Universidad Alfonso X El Sabio

Grado en Ingeniería Matemática

GESTION DE DATOS

Caso número 1 GD

Integrantes:

Sánchez Escribano, José Antonio

29 de marzo de 2025

Índice

Índice **1**

1. Introduction **1**

2. Metodología **1**
2.1. Análisis inicial del Data Lake en Azure 1
2.2. Análisis de Relaciones entre Tablas 2
2.3. Estructura Final de Datos 2

3. Transformación del Modelo de Datos **3**
3.1. Consolidación de Tablas 3
3.2. Procesamiento Técnico de Datos 3
3.3. Migración Automatizada 4

4. Construcción del Modelo Predictivo **4**
4.1. Creación de la Variable Objetivo (Churn) 4
4.2. Resultados del Modelo 4

5. Análisis del Valor del Cliente (CLTV) **5**
5.1. Relaciones Clave 5
5.2. Fórmula Básica 5
5.3. Interpretación Práctica 5

1. Introduction

2. Metodología

2.1. Análisis inicial del Data Lake en Azure

Se realizó un exhaustivo análisis de las 19 tablas disponibles en el Data Lake de Azure, identificando su estructura y relaciones. A continuación se detallan las tablas más relevantes:

Cuadro 1: Principales tablas del Data Lake

Tabla	**Descripción**
001_sales	Datos principales de ventas
002_date	Dimension temporal
003_clientes	Información de clientes
004_rev	Datos de revisiones
005_cp	Códigos postales
006_producto	Catálogo de productos

2.2. Análisis de Relaciones entre Tablas

Tras examinar detenidamente las 19 tablas del sistema, identificamos las siguientes conexiones clave:

- **Relaciones 1 a N (Uno a Muchos)**:
  - *Ejemplo*: Un cliente puede tener múltiples ventas registradas, pero cada venta pertenece a un único cliente
  - *Implementación*: Mediante campos como Customer_ID que vinculan tablas
- **Relaciones N a M (Muchos a Muchos)**:
  - *Ejemplo*: Productos que aparecen en múltiples ventas y ventas que contienen varios productos
  - *Solución*: Implementadas mediante tablas puente especiales
- **Casos especiales**:
  - Campos opcionales (valores nulos) en algunas relaciones
  - Conexiones indirectas a través de tablas intermedias
  - Relaciones condicionales basadas en criterios de negocio

Resultado del análisis:

- Todas las tablas quedaron interconectadas mediante relaciones directas o indirectas
- Se normalizaron las conexiones para evitar redundancias

2.3. Estructura Final de Datos

Después de analizar toda la información, creamos un sistema organizado donde:

- **Tabla principal de ventas**:
  - Contiene todos los registros de transacciones
  - Conecta con las demás tablas mediante códigos únicos
- **Tablas complementarias**:
  - Datos de clientes (quién compra)
  - Catálogo de productos (qué se vende)
  - Información de tiendas (dónde se vende)
  - Calendario (cuándo se vende)
- **Características clave**:
  - Todas las fechas siguen el mismo formato
  - Los productos y clientes tienen identificadores únicos
  - Diseño optimizado para generar informes rápidamente

Este sistema nos permite:

- Consultar cualquier venta con todos sus detalles
- Analizar tendencias por producto, tienda o período
- Mantener la información actualizada sin duplicados

3. Transformación del Modelo de Datos

3.1. Consolidación de Tablas

Una vez definido el modelo entidad-relación, simplificamos la estructura original mediante:

- **5 tablas principales**:
  - `FACT_SALES`: Tabla maestra con todas las transacciones comerciales
  - `DIM_CLIENTE`: Información unificada de clientes
  - `DIM_PRODUCTO`: Catálogo completo de productos
  - `DIM_LUGAR`: Datos geográficos y de tiendas
  - `DIM_TIEMPO`: Calendario analítico
- **Reducción de complejidad**:
  - De 19 tablas originales a 5 tablas optimizadas
  - 112 campos reorganizados lógicamente

