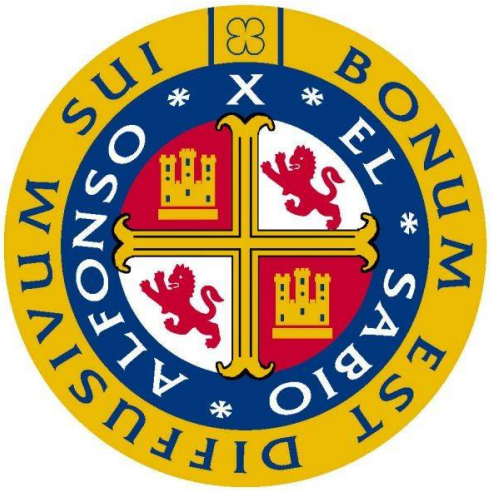
Universidad Alfonso X El Sabio

Grado en Ingeniería Matemática

Gestión Datos

Construcción DWH y Cálculo CLTV



Martina García González

Índice

1. Contexto

En el sector automotriz, la capacidad de analizar eficazmente la información relativa a clientes, ventas y servicios postventa es esencial para la toma de decisiones estratégicas. Sin embargo, muchas empresas enfrentan dificultades al contar con datos distribuidos en múltiples sistemas, lo que dificulta considerablemente su explotación analítica.

Este proyecto aborda dicho reto mediante la consolidación y optimización del modelo de datos de una empresa automotriz, que originalmente disponía de información distribuida en 19 tablas almacenadas en Azure. Estas tablas contenían información clave para el negocio, agrupadas en cuatro áreas principales:

- **ERP**: Información sobre ventas, costos, productos e inventarios.
- **CRM**: Datos sobre clientes, incluyendo sus perfiles, interacciones y comportamientos.
- **Logística**: Información acerca del transporte, distribución y ubicación de productos.
- **Postventa**: Historial de revisiones, mantenimiento y reclamaciones de clientes.

El objetivo principal del proyecto fue transformar esta estructura relacional inicial, compleja y poco eficiente en términos de rendimiento, en un modelo dimensional más ágil y optimizado mediante un proceso ETL (Extract, Transform, Load) con Python.

2. Modelo Relacional Inicial

Inicialmente, se diseño un modelo relacional compuesto por todas las tablas interconectadas mediante claves primarias y foráneas.(El modelo ERD se encutra adjunto en los archivos envíados)

El análisis de este modelo permitió entender las relaciones existentes, pero también reveló desafíos importantes que debían gestionarse cuidadosamente durante el proyecto:

- **Relaciones muchos a muchos**: Dado que estas relaciones pueden generar referencias circulares y problemas de rendimiento, se tuvo especial cuidado desde el inicio para evitar la creación de este tipo de relaciones, simplificando así al máximo las conexiones entre las entidades.
- **Riesgo de referencias circulares**: Se definieron con atención las claves y relaciones

para asegurar que no existiesen dependencias cíclicas, que dificultan tanto el rendimiento como el mantenimiento.

- **Optimización estructural**: Se optimizó la estructura del modelo relacional mediante la consolidación de información redundante, asegurando relaciones claras y directas para facilitar su explotación analítica.

Estos desafíos subrayaron la importancia de transformar el modelo relacional en un modelo dimensional más simple y eficiente, facilitando así la generación rápida de información estratégica.

3. Transformación del Modelo: Proceso ETL

Para alcanzar una estructura óptima de análisis, se llevó a cabo un proceso ETL, dividido en tres fases claramente definidas:

3.1. Extract

En la fase de extracción, se establecieron conexiones seguras entre Azure SQL y SQL Server. Para ello, se utilizaron scripts en Python que permitieron acceder de manera segura a ambas bases de datos.

Una vez establecidas estas conexiones, se procedió a leer archivos SQL previamente preparados, que contenían las consultas específicas necesarias para crear cada tabla del modelo dimensional. Estas permitieron extraer los datos directamente desde Azure , almacenándolos inicialmente en estructuras intermedias (dataframes), preparándolos para la siguiente fase del proceso ETL.

