UNIVERSIDAD
ALFONSO X EL SABIO

Caso 01 Modelo Relacional y construcción de DWH Analítico + CLTV

Alejandro Serrano Catalina

29 de marzo de 2025

Introducción

En el ámbito de la gestión de datos empresariales, el modelado dimensional y la migración de datos son fundamentales para obtener insights significativos y tomar decisiones estratégicas informadas. Este proyecto tiene como objetivo principal el diseño e implementación de un modelo dimensional utilizando un esquema estrella, para lo cual se realiza una migración de datos desde una base de datos en la nube a un entorno local.

La implementación de un proceso ETL eficiente y la creación de un modelo de datos coherente son pasos clave para asegurar la integridad de los datos y optimizar su posterior análisis. En este contexto, se analiza el comportamiento de los clientes mediante el uso de regresión y la estimación del Customer Lifetime Value (CLTV), con el fin de prever la rentabilidad a largo plazo de cada cliente.

Este documento presenta una descripción detallada de los pasos seguidos en el proyecto, desde el análisis inicial de los datos hasta la implementación de modelos predictivos y el cálculo de métricas clave como el CLTV, esenciales para las decisiones de negocio.

Desarrollo y Ejecución del Proyecto

1. Análisis y Comprensión de las Tablas y sus Relaciones

El primer paso ha sido analizar la estructura de las 19 tablas en la base de datos, identificando sus atributos, claves primarias y foráneas para entender las relaciones entre ellas. Este análisis permitió definir cómo fluyen los datos y sentar las bases para el diseño del modelo dimensional.

2. Diseño del Diagrama Entidad-Relación (DER)

Con el análisis previo, se creó un Diagrama Entidad-Relación (DER) que representa visualmente la estructura de la base de datos. Este diagrama permite validar la integridad de los datos y optimizar futuras consultas.

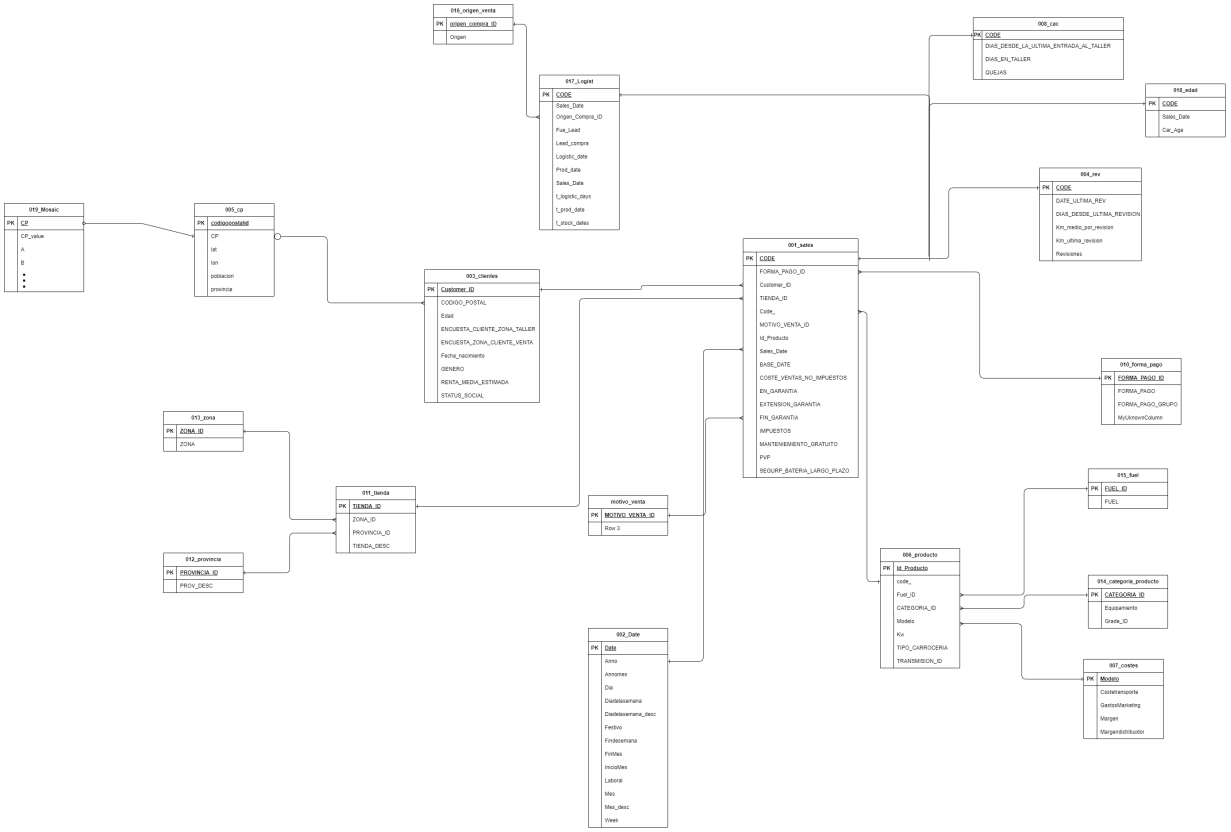

Figura 1: Diagrama Entidad-Relación (DER) del proyecto.

