Segmentación de clientes y predicción de Churn Valioso

Proyecto de analítica avanzada — Superstore

- Segmentación con K-Means para entender perfiles de clientes.
- Modelo de clasificación para anticipar el **churn valioso**.
- Enfoque orientado a decisiones de negocio, no solo a métricas técnicas.

Contexto y problema de negocio

- Retail tipo "Superstore" con miles de clientes que compran por web, catálogo y tienda física.
- Cada año, una parte de los clientes deja de comprar, pero **no todos tienen el mismo impacto en ingresos**.
- El reto no es solo saber quién se va, sino **quién se va y es realmente valioso para el negocio**.
- Necesidad: priorizar recursos comerciales (descuentos, llamadas, campañas) hacia los clientes cuyo abandono "duele" más.

Decisión clave: Definir "churn valioso" como métrica de negocio, no solo churn técnico. Esto alinea el proyecto con el impacto económico real.

Objetivo del proyecto

01

Definir cuantitativamente el concepto de Churn Valioso

Combinando inactividad + valor económico.

02

Segmentar la base de clientes

En **perfiles accionables** mediante K-Means.

03

Entrenar un modelo supervisado

Que entregue una **probabilidad de churn valioso por cliente**.

04

Diseñar una lógica de toma de decisiones

Qué hacer, con quién y en qué orden.

> **Pipeline de ML:** Este proyecto sigue un enfoque end-to-end desde la definición del problema hasta la implementación de reglas de decisión accionables.

Datos y variables utilizadas

Variables numéricas principales

- **Recency** (días desde la última compra).
- **MntTotal** y **TotalPurchases** (importe y frecuencia histórica).
- **CLV_log**: transformación logarítmica del valor de vida del cliente.
- **NumWebVisitsMonth**, **Perc_CatalogPurchases**, Income.

Variables categóricas y preparación

- Canal predominante
- Participación en campañas
- Estado civil, educación

Preparación: limpieza básica, creación de CLV_log, codificación de categorías para el modelo.

Base histórica de clientes con información demográfica y transaccional.

Decisión de feature engineering: CLV_log normaliza la distribución sesgada del valor del cliente, mejorando el desempeño de K-Means y el modelo supervisado.

EDA — Comportamiento de los clientes

Separación por Recency

Recency separa de forma clara clientes activos vs inactivos.

Distribución tipo Pareto

Pocos clientes concentran gran parte del gasto: patrón tipo Pareto en MntTotal y TotalPurchases.

Los clientes con CLV_log alto:

- compran más veces y gastan más,
- usan más de un canal (web, catálogo, tienda),
- suelen tener mayor interacción digital (más visitas web mensuales).

> Esta lectura confirma que no todos los clientes inactivos tienen el mismo peso para el negocio.

Insight del EDA: La heterogeneidad en valor justifica un enfoque de segmentación antes de modelar. No podemos tratar todos los churners igual.

Definición cuantitativa de Churn Valioso

Se aplica **K-Means sobre Recency** para encontrar grupos de inactividad sin imponer un umbral arbitrario.

