Universidad Alfonso X El Sabio

Grado en Ingeniería Matemática

Gestión de Datos

Caso DW

Integrantes:
Calviño García, Jacobo

29 de marzo de 2025

1. Introduction

En este trabajo se lleva a cabo un proyecto completo enfocado en la gestión y análisis de datos. Comienza con un modelo entidad-relación que incluye 19 tablas, las cuales representan distintos aspectos importantes del negocio. Este modelo se simplifica en uno dimensional, compuesto por cinco tablas principales: la tabla central de Hechos, y cuatro tablas adicionales sobre Clientes, Tiempo, Productos y Zonas.

Luego, para trabajar cómodamente en un entorno local, se realiza un proceso de ETL (Extracción, Transformación y Carga), con el que se extraen los datos originales, se transforman y se cargan en el *localhost*, dejando todo listo para análisis posteriores.

Con los datos preparados, se crea la `Tabla_Clientes`, que permite tener información clara y ordenada sobre los clientes. A partir de esto, se realiza una regresión lineal, una técnica estadística que ayuda a entender mejor los patrones y predecir comportamientos futuros.

Finalmente, con los resultados de la regresión lineal, se calcula el valor del tiempo de vida del cliente (CLTV). Este cálculo permite conocer cuánto valor aporta cada cliente al negocio a lo largo del tiempo, ayudando así a tomar decisiones para mejorar la rentabilidad y mantener a los clientes satisfechos.

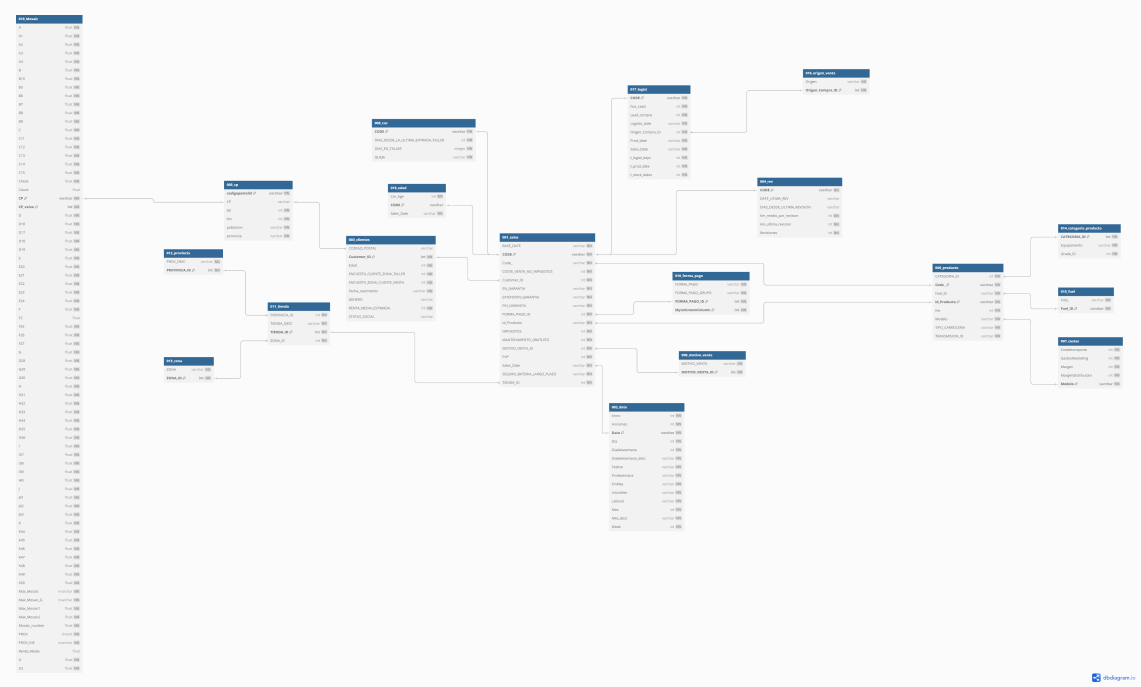
2. ER

El modelo entidad-relación (ER) desarrollado en este trabajo está compuesto por 19 tablas que representan diferentes aspectos fundamentales del negocio. La tabla principal es `001_sales`, que registra información clave sobre las ventas realizadas, siendo el núcleo alrededor del cual giran todas las demás tablas relacionadas.