El DER del proyecto se encuentra en la carpeta `Database/ER`, dentro del archivo adjunto, y está disponible en los formatos `.drawio` y `.png`.

3. Creación del Modelo Dimensional

El modelo dimensional se estructuró en un esquema estrella, compuesto por varias dimensiones y una tabla de hechos central. Las **tablas dimensionales** son aquellas que contienen información descriptiva o atributiva, como detalles de clientes, productos, tiempo y ubicación. Estas tablas permiten desglosar los hechos y realizar análisis detallados. La **tabla de hechos**, por su parte, contiene las métricas y cifras clave del negocio, y se conecta a las tablas dimensionales mediante claves foráneas. A continuación, se describen las tablas del modelo dimensional:

- **Dimensión Tiempo (Dim_Time)**: Esta tabla contiene información sobre las fechas, permitiendo analizar las ventas en función de distintos periodos temporales como años, meses, días y semanas. Los campos clave de esta tabla son:
  - **Date**: Fecha específica.
  - **Anno, Annomes**: Año y año/mes en formato numérico.
  - **Dia, Diadelasemana**: Día del mes y del semana.
  - **Mes, Mes_desc**: Mes y su descripción.
  - **Laboral, Festivo**: Indicadores sobre si el día es laboral o festivo.
  - **Semana**: Número de la semana del año.
- **Dimensión Cliente (Dim_Cli)**: Esta tabla almacena datos demográficos y socioeconómicos de los clientes. Incluye atributos como:
  - **Customer_ID**: Identificación única del cliente.
  - **Edad**: Edad del cliente.
  - **Fecha_nacimiento**: Fecha de nacimiento del cliente.
  - **GENERO**: Género del cliente.
  - **CP, Población, Provincia**: Información geográfica de los clientes.
  - **RENTA_MEDIA_ESTIMADA**: Renta media estimada del cliente.
  - **ENCUESTA_ZONA_CLIENTE_VENTA, ENCUESTA_CLIENTE_ZONA_TALLER** Resultados de encuestas relacionadas con la zona de venta y taller del cliente.
- **Dimensión Producto (Dim_Product)**: Esta tabla contiene información detallada sobre los productos vendidos, como su categoría, modelo y costos asociados. Los campos principales incluyen:
  - **Id_Producto**: Identificador del producto.
  - **Code**: Código del producto.
  - **CATEGORIA_ID**: Identificador de la categoría del producto.
  - **Modelo**: Modelo del producto.
  - **FUEL**: Tipo de combustible asociado al producto.
  - **Grade_ID, Equipamiento**: Detalles del producto relacionados con su categoría y equipamiento.
  - **Costos**: Información sobre los costos asociados al producto, como coste de transporte y marketing.
- **Dimensión Geográfica (Dim_Geo)**: Esta tabla describe la ubicación geográfica de los clientes y las tiendas. Incluye los siguientes atributos:

- **TIENDA_ID**: Identificación de la tienda.
- **PROVINCIA_ID, ZONA_ID**: Identificadores de la provincia y zona.
- **TIENDA_DESC**: Descripción de la tienda.
- **PROV_DESC**: Descripción de la provincia.
- **ZONA**: Nombre de la zona.

- **Tabla de Hechos (Dim_Fact)**: La tabla de hechos centraliza métricas como ventas, márgenes y otros valores clave que se relacionan con las dimensiones mencionadas. Algunos de los campos relevantes son:
  - **CODE**: Identificador único de la transacción de venta.
  - **TIENDA_ID**: Identificación de la tienda donde se realizó la venta.
  - **Customer_ID**: Identificación del cliente.
  - **Id_Producto**: Identificación del producto vendido.
  - **Sales_Date**: Fecha de la venta.
  - **PVP**: Precio de venta del producto.
  - **MANTENIMIENTO_GRATUITO, SEGURO_BATERIA, COSTE_VENTA, IMPUESTOS, GARANTIA**: Detalles sobre los costos adicionales relacionados con la venta.
  - **Margen**: Margen calculado basado en la venta y otros costos.
  - **Churn**: Indicador de cancelación de la venta (churn) basado en la última revisión del producto.