3.2. Transform

Durante esta fase se llevaron a cabo tareas críticas para asegurar la calidad y coherencia de los datos:

- **Gestión de valores nulos**: Los valores vacíos en columnas categóricas se sustituyeron por la etiqueta 'NA', evitando confusiones en futuros análisis.

- **Conversión de formatos**: Las fechas se ajustaron a un formato estándar (YYYY-MM-DD), facilitando análisis temporales precisos. Adicionalmente, las columnas nu-

méricas se optimizaron para reducir su tamaño, transformando tipos int64 a int32 y float64 a float32.

- **Estandarización de categorías**: Se consolidaron nombres y códigos para evitar duplicaciones, especialmente en productos, clientes y regiones geográficas.

- **Tratamiento específico en variables críticas**: Variables clave como días desde la última revisión fueron revisadas cuidadosamente para asegurar su correcta interpretación durante cálculos críticos, como el Churn.

3.3. Load

Finalmente, los datos transformados fueron cargados en SQL Server. Durante este proceso, se implementaron mecanismos de validación para garantizar que la información se insertara correctamente sin pérdidas de registros. Realizando comparaciones entre la cantidad de registros extraídos y los registros cargados en la base de datos final para asegurar la integridad

4. Modelo Dimensional

Un modelo dimensional es una técnica de diseño de bases de datos orientada al análisis, específicamente pensada para sistemas OLAP (Online Analytical Processing), cuyo objetivo principal es facilitar consultas rápidas y multidimensionales sobre grandes volúmenes de información. Este tipo de modelo permite analizar los datos desde diferentes perspectivas mediante una estructura en forma de cubo.

El modelo dimensional se basa en dos componentes fundamentales:

- **Tablas de hechos (Fact Tables)**: Son tablas que almacenan métricas cuantitativas del negocio, como ventas, costos, márgenes, entre otras. Constituyen el núcleo central del análisis.

- **Tablas de dimensiones (Dimension Tables)**: Representan las distintas perspectivas o categorías mediante las cuales se analizan las métricas almacenadas en la tabla de hechos.

Basándonos en esta metodología, se diseñó un modelo dimensional adaptado específicamente a las necesidades analíticas de la empresa automotriz, con la siguiente estructura:

- **Fact**: Contiene el histórico de transacciones comerciales y métricas fundamentales.

- **Cliente**: Almacena información demográfica y comportamental de los clientes.
- **Producto**: Registra características técnicas específicas de los vehículos vendidos.
- **Tiempo**: Permite realizar análisis temporales detallados de ventas y tendencias.
- **Geografía**: Contiene datos sobre la ubicación de clientes y puntos de venta para un análisis geográfico profundo.

5. Cálculo de Métricas Relevantes en la Tabla Fact

Con el modelo dimensional optimizado, se procedió al cálculo de métricas fundamentales para la empresa, permitiendo evaluar la rentabilidad y el comportamiento de los clientes.

5.1. Margen Bruto

El margen bruto representa la rentabilidad obtenida antes de aplicar costos adicionales. Se calcula en función del precio de venta, ajustado por el margen de contribución y los impuestos asociados.

5.2. Margen Neto

A diferencia del margen bruto, el margen neto refleja la ganancia real después de descontar costos operativos, como gastos de distribución, marketing, transporte y comisiones.

5.3. Cálculo del Churn

El churn, o tasa de abandono de clientes, es un indicador clave para evaluar la fidelidad y retención de clientes. En este análisis, se establecieron reglas claras para su cálculo:

- Clientes sin revisiones en más de 401 días se consideran en alto riesgo de abandono ($Churn = 1$).
- Clientes con vehículos de más de un año y sin revisiones recientes también se clasifican como de alto riesgo ($Churn = 1$).
- Clientes con vehículos recientes y revisiones activas se consideran activos ($Churn = 0$).

Durante la implementación del modelo, se detectó que en gran parte los registros no se disponía de información de los días desdes la última revisón del cliente. Para solucionar este

problema, se aplicó un criterio conservador: si no se tenía información sobre la última revisión y el vehículo tenía más de dos años de antigüedad, se consideró como un cliente en riesgo de churn ($Churn = 1$).

Sin embargo, esta decisión se basó en una interpretación de los datos . Para mejorar la exactitud en el cáclulo del Churn, se recomienda analizar los patrones históricos de revisiones para lograr una adecuada estimación de estos valores nulos .