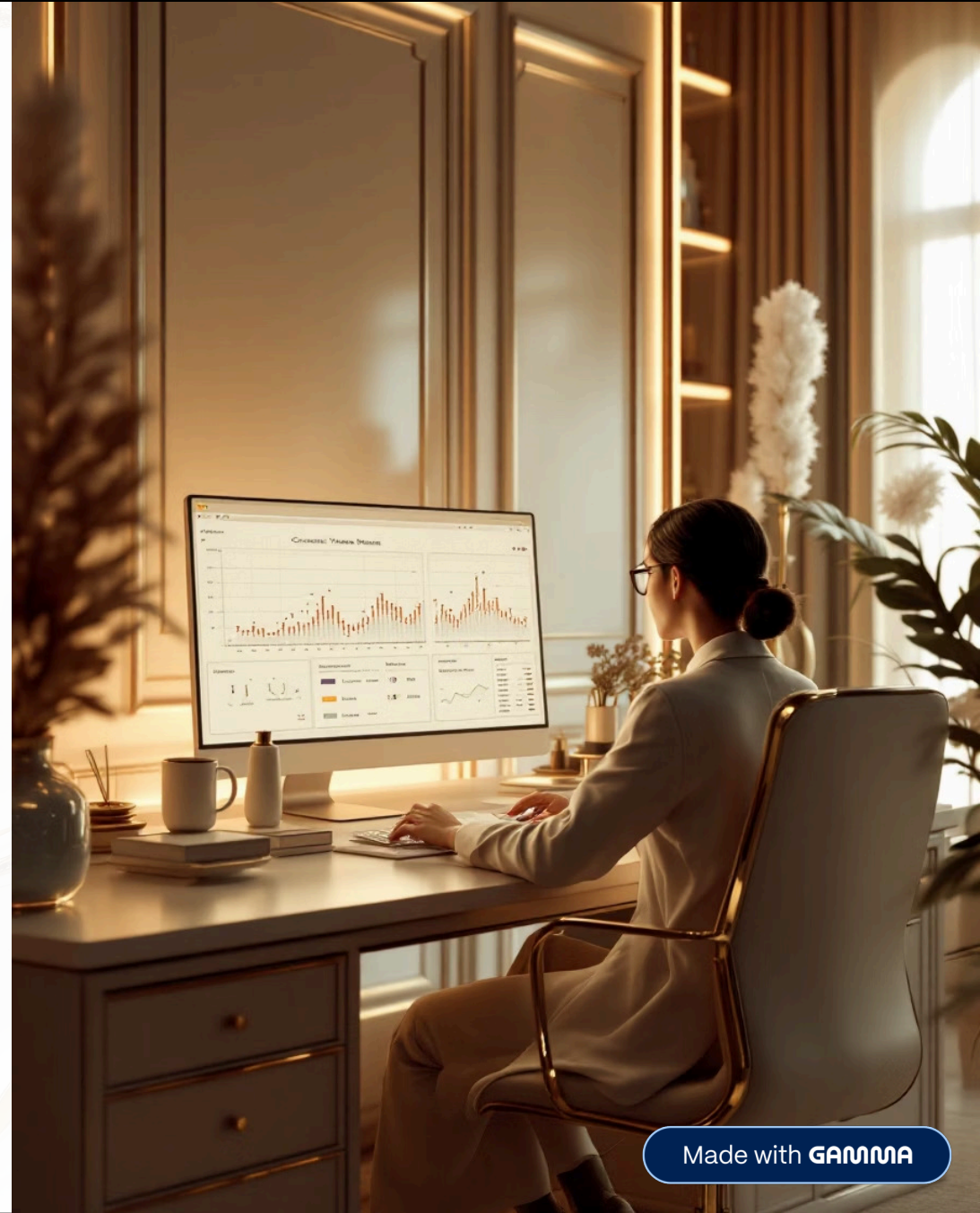
El modelo identifica un salto claro a partir de **Recency ≥ 83 días** → se considera **cliente inactivo**.

Se calcula la **mediana de CLV_log** para distinguir clientes de mayor valor económico.

Definición final de Churn Valioso = 1:

- Recency ≥ 83 días
- Y CLV_log ≥ mediana de CLV_log de la base.

Así, solo se marca como crítico a quien está inactivo **y** representa un valor relevante.



Segmentación de clientes con K-Means

Se entrena K-Means con variables de valor y comportamiento (Recency, CLV_log, MntTotal, TotalPurchases, uso web/catálogo, etc.).