Las tablas dimensionales permiten desglosar las métricas almacenadas en la tabla de hechos, ofreciendo un análisis detallado según diferentes atributos (como cliente, producto, tiempo y geografía).

Este esquema estrella facilita la realización de consultas complejas y la creación de reportes detallados, lo cual es esencial para la toma de decisiones estratégicas.

Las tablas de las dimensiones y la tabla de hechos se encuentran en la carpeta `Database/Dimensional` dentro del archivo adjunto.

4. Implementación del Proceso ETL

Para trasladar los datos desde una base de datos en la nube (Azure SQL Server) a un entorno local, se desarrolló un proceso ETL optimizado. Este proceso consta de las siguientes etapas:

Extracción de Datos

- Se establecieron conexiones con las bases de datos origen y destino mediante `pyodbc`.
- Se utilizaron consultas SQL almacenadas en archivos externos (`Dim_Geo.sql`, `Dim_Product.sql`, etc.) para extraer datos de forma modular.
- Los resultados fueron almacenados en `DataFrames` de `pandas` para facilitar su manipulación.

Transformación de Datos

- Se eliminaron tablas preexistentes en la base de datos local para evitar inconsistencias.
- Se limpiaron datos eliminando duplicados y manejando valores nulos:
  - Datos numéricos: reemplazo por `-1`.
  - Datos categóricos: reemplazo por `"N/A"`.
  - Fechas: imputación con el valor más frecuente.
- Se verificó la integridad referencial asegurando que todas las claves primarias y foráneas estuvieran correctamente establecidas.
- Se ajustaron tipos de datos para optimizar consultas y garantizar la compatibilidad con el modelo dimensional.

Carga de Datos

- Se crearon dinámicamente las estructuras de las tablas en la base de datos local.
- Los datos transformados fueron insertados mediante `bulk insert`, optimizando la carga masiva.
- Se validó la carga comparando la cantidad de registros insertados con los datos extraídos y transformados.

El proceso ETL detallado se encuentra en el archivo adjunto, dentro de la carpeta `notebooks`, en un archivo llamado `code_mover_entornos.ipynb`.

5. Cálculo de la Regresión

Tras la migración de los datos, se implementó un modelo de regresión lineal para analizar la relación entre diferentes variables y predecir el comportamiento del cliente.

- Se seleccionaron variables clave como PVP, Edad Media del Coche, Km Medio por Revisión, Margen Medio y Revisiones Medias.
- Se ejecutó una consulta SQL para agrupar datos por PVP y calcular medias de otras métricas.

- Se entrenó un modelo de regresión lineal con las variables seleccionadas, obteniendo coeficientes para predecir el porcentaje de churn.
- Se evaluó el modelo utilizando métricas como R^2 y *MSE*.

Evaluación del Modelo

Para evaluar la calidad del modelo, se analizaron las métricas de rendimiento:

- **R^2 (coeficiente de determinación)**: Indica qué porcentaje de la variabilidad en la variable dependiente es explicada por el modelo. En este caso, se obtuvo un valor de **0.6257**, lo que sugiere que el modelo explica aproximadamente un **62.57 %** de la variabilidad del churn.
- **MSE (error cuadrático medio)**: Mide la magnitud media de los errores. Se obtuvo un valor de **0.0317**, lo que indica que el modelo tiene un error relativamente bajo.

	Métrica	Valor
0	R^2 (coef. de determinación)	0.625758
1	MSE (error cuadrático medio)	0.031737

Figura 2: Evaluación del modelo de regresión: R^2 y MSE.

Coeficientes Estimados

Los coeficientes obtenidos en la regresión indican el impacto de cada variable en la probabilidad de churn:

	Variable	Coeficiente
0	PVP	-0.000008
1	Edad_Media_Coche	0.102158
2	Km_Medio_Por_Revision	-0.000009
3	Margen_Medio	-0.001788
4	Intercepto	0.911341

Figura 3: Coeficientes estimados en la regresión.

- **PVP (-0.000008)**: Un precio de venta más alto reduce ligeramente la probabilidad de churn.

- **Edad Media del Coche (+0.102158)**: Cuanto mayor es la edad media del coche, mayor es la probabilidad de churn.
- **Km Medio por Revisión (-0.000009)**: Un mayor número de kilómetros entre revisiones reduce ligeramente el churn.
- **Margen Medio (-0.001788)**: Un mayor margen medio disminuye la probabilidad de churn.
- **Intercepto (0.911341)**: Representa el valor base del modelo cuando todas las variables independientes son cero.

