV_180 más alto.
    - *aov (Average Order Value)*: muestra también una tendencia ascendente, reforzando que los pedidos de mayor valor contribuyen a un mayor valor de vida del cliente.
    - *gasto_mensual_promedio*: la curva es prácticamente plana, lo que indica que, más allá de un cierto umbral, el gasto promedio mensual aporta muy poca información extra para predecir el LTV_180.
    - *mes_primera_sesion:* no se aprecia variación significativa a lo largo de los diferentes meses, sugiriendo ausencia de un efecto estacional marcado en el valor a seis meses.
    - *días_primera_sesion_a_primera_compra:* la línea resultante es estable y casi horizontal, lo que implica que el tiempo hasta la primera conversión no discrimina de manera relevante el LTV dentro del horizonte considerado.

- En **CAC_source_30**:

    - *LTV_180*: presenta una ligera pendiente ascendente, indicando que un LTV_180 mayor se asocia con un CAC_source_30 ligeramente más alto, posiblemente porque clientes de alto valor requirieron mayor inversión inicial.
    - *días_entre_compras:* curva ascendente clara, lo que sugiere que intervalos más largos entre compras incrementan el costo de adquisición medio.
    - *frecuencia_compras*: muestra una tendencia creciente, señalando que los compradores más frecuentes también implican un CAC más elevado, quizá por campañas continuas de retención.
    - *n_dispositivos_distintos*: la pendiente es marcadamente ascendente, reflejando que la omnicanalidad (usuarios que usan varios dispositivos) eleva el coste de captación.
    - *revenue_total:* curva descendente, lo que indica que a mayor ingreso total, menor CAC_source_30 predicho, reflejando eficiencia en la adquisición de usuarios rentables.
    - *días_activo*: muestra un leve descenso, sugiriendo que los usuarios con mayor tiempo activo en la plataforma costaron menos en marketing.

Estos PDP confirman y matizan la interpretación de las importancias y coeficientes: por un lado, validan las relaciones lineales principales (ingresos y valor de pedido para LTV; engagement y valor histórico para CAC) y, por otro, revelan cuándo ciertas variables (gasto mensual o mes de adquisición) tienen un efecto marginal o nulo

**Análisis de errores sistemáticos ¿Dónde falla el modelo?**

**LTV_180 (XGBoost)**

Agrupamos las predicciones y los valores reales por número de compras (n_compras) y calculamos el MAE medio en cada grupo; además segmentamos la variable revenue_total en cuartiles y obtuvimos el MAE en cada uno.

- Resultados:
    - Clientes con entre 1 y 4 compras mostraron un MAE inferior a 5, lo que indica alta precisión en usuarios de baja frecuencia.
    - Se detectó un pico extremo en n_compras = 19 con MAE superior a 35, y picos aislados en 12, 40 y 117 compras, señalando subgrupos atípicos donde el modelo comete errores muy elevados.
    - Al dividir revenue_total en cuartiles, el cuartil más alto (ingresos > 4.89) presentó un MAE casi el doble que el resto, lo que revela mayor incertidumbre al predecir clientes de muy alto gasto.
- Interpretación: el error crece para comportamientos extremos no representados adecuadamente en el entrenamiento, y el modelo tiende a fallar con usuarios de alta frecuencia de compra o de ingresos muy elevados.

**CAC_source_30 (Regresión Ridge)**

Graficamos los residuos (diferencia entre predicción y valor real) frente a las predicciones para identificar patrones de error, y construimos el histograma de esos residuos.

- Resultados:
    - En el scatter residuos vs. predicción se observan bandas horizontales, lo que indica que el modelo produce valores discretos repetidos en lugar de una salida continua.
    - Existe una ligera asimetría hacia errores positivos, lo que significa que el modelo tiende a sobreestimar el CAC en varios rangos.
    - El histograma de residuos resultó bimodal, con dos picos alrededor de –0.22 (subestimaciones) y 0.09 (sobreestimaciones), evidenciando dos clusters de observaciones con errores sistemáticos.

- Interpretación: la regularización fuerte o la escala de la variable objetivo puede estar originando predicciones agrupadas; el modelo muestra un sesgo al alza en ciertos segmentos y subestima en otros, lo que sugiere la existencia de grupos de usuarios con patrones de coste muy distintos.

Estos análisis permiten identificar claramente los segmentos donde cada modelo falla, orientar tratamiento de outliers y plantear soluciones de modelado segmentado o ajustes de hiperparámetros para mejorar la robustez de las predicciones.