Las uniones desde esta tabla principal hacia las otras tablas permiten detallar aspectos específicos relacionados con cada venta, como los clientes involucrados, las características de los productos vendidos, la dimensión temporal (fechas y períodos específicos), y las zonas geográficas correspondientes.

Algunas de estas relaciones aparecen gráficamente como *"0..1 a 1 o varios"*, sin embargo, en la práctica, estas relaciones siempre tienen al menos una instancia relacionada, por lo que en realidad se interpretan como relaciones obligatorias o necesarias, es decir, de *"1 a varios"* o *"1 a 1"*, según sea el caso.

Este modelo ER facilita una gestión organizada de los datos del negocio y permite realizar análisis efectivos, aportando claridad y estructura a la información utilizada en decisiones estratégicas posteriores.



3. Modelo Dimensional

Tabla de Hechos: Ventas

Esta es la tabla central del modelo dimensional. Registra cada transacción de venta y enlaza con las dimensiones de producto, cliente, zona, y tiempo. Incluye variables de negocio como el precio de venta, costes, márgenes y una variable calculada de churn, útil para análisis de retención de clientes.

Dimensión Clientes

Contiene información demográfica y socioeconómica del cliente, además de su localización geográfica y clasificaciones del sistema Mosaic. Permite analizar ventas por perfil de cliente, nivel de renta o provincia.

Dimensión Productos

Incluye atributos del producto como modelo, tipo de combustible, categoría, transmisión y carrocería. Se agregan también métricas como cantidad vendida y total facturado.

Dimensión Tiempo

Proporciona atributos temporales como día, mes, año, semana, tipo de día (laboral/festivo) y descripciones útiles para análisis estacionales o por periodo.

Dimensión Zonas

Contiene la jerarquía de ubicación de la tienda: zona, provincia y descripción de tienda. Es útil para segmentar las ventas por región geográfica.

4. Tabla Clientes

La **Tabla_Clientes** es una tabla derivada que se construye mediante la unión entre la dimensión de clientes (`dim_client`) y la tabla de hechos de ventas (`fact_sales`). Esta tabla tiene como objetivo integrar la información demográfica, socioeconómica y de comportamiento del cliente con los detalles de sus transacciones.

Gracias a esta fusión, se pueden realizar análisis enriquecidos que combinan:

- Características del cliente (edad, género, renta, perfil Mosaic, encuestas, etc.).
- Datos de compra (producto, precio, garantías, márgenes, impuestos).
- Métricas calculadas (como el total de leads asociados a un cliente).

Este tipo de tabla es muy útil en estudios de segmentación, modelos predictivos como CLTV (Customer Lifetime Value), churn o scoring de clientes.

5. Regresión Lineal

Regresión Lineal: Predicción de Churn

Esta sección se divide en dos partes: la construcción de una tabla agregada que sirve como entrada para el modelo de regresión y el análisis de los resultados obtenidos del modelo.

1. Tabla Agregada para la Regresión

Se construye una tabla agregada a partir de la `fact_sales`, agrupando por el precio de venta (`PVP`). Esta tabla resume variables relevantes como edad media del coche, kilómetros medios hasta revisión, porcentaje de churn, número medio de revisiones por cliente, margen medio por revisión, y los días desde la última revisión.

Esta tabla sirve como base para alimentar el modelo de regresión que busca explicar o predecir la probabilidad de churn en función de las variables mencionadas.