6. CLV

El *Customer Lifetime Value* (CLV) es una métrica clave para estimar los ingresos netos que un cliente generará a lo largo de su relación con la empresa. Comprender esta métrica permite optimizar estrategias de fidelización y segmentación de clientes.

Resumidamente,

- **CLV alto**: Indica que el cliente es rentable y genera ingresos significativos.
- **CLV bajo o negativo**: Puede significar que el cliente no aporta beneficios o incluso que la empresa está incurriendo en pérdidas con él.

6.1. Fórmula del CLV a 5 años

El CLV a cinco años se calcula mediante la siguiente expresión:

$$CLV_{5\_anos} = \text{Margen\_eur\_Medio} \times \sum_{t=1}^{5} \frac{\text{Retencion}_t}{(1+i)^t} \tag{1}$$

Donde:

- Retencion_t: Probabilidad de que el cliente continúe comprando en el tiempo.
- Margen_eur_Medio: Beneficio neto por cliente, calculado como la diferencia entre los ingresos generados por el cliente y los costos asociados a su adquisición y mantenimiento.
- $i = 7\,\%$: Tasa de descuento aplicada para ajustar el valor del dinero en el tiempo.

6.2. Relevancia de la Retención en CLV

La retención es un factor determinante en el cálculo del CLV, ya que impacta directamente en la rentabilidad del cliente a lo largo del tiempo. Se pueden presentar varios escenarios:

- **Alta Retención**: Un cliente con alta retención seguirá siendo rentable a largo plazo, incluso si su margen es moderado.
- **Retención Cero**: Si la retención es cero, el cliente deja de generar ingresos en el futuro, sin importar su margen actual.
- **Alta Retención pero Margen Negativo**: Cuando la retención es alta pero el margen es negativo, el cliente continuará generando pérdidas con el tiempo.

6.3. Estimación del CLV

Para calcular el CLV, primero fue necesario estimar la probabilidad de *Churn* mediante un modelo de regresión lineal. En este proceso, se calcularon diversas métricas agregadas a partir del precio de venta promedio (PVP) y otras características de los clientes y vehículos (Número de Compras, Edad Media del Cliente, Edad Media del Coche,Margen Bruto Medio...)

Tras evaluar la correlación de estas variables con el churn, se identificaron aquellas con mayor impacto en la predicción del abandono de clientes. La ecuación de regresión para el churn se expresa como:

$$\begin{aligned} \text{Churn} = \text{Intercepto} + a_1 \cdot PVP + a_2 \cdot \text{Edad\_Media\_Coche} \\ + a_3 \cdot \text{Km\_medio\_por\_revision} + a_4 \cdot \text{Revisiones\_Medias} \end{aligned} \tag{2}$$

Los coeficientes obtenidos han sido almacenados en una tabla local para automatizar la gestión de los datos y el cálculo del CLV para cada cliente.

6.4. Gráficas CLV

Aunque se ha realizado un análisis más detallado del CLV en Streamlit, donde se calculan múltiples métricas adicionales, a continuación se presenta la distribución general del CLV calculado a 5 años:

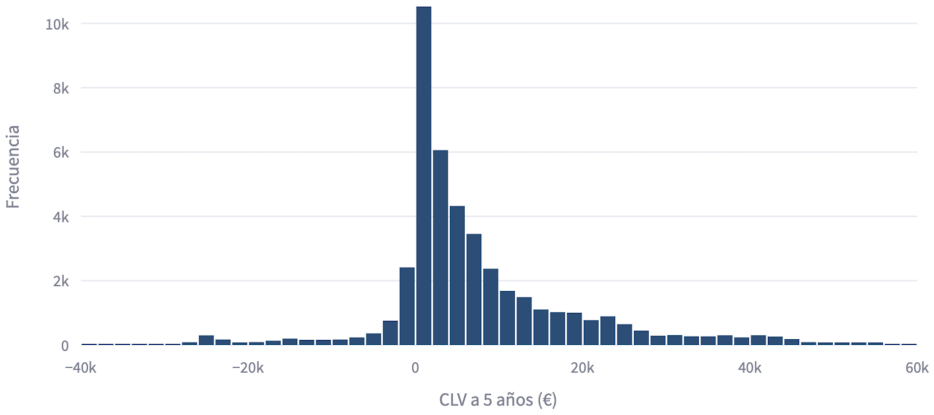


Figura 1: Distribución del CLV a 5 años (€)

Análisis Distribucción

- Distribución sesgada positiva: Mayoría cercana a 0€, pocos clientes altamente rentables.
- Colas largas: Clientes negativos (pérdidas) y positivos (altamente rentables).
- Pico en 0€: Alta proporción de clientes sin beneficios netos claros.