***Despliegue (simulación)***

Para definir la mejor estrategia de marketing en un entorno simulado, se tomó como punto de inicio los mejores modelos, tanto para LTV como CAC, resultantes del proceso de entrenamiento y validación previamente descrito. Se analizó la métrica MAE en validación y aquellos con el menor valor fueron los siguientes:

| **LTV** | | **CAC** | |
|---|---|---|---|
| *Blending* | | *Ridge* | |
| MAE | 0.64 | MAE | 0.13 |
| MAPE | 5.11% | MAPE | 42.43% |
| RMSE | 3.26 | RMSE | 0.15 |

MAE mide el error medio absoluto entre las predicciones y el valor real, un valor bajo asegura mayor ganancia, lo que se alinea con el objetivo de la estrategia de marketing a proponer. El modelo blending cumplió con los umbrales de decisión establecidos en la parte de entrenamiento, a excepción del RMSE que sobrepasa el límite con 1.66 puntos. Las métricas del modelo Ridge sobrepasaron el nivel de tolerancia elegido, un comportamiento que era de esperarse según su desempeño en el entrenamiento; no obstante, este es el modelo con menor MAE y resulta pertinente.

Se cargaron los modelos entrenados con Joblib y sus listas de características en formato de Json para el cálculo de valores de predicción para LTV_180 y CAC_30. Una vez obtenidos estos valores, se calcularon las siguientes métricas agrupadas según la columna ‘primer_source’ (fuente que llevó a los usuarios al primer contacto con el negocio):

- avg_LTV: media de los valores predichos de LTV_180 para todos los usuarios de ese canal.
- avg_CAC: media de los valores predichos de CAC_30 para todos los usuarios de ese canal.
- n_users: conteo de cuántos usuarios llegaron a través de ese canal.

Estas métricas permiten el cálculo de:

- ROMI

$$ROMI = \frac{avgLTV}{avgCAC}$$

- Presupuesto

$$budget = avgCAC \times n\,users$$

Estos indicadores agregados permiten comparar canales entre sí en términos de costo, ingreso y eficiencia y tomar decisiones informadas sobre dónde incrementar o redistribuir el presupuesto para maximizar los ingresos.

| | primer_source | avg_LTV | avg_CAC | n_users | ROMI | budget |
|---|---|---|---|---|---|---|
| 0 | 1 | 9.796521 | 0.330681 | 2899 | 29.625266 | 958.645055 |
| 1 | 2 | 13.000894 | 0.329137 | 3506 | 39.499962 | 1153.953870 |
| 2 | 3 | 5.984117 | 0.331510 | 10473 | 18.051077 | 3471.906927 |
| 3 | 4 | 5.898427 | 0.330774 | 10296 | 17.832208 | 3405.646750 |
| 4 | 5 | 7.917749 | 0.331241 | 6931 | 23.903262 | 2295.833821 |
| 5 | 7 | 2.150563 | 0.343806 | 1 | 6.255168 | 0.343806 |
| 6 | 9 | 5.306426 | 0.325173 | 1088 | 16.318796 | 353.787873 |
| 7 | 10 | 4.162009 | 0.331131 | 1329 | 12.569061 | 440.073437 |

Se establecieron dos escenarios de análisis. El primero propone incrementar un 10% el presupuesto asignado a una fuente X, el segundo redistribuye proporcionalmente este incremento entre todas las fuentes. Los valores de referencia para la simulación fueron el total de presupuesto, sumando los budget de todas las fuentes, y el ingreso total, multiplicando el gasto actual de cada fuente por su ROMI y sumándolo todo.

| source | base_rev | rev_+10%_solo | delta_+10% | rev_redistrib | delta_redistrib |
|---|---|---|---|---|---|
| 3 | 263567.87993 | 269835.045883 | 6267.165953 | 271142.951443 | 7575.071513 |
| 4 | 263567.87993 | 269640.900036 | 6073.020106 | 270998.383750 | 7430.503820 |
| 5 | 263567.87993 | 269055.671746 | 5487.791815 | 268576.973129 | 5009.093199 |
| 2 | 263567.87993 | 268125.993354 | 4558.113423 | 266085.598081 | 2517.718151 |
| 1 | 263567.87993 | 266407.891402 | 2840.011472 | 265659.469654 | 2091.589724 |
| 9 | 263567.87993 | 264145.219128 | 577.339198 | 264339.780934 | 771.901004 |
| 10 | 263567.87993 | 264121.010898 | 553.130968 | 264528.040392 | 960.160462 |
| 7 | 263567.87993 | 263568.094986 | 0.215056 | 263568.630052 | 0.750122 |

***Resultados***

De acuerdo a las tablas anteriores, la fuente 3 es la que más usuarios trae a Showz, por lo que si se invierte más capital en una fuente sería en este. En el escenario 1, invertir 10% extra en

la fuente 1 le traería a la empresa una ganancia de $6267.17. En el escenario 2, redistribuir este 10% según el peso de cada budget en el total, hace que Showz gane más en casi todos los canales (a excepción de 1, 2 y 5), por ejemplo, en el canal 3, gana $1307.90 más que en el escenario 1.