2. Análisis del Modelo de Regresión

EDA del modelo

La gráfica muestra que la mayoría de los registros tienen un valor de churn cercano a 0, evidenciando que la tasa de abandono es baja para la mayoría. Sin embargo, se observa un grupo significativo con valores intermedios, e incluso algunos casos alcanzan el valor máximo de 1. Esto indica variabilidad en la retención, donde la mayoría se mantienen activos mientras que una fracción notable experimenta abandono. La dispersión en estos valores resalta la importancia de analizar los factores que influyen en el comportamiento de churn.

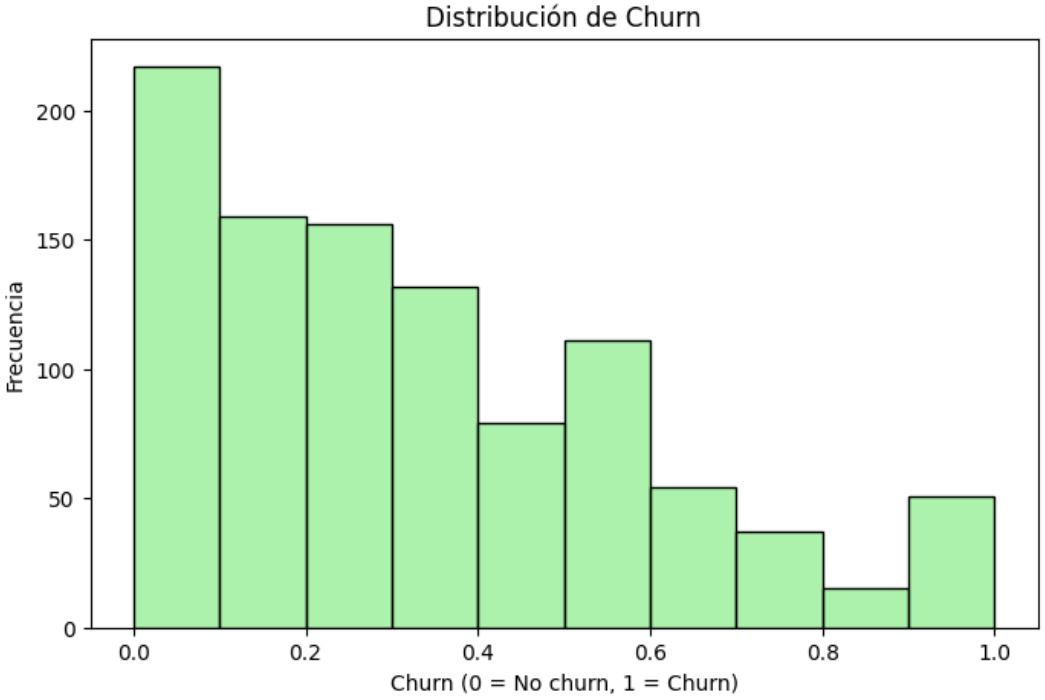


Figura: Distribución de la tasa de Churn en la muestra.

La gráfica revela que la mayoría de los registros se concentran alrededor de los 2 años, lo que indica que la muestra está compuesta en gran parte por elementos relativamente jóvenes. Se observa una disminución gradual en la frecuencia a medida que se aleja de este valor central, mostrando pocos casos en extremos inferiores o superiores. Este patrón sugiere homogeneidad en la edad, con la mayoría dentro de un rango estrecho. La distribución centralizada puede facilitar segmentaciones o comparaciones en análisis posteriores.