La tabla para calcular la regresión puede encontrarse en la carpeta `Extras`, en el archivo `Crear_tabla_para_regresión.sql`.

Comparación del Churn Real vs. Predicho

Finalmente, se compararon los valores de churn reales con los predichos por el modelo para evaluar su precisión visualmente.

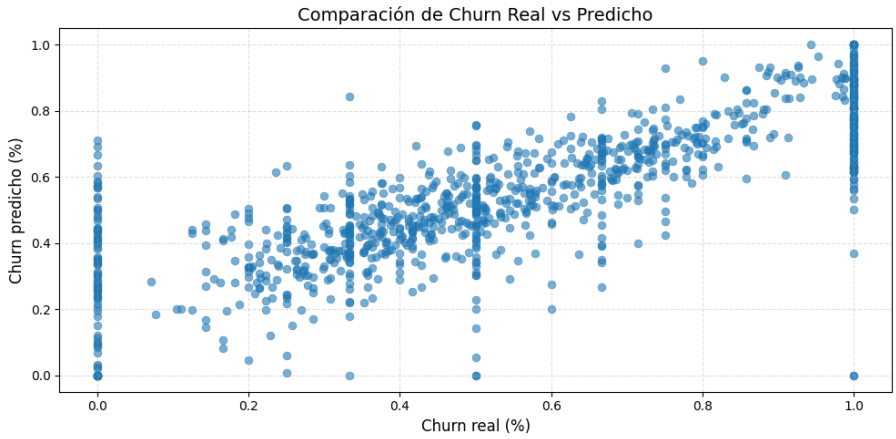

Figura 4: Comparación del Churn real vs. predicho.

Los resultados de la regresión calculada están disponibles en el archivo `regresion.ipynb`, que se encuentra en la carpeta `notebooks`.

6. Cálculo del Customer Lifetime Value (CLTV)

El CLTV se calculó para estimar el valor total que un cliente generará en un período de 5 años, siguiendo estos pasos:

- Se utilizaron los coeficientes del modelo de regresión para estimar la probabilidad de retención de cada cliente.
- Se calculó el CLTV considerando un descuento del 7 % sobre los ingresos futuros esperados.
- Se proyectó el valor del cliente a lo largo de 5 años, sumando sus márgenes anuales ajustados por la probabilidad de retención.

Esta métrica proporciona información clave para estrategias de fidelización y optimización de ingresos.

Encontraremos el archivo del calculo del cltv en la carpeta `Extras` llamado como `cltv.sql`

Resultados y gráficas

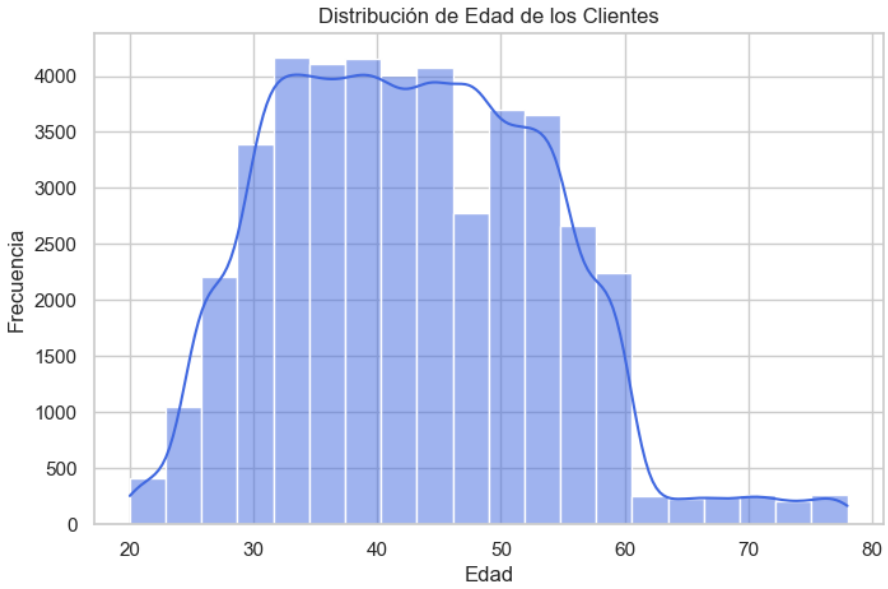

Figura 5: Distribución edad clientes

El análisis de la gráfica muestra que el CLTV disminuye en clientes de mayor edad, mientras que aquellos entre 30 y 60 años mantienen valores relativamente similares