***Recomendaciones***

Posterior a la simulación, se puede concluir que el escenario óptimo es quitar un 10% del presupuesto de la Fuente 3 y redistribuir ese extra proporcionalmente entre todas las fuentes, esto produce la mayor ganancia global ($7575.07). La distribución de presupuesto se vería de la siguiente manera:

Fuente óptima para redistribuir: 3

| primer_source | presupuesto_original | extra_redistrib | presupuesto_final |
|---|---|---|---|
| 1 | 958.645055 | 27.551934 | 986.196989 |
| 2 | 1153.953870 | 33.165206 | 1187.119076 |
| 3 | 3471.906927 | 99.784326 | 3571.691253 |
| 4 | 3405.646750 | 97.879976 | 3503.526726 |
| 5 | 2295.833821 | 65.983402 | 2361.817223 |
| 7 | 0.343806 | 0.009881 | 0.353687 |
| 9 | 353.787873 | 10.168039 | 363.955912 |
| 10 | 440.073437 | 12.647929 | 452.721366 |

Donde el ingreso total pasa de $263,567 a $271,143, un aumento de $7575.

Para asegurar que esta estrategia de marketing va a ser efectiva, los analistas de datos de Showz podrían dividir por cohortes y repetir simulaciones para ver si ciertos segmentos responden mejor a incrementos de presupuesto. Asimismo, podrían repetir el estudio después de un periodo de tiempo (mensual, trimestral, semestral) y estar al día sobre el desempeño de cada una de sus fuentes.

Siempre se debe priorizar la fuente de mayor ROMI para reducir el costo de oportunidad, acelerar el crecimiento reinvirtiendo en canales de alta rentabilidad, maximizar la eficiencia del presupuesto y simplificar la toma de decisiones.

De igual manera, en el futuro podría integrar modelos de reinforcement learning que, a lo largo de campañas, aprendan sobre la asignación óptima entre canales para maximizar ROMI acumulado. Incluir análisis de churn que anticipe usuarios inactivos o propensos a no volver, disparando campañas de retargeting antes de perderlos.

***Limitaciones***

Algunos de los obstáculos que enfrenta este análisis son:

- Se basa únicamente en la fuente de primer contacto del usuario. No considera otras interacciones posteriores que puedan haber influido en la compra, por lo que el CAC y el LTV podrían estar sesgados.
- No considera efectos de saturación o decrecimiento en interacciones.
- Las estimaciones de LTV y CAC provienen de modelos que tienen un margen de error. Si los errores de predicción son elevados, los cálculos de ROMI y las simulaciones pueden dar recomendaciones equivocadas.
- El análisis no segmenta a los usuarios (edad, ubicación, dispositivo, tipo de evento). Puede que ciertos grupos respondan muy distinto a los cambios de presupuesto, y el análisis global oculta esas diferencias.
- El LTV se calcula a 180 días y el CAC a 30 días. No considera el valor más allá de esos plazos, ni el impacto en retención a largo plazo ni el churn después de un tiempo.
- Se simula un único movimiento de presupuesto (+10 % o redistribución focalizada). No contempla ajustes dinámicos basados en resultados reales de campañas.
- Depende de la calidad de los datos históricos de sesiones, conversiones y costos. Cualquier duplicado, dato faltante o error de tracking impacta directamente en las métricas y simulaciones.

***Conclusiones***

Adoptar un enfoque data-driven en la gestión de marketing es una necesidad estratégica. Al fundamentar cada decisión en métricas objetivas y simulaciones cuantitativas, la empresa maximiza el retorno de cada dólar invertido, minimiza el riesgo puesto que se basa en evidencia palpable para tomar decisiones, agiliza la respuesta a cambios de mercado refinando continuamente los modelos, fomenta la colaboración entre equipos al disponer de KPIs (métricas que evalúan el éxito de una organización)  y dashboards en tiempo real.

Las estrategias data-driven transforman cualquier estrategia en un motor de crecimiento sostenible: no solo optimizan el gasto, sino que generan aprendizaje constante y permiten anticipar tendencias, maximizando ingresos y construyendo una ventaja duradera en un entorno cada vez más competitivo.

El entendimiento del negocio permitió identificar que el valor de cada cliente (LTV) y el costo asociado a su adquisición (CAC) son métricas críticas para la rentabilidad sostenida de Showz en el competitivo sector de eventos. Este enfoque estratégico guió la formulación del problema hacia la predicción y optimización del retorno de inversión en marketing. Desde el punto de vista de los datos, se evidenció una excelente calidad desde el inicio, sin valores nulos ni duplicados, lo cual facilitó una preparación eficiente y enfocada en el análisis. Las tres fuentes (visits, orders y costs) capturaron dimensiones clave del comportamiento del usuario y la inversión en marketing, incluyendo patrones estacionales, tipos de dispositivos, canales de adquisición y métricas temporales. La decisión de conservar outliers relevantes permitió preservar señales valiosas asociadas a clientes de alto valor o campañas intensivas.