Se obtienen 4 clusters con perfiles diferenciados:

Cluster 0 - Bajo valor/ Baja interacción

Gasto histórico min, pocas transacciones, baja interacción digital. CV:1.5%

Cluster 1 - Valor medio-Alto / Riesgo Emergente

Valor medio-alto, recency en aumento, uso mixto de canales. CV:14%

Cluster 2 - Alto valor / Alta inactividad

Alto valor histórico, alta intensidad de compra, pero señales claras de inactividad. Segmento critico: CV: 17%

Cluster 3 - Valor bajo / Aún activo

Bajo valor pero con comportamiento aún activo o poco histórico. Con potencial o compradores ocasionales. CV: <2%

Clusters 1 y 2 concentran la mayor parte del **Churn Valioso (~14–17%)**; clusters 0 y 3 aportan <2%.

Conclusión: el churn valioso **no es aleatorio**, se concentra en perfiles muy concretos.

Justificación del unsupervised learning: K-Means permite descubrir patrones naturales antes del modelo supervisado. Los clusters informan la estrategia de retención y validan que el churn valioso tiene estructura identificable.

Modelo supervisado de Churn Valioso

Configuración del modelo

- Problema: **clasificación binaria** (Churn Valioso = 1, No Churn Valioso = 0).
- Esquema: separación train/test **80/20**, estratificada por la etiqueta Churn Valioso.
- Modelo de tipo **ensamble de árboles**, adecuado para variables numéricas y categóricas mezcladas.

Métricas en test

0.9933

ROC-AUC

Excelente capacidad de separar churn valioso vs no churn.

0.9955

Accuracy

Muy pocos errores globales.

0.9722

Precision

Casi todos los marcados como churn valioso realmente lo son.

0.9722

Recall

El modelo captura casi todos los churners valiosos.

Resultado: tenemos un score de probabilidad fiable para usar en decisiones.

Elección del modelo: Ensambles de árboles (Random Forest/Gradient Boosting) manejan bien features mixtas, no requieren escalado, y capturan interacciones no lineales. La estratificación en train/test preserva la proporción de churn valioso (~10%), evitando sesgo en la evaluación.

Toma de decisiones basada en clusters y riesgo

Cada cliente queda con:

- Un **cluster** (perfil de valor/comportamiento).
- Una **probabilidad de Churn Valioso** (score entre 0 y 1).

Niveles de riesgo basados en el score del modelo:

Bajo riesgo

probabilidad < 0.30

Riesgo medio

0.30–0.60

Alto riesgo

> 0.60

Lógica de priorización (árbol de decisiones):

Prioridad 1

Clientes en clusters 1–2 (más valiosos) **y** alto riesgo → acciones intensivas (llamada, oferta personalizada, cupón fuerte).

Prioridad 2

Clusters 1–2 con riesgo medio → recordatorios, campañas segmentadas, beneficios moderados.

Prioridad 3

Otros clusters con riesgo alto → acciones más ligeras (email, campañas masivas).

> Esto permite **alinear el esfuerzo comercial** con el impacto esperado en CLV recuperado.

Conclusiones y siguientes pasos

Conclusiones

- La definición de Churn Valioso (Recency ≥ 83 + CLV_log ≥ mediana) se ajusta al impacto económico real, no solo a la inactividad.
- La combinación de **segmentación (clusters) + modelo de clasificación** genera un mapa claro de **quién es quién** en la base de clientes.
- El modelo ofrece un score de riesgo robusto que se traduce en una **regla de decisión concreta** para el equipo comercial.

Próximos pasos propuestos

1. Implementar estas reglas en un dashboard / app interna (por ejemplo, Streamlit).
2. Lanzar campañas diferenciadas por cluster y nivel de riesgo, y medir resultados con pruebas A/B.
3. Reentrenar y recalibrar el modelo con datos nuevos para mantener su desempeño en el tiempo.

Pipeline completo: Problema de negocio → EDA → Feature engineering → Unsupervised learning (segmentación) → Supervised learning (clasificación) → Reglas de decisión → Implementación y monitoreo continuo.