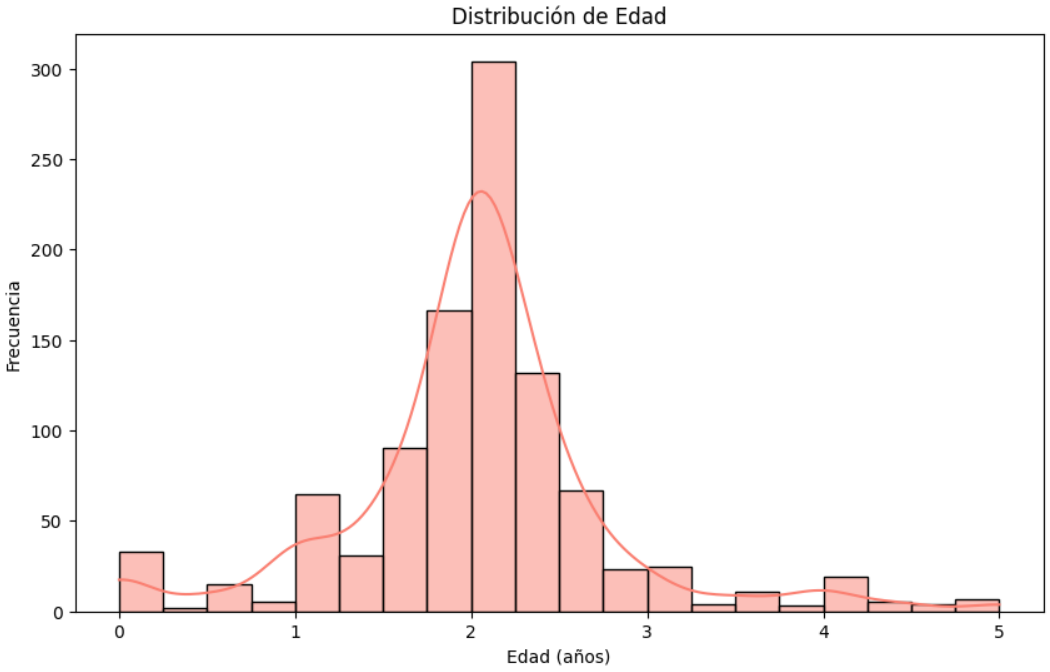


Figura: Distribución de Edad de la muestra.

Evaluación del Modelo

El modelo de regresión lineal obtenido fue evaluado con las siguientes métricas:

- **R^2 (coeficiente de determinación):** 0.638240
  Indica que aproximadamente el 65.4 % de la variabilidad en la variable objetivo (churn) es explicada por el modelo.
- **MSE (error cuadrático medio):** 0.025862
  Mide la media de los errores al cuadrado. Cuanto menor, mejor es el ajuste del modelo.

Coeficientes Estimados

Los coeficientes estimados por el modelo son los siguientes:

- `PVP`: -0.000007
- `Edad_Media_Coche`: 0.222160
- `Km_Medio_Por_Revision`: -0.000008
- `Revision`: -3.431966
- `Margen`: -0.001338
- `Intercepto`: 3.817628

La interpretación de estos coeficientes permite entender cómo cada variable afecta al porcentaje de churn. Por ejemplo, a mayor edad media del coche, mayor probabilidad de churn, mientras que más revisiones están asociadas a menor churn.

Gráfica: Churn Observado vs. Churn Predicho

La gráfica de **churn real vs. predicho** permite visualizar la capacidad del modelo para aproximarse a los valores reales. En ella se comparan los valores observados con los estimados por la regresión, lo cual es útil para identificar posibles patrones de error, evaluar la precisión del modelo y detectar sesgos.

Una alineación cercana entre los puntos y la línea de identidad indica un buen ajuste. Esta visualización es fundamental para validar visualmente la calidad de la predicción realizada por el modelo.