Este entendimiento conjunto estableció una base sólida para la construcción de modelos predictivos robustos y alineados con los objetivos estratégicos de la empresa.

La ingeniería de características implementada fue fundamental para transformar datos transaccionales, temporales y de marketing en señales predictivas valiosas. Al construir variables como AOV, retención, frecuencia de compra, cohorte de adquisición y tendencia de gasto, se logró capturar comportamientos reales del cliente con un alto nivel de granularidad. Esta preparación permitió a los modelos identificar patrones relevantes para anticipar ingresos futuros y comportamientos de adquisición. En términos prácticos, esta etapa estableció la base que hizo posible predecir el valor futuro de un cliente con gran precisión, habilitando decisiones más inteligentes y rentables para la empresa.

Los resultados muestran que los modelos desarrollados, especialmente Random Forest y XGBoost, permiten predecir el Lifetime Value a 180 días con un error medio inferior a $1.40 y una precisión superior al 97%, lo que representa una herramienta altamente confiable para decisiones estratégicas de personalización y asignación de presupuesto. En contraste, la predicción del CAC presentó errores porcentuales elevados (>40%), lo que limita su aplicabilidad inmediata. Esta diferencia evidencia que, mientras los datos disponibles capturan bien el comportamiento del cliente, aún se necesita enriquecer la información de marketing para estimar costos de adquisición con la misma precisión. En conjunto, la evaluación valida que el modelo de LTV está listo para uso empresarial, mientras que el de CAC requiere iteración y mejora.

El modelo XGBoost para LTV_180 se fundamenta principalmente en variables de valor monetario y comportamiento de compra. El Average Order Value (AOV) y el revenue_total concentran la mayor parte del gain, lo que confirma que tanto el importe medio de las órdenes como el gasto acumulado son los mejores predictores del valor de vida del cliente a 180 días. Los Partial Dependence Plots corroboran esta relación lineal positiva y muestran que otras variables como el gasto mensual promedio o el mes de primera sesión aportan muy poca información adicional.

El análisis de errores revela que el modelo predice con gran precisión a los usuarios de baja frecuencia (1–4 compras), pero presenta errores muy elevados en segmentos extremos, especialmente en clientes con 19 compras o en el cuartil de ingresos más altos. Esto indica que el error aumenta en comportamientos atípicos no representados adecuadamente durante el entrenamiento.

Por su parte, la regresión Ridge para CAC_source_30 combina indicadores de valor (LTV_180, revenue_total) y de engagement (frecuencia, sesiones, dispositivos). Los coeficientes positivos de LTV_180, días entre compras y frecuencia de compras apuntan a que clientes de alto valor o con ciclos de compra más espaciados exigen mayor inversión inicial. En cambio, la relación inversa con revenue_total y días activos sugiere eficiencia en la adquisición de usuarios rentables y fieles. Los PDP confirman estas direcciones de efecto de forma suave y lineal, sin evidenciar interacciones complejas.

El estudio de residuos muestra un patrón bimodal y bandas horizontales en el scatter, lo que indica predicciones discretas y dos clusters de error: subestimaciones alrededor de –0.22 y sobreestimaciones cerca de 0.09. Además existe un leve sesgo a sobreestimar el CAC en algunos rangos. Esto apunta a un efecto de regularización fuerte o a agrupaciones de valores en la variable objetivo que el modelo no discrimina finalmente.

En conjunto, estos hallazgos permiten proponer mejoras concretas: aplicar modelado segmentado o técnicas robustas en los subgrupos con error elevado (usuarios de alta frecuencia o ingreso extremo en LTV_180; clusters de residuos en CAC_source_30), revisar el tratamiento de outliers y la escala de las variables objetivo, y establecer un monitoreo continuo de MAE y distribución de residuos por cohorte, con reentrenamientos periódicos para garantizar la adaptabilidad y robustez de ambos modelos.

**Referencias**

Kumar, A. (18 de agosto de 2024). *MSE vs RMSE vs MAE vs MAPE vs R-Cuadrado: ¿Cuándo usarlo?*. https://vitalflux-com.translate.goog/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Santa Clara Leavey School of Business. (6 de noviembre de 2023). *Exploring Data-Driven Marketing Strategies*. https://onlinedegrees.scu.edu/media/blog/exploring-data-driven-marketing-strategies

ScikitLearn. (s.f.). *Supervised Learning*. https://scikit-learn.org/stable/supervised_learning.html