3.2. Procesamiento Técnico de Datos

- **Tratamiento de valores faltantes (análisis predictivo)**:
  - **¿Por qué**: Los algoritmos de machine learning no pueden trabajar con valores nulos. Para variables numéricas clave usadas en regresión, sustituimos los vacíos por la mediana (no por la media) porque:
    - La mediana es más robusta contra valores extremos
    - Mantiene mejor la distribución original de los datos
    - Evita sesgos en modelos lineales
  - **Implementación**: Usamos consultas SQL condicionales que solo aplican este tratamiento a campos específicos seleccionados para modelado.
- **Definición de restricciones de integridad**:
  - **¿Por qué?**: Las Primary Keys (PKs) y Foreign Keys (FKs) son el esqueleto de cualquier base de datos porque:
    - PKs: Garantizan que cada registro es único (evitan duplicados)
    - FKs: Mantienen las relaciones lógicas entre tablas (evitan datos huérfanos)
  - **Cómo**: Implementamos mediante ALTER TABLE con verificaciones en dos fases:
    1. Primero comprobamos que los datos cumplen las reglas
    2. Luego aplicamos las restricciones definitivas
- **Selección de columnas no nulas (churn)**:
  - **¿Por qué?**: Para predecir abandono de clientes necesitamos:
    - Datos completos en variables críticas (ej: frecuencia de compra)
    - Consistencia temporal en series históricas
  - **Implementación**: Creamos vistas materializadas con:
    - Clausulas WHERE que filtran registros incompletos
    - COALESCE para campos opcionales no usados en el modelo

3.3. Migración Automatizada

Desarrollamos un proceso en Python que:

- Extrajo los datos limpios de Azure Data Lake
- Transformó las relaciones complejas en joins optimizados
- Cargó la estructura final en nuestra base de datos local
- Estableció automáticamente las claves primarias y foráneas

4. Construcción del Modelo Predictivo

4.1. Creación de la Variable Objetivo (Churn)

- **Definición técnica**:
  - Variable binaria: $churn \in \{0, 1\}$
  - Criterio: Cliente inactivo si no realiza revisión en últimos 400 días o si no ha realizado nunca revisión.
  - Lógica: Umbral basado en ciclo de mantenimiento promedio del sector
- **Variables clave**:
  - `PVP`: Precio como variable continua
  - `Car_Age`: Antigüedad del vehículo (normalizada)
  - `Km_medio`: Kilometraje entre revisiones
  - `Revisiones`: Frecuencia de mantenimiento

4.2. Resultados del Modelo

Cuadro 2: Métricas de Rendimiento

Métrica	**Valor**
MSE	0.0276
R^2	0.6774

Cuadro 3: Coeficientes de Regresión

Variable	**Coeficiente**
PVP	0.000008
avg_car_age	-0.070765
avg_km_revision	0.000016
avg_revisiones	0.118480

Interpretación técnica:

- red$R^2 = 0.677$: El modelo explica el 67.74 % de la varianza

- redMSE bajo (0.0276): Error cuadrático medio mínimo
- **Relaciones significativas**:
  - Antigüedad del coche ↑ ⇒ Churn ↑ (coef. negativo)
  - Revisiones frecuentes ↑ ⇒ Fidelidad ↑ (coef. positivo)

5. Análisis del Valor del Cliente (CLTV)

5.1. Relaciones Clave

- **Retención (R)**:
  - Probabilidad anual de que el cliente siga activo
  - Directamente proporcional al CLTV
  - R ¿0.7 indica clientes altamente leales
- **Margen (M)**:
  - Beneficio promedio anual por cliente
  - Impacta linealmente en el CLTV

5.2. Fórmula Básica

$$CLTV = \sum_{n=1}^{5} \frac{M \times R^n}{(1{,}07)^n}$$

5.3. Interpretación Práctica

- **Clientes de Alto Valor**:
  - Combinan R alta (mayor de 0.6) y M elevado
  - Priorizar retención sobre captación
- **Estrategias Óptimas**:
  - Inversión en fidelización 30 % del CLTV
  - Descuentos controlados para R menor de 0.5

Cuadro 4: Perfiles de Cliente

Tipo	**Rango CLTV**	**Acción**
Premium	Alto	Retención VIP
Estándar	Medio	Upselling
Riesgo	Bajo	Reactivación