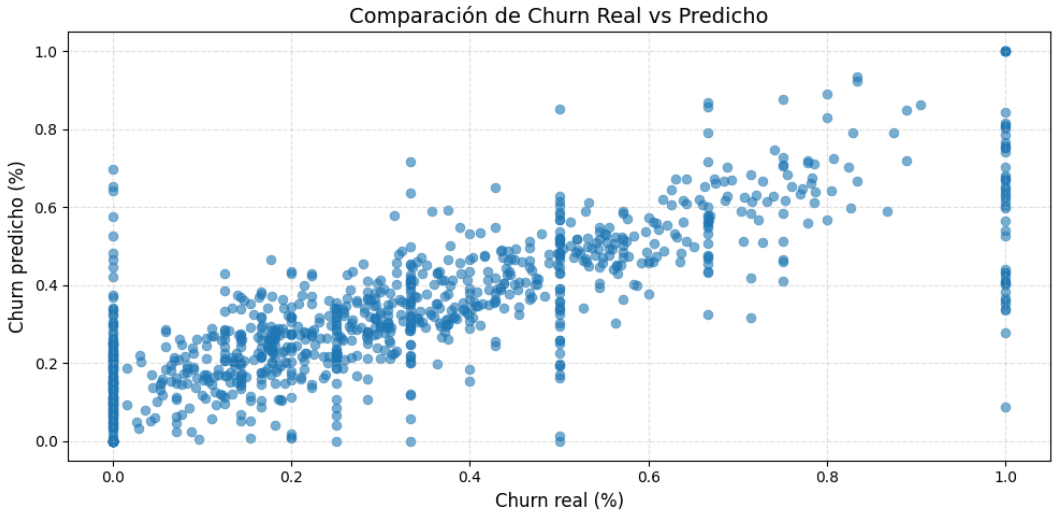


Figura: Comparación entre el churn observado y el churn predicho por el modelo.

6. CLTV (Customer Lifetime Value)

El **Customer Lifetime Value (CLTV)** es una métrica clave en marketing y análisis de clientes. Representa el valor estimado que un cliente aportará a la empresa a lo largo de su relación con ella. En este proyecto, el CLTV se calcula combinando la probabilidad de retención del cliente (estimada con un modelo predictivo) y la media de revisiones asociadas a ese cliente.

Propósito del Código

Este procedimiento en SQL permite, a partir del modelo de churn previamente entrenado, predecir el comportamiento futuro de los clientes y estimar su valor futuro en función de la retención y las revisiones realizadas. El cálculo está diseñado para proyectar el valor del cliente durante 1 a 5 años, teniendo en cuenta una tasa de descuento del 7 % anual.

Qué se realiza en el código

1. **Declaración de variables:** Se inicializan variables para almacenar los coeficientes del modelo de churn entrenado previamente.
2. **Carga de coeficientes:** Se extraen los coeficientes estimados del modelo (como intercepto, edad, PVP, etc.) desde una tabla llamada `churn_coef`.
3. **CTE de retención:** Se calcula para cada cliente la **probabilidad de retención** estimada utilizando la ecuación del modelo lineal de churn:

   $$\text{retención estimada} = 1 - (b_0 + b_1 \cdot PVP + b_2 \cdot Edad + \ldots)$$

   Se aplica un `LEAST(1, GREATEST(0, ...))` para limitar los resultados al rango válido [0, 1].
4. **Consulta final:** Se obtiene para cada cliente:
   - **churn_estimado:** 1 - retención estimada.
   - **retención_estimado:** probabilidad de que el cliente continúe.
   - **CLTV a 1, 2, 3, 4 y 5 años**, calculado con la fórmula de valor presente (descuento anual del 7 %):

     $$\text{CLTV}_n = \sum_{t=1}^{n} \frac{(\text{retención})^t \cdot \text{Margen}}{(1 + 0{,}07)^t}$$

   Esto permite valorar a los clientes no solo por sus compras actuales, sino también por su potencial de generar ingresos en el futuro.

   Clientes con alta retención y muchas revisiones tendrán un CLTV alto, lo que los convierte en candidatos ideales para fidelización o estrategias personalizadas.

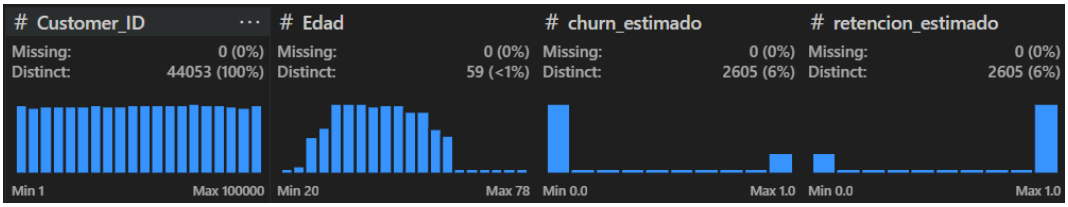
Aplicaciones del CLTV

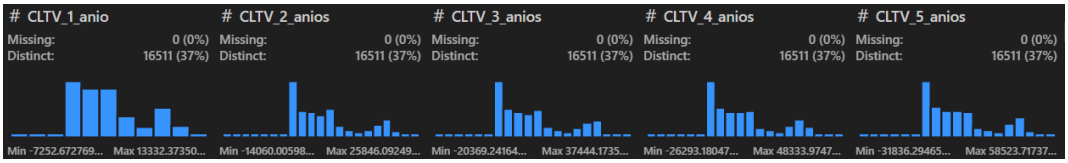
- **Segmentación de clientes por valor futuro.**
- **Diseño de estrategias de retención y marketing personalizado.**
- **Asignación de recursos comerciales según el valor del cliente.**
- **Evaluación del retorno de inversión en adquisición de clientes.**

En estas gráficas se observan diversas columnas que ofrecen una perspectiva integral del conjunto de datos. Por un lado, `Customer_ID` muestra un gran número de registros únicos sin valores faltantes, lo que indica una base de clientes extensa y completa. La variable `Edad`, con un máximo cercano a 20, sugiere una distribución concentrada en rangos relativamente jóvenes, mientras que `churn_estimado` y `retencion_estimado` revelan estimaciones de abandono y permanencia con distintos valores discretos, sin huecos en la información.

En la segunda parte, se muestran las variables de `CLTV` (Customer Lifetime Value) proyectadas a distintos horizontes de 1 a 5 años. Todas presentan rangos amplios, con valores mínimos y máximos considerables, y mantienen un 0 % de datos ausentes. Esto sugiere que el conjunto está completo para cada periodo de proyección, lo que facilita el análisis del valor potencial de cada cliente a corto, mediano y largo plazo.

De manera general, la ausencia de datos faltantes y la variedad de valores en cada columna evidencian la riqueza del dataset para realizar estudios de segmentación, predicción de churn y estimaciones de ingresos futuros. Esta calidad de información garantiza una base sólida para construir modelos de marketing orientados a la retención y maximización del valor del cliente.





Conclusión

El desarrollo del modelo predictivo de churn y el cálculo del Customer Lifetime Value (CLTV) han permitido obtener una visión estratégica y orientada al cliente que va más allá del análisis tradicional de ventas. A partir de variables clave como la edad del coche, los kilómetros recorridos, el margen por revisión y la frecuencia de mantenimiento, se puede anticipar el comportamiento futuro del cliente y estimar su valor económico a lo largo del tiempo.

Esta capacidad predictiva resulta esencial para el área de marketing, ya que permite segmentar a los clientes en función de su riesgo de abandono y de su rentabilidad futura. Por ejemplo, los clientes con alta probabilidad de churn pero alto CLTV pueden ser priorizados en campañas de fidelización personalizadas, mientras que aquellos con bajo valor esperado podrían gestionarse con estrategias de bajo coste.

Además, el modelo permite asignar recursos de manera más eficiente, maximizando el retorno de inversión en acciones de retención y optimizando la planificación comercial. También facilita el diseño de programas de mantenimiento, abonos o promociones que estén alineados con el comportamiento real del cliente.

En definitiva, el uso combinado de modelos predictivos y métricas de valor proporciona una base sólida para una estrategia de marketing basada en datos, centrada en el cliente y orientada a la sostenibilidad y crecimiento del negocio a largo plazo.