7. Segmentación de Clientes y Estrategias de Fidelización

Una vez obtenida la probabilidad de *churn*, se aplicaron técnicas de reducción de dimensionalidad y segmentación no supervisada para identificar patrones de comportamiento en los clientes.

- Se utilizó **Análisis de Componentes Principales (PCA)** para reducir la dimensionalidad de los datos y mejorar la interpretabilidad de los segmentos.
- Posteriormente, se aplicó el algoritmo **K-Means**, agrupando a los clientes en diferentes segmentos en función de sus características de compra y comportamiento.

Cada segmento se asoció con estrategias de fidelización personalizadas, permitiendo optimizar la retención la rentabilidad de cada grupo.

8. Estructura Proyecto

8.1. Notebooks

- **01_modelo_dimensional.ipynb**: Creación del modelo dimensional en SQL Server (SSMS).
- **02_vision_cliente.ipynb**: Generación de métricas agregadas por cliente.
- **03_regresion_cliente.ipynb**: Modelo de regresión logística para predecir el churn.

8.2. Consultas SQL

Consultas diseñadas y comentadas para estructurar las tablas dimensionales:

- **dim_cliente.sql**: Tabla de clientes.
- **dim_fact.sql**: Tabla de hechos.
- **dim_geog.sql**: Tabla de geografía.
- **dim_prod.sql**: Tabla de productos.
- **dim_tiempo.sql**: Tabla de tiempo.

Análisis y Insights:

- **vision_cliente.sql**: Vista 360° del cliente con métricas clave.
- **regresion_cliente.sql**: Consulta de para regresión lineal que estima Churn.
- **bi_cliente.sql**: Consulta final con cálculo de CLV y análisis avanzado.

8.3. Arhivos Adjuntos

- **Diargrama ERD**
- **Diargrana Dimensional**.
- **Arquitectura**
- **Despliegue en Streamlit**: Enlace a la app de Streamlit

9. Conclusiones del Proyecto

Disponer de un **Data Warehouse (DWH)** es fundamental para gestionar eficazmente grandes volúmenes de datos en una empresa. En nuestro caso, partimos de un DWH que consolidaba datos procedentes de múltiples fuentes funcionales dentro

Posteriomente, realizamos una transición clave, pasando de un complejo modelo relacional con 19 tablas a un modelo dimensional. Este cambio simplificó significativamente la estructura de datos y mejoró sustancialmente la rapidez y eficiencia de las consultas analíticas, facilitando la obtención ágil de resultados.

Finalmente, efectuamos un análisis profundo del **Customer Lifetime Value (CLV)**, una métrica crucial para identificar el tipo de clientes dentro de una empresa.

En definitiva, en el ámbito empresarial es fundamental contar con datos bien estructurados, organizados y accesibles, ya que esto permite realizar cálculos precisos y confiables del CLV y otras métricas clave. Disponer de información clara y actualizada garantiza la toma de decisiones estratégicas acertadas, lo que contribuye directamente a mejorar la competitividad y rentabilidad de la empresa.

Puntos a mejorar:

1. Implementar un modelo de **regresión logística** para la estimación del churn, asegurando la estandarización previa de los datos para evitar el overfitting y mejorar la precisión del modelo.

2. **Automatizar la orquestación** de todos los procesos analíticos para generar reportes y dashboards actualizados diariamente, facilitando la monitorización continua de las métricas clave.

3. Desarrollar **simulaciones estratégicas** del CLV para evaluar diferentes escenarios, como ajustes en costos o estrategias de fidelización, optimizando continuamente la rentabilidad